

HEALTH SYSTEMS DATA INTEROPERABILITY AND IMPLEMENTATION

by

MANDLENKOSI NGWENYA

Submitted in accordance with the requirements

for the degree of

MASTER OF SCIENCE

in the subject

COMPUTING

at the

UNIVERSITY OF SOUTH AFRICA

SUPERVISOR: PROF F O BANKOLE

FEBRUARY 2018

DECLARATION

1. I declare that “**Health Systems and Data Management** ” is my own work and that all the sources that I have used or quoted have been indicated and acknowledged by means of complete references.
2. This study has never been submitted to any other institution or organization before.
3. I have used the 6th Edition of APA (American Psychological Association) for citation and referencing.
4. I have not allowed and will never allow anyone to copy my work for any intention.

M. A. W Ngwenya

Student No: 43615554

January 2018

ACKNOWLEDGEMENTS

- I would like to thank and appreciate my supervisor Professor Felix Bankole for his continuous support, guidance and the patience he displayed throughout the progress of my study.
- I would like to thank my great friend Nomalungelo for the support and encouragement I received from her.
- My gratitude goes to the language editors Genevieve Wood, and Sandy Tolosana for their effort in editing my work.
- Lastly, I want to thank my God for the opportunity to study, the ideas and the passion He instilled in me for the field of science.

ABSTRACT

Objective

The objective of this study was to use machine learning and health standards to address the problem of clinical data interoperability across healthcare institutions. Addressing this problem has the potential to make clinical data comparable, searchable and exchangeable between healthcare providers.

Data sources

Structured and unstructured data has been used to conduct the experiments in this study. The data was collected from two disparate data sources namely MIMIC-III and NHanes. The MIMIC-III database stored data from two electronic health record systems which are CareVue and MetaVision. The data stored in these systems was not recorded with the same standards; therefore, it was not comparable because some values were conflicting, while one system would store an abbreviation of a clinical concept, the other would store the full concept name and some of the attributes contained missing information. These few issues that have been identified make this form of data a good candidate for this study. From the identified data sources, laboratory, physical examination, vital signs, and behavioural data were used for this study.

Methods

This research employed a CRISP-DM framework as a guideline for all the stages of data mining. Two sets of classification experiments were conducted, one for the classification of structured data, and the other for unstructured data. For the first experiment, Edit distance, TFIDF and JaroWinkler were used to calculate the similarity weights between two datasets, one coded with the LOINC terminology standard and another not coded. Similar sets of data were classified as matches while dissimilar sets were classified as non-matching. Then soundex indexing method was used to reduce the number of potential comparisons. Thereafter, three classification algorithms were trained and tested, and the performance of each was evaluated through the ROC curve. Alternatively the second experiment was aimed at extracting patient's smoking status information from a clinical corpus. A sequence-oriented classification algorithm called CRF was used for learning related concepts from the given clinical corpus.

Hence, word embedding, random indexing, and word shape features were used for understanding the meaning in the corpus.

Results

Having optimized all the model's parameters through the v-fold cross validation on a sampled training set of structured data ($m = 2483$), out of 24 features, only ($n = 8$) were selected for a classification task. RapidMiner was used to train and test all the classification algorithms. On the final run of classification process, the last contenders were SVM and the decision tree classifier. SVM yielded an accuracy of 92.5% when the C and γ parameters were set to $C = 200.0008$ and $\gamma = 0.0900811$. These results were obtained after more relevant features were identified, having observed that the classifiers were biased on the initial data. On the other side, unstructured data was annotated via the UIMA Ruta scripting language, then trained through the CRFSuite which comes with the CLAMP toolkit. The CRF classifier obtained an F-measure of 94.8% for "nonsmoker" class, 83.0% for "currentsmoker", and 65.7% for "pastsmoker". It was observed that as more relevant data was added, the performance of the classifier improved. The results show that there is a need for the use of FHIR resources for exchanging clinical data between healthcare institutions. FHIR is free, it uses: profiles to extend coding standards; RESTful API to exchange messages; and JSON, XML and turtle for representing messages. Data could be stored as JSON format on a NoSQL database such as CouchDB, which makes it available for further post extraction exploration.

Conclusion

This study has provided a method for learning a clinical coding standard by a computer algorithm, then applying that learned standard to unstandardized data so that unstandardized data could be easily exchangeable, comparable and searchable and ultimately achieve data interoperability. Even though this study was applied on a limited scale, in future, the study would explore the standardization of patient's long-lived data from multiple sources using the SHARPn open-sourced tools and data scaling platforms such Hadoop.

Table of Contents

DECLARATION	i
ACKNOWLEDGEMENTS.....	ii
ABSTRACT	iii
LIST OF FIGURES	viii
LIST OF TABLES.....	x
LIST OF ABBREVIATIONS AND ACRONYMS.....	xii
1. INTRODUCTION	1
1.1 BACKGROUND TO THIS STUDY	2
1.2 AN OVERVIEW OF DATA MANAGEMENT AND DATA PROCESSING METHODS.....	4
1.3 DEFICIENCIES IN PAST LITERATURE	5
1.4 THE SIGNIFICANCE OF THE STUDY	7
1.5 THE OBJECTIVE OF THE STUDY	8
1.5.1 RESEARCH QUESTIONS.....	9
1.5.2 HYPOTHESIS DEVELOPMENT	10
1.6 RESEARCH DESIGN	10
1.7 THE OUTLINE OF THE STUDY	13
2. LITERATURE REVIEW.....	14
2.1 INTRODUCTION.....	14
2.1.1 BACKGROUND AND FUTURE OF BIG DATA IN HEALTH CARE.....	14
2.2 THE CHARACTERISTICS OF BIG DATA.....	16
2.2.1 VOLUME	16
2.2.2 VARIETY	19
2.2.3 VELOCITY	22
2.2.4 VERACITY	23
2.2.5 THE RISKS OF BIG DATA	25
2.3 HEALTH SYSTEM APPLICATIONS AND THE INFLUENCE OF BIG DATA.....	26
2.3.1 HEALTH INFORMATION SYSTEMS	27
2.3.2 COLLECTION OF HEALTH BIG DATA THROUGH GIS LAYERS	29
2.3.3 M-HEALTH AND TELEHEALTH	31
2.4 STANDARDS FOR SYSTEMS INTEROPERABILITY	33
2.4.1 CODING STANDARDS	39
2.5 DATA PRE-PROCESSING.....	41

2.5.1	DATA CLEANING	42
2.6	DATA INTEGRATION	45
2.6.1	FEATURE SELECTION	45
2.6.2	SIMILARITY MEASURE	47
2.6.3	INDEXING TECHNIQUES	51
2.6.4	DATA SET MATCHING	52
2.6.5	TEXT CLASSIFICATION ALGORITHMS.....	55
2.6.6	STORAGE MECHANISMS FOR BIG HEALTH DATA.....	58
2.6.7	CONCLUSION	60
3.	RESEARCH METHODOLOGY	62
3.1	INTRODUCTION.....	62
3.2	THEORETICAL PERSPECTIVE TO THE PROPOSED SOLUTION	62
3.3	FORMULATION OF RESEARCH QUESTIONS	67
3.4	CRISP-DM FRAMEWORK AND DESIGN SCIENCE RESEARCH.....	67
3.5	DATA UNDERSTANDING.....	70
3.5.1	DATA SOURCES.....	72
3.5.2	DATA EXPLANATION.....	73
3.6	DATA PREPARATION	77
3.6.1	DATA PRE-PROCESSING FOR STRUCTURED DATA	77
3.6.2	DATA PRE-PROCESSING FOR UNSTRUCTURED DATA	80
3.7	NOTATION USED.....	82
3.8	CLASSIFIERS AND PROBABILISTIC GRAPHICAL MODELS USED.....	83
3.9	TOOLS AND DATABASES.....	93
3.10	ETHICAL CLEARANCE.....	96
3.11	CONCLUSION	96
4.	MODELLING.....	98
4.1	INTRODUCTION.....	98
4.2	FEATURE ENGINEERING AND SELECTION	99
4.3	FEATURE SELECTION FOR MATCHING SOURCE TO TARGET	108
4.4	ANNOTATING THE CLINICAL CORPORA.....	110
4.5	MODEL SELECTION AND OPTIMISATION.....	115
4.6	EXPERIMENTAL PROCEDURES	127
4.6.1	SYSTEMS SET UP	127
4.6.2	EVALUATION MEASURES	127
4.7	CONCLUSION	131

5. EVALUATIONS	132
5.1 INTRODUCTION.....	132
5.2 RESULTS FROM SIMILARITY MEASURES	133
5.2.1 MATCHING DISCUSSION	134
5.3 FIRST EXPERIMENT: STRUCTURED DATA	136
5.3.1 SUPPORT VECTOR MACHINES.....	136
5.3.2 MULTIPLE MODEL PERFORMANCES	138
5.4 SECOND EXPERIMENT: UNSTRUCTURED DATA	142
5.5 FIRST EXPERIMENT DISCUSSIONS	145
5.6 SECOND EXPERIMENT DISCUSSIONS	150
5.7 CONCLUSION	160
6. IMPLICATIONS OF THE FINDINGS.....	161
6.1 INTRODUCTION.....	161
6.2 IMPLICATION OF THE FINDINGS BASED ON CRISP-DM PROCESS	161
6.3 IMPLICATION OF THE FINDINGS	163
6.3.1 INTEROPERABILITY FOR STRUCTURED DATA	164
6.3.2 INTEROPERABILITY FOR UNSTRUCTURED DATA	165
6.4 LIMITATIONS, FUTURE AND ADVICE	167
6.5 CONCLUSION	170
REFERENCES.....	171
APPENDIX A-1: ETHICAL CONSENT LETTER FROM UNISA.....	195
APPENDIX A-2: REPORT TO AUTHORIZE THE USE OF MIMIC-III DATABASE FOR RESEARCH.....	197
APPENDIX B: PROCESS FLOW FOR ROC RESULTS COMPARISON BETWEEN SVM, DECISION TREES AND LOGISTIC REGRESSION.....	199
APPENDIX C: DECISION TREE, SPLITTING CRITERION EVALUATION	200
APPENDIX D: SCREENSHOT OF THE PROGRAM THE RESEARCHER WROTE FOR THE PURPOSE OF EXTRACTING SMOKING INFORMATION FROM A LARGE TEXT FILE.....	201
APPENDIX E: SETUP FILES AND RESULTS FROM EXPERIMENT 1 AND EXPERIMENT 2.....	202

Author keywords: Supervised Machine Learning Algorithms; Big Data; Healthcare Coding Standards; Record Linkage; Natural Language Processing; Health Informatics; Unstructured Information Management Architecture; Health Data Standardization; CRISP-DM, Structured and Unstructured data

LIST OF FIGURES

FIGURE 1. 1: CRISP-DM PROCESS FLOW (SOURCE: (OLSON & DELEN, 2008))	12
FIGURE 2. 1: DIMENSIONS OF BIG DATA (SOURCE: (FELDMAN ET AL., 2012))	16
FIGURE 2. 2: INTERNET OF THINGS PRODUCTS AND PROTOTYPES (SOURCE: ISLAM, KWAK, KABIR, HOSSAIN, & KWAK, 2015)	18
FIGURE 2. 3: UNSTRUCTURED DATA WORD CLOUD	21
FIGURE 2. 4: FACTORS AFFECTING DATA QUALITY (SOURCE: JERRY GAO ET AL., 2016)	25
FIGURE 2. 5: GEOGRAPHIC INFORMATION SYSTEM OF A HUMAN BEING (SOURCE: TOPOL, 2014)	29
FIGURE 2. 6: ITRIAGE MOBILE HEALTH APPLICATION (SOURCE: HTTP://HISTALKMOBILE.COM)	32
FIGURE 2. 7: LEADING STANDARDS EVALUATION MATRIX (SOURCE: (CSIR & NDOH, 2014))	35
FIGURE 2. 8: CROSS-ENTERPRISE DOCUMENT SHARING ARCHITECTURE AND DATA FLOW SOURCE: (NOUMEIR, 2011)	38
FIGURE 2. 9: FHIR OBSERVATION EXAMPLE FOR REPRESENTING PATIENT’S RESPIRATORY RATE USING FHIR RESOURCES (SOURCE: FHIR, 2011)	39
FIGURE 2. 10: A UNIFIED VIEW OF A FEATURE SELECTION PROCESS (SOURCE: LIU, MOTODA, SETIONO & ZHAO, 2010)	46
FIGURE 2. 11: TRAINING A SUPERVISED ALGORITHM	55
FIGURE 2. 12: EXAMPLE OF TRAINING EXAMPLES FOR DETERMINING WHETHER TO GRANT LOAN TO THE APPLICANT OR NOT (SOURCE: (GORUNESCU, 2011))	56
FIGURE 2. 13: RULES THAT USE CONJUNCTIONS AND DISJUNCTIONS TO DETERMINE WHETHER TWO RECORDS MATCH OR NOT (SOURCE: (CHRISTEN, 2012))	57
FIGURE 3. 1: HYPOTHESIS EVALUATION PROCESS	63
FIGURE 3. 2: HIGH VARIANCE AND HIGH BIAS (SOURCE: (HAMEL, 2009))	65
FIGURE 3. 3: CRISP-DM FOR DATA UNDERSTANDING AND DATA PREPARATION (SOURCE: (OLSON & DELEN, 2008))	70
FIGURE 3. 4: SIGMOID FUNCTION OR LOGISTIC FUNCTION	84
FIGURE 3. 5: SVM DECISION BOUNDARIES (SOURCE: (NASIEN ET AL., 2010))	86
FIGURE 3. 6: MAPPING INPUT SPACE TO FEATURE SPACE (SOURCE:(HOFMANN, 2006))	87
FIGURE 3. 7: RISK PREDICTION BASED ON THE TYPE OF CAR AND THE DRIVER’S AGE (SOURCE: (GORUNESCU, 2011))	89
FIGURE 3. 8: HIDDEN MARKOV MODEL GRAPH FOR ESTIMATING THE ATMOSPHERIC PRESSURE (SOURCE: (KOLLER & FRIEDMAN, 2009))	91
FIGURE 3. 9: A LINEAR CHAIN CRF MODEL SHOWING OBSERVABLE STATES DENOTED SHOWN AS GREY NODES, AND HIDDEN STATES SHOWN AS CLEAR NODES (SOURCE: (KOLLER & FRIEDMAN, 2009))	93
FIGURE 3. 10: CTAKES PROCESSING OF A CLINICAL TEXT DOCUMENT (SOURCE: SAVOVA ET AL. (2010))	95
FIGURE 4. 1: CRISP-DM FRAMEWORK FOR MODEL SELECTION (SOURCE: (OLSON & DELEN, 2008))	99
FIGURE 4. 2: INITIAL RULES FOR DETERMINING IF TWO RECORDS MATCHES OR NOT.	109
FIGURE 4. 3: A SAMPLE RULE FOR DETECTING A SMOKING STATUS OF A PATIENT FOR A GIVEN CLINICAL NOTE.	113
FIGURE 4. 4: FEATURE EXTRACTION FROM CLINICAL TEXT USING NAMED ENTITY RECOGNITION	115
FIGURE 4. 5: FEATURE EXTRACTION FROM CLINICAL TEXT USING NAMED ENTITY RECOGNITION	115

FIGURE 4. 6: NUMBER OF TRAINING EXAMPLES FOR DATASETS	116
FIGURE 4. 7: TESTED MODEL WITHOUT CROSS-VALIDATION	117
FIGURE 4. 8: PREDICTED MODEL WITHOUT CROSS-VALIDATION	118
FIGURE 4. 9: PICTORIAL VIEW OF A 10-FOLD CROSS-VALIDATION (SOURCE: (D. L. OLSON & DELEN, 2008))	119
FIGURE 4. 10: ANNOTATION PROCESS (SOURCE: (PUSTEJOVSKY & STUBBS, 2013))	123
FIGURE 4. 11: SHOWS THE USAGE RESULTS OF TWO SETS OF NGRAMS BETWEEN THE YEAR 1800 AND 2008, THIS IS A COMPARISON BETWEEN 3-GRAM WHICH IS (DOES NOT SMOKE) AND A 2-GRAM (NEVER SMOKED).	124
FIGURE 4. 12: WORD CLUSTERED BASED ON CONTEXT AND RELATEDNESS FROM AN INPUT OF 260 741-WORD VOCABULARY (SOURCE: (BROWN ET AL., 1992))	125
FIGURE 4. 13: ROC CURVE WITH MULTIPLE CLASSIFIERS (SOURCE: (OLSON & DELEN, 2008))	131
FIGURE 5. 1: CRISP-DM FRAMEWORK FOR MODEL EVALUATION (SOURCE: (OLSON & DELEN, 2008))	133
FIGURE 5. 2A: RESULTS FROM RUNNING A 10-FOLD CROSS-VALIDATION AND GRID-SEARCH FOR PARAMETER OPTIMISATION OF SVM	137
FIGURE 5. 2B: RESULTS FROM RUNNING A 10-FOLD CROSS-VALIDATION AND GRID-SEARCH FOR PARAMETER OPTIMISATION OF LOGISTIC REGRESSION	140
FIGURE 5. 2C: RESULTS FROM RUNNING A 10-FOLD CROSS-VALIDATION AND GRID-SEARCH FOR PARAMETER OPTIMISATION OF DECISION TREE	141
FIGURE 5. 3: ROC CURVE FOR DECISION TREE, LOGISTIC REGRESSION AND SVM	142
FIGURE 5. 4: SVM MODEL CLASSIFICATION ERROR	147
FIGURE 5. 5 A: ANNOTATIONS FROM CLAMP'S PREDEFINED RULES	151
FIGURE 5. 5 B: ANNOTATIONS FROM THE CUSTOM DEVELOPED RULES	151
FIGURE 5. 6: A DOUBLE CLASS ANNOTATION WHERE A RECORD IS CLASSIFIED AS "CURRENTSMOKER" AND "NONSMOKER" AT THE SAME TIME.	151
FIGURE 5. 7: PREDICTED NAMED ENTITIES FROM THE CRF CLASSIFIER	156
FIGURE 6. 1: CRISP-DM PROCESS FLOW (SOURCE: (OLSON & DELEN, 2008))	162

LIST OF TABLES

TABLE 1. 1: RESEARCH SUB-QUESTIONS FOR THIS RESEARCH STUDY.....	9
TABLE 2. 1A: DATA SET FROM PROVIDER A.....	19
TABLE 2. 1B: DATA SET FROM PROVIDER B.....	19
TABLE 2. 2: MAPPING FUNCTIONS TO PROFILES AND STANDARDS.....	37
TABLE 2. 3 MAPPING FUNCTIONS TO PROFILES AND STANDARDS.....	40
TABLE 2. 4A WORD STEMMING FOR SOURCE DATA.....	50
TABLE 2. 4B TARGET VITAL SIGNS FEATURES.....	50
TABLE 2. 4C TARGET VITAL SIGNS FEATURES.....	51
TABLE 2. 5 MAPPING LEGACY DATA SET AND ATTRIBUTES TO FHIR RESOURCE.....	53
TABLE 3. 1 VARIABLES OF THE STUDY.....	66
TABLE 3. 2 RESULTS OF RECORDS TO BE COMPARED WITH BLOOD PRESSURE RECORD.....	66
TABLE 3. 3 SIMILARITIES BETWEEN DSRM AND CRISP-DM.....	68
TABLE 3. 4A STANDARDIZED TESTS FROM LABEVENTS AND D_LABITEMS MIMIC TABLES.....	71
TABLE 3. 4B UNSTANDARDIZED OBSERVATIONS FROM CHARTEVENTS AND D_ITEMS MIMIC TABLES.....	71
TABLE 3. 5 MECHANISMS FOR MINING ABBREVIATION EXPANSIONS.....	78
TABLE 3. 6 EXAMPLE ABOUT RELATION EXTRACTION TO SHOWCASE THE SHORTCOMINGS OF HMM.....	91
TABLE 4. 1 MIMIC-III SOURCE OBSERVATION DATASET.....	100
TABLE 4. 2A SOURCE FEATURES WITH SOUNDEX BLOCKING KEYS.....	101
TABLE 4. 2B TARGET FEATURES WITH SOUNDEX BLOCKING KEYS.....	103
TABLE 4. 3 BLOCKING KEY VALUES.....	103
TABLE 4. 4 OBSERVATIONS WITH SOUNDS THAT DIFFER FROM LOINC OBSERVATION.....	104
TABLE 4. 5 CLINICAL TEXT ABOUT PATIENT’S SMOKING INFORMATION AND THE MEANING.....	106
TABLE 4. 6 ATTRIBUTES SIMILARITY COMPARISON.....	108
TABLE 4. 7 SMOKING STATUS EXAMPLES.....	111
TABLE 4. 8 MODEL SELECTION CRITERIA.....	120
TABLE 4. 9 EVALUATION METRICS FOR THE CLASSIFIER.....	128
TABLE 5. 1A EDIT DISTANCE SIMILARITY RESULTS FOR OBSERVATION NAME AND UNIT OF MEASURE.....	133
TABLE 5. 1B JARO-WINKLER SIMILARITY RESULTS FOR OBSERVATION NAME AND UNIT OF MEASURE.....	134
TABLE 5. 2A CONFUSION MATRIX FOR SUPPORT VECTOR MACHINES.....	136
TABLE 5. 2B CONFUSION MATRIX FOR A LOGISTIC REGRESSION CLASSIFIER.....	138
TABLE 5. 2C CONFUSION MATRIX FOR A DECISION TREE CLASSIFIER.....	140
TABLE 5. 3 RESULTS FOR SMOKING STATUS DETECTION PRODUCED BY A CRF SEQUENCE CLASSIFIER.....	143
TABLE 5. 4A CONFUSION MATRIX FOR A DECISION TREE CLASSIFIER WITH ACCURACY: 91.00% +/- 5.83% (MIKRO: 91.00%).....	149

TABLE 5. 4B CONFUSION MATRIX FOR THE SVM CLASSIFIER WITH ACCURACY: 92.50% +/- 5.12% (MIKRO: 92.50%)	149
TABLE 5. 5 AN EARLIER TEST RESULTS FOR THE NAMED ENTITY EXTRACTION FOR THE PATIENT'S SMOKING STATUS AND OTHER RELEVANT INFORMATION THROUGH A CRF SEQUENCE CLASSIFIER.....	154
TABLE 5. 6 LATER TEST RESULTS FOR THE NAMED ENTITY EXTRACTION FOR THE PATIENT'S SMOKING STATUS AND OTHER RELEVANT INFORMATION THROUGH A CRF SEQUENCE CLASSIFIER	155
TABLE 5. 7 RESULTS FROM EXECUTING RULES ON CLINICAL TEXT DATA	157
TABLE 5. 8 SNOMED-CT CODING INFORMATION ACCORDING TO THE UMLS METATHESAURUS	157

LIST OF ABBREVIATIONS AND ACRONYMS

CDA	Clinical Document Architecture
CMS	Council of Medical Schemes
CPT	Current Procedural Terminology
CRISP-DM	Cross-Industry Standard Process for Data Mining
CSIR	Council for Scientific and Industrial Research
EHR	Electronic Health Record
EMR	Electronic Medical Record
HIS	Health Information Systems
HL7	Health Level Seven Standard
HNSF	Health Normative Standards Framework
ICD	International Classification of Diseases
LOINC	Logic Observation Identifiers Names and Codes
MIMIC-III	Medical Information Mart for Intensive Care
m-Health	Mobile Health
NDoH	National Department of Health
NDP	National Development Plan
NHANES	National Health and Nutritional Examination Survey
NHI	National Health Insurance
PHR	Personal Health Record
REST	Representative State Transfer
SDK	Software Development Kit
SNOMED	Systematic Nomenclature of Medicine Clinical Terms
SVM	Support Vector Machines
TFIDF	Term Frequency Inverse Document Frequency
UMLS	Unified Medical Language Systems
UNISA	University of South Africa

CHAPTER 1:

Introduction

1. INTRODUCTION

Health care facilities in South Africa still find it difficult to share, trace and efficiently search for patients' medical data on their health information systems. According to (Masilela, Foster, & Chetty, 2013) health information systems are characterised by fragmentation and a lack of coordination. The report further adds that there is prevalence of manual systems and the lack of automation in health care, and between those systems that have been automated, there is a lack of interoperability.

(Mxoli, Mostert-Phipps, & Gerber, 2014) have defined interoperability in health care systems as the ability of information and communication technology (ICT) systems to share and exchange patients' health data. In health care, standardization concepts have been considered to be the potential solution to the fragmented and *siloed* health systems (Smith, Fridsma, & Johns, 2014). Data management standards have enabled seamless exchange of information and have reduced the complexity when sharing data between multiple systems (Adebesin, Kotzé, Greunen, & Foster, 2013; Gruenheid, Dong, & Srivastava, 2014; Nagy, Preckova, Seidl, & Zvarova, 2010).

Even though there are standards in place designed to ensure consistency and interoperability between systems, the adoption rate in South Africa remains low. This has been attributed to the lack of human resources for implementing the standards, lack of implementation guidelines, a limited participation in standards development, and a lack of standards' development prioritisation (Adebesin, Kotzé, et al., 2013). Another problem is that standards evolve and change over time, for instance HL7 health Version 2 standard organises data in a "comma separated value" file system, while Version 3 uses a complex XML file format. FHIR is the latest version of HL7 standards, it is resource-based and organizes information in XML, JSON, and turtle syntax. (Smits & Cornet, 2014) in their findings have reported FHIR to be completely different and not compatible with the previous versions of HL7 standard. Therefore, the researcher

claims that as the standards evolve, the health systems implementing those standards would need to adapt to that change. Now the problem comes when the data in system A is not easily retrievable, or comparable, or exchangeable with system B. Therefore, this study addresses the problems mentioned through health standards and machine learning.

As such, this research study addresses the data interoperability problem that is currently experienced by the health care industry in both developing and developed countries. In the United States of America, they introduced the Meaningful Use programme, aimed at improving quality, safety, and the efficiency of Electronic Health Records (EHR) systems, and thus reducing health disparities (D'Amore et al., 2014). Here in South Africa, the National Department of Health (NDoH) has introduced the Health Normative Standards Framework (HNSF), which is an interoperability guideline that provides guidance for eHealth standards implementation between information technology systems (CSIR & NDoH, 2014). These are some of the items that this study aims to address. Below is an overview of the current chapter.

In section 1.1 the researcher gives the background of this study, and a brief detail about data management is covered in section 1.2. Deficiencies in past literature and the significance of the study is then covered in section 1.3 and 1.4 respectively. While the objective and research questions are detailed in section 1.5 and in its subsections. Section 1.6 gives an outline and the proposed research methods for this study, and later in this section, the limitations of the study are identified, proposed tools and instruments are mentioned, and the validity of the instruments and data analysis methods are described. Section 1.7 gives a summary of what lies beyond this chapter.

1.1 BACKGROUND TO THIS STUDY

It could be said with great confidence that standards are put in place to format and organise data, regardless of the industry. In health care as well, standards can be used to achieve data exchange (Gay & Leijdekkers, 2015), however, health data comes in many forms. Some of the data is produced from wearable devices, and does not follow

a certain health standard, yet wearable device data is considered to be a treasure trove when it comes to health care (Topol, 2015). If it can be possible to integrate this data to the clinicians' or hospitals' health systems, then it can be possible to achieve high quality health care, due to the availability of useful data that is passively generated. Patients are able to generate their own data from their devices (smartphones and sensors), and are taking advantage of m-Health applications to improve and assist their health, said (Paschou, Sakkopoulos, Sourla, & Tsakalidis, 2012).

According to (Swan, 2012), data from patients' devices can be treated as personalised preventative medicine and can be used to prevent, diagnose and treat diseases. Personalised preventive medicine does not only focus on disease management, but has the following advantages: reduction of patients' hospital readmission rates; extension of the patients' lifespan and reduction of disability; and also prevent conditions from rising.

Patients with chronic diseases are constantly required to monitor their health, and some use their smartphones, while others, especially diabetics, use a glucose tracker device to monitor their health. Data produced from these devices cannot be easily combined with data at the clinician's office. Data from wearable sensors is said to be heterogeneous, unstructured, and noisy (Chen, Mao, Zhang, & Leung, 2014), and as a result, it is difficult to integrate, and is costly to manage and exchange.

To make the data interoperability picture clearer, in the Eastern Cape, the South African Society of Cardio-vascular Intervention has observed that different doctors are not able to share their medical notes. As a result, they don't know the history of the patients' treatments and often during consultations, patients would be requested to do lab scans, lab tests, and be prescribed to medicine that another doctor previously prescribed but that did not work (The Competition Commission South Africa, 2016). The current health data management system is costly and inefficient. Hence, this study is targeted at collecting patient's data from multiple data sources, then classifying it based on health standards, such that patients' information can be easily searchable, shareable and comparable for patterns.

1.2 AN OVERVIEW OF DATA MANAGEMENT AND DATA PROCESSING METHODS

The South African health information systems policy states that information that can be gathered in health facilities includes the following:

- **Health status information:** which includes morbidity, mortality, births, deaths, injuries and disease burden;
- **Health related information:** which includes demographic, social economic, residential and other related information;
- **Health service information:** which is about utilisation of health services; and
- **Health management information:** which is about the administrative services.

Even though the policy clearly categorises the types of information in health facilities as listed above, the type of information collected in the private health sector is not similar to that collected in the public sector (Matshidze & Hanmer, 2007). In the private sector, the Council of Medical Schemes (CMS) has developed a minimum data set that stipulates which information the medical aid scheme ought to collect.

There is also disparity in health services between health facilities in rural and in urban areas. (Coleman, Herselman, & Potass, 2012) have found that in urban areas, internet connection is much faster and more reliable than it is in rural areas, even though the ICT infrastructure and systems are not integrated. Furthermore, (Coleman et al., 2012) have also stated that the PAAB system is used to collect and send patient demographic information to the head office of the North-West Health Department on a monthly basis. Urban hospitals in Rustenburg and Klerksdorp are able to share x-ray images electronically.

The ability of certain hospitals to share data amongst themselves does not remove the interoperability issues. The National Health Insurance (NHI) Plan aims to achieve interoperability between health systems by implementing the health information exchange middleware, while clinically-generated data will be shared and exchanged

using the middleware (South Africa Department of Health, 2015). However, wearable device data, such as heart rate, blood pressure, glucose measurements, sleep patterns, activity measurements, and so forth cannot be easily integrated into the national patient-based information system because at the moment there is no standard that stipulates how wearable data should be stored (Li et al., 2017). If data stored or shared uses a similar standard, then it would be similar in structure, thus making it easier to manage.

The health care industry is flooded with both structured and unstructured data, and when structured data is shared between health care organizations, the original data ends up being semi-structured, due to the lack of standardization. In health care, unstructured data comes in the form of medical reports, medical scans, doctors' notes, and more. (Sarawagi, 2007) suggests that structure could be given to unstructured data through information extraction methods. Extraction methods include: rule-based learning and statistical methods. A number of statistical models have been used to assign labels to tokens in a sentence. Sarawagi also stated that Support Vector Machines (SVM) have been used for classifying each token to an entity type, e.g. a person's name would be classified to a "person" entity, depending on a list of available entities. Classification helps with the task of choosing the correct target class for a given input.

SVMs are not only useful for classifying sentences into entities, but other researchers such as (Cheng, Zhang, Xie, Agrawal, & Choudhary, 2012; Zhao, Wang, Bi, Gong, & Zhao, 2011) have used SVM classifiers for classifying hierarchical data, such as web pages and xml documents. Therefore, the researcher proposes the use of machine learning algorithms and data coding standards for achieving data interoperability across manifold datasets that are not standardized, or that are fragmented.

1.3 DEFICIENCIES IN PAST LITERATURE

Fragmented and disparate health care systems in South Africa can achieve interoperability through the standardization of health care systems (Adebesin, Foster, Kotzé, & Van Greunen, 2013; Orgun & Vu, 2006). To ensure that health standards are

developed in health care, the National Department of Health (NDoH) has commissioned the Meraka Institute of CSIR to develop a Health Normative standards framework (Masilela et al., 2013). The framework will provide guidance and the know-how of the eHealth standards to the Health Department. However, existing systems would have to comply with the framework, which means redevelopment of these systems, and begs the question as to how the old data before the enrolment of the standard ought to be standardized.

The old data would have to be captured, or exported, and be structured based on the standard that was implemented, if data is sourced from different providers; which use different standards, then the problem of interoperability resurfaces. For example, (Ding, Yang, & Wu, 2011) have stated that different sources of health data, such as data from a wearable sensor, can have different semantics and data structures, which increase the difficulties in data processing. Hence, previous literature on data management shows that the focus has been on implementing a standard at systems-level and not at data-level. The standardization of structured data solves a fraction of the interoperability problem, however as it has been stated, 80% of organizations' data is unstructured (Barrett, Humblet, Hiatt, & Adler, 2013). Unstructured data is also dormant in health care. In a hospital setting, vital clinical information is recorded in a human-readable language such as English. Recording the information in a human readable language makes it easier and faster for clinical personnel to record into the EHR (Electronic Health Record) than to record the data in a structured format. Unstructured data is often easier to read by humans but it is much more difficult to manage via computers (Barbulescu et al., 2013; Rosenbloom et al., 2010). Even in such a case, the volume of this data is overwhelming for clinicians to manage manually, and to organize via computers. Therefore one would have to apply Natural Language Processing (NLP) algorithms in order to easily manage this data. Therefore, there is a need to also standardize unstructured data.

Another method of standardizing data is through the use of SDKs. APPLE provides an SDK HealthKit to third party devices and application developers. The SDK is aimed at

making patients' data collatable and shareable between applications and devices. Nevertheless, in order to use the SDK, developers or data users must own an iPhone smartphone, and they would also have to redevelop their applications using Apple's SDK HealthKit, which runs on the Macintosh operating system (Hattersley, 2014).

Other data management techniques include Data Fusion, which has been used to combine multiple data sources in order to ensure data management. Another technique is using record-linkage algorithms that are aimed at finding attributes that are shared between data sources, where they can be used to match records across different sources. (Hassanzadeh et al., 2013) used a record-linkage algorithm to create a framework for discovering linkage points over large semi-structured web data. This framework was only focused on web data sources, and they saw a need to extend their framework to accommodate syntactic, semantic, and lexical matching functions. Other researchers such as (Viangteeravat et al., 2011) have presented a prototype for the implementation of HL7 Reference Information Model mapping for data integration of distributed clinical data sources. These researchers have recognised the need for an automatic mapping service that uses semantic mapping, pattern matching and machine learning techniques for mapping traditional health data to an appropriate RIM-based classes and attributes.

1.4 THE SIGNIFICANCE OF THE STUDY

The study of the integration and management of patients' data in South Africa is of paramount importance, because without relevant data, it is impossible to make correct decisions. It is mentioned by (Mayosi et al., 2012) that, "detection, management, and outcomes of care for individuals with non-communicable diseases are suboptimum" (p. 10). With the introduction of the National Health Insurance (NHI) plan, there is a need to standardize clinical data so that data can be easily sharable between health care institutions. Standardizing data ensures that common reimbursement codes are used for the clinical services being provided, and hence preventing fraud by overcharging the services provided in health care.

More relevant data is needed to understand the patient so that decision-making in health care can be improved through the use of integrated patient information. If the data is integrated and is easily retrievable, then it can be easy to extract useful information (Hinssen, 2012). Integrated patients' data would allow the physician to search for treatments that worked for a similar patient to the one being treated; and data can be filtered by age, gender or any other relevant characteristics (Barbarito et al., 2015). However, currently health data is not integrated, and a large portion of it cannot be used for secondary purposes because it is not structured in the same way and is stored in different locations (Rea et al., 2012).

Data is difficult to manage manually, which is why this study proposes the use of computer algorithms to organise patients' data. In addition, (Fu, Christen, & Boot, 2010) suggest that linked information facilitates improved retrieval of information, and it also improves the quality of the data, which in return offers more value and opportunities in data usage for further analysis. (Porter & Lee, 2013) have suggested that in order to enable universal comparison of health outcomes, and for stimulating improvements in health care, it is vital to measure outcomes by conditions, and the researcher suggests that standardizing health data would improve how outcomes are measured in health care.

1.5 THE OBJECTIVE OF THE STUDY

The objective of this study is to use standardized clinical observation data as input on a learning algorithm, where the algorithm would learn a function f for identifying patterns in the input data, such that when the algorithm is given new but related unstandardized observation data, it would be able to classify the data to the related standard. The researcher has planned to use the SVM classifier as the learning algorithm, and laboratory data that is standardized (also known as gold standard), based on the LOINC standard. This objective is meant to address the problem of data interoperability, by ensuring that clinical observation data is searchable, comparable and exchangeable between health care facilities. Clinical observation data includes but is not limited to vital signs, laboratory data, and social history such as tobacco use.

1.5.1 RESEARCH QUESTIONS

The following research questions have been identified, and are aimed at addressing clinical observation data interoperability across health care facilities. The researcher has identified two main questions for this study, and sub-questions are extracted from these main questions:

- 1. When will health information systems in South Africa be standardized in order to be able to seamlessly exchange and share consolidated patients' data?**
- 2. How can the process of data compliance across health care providers be automated through machine learning concepts?**

In order for these questions to be answered, the following sub-questions have been identified as mentioned in Table 1.1.

#	Research sub-question
i.	What type of health-related data sets will this research study focus on?
ii.	What methods are being used to classify objects accordingly in other industries, and how can those methods be applied in healthcare in order to achieve semantic and syntactic interoperability?
iii.	How were features selected for structured data?
iv.	How were features selected for unstructured data
v.	What methods are used to automatically map source dataset (unstandardized) to the target dataset (standardized) with high level of accuracy? And which one is appropriate for health-related data?
vi.	What features will be used to determine similarity between two records?
vii.	How will the correctness of the results be evaluated?

Table 1. 1: Research sub-questions for this research study

1.5.2 HYPOTHESIS DEVELOPMENT

From accomplishing the objective of the study as stated in section 1.5, the researcher claims interoperability will be achieved, therefore the following hypothesis statements have been developed:

- Patients' data is not easily exchangeable, searchable and comparable because it's not structured; therefore, in order to give it a structure, one must apply a working standard, and to automate the process of data standardization one can use a learning algorithm.
- Support Vector Machine algorithm can learn better than logistics regression and Decision Tree algorithms because they are sensitive to outliers, and it maximizes the margin that separates the positive and negative training examples.

On the discussion section of this study, the researcher proves the hypothesis he has developed.

1.6 RESEARCH DESIGN

The output of this research study is evaluated through a design science research (DSR) approach, whereby a model prototype is developed in order to test if the learning algorithm is able to make correct predictions on unknown data. The DSR approach helps design research experiments that can be reproduced by other researchers. In addition, the researcher uses Knowledge Discovery and Data (KDD) mining process models as a guideline for implementing data mining projects. Few of the KDD process models are: Sample Explore Modify Model Assess (SEMMA), Cross-Industry Standard Process for Data Mining (CRISP-DM), and Integrated Knowledge Discovery and Data Mining (IKDDM). SEMMA was developed by SAS, and it uses an iterative experimental cycle of five steps which makes up its name. The SEMMA data mining process is as follows: firstly, the data is *sampled* where training set, cross-validation set and test set are selected and partitioned; selected data is then *explored* for anomalies and outliers; thereafter *modified* through the identification of additional features and removal of redundant features; then the *model* is built by using modelling techniques such Decision Trees, Support Vector Machines (SVM), and more; then lastly, the selected model is

assessed in order to predict its performance on test data (Olson & Delen, 2008). Alternatively, CRISP-DM and IKDDM consists of six phases namely: business understanding, data understanding, data preparation, modelling, evaluation, and deployment (Rivo et al., 2012) see Figure 1.1 for a process flow.

The researcher starts by defining the phases in relation to CRISP-DM. The business understanding phase is meant to assess the need, significance and the objective of a DM and KD project. From section 1.1 through to 1.5, the researcher provides the business understanding for this research study. The second step of the CRISP-MD methodology is data understanding, which includes the process of data collection, data defining, data review and exploration, and the verification of the authenticity of the data. The third step is data preparation, where, during this step, the collected sampled dataset is cleaned of redundant data values, missing values are filled, and outliers are identified and fixed. Part of data preparation includes data normalization, indexing, attribute and record comparisons, feature selection, feature preparation, and feature weighing and vectorisation.

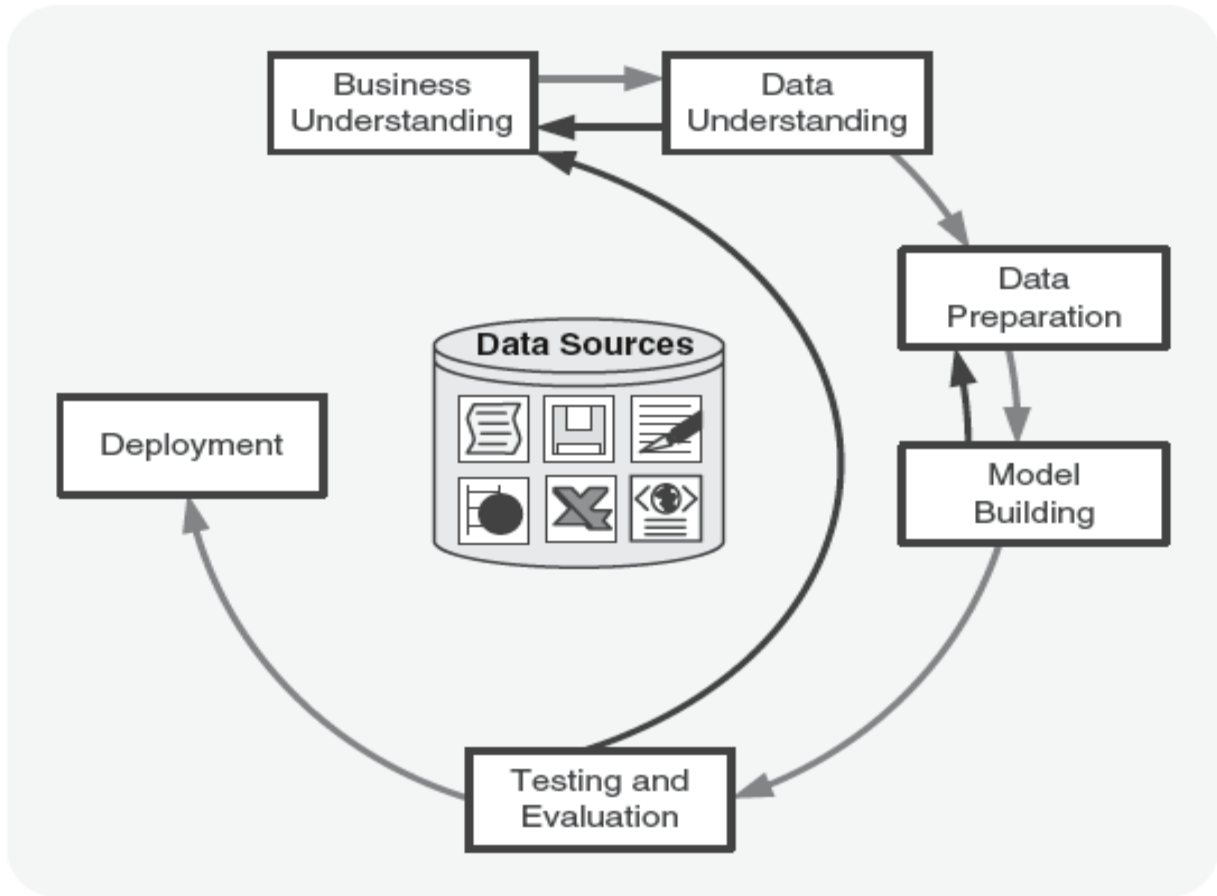


Figure 1. 1: CRISP-DM process flow (Source: (Olson & Delen, 2008))

The fourth step of CRISP-DM is data modelling, where the data is loaded into software such as RapidMiner, MATLAB, Octave, or R for visual exploration of the data points. During this step, the data is partitioned into three sets, namely training data, cross-validation data and the testing data. Thereafter the data mining technique is identified, where classification algorithms such as Decision Trees, SVM, and Logistic Regression are identified. The data is then evaluated during the fifth step of CRISP-DM, where the evaluation is based on recall, precision and accuracy. Finally, the last step is deployment, which involves applying the results of the learned model on a live system, and observing the performance (Olson & Delen, 2008). Similarly, IKDDM also defines the same phases as CRISP-DM. IKDDM is an integrated version of CRISP-DM whereby links are formed between tasks within a phase and between phases. The IKDDM

approach has been reported to provide an efficient and an effective implementation of DM and KD processes. Furthermore, IKDDM was designed to address the fragmented approach of CRISP-DM (Mansingh, Osei-Bryson, & Asnani, 2016; Sharma & Osei-Bryson, 2010). However, with all the features that IKDDM offers, the researcher has struggled to find documentation on the uses of IKDDM. Therefore, the researcher has considered the use of CRISP-DM since its documentation is easily accessible, even (Krzysztof Cios, Witold Pedrycz, Roman Swiniarski, & Lukasz Kurgan, 2007) declared that its documentation is good and easy to follow. In addition (Kurgan & Musilek, 2006) have reported that CRISP-DM can be used by novice data miners, it is suitable for industrial projects, and has been regarded as a successful and extensively applied framework in multiple industries. With the CRISP-DM base set, in Chapter Three, the researcher shows the relationship between the DSR approach and CRISP-DM.

1.7 THE OUTLINE OF THE STUDY

Chapter Two presents a review of literature, and is aimed at giving a summary of the studies consulted when conducting this study. During this chapter, the researcher will define the components of this study such as health data, health care data standards, and data mining concepts. Chapter Three will provide research methodologies, approaches and strategies.

Chapter Four will analyse the data collected in this study. Chapter Five will provide the results and discuss the findings, and finally Chapter Six will present the applicability, impact and the implication of the findings.

CHAPTER 2: Health Data, Coding Standards and Data Integration Techniques

2. LITERATURE REVIEW

2.1 INTRODUCTION

This chapter builds the foundation for this study by establishing the attributes that makes up big data, and discussing how to manipulate this data using computer algorithms and international standards, for health data. The domain of this study is health care, specifically data about clinical observations, which include(s) vital signs, laboratory data, device measurements, and social history, such as tobacco usage. The researcher starts section 2.2 by describing the characteristics of big data. Section 2.3 is focused on big data applications, information systems in health care, data sources from which to collect data from, and lastly, mobile health care delivery systems, are discussed in section 2.3.3.

Section 2.4 discusses the health standards to be used for this study; standards ensure interoperability between disparate systems in different health care facilities. Section 2.5 describes the details about data cleaning, for structured, and semi-structured data. In section 2.6 the researcher provides methods for preparing the collected data sets so that machines are able to read the contents of this data, thereafter, also in this section schema and attribute mapping methods are covered. Data storage, querying, and data exchange are covered in section 2.6.

2.1.1 BACKGROUND AND FUTURE OF BIG DATA IN HEALTH CARE

Big data has been deemed as the key driver for creating value and transforming health care providers, however, health care providers have been reported to discard 90% of the data that they generate (Hinssen, 2012). Even though health care providers collect massive amounts of data, however, this collection has been motivated by patients' care, compliance, regulatory requirements, and record-keeping (Raghupathi & Raghupathi, 2014). Apart from what big data has been used for previously, it can also be used for managing decision support systems, disease surveillance, and population health

management, but then, in order to achieve all of this, health organizations need to be data-driven. According to a report by (IBM, 2013),

To thrive, or even survive, in this time of massive change, health care organizations [sic] must become data driven. They must treat data as a strategic asset and put processes and systems in place that allow them to access and analyse the right data to inform decision-making processes and drive actionable results (p. 2).

In support of this statement (Groves, Kayyali, Knott, & Van Kuiken, 2013; Shah & Tenenbaum, 2012) emphasise that data-driven medicine will enable the discovery of new treatment options, discover hidden trends in data, identify patterns related to readmissions and drug side-effects, deliver patient-centered care, and reduce health care costs. Data-driven health care systems have a strong focus on big data, but what does big data entail?

Big data definition

(Villars & Olofson, 2011) have defined big data as “a growing challenge that organizations [sic] face as they deal with large and fast-growing sources of data or information that also present a complex range of analysis and use problems” (p. 2). (Kaisler, Armour, Espinosa, & Money, 2013) define big data as the amount of data that is beyond the current computer storage and the processing power, and regard big data as a moving object, because it constantly changes in structure. The data is difficult to store and process, because some is correctly ordered (structured), while some is without order (unstructured). The properties of big data includes:

- the lack of available computer processing for ingesting, validating and analysing large volumes of data;
- the lack of methods to deal with unstructured or schemaless data; and
- the lack of methods to deal with real-time collection and analysis of data.

While On the other hand, (Feinleib, 2014) defines big data based on the impact that this data has, noting that it is the ability to capture and analyse data and gain actionable

insight from that data at a much lower cost than was historically possible, that makes it valuable.

In a similar way to (Feinleib, 2014), (Feldman, Martin, & Skotnes, 2012) have defined big data as a natural resource with a high value by stating that “big data is the fuel, it is like oil. If you leave it on the ground it does not have a lot of value” (p. 7). A succinct definition that covers all the aspects of big data is that one of (Demchenko, De Laat, & Membrey, 2014), they have defined it as: “high-volume, high-velocity and high-variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making” (p. 9). Common dimensions for big data are: volume, variety, velocity, and veracity as it is shown in Figure 2.1.

2.2 THE CHARACTERISTICS OF BIG DATA

1	Volume	=	quantity, from terabytes to zettabytes
2	Variety	=	structured, semi-structured and unstructured
3	Velocity	=	from any-time batch processing to real-time streaming
4	Veracity	=	quality, relevance, predictive value, meaningfulness

Figure 2. 1: Dimensions of big data (Source: (Feldman et al., 2012))

Figure 2.1 summarises the dimensions of big data, where in this case, the ‘Four Vs’ that are used to define big data.

2.2.1 VOLUME

The first “V” in Figure 2.1 is the volume of the data, which indicates its size or quantity. The greater the number of electronic devices there are, the bigger the volume of data

produced from those devices, where it can therefore be deduced that the rate of growth of the number of electronic device is directly proportional to the growth of data from those devices. In 2011, (Friess & Vermesan, 2011) predicted that the growth of connected devices will reach 16 billion by 2020, and a year later (Swan, 2012) predicted 50 billion to be reached by 2020. With the expanding production of electronic devices, data is growing in immeasurable quantities as well.

In 2011, (Hinssen, 2012) estimated that the size of big data was 150 exabytes, which is equivalent to 250 million DVDs of data, noting that this data is growing at a rate of 1.2 to 1.4 exabytes per year. (Chen et al., 2014) cements the relationship between data and electronic devices by stating that, the growth of IoT (internet of things or connected electronic devices) and cloud computing promotes a sharp growth of data. Figure 2.2 gives a glimpse of IoT products that are currently being used in delivering care to patients, noting that all these devices generate exabytes of data that could be useful to health care facilities if they could be collected and analysed. This proliferation of data is caused by the fact that these devices provide more processing power; more storage; more value for money; and are smaller in size, in such a way that their users can carry them around. This is in keeping with Moore's Law, which states that the overall processing power of computers doubles every two years. Increased processing power means more transistors can be fitted into the device's microchips, and there is and more memory storage, and ultimately more data, which can be generated quicker than it can be stored. As a result, it is termed 'big data'. In addition (Philip Chen & Zhang, 2014; Villars & Olofson, 2011) supposed that the rate of growth of data has extended beyond Moore's Law.



Figure 2. 2: Internet of Things products and prototypes (Source: Islam, Kwak, Kabir, Hossain, & Kwak, 2015)

The effective use of big data has the potential to transform economies and to deliver production growth, however, big data includes data that is inconsistent, incomplete, lacks privacy, is semi-structured, and unstructured (Philip Chen & Zhang, 2014). Hence, (Friess & Vermesan, 2011) advise caution, saying that the data generated will only be of value if it can be collected, analysed and interpreted. In many instance it remains difficult to obtain value from big data.

2.2.2 VARIETY

Similar data sets that are collected from different devices and from different data sources have a high probability of becoming unstructured during data integration, particularly when the target data set and the source data set do not conform to a similar data acquisition standard, data exchange standard, or data storage standard. On the basis of varying data, variety in big data can be defined as data acquired from diverse data sources, and from multiple data sets. The attribute of variety means that the data is extremely heterogeneous at the data set (schema) level as well as at the metadata level (Dong & Srivastava, 2013). Big data varies in structure because it holds the qualities of being structured, semi-structured and unstructured (Demchenko et al., 2014; Raghupathi & Raghupathi, 2014). Structured data can be easily stored, queried, analysed, recalled and manipulated (Feldman et al., 2012), whereas semi-structured data is defined as being neither raw, nor of a strict type or characteristic.

Incomplete columns in the data sets might even have extra information such as annotations, and similar information that is stored differently in multiple tables (Abiteboul, 1997). Figure 2.1A and Figure 2.1B reveal some of the properties of semi-structured data as defined by (Abiteboul, 1997). Data in both features represent a single patient, however the patient's laboratory data is stored differently across health care provider A and B.

Table 2. 1A: Data set from Provider A

HOSPITAL	ITEM	VALUE	UOM	RESULTSTATUS
Medico	Glucose (serum)	121		Normal
Medico	Blood Pressure systolic	137	mmHg	
Medico	Blood saturation	95	%	Normal

Table 2. 1B: Data set from Provider B

HOSPITAL NAME	LABEL	Sample	OBS_VALUE	UNIT OF MEASURE	OBSERVATION
Steve Lancet	Manual BP [Systolic]		20	Mmol/ml	True
Steve Lancet	Glucose	Serum	-		
Steve Lancet	SpO2	Blood	137	percentage	

Both Table 2.1A and Figure 2.1B can be distinctly classified as structured data, but if the data from Table 2.1A and Table 2.1B were to be integrated, the data would then be semi-structured. The process of exchanging the data between provider A and provider B would be difficult due to the following issues in the data:

- Metadata integrity: Observation name and sample is concatenated into one field called Item in provider A, whereas for provider B there is a column for the sample called sample.
- Metadata and data inconsistency: Both providers have similar database attribute names, where even the method of measuring test units is not the same. Provider A uses mmHg for blood pressure unit of measure, while Provider B uses mmol/ml.
- Data integrity: Provider A stores the full observation name for oxygen saturation, while Provider B only stores an abbreviated (spO2) version of the analyte.
- Missing data: The sample attribute in Provider B does not have a value and this is an important attribute when managing laboratory tests.

Therefore, it can be deduced that the above data is dirty or messy, it is filled with conflict, and this sort of data can mislead data analysis if the data cleaning process is not carried out correctly (Do, 2009). More details about data cleaning are covered in the current chapter in section [2.5.1](#), where the researcher will delve deeper into data cleaning strategies, including which ones are appropriate for clinically-based data. Thus far, the focus has been solely on data that is stored in tabular form such as tables, relations, arrays and spreadsheets, and this form of data storage is suitable for structured and semi-structured data.

Meanwhile, unstructured data is a lot more complex, because it is difficult to acquire, to store, to analyse, and to visualise. This sort of data is collected from different sources, at different intervals, and as a result, the data has a high possibility of becoming unstructured. Figure 2.3 presents a “word cloud” of unstructured data, which is aimed at giving an overview of what type of data is unstructured in the health care industry.



Figure 2. 3: Unstructured data word cloud

Acquiring unstructured data means: getting data from physical file processing systems; scanned files using text-extraction algorithms; organizational email servers; or through voice input. All these inputs require extra processing power and intelligence in order to extract and transform the data into a machine-understandable format. Hence, (Barbulescu et al., 2013) roughly defined unstructured data as the type of data that is easily understood by humans, but least understood by computers. Once the data is acquired, it is then to be stored in a format that allows it to be easily retrieved using a suitable query language.

It is reported that 80% to 85% of business information exists as unstructured data, which includes: organizational documents, images, emails, reports and more (Abdullah & Ahmad, 2013; Jing Gao & Koronios, 2015; Gharehchopogh & Khalifelu, 2011). It is thus a challenge for organizations to create value from this data, because it is not structured in a manner that would allow for accurate data analysis. Storing big data in the popularly used Relational Database Management Systems (RDBMS) would be a challenge, and more details about big data storage is covered in Section [2.6.6](#).

Once the unstructured data has been stored in a format that allows it to be queried and retrieved, then data analysis and even visualization can be performed on this data. Inasmuch as unstructured data poses a lot of challenges, it also presents a lot of

opportunities for organizations that are prepared to use efficient computer algorithms to analyze the data and create value from it. Algorithms that were fed lots of messy but relevant data performed better than the same algorithms with less but accurate data.

In that regard, IBM and Google's language translation algorithms were compared against each other for performance. IBM is said to have fed their translation algorithm lots of accurate data sets for translation between English and French languages, where the algorithm performed fairly well. However, Google later fed their messy data sets from multiple and various data sources, including voice as input. At first, the translation was accurate to some degree, but with glitches. Over time, however, it performed better translating more than 60 languages, where even uncommon translations, such as from Hindi to Catalan, proved possible (Mayer-Schönberger & Cukier, 2013).

2.2.3 VELOCITY

It can be assumed that measuring a phenomenon gives one an advantage in gaining valuable information about that phenomenon. This emanates from observing environments where data is constantly being collected in huge quantities, over short time periods, and from various data sources, with the aim of identifying areas that could cause problems, or that could create value for organizations.

A Controller Area Network (CAN) is a valuable asset in mobile vehicles, because it constantly collects data while monitoring every state of the vehicle. The same can be said about a critically ill patient in a hospital bed, where the patient can be monitored at different intervals by machines and even by humans; and data is collected in real-time, with the objective of improving the patient's outcomes. Patients with heart conditions can be given a wearable electrocardiogram (ECG) device that they can wear while at home, and this device constantly streams ECG measurements to the patient's electronic health record system.

In 2004, Google started a project of digitising the world's textbooks, and by 2012, 15% of those books had already been digitised. An estimated 130 million distinct books have been published since the invention of the printing press between 1440 and 1450, so Google was able to digitise 20 million unique textbooks in eight years. Not only was the content digitised, but it was also transformed into usable data that can be indexed, and comparable for analysis. The indexing of books reveals the need to use the data generated by hospital telemetry devices, where, as it stands, the ECG telemetry device connected to a patient is able to generate 1000 readings per second (Mayer-Schönberger & Cukier, 2013). However, this data is underused and thus wasted (Belle et al., 2015).

Digitising and indexing data raises challenges for the privacy and security of the organizations collecting data, where data is classified, and some has limited availability. The next section introduces data quality assurance as big data is acquired, stored, analysed and visualised.

2.2.4 VERACITY

Veracity results from collecting large sums of data. Where the data has been collected, the following questions ought to be asked to ensure data quality:

- How accurate is the data?

This quality indicator measures the correctness of data values stored for an object. e.g., a short date format for a South African date format is as follows "yyyy/MM/dd", where storing date values as "2016/13/10" is not accurate, because the maximum month value is 12.

- Is the data recent?

This measures how up-to-date is the data. When it comes to health, doctors require relevant data to make informed decisions, where they need to measure the amount of cholesterol in the patient's blood, the most recent information would be more relevant than old one, because the patient's body changes over time, hence old information becomes irrelevant over time.

- Is the data consistent?

This measures the data uniformity, where, when sharing data, the data values and the metadata must always be consistent. If a “gender” field name is used to store gender information, at any point in the future, the same field must be used instead of alternating to a “sex” field name.

- Is it accessible, or is it private?

This measures how easily accessible the data is, and whether the data can be searched and retrieved. Other data is confidential, and should therefore always be treated as private data. This is sometimes encrypted, therefore, it should be possible to decrypt encrypted data.

- Can organizations trust this data?

This measures the integrity of the data in relation to where the data originates, and whether the data provider can be trusted. This question is even asked of the data manager themselves.

Since organizations share information between one another, it is of paramount importance that measures are taken to determine the trustworthiness of the data as well as the data providers (Dai, Lin, Bertino, & Kantarcioglu, 2008).

Answering these questions about the data ensures data quality assurance, which is defined as: the process of profiling the data to discover inconsistency, inaccuracy, incompleteness, and other anomalies in the data (Gao, Xie, & Chuanqi, 2016). Data cleaning, extraction, aggregation, transformation, and loading are all part of the data quality process.

When big data is collected, there is a high chance that the data will be unstructured as it was mentioned in Section [2.2.2](#), and some of the data might be redundant. Therefore, applying redundancy reduction and data compression can reduce redundancy, without affecting the validity of the values, thereby compressing the magnitude of the data for efficient data storage (Chen et al., 2014). In addition, there are other data issues that must be eliminated in order to improve data quality (see Figure 2.4). (Gao et al., 2016) suggested that organizations do not understand their data quality, and have difficulty understanding the reasons to invest in data quality.

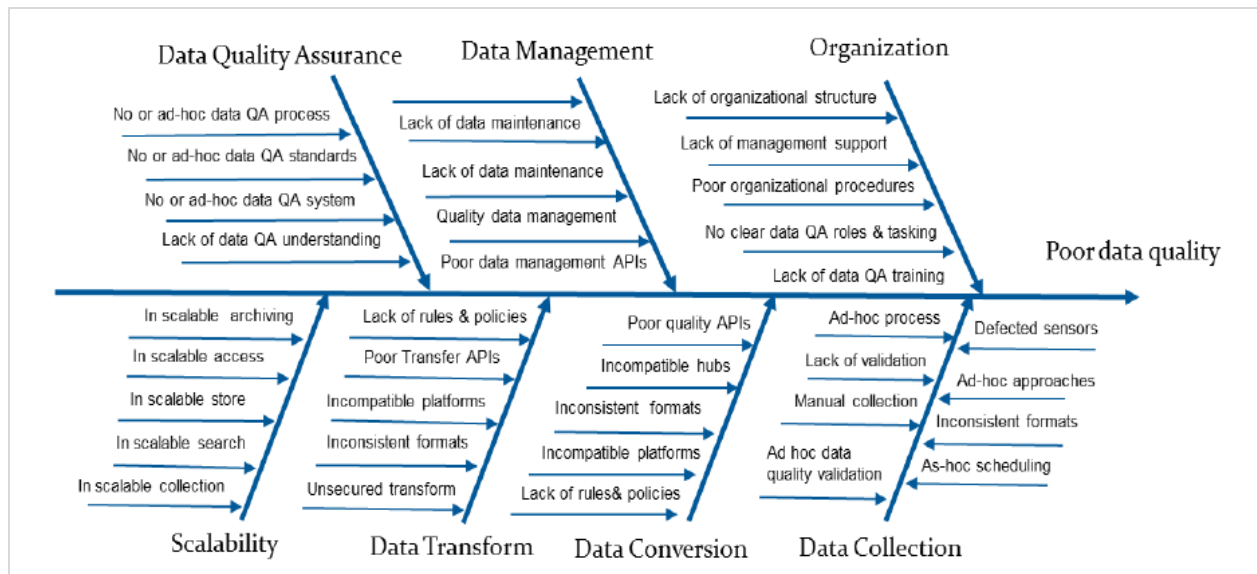


Figure 2. 4: Factors affecting data quality (Source: (Jerry Gao et al., 2016))

In other instances, it is difficult to achieve data quality, because the data is deliberately messy, and is encrypted to ensure that confidentiality is not compromised. Health data has a high chance of being private, anonymous and secured, because it is sensitive, and an incorrect change to it could lead to wrong prescriptions being given to incorrect patients. According to (Kleynhans, 2011), in South Africa health information is not easily accessible because majority of the health institutions record patient information on a paper-based filing system. While South Africa lags behind in the digitisation of health records, in the United States of America (USA), patients are able to download their data using the blue button programme, so that patients and doctors can easily access this information (Turvey et al., 2014). Once the data has been downloaded, it can be accessible to authorised personnel, but, what are the risks to be mitigated that comes with big data?

2.2.5 THE RISKS OF BIG DATA

The availability of relevant information gives companies a competitive edge in business. Amazon can recommend ideal books to its users, while Google can rank and list the most relevant websites to its users, Facebook gives you the platform to find your long-lost friends, and governments use the census data to improve service delivery to its citizens. However, the government can turn this information into a system of repression,

where for example in 2013, Edward Snowden exposed mass data collection by the National Security Agency, who uses this data to spy on American citizens (Bilbao-Osorio, Dutta, & Lanvin, 2013).

Companies can misuse the data that users share on their platforms, selling it for revenue to advertisers without the platform users knowing that their personal information is being shared. It is not only the users' privacy that is a concern with the use of big data, but in addition, there are the dangers of predictive analysis, when big data is used with algorithms to predict in advance whether a certain person is culpable for future actions. Actuaries use data predictively, and can calculate subjects such as, "men over 50 are prone to prostate cancer", and therefore, any man that is over 50 years of age may pay more for health insurance, irrespective of their state of health. Predicting events before they happen could lead to discrimination against certain groups of people and also lead to guilt by association. In the US, the Department of Homeland Security uses big data to try and identify potential terrorists by monitoring body language, and other physiological patterns, and this could turn into a weapon of dehumanisation if big data and algorithms are used inappropriately (Mayer-Schönberger & Cukier, 2013).

It remains crucial to acknowledge that big data can offer incredible benefits to governments, companies and individuals, and contrarily, incorrect uses of big data pose privacy risks, discriminatory predictions, and overreliance on data. In order to minimise these risks, government policy makers should assess the value of data usage against the risks. A risk matrix framework can be developed to measure the use of data against potential risks, and they can also develop methods aimed at evaluating the practicality of obtaining true and informed consent to use the data. The most important societal values in communities are: public health, national security, environmental protection, and economic efficiency, and therefore, the ideas of privacy and data protection should be geared towards these areas (Tene & Polonetsky, 2012).

2.3 HEALTH SYSTEM APPLICATIONS AND THE INFLUENCE OF BIG DATA

In this section of the study the researcher attempts to show which systems are used to manage data in health care. On the following subsection, the researcher covers the

categories of data that is collected in Health Information Systems (HIS). These categories are represented in a form of layers that are intended to cover all health aspects of a human being. Then lastly the researcher looks at mobile health and tele-health. All these sections are aimed at showing the different forms of data that is collected in health care, although this study only covers observable patients' data.

2.3.1 HEALTH INFORMATION SYSTEMS

HIS can technically be defined as a socio-technical subsystem of an institution, which comprises all information processing as well as associated human or technical actors. In simple terms, HIS deals with processing data, information, and knowledge in health care environments (Winter et al., 2011). HIS has four key functions, known as: data generation, capturing, analysis and synthesis, and visualisation. Ultimately, data is converted into information for making health-related decision in the health care environment (World Health Organization, 2008).

A widely regarded paper by (Haux, 2006) argues that HIS systems were intended to support health care professionals, and administrative staff in hospitals, where the primary component in HIS is the patient, such that HIS systems should be aimed at contributing to a high-quality, and efficient patient care.

Currently, health care delivery is mass-focused, but in the future, it will be increasingly individualised and patient-driven, because more data will be available that distinguishes each patient from the rest. (Topol, 2015) has noted the following about the physicians of the future:

More importantly, they will incorporate sharing your data, the full gamut from sensors, images, labs, and genomic sequence, well beyond an electronic medical record. We are talking about lots of terabytes of data about you, which will someday accumulate, from the womb to tomb, in your personal cloud, stored and ready for ferreting out the signals from the noise, even prevent an illness before it happens.

Before Topol wrote about patient-driven health care, (Haux, 2006) wrote on the future of medicine to say that in the next 10 years, information technology (IT) will be the catalyst in transforming health care into becoming patient-driven. More types of data will be captured, such as genome and proteins, where technologies will emerge such as wearable devices that continually measure and track patient's health non-invasively, which means the ability to monitor and measure non-invasively. This means that data about the patient will not be solely generated at the health care facility, but from the various patients' points of interaction.

Data generated at the health care facility is known as clinically-generated data. This data is collected from HISs such as Electronic Health Records (EHR) and Electronic Medical Records (EMR). EMR contains medical information and treatment history of a patient gathered in one practice. While EHR contains a patient's lifelong data collected from more than one practice, both EMR and EHR may include data about the patient's demographics, test results, medical scans, prescription data, doctors' medical notes, medical reports and more (Ebadollahi et al., 2006; Mxoli et al., 2014). EMR and EHR systems are managed by the health practice, whether in a hospital or a clinician's office.

Nowadays there are also other systems, known as Personal Health Record (PHR) systems. PHR systems allow patients to create, store and maintain information related to their health, where the information could be collected from multiple sources, and where the goal is to allow the patient to centrally manage their own health. They can therefore share their health information with relevant parties. PHRs can improve doctor-patient relationship, as well as health knowledge for both patients and clinicians, and allow for better management of chronic diseases (Luo, Tang, & Thomas, 2012; Mxoli et al., 2014; Mxoli, Mostert-Phipps, & Gerber, 2015). With all these advantages that PHR systems offer, at the moment, there is no PHR aimed at the South African population (Mxoli et al., 2014).

2.3.2 COLLECTION OF HEALTH BIG DATA THROUGH GIS LAYERS

EHR, EMR and PHR systems collect data from multiple sources, where Figure 2.5 gives an overview of the layers that make up the multiple health data sources, known as human GIS (Geographic Information System). In a short summary, the layers include information that deals with an individual's demographic, physiologic, anatomic, biologic and environmental data.

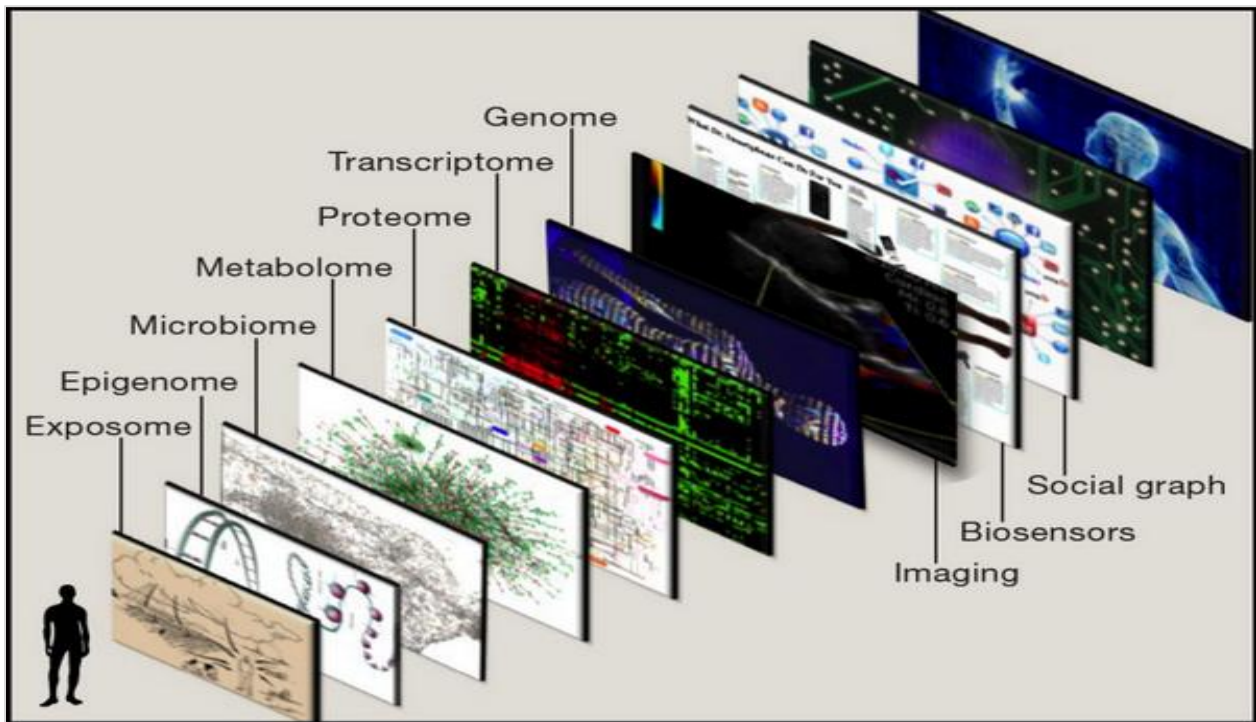


Figure 2. 5: Geographic Information System of a human being (Source: Topol, 2014)

The first layer is called the phenome, which is meant to collect information such as age, gender, occupation, family history, medications, and more. The physiome and biosensor layers work in conjunction, where physiologic data is captured using wearable sensors, and other physiologic tracking devices, such as blood pressure gauges or devices. Physiome data includes blood pressure, heart rhythm, respiratory rate, blood glucose and other metrics (Omholt & Hunter, 2016).

There are other instances when a patient's need exceeds what their clinicians or health care provider could offer, for example, when a patient requires blood or organ donation. In the health space, there are social networks that have been launched internationally in the last few years including PatientsLikeMe, CureTogether, DailyStrength, and MedHelp just to mention a few. Health social networks serve a great purpose in connecting patients, and the social aspect of health is covered by the Social Graph layer (Swan, 2009).

Health social networks have made it possible to bring people with shared interests together, even when the people are separated by geographical boundaries. It can then be inferred that through these connections: patients can find suitable organ donors; clinicians can share knowledge with other clinicians or patients through social connections; and patients can ask physicians questions and get responses at a low fee, without the need to visit the doctor's office in person. Hence pharmaceutical companies, industry analysts, policy architects, and other interested parties can easily assess the demand and the market size during clinical trials (Christakis & Fowler, 2009).

The next GIS layer is imaging and anatome, aimed at collecting data about medical scans such as x-ray, CT scans, and MRI scans. Other layers include: genome, transcriptome, proteome, metabolome, microbiome and epigenome layer, all these layers represent the levels at which one may collect microbiology data. Lastly, the exposome layer, where data is collected regarding an individual's exposure from internal to external environment, from the time they are born to the time they die (Omholt & Hunter, 2016).

Internal exposure refers to when the body's metabolism, physical activity, ageing and more are studied, after which an individual's financial status, social capital, education, climate and more are taken into consideration. Then, there is a specific external exposure that deals with matters such as radiation, environmental pollutants and chemical contaminants, occupation and medical interventions, diets, food, lifestyle factors such as alcohol or tobacco, and infectious agents (Wild, 2012).

A patient's historic data should always include environmental information, which serves to help clinicians understand the underlying causes of the patient's diseases and sicknesses. On the ground, when collecting various data sets, value could be added if data about the heart rate, blood pressure, respiratory and more could be aggregated. Aggregating this data is helpful because one could: develop deep knowledge about patients; discover proactive practice of individually-based medicine; and provide disease risk profiles for individual patients, empowering them through data (Belle et al., 2015; Chawla & Davis, 2013).

2.3.3 M-HEALTH AND TELEHEALTH

Smartphones can be used to monitor virtually any psychological metric from any place, any time, or even all the time, and such are the attributes of m-Health technology. M-Health is the use of mobile devices to provide health care services to communities, and is fully focused on delivering care to patients via mobile software applications. According to (Malvey & Slovensky, 2014), m-Health has the following advantages for the health care industry:

- it allows care to be provided at a personal level for patients;
- it improves patient's participation throughout the arc of a sickness;
- it provides preventive measures; and
- it is less expensive to implement.

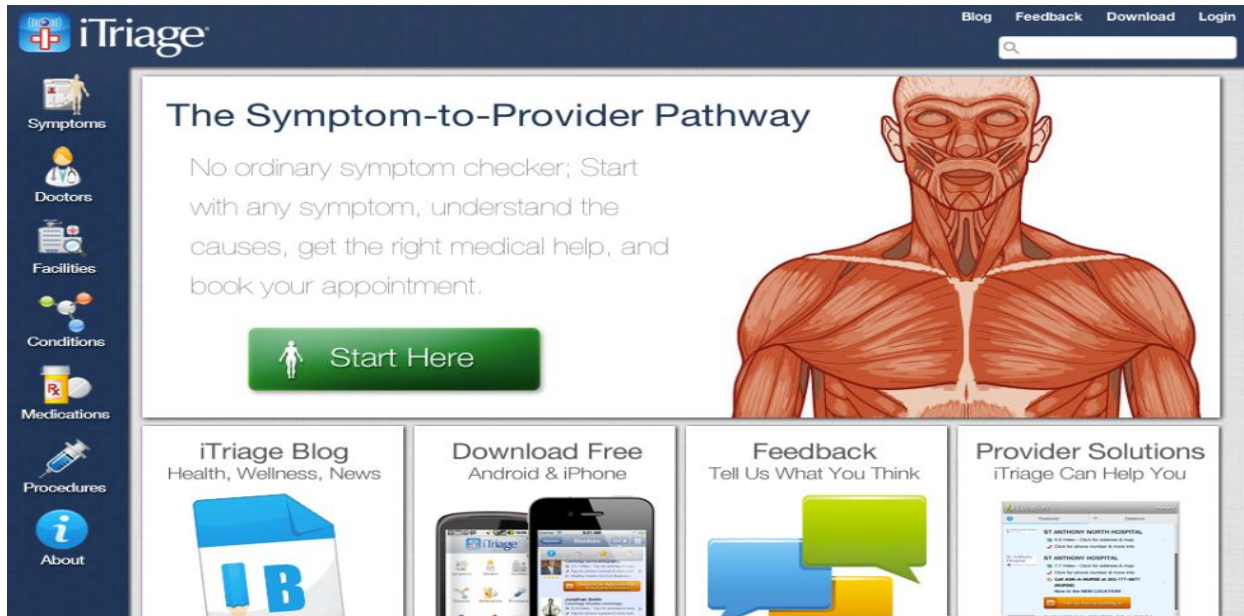


Figure 2. 6: iTriage mobile health application (Source: <http://histalkmobile.com>)

Figure 2.6 shows a mobile health application called iTriage, a free PHR App that allows patients to get answers to their health questions. It help patients to find nearby and appropriate help, securely stores patient’s health information, and allows it to be accessed remotely. Wireless communications technology have overcome geographical and organizational barriers (Poon, Zhang & Bao, 2006), where specialists such as gynaecologists spend a great deal of time travelling between multiple hospitals treating patients, some of whom could be assessed remotely.

There are other mobile applications that have helped doctors to monitor patients remotely. Monitoring of patients remotely is known as Telehealth or Telemedicine, and this improves patient’s access to health care services. Other methods of improving access include the primary health care (PHC) service, which is aimed at providing care as close as possible to where people live and work. This is an essential form of health care based on practical and scientific methods that have been made to be universally accessible to individuals and families at an affordable cost (National Department of Health, South Africa, 2015). Previously, (Porter & Lee, 2013) released a paper supporting the creation of value for a patient through integrated practice units (IPU).

IPUs are organised around the patients' medical condition, and health providers see themselves as part of a common organizational unit. With the South African NHI system, IPU can be seen as a PHC service focused in districts or municipalities or wards. There is no doubt that m-Health, Telehealth and PHC will improve health care accessibility in South Africa. However, at the moment, there are still issues such as the ones listed below:

- lack of national eHealth strategy;
- differing eHealth strategies across and within provinces;
- expensive broadband connectivity; and
- lack of interoperability and communication between health systems.

These issues were listed on the eHealth strategy document by (Masilela et al., 2013), who, in their report, suggested that a Health Normative Standards Framework could solve the interoperability problem between health systems, which is the focus of this research study. The next section presents the use of standards in health care to improve interoperability between health systems.

2.4 STANDARDS FOR SYSTEMS INTEROPERABILITY

This section partly addresses the following research sub-question:

#	Research sub-question
iii.	What methods are being used to classify objects accordingly in other industries, and how can those methods be applied in health in order to achieve semantic and syntactic interoperability?

One of the strategic priorities of the eHealth strategy for South Africa is how to achieve interoperability through standards in the delivery of care. This strategy is aimed at solving the interoperability problem that exists between heterogeneous systems when exchanging data, or when sharing health information (Masilela et al., 2013). According

to (CSIR & NDoH, 2014; Lopez & Blobel, 2009), interoperability is categorised as follows:

- Technical interoperability: covers matters of connecting systems and services through interfaces, protocols and more, for example, the IPv4 router is not compatible with IPv6 router, and there is no interoperability between the two.
- Syntactical interoperability: the exchange of messages from one system to the other, where messages must have a well-defined syntax, vocabulary, and encoding. This follows the same example that was made earlier in Section [2.2.2](#) regarding similar variable names that are written differently, or which store data in different formats.
- Semantic interoperability: concerned with the meaning of the content which is agreed upon by human rather than computer interpretation, where in health care, it is focused on coding standards. For example, application developers from organizations that exchange data should be open about the medical coding schemes that they use in their software programmes to achieve data sharing.
- Organizational interoperability: the ability for organizations to effectively communicate and transfer data or information to other organizations that are not using the same infrastructural architecture, dependent on the success of technical, syntactical, and semantic interoperability.

In this study, a great deal of focus will be paid to syntactical and semantic interoperability of patients' data. The goal of this research is to use a learning algorithm and health standards to format data so that it follows the desired structure as per the directive of the standard. Using a standard to format the structured and unstructured data ensures that the resultant data is FAIR, viz.: findable, accessible, interoperable, and reusable (Nickerson et al., 2016).

In South African health care facilities, interoperability between health systems remains a problem requiring higher priority. In light of the interoperability problem at hand, a report by National Department of Health (NDoH) compiled by CSIR shows that more than 70% of Health HIS used in hospitals do not comply with interoperability standards. Some of those that do comply, are not able to exchange health records because the hospital to

exchange with uses a different HIS, and does not comply with the standard from other hospitals (CSIR & NDoH, 2014). It is not only HIS vendors in South Africa who are less eager to implement health standards, but software vendors in other countries as well (Jian et al., 2007).

These are some of the reasons why HIS vendors drag their feet in implementing standards: 1) there are several conflicting and overlapping standards; 2) it is difficult to combine standards from different Standard Development Organizations (SDO); 3) there is limited participation in standards development process; 4) governments do not understand the importance of standards development; 5) and there is a lack of implementation guidelines and the well-skilled standards developers (Adebesin, Kotzé, et al., 2013).

Health Normative Standards Framework

The standards implementation problems have sparked the development of the Health Normative Standards Framework (HNSF) by the National Department of Health (NDoH). HNSF aims to set a foundational basis for interoperability between health systems in South Africa. On the HNSF report, three of the leading standards in health care were compared against each other with the aim of assessing which one would be suitable for implementation in South Africa. See Figure 2.7 for a summary of attributes that were weighted to reach a decision about which one to implement.

Criteria	HL7 V3	ISO 13606	IHE
Scalability	●	●	●
Implementability			●
Conformance testable			●
Market acceptance			●
Economically feasible	●	●	●
Technical capacity			●
Maturity	●		●
Extensibility and flexibility	●	●	●
Support clinical and healthcare initiatives	●	●	●

Figure 2. 7: Leading standards evaluation matrix (Source: (CSIR & NDoH, 2014))

According to Figure 2.7, the results of the report by the (CSIR & NDoH, 2014) favoured the Integrating the Health care Enterprise (IHE) option. The report said IHE has low risks to implement, has a huge market acceptance in first world countries, and there is availability of technical workforce that can implement the standard. However, IHE is not a standard per se, but an initiative by health care professionals that uses established standards such as Health Level 7 (HL7) or documents imaging and communications in medicine (DICOM) to accomplish medical workflows (Adebesin, Kotzé, et al., 2013; Vreeland et al., 2016). HL7 Version 3 has been regarded as being too technical and complex to implement, while ISO 13606 comes with high implementation risks.

HL7 is a non-profit American National Standards Institution (ANSI) accredited organization that develops standards aimed at exchanging clinical and administrative data from multiple systems (Adebesin, Kotzé, et al., 2013). On the other hand, DICOM, is used for storing and communicating medical images in Radiology, Cardiology, Ophthalmology and other departments that use Ultrasound Imaging. IHE's XDS (Cross-Document Sharing) and XDS-I (Cross-Document Sharing for Imaging) leverages on DICOM, HL7, ebXML RIM, and other standards that aim to structure and mark-up clinical content for the purposes of data exchange between institutions (Viana-Ferreira, Ribeiro, & Costa, 2014; Vreeland et al., 2016).

Table 2.2 provides a list of functions performed in a health care facility in relation to the objective of this study, hence the focus is on searching for patients' records, exchanging records between health care facilities, and tracing for patterns in patients' health records from any HIS.

IHE's XDS Architecture

When using the IHE option, every action or function in health care is linked to a profile. A profile is a detailed specification for any action to be performed in a health care facility, which is then linked to standards that can be used in conjunction with one another to carry out a given action. One of the most used profiles to share health information between disparate health systems is the IHE's XDS. XDS uses XML to store

information in ebXML repositories and registries. Figure 2.8 highlights the architecture of XDS. Repositories are used to store the physical XDS documents in a file system or a database server; registries are used to store the metadata that builds up the XDS files; the document source is used to publish the XDS document; and the consumer (clinician or patient) queries for patient's information from this document registry (Eichelberg, Aden, & Riesmeier, 2005; Noumeir, 2011).

Table 2. 2: Mapping functions to profiles and standards

Function	IHE Profile	Standards
Searching and retrieval of patient's record across multiple HIS.	Retrieve information for display (RID) and Patient identifier cross-referencing (PIX) Cross-enterprise document sharing (XDS) Multi-Patient Queries (MPQ)	- HL7 V 2.3.1 - HL7 V3 CDA release 2.0 - RIM - DICOM - ebRIM, ebMS, ebRS - OWL
Exchanging or sharing patients' electronic, media or record	Cross-enterprise document sharing (XDS) Cross-enterprise document sharing for imaging (XDS-I.b) Cross-community access (XCA)	- HL7 V3 CDA release 2.0 - RIM - DICOM - ebRIM, ebMS, ebRS
Trace for patterns in medical health data	Cross-enterprise document sharing (XDS) Patient identifier cross-referencing (PIX) Cross-enterprise document sharing for images (XDS-I.b) Cross-community access (XCA)	- HL7 V 2.3.1 - HL7 V3 CDA release 2.0 - RIM - DICOM - ebRIM, ebMS, ebRS - OWL

HL7 Clinical Document Architecture

HL7 standards were introduced in order to fix the interoperability problem in the health care industry, where this fix was mainly based on the messaging standards applied on

Version 2 and Version 3 of the standards. The Clinical Document Architecture uses XML to represent medical concepts using Reference Information Model (RIM) standard. RIM is used to define the metadata and the structure of Clinical Document Architecture (CDA) of the HL7 standards. CDA is a document mark-up standard that specifies the structure and the semantics of clinical documents. The top hierarchy of RIM contains the core attributes of RIM, which are as follows: entity, role, participation and act.

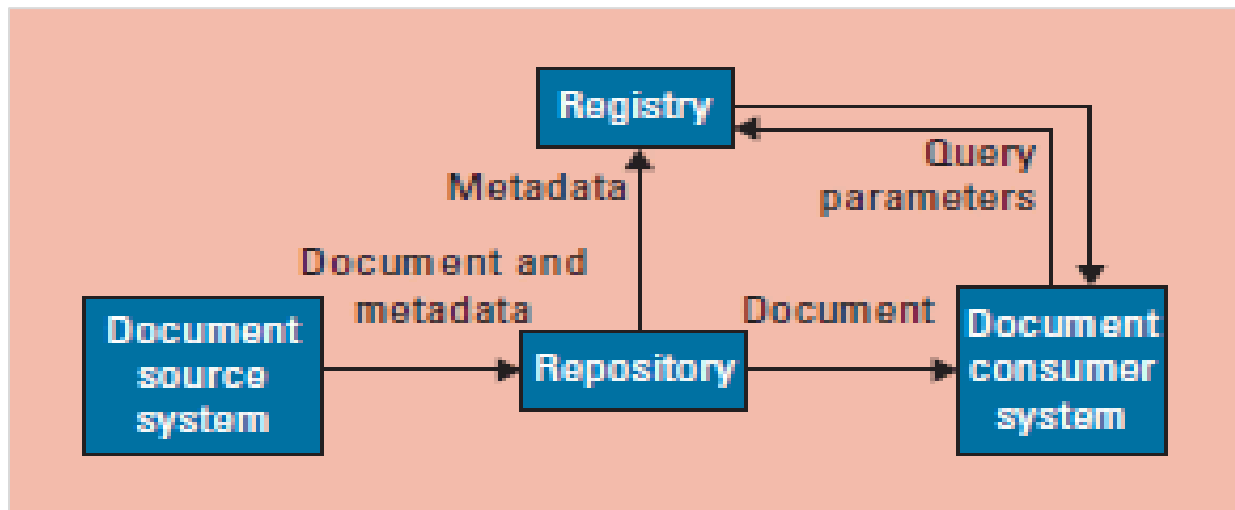


Figure 2. 8: Cross-Enterprise Document Sharing Architecture and data flow
(Source: (Noumeir, 2011))

A new standard called HL7 FHIR has been developed to eradicate the complexities of CDA, but is however still based on the ideas of HL7 RIM. FHIR is more specific, and uses resources to categorise medical concepts, for instance the observation resource is used for managing and capturing demographic characteristics, monitoring progress, and for supporting diagnostics. FHIR uses resources to represent health data, where FHIR does not only represent the data in XML, but also uses JSON, and Turtle syntax. One can observe from Figure 2.9 how respiratory information is encoded in FHIR. Some of the information on Figure 2.9 has been discarded for reasons of brevity. For users of the data to understand the data, FHIR encodes the human-readable data on an HTML tag, while other contents of the file use alternative formats mentioned above. When it comes to exchanging messages between one health care institution to the next, FHIR uses RESTful API for sending, receiving, and querying messages. The contents of the FHIR message contains even more coding standards, where in Figure 2.9, there is LOINC

code, which is embedded on the file. Now, when data is shared between health care institutions, the receiving institution and the sending institution should be able to understand and interpret the contents of the file. The next section covers the coding standard that is embedded within the FHIR resource file.

```
<category>
  <coding>
    <system value="http://hl7.org/fhir/observation-category"/>
    <code value="vital-signs"/>
    <display value="Vital Signs"/>
  </coding>
  <text value="Vital Signs"/>
</category>
<code>
  <coding>
    <system value="http://loinc.org"/>
    <code value="9279-1"/>
    <display value="Respiratory rate"/>
  </coding>
  <text value="Respiratory rate"/>
</code>
<subject>
  <reference value="Patient/example"/>
</subject>
<effectiveDateTime value="1999-07-02"/>
<valueQuantity>
  <value value="26"/>
  <unit value="breaths/minute"/>
  <system value="http://unitsofmeasure.org"/>
  <code value="/min"/>
</valueQuantity>
</Observation>
```

Figure 2. 9: FHIR Observation example for representing patient’s respiratory rate using FHIR resources (Source: (FHIR, 2011))

2.4.1 CODING STANDARDS

LOINC

(Fidahusseini & Vreeman, 2014) have defined LOINC as a universal coding system for identifying clinical laboratory observations such as patients’ vital signs, laboratory data, device measurements, microbiology, social history such as tobacco examination usage. (Abhyankar, Demner-Fushman, & McDonald, 2012) have meanwhile defined LOINC to have the following features:

- allows redundant laboratory codes to be grouped into one common code;

- used to exchange clinical documents and messages between disparate health systems using Health Level (HL) 7 FHIR or the previous HL7 standards; and
- allows data from multiple different sources, and data recorded in different time intervals to be commonly coded as a unit.

The LOINC coding standard contains six major elements, and the headings of Table 2.3 display those elements.

Table 2. 3 Mapping functions to profiles and standards

Component	Property	Time Aspect	System	ScaleType	Method
Body temperature	Temp	PT	Mouth	Qn	
Breaths	NRat	PT	Respiratory system	Qn	
Heart rate	NRat	PT	Arterial system	Qn	
Cholesterol	MCnc	PT	Ser/Plas	Qn	
Glucose	MCnc	PT	Bld	Qn	Glucometer

The component is the name of the physiologic measure, the property distinguishes between different quantities for the same substance, e.g. mass ratio code for items with mg/g as unit of measure. Time aspect specifies when the property is measured, whether at a moment in time or over a time interval, for instance, an amount over interval is expressed as mass rate (MRat, e.g. mg/24h). System is also known as the sample (e.g. blood sample), which could be urine, blood, or even the patient who is being tested can be regarded as a sample. Scale type is the scale of measurement, for instance an Albumin test could be written as follows “Albumin(>3.2)”, where “>3.2” indicates the scale type. The last part of the LOINC element is the method, which specifies the method of performing the test (Kim, El-Kareh, Goel, Vineet, & Chapman, 2012).

CPT

In full this is called Current Procedural Terminology (CPT-4), it is a five-digit code that is used to describe diagnostic procedures and other medical services such as medical billing. This code was established by the American Medical Association (AMA) and its sole purpose is to provide a standard that defines medical, surgical, and diagnostic

services. Even LOINC could be mapped to CPT codes, however, one LOINC observation name could be mapped to multiple CPT codes, and a broad CPT code to be mapped to more than one LOINC code (Vreeman & McDonald, 2005). (Matshidze & Hanmer, 2007) have reported that there is also a South African version of CPT called Complete CPT, which has extra South African codes, however this code is often used by medical schemes and providers and it was also mentioned that adoption by the public sector was tied to the CPT's proprietary nature.

SNOMED-CT

SNOMED-CT coding standard is also known as Systematised Nomenclature of Medicine-Clinical Terminology, and is maintained by the International Health Terminology Standards Development Organization (IHTSDO). It is used for representing clinically relevant information with consistency, where the developers of SNOMED claim that it is the most comprehensive health care terminology system in the world (Aouicha, Ali, & Taieb, 2016). Other researchers (Melton et al., 2006) have deemed SNOMED-CT to be an information-rich framework that has a good clinical concept coverage and also with a rich structure of relationship between the concepts. SNOMED-CT has been reported as having more than 361 800 concepts since 2004. It has 46 semantic relationships, which define the type of relationship between concepts.

RXNORM

It is a coding standard for controlled-medical terminologies, where this standard was developed by Unified Medical Language System (UMLS) in order to integrate and map competing medical terminologies with an aim of achieving interoperability. This standard is built up of the following elements: medication name; dosage; route of administration; ingredients; and common dose forms. The use of RxNorm has become even more important, due to the Meaningful Use programme, which is focused on improving the quality of delivering care in United States (Bennett, 2012).

2.5 DATA PRE-PROCESSING

The causes of unstructured data were mentioned in Section [2.2.2](#). To review, structured data becomes unstructured as more data from multiple and various sources are brought

together. In Figure 2.3, types of unstructured data were shown, such as medical reports, scans, doctor's notes, and so forth. In this section, the researcher will cover data pre-processing methods such as data cleaning, as one of the processes of data mining.

2.5.1 DATA CLEANING

According to (Natarajan, Li, & Koronios, 2009), data mining consists of a list of methods for discovering useful information in the data, and extracting hidden data from a collection of data sets. Once the data has been collected, it must first go through a data cleaning process before value can be created from that data. Data cleaning is the process of identifying errors within messy data, such as missing, duplicate, inconsistent, incomplete, or unreasonable data.

According to (Tang, 2014), errors in the data are removed by following a three-step process: error detection, data repair, and data cleaning systems. In addition (Chen et al., 2014) suggests a more rigorous approach which involves: (a) identifying error types and categorising them; (b) searching for and identifying actual errors; (c) documenting error examples and error types; and (d) modifying data entry procedures to reduce future errors.

Data cleaning plays a pivotal role in data analysis, because it dictates what should happen to the incoming data before it is integrated with other data sets for analysis. The main problem with incorrect data is that it may lead to incorrect analysis, and ultimately provide detrimental conclusions to the consumers of this information. When data cleaning is performed incorrectly, it may lead to accidentally introducing bias during modifications, and can even remove important fields and values from those data sets, where any change to the data set impacts data analysis (Malley, Ramazzotti & Wu, 2016; Taleb, Dssouli & Serhani, 2015). One of the inevitable problems of data integration is when there is missing, erroneous and inconsistent data.

Missing data occurs when no value is stored for the variable in an observation, or the data set attributes do not exist, whereas in other databases they exist, or attribute

names do not match across databases. Below is a list of potential solutions for how to handle missing data as per (Christen, 2012b; Kuhn & Johnson, 2013).

- remove records without data, in SQL databases those are attributes with empty values, not “null” value, where “null” indicates a value that is not required. The records to be removed should be assessed, firstly if they do not relate to other records from a different table that has values, or if crucial fields such as addresses and names are missing from a contact table, that record can be removed because the sole information is missing;
- remove non-identifying attributes (non-primary key) that are missing values, where identifying attributes by default ought not to allow null or empty values. However, an identity field should not allow empty or “null”, for example, because it is an identifying and a crucial field; and
- if the missing value is a postal code, and the name of the province and the city appear, then one could use the available data to search for the missing value in other databases. Also, a gender value could be extracted from the person’s specified identity number.

Missing data can be managed better with tree-based techniques, such as Decision Trees, which do not require attributes or values to be removed or altered. However, k-nearest neighbour, feature extraction and linear regression perform better when dealing with missing data. Hidden Markov has been found to be useful when segmenting attributes into well-defined and consistent attributes (Kuhn & Johnson, 2013).

Noisy data is defined as data that is mislabelled. At this stage, noisy data is different from an outlier, where an outlier is an abnormality, anomaly, discordant, or deviant. In health care, outliers are the result of equipment malfunction, human error, or anomalies arising from patient’s behaviour, due to unusual samples combined with other samples in a multidimensional space. An example of this would be having a value where the date of birth is recorded as “01/01/2016” but the recorded age is 20 years old, or a male record that also has an attribute with unexpected values about “number of times pregnant”.

To fix these issues, Logistic Regression algorithms and clustering algorithms can be used to detect outliers by grouping a set of values, such that those in the same group are more similar than those from other groups. As for fixing noisy data, binning methods can be used to smoothen a sorted data value according to their neighbouring and values around that data (Malley et al., 2016).

Some of the data in data sets is inconsistent or duplicated, and such data is costly to maintain and manage. This incurs great expense, in terms of money, as well as the computer storage and processing speed. Often this is caused by allowing free-text, instead of allowing users to choose from a list. An example of that is when exchanging data between two databases, one uses “F” to store a female value and another uses “Female”. Inconsistencies could be identified through the use of object identification and linkage through multiple sources, where linked data sets help to remove inconsistencies. Linked data sets work with the context and the data usage pattern, where context is used to identify similar data items between data sets, and the data usage pattern is used to identify data that is grouped together even when it is not similar (Liu, Kumar, & Thomas, 2015).

Other research studies have found that Functional Dependency (FD), and its extension Conditional Functional Dependency (CFD) integrity constraints yields better outcomes for detecting inconsistencies in data sets. Inconsistencies can be repaired by partitioning data sets either vertically or horizontally, however in distributed systems such as Hadoop and MapReduce, it is much harder to detect errors in the data (Fan, Li, Tang, & Yu, 2014).

The methods discussed above are sufficient for data cleaning only, and once that has been achieved, features should be extracted or selected from the cleaned data sets. Thereafter, more algorithmic processing can be performed from this data, and details about this processing is covered on the following section.

2.6 DATA INTEGRATION

Data integration is about building relationships between one or more data sets, and relations create value for data consumers, which in the case of this study are health care clinicians and patients. Through these relations, predictive models can be created. Predictive models work based on probabilities and not on exactitude, an example of which is Google's Flu trends model, which was used to forecast in real-time the potential number of influenza cases in various geographical location based on what people searched for on Google's Search engine. This model helped medical centres to respond timeously to pandemic outbreaks (Huang et al., 2015; Mayer-Schönberger & Cukier, 2013).

2.6.1 FEATURE SELECTION

Dimensional reduction is one of the data preparatory concepts that comes post data cleaning step. There are two forms of dimensional reduction, namely feature extraction and feature selection. Feature extraction is used to transform the data from its original space into a new one with lower dimensionality that cannot be linked back to the original space. Subsequently, feature selection aims to select a subset of features that minimise redundancy and maximise relevance to the target, which is known to have better readability and interpretability features (Aggarwal & Reddy, 2013). The most popular feature selection models are as follows:

- Filter model: there are three types of this model, which work without classifiers, and these are: Relief, Information Gain, Fisher Score, CFS and FCBF.
- Wrapper model: this uses a classifier as a selection criteria, and it requires cross-validation, even though it is computationally expensive, but it offers greater accuracy. Classifiers that could be used include Support Vector Machines and K-Nearest Neighbour.
- Hybrid model: this combines the best functionalities of both Filter and Wrapper models, however, it employs the BBHFS and HGA algorithms.
- Embedded model: this is known for achieving model fitting and feature selection simultaneously, these are regularisation methods such as Lasso, C4.5, BlogReg and SBMLR.

Based on the type of features to be used, the researcher will either extract features from data attributes or will select features from data attributes. Feature selection methods offer: a high learning accuracy; a better learning performance; and a lower computational cost. Feature selection is also able to discriminate samples that belong to different clusters; and ultimately it allows for human supervision (Charu, Aggarwal & Reddy, 2013). A learning algorithm that works with health-based data should allow for human input, and hence, ought to feature selection methods which allow for human intervention in order to improve the accuracy of the classifier, this procedure is part of supervised learning.

Feature selection criteria can be implemented using the processes as shown in Figure 2.10, where the whole process is divided into two phases, namely, Feature Selection and Model Fitting and Performance Evaluation.

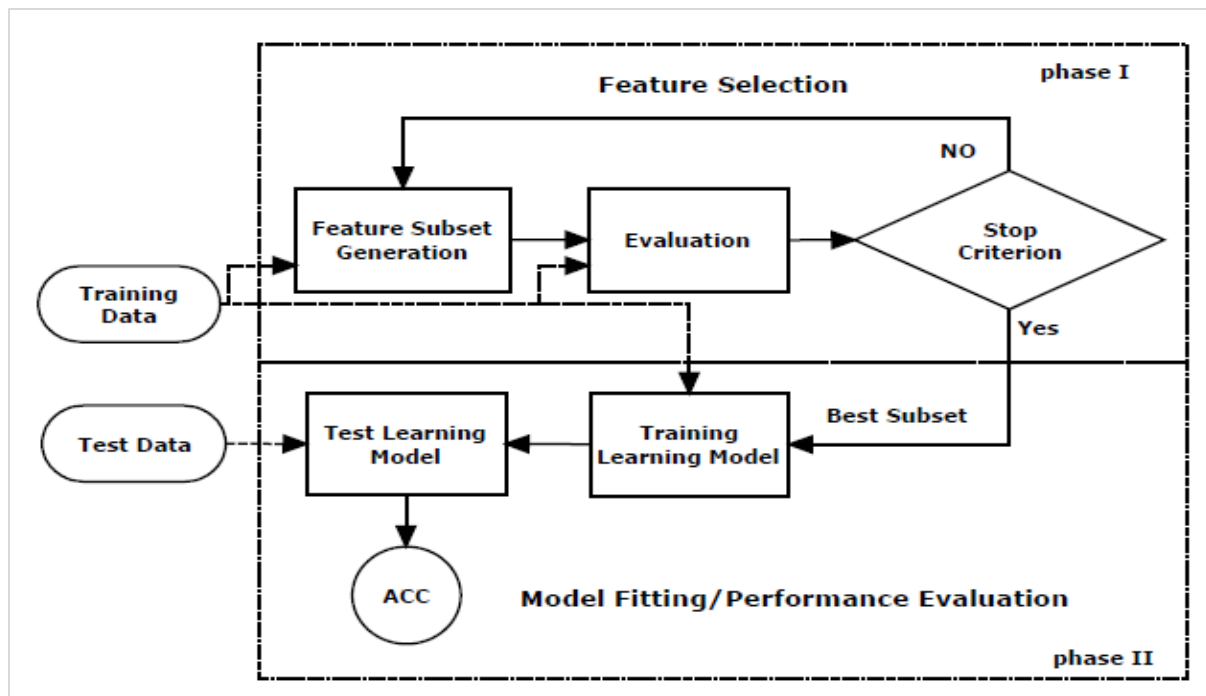


Figure 2. 10: A unified view of a feature selection process (Source: Liu, Motoda, Setiono & Zhao, 2010)

During the Feature Selection phase, a candidate training set that contains a subset of the original training data (sample) is generated, after which the candidate set is evaluated by discarding or adding features based on relevance. Lastly, using a stopping criterion, optimum features are determined and selected for the learning model, if they

are not good enough or do not satisfy the stopping criterion, the whole process is repeated (Setiono et al., 2010).

Furthermore, (Setiono et al., 2010) have noted that, once the optimum features have been selected, they can then be used to filter the training and test data for model fitting and predictions. One other thing to take note of is the results of the model on the test data, which could be used to evaluate the effectiveness of the feature selection algorithm for the learning model. However, before features can be fed into a learning algorithm, they should be presentable mathematically. According to (Zhao et al., 2011) Vector Space Model (VSM) can be used to represent features by taking term occurrence statistics as feature vectors from a plain text document.

Although VSM works best with flat files, it has been recorded to perform poorly when handling structured data sets. Structured data sets include RDF graphed data sets, XML data sets, and JSON-based data sets. (Asghari & Keyvanpour, 2015; Zhao et al., 2011) have proposed the use of a Structured Link Vector Model (SLVM) which extends VSM, and can be used to represent the structure and the contents of XML files for the learning algorithm.

2.6.2 SIMILARITY MEASURE

In the case of this study, the classification is based on whether the source record matches the target record. However, in order to determine these matches, a similarity measure is used for each attribute value. The similarity measure function outputs a weight of how much similar one string is from another, where, if the source string matches the target string then a weight of “1” is given otherwise it is given a weight of a “0”. The following section serves to provide details about similarity measure algorithms used in string comparisons. Similarity algorithms are not limited to the three covered in this study, but an interested reader may find more from a studies by (Christen, 2012; Doan, Halevy, Ives, et al., 2012).

Jaro-Winkler

(Doan, Halevy, Ives, et al., 2012) have reported that both Jaro and Winkler's techniques have the highest average similarity values for short strings, and therefore that makes these techniques suitable for calculating the similarity between the source and the target labels. Firstly the Jaro function aims to find the common character between the instance of source label and the target label, in Equation (1) c represents instances of common characters between the source label and the target label, and t is the transposition character which represents the instance of both the source label and the target label that are not matching even though they are common (Han, Kamber & Pei, 2012).

$$Jaro(x_i, x_j) = 1/3 [c/|x_i| + c/|x_j| + (c - t/2)/c] \quad (1)$$

The Jaro formula was then modified. Equation (2) shows the modification to the equation. This modification is meant to improve similarity between strings that are similar in the beginning of the string and differences are found towards the middle and in the end of the two strings (Christen, 2012).

$$sim_Winkler(x_i, x_j) = Jaro(x_i, x_j) + (1.0 - Jaro(x_i, x_j))p/10 \quad (2)$$

The "p" variable represents the first four matching characters at the beginning of two strings, for instance comparing the strings "Mandla" and "Mandela" would yield the results "p"=4.

$$sim_{Winkle_long}(x_i, x_j) = sim_{Winkler}(x_i, x_j) + (1.0 - sim_{Winkler}(x_i, x_j))c - \frac{p+1}{|x_i|+|x_j|-2(p-1)} \quad (3)$$

The optimisation objective of the Jaro and Winkler formula is the output of a value between “0” and “1”, where any value that is closer to “1” indicates that the similarity between the source label and the target label is high, and closer to “0” indicates items with low similarity weight.

Edit distance

The Edit distance algorithm is also known as the Levenshtein distance, and it measures the minimum cost of transforming one string to the other. The process of transforming involves inserting, deleting, and substituting characters from one string to the other. This process can be applied to either string and the effect is the same, and this method is mostly used where data is captured manually, where people could make typographical errors (Doan, Halevy, & Ives, 2012).

$$s(x_i, x_j) = 1 - \frac{d(x_i, x_j)}{\max(\text{length}(x_i), \text{length}(x_j))} \quad (4)$$

An example of Edit distance is shown below where a misspelt “blood pressure” is compared against a correctly spelled laboratory name.

$$s(\text{blood presure}, \text{blood pressure}) = 1 - \frac{1}{\max(14)} = 0.9286$$

The Edit distance measure compares each character of x_i instance with each and every character of the x_j instance, and this has a computation cost of $O(|\text{length}(x_i)||\text{length}(x_j)|)$. With this setup, (Perkins et al., 2011) suggests that there will be $\frac{n(n-1)}{2}$ comparisons, making this function highly computationally expensive.

Term Frequency and Inverse Document Frequency

This method is abbreviated as TFIDF and is used to evaluate how important a word is as well as the absence of a word from a document. This method uses a Vector Space Model, which is used for converting the occurrence of a string from document into a

numerical value. The conversion process checks for occurrence of a word from a dictionary, and assigns a “1” if the value exists, and “0” if it does not exist in the dictionary. This method is often used in spam classification problems, where the words that are expected to be contained on a random spam message will be recorded in the dictionary. The input text is first normalized to preferred text casing, then tokenised and, then stemmed, by removing stop words. Stop words are words that occur often on an English text, such as *the, is, this, on* etc. (Manning, Raghavan, & Schütze, 2009; Lan, Tan, Su, & Lu, 2009).

$$tf(t, d) = \sum_{x \in d}^n fr(x, t) \tag{5}$$

$$fr(x, t) = \{1, x = t; 0, otherwise\} \tag{6}$$

The function $tf(t, d)$ returns the number of times that term t is present in document d , where the function $fr(x, t)$ assigns a “1” if the compared terms are the same, and “0” if otherwise. Table 2. 4A represents the doctor’s notes about the patient’s vital signs, where the text has been lowered for cases and stems from common English words. Then Table 2. 4B records the document vector for vital signs, where the headers on Table 2. 4B indicates the dictionary used, and the given text is checked for whether it is contained in the defined dictionary. While Table 2. 4C is similar to Table 2. 4B, it checks for unit of measures used in vital signs.

Table 2. 4A Word stemming for source data

Source not stemmed	patient’s weight 3.115 kg. length 50 cm. head circumference 31.5 cm. large for gestational age
Source stemmed	patient s weight 3.115 kg length 50 cm head circumfer 31. 5 cm larg for gestat ag

Table 2. 4B Target vital signs features

height	weight	respiratori	head circumfer	Blood pressure	oxygen	length
0	1	0	1	0	0	1

Table 2. 4C Target vital signs features

metre	m	Centimeter	cm	kg	litre	Mmol/L
0	0	0	1	1	0	0

2.6.3 INDEXING TECHNIQUES

Indexing, also known as blocking, involves limiting the number of data objects and therefore comparisons in a feature space. If one were to compare the similarity of data objects from the source data with 100 records and the target data with 100 records, there would have to be 10 000 comparisons, which affects the algorithm's running time. The number of comparisons grows quadratically with the training set. This problem is fixed through blocking strategies whereby blocks are created based on similar characteristics of the data, for instance, records that have the same postal code would be blocked together, or laboratory names that sound the same, or the first three characters of lab names that are similar. (Bilenko, 2006) has reported that blocking is more critical in the scaling of record linkage systems and data clustering algorithms. With the emergence of data integration, it becomes even more important to apply automatic blocking, and (Bilenko, 2006) has used an approximation algorithm to construct blocking functions automatically. (Christen, 2012) suggests that indexing should be applied to attributes or attributes that do not have missing values, and with a uniform frequency distribution between the values. He further advised that phonetic coding was specifically designed for the English language, and therefore, it ought to be used cautiously when considering South African names. An example of an indexing technique includes *soundex*, which looks at the similar sounds between two words, and *phonex* which is a variation of *soundex*; however these apply punctuation, such as removing characters prior to the word/sound comparison. There is another indexing

method called *phonix*, which has more punctuation rules, and is said to run slower than *soundex* (Manning et al., 2009).








2.6.4 DATA SET MATCHING

Differences and mismatches between heterogeneous data formats can be solved by dataset and attribute mapping systems. Dataset mapping is important for handling problems experienced during data integration, data exchange, peer-to-peer data sharing, and dataset evolution (Fagin et al., 2009). For illustration, see Table 2.5, which shows the mapping process from the source to the target dataset. For the purposes of this study, a dataset is defined as the organization of data according to a blueprint of how databases are constructed and can be viewed as a set of repositories in the form of database tables or XML or ontologies. In Table 2.5, a database table “TbILabs” is mapped to the Observation FHIR resource, where the table’s attributes are also mapped to the FHIR valuesets. The example in Table 2.5 satisfies the definition of what dataset mapping is (Bonifadi, Mecca, Papotti, & Velegrakis, 2011), defining it as “expressions that specify how an instance of the source repository should be *translated* into an instance of the target repository” (p.112).

Conceptual mapping of FHIR resources in a clinical setting

FHIR resources could be understood through scenarios in the health care environment. A patient’s visit to a clinician or a hospital could be systematically recorded, based on the following set of variables, but not limited to this list: patient information, demographics, providers, health care procedures, utilisation data (e.g., length of stay in hospital, charges), and more. According to FHIR resources, a patient’s visit to a health care facility is classified as an *Encounter*, and a patient is defined through the *Person* and *Patient* resources. A *Person* allows for a variety of roles in delivering care, where, for example, a patient being treated is handled differently from an organ donor, while *Patient* resource includes patient’s attributes.

Table 2. 5 Mapping legacy data sets and attributes to FHIR resource

Source			Target	
Database attribute	Database table	Match	FHIR ValueSet	FHIR Resource
Glucose	TblLabs		LOINC: 15074-8	Observation: Glucose [moles/volume] in blood
Weight	TblPatient		LOINC: 29463-7	Observation: Body Weight
CellNo	TblContact		Telecom	Person
Temp	TblPatient		SNOMED- CT: 56342008	Observation: Temperature taking
Chol hdl	TblLabs		LOINC: 2085-9	Observation: Cholesterol in HDL Serum or Plasma
Gender	TblPatient		Gender	Patient
Drank contaminated water, tested +ve for lead exposure	TblSummary		LOINC: 10368-9	Observation: Lead in Capillary blood

Now, the doctor treating the patient is classified as *Practitioner*, where the patient's complaints are termed *Condition*, tests to be done are termed *Observation*, the doctor's findings are termed *Diagnosis*. The resources are not limited to the few mentioned above, this is only to give an example of how the classification is done, and each resource also has attributes, which are also used for classification purposes.

Methods of mapping source dataset to target dataset

There are various methods used to map the source repository to the target repository. One of the methods is attribute correspondence. Attribute correspondence is used for associating attributes from different datasets, which also helps limit the query search space during a mapping activity. Attribute correspondence has been extended to include contextual, semantic, and probabilistic attribute correspondence. Contextual attribute correspondence maps object A_i to object A_j , based on condition c , and it is interpreted in this triple form (A_i, A_j, c) . Contextual attribute correspondence is deterministic and would prove highly expensive to thoroughly search for attributes in data that is hierarchically structured. Semantic attribute correspondence fixes the issue of hierarchical search, while probabilistic methods allow for attributes to be matched through machine learning concepts, and the combination of these three methods is thought to have the potential of an even more powerful model (Bonifadi et al., 2011).

Classification methods for dataset mapping

Datasets, attributes and values can also be mapped to the target data set using the following classification methods:

- threshold-based classification
- probabilistic classification
- cost-based classification
- rule-based classification
- supervised classification methods

Classification is used for predicting a class or a category on a given set of training examples. The threshold method is the simplest way to classify whether candidate records pairs are a match, a non-match, or a potential match, through the use of the similarity threshold. The probabilistic method uses the dataset attributes as well as the values stored on that attribute to determine a match, a non-match or a potential match, and the threshold-based method lacks this functionality. While the cost-based method can be applied in all classification methods, it aims to minimise misclassifications and it is a suitable method for classifying sensitive data. Lastly, the rule-based method applies rules that classify the candidate record pairs into a match, non-match and potential matches (Christen, 2012).

2.6.5 TEXT CLASSIFICATION ALGORITHMS

Classification is when a machine learning algorithm receives input data for the task of predicting a class or a category to which the input data should be classified. Now the received input data might have labels or it might not, so when the data with labels is trained through the learning algorithm, the algorithm is considered supervised because it is given clues about the classes to which the input data should be classified. The opposite is unsupervised classification, when the algorithm learns patterns from the input data and it creates clusters based on the unlabelled data (Wang & Domeniconi, 2008). The model is first built from the training data using the learning algorithm, the produced model is then applied to the new or unknown (test) data for making predictions (Figure 2.11).

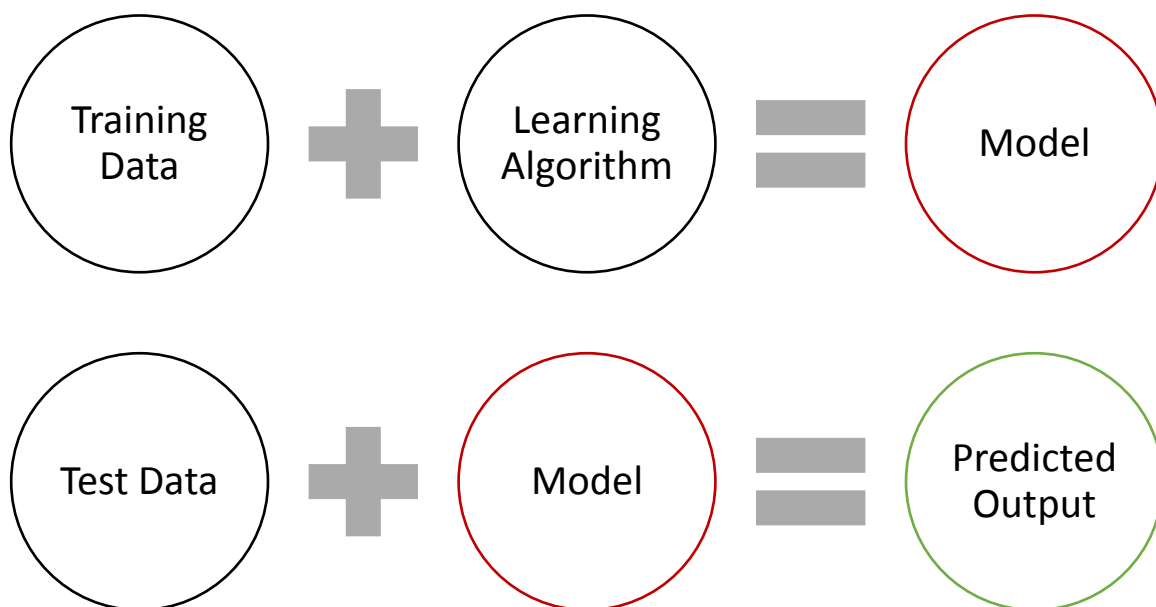


Figure 2. 11: Training a supervised algorithm

Partitions of the data are created prior to training, where there is a training set, cross-validation set and test set. The training set for a supervised classification contains both positive and negative training examples. In Figure 2.12, the training data contains five training examples m , with four features n that can be used to build the model, the target feature or class is the “Secured loan” attribute from the given training set. It is

observable that the target feature outputs a value of “No” or “Yes”, this is called a binomial classifier, there is also a multiclass classifier which output a range of values e.g. (“1”, “2”, “3”, “4”).

#	Credit history	Debts	Secured loan	Taxable income	Risk
1	bad	many	No	9,600-18,000	high
4	unknown	few	No	9,600-18,000	high
7	bad	few	No	9,600-18,000	high
19	good	few	Yes	9,600-18,000	low
20	bad	many	No	9,600-18,000	high

Figure 2. 12: Example of training examples for determining whether to grant loan to the applicant or not (Source: (Gorunescu, 2011))

The test set is loaded once a good model has been built. During the building process of the model there are parameters that are optimized using the cross-validation set. The best model is then built based on the optimized parameters, and such a model is one with a smaller cost function value. The produced model is then tested on the test set. The test set contains data that was not used during the training or cross-validation process, and it should be noted that each set of data is applied on a specific process. For instance, the training set is applied during the training process, the cross-validation set is applied during the cross-validation process, and the test set is applied during the testing process. The testing process is the final output because it tests the classifier’s (or learning algorithm) predictions on the given data, if the classifier is able to predict correctly, then it is regarded as being able to learn, and therefore, it is also considered being able to generalise from unknown data (Han et al., 2012). (Christen, 2012) further added that test data and training data should be in the same format and structure,

however, test data should not have the same data that was used on the training set, or on the cross-validation set.

There are a myriad of examples where a classification-based algorithm can be applied namely: an email spam classification where the input are words on an email message, and the classifier has to predict whether the email is spam or not spam (Kuhn & Johnson, 2013); predicting patients who are eligible for palliative care by collecting EHR data from different clinical systems (Avati et al., 2017); another example is the prediction of whether a patient's tumour is benign or malignant, this is classified from an input of an electronic radiograph image. Another well-known method of classification is one that is rule-based, it uses a set of IF-THEN rules in order to achieve classification. The rules use conditions which include disjunctions (logical OR (\vee)) and conjunctions (logical AND (\wedge)) to determine when to classify the given input data into a corresponding class (Christen, 2012). For instance, Figure 2.18 uses logical rules for determining if two distinct records match or not, in the case of Figure 2.18 it is record i and record j , and therefore the process of classification is determined by the conditions in each rule.

$$\begin{aligned}
& (s(\text{GivenName})[r_i, r_j] \geq 0.9) \wedge (s(\text{Surname})[r_i, r_j] = 1.0) \\
& \quad \wedge (s(\text{BMonth})[r_i, r_j] = 1.0) \wedge (s(\text{BYear})[r_i, r_j] = 1.0) \Rightarrow [r_i, r_j] \rightarrow \text{Match} \\
& (s(\text{GivenName})[r_i, r_j] \geq 0.7) \wedge (s(\text{Surname})[r_i, r_j] \geq 0.8) \\
& \quad \wedge (s(\text{BDay})[r_i, r_j] = 1.0) \wedge s(\text{BMonth})[r_i, r_j] = 1.0 \\
& \quad \quad \quad \wedge (s(\text{BYear})[r_i, r_j] = 1.0) \Rightarrow [r_i, r_j] \rightarrow \text{Match} \\
& (s(\text{GivenName})[r_i, r_j] \geq 0.7) \wedge (s(\text{Surname})[r_i, r_j] \geq 0.8) \\
& \quad \wedge (s(\text{StrName})[r_i, r_j] \geq 0.8) \wedge (s(\text{Suburb})[r_i, r_j] \geq 0.8) \Rightarrow [r_i, r_j] \rightarrow \text{Match} \\
& (s(\text{GivenName})[r_i, r_j] \geq 0.7) \wedge (s(\text{Surname})[r_i, r_j] \geq 0.8) \\
& \quad \wedge (s(\text{BDay})[r_i, r_j] \leq 0.5) \wedge (s(\text{BMonth})[r_i, r_j] \leq 0.5) \\
& \quad \quad \quad \wedge (s(\text{BYear})[r_i, r_j] \leq 0.5) \Rightarrow [r_i, r_j] \rightarrow \text{Non-Match} \\
& (s(\text{GivenName})[r_i, r_j] \geq 0.7) \wedge (s(\text{Surname})[r_i, r_j] \geq 0.8) \\
& \quad \wedge (s(\text{StrName})[r_i, r_j] \leq 0.6) \wedge (s(\text{Suburb})[r_i, r_j] \leq 0.6) \Rightarrow [r_i, r_j] \rightarrow \text{Non-Match}
\end{aligned}$$

Figure 2. 13: Rules that use conjunctions and disjunctions to determine whether two records match or not (Source: (Christen, 2012))

For instance, the first rule of Figure 2.13 states if the attribute “GivenName” for record i has a similarity weight of “0.9” or more when compared with the j record, and also considering the other attributes and weights, if the first rule is met then the two records are considered to match. The similarity weight is calculated using methods discussed in section 2.6.2.

2.6.6 STORAGE MECHANISMS FOR BIG HEALTH DATA

Currently, Relational Database Management Systems (RDBMS) are being used on a daily basis to store data in a structured format, and RDBMS are easily queried through “Structured Queried Languages” (SQL). RDBMS require tables and columns to be defined first before data can be stored. However, the nature of unstructured data makes it impossible to have predefined table names and columns, and therefore, relationships between the data cannot be established in similar formats as with RDBMS (Leavitt, 2010). It can then be concluded that RDBMS are powerless when storing unstructured data (Liu, Lang, Yu, Luo, & Huang, 2011).

Distributed file systems

Big data storage can be classified into three mechanisms, namely, distributed file systems, databases, and programming models. An example of a distributed file systems is a cluster-based Hadoop Distributed File System (HDFS), which was derived from Google File Systems (GFS). HDFS is a data storage platform for a MapReduce Framework, and both these technologies are a part and parcel of Hadoop. Parallel computing for MapReduce Framework is achieved when the HDFS cluster uses a single NameNode for managing the metadata of files, while the data nodes are used for storing the actual data (Chen et al., 2014; Huang et al., 2015).

Big data databases

Another form of storage systems are databases, NoSQL databases have been designed for managing huge heterogeneous data sets, as well as to scale to thousands or millions of users who are performing updates and reads, almost at the same time,

instead of guaranteeing data integrity through ACID (atomicity, consistency, isolation and durability) transactions like RDBM. NoSQL databases ensure that there is strong consistency, high availability, and that there is partition tolerance, which is transparent to the user, and is done across different servers (Moniruzzaman & Hossain, 2013).

XML has become a heavily used data format for achieving interoperability between disparate organizations. The HL7 standard and IHE XDS achieves interoperability between data through the use of XML-based technologies, such as ebXML and HL7 CDA. Therefore, in health care, there is a need to implement NoSQL-based database for XML documents, because of scalability and performance issues. An Italian hospital has used an open-source version of MongoDB database for managing large CDA documents, with a repository that contains about 22 million CDA documents, and with 50K admissions per year, and 2.5 million outpatient visits in a year (Adrián et al., 2013).

MongoDB, Cassandra, BigTable, and HBase are various forms of NoSQL databases, HBase is Google's open-sourced version of BigTable. These databases can be categorised into three forms, based on how they store data, and are: Key-value databases, Column-Oriented databases, and Document databases. MongoDB is a Document-based database, which uses Binary JSON (BSON) objects to store data, BSON and is derived from Javascript Object Notation (JSON) (Chen et al., 2014).

If one plans to store XML documents in MongoDB, a translator would be needed to translate XML elements to JSON objects in order to store and query the data in a supported language syntax. On the other hand, in section 2.4, the architecture of ebXML was explained, and it was mentioned that ebXML stores XDS documents to repositories and the document's metadata to the registry. The question then arises as to how to integrate NoSQL databases with ebXML for managing XDS documents. (Messina, Storniolo, & Urso, 2016) have proposed the use of a Multi-Model NoSQL database called OrientDB. OrientDB is an open-source database that supports Graph

databases, Document databases, Key-Value and Object models, and allows for data to be queried through “SQL” related queries.

Programming models for big data

The processing of big data comes with a lot of challenges for organizations, one of which is querying data from heterogeneous distributed sources. Even if the data uses the same standard, querying such sparse data requires distributive and parallel computing services. Exchanging and sharing big health data between multiple organizations presents scalability and performance problems.

With these problems in mind, in order to fix them, programming models like MapReduce may be appropriate. A well-known MapReduce framework is Apache Hadoop, which has two operational bases, namely, Map and Reduce. The ***Map()*** step uses the master node to take the input and recursively divides it into smaller sub-problems then distribute it to slave nodes. During the ***Reduce()*** step, the master node collects the answers and combines them together to form the final answer to the actual problem to be solved (Philip Chen & Zhang, 2014). The map and reduce are a part of analysis or interacting with the stored data from multiple nodes in Hadoop.

There are other programming models that can be used with NoSQL databases such as Dryad, All-pairs and Pregel. These models have become the foundation of analysis for big data, because they effectively improve the performance of NoSQL databases by reducing the performance gaps between relational databases (Chen et al., 2014).

2.6.7 CONCLUSION

This chapter commenced by indicating the characteristics of big data. It has been shown that big data can have four or more characteristics, the most common of which are: volume, variety, velocity and veracity. More emphasis was placed on the “variety” attribute of big data, because most data in organizations are not structured, however, this data is often deleted, or not used, because it is difficult to create value from this type of data.

The veracity attribute addressed the issue about the accuracy, relevance, consistency, security, and the ownership of the data. Often, machine learning and big data have

been regarded as a silver bullet for problems encountered in various industries. However, in this chapter, it has been shown that such data comes with risks. To mitigate these risks, a risk matrix framework should be developed in order to protect the consumers and the producers of big data.

In section 2.3, Health Information Systems such as EHR, EMR, and PHR were defined. Then in section 2.3.2 one of the research questions was answered by providing detailed layers that constitute the sources of data for this study. One of the layers is the exposome, which concerns capturing environmental data and including it as part of the patient's profile. In section 2.3.3, technologies for delivering care remotely to patients were identified and defined. One of the core research themes for this study is data standardization, by means of which to achieve interoperability amongst disparate health care facilities. In section 2.4, the researcher communicated about the use of standards such as HL7, LOINC, SNOMED-CT, ICD-10, and more. Section 2.5 briefly identifies methods for cleaning the data. In section 2.6, the researcher identified feature selection methods, similarity measures, indexing techniques, data matching and classification algorithms, and lastly storage mechanism for big data.

CHAPTER 3: Research Design and Methodology

3. RESEARCH METHODOLOGY

3.1 INTRODUCTION

The previous chapter gave a detailed view of relevant literature that constitutes this study. In this chapter, the researcher will discuss research planning. In section 3.2 the researcher addresses the theoretical perspective of the use of SVM. Then, in section 3.3 the researcher lists the research questions and shows how these questions are addressed. In section 3.4, CRISP-DM and DSRM is discussed, then in section 3.5 the researcher lists the datasets to be used to conduct this study. Then section 3.6 shows the methods of data preparation and section 3.7 lists the notation that is used in the study. Section 3.8 talks about the supervised classification methods, while section 3.9 addresses clinical tools and medical thesaurus to be used in this study. Lastly, section 3.10 speaks about ethical clearance.

3.2 THEORETICAL PERSPECTIVE TO THE PROPOSED SOLUTION

The objective of this study is to use standardized clinical observation data as input on a learning algorithm, where the algorithm would learn a function (f) for identifying patterns in the input data, so that when the algorithm is given new but related unstandardized observation data, it would be able to classify the data to the related standard. The researcher has planned to use the SVM classifier as the learning algorithm, and clinical observation data that is standardized, based on the LOINC standard. The standardized data is sometimes referred to as the gold standard, and in this study it is also termed as such. The solution that the researcher proposes is based on the Statistical Learning (SL) theory. According to a paper by (Vapnik, 2013), the theory was developed by Vapnik and co-workers more than 30 years ago. (Bousquet, 2004) noted that the SL theory is used for studying the problem of inferences by focusing on learning, generalisation, regularisation and the characterisation of the performance of a learning algorithm. In simple terms, the theory formalizes the process

of: (1) observing a phenomenon; (2) constructing a model for that phenomenon; (3) then making predictions using the constructed model. Machine learning therefore allows the steps mentioned above to be automated (Bousquet, 2004).

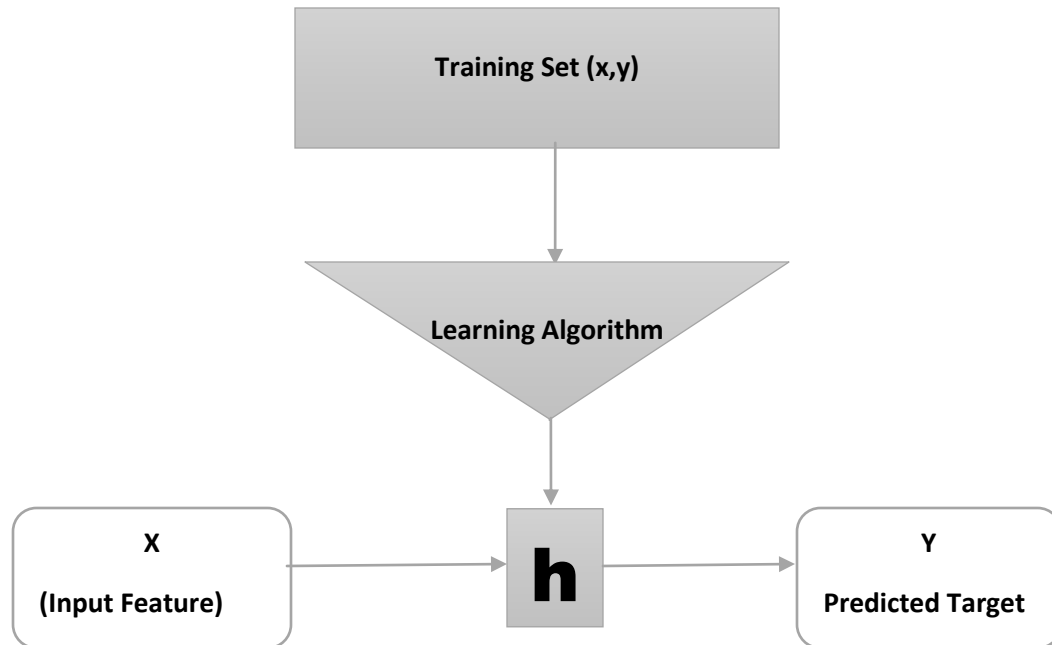


Figure 3. 1: Hypothesis evaluation process

Figure 3.1 shows the process of evaluating the hypothesis value, which is part of observing the phenomenon. The *training set* consists of input variables (X) and target variables (Y), the model is created by learning the hypothesis function ($h: X \rightarrow Y$), which is also used to predict the target variable from the given input variable. A linear classifier can be used to predict the hypothesis whereby θ_1 represents the slope of the line and θ_0 represents a point that crosses, see Equation (7).

$$h_{\theta}(X) = \theta_0 + \theta_1 \times X \tag{7}$$

The SL theory is focused on three learning problems namely, pattern recognition, regression estimation, and density estimation. Pattern recognition is used in object categorisation problems, whereby an object is categorised to a certain class based on

its properties. The SL theory is supported by a highly used classifier called Support Vector Machines (SVM), which is supervised, because it learns patterns from predefined training examples. Therefore, SVM has its roots from the statistical learning theory (Nasien, Yuhaniz, & Haron, 2010).

The principles of the SL theory has made way for SVM to be applied to classification and regression problems. Some of the notable uses of SVM include: sentiment classification (or market prediction); spam classification; bioinformatics; image retrieval; face detection; and text categorisation (Moraes, Valiati, Gavião Neto, & Neto, 2013; Tian, Shi, & Liu, 2012). SVM provides a highly accurate classification capability, and (Xu, Zhen, Yang, & Wang, 2009) have further added that SVM provides a high performance generalisation of data.

In machine learning, one of the requirements is an accurate generalisation, meaning that there is a quest to find a function (f) that is able to correctly classify previously unseen examples. Therefore, the key variables in statistical learning theory are the ability for the classifier to learn from feature sets (x, y) , generalise unseen examples, and regularise by preventing high variance (over fit) and high bias (under fit) on the training set (Ng, 2011). An example made by (Hamel, 2009) drives the bias-variance point home, where the author notes that high bias occurs when the learning algorithm cannot fit the training data, and high variance occurs as the result of fitting all the data points accurately, such that it fails to regularise (make correct predictions) on new or unknown input data. Figure 3.2 illustrates this point.

According to Figure 3.2, underfitting is shown by the linear graph, while overfitting is shown by the polynomial graph. Regularisation problems occur as the result of fewer features as well as unnecessary features, where this shows the relationship between the process of learning and the features that were selected.

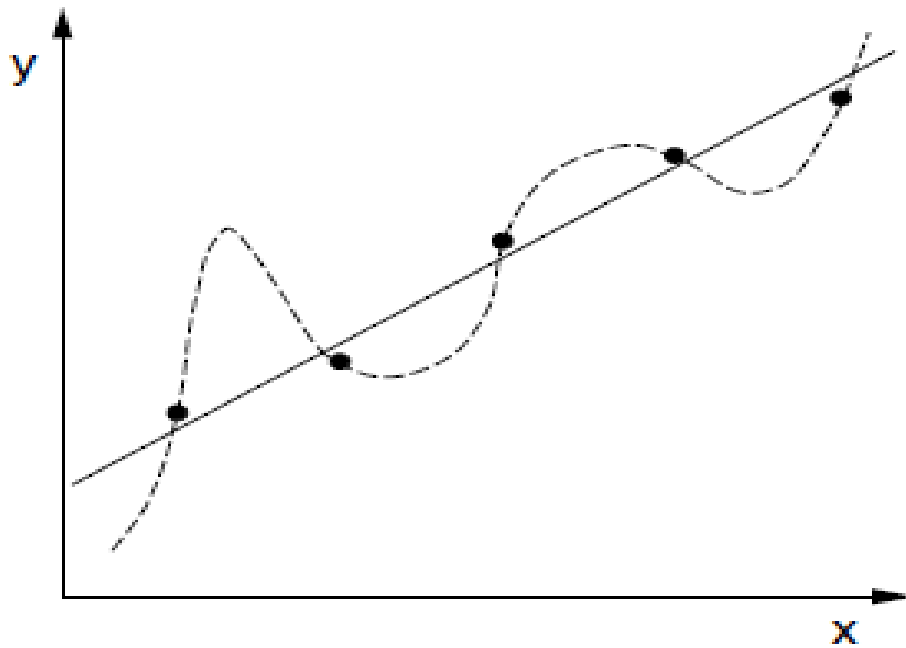


Figure 3. 2: High variance and high bias (Source: (Hamel, 2009))

A relationship that makes up a learning algorithm is guided by three variables, according to an expert in machine learning (Mitchell, 1997). Mitchell defines a learning algorithm as being able to: “learn from experience E with respect to some task T and performance P , if its performance at tasks in T , as measured by P , improves by experience E ” (1997:2). These machine learning variables can further be mapped to the SL theory, whereby a supervised learning algorithm will be able to generalise if it is able to learn from experience. The task is the action being done by the classifying algorithm, and the learning performance of the algorithm improves when the generalization error is minimized, and therefore achieving regularization.

From the objective of this study, which was defined in section [1.5](#), the researcher has extracted variables shown in Table 3.1, also revealing how the machine learning definition by Mitchell provided above influences the construction of these variables.

Table 3. 1 Variables of the study

Independent variables	Dependent variables	Mediating variables
<ol style="list-style-type: none"> 1. Similarity Measures 2. Indexing algorithms 3. Classifiers 	<ol style="list-style-type: none"> 1. Similarity weights 2. Set match 3. Classification 4. Performance measures 	<ol style="list-style-type: none"> 1. LOINC mapped dataset 2. Observation dataset

Independent variables affect the outcomes of the study, and dependent variables are the outcomes of the study, while the mediating variables are actually an independent variable that directs the outcome of the study (Creswell, 2014). The mediating variable is used as a supervision method for unstandardized data, where in simpler terms, it is an example that is emulated by the classifier by following the gold standard so as to standardize the unstandardized data. The independent variables are functions that manipulate the mediating variables in order to get the dependent variables. For instance, an indexing function is applied on the observation dataset and LOINC-mapped datasets, in order to obtain records that are compatible with one another in terms of sound, where for example, records about “blood pressure” would be compared against records that sound the same, such as shown in Table 3.2 below.

Table 3. 2 Results of records to be compared with blood pressure record

LOINC Code	Observation name
10389-5	Blood product.other
9855-8	Blood pressure special circumstances^*
79965-0	Blood velocity-time integral.systole

The indexing algorithm such as soundex limits the number of potential target records to be compared against the source record. Then, a similarity weight function such as Jaro-Winkler, Edit distance or TFIDF is applied in order to calculate how similar the two records are, and these weights are then calculated for each record-distinguishing field in a record set. The weights then are loaded into a classifier, which determines whether

the two records matches on not. Therefore the researcher's goal is to use the independent variables and mediating variables as inputs in order to determine if two records match. If they do match then that record can be standardized to the selected LOINC code. Tests for these matches are used to evaluate whether a standard can be learned through a machine learning classifier or not.

3.3 FORMULATION OF RESEARCH QUESTIONS

The researcher used the objective of the study to draw out the two main research questions, where question (a) addresses the nature of society that hinders data to be interoperable; question (b) speaks to the use of methods from science and technology in order to standardize patients' data. See questions below:

- a) When will health information systems in South Africa be standardized in order to be able to seamlessly exchange and share consolidated patients' data?
- b) How can the process of data compliance across health care providers be automated through machine learning concepts?

These are the core questions driving the research, and were used to develop the sub-questions listed in section [1.5.1](#). The objective of this study further suggests that a functionalist paradigm is used in this study because of the combinational use of why and how as main questions. (Cronje, 2014) advises that if the researcher wants to develop a prototype solution for the research problem, then the questions to ask are arranged in the following format "why is the current method not working?" and "how should it be fixed?" The questions raised suggest that a prototype will be created in order to reach the objective of the study, therefore the researcher will follow the CRISP-DM framework and Design Science Research as guidelines for purposes of this study.

3.4 CRISP-DM FRAMEWORK AND DESIGN SCIENCE RESEARCH

The researcher has chosen to use the CRISP-DM framework as a guide to reach the expected output for this study. The CRISP-DM framework and other alternative frameworks were discussed in Chapter One. CRISP-DM in full is called CRoss-Industry Standard Process for Data Mining. This framework is not only for guidance purposes, it also allows projects to be replicated, and encourages best practices of data mining in order to get correct results (Clifton, 2004). CRISP-DM consists of six steps for

conducting data-mining projects. As shown in Table 3.3, these steps include: business understanding, data understanding, data preparation, modelling, evaluation and deployment. Therefore, based on the manner that CRISP-DM framework uses to address a problem, it can be said that CRISP-DM is a framework artefact that might have been developed using Design Science Research Methodology (DSRM). This is because DSRM approach aims to define a solution to a business requirement by building an IT artefact (Lapão, da Silva, & Gregório, 2017) of which in this case is CRISP-DM.

Table 3. 3 Similarities between DSRM and CRISP-DM

DSRM Activity	CRISP-DM Phase	Tasks
Identify problem and motivate	Business understanding	Health care Information systems in South Africa are operated in silos, a large portion of these systems are not implementing health standards. Those that do implement cannot share that data because the receiving system would not be able to interpret this data.
Define objectives of a solution	Data understanding	Collect structured and unstructured relevant health data from multiple sources so as to replicate the problem being experienced.
Design and development	Data preparation	Prepare the data processing, design methods that would make the data easily computable through feature selection and vectorization.
Demonstration	Modelling	From the selected features, split data into training and testing set. Build a predictive model from the training set.
Evaluation	Evaluation	Test if the model built can make correct predictions.
Communication	Deployment	Deploy the model on live environment.

In this study the researcher follows CRISP-DM for the application of supervised machine learning algorithms on structured data. However, the same framework was applied on unstructured data, where natural language processing (NLP) techniques and unsupervised machine learning algorithms (e.g. brown clustering) were used. Therefore, data mining and knowledge discovery applications can use CRISP-DM as a guideline for achieving the desired outcome for a given data mining problem. CRISP-DM is considered finished if it solves the relevant problem at hand (Weber, 2010), therefore, the final phase cannot be reached until a model that satisfies the business requirement is built, thereafter it can be deployed. In Table 3.3 the researcher shows the relationship that can be drawn between CRISP-DM and the DSRM approach.

However, it is worth mentioning that there are also differences between the CRISP-DM and DSRM. On DSRM, the iterative process of building is running concurrently with the process of evaluating. While with CRISP-DM the modelling phase would have to be finished before the model could be evaluated. However there is flexibility, because the process could be refined and restarted from business understanding in case the model built is not satisfactorily. This chapter only covers two processes from the CRISP-DM framework as highlighted in Figure 3.3, however the researcher also gives an outline of the modelling phase in section 3.7 of this chapter.

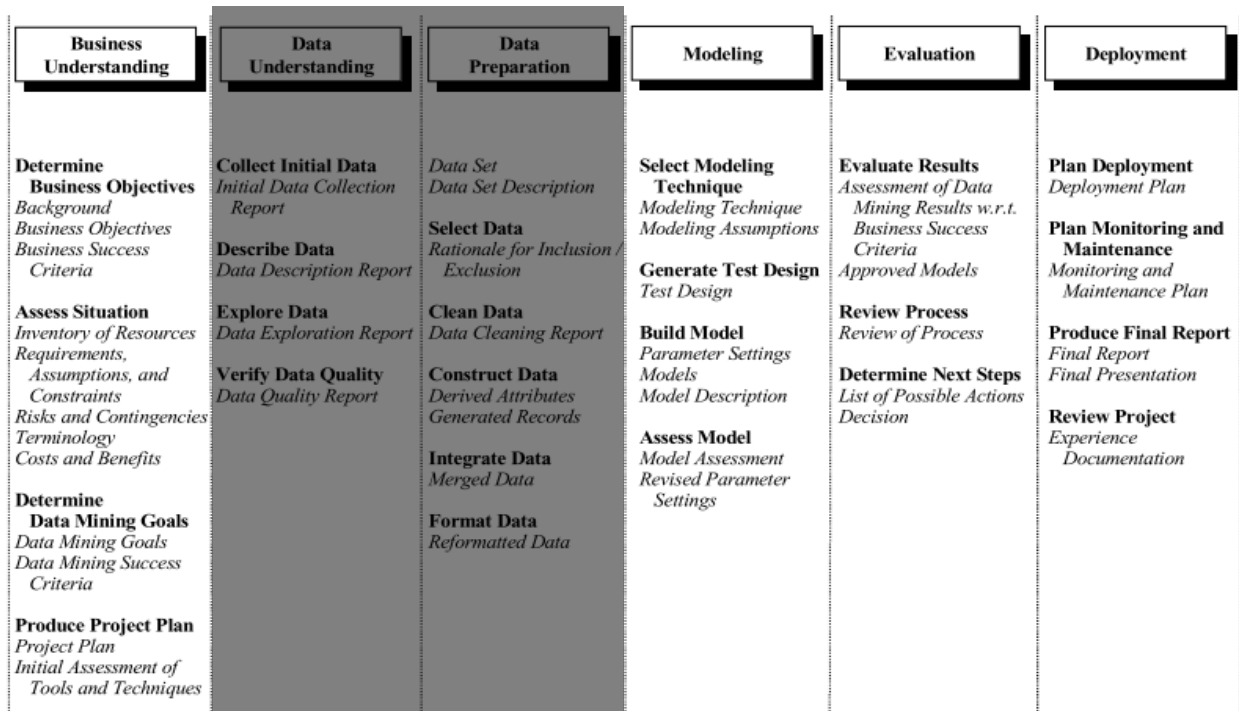


Figure 3. 3: CRISP-DM for data understanding and data preparation (Source: (Olson & Delen, 2008))

3.5 DATA UNDERSTANDING

This section addresses research sub-questions (i) and (ii) of this study:

#	Research sub-question
i.	What type of health-related data sets will this research study focus on?

The problem at hand is that health facilities are not able to exchange health records between themselves, because there is no common coding standard for data management, hence the systems are operated in silos. Therefore, the researcher has proposed a solution that allows a classifier to learn patterns of standardized data (Table 3.4A) so that the generated model can be applied to the unstandardized data (Table 3.4B), and hence learn the factors from standardized data. Standardized data acts as a gold standard that the researcher uses as a base to standardize other data.

Table 3. 4A Standardized tests from LABEVENTS and D_LABITEMS MIMIC tables

LABEL	FLUID	CATEGORY	VALUE	VALUEUOM	LOINC_CODE
Cholesterol, LDL, Calculated	Blood	Chemistry	101	mg/dL	2090-9
Hematocrit	Blood	Hematology	42.8	%	4544-3
Hemoglobin	Blood	Hematology	12.6	g/dL	718-7
Cholesterol, LDL, Measured	Blood	Chemistry	140	mg/dL	18262-6
Cocaine, Urine	Urine	Chemistry	NEG		3397-7
Oxygen Saturation	Blood	Blood Gas	95	%	20564-1
pCO2	Blood	Blood Gas	33	mm Hg	11557-6
Urine Appearance	Urine	Hematology	Cloudy		5767-9
Urine Color	Urine	Hematology	Amber		5778-6

Table 3. 4B Unstandardized observations from CHARTEVENTS and D_ITEMS MIMIC tables

LABEL	FLUID	CATEGORY	VALUE	VALUEUOM	DBSource
Cholesterol		Labs	173	mg/dL	MetaVision
Cholesterol (<200)		Chemistry	252	mg/dl	CareVue
Hematocrit (serum)		Labs	36.9	%	MetaVision
O2 saturation pulseoxymetry		Respiratory	91	%	MetaVision
Mixed Venous O2% Sat		Blood Gases	55		CareVue
SaO2		ABG's		%	CareVue
pCO2		ABG'S	43		CareVue
Urine pH			5	kg	CareVue

The proposed approach is derived from the statistical learning theory. To test the proposal, the researcher has collected data that is standardized based on the LOINC coding standard, as well as data that is not standardized. (Bousquet, 2003) has suggested that initially, sampled data that is used to train the model should be somehow related to the future data (or unseen data) in order to be able to make correct predictions on the new data, otherwise it would not be possible to solve the prediction problem. Therefore, the researcher has collected health data from two databases,

namely MIMIC-III and NHanes. These databases were used to exemplify the lack of interoperability between two disparate systems. This is a similar problem as the one experienced by Health Information Systems (HIS) in South Africa (CSIR & NDoH, 2014). The MIMIC-III database stores the same clinical observation data as the observation dataset on the NHanes database. However, the data in these databases was collected on different setups, MIMIC-III contains data from the hospital, while NHANES contains data from a mobile centre. These databases store this information differently in terms of data values, data types, and attribute names. It should be noted that this study only covers the variety aspect of big data, whereby the researcher looks at two disparate databases. In addition, MIMIC-III also contains unstructured clinical data, and the researcher intends to standardize this data in order to make it easier to query or retrieve, to make it comparable and to make it ready for exchange purposes. The data from MIMIC-III is sourced from two separate information systems namely Philips CareVue Clinical Information System and IMDSoft MetaVision ICU (Johnson et al., 2016). Other details about these systems are covered in the following sections, and throughout this study, these systems will be referred to as CareVue and MetaVision.

3.5.1 DATA SOURCES

- MIMIC-III Databases

This database contains patients' data and not only limited to that, but it also includes laboratory tests, medications, ICD9 diagnoses, admitting notes, discharge summaries and pharmacotherapy, demographics, and a medical history dictionary. This database consists of data collected from the following technologies: Electronic Medical Record (EMR), free text format, medical record, medical coding process document and electronic bill system. This data is not open-source data, however it is accessible to researchers under a data usage agreement (Johnson et al., 2016) and MIMIC-III database is accessible on the <http://mimic.physionet.org> website.

- NHANES

NHANES in full is known as the National Health and Nutrition Examination Survey. This project is meant to assess the health and nutritional status of adults and children in the United States. The survey has been defined to be unique, because it conducts interviews and also

collects data about the patients' physical examination. In addition, the NHANES program is publicly accessible, however there is also restricted data that may be accessed upon request. The laboratory tests for NHANES take place at a mobile examination centre (MEC), and the interviews conducted include demographic, dietary, socioeconomic, and health-related questions (Patel et al., 2016).

3.5.2 DATA EXPLANATION

MIMIC-III structured data tables

There were five database tables that were recognized for the structured data and the list is as follows:

- LABEVENTS

This table records laboratory information for all inpatients and outpatients, there are 27 million records in this table. The table uses eight attributes for recording the data namely subject_id, hadm_id, itemid, charttime, value, valuenum, valueuom and flag. The subject_id is an identifier for the patient and hadm_id is an identifier for patient's stay in hospital and records without a value for this field are meant to represent an outpatient. Then itemid is a foreign key from the D_LABITEMS, which is a code-list for all the observation names contained in the LABEVENTS table. The charttime is the time when the observation was charted, and it is the closest time to when the test was actually taken; then the value is the recorded value for the test and valuenum stores the same value as recorded value attribute provided it is a numeric value. The valueuom attribute is the unit of measure for the test, and then the flag records whether or not the test value is abnormal or not.

- PATIENTS

This table contains 46 520 records for patients whose data is sourced from the MetaVision and CareVue Health Information Systems (HIS). There are seven attributes used to store the data, subject_id is the unique identifier for the patient as mentioned above, there is a gender attribute, and dob which is used for recording the patient's date of birth. Patients whose age is older than 89 have had their date of

birth shifted, with the aim of obscuring their age and hence complying with the HIPPA (Health Insurance Portability and Accountability Act) regulations. Then `dod` is the date of death for the given patient while `dod_hosp` is the date of death as recorded in the hospital database, and then `dod_ssn` is the date of death from the social security database, which is not part of the MIMIC-III database.

- CHARTEVENTS

This table is also sourced from the MetaVision and CareVue HIS, there are 330 million records in ChartEvents table. It mainly contains patient's stay while in the ICU. The table stores information such as vital signs, ventilation settings, mental status, laboratory values, and patients' additional information. Some of the table's attributes are similar to ones mentioned before, `subject_id`, `hadm_id`, `charttime`, `value`, `valuenum`, `valueuom`. Attribute `icustay_id` is a unique identity per patient stay at the ICU, `item_id` is sourced from a different code-list table `D_ITEMS`, then the `storetime` attribute stores the time when the record was manually validated by the member of the clinical staff. The `dgid` stores the unique identifier of the caregiver, then `warning` and `error` are MetaVision specific fields, which record whether a warning for a value was raised, and if an error occurred during a measurement. The CareVue HIS uses `result_status` to determine whether the type of measurement was automatic or manual, and the `stopped` attribute specifies whether the test was stopped or not.

- D_LABITEMS

This is code-list table and sometimes it is referred to as a definition table, `D_LABITEMS` contains 753 unique records about the definition of laboratory tests, data from this table is linked to `D_LABEVENTS` through the `itemid` attribute. The data contained in this table includes data from hospital wards and clinics outside the hospital. There are 585 records that have been standardized and mapped to LOINC and 168 have not been mapped, out the 585 that have been mapped there were 565 active LOINC codes. This table uses: `itemid` as a unique identifier; the `label` attribute represents the observation name; `fluid` attribute stores information about the sample; and the `category` attribute gives information about the type of measurement being done.

- D_ITEMS

This is also a definition table with 12 487 records, the table's data is sourced from the CareVue and MetaVision HISs. It contains itemid which is different from the one in the D_ LABITEMS table however it is used for the same purpose, the label attribute is the same as one described on D_ LABITEMS. There is also an abbreviation attribute, and the dbsource attribute which specifies the data source database name, this is either the hospital, or CareVue or the MetaVision HIS. Then the category attribute from D_ITEMS is used for storing the type of test, and the unitname stores the unit of measure values.

The above listed tables are the main tables used in this study for working with structured data, also the LAB_EVENTS table contains LOINC-standardized data which will be used as the target dataset or the table that defines the gold standard. The mapping for this table was done by a fourth-year medical student and an informatics fellow using the RELMA mapping tool. Then an expert reviewer assessed the mappings made by the student and the informatics fellow (Abhyankar et al., 2012). There are cases where the LAB_EVENTS table would not be sufficient as the target dataset table. In an instance when the data to be mapped to is not available on that table, the researcher would therefore use the LOINC database table. (McDonald et al., 2017) give more information about the structure of the LOINC database table.

The information contained in the mentioned tables was regarded relevant to this study because: it contained information that has been captured from different health systems; the data contained duplicates information, missing values, outliers and more; MetaVision and CareVue HIS do not record the same information based on the same itemid, meaning that one could get a heart rate using itemid of 212 for CareVue, whereas the same test uses a different itemid on the MetaVision system; and the CareVue system has been reported to be the source of duplicates, because some of the data entry allows for free text. Therefore, the fact that this data is not organised in the same order makes it a good candidate for the objective of this study.

MIMIC-III unstructured data table

For the unstructured data the researcher used the NOTEEVENTS table, this table's data is sourced from the hospital database which is different from the CareVue and MetaVision HIS and in total this table contains 2 million records. These records consist of medical reports, ECG reports, social work reports, discharge summaries, respiratory reports, nutritional reports and more unstructured text data. The NOTE_EVENTS table has 8 database attributes namely subject_id, hadm_id, chartdate, category, description, cgid, iserror, and text. The rest of these attributes store the same type of values as indicated above, however category and description define the type of note recorded, for instance a category could be "nutritional" and the description could be the "summary". Then the iserror attribute is used to indicate that the physician has identified an error on the clinical note, while the text attribute contains the actual patient's note in a textual format compiled by a nurse or a clinician. According to (Pustejovsky & Stubbs, 2013) the data contained in the text fields is referred to as corpora, and once a single note from this set is annotated then the annotated one is then referred to as the corpus, therefore in this study this type of data will be referred to as such. The researcher had sampled 195 unique records based on the subject_id, these records were filtered by the "discharge summary" category and by whether they contained behavioural data such as the patient's smoking status. Additional filters were applied to exclude: deceased patients, patients younger than 18 years of age, and to exclude records with a true flag for the iserror attribute. The researcher ensured that the retrieved results for all the queries are unique based on every sample that was selected, the uniqueness of a record was based on the subject_id which is unique per patient on the MIMIC-III database.

NHANES dataset

The researcher has collected the 2011-2012 NHANES publicly accessible health data. This data consisted of 9338 examined participants, however the data that is of interest to this study is the laboratory and physical examination data. The NHANES data uses data that is stored in multiple datasets, for instance, data about the participants' age and gender is stored on the demographic dataset which is separate from the laboratory

dataset and the physical examination dataset. There were 860 unique observation names that have been identified for the purpose of this study. The data was arranged based on the test names, for example the cholesterol HDL observation had 7821 records, and each record represents the number of participants for that test. Environmental data such as the presence of lead in the blood is included in the lab data. Capturing environmental data is part of the effort to store patient's living conditions. When it comes to the physical examination data, the researcher only covered blood pressure and body measures surveys with results. The blood pressure data file had 27 variables, however, variables that captured comments were excluded from the rest of the data. The dataset consisted of the heart rate, radial pulse, and blood pressure measurements. The NHANES data is in line with this study, because it captures more details about an observation, wherein with this type of data, the researcher will be able to apply a supervised classification algorithm in order to learn how to standardize laboratory observations using a coding standard.

3.6 DATA PREPARATION

One of the underlying steps that should be carried out before feature selection commences is data pre-processing, where data from the identified datasets is cleaned of errors, duplicates are removed, and outliers are identified. As a matter of fact, (Doan, Halevy, Ives, et al., 2012) have suggested that it is useful to perform feature standardization before applying similarity measures between the source and the target dataset values. Therefore, the researcher centralized the data to be pre-processing by firstly loading it from the flat file format into a Postgre SQL 9.3 database. More information about how to load the MIMIC-III database can be found on the following web address <http://mimic.physionet.org>.

3.6.1 DATA PRE-PROCESSING FOR STRUCTURED DATA

Since MIMIC-III database is a relational database, data is stored in different related tables, however in order to create value from the identified tables one needed to join the data through SQL join statements. The D_ITEMS table was joined with the

CHARTEVENTS using the item_id, and also the PATIENT table was joined to the CHARTEVENTS table through the subject_id and this join was treated as the source dataset. The resulting dataset contained more than 300 million records, the researcher randomly took 50 000 records and also applied filters so that only CareVue and MetaVision data was retrieved, also filtered patients whose age was less than 18 and those that were diseased. These filters were done specifically for MIMIC-III, however some were also applicable to the NHANES database. For MIMIC-III the researcher filtered out data that contained different categories other than: respiratory, routine vital signs, hemodynamics, laboratory data, cardiovascular (pacer data), and the general category. From MetaVision, there were 27402 records, where the same setup was applied on the CareVue HIS, but without filtering categories because only 7% of the CareVue data had a category value specified. There were also laboratory observations such as AST which needed to be expanded in order to make sense of the acronym.

Abbreviation expansion

Abbreviation expansion is a technique used for identifying corresponding and relevant long forms of an abbreviation. In this context, abbreviations also cover acronyms, and therefore for the duration of this study, the researcher will use abbreviations to represent both terms. The abbreviation's long form can be illustrated as follows: "DOD" is an abbreviation of "Date of death" which is its long form. In addition, (Moffat et al., 2008) have noted that abbreviations often cover multiple long forms, which makes it difficult to identify the relevant long form, e.g. the short form "DOD", which could be expanded to "Department of Defence", or "dead of disease", or "date of discharge". The representation of abbreviations is sometimes confusing and unclear, and therefore, (Hill et al., 2008) have devised a method of handling abbreviations using regular expressions for various patterns (see Table 3.5).

Table 3. 5 Mechanisms for mining abbreviation expansions

Pattern	Regular expression	Short form	Long form
Acronym	$c_0[a - z] + c_1[a - z] + \dots + [a - z] + c_n$	ICU	I ntensive C are U nit
Prefix	$sf[a - z] +$	Lab	L aboratory
Dropped letter	$c_0[a - z] c_1[a - z] \dots [a - z] c_n$	Dept	D epartment
Combined word	$c_0[a - z] ? c_1[a - z] ? \dots [a - z] ? c_n$	Rcpt	R eceptor

Other standardization procedures

The dataset that the researcher is using contains data that is arranged based on international localization and globalization standards, and therefore, some of the data needs to be standardized to a South African localization and globalization standard. The following labels have been identified for localization.

- Dates: US format is *mm/dd/yyyy*, and the South African format is *yyyy/mm/dd*.
- Temperature values: the standard temperature unit in South Africa is Celsius (C) therefore any unit that is in Fahrenheit will be converted to Celsius.
- Units of measurement: pounds are converted to kilograms.

All the dataset attributes were converted into lower cases for both the source and the target attributes. Observation test names that start with any alphanumeric characters were normalized by removing the pretext, for instance, tests such as “% hemoglobin a1c”, the leading “%” sign was removed. However based on the guide stipulated by (Regenstrief Institute, 2016), the pretext “%” provided more information about the unit of measure for the test. This showed that other health organizations store laboratory observation names with the unit of measure in one field. This was similar to the NHANES observation names, for instance, an observation name would be structured in this format (*albumin, ser*), which provided an indication that the same field is also used to store the observation name (albumin) and the sample (serum).

It was noted that this would cause data reading issues, because the researcher would pre-process the data and store it on a csv file format, and because of the comma an extra column would be added, causing the record to lose its structure. Therefore, all the observation names were pre-processed so that commas are replaced by underscores.

It was also mentioned previously that the NHANES used multiple separate files to store data, therefore the researcher created joins that joined the demographic dataset with both laboratory dataset and physical examination dataset.

3.6.2 DATA PRE-PROCESSING FOR UNSTRUCTURED DATA

The unstructured data from the NOTEEVENTS table is purely textual data, and the structure of the data is for human-readability purposes only because it is not organized in a computer-readable format such as XML or JSON. However, indentation, spacing and letter capitalization was used to format the contents of these files in order to indicate section headers. Part of the contents contained in the files is categorized by allergies, major procedure, history of present illness, and past medical history just to mention a few. From the Postgre database, the researcher ran an SQL query to filter the results from the NOTEEVENTS table, there were 288 unique records that were sampled based on the subject_id, these records were filtered by the “discharge summary” category. The researcher further filtered the records by selecting only records that contained living patients, patients older than 18 years of age, and records with a false flag for the iserror attribute. Then each of the selected records was saved into a separate text file renamed by the subject_id. It is worth mentioning that the files were initially uploaded to Postgre database table for easier searching capabilities, otherwise the researcher would have to manually search file-by-file in order to find the relevant content. Once the files were saved, then 80% of the files were loaded into the CLAMP training corpus folder, this folder already had 388 pre-annotated clinical notes as per (Soysal et al., 2017). The remaining 20% from the selected files were loaded into the test corpus. Thereafter the text was transformed into lower cases, it was tokenized, and stemming was also applied. According to (Manning et al., 2009) the input below is referred to as the document in NLP, so therefore it will be referred to as such

throughout this study. Below are the examples of input text where lower-casing, tokenization and stemming were applied:

Input: The patient denies smoking.

Function: transform cases

Output: the patient denies smoking

The transform cases function converts upper case into lower cases if it is specified as such, this ensures that document similarity function compares documents that are based on the same case style.

Function: tokenization

Output:

the	Patient	Denies	Smoking
-----	---------	--------	---------

The tokenization function breaks the document down into tokens, (Manning et al., 2009) defines a token as an instance of a sequence of characters from a document that are useful for understanding the building blocks of the document for further processing. One can choose a breaking point for documents, where a regular expression can be used so that words are broken based on the rules defined, whereas a common method breaks the document based on spaces found between words on a document.

Function: Stemming and lemmatization

Output:

Patient	banana	supplier	Deni	Smoke
---------	--------	----------	------	-------

Stemming is used for removing multiple derivations of words, e.g. smoke has multiple grammatical forms such as smoked, smoking, smokes and others. Therefore, stemming reduces the word to its base form. Lemmatization performs the same function as stemming, however lemmatization reduces the word into its

canonical form (Biba & Xhafa, 2011), unlike stemming which reduces to an extent where it removes meaning. An example of this is the stemming of the word *saw*, stemming would produce *s* as an output while lemmatization would produce a lemma such as *see* or *saw* based on grammatical meaning on a document.

These are few of the well-known Natural Language Processing (NLP) pre-processing techniques. The CLAMP software comes with dictionaries for identifying temporal features (such as dates) through the temporal recognizer, and negation keywords (such as “not”) in the following statement “patient does not drink”, an assertion classifier was used for negation detection in clinical statements. However, the dictionaries were limited in word coverage, therefore the researcher modified the dictionaries and added more words for both temporal and negation detection functions.

3.7 NOTATION USED

This is the common notation and this section is only meant to be a guideline for understanding symbolism that is used throughout this study.

$x^{(i)}$: Input i^{th} variable or feature

$y^{(i)}$: Target or output i^{th} variable

X, Y : Training example

m : Number of training examples

n : Number of training features

$(x^{(i)}, y^{(i)})$: i^{th} training example

$h^{(i)}$: i^{th} Hypothesis function (Maps input feature to output feature)

θ : Parameter of the model

p : Length of projection between vectors

C : Regularisation constant

\wedge : Conjunctions (Logical AND)

\vee : Disjunctions (logical OR)

s : Similarity value

Term^T: Transposed term or parameter

$l^{(i)}$: l^{th} landmark

σ^2 : Gaussian Kernel parameter

3.8 CLASSIFIERS AND PROBABILISTIC GRAPHICAL MODELS USED

There are two types of classifiers that will be employed to model structured data in this study; a rule-based (such as Decision Trees) and a kernel-based classifier (such as SVM or Logistic Regression). However, for the task of modelling unstructured data for semantic and standardization purposes a different method is used which is probabilistic graphical modelling. Therefore, in this section the researcher gives details about these methods.

Logistic Regression

Logistic Regression is a classification algorithm that is used for classifying data into discrete classes, this is different from linear regression which attempts to fit a straight line to the training data. With Logistic Regression one could perform a binary classification which outputs a binary output (y) and $y \in \{0, 1\}$, or in cases of a multiclass classification it outputs $y \in \{0, 1, \dots, n\}$ classes (Ng, 2000; Ng & Jordan, 2002). An example of problems that have been solved using a Logistic Regression classifier are as follows: an email spam classification where the input are words on an email message, and the classifier has to predict whether the email is spam or not spam; a loan application problem whereby the classifier receive as input details about the applicant's spending behaviour, and the classifier predicts a binary value of whether to give or not to give a loan; another example is the prediction of whether a tumour is benign or malignant, this is classified from an input of an electronic radiograph image. The Logistic Regression classifier uses sigmoid function as shown in Figure 3.4, and its hypothesis function is shown in Equation (8).

$$h_{\theta}(x) = \frac{1}{1+e^{-\theta^T x}} \quad (8)$$

g(z)

- z +
Figure 3. 4: Sigmoid function or logistic function

The sigmoid function is also known as the logistic function and it asymptotes at value “0” and value “1”, when the value of z approaches negative infinity it can be observed that the sigmoid function $g(z)$ is less than “0.5”, when the value of z approaches positive infinity, then the sigmoid function $g(z)$ becomes greater than “0.5” Figure 3.4. With this said, Logistic Regression predicts $y = 0$ when the hypothesis function is $h_{\theta}(x) < 0.5$, and $y = 1$ when it is $h_{\theta}(x) \geq 0.5$ (Ng, 2000). Therefore, when the hypothesis outputs a value of “0.8”, it is interpreted as that there is an 80% probability that the evaluated condition is true. However, one of the most important steps is the calculation of the cost function J , the cost function J measures how close the predicted hypothesis $h_{\theta}(x)$ is from the corresponding given output y value as shown in Equation (9).

$$J(\theta) = \frac{1}{2} \sum_{i=1}^m (h_{\theta}(x^{(i)}) - y^{(i)})^2 \quad (9)$$

Now the cost function should be minimized in order to get an accurate output, and it is minimized by applying a batch gradient descent whereby the parameter θ is simultaneously updated so as to get an optimized parameter value. Logistic Regression is also known as a discriminative classifier, and (Ng & Jordan, 2002) have found that these types of classifiers outperform the generative ones such as Naïve Bayes. Another discriminative classifier is Support Vector Machines (SVMs) which is covered in the following sub-section.

Support Vector Machines

SVM is a classifier that attempts to find an optimal hyperplane to separate positive training examples from the negative ones. This classifier is built on the principle of Structural Risk Minimisation (SRM), where positive and negative training examples are separated by a hyperplane and SRM helps maximise the margin between the hyperplane and the training examples (Nasien et al., 2010). SRM is comparative to Artificial Neural Network's (ANN) empirical risk minimisation principle, and in addition to that (Olson & Delen, 2008) have reported in favour of SVM as: it is less prone to overfitting; it always finds the global minimum; and the complexity of the SVM's model is not controlled by keeping the size of the features small as with ANN. However, SVM minimization function is similar to that of Logistic Regression, which is a classifier that outputs a probability, in contrast SVM outputs a prediction of either "1" or "0".

The SVM separating planes can be seen from the equations given below, Equation (10) is the central hyperplane, and in Figure 3.5 it is represented by the solid blue line. Equation (11) represents the margin from the central hyperplane to the positive plane (dotted), in other words the value of "1" in Equation (11) represents a threshold that should be met for the classification of positive training examples. For illustration purposes, positive examples can be viewed as the orange dots in Figure 3.5, while negative examples are represented by the black squares, and in terms of the equations this is shown by Equation (12) which are training examples with negative classes.

$$\theta^T x + b = 0 \tag{10}$$

$$\theta^T x = +1 \tag{11}$$

$$\theta^T x = -1 \tag{12}$$

The value of x in these equations is the input feature, whereas the value of θ is the distance from the central hyperplane to the support vectors which are training examples that lie closer to the hyperplane, and in Figure 3.5 they are shown in a yellow colour. Then b is the bias which controls the displacement of the hyperplane from the origin point.

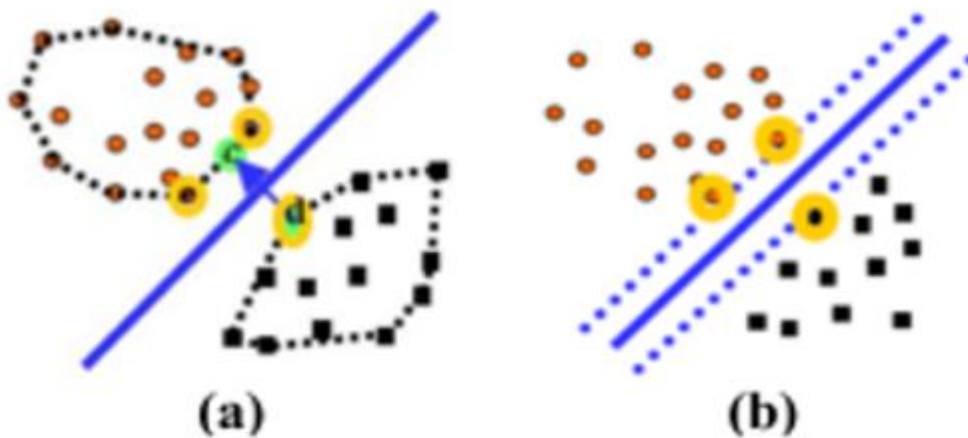


Figure 3. 5: SVM decision boundaries (Source: (Nasien et al., 2010))

One can observe from Figure 3.5 that the training data is linearly separable, however there are other instances where the training set is not linearly separable (see first graph from Figure 3.6). In such cases, kernels are used to transform the input space into feature space as shown in Figure 3.6, this arranges the training examples in a manner that is easily understandable by the classifier (Harrington, 2012). According to (Kumar, 2015), kernels are suitable for classification tasks when the number of training

examples are in the range ($10 \leq m \leq 10\,000$) and when the number of features are ($1 \leq n \leq 1000$).

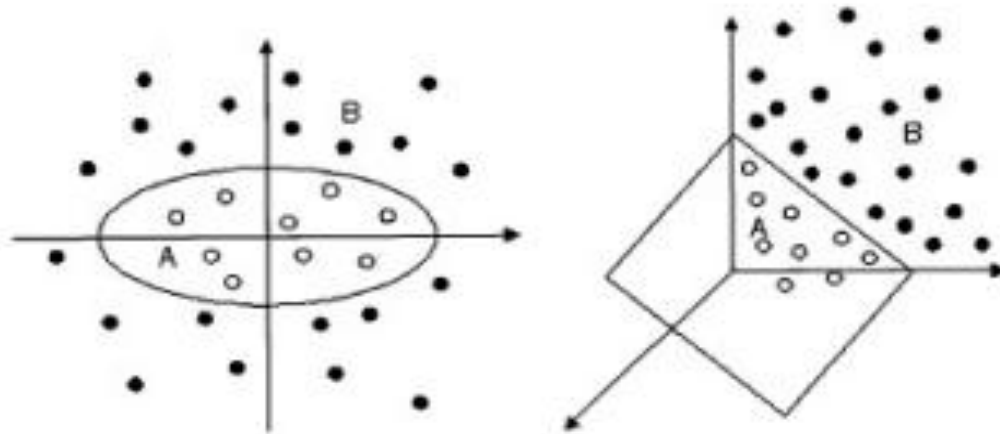


Figure 3. 6: Mapping input space to feature space (Source:(Hofmann, 2006))

By definition kernels are known as similarity measures, they provide a functionality for calculating the similarity between a high dimensional input feature $\varphi(x)$ and the new input represented as x feature. The phi (φ) function is useful for mapping the original data attribute to a high dimension feature. Below are the two of the kernel similarity measures that can be implemented with nonlinear SVM, Equation (13) represents a polynomial kernel which allows for features to be constructed in a joint format up to the order of polynomial (such as quadratic, or cubic order). Then Equation (14) is a radial basis function (RBF) kernel is also known as Gaussian kernel, it is one of the well-known kernel functions and it maps the data into an infinite dimensional space (Manning et al., 2009).

$$K(x_i, x_j) = (\gamma x_i^T x_j + r)^d, \gamma > 0 \quad (13)$$

$$K(x_i, x_j) = \exp(-\gamma \|x_i - x_j\|^2), \gamma > 0 \quad (14)$$

One of the differences between polynomial kernel and RBF kernel is that polynomial has more hyperparameters as compared to RBF, with RBF there are two important parameters, gamma (γ) and the penalty or cost parameter C . More details about these parameters and how they are optimized is covered in section 4.5.

Decision Trees

The Decision Trees are one of the most well-known data mining techniques, they can be used for data regression or data classification or even both through Classification and Regression Trees (CART). CART is one of the algorithms used for implementing Decision Trees, there are other algorithms such as ID3, C4.5, CHAID and more which are built for decision tree implementations. (Mitchell, 1997) who is an expert in Machine Learning have defined decision tree learning as one of the practical methods for inductive inference, in support of that statement (Gorunescu, 2011) said a decision tree is built through an inductive process called “tree induction”. Decision Trees offer more benefits because they are easy to implement, they help define rules that are governing the dominant attributes in a dataset, and they can be easily visualized. They are also easy to convert into a set of rules, the tree can be generated from the training set, and each tree node represents a condition that tests a rule. The leaf nodes are a possible outcome of the rule, whether the condition is true or false (Christen, 2012; Doan, Halevy, & Ives, 2012).

Rules are expressed as the testing of a condition, which yields a certain conclusion, following the expression as show below:

IF condition THEN conclusion

From Figure 3.7 it can be seen that the dataset consists of three variables, two input variables and one output variable, the “Age” represents continuous values, “Car type” is categorical value and “Accident risk” is a binary value.

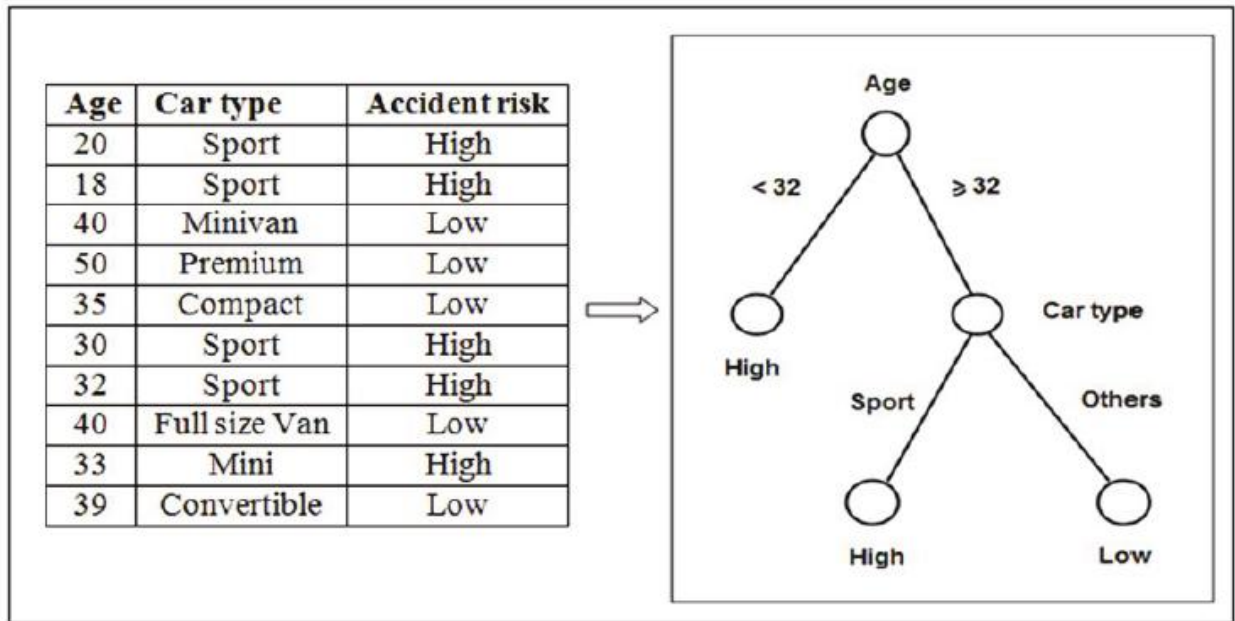


Figure 3. 7: Risk prediction based on the type of car and the driver’s age (Source: (Gorunescu, 2011))

The tree in Figure 3.7 has conditions that checks if the “Age < 32”, if this is true then the risk is high, whereas if “Age ≥ 32” and “Car Type = Other” then the risk is considered low. One can observe that “Age” is at the root of the tree, which means that it is the splitting attribute, as mentioned above that Decision Trees help identify the dominant attributes. There are various methods that can be used for achieving the splitting criterion, (Gorunescu, 2011) lists some of the few methods:

- GINI INDEX: It is an impurity measure often used with the CART algorithm, and it measures the frequency of a randomly selected attribute from the training set that could be incorrectly labelled if it was randomly labelled according to the distribution of labels in the dataset.
- Information Gain: It uses the concept of an entropy for deciding on which feature to split at during each step of building the tree, it is mainly implemented by ID3, C4.5 and C5.0.
- Chi-square measure: It is a statistical hypothesis test that is commonly used in inferential statistics, this method tests for the goodness of fit on an observed distribution, and it is also commonly used with CHAID trees.

Up to so far the machine learning algorithms that were described are used for the classification of structured data, and these algorithms fall short in extracting relational clinical concepts and identifying related sequences in the clinical text. A study by (Li, Kipper-Schuler, & Savova, 2008) has found that Conditional Random Fields (CRFs) outperform SVM for named-entity recognition tasks. This type of classifier is specifically designed for identifying sequences in various forms of data, hence this classifier is called a sequence classifier.

Conditional Random Fields

Conditional Random Fields (CRFs) is a task-specific type of a probabilistic graphical modelling framework, it is used for classifying sequential data through segmentation and annotation. CRFs trains a model discriminatively, a discriminative model learns to make a conditional prediction of a class (or hidden state) from the given features (or observable states) and it is represented as follows: $P(Y|X)$. Other than that, there are generative models which learn the features (or observable states) that would result in predictions that favour the given class (or hidden state). Examples of classifiers that apply generative models are Naïve Bayes and Hidden Markov Models (HMMs) (Sutton & Mccallum, 2011). (Ng & Jordan, 2002) have found that generative-based classifiers are easier to implement and give a good performance on a small training set, however on a larger training set, the discriminative models are preferred because they provide a better performance. It is also worth mentioning that CRFs are an extension of a generative HMM and the discriminative Maximum Entropy Markov Models (MEMMs), because CRFs are a type of a graphical model, which also has observable states X , hidden states Y and state transitions (that is edges between the hidden states) as shown in Figure 3.8.

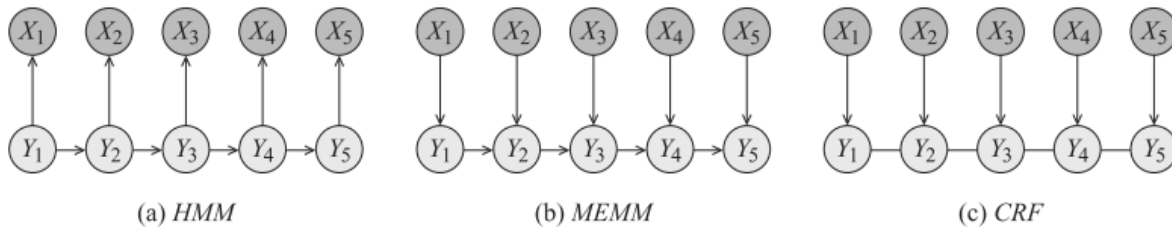


Figure 3. 8: Hidden Markov Model graph for estimating the atmospheric pressure (Source: (Koller & Friedman, 2009))

The first model of Figure 3.8 represents HMM, HMM is a probabilistic finite state machine that consists of observable and hidden states, state transitions, observation symbol and initial state. HMMs have been used previously for a speech recognition task and in natural language processing tasks such as part-of-speech tagging, Named-Entity recognition (NER), and chunking (Christen, 2012; Marszalek, 2009; Ponomareva, Rosso, Pla, & Molina, 2007). However, HMMs employ a direct graphical model which means that they are tied to a linear sequence structure, and as thus (Xing, 2007) has reported that HMMs have a dependency weakness. They fail to capture related items from the given input. To illustrate this point, Table 3.5 gives an example of a given clinical note as input, and tokens extracted from the input and the part-of-speech (POS) tokens, the purpose here is to extract entities and relationships between these entities.

Table 3. 6 Example about relation extraction to showcase the shortcomings of HMM

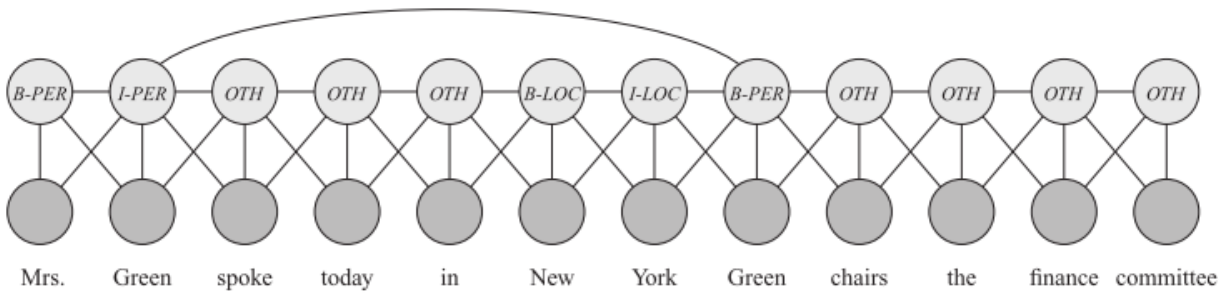
Input	Laboratory data revealed Hematocrit of value 32.4.							
Tokens	Laboratory	Data	Revealed	Hematocrit	of	value	32.4	.
POS	NN	NNS	VBD	NN	IN	NN	CD	.

Since HMMs only capture dependencies between a specific hidden state and its observed state, therefore it would fail to see:

- That “Hematocrit” and Laboratory are capitalized

- The end of the sentence where there is a punctuation mark “.”
- That “32.4” is the decimal value, and that it is a value of “Hematocrit”

These are few of HMMs limitations, however, all the limitations of HMMs are addressed by MEMM, hence MEMMs provides the freedom of choosing features for representing the observable states. MEMMs is modelled discriminatively unlike HMMs, and it uses a conditional probability to predict the state sequence from the observable state sequences (Siddiqi, Alam, Hong, Khan, & Choo, 2016). In addition, MEMM share the same applications with HMM, also (Siddiqi, Alam, Hong, Khan, & Choo, 2016) have further applied it in human facial expression detection. Although MEMM is better than HMM, MEMMs also suffer from a label-bias problem, and (Sutton & McCallum, 2011) have described it as the inability for future observable state to provide information about the currently observable state, and (Koller & Friedman, 2009) have described it as the failure of the model to go back and change its predictions about the first few observable states. Therefore, CRF has capabilities that address the label-bias problem experienced by MEMM, although CRFs were defined in the beginning of this sub-section, the researcher will further add more details about this framework. CRFs are modelled as undirected graphs and are used in applications similar to those mentioned for HMM and MEMM. In this study, the researcher has aimed to use CRF for extracting clinical entities such as: the smoking status and the negation status. The CRF framework makes the aim possible to achieve because it employs the “BIO notation” whereby the “B” indicates the beginning of the named-entity phrase, “I” indicates the inside or the end of the named-entity phrase and “O” is other, which indicates that the word is not part of the named-entities (Koller & Friedman, 2009), see Figure 3.9 for use of the “BIO notation”.



KEY

<i>B-PER</i>	Begin person name	<i>I-LOC</i>	Within location name
<i>I-PER</i>	Within person name	<i>OTH</i>	Not an entity
<i>B-LOC</i>	Begin location name		

Figure 3. 9: A linear chain CRF model showing observable states shown as Grey nodes, and hidden states shown as clear nodes (Source: (Koller & Friedman, 2009))

3.9 TOOLS AND DATABASES

Model creation for structured data

The researcher has used a free version of RapidMiner studio for creating the models, and for visualisation of the training data (see Appendix B). RapidMiner is a software platform for data science teams that unites data through data preparatory processes, which allows for the application of machine learning, and the predictive model deployment. The researcher used version 7.6.001 of the software. However, prior to model creation, similarity weights were calculated using Jaro-Winkler, Edit distance, and functions from MATLAB.

Unstructured data annotation methods

Unstructured data is said to be difficult to search, summarize, and to apply in decision support systems. This difficulty is fuelled by the data having been captured on a free-text basis, and this data is prone to spelling mistakes where in health care this data is captured by multiple health personnel which increases the number of mistakes. Therefore, the researcher has assessed the usefulness of the UIMA component for the

purpose of this study. UIMA in full is called Unstructured Information Management Architecture, it is an open source framework that was originally developed by IBM for processing text, sound and video. For text, UIMA uses analysis engines in order to annotate documents. The user or implementer defines these engines through a type system by using a structure for a possible markup, and this markup in turn helps to achieve interoperability (Wu et al., 2013). UIMA is scalable and extensible, and could be used for processing any type of document, IBM has used it to showcase this framework's ability to understand complex natural language questions on the Jeopardy competition, and giving correct answers from the Wikipedia corpus (Pablo, 2014). However, in the case of this study, the researcher wants to use this framework for processing unstructured clinical data.

Two systems which use UIMA as an underlying framework are cTakes and CLAMP. cTakes is also known as clinical Text Analysis and Knowledge Extraction System, while CLAMP is Clinical Language Annotation, Modelling and Processing. The cTakes system has been defined as an open source system that helps discover codable entities, events, properties and relations. Figure 3.10 gives an overview of the features that are found in cTakes, observe the type the input that cTakes receives below.

Input: Fx of obesity but no fx of coronary artery diseases

Tokenizer output – 11 tokens found:

Fx of obesity but no fx of coronary artery diseases .

Normalizer output:

Fx of obesity but no fx of coronary artery disease .

Part-of-speech tagger output:

Fx of obesity but no fx of coronary artery diseases .
 NN IN NN CC DT NN IN JJ NN NNS .

Shallow parser output:

Fx of obesity but no fx of coronary artery diseases .
 NP PP (NP) (NP) PP (coronary artery diseases) NP

Named Entity Recognition – 5 Named Entities found:

Fx of obesity but no fx of coronary artery diseases .
obesity (type=diseases/disorders, UMLS CUI=C0028754, SNOMED-CT codes=308124008 and 5476005)
coronary artery diseases (type=diseases/disorders, CUI=C0010054, SNOMED-CT=8957000)
coronary artery (type=anatomy, CUI(s) and SNOMED-CT codes assigned)
artery (type=anatomy, CUI(s) and SNOMED-CT codes assigned)
diseases (type=diseases/disorders, CUI = C0010054)

Status and Negation attributes assigned to Named Entities:

Fx of obesity but no fx of coronary artery diseases .
obesity (status = family_history_of; negation = not_negated)
coronary artery diseases (status = family_history_of, negation = is_negated)

Figure 3. 10: cTakes processing of a clinical text document (Source: Savova et al. (2010))

The input shown above is an excerpt of a large clinical note file, so one of the functions in cTakes is the sentence boundary detector. This function detects the beginning and the end of a sentence, then the tokenizer has two sub-functions, firstly it breaks the sentences into tokens that can be analysed further, then it merges tokens in order to create date, fraction, measurement, person title, range, roman numerals, and time-based tokens. The normalizer also produces tokens, but now based on punctuation, spelling variants, stop words, and symbols just to mention a few. Part of speech (POS) functionality detects the type of grammar used on the text data, it assigns tags of tokens such as “patient” to a noun tag, then the shallow parser or chunker is used for tagging noun phrases, verb phrases and more. The Named Entity Recognition (NER) extracts entities from the given text through rule-based techniques and machine learning, this is one of the most important functions because it is a building block for understanding the semantics of a language (Savova et al., 2010).

There are also other useful functions that one can use within cTakes. Apart from the initial functions, (Garla et al., 2011) have extended the functionality of cTakes by introducing YTEX which is also an open source component built on top of cTakes and UIMA. The component was aimed at improving and simplifying feature extraction and applying the latest Negex algorithm for detecting negation (which determines if a medical condition exists or not) in a clinical note said (Mehrabi et al., 2015). YTEX also stores annotations to a relational database using DBConsumer analysis engine. In addition, the functionality of cTakes is similar to that of CLAMP, however CLAMP has a distinct functionality to disambiguate and reorganise abbreviations in clinical text. CLAMP provides a graphic user interface (GUI) which simplifies the process of annotating clinical text, and the output from annotating is a UMLS Concept Unique Identifier (CUI) code that can be used to map to a coding standard such as LOINC, RXNorm or SNOMED-CT.

Medical tools and databases

The UMLS database was used for searching for acceptable medical terms so that they can be used to replace the ones that are abbreviated, and those that are incorrectly written from the MIMIC-III and NHanes databases. Additionally, the UMLS database is used for integrating and distributing key terminologies, find related medical terms, classifications and coding standards in order to promote the creation of more effective and interoperable biomedical information systems and services. UMLS is a non-fee service, although its users are required to fill in an annual report on how they use the service (Hassanpour & Langlotz, 2016). The researcher will thus use these tools and data in order to load, clean the dataset and make it compatible to acceptable medical terms.

3.10 ETHICAL CLEARANCE

The researcher has applied for ethical clearance before conducting this study, and the application was approved with a (040/MN/2017/CSET_SOC) reference number (see Appendix A-1). In preparation for the use of the MIMIC-III database the researcher had to complete a prerequisite course called *data and or specimen research*, after which a certificate from the CITI programme (<http://www.citiprogram.org>) was obtained under the affiliation Massachusetts Institute of Technology Affiliates (ID: 1912), the certificate is attached in Appendix A-2.

3.11 CONCLUSION

This chapter introduced the methodology that was followed in conducting this study. Firstly, a theoretical perspective was given in section 3.2 where the idea about SVM was introduced and how the SVM classifier works. The researcher then showed how the research questions were generated based on the objective of the study in section 3.3.

In section 3.4, the researcher spoke about the relationship between CRISP-DM and DSRM, and section 3.5 covered data understanding. Section 3.6 listed methods for data preparation, where the researcher dealt with abbreviations and compound nouns used

in defining data objects. Section 3.7 listed the notations used in this study, 3.8 covered data classification methods. Section 3.9 covered clinical tools and medical thesaurus to be used in this study. Then lastly, section 3.10 presented information on ethical clearance required for the use of clinical data.

CHAPTER 4:

Data Modelling

4. MODELLING

4.1 INTRODUCTION

In this chapter, the researcher aims to address the practical aspects of model design, where supervised classification methods are used for data modelling. The previous chapter focused more on the theoretical understanding of statistical learning theory, and how can it be applied to address the researcher's proposed solution to the research problem. The researcher starts this chapter by feature engineering from the collected raw datasets in section 4.2, whereas section 4.3 addresses feature selection methods for structured data. Thereafter in section 4.4 the researcher shows how features are selected for unstructured data. In section 4.5 model selection is covered, and thereafter in section 4.6 the researcher addresses environmental setup for experiments, and how the performance of the model will be measured. Figure 4.1 has highlighted what will be covered on this chapter based on the CRISP-DM model.

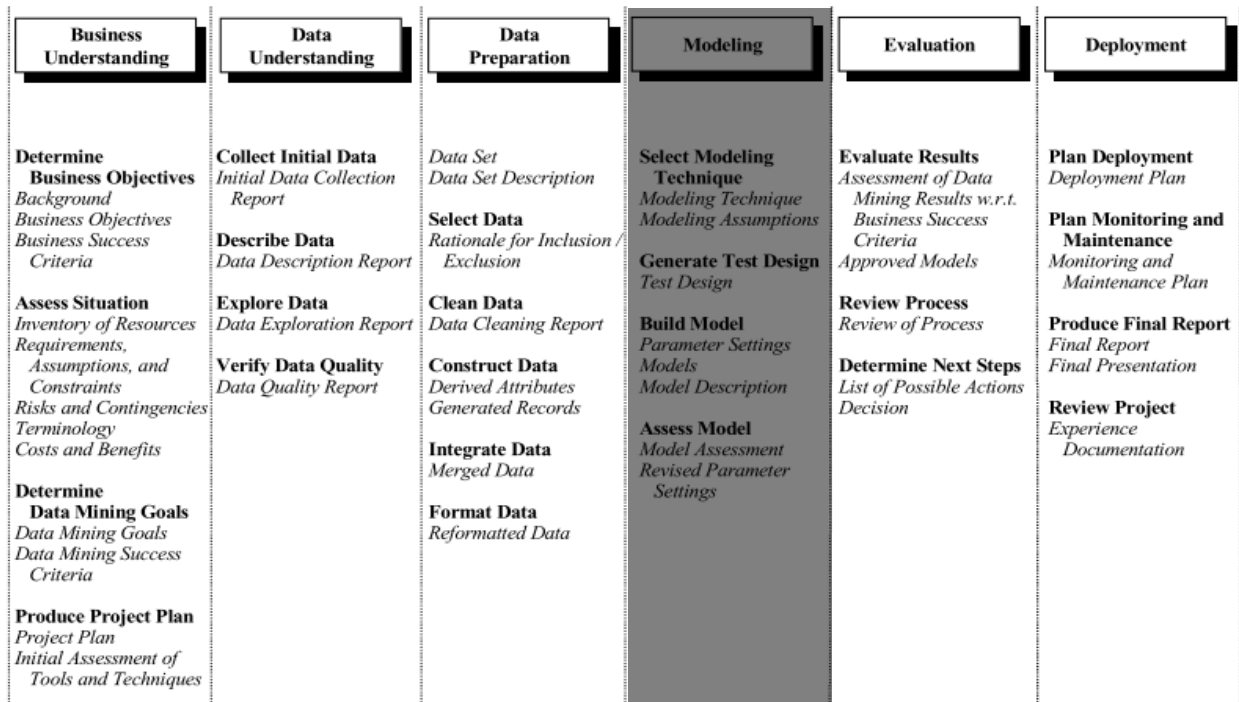


Figure 4. 1: CRISP-DM framework for model selection (Source: (Olson & Delen, 2008))

4.2 FEATURE ENGINEERING AND SELECTION

#	Research sub-question
lii	How were features selected for structured data?

Structured and semi-structured data

Features were extracted and selected from dataset attributes. A clear view of the observation attributes are shown in Table 4.1, which gives an overview about how source dataset is structured. Not all datasets are structured in this format, even though Table 4.1 represents real database objects. At this moment, it is only used for illustration purposes. From these source attributes, the researcher manually selects features that would have a high impact when comparing attributes with those on the target dataset.

Now, when developing a model, it is common to use few and significant features for predicting a phenomenon (Holzinger, 2016). Also in this study, the researcher first selects a few features and then tests their influence on the predictions. The selected features include the “observation name”, “category” and “uom” for both the source and the target database.

Table 4. 1 MIMIC-III Source Observation Dataset

Observation name	Category	UoM	Charttime	Value	Flag
Respiratory Rate	Alarms	BPM	2106-03-02 03:00:00.000	32	
Arterial Blood Pressure diastolic	Routine Vital Signs	mmHg		200	
Arterial Blood Pressure mean	Routine Vital Signs	mmHg			
Alkaline Phosphate	Labs	IU/L	2175-07-24 08:00:00.000	106	
SpO2		%			
ALT	Labs	IU/L			
Anion Gap	Chemistry	mEq/L	2104-08-08 04:15:00.000	28	abnormal
Fingerstick Glucose	Chemistry				
Gentamicin (Trough)	Labs				
Glucose	Chemistry	mg/dL	2134-10-01 14:50:00.000	21	Normal
AST	Labs	IU/L			

The “observation name” and “uom” were expanded in order to gather meaningful comparisons, for instance “BP” was expanded to “blood pressure”, and “mmHg” to “millimetres of mercury” unit of measure. Thereafter, the researcher applied a blocking strategy using the “observation name” and the “uom” as the blocking keys in order to minimise the number of comparisons between the source data and the target data. Blocking was part of data normalization, and details about it were covered in section 3.5. The researcher used the *soundex* algorithm for blocking purposes, where *soundex* uses the sound of words to generate a code that can be used to identify a word, e.g. “activity” and “activated” would be assigned the same code.

Table 4. 2A Source features with soundex blocking keys

RecId	SourceObservation name	SourceUOM	SourceUOMFull	SourceCat	SNDX(SLName)	SNDX(SUOM)
A1	Access Pressure	mmHg	millimeters of mercury	Dialysis	A220	M453
A2	Activity oxygen sat - Aerobic Activity Response(O2)	%	Percentage	OT Notes	A231	P625
A3	Activated Clotting Time			Labs	A231	0000
A4	Acetylcysteine				A234	0000
A5	Albumin	g/dL	grams per deciliter	Labs	A415	G652
A6	Fibrinogen	mg/dL	milligrams per decilitre	Labs	F165	M426
A7	Fibrinogen	mg/dL	milligrams per decilitre	Labs	F165	M426
A8	Glucose (serum)	mg/dL	milligrams per decilitre	Labs	G422	M426
A9	glucose by glucometer (Fingerstick Glucose)			Chemistry	G422	0000

It can be observed from both Table 4.2A and Table 4.2B that “observation name” and expanded “uom” were used as blocking keys for both the source and the target datasets. The researcher used query Q1 to obtain the results for both the source and the target datasets, where the query is aimed at retrieving observation names that have ten similar characters between the source and the target observation names. The starting characters were also checked.

Query 1: Similar lab names with ten characters

```
Q1: SELECT distinct s.[Observation name] as SourceObservation name, s.[UOM] as
SourceUOM
      ,s.[UOMFull] as SourceUOMFull,s.[Category] as SourceCat
      ,SOUNDEX(s.Observation name) as 'SNDX(SourceObservation name)'
      ,SOUNDEX(s.[UOMFull]) as 'SNDX(SourceUOM)'
      ,t.[COMPONENT] as TargetObservation name, t.[UOM] as TargetUOM
      ,t.[UOMFULL] as TargetFullUOM, t.[SYSTEM] as TargetSYSTEM
      ,SOUNDEX(t.COMPONENT) as 'SNDX(TargetObservation name)'
      ,SOUNDEX(t.UOMFULL) as 'SNDX(TargUOM)'
FROM [ResearchTestData].[dbo].[SourceData] s
INNER JOIN [ResearchTestData].[dbo].TargetDataSet t
ON SOUNDEX(s.Observation name) = SOUNDEX(t.COMPONENT)
ORDER BY 'SNDX(TargetObservation name)' ASC
```

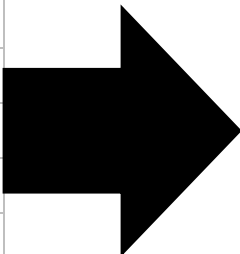
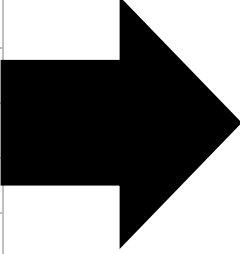
Table 4.3 shows comparisons that will be made based on lab attributes e.g., “Observation name”, for example records that have been recorded as *A231* will be compared with records *B2* from the target dataset and so on, and block *M426* of the “uom” attribute will be compared with the “uom” in record *B8* and *B9*. Thereafter, a similarity measure between the source and the target attribute were achieved through the use of a Jaro-Winkler algorithm. There are cases where the similarity measure algorithm failed to yield a correct similarity weight, because of the manner in which the attribute value is structured. For instance, when comparing “body weight” and “weight”, the Jaro-Winkler similarity measure outputs a weight of “0”, while Edit distance outputs a weight of 0.63 reflecting that the values being compared do not match. Although the algorithm failed to identify the attributes as similar, a human would know that “body weight” and “weight” refer to that of a person. Therefore, as part of a learning algorithm supervision process, an extra feature x_5 was added, so as to check reversed characters of the lab test names. When the reverse algorithm is applied to “body weight” and “weight”, the following results were obtained: “body weight” becomes “thgiew ydob”, then “weight” becomes “thgiew”, and when the Jaro-Winkler similarity measure was applied, a weight of 0.89 was obtained.

Reversing strings is important in cases where there are two strings and the first string is used to describe the other string, however the reversed weight feature will only be used when the output of the “observation name” weight is less than 0.8.

Table 4. 2B Target features with soundex blocking keys

Recl d	TargetObservation name	TargetUOM	TargetFullUOM	TargetSYSTEM	SNDX(TargetObservation name)	SNDX(TUOM)
B1	Accessory nerve (CN XI) exam			Nerves.cranial	A226	M453
B2	Activated clotting time	ratio	Ratio	PPP	A231	R300
B4	Acetylcarnitine (C2)	umol/L	micromoles per litre	Amnio fld	A234	M265
B5	Albumin	g/24 H	grams per 24H	Urine	A415	G652
B6	Fibrin D-dimer	ug/L	micrograms per litre	PPP	F165	M262
B7	Fibrinogen	g/L	grams per liter	PPP	F165	G652
B8	Glucose	umol/L	umol/L	Bld	G422	M265
B9	Glucose^pre dialysis	mg/dL	milligrams per deciliter	Dial fld prt	G422	M426

Table 4. 3 Blocking key values

A231		B2		M453	B1
A234		B4		G652	B5
A415		B5		M426	B8, B9
F165		B7			
G422		B8, B9			

Block A: Observation name

Block B: Unit of measure

For the unit of measure the researcher used Edit distance, because the source unit of value often differs by the prefix of the matrix of the target e.g., source might be “ug/ml” while the target unit of measure uses “ng/ml” matrix, where the difference is that one uses micro, while the other uses nano as a prefix. Therefore, Edit distance calculates the minimum cost of transforming string “y” to string “x”. On the Metavision dataset, there were 214 unique tests that were identified for mapping, however, only 97 had the same sound and similar starting characters with the LOINC dataset, and the 117 that did not pair with LOINC were recorded on an observation dictionary database table. The researcher used detailed laboratory information from the following sources (Mayo Clinic, 2015; Regenstrief Institute, 2016) for extracting more information about the remaining tests. Thereafter each test that was found after searching was inserted into the observation dictionary table and a matching LOINC code was entered on this table as well.

Table 4. 4 Observations with sounds that differ from LOINC observation

Analyte Id	Observation name	LOINC	LOINCAAnalyte	LongUOM	Datasource
39	SVO2SAT	56875-8	Mixed venous oxygen saturation monitoring	Mixed venous saturation (SVO2)	CareVue
1	ALT	1742-6	Alanine aminotransferase [Enzymatic activity/?volume] in Serum or Plasma	Alanine transaminase	MetaVision
10	O2 Consumption	60842-2	Oxygen consumption (VO2)	oxygen consumption	MetaVision

Goal definition and information extraction from the selected corpora

The unstructured data does not follow any kind of schema, and as a result it does not have structure at all, except that it has headings that are meant to make the documents easier to read for a human. In this study, structure is inferred through the use of sentence identification and section header detector, and these form part of the features in this study. However, in other applications like spam classification, or sentiment analysis, each word is treated as a feature and features are selected through Mutual Information technique, chisquare (X^2) feature selection, and frequency-based system (such as TFIDF) feature selection methods (Manning et al., 2009). Therefore, before features could be selected or extracted, it is worth mentioning the purpose for extracting meaning from clinical text and how it links with the overall study. The researcher is interested in extracting and standardizing smoking status from clinical text. The patient's smoking information is a behavioural factor which forms part of an external environmental exposure that was discussed in section [2.3.2](#). This task is sparked by the fact that other important clinical information is not easily recorded in a structured format, and according to (Wu et al., 2013), health professionals such as nurses use a human language to record information in a more detailed format. However, this kind of information is not easy to search for because it is not structured, hence the goal is not simply about extracting text contained in the clinical note, but it is about extracting

meaning and context from these notes and also standardize it so that it becomes a common way of representing smoking information from clinical text. Therefore, in order to be able to extract meaning and context from clinical text, data, methods, and tools that are dedicated to that identified task are needed. Hence the researcher has identified open source tools such cTakes and CLAMP for processing the identified clinical corpora, and these NLP tools were discussed in detail in section 3.9. In this section the researcher will talk about the methods that were used for extracting features from a given clinical text such as the one in Table 4.5.

Table 4. 5 Clinical text about patient’s smoking information and the meaning

Text	Meaning
<p>Social history Miss. CM is an energetic young woman who has had bouts with sleeplessness for the past year or so. She said that her insomnia began with the death of her father who was killed in a train accident last year. Patient is 25 and claims she has <u>smoked for the last five years</u> or so. She <u>used</u> to smoke about <u>half a pack a day</u>, but for the last month she has been down to about <u>3-5 cigarettes a day</u>. She is having trouble stopping altogether.</p>	<p>Current Light Smoker</p>

The clinical note in Table 4.5 covers the patient’s social history, however the rest of the note has multiple sections such as medication report, discharge summary, ECG report, and physical examination just to mention a few. It should be observed from Table 4.5 that the underlined words are key in determining the patient’s smoking status, therefore the researcher will use a clinical pipeline in order to extract information from the clinical notes, and the pipeline includes the following components:

- Sentence detector: A specific DF_CLAMP_Sentence_Detector was used, this is a default sentence detector within the CLAMP software, it was specifically built to process clinical text by determining where a sentence on a clinical note ends.
- Rule-based tokenizer which segments raw text into tokens, in this study DF_CLAMP_Tokenizer was used.
- POS tagger: A DF_OpenNLP_POS_tagger which is used to tag parts of speech on the tokens of data was used as the default NLP Part of speech.

- Section Identifier: This feature was used to identify a section on a clinical text, meaning that there is no need to manually specify which section deals with laboratory data or medication data. For the section identifier the researcher used DF_Dictionary_based_section_identifier.
- Assertion Identifier: This checks if there is a negation associated with a clinical concept, the negation function uses DF_NegEx_assertion to check for the absence or opposite of a positive observation, e.g. “Patient’s father has history of alcohol abuse, but patient does not drink alcohol”, in this case “patient does not drink alcohol” is negated, while the first passage about patient’s father is not negated.
- UMLS encoder: The encoder is used to match the clinical concept terms into UMLS Concept Unique Identifier (CUI) code, once a term has been mapped to a CUI code it is then easier to map that term to LOINC or SNOMED or to any coding standard. For instance, heart rate is mapped to the CUI code of C0018810 which has a LOINC code of 8861-7.
- Named Entity Recognizer: The researcher used the DF_CRF_based_named_entity_recognizer which identifies three types of clinical concepts namely problems, treatments and tests.
- Ruta Rule Engine: This is also known as UIMA rule engine, this was used for identifying, creating and modifying annotations, and the identified annotations are treated as features, one example where the rule engine is used is the identification of lab tests and their corresponding values and unit of measures.
- Temporal recognizer and relation: For the recognizer, a CRF-based temporal was used, a temporal is able to extract time-specific information such “last month”, “3rd of August”, “2011-01-02” and more. Then temporal relation is used for creating relations between the event and the time, e.g. “smoked” is the event, and “five month” is the temporal recognized.

These components help with the task of annotating clinical notes, and annotation helps provide more information about a text, it is like the metadata of the whole text. Annotation is similar to the process of supervising a machine learning algorithm, this gives the machine learning algorithms clues about the data. Therefore, in the following

section the researcher will show rules written in UIMA RUTA (Kluegl, Toepfer, Beck, Fette, & Puppe, 2016) language for creating an annotated corpus.

#	Research sub-question
iv	What features will be used to determine similarity between two records?

4.3 FEATURE SELECTION FOR MATCHING SOURCE TO TARGET

Based on Table 4.6, one can observe that there are four features with weight attached on each, $x_1 = 0.77, x_2 = 0, x_3 = 1, x_4 = 1$ and the indication of whether it's a match or non-match is represented by y as an output feature. In the case of this study, the process of feature selection is aimed at selecting features that contribute to the decision of determining if record A matches record B, and the all the features are scaled between values "0" and "1".

Table 4. 6 Attributes similarity comparison

	Observation name	Category	ShortUOM	LongUOM	Match
Record A - Source	Intra Cranial Pressure #2	Hemodynamics	mmHg	millimeters of mercury	
Record B - Target	Intracranial systolic	Skull	mmHg	millimeters of mercury	
Weight	0.77	0	1	1	?

For the output variable, the researcher uses the feature called "match" as shown in Table 4.6, where at the moment it is not known if the two records match or not. Table 4.6 shows record "A" as the source dataset and record "B" as the target dataset, where each attribute is compared and weighted using Jaro-Winkler and Edit distance similarity algorithm. It can also be observed from Table 4.6 that the weight label is used to record the similarity output. If the output is 0.77, then this is interpreted a 77% match between the source and the target. However, the researcher has set a 75% threshold for

matching laboratory observations except for the “unit of measure” which uses a threshold of 80%. The “unit of measure” observation is short, and the majority of the “unit of measure” characters should match in order to determine if the compared records match or not. A match is represented by a “1” and a non-matching record by a “0” on the “match” attribute.

$$Z = \{(i, j); i = j, i \in X_{source}, j \in X_{target} \} \quad (15)$$

$$U = \{(i, j); i \neq j, i \in X_{source}, j \in X_{target} \} \quad (16)$$

Furthermore, a match is represented as shown in Equation (15) and a non-match in Equation (16), where the variable i and j represent the compared attribute instances, X represents the source or the target dataset. Features are numerical representations of raw data, or data that could be understood by the classifier for model building. However, the researcher has proposed the use of rules to determine matching records. This method of record comparison is not new, it is often used in record matching system as illustrated by (Doan, Halevy, Ives, et al., 2012). The rules in Figure 4.2 were defined in order to determine if two records match, the training data was loaded into a decision tree model, and the rule model was produced from executing the decision tree model.

$(s(ObservationName)[x_i, x_j] \geq 0.75) \wedge (s(UoM)[x_i, x_j] \geq 0.8) \wedge (s(ExpUoM)[x_i, x_j] \geq 0.75) \wedge (s(Category)[x_i, x_j] \geq 0.75) \Rightarrow [x_i, x_j] \rightarrow Match$ $(s(ObservationName)[x_i, x_j] < 0.75) \wedge (s(UoM)[x_i, x_j] \geq 0.8) \wedge (s(ExpUoM)[x_i, x_j] \geq 0.75) \wedge (s(Category)[x_i, x_j] \geq 0.75) \Rightarrow [x_i, x_j] \rightarrow Non - Match$ $(s(ObservationName)[x_i, x_j] \geq 0.75) \wedge (s(UoM)[x_i, x_j] < 0.8) \wedge (s(ExpUoM)[x_i, x_j] \leq 0.75) \wedge (s(Category)[x_i, x_j] \geq 0.75) \Rightarrow [x_i, x_j] \rightarrow Non - Match$ $(s(ObservationName)[x_i, x_j] \geq 0.75) \wedge (s(UoM)[x_i, x_j] \geq 0.75) \wedge (s(ExpUoM)[x_i, x_j] \geq 0.8) \wedge (s(Category)[x_i, x_j] < 0.75) \Rightarrow [x_i, x_j] \rightarrow Match$
--

Figure 4. 2: Initial rules for determining if two records match or not.

According to (Kim, El-Kareh, Goel, Vineet, & Chapman, 2012), enhancing or expanding local observation names improved the chances of correctly identifying the matching LOINC observation name, where also adding the unit of measure was found to reduce the number of false positive matches. (Mcdonald et al., 2017) also suggests that the “unit of measure” is important when matching local observation names to LOINC observations. In addition, the “ObsevationName” shown in Figure 4.2 was already expanded from a short “ObservationName”, e.g “O2 sat” was expanded to “Oxygen Saturation”, “Temp” to “Temperature”.

For laboratory tests, there are other cases where the name of the sample is included on the observation name, for example, bicarbonate serum or base excess arterial. Such tests give extra information and therefore a laboratory sample information can be extracted from the observation name when the sample name has not been provided.

4.4 ANNOTATING THE CLINICAL CORPORA

#	Research sub-question
iv	How were features selected for unstructured data?

Annotating the text is part of extracting and selecting features for the given corpora. A simple form of annotation in web design is the enclosing of text such as the following: “Text annotation” and the browser would interpret this as a bolded text “**Text annotation**”. Therefore, also with clinical data annotation, the goal is to teach the algorithm how to identify smoking-based named entities and how these entities relate to one another. Therefore, in this study, annotation rules are used to annotate the corpora, and once it has been annotated then the annotated corpus will act as input to the learning algorithm for the purpose of training the algorithm. However, before training the algorithm, it is good to clearly define what the goal of extracting smoking information

entails. The researcher has followed the guidance of a study by (Uzuner, Goldstein, Luo, & Kohane, 2008) to identify smoking status, and this is as follows:

- Current smoker: This is a patient whose discharge summary states that for the past year the patient was a smoker.
- Smoker: A patient who can be regarded as a current smoker or non-smoker, however the discharge summary does specify that the patient has history of smoking although it does not mention whether the patient did quit or not.
- Past smoker: The discharge summary states that the patient has a history of smoking, however has not smoked for the past year.
- Non-smoker: The discharge summary states that the patient never smoked before.
- Unknown: The discharge summary of the patient does not state whether the patient smokes or not.

One would also note that Table 4.7 has CUI codes which help identify each smoking status, once the CUI code is defined it becomes easier to standardize the clinical term using LOINC or SNOMED-DT coding standards.

Table 4. 7 Smoking status examples

Clinical note	Smoking status category	CUI	SNOMED-CT	LOINC
Former 2 pack per day smoker x 28 years, now smokes a pack every other day.	Current Smoker	C3241966	428071000124103	64234-8
The patient's coronary artery disease risk factors include, hypertension, hypercholesterolemia and a cigar smoker for thirty years. The patient has no history of diabetes.	Smoker	C0337664	449868002	
She quit smoking >10years ago, but prior to that had approx. 30 packyear h/o tobacco.	Past smoker	C0337671	8517006	
Patient is an accountant. He does not consume alcohol or smoke cigarettes.	Non-Smoker	C0425293	266919005	
Binge drinking (6-pack x2 per week). He uses cocaine via inhalation once or twice per month. He also uses marijuana and has a history of IV drug use, heroin and cocaine, approximately 10 years ago.	Unknown	C0425306	266927001	

Now, based on the above defined smoking statuses, a rule-based system known as UIMA Ruta (Rule-based scripting language) was used for creating rules for extracting information that could be standardized. The RUTA rule allows for execution of conditional statements, control structures and the declaration of variables. Figure 4.3 indicates the rules written in a RUTA language for determining the smoking status of the patient from a given clinical text. The rules are executed in a linear order, and before the rules can be executed, the CLAMP pipeline is executed first. During the running of the CLAMP pipeline each token or word from a clinical text is tagged using the “semanticTag”, so the first rule in Figure 4.3 states that if the “history” tag is followed by the “smoker” tag then a new tag “PastSmoker” is created as a feature. This is only a sample rule and it is not explicitly defined, however the rules that the researcher has used are accessible and be opened via Notepad++, the path to access the file is shown in Appendix E.

```

TYPESYSTEM ClampTypeSystem;

// 1. rules to parse past smokers;
BLOCK(ForEach) Sentence{} {
// pattern: history of smoking;
ClampNameEntityUIMA{ FEATURE( "semanticTag", "History") }
    ClampNameEntityUIMA{ FEATURE( "semanticTag", "Smoker") -> SETFEATURE(
"semanticTag", "PastSmoker" ) };

// 2. rules to parse non-smokers;
BLOCK(ForEach) Sentence{} {
    ClampNameEntityUIMA{ FEATURE( "semanticTag", "Smoker"), FEATURE(
"assertion", "absent" )
        -> SETFEATURE( "semanticTag", "Non-smoker" ) };
}

// 3. rules to parse current smokers;
BLOCK(ForEach) Sentence{} {
// currently smokers
ClampNameEntityUIMA{ FEATURE( "semanticTag", "TimeModifier") }
    ClampNameEntityUIMA{ FEATURE( "semanticTag", "Smoker") -> SETFEATURE(
"semanticTag", "CurrentSmoker" ) };
}

```

Figure 4. 3: A sample rule for detecting a smoking status of a patient for a given clinical note.

The second rule states that if there is a “smoker” tag followed by a negated tag “absent”, then create a new feature and label its tag as “Non-smoker”. The “absent” keyword indicates that the tag is negated, for instance, when the clinical note states that “the patient denies tobacco use”, therefore because of the keyword “denies” next to the “tobacco” keyword then the phrase is said to be negated, and the opposite of this is the “present” keyword which means the phrase is not negated. The third rule extracts information about the current smokers, the rule was constructed by first identifying a temporal which in this case are time-based adverbs such as currently, momentarily, in the meantime, presently, time being, present moment and more. This rule checks if the temporal is followed by a “smoker” tag, and if this condition is true, then a “CurrentSmoker” tag is set as a feature. The rules used for detecting patient’s smoking status were derived from the studies by (Sohn & Savova, 2009; Uzuner, Goldstein, Luo, & Kohane, 2008).

Figure 4.4 shows the results of a tagged clinical note using as input the text from Table 4.5, from these tags the next task is to add rules that specifically would make an annotated corpus. Figure 4.5 shows the result of annotated corpus; which is an XML file that uses an XML Metadata Interchange (XMI) structure, and the CRF learning algorithm expects input such as one in “.xmi” format. The “.xmi” file that is produced uses a Stand-off Annotation by Character Location type of annotation, this method records the start and an end of the annotated text. One can observe from Figure 4.4 that the text “the past year” is annotated as a “temporal”, meaning it is a time-specific text, then Figure 4.5 is the representation of the “.xmi” file whereby the location of the annotated text is recorded, the start of the “the past year” starts at character location “77” and ends at location “90”. Other information such as the concept unique identifier (CUI) code can also be added as part of the “.xmi” file. The start and the end locations are helpful for the computer to find important information on an annotated text which is necessary for the application of machine learning algorithms (Pustejovsky & Stubbs, 2013).

1 Miss. CM is a energetic young woman who has had bouts with sleeplessness for the past year or so. She said that her insomnia began with the death of her father who was killed in a train accident last year.

Patient is 25 and claims she has smoked for the last five years or so. She used to smoke about half a pack a day, but for the last month she has been down to about 3-5 cigarettes a day. She is having trouble stopping altogether.

Figure 4. 4: Feature extraction from clinical text using named entity recognition

Start	End	Type	Value	Other	Text
77	90	temporal			the past year
196	205	temporal			last year
240	270	PastSmoker			smoked for the last five years
240	246	Smoker			smoked
260	270	temporal			five years
260	264	snum	present		five
265	270	year			years
290	295	Smoker			smoke
314	319	temporal			a day
329	343	temporal			the last month
375	385	Smoker			cigarettes
386	391	temporal			a day

Figure 4. 5: Results from the extracted entities

4.5 MODEL SELECTION AND OPTIMISATION

This study is aimed to classify patient’s observation data into a standard that would make the data easy to search, trace and share. Therefore, the researcher has used multiple datasets from different data sources. There is a physical activity dataset, environmental dataset, laboratory dataset and a vital signs dataset. As stated in Chapter Three, these datasets are from different sources, and the researcher will use an SVM classifier to learn to classify this data according to the LOINC coding standard.

Data visualisation

The researcher started off by loading the training data into the RapidMiner tool, and the visual exploration of the data can be seen in Figure 4.6. The training data consisted of laboratory data, examination data, and vital signs from four different data sources. Three of the datasets were *LabEvents*, *CareVue* and *MetaVision*, and all of these were from the MIMIC-III database. There are also datasets from *NHanes*, which included laboratory and examination data.

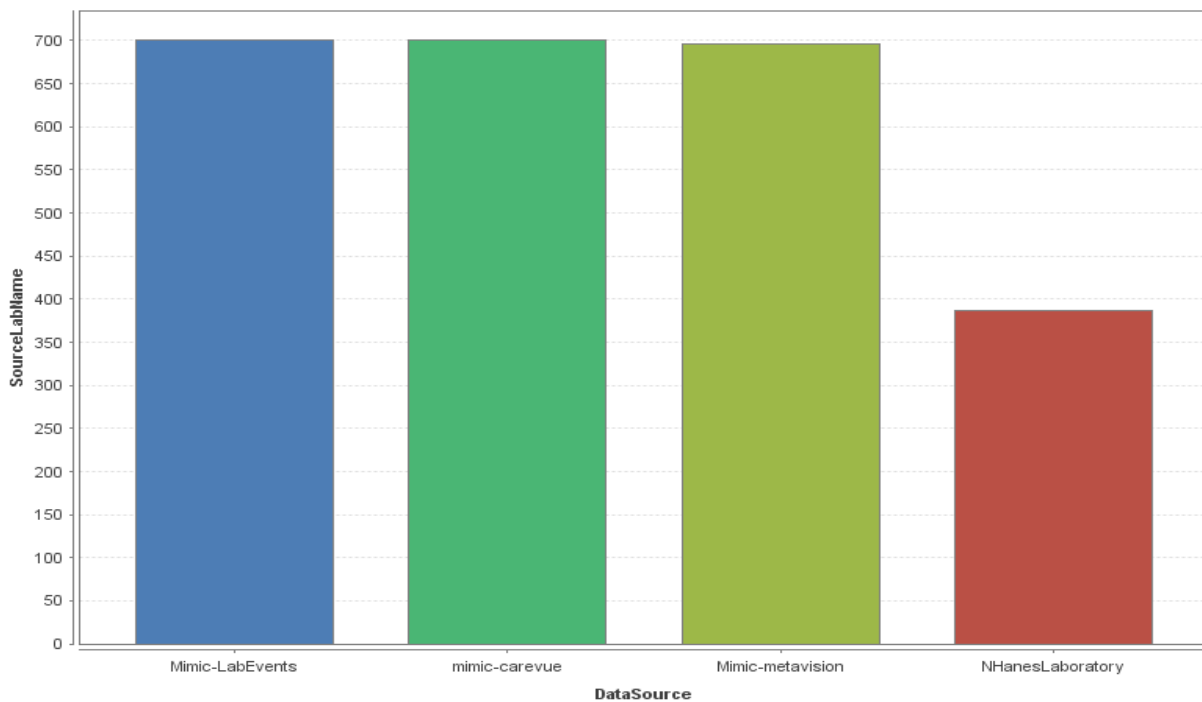


Figure 4. 6: Number of training examples for datasets

Figure 4.6 shows the number of training examples for each data set, where there were 2483 training examples, with 629 missing data values for output feature. Therefore, only 1845 were used for training the model. There were also 29 features, included in which were eight attribute similarity weighting features. Weighting features were used for evaluating whether the source matched the target attribute.

Model testing

Having selected the training data with all the weight attributes, the researcher then loaded the training data so as to get a pictorial view of the initial model. Figure 4.7 shows results of the generated model. The model was trained firstly with the labelled 1845 training set, and a model was generated, then the generated model was applied onto the 629 of the unlabelled training set and a prediction with the accuracy of 90.88% was achieved, a model with a predicted output can be seen in Figure 4.8.

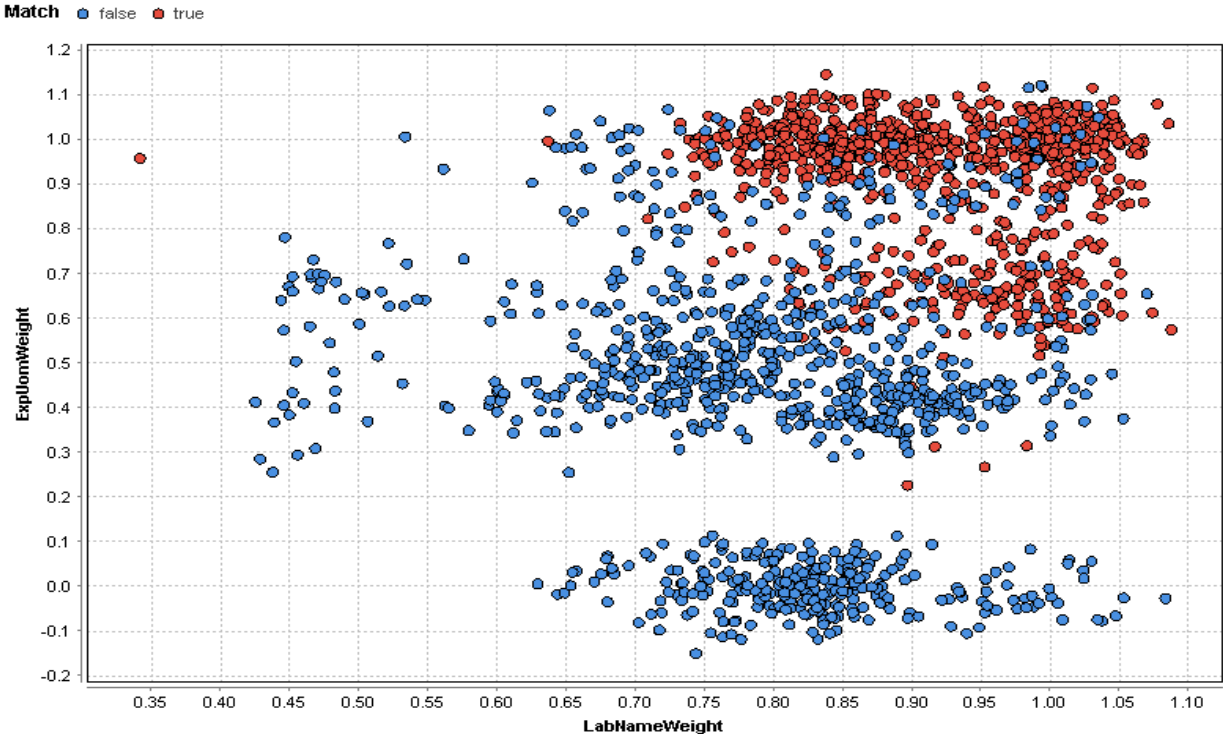


Figure 4. 7: Tested model without cross-validation

However, the results shown above are there to get started with model building, this helps address complexities such as variance and bias when working on a machine learning problem. The model generated above used a decision tree classifier, and building a decision tree helps with the understanding of which feature has a high splitting criterion. A decision tree classifier also gives a vivid picture of the rules

governing the training data. Further on, the researcher added an SVM classifier, and model validation criterion was applied.

Model validation and parameter optimisation

The model is evaluated in order to avoid underfitting, therefore it is of paramount importance to validate the model. The model is evaluated by splitting the dataset into a training set, cross-validation set, and a testing. A training set consists of all the examples from the selected dataset for fitting the model, a cross-validation is a set of examples also from the same dataset, however, its purpose is for tuning parameters during the process of training. Finally, a test set is used for testing the performance of a classifier (Kuhn & Johnson, 2013; Meyer, 2009; Ng, 2016). As for sampling the data, the researcher used automatic sampling which uses a stratified sampling for either polynomial or binomial class labels, and if the class labels are neither polynomial or binomial then it uses random shuffling.

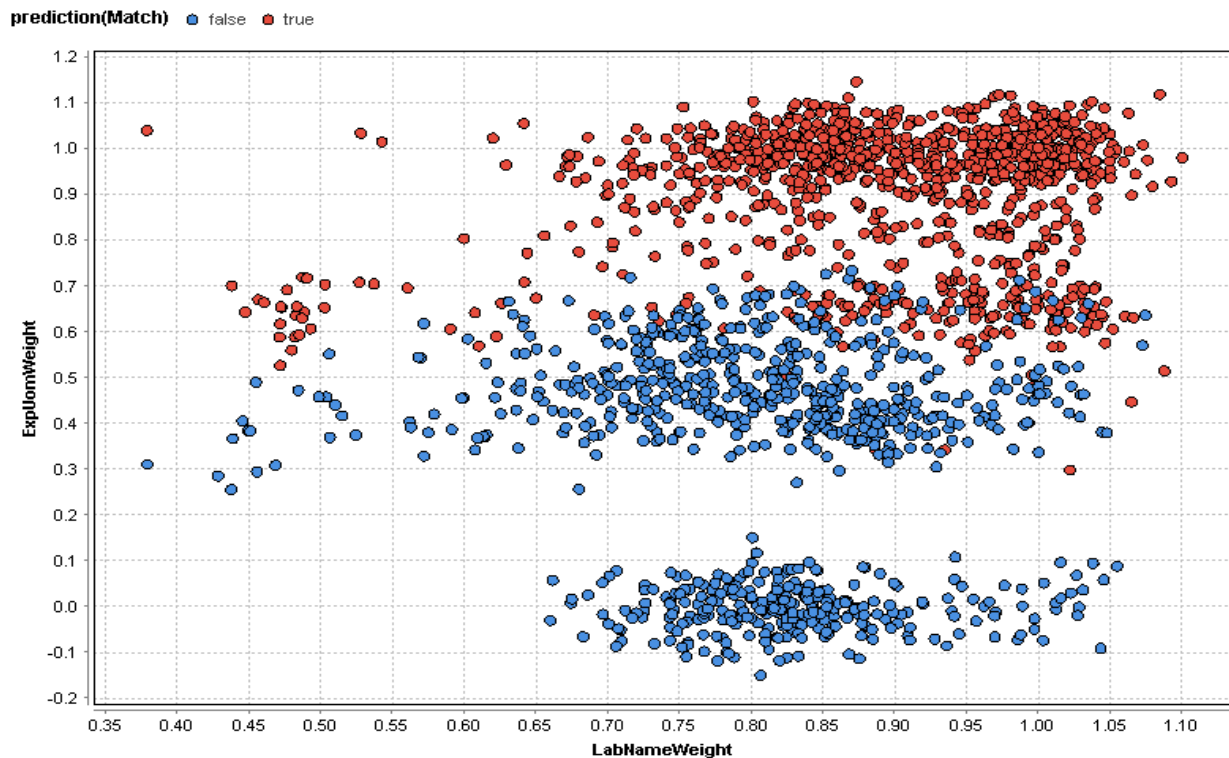


Figure 4. 8: Predicted model without cross-validation

There are many different methods for performing cross-validation, however in this study the researcher will use V-fold cross-validation (CV). (Arlot & Celisse, 2010) have noted that this procedure is the most popular, due to its mild computational cost. The V-fold CV works by partitioning the training data into “k-1” number of folds, where only a single “k” fold will be used for testing the model. In the case of this study, 10 sets of fold were selected for training the model, and only one set was used for testing, and Figure 4.9 shows how V-fold CV works.

Having selected the cross-validation procedure, the researcher has also added the LibSVM classifier, which makes it effortless to optimise parameter θ during the cross-validation procedure.

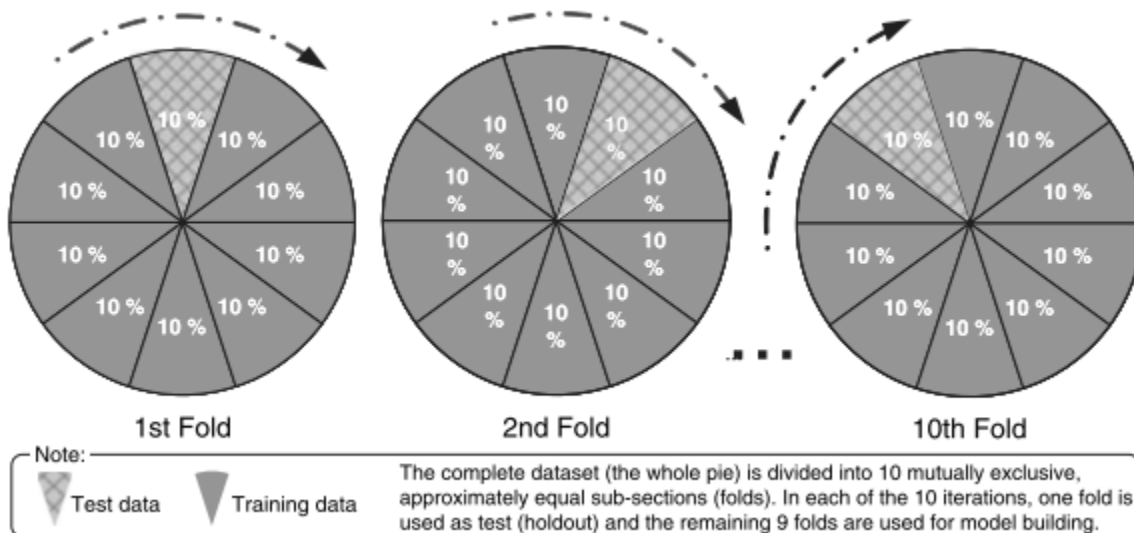


Figure 4. 9: Pictorial view of a 10-fold cross-validation (Source: (Olson & Delen, 2008))

LibSVM is a non-linear library for solving classification, regression and distribution problems. Before LibSVM could be implemented, the user should first select the value of C (which is a penalty parameter for error term). Equation (17) shows this parameter.

LibSVM also requires a kernel function to be chosen from a set of available ones which were mentioned in section 3.8, and radial basis function (RBF) kernel has been selected for the purposes of this study. This function uses two parameters, namely the cost parameter C and γ (gamma) parameter, and these parameters were optimised through the execution of a grid-search process (Chang & Lin, 2011).

$$\text{Min}_{\theta} C \left[\sum_{i=1}^{m_{test}} y_{test}^{(i)} \text{Cost}_1(\theta^T x_{test}^{(i)}) + (1 - y_{test}^{(i)}) \text{Cost}_0(\theta^T x_{test}^{(i)}) \right] + 1/2 \sum_{j=1}^n \theta_j^2 \quad (17)$$

The grid-search process was made part of the CV process and Table 4.8 illustrates the process of grid-search. While the classifier was being trained on 10 folds, the C and γ parameters were selected and printed out for performance evaluation purposes. There were 242 iterations when the number of folds were set to 10; parameter C within a range of 0.001 to 1000; the number of steps set to 100; γ with a range of 0.001 to 1; and number of steps set to 10. Both parameters were set to a logarithmic scale of log base 2.

Table 4. 8 Model selection criteria

Degree of polynomial	Hypothesis function	Model
d=1	$h_{\theta}(X) = \theta_0 + \theta_1 \times X$	Linear
d=2	$h_{\theta}(X) = \theta_0 + \theta_1 \times X + \theta_2 \times X^2$	Quadratic
d=3	$h_{\theta}(X) = \theta_0 + \theta_1 \times X + \dots + \theta_3 \times X^3$	Cubic
⋮		
d=10	$h_{\theta}(X) = \theta_0 + \theta_1 \times X + \dots + \theta_{10} \times X^{10}$	10 th order Polynomial

Table 4.8 starts off with a simple linear problem whereby a linear boundary can be used to separate the positive and negative training examples. As more features are added, the model changes with the degree of polynomial as shown in Table 4.8. The variable

“d” denotes the order of polynomial for the θ parameter, therefore this parameter value is used as input in the calculation of the cross-validation error. Then the cross-validation function that yields the smallest cross-validation error will be used to test for the generalisation errors during the testing phase. The cross-validation phase is used to tune parameters, and for selecting new or discarding features so that the learning algorithm can produce a near-accurate classifier. During the testing phase, the researcher used training data that was not previously fed to the model, or data that is not known to the model, in order to produce a generalisation error. The produced generalisation error is used for evaluating if the classifier is able to predict that the source observation data is similar to the target observation data. When the classifier makes a correct prediction, it confirms the classifier is capable of learning from the training data, and therefore that source data was standardized in the same manner as the target data. In Chapter Five the researcher will show that even when the classifier has made a correct prediction, that prediction might have a high bias, or alternatively, the results might have a high variance, or may be overfitted. Overfitting is a result of a model’s inability to make correct predictions on unseen training data. Both high variance and high bias are major problems when designing a learning system, if left undiagnosed then one might be optimistic about the model’s performance even though it is highly biased, and such a model would lead to incorrect results on a production system. Overfitting occurs also with NLP algorithms when unstructured data is used, and with NLP applications it is often caused when many features are used to train the algorithm, and an algorithm that is given many features fails to correctly predict a class for new training examples.

Gold standard and feature selection for the corpus

Model selection for structured data is similar to the process of annotation creation for unstructured data which means that the annotation process is based on expert advice on how to differentiate between different smoking statuses. Therefore, in the absence of a health informatician expert or a clinical data annotator, the researcher has resorted into using the suggestions of (Sohn & Savova, 2009) in order to create the gold

standard. The gold standard is also known as the benchmark. Once this is properly defined then the resulting corpus is ready to be used to train a machine learning algorithm. A good performance by the algorithm would mean that the generated model could thus be used as the common annotator for the detection of the smoking status across multiple clinical corpora. It should be noted that the end-goal of the annotation task is not about the smoking status annotation, but it is about standardizing the smoking status through a coding standard such as LOINC or SNOMED-CT. In fact (Pustejovsky & Stubbs, 2013) said that these days annotations are done so as to get data to train a machine learning algorithm, which is the case with this study as well. For the purposes of generating the gold standard, the researcher has identified three smoking statuses to be extracted from the corpora. These are referred to as classes or tags and they include: the current smoker, the non-smoker and the past-smoker class. The other two classes (unknown and smoker) were excluded because the findings by (Sohn & Savova, 2009) have indicated that it is least challenging to predict if a document should be annotated as an “unknown” or “smoker” class, hence they have obtained a high F-measure score for both, and as a result the researcher will not cover these classes.

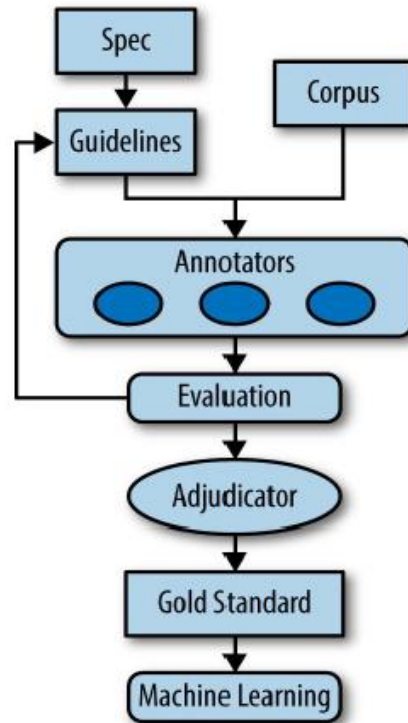


Figure 4. 10: Annotation process (Source: (Pustejovsky & Stubbs, 2013))

Since there are only three smoking classes that the researcher is interested in, it is important to determine how these statuses are often phrased in the English language. For instance, determining if whether the patient never smoked could be phrased like this: “Patient does not have history of smoking”; “Smoking is the least thing the patient has considered in his life”; “He never smoked”, these are a few of the examples of a non-smoking class. The problem is now that there is no common method that nursing professionals use to phrase that a person does not smoke. However there are programs that can be used to identify common words used to determine an event or action. The researcher therefore used the Google’s Ngram Viewer (Google, 2013) to learn about the common use of words in books to indicate non-smokers. The Google Ngram Viewer is a website which allows for the search of common words or phrases using parts of speech and wildcards for querying the desired information in Google books. Data is derived from a subset of 5 million books out of a total of 15 million that have already been digitized (Michel et al., 2011). From the Ngram Viewer website the researcher was

able to observe common use of words in relation to time, for instance Figure 4.11 gives a graph of the word usage in books regarding patient's non-smoking status, and from Figure 4.11 it can be observed that the phrase "does not smoke" was more common in books than the phrase "never smoked" around the year 1947 and 2007. This information could help determine common words that should be tagged from the corpora to determine the three smoking classes.

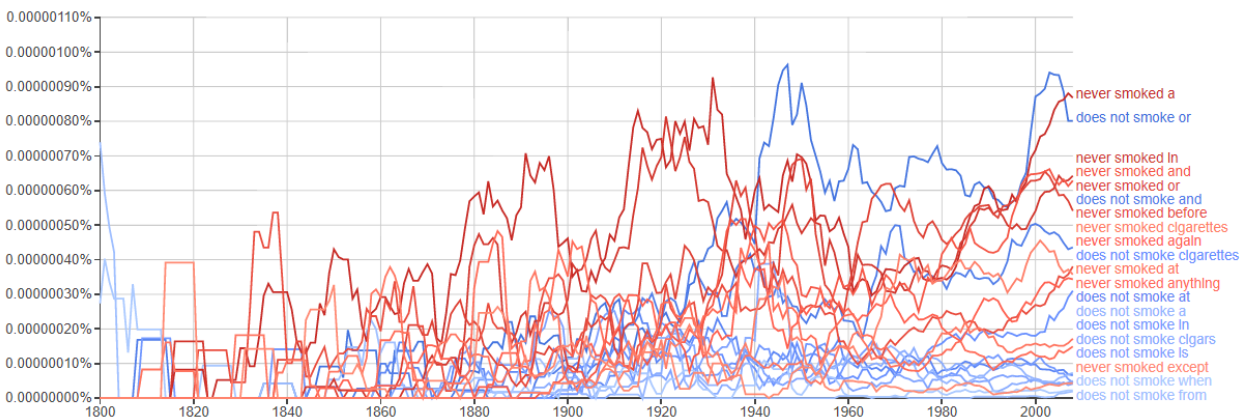


Figure 4. 11: Shows the usage results of two sets of ngrams between the year 1800 and 2008, this is a comparison between 3-gram which is (does not smoke) and a 2-gram (never smoked).

Once the researcher was satisfied with the gold standard, the next task is to select features that will be used with the learning algorithm. The CLAMP software allows one to select features for named-entity extraction tasks, these are word representation (WR) features and they arranged as follows: (1) clustering-based feature; (2) distributional feature; (3) word embeddings features. These features are described in brief below:

- Brown Clustering

This is a clustering-based word representation algorithm that groups related words into clusters based on the context that these words are in. The brown clustering receives either a corpora of words or an annotated corpus, then the algorithm partitions the words and thereafter outputs the partitions into clusters of words (Figure 4.12). Lastly, it

generates an agglomerative hierarchical cluster which is a cluster that implements a bottom up approach (Collins, 2011). Figure 4.12 shows the results that were obtained from a classic study done by (Brown, Della Pietra, deSouza, Lai, & Mercer, 1992), where their findings showed that words can be grouped together based on the surrounding words and their contexts.

Friday Monday Thursday Wednesday Tuesday Saturday Sunday weekends Sundays Saturdays
June March July April January December October November September August
people guys folks fellows CEOs chaps doubters commies unfortunates blokes
down backwards ashore sideways southward northward overboard aloft downwards adrift
water gas coal liquid acid sand carbon steam shale iron
great big vast sudden mere sheer gigantic lifelong scant colossal
man woman boy girl lawyer doctor guy farmer teacher citizen
American Indian European Japanese German African Catholic Israeli Italian Arab
pressure temperature permeability density porosity stress velocity viscosity gravity tension
mother wife father son husband brother daughter sister boss uncle
machine device controller processor CPU printer spindle subsystem compiler plotter
John George James Bob Robert Paul William Jim David Mike

Figure 4. 12: Word clustered based on context and relatedness from an input of 260 741-word vocabulary (Source: (Brown et al., 1992))

Therefore, when one is working on a named-entity recognition (NER) problem, the search for words that are not familiar or not in the dictionary defined for NER, can be inferred through Brown clustering methods by finding their surrounding words. This helps identify meaning from phrases that are not structured in the same way, because people use language differently even though the concept being addressed might be the same. Therefore the researcher used 34066 words that had already been arranged into a hierarchical structure. In fact, this is the default setting from the CLAMP software, and with the Word embedding feature the researcher also used the default list provided by CLAMP.

- Discrete Word Embedding

CLAMP differentiates between word embeddings and discrete word embedding. However in this section the researcher groups the two and give the underlying idea

behind the word embedding feature. The word embedding feature has the capability to represent words as vectors, and words that are contextually related to one another are represented closer while nonrelated words appear far apart from each other, for instance words such as “king” and “queen” are paired closer to each other, as are “dog” and “cat”. However, these two sets appear far apart from one another on a vector space. Word embeddings could be employed through techniques such as Word2vec and glove. Both these techniques use a neural networks and matrix factorization so that it learns to predict a word when given a set of a context word (Zamani & Croft, 2016), e.g. “Patient ? cigarette”, the algorithm would predict the probable word to replace the question mark “?” based on the context and other related words, could be “smokes”, “hates” or any other word that addresses a similar concept.

- Random Indexing

This is a form of a distributional word representation technique that has been reported to have human cognitive features such as the ability to make judgements about the quality of an essay or any text-based material that one wants to analyse. (Higgins & Burstein, 2007) have used Random Indexing (RI) for assessing the coherence of words used in a student’s essay. The RI technique addresses the drawbacks of latent semantic analysis (LSA), LSA is also a technique that uses statistical computing in order to extract information from a large text corpora and represents the meaning of words, passages, and sentences in different contexts as context vectors. Therefore the idea is that words that have similar meaning have similar context vectors and those that are not similar have dissimilar context vectors (Kanerva, 2009). Since the RI feature helps to extract meaning from the given corpora or corpus, it is therefore considered as one of the features to be used to train CRF sequence classifier.

The researcher has attached all documents and dictionaries that were used with the identified features, and these dictionaries use “.txt” and therefore can be opened with Notepad, and the path to access these files is shown in Appendix E. Other preparations for the data included the splitting of the training data and test data and the test data was made up of 20% of the annotated corpus. A 5-fold cross validation was then

selected for parameter optimization, and 5-folds were selected because CLAMP only allowed this number of folds to be selected.

4.6 EXPERIMENTAL PROCEDURES

4.6.1 SYSTEMS SET UP

The experiments will be run on a Windows 10 Lenovo machine, with the following specifications: Processor: Intel (R) Core (TM) i7 7500U CPU at 2.90GHz; RAM: 8GB System type: 64-bit Operating System.

4.6.2 EVALUATION MEASURES

#	Research Question
vii	How will the correctness of the results be evaluated?

This section of the study claims to answer the research sub-question as shown above, where the researcher gives details about machine learning model evaluation methods. The results of applying these methods are discussed in Chapter Five. Building a model comes with many challenges that should be addressed before results can be considered correct or accurate. Recall, Precision and F1-score are few of the methods used for evaluating the performance of a classifier on the given test data. Recall is also known as sensitivity and it is the measure of completeness or coverage. A simple example that helps with the understanding of recall and precision is the diagnosis (prediction) of cancer patients, in this case recall would be the proportion of patients that had cancer which were diagnosed (or predicted) by the oncologist (or algorithm) as having cancer. Then precision (also known as specificity) is the proportion of patients that were diagnosed (or predicted) as having cancer who in fact had cancer. There is a trade-off between recall and precision. An algorithm with a very high precision has low recall because it would fail to cover patients without cancer, also with a very high

coverage the algorithm would have low precision, therefore F1-score or F-Measure is used to combine the measures of both precision and recall (Gorunescu, 2011).

(Han et al., 2012) has described accuracy as measuring the recognition rate on a test set, that is, how accurate in terms of percentages the classifier can be in identifying correct matches between records. Four different tests are identified in order to test for performance of the algorithm, these are as follows: true positives, false positives, false negatives, and true negatives. In order to define these four tests with regards to this study, the researcher uses the following illustration: set “B” represents an i^{th} training example from the source dataset, and set “C” an i^{th} training example from the target dataset. The true positives measure signifies that the predicted class matches the actual class, set “B” matches set “C” in terms of similarities measures (see Table 4.9). Then, false positives is when the classifier incorrectly predicts that set “B” matches set “C”, while false negatives refer to when the classifier fails to predict that set “B” matches set “C” when it ought to match. Lastly, true negatives refer to when the classifier correctly predicts that set “B” and “C” do not match.

Table 4. 9 Evaluation metrics for the classifier

		Actual Class	
		1	0
Predicted Class	1	True Positive (TP)	False Positive (FP)
	0	False Negative (FN)	True Negative (TN)

The “1s” in Table 4.9 represent a positive outcome such as “is a match”, and alternatively “0” represents “is not a match”. The researcher will thus use this method to measure the accuracy, precision (Equation 18), recall (Equation 19), and the F1-score (Equation 20).

$$\mathbf{Precision} = \frac{TP}{TP+FP} \quad (18)$$

$$\mathbf{Recall} = \frac{TP}{TP+FN} \quad (19)$$

$$\mathbf{F1 - score} = \frac{2 * \mathbf{Precision} * \mathbf{Recall}}{\mathbf{Precision} + \mathbf{Recall}} \quad (20)$$

These evaluation methods are suitable for binary classification problems where the prediction is either “yes” or “no”, “true” or “false”, however there are other cases where one needs a multiclass classifier where the predictions could be a range of classes. Therefore in such cases there are measures for averaging the performance of the predictions across classes, these measures take the average of both precision and recall. (Barrett, Levell, & Milligan, 2013) have defined macro-average precision as the average precision from all the classes (see Equation 22), and macro-average recall and Equation (21) was derived for simplicity and it is exactly the same as Equation (18), the same is applicable for Equation (19) and Equation (23). Equation (24) calculates the macro-average recall which is basically the average of all the “recall” measures for the given classes.

$$P = \frac{TP}{TP+FP} \quad (21)$$

$$\mathbf{Precision}_M = \frac{\sum_{j=1}^n P_j}{n} \quad (22)$$

$$R = \frac{TP}{TP+FN} \quad (23)$$

$$\mathbf{Recall}_M = \frac{\sum_{j=1}^n R_j}{n} \quad (24)$$

$$\mathbf{Precision}_\mu = \sum_{j=1}^n P_j \quad (25)$$

$$\mathbf{Recall}_\mu = \sum_{j=1}^n R_j \quad (26)$$

$$\mathbf{F1 - Score}_M = \frac{\sum_{j=1}^n f1-score}{n} \quad (27)$$

$$\mathbf{F1 - Score}_\mu = \sum_{j=1}^n f1 - score_j \quad (28)$$

Now micro-averaging is used for summing up all the true positives, false positives, and false negative for each class or tag, and this sum is further computed for effectiveness on large classes on the test data (Manning et al., 2009). Then Equation (27) is calculated by taking an average of all the F1-scores for multiple classes, whereas Equation (28) uses a harmonic mean (as shown in Equation 20) of all the used classes.

The researcher will use the Receiver Operating Characteristic (ROC) curves to compare the performance of SVM and decision tree on the given training and test data. ROC, as shown in Figure 4.13, is one of the methods used to quantitatively evaluate the performance of machine learning models. The ROC curves are used in classification problems, and these curves show the relationship between the sensitivity of the classification model and the rate of false positives that were yielded by the evaluation metrics. When using ROC, there are four possible outcomes: when x-axis represents the level of false positives, and y-axis the level of true positives. It can therefore be said that point (0;0) represents a level where there are no true positives and no false positives, and point (0;1) is a perfect classification, when there are true positives and no false positives. The upper right point (1;1) reveals both high levels of true positives and high levels of false positives; point (1;0) reveals high false positives with low true positives. A classifier that leans towards the North-West of the curve is a good classifier, because it has more true positives than false positives. A liberal classifier is one whose curve is towards that north-eastern direction, where this classifier is able to classify true positives. However, there may be an abundance of errors, because it also has high false positives (Olson & Delen, 2008; Saliccioli, Crutain, Komorowski, & Marshall, 2016).

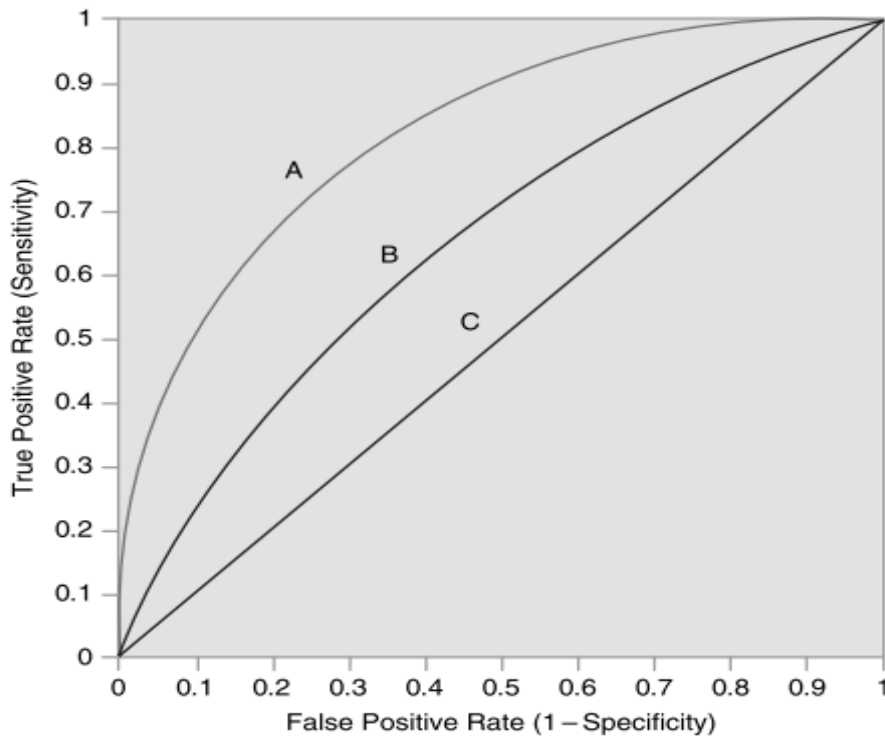


Figure 4. 13: ROC curve with multiple classifiers (Source: (Olson & Delen, 2008))

4.7 CONCLUSION

This chapter introduced the design of experiments for a machine learning problem, where the researcher started off by identifying what constitutes features in machine learning. During this process, features were manually identified from source data's attributes. It was then shown how features are converted into a numerical format that can be received as input and manipulated by a learning algorithm, and all of this was covered in section 4.2. As features presentation was discussed, the researcher in section 4.3 covered features selection, section 4.4 covered feature selection for unstructured data using data annotations. Section 4.5 covered model selection and how the classifier's parameters were optimized. Section 4.6 identified the computer setup for experimental purposes and also identified how the classifier's performance will be evaluated.

CHAPTER 5: Evaluations

5. EVALUATIONS

5.1 INTRODUCTION

This chapter presents the results of the experiments that were carried out throughout the progress of this study, and covers both results from the structured and the unstructured data. As has been shown in the previous chapters, Figure 5.1 is a continuation of the CRISP-DM framework, and in this chapter, the researcher will present and discuss the results of the experiments that were carried out based on the procedures outlined in Chapter Four. Firstly, the researcher evaluated the similarity measures that were used in this study, which are Jaro-Winkler, Edit distance, and (Term frequency and inverse document frequency). Section 5.2 presents the results of the similarity measure that were used. Section 5.3 gives results for the standardization of structured data, while section 5.4 covers the results of unstructured data. Section 5.5 to 5.6 are the discussions for both experiments.

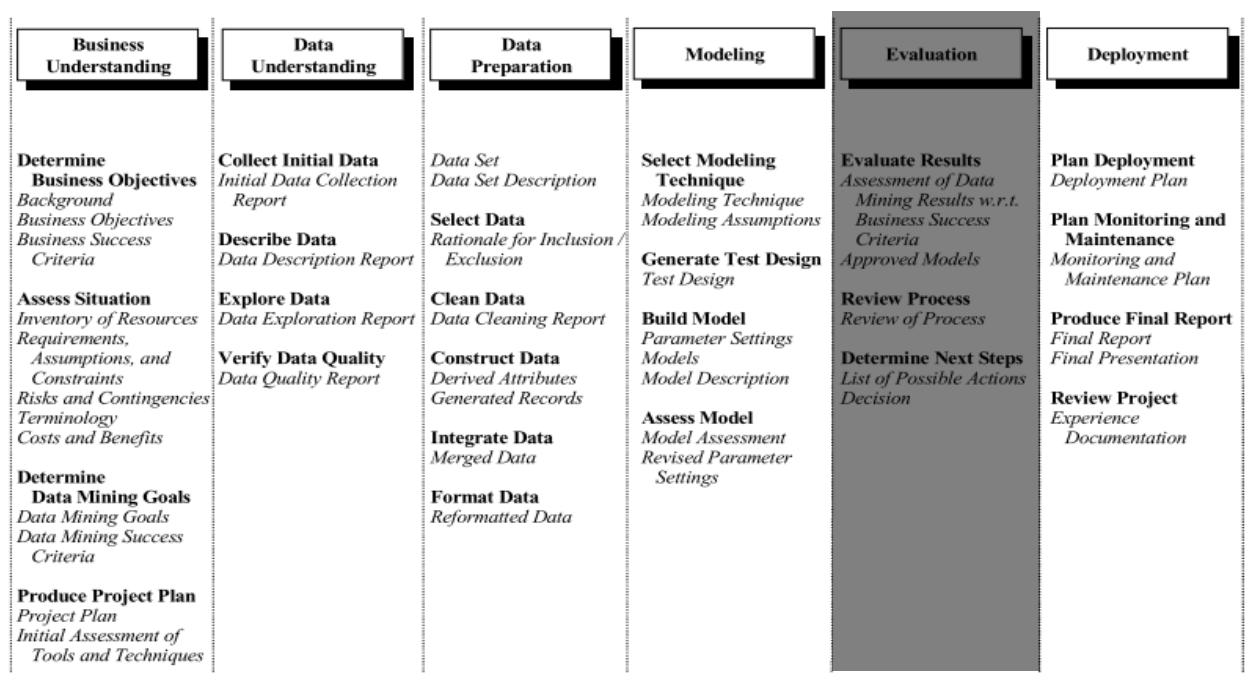


Figure 5. 1: CRISP-DM framework for model evaluation (Source: (Olson & Delen, 2008))

5.2 RESULTS FROM SIMILARITY MEASURES

The objective of the first experiment was to develop a solution that uses a Support Vector Machines (SVM) classifier to determine how to classify clinical observation data sets from multiple data sources through the prescription of health data coding standards. The researcher has attempted to fulfil this objective by collecting clinical observation data from multiple data sources. Similar data items are then matched using the learning algorithm, after which the algorithm is taught how to distinguish between similar tests that could not initially be detected. Firstly, results from Jaro-Winkler and Edit distance similarity algorithms are shown in Table 5.1A and Table 5.1B. Both tables record the same attributes. Attribute “S_obsname” is the source observation name, and the “S_UOM” is the source unit of measure, while the ones prefixed by “T” are the targets, and weights are calculated using Jaro-Winkler and Edit distance. From these results, it is evident that Jaro-Winkler is the best performing similarity measure when compared with Edit distance for clinical laboratory observation names and the expanded unit of measures. Edit distance did not perform poorly simply because the source observation name is syntactically dissimilar to the target observation name. It did so because the lengths of the strings have a negative impact on the performance of the similarity function.

Table 5. 1A Edit distance similarity results for observation name and unit of measure

Row#	S_obsname	T_obsname	Weight	S_UOM	T_UOM	Weight
1	mean blood pressure non invasive	mean platelet dry mass	0.3438	millimeters of mercury	picograms	0.1364
2	mean blood pressure arterial	mean platelet dry mass	0.3214	millimeters of mercury	picograms	0.1364
3	blood urea nitrogen	blood flow.mean	0.3684	millimeters of mercury	milli liters per second	0.5652
4	mean blood pressure non invasive	mean platelet component	0.3125	millimeters of mercury	grams per deciliter	0.1818
5	mean blood pressure arterial	mean platelet component	0.2857	millimeters of mercury	grams per deciliter	0.1818
6	mean blood pressure non invasive	mean sphered cell volume	0.3438	millimeters of mercury	fluid	0.0909

Table 5. 1B Jaro-Winkler similarity results for observation name and unit of measure

Row#	S_obsname	T_obsname	Weight	S_UOM	T_UOM	Weight
1	mean blood pressure non invasive	mean platelet dry mass	0.7395	millimeters of mercury	picograms	0.5148
2	mean blood pressure arterial	mean platelet dry mass	0.8029	millimeters of mercury	picograms	0.5148
3	blood urea nitrogen	blood flow.mean	0.765	millimeters of mercury	milli liters per second	0.7851
4	mean blood pressure non invasive	mean platelet component	0.7249	millimeters of mercury	grams per deciliter	0.6302
5	mean blood pressure arterial	mean platelet component	0.7405	millimeters of mercury	grams per deciliter	0.6302
6	mean blood pressure non invasive	mean sphered cell volume	0.7618	millimeters of mercury	Fluid	0.3303

5.2.1 MATCHING DISCUSSION

The objective of applying the similarity measures was to calculate how similar each source item is to the target item. The results presented in Table 1B gives evidence that Jaro-Winkler is the best performing algorithm for measuring the similarity of observation names. The Jaro-Winkler algorithm has been used previously to match people's names, street names, and surnames; and according to (Taburt, 2011) it performed better than Edit distance for short strings. Although Jaro-Winkler performed better than Edit

distance for observation names, the latter algorithm is also powerful when comparing two strings that have approximately the same length and fewer spelling mistakes (Doan, Halevy, Ives et al., 2012). As a result, the researcher has used Edit distance for short observation names that could not be matched using the soundex indexing algorithm. An example of this is when the soundex yielded a code of “B300” and “T510” for “bdy temp” and “temp” respectively, and Jaro-Winkler yielded a similarity weight of “0”. When using soundex, non-matching codes are an indication that the two tests are not similar, and therefore there is no need to compare other features since the observation names do not match. However, it is a given that similar tests will not always sound the same such as “bdy temp” and “temp”. For this reason, the researcher supervised the comparisons by applying another algorithm, in this case, edit distance. When Edit distance was applied to the tests mentioned above, a match of 50% was achieved, this was a good sign for the rest of the features to be compared between the source and the target. Using the rest of the features proved that “bdy temp” and “temp” are a match. Nevertheless, there were also cases where Edit distance failed to find a match between short observation names. The source short observation name was “art bp sys” and the target short observation name was “sys bp”. The “art bp sys” observation name is from the MetaVision dataset (source), while “sys bp” is from a LOINC data table (target). When comparing the similarities between these observation names, Jaro-Winkler yielded a similarity weight of 0.6, while Edit distance yielded a weight of 0.3. Therefore, in such cases where tokens of the observation name are similar although structured differently, the researcher used the Term Frequency Inverse Document Frequency (TFIDF) algorithm. This algorithm produced a similarity weight of 0.707, meaning that “sys bp” has a more than 70% chance of being similar to “art bp sys”. Since the short observation name could not be assessed with one similarity measure, the researcher has implemented three methods for measuring similarity between short observation names, namely, Edit distance, Jaro-Winkler and TFIDF. The one with the highest weight was used as a feature for short observation names.

5.3 FIRST EXPERIMENT: STRUCTURED DATA

The weights produced by the similarity measures are used as input features for the classifier to make predictions. This section is expected to show the reader how the data was trained, tested and how the predictive model was generated. The training data consisted of data that had already been mapped to a LOINC standard. Data that was already mapped to a coding standard was treated as the target dataset, while the one to be mapped was the source dataset. The researcher's task was to apply the model produced during the cross-validation process to the unstandardized data and predict whether the model could correctly classify the new data. Firstly, SVM's performance is covered, then thereafter the performance of SVM is compared with the one from Logistic Regression and later compared with the decision tree classifier. All these classifiers went through a process of parameter optimisation, cross-validation, training, testing, and prediction.

5.3.1 SUPPORT VECTOR MACHINES

The process of optimising hyperparameters gamma (γ) and cost (C) took approximately 30 minutes for the SVM classifier. The cost (C) parameter was set to iterate 10 times in the search for an optimum value, while gamma (γ) was set to iterate 10 times. Therefore, the pair of parameters executed for around 100 steps multiplied by the number of folds selected for the cross-validation, which was 10, the optimization process resulted to approximately 1000 models created.

Table 5. 2A Confusion matrix for Support Vector Machines

	true false	true true	class precision
pred. false	890	86	91.19%
pred. true	75	879	92.14%
class recall	92.23%	91.09%	

Table 5.2A shows a confusion matrix as a summary of the results shown in Figure 5.2A. These results were produced during the search for optimum parameters for the SVM

classifier. This output is from the RapidMiner software, and some of the performance variables in Figure 5.2A were repeating. Thus, for reasons of brevity, the researcher has discarded the rest of the iterative steps and only included the ones that select optimum performance for the model and the parameters. One should note that the final output of SVM classifier output in Figure 5.2A produced an accuracy of 91.63%, with C parameter set to 200.0008 and the γ parameter set to 0.0900811. Tuning the parameters is necessary for controlling overfitting and underfitting. A large value for the C parameter ensures that the positive examples are separated from negative examples and vice versa through the decision boundary. However, a large value for the C leads to overfitting, because the model tries to perfectly fit all the training examples, and when new examples are added, it then becomes difficult for the model to generalise to new examples if all it knows is to fit the training data accurately. When C is too small, the model underfits the data, which is called high bias, where the goal that the researcher is trying to reach is to have a value of C that is not too small and not too big.

```

PerformanceVector:
accuracy: 91.63% +/- 1.74% (mikro: 91.66%)
ConfusionMatrix:
True:   false   true
false:  890     86
true:   75      879
classification_error: 8.37% +/- 1.74% (mikro: 8.34%)
kappa: 0.832 +/- 0.035 (mikro: 0.833)
AUC (optimistic): 0.926 +/- 0.011 (mikro: 0.926) (positive class: true)
AUC: 0.927 +/- 0.011 (mikro: 0.927) (positive class: true)
AUC (pessimistic): 0.927 +/- 0.011 (mikro: 0.927) (positive class: true)
precision: 92.10% +/- 2.64% (mikro: 92.14%) (positive class: true)
recall: 91.11% +/- 1.67% (mikro: 91.09%) (positive class: true)
f_measure: 91.58% +/- 1.69% (mikro: 91.61%) (positive class: true)
sensitivity: 91.11% +/- 1.67% (mikro: 91.09%) (positive class: true)
positive_predictive_value: 92.10% +/- 2.64% (mikro: 92.14%) (positive class:
true)
true:   75      879
negative_predictive_value: 91.14% +/- 2.00% (mikro: 91.19%) (positive class:
true)

SVM.C = 200.0008
SVM.gamma = 0.0900811

```

Figure 5. 2A: Results from running a 10-fold cross-validation and grid-search for parameter optimisation of SVM

It should be noted as well that changing C has a direct impact to the gamma parameter (Ben-Hur & Weston, 2010), and this change affects the performance of the model.

Therefore, big and small values of C are relative to the data, as well as to the gamma parameter. Equation (13) and Equation (14) shows how the gamma parameter influences the x_j and x_i values. If x_j is a support vector while γ holds a small value, then the class (positive or negative) of the support vector will determine how x_i should be classified. Support vectors are data points that are closest to the decision boundary of the SVM, and are thus helpful for determining whether a new training data point should be classified on a positive or a negative class (Ben-Hur & Weston, 2010; Chih-Wei Hsu, Chih-Chung Chang, 2008). One should observe that there were a total of 965 positive training examples and 965 negative training examples, and that the total number of training examples was 1930. This number was reduced from 2483 to 1930, with the goal of balancing the number of positive examples to negative examples. Without balancing these numbers, the model would produce inaccurate output, and (Longadge & Dongre, 2013) have warned about the danger of imbalanced or skewed classes. When there are imbalanced classes the minority classes have a high chance of being misclassified. Therefore, the researcher had to select the maximum number of positive examples, since they were the minority classes and the proportions were decided based on the minority classes. For testing the model, 800 testing examples without the output value were selected through the stratified sampling methods in order to ensure equal set distribution between the data. All the 1930 records were used for training and testing the SVM, Logistic Regression and the decision tree classifier.

5.3.2 MULTIPLE MODEL PERFORMANCES

As part of parameter optimisation, Logistic Regression classifier ran for less than one minute with the same setup as SVM regarding the number of iterations for the selection of the parameters. The grid-search yielded a lambda parameter value of 0.001 and the alpha parameter value of 0.7. The prediction accuracy of the model sat at 88.89% as shown in Figure 5.2B, and the classification error was 11.11%.

Table 5. 2B Confusion matrix for a Logistic Regression classifier

	true false	true true	class precision
pred. false	864	113	88.43%
pred. true	101	852	89.40%
class recall	89.53%	88.29%	

When compared with the SVM performance, the Logistic Regression classifier had a total of 1716 correct predictions, while SVM had 1769 correct predictions. Correct predictions are the diagonal measures from the top position of the confusion matrix, in the case of Logistic Regression the values are 864 and 852. The 864 value represents the number of negative examples that have been correctly predicted by the classifier to be negative, and the bottom value of 852 represents the number of positive examples that have been predicted to be positive by the classifier. The value 101 represents the number of negative examples that the classifier failed to predict as negative, and the value 113 are positive examples that the classifier failed to predict as positive. A study by (Kim, XuYu, & Unland, 2011) specified that the accuracy of the model can be calculated as shown in Equation (29), whereby TN are the total number of true negatives, TP is the total number of true positives, FN is the total number of false negatives, and FP is the total number of false positives.

$$Accuracy = \frac{TN+TP}{TP+TN+FN+FP} * 100 \quad (29)$$

The recall value of 88.29% reveals that out of all positive training examples, the classifier was only able to predict 88.29% as positive, and when it came to negative examples, the classifier was able to predict 89.53% as negative. The recall measure looks at the actual examples and calculates how much of the actual examples were predicted correctly, which is unlike precision, which examines what has been predicted, and then calculates how much of the predicted examples are actually true.


```

PerformanceVector:
accuracy: 88.89% +/- 2.54% (mikro: 88.91%)
ConfusionMatrix:
True:   false   true
false:  864     113
true:   101     852
classification_error: 11.11% +/- 2.54% (mikro: 11.09%)
kappa: 0.777 +/- 0.051 (mikro: 0.778)
AUC (optimistic): 0.912 +/- 0.016 (mikro: 0.912) (positive class: true)
AUC: 0.912 +/- 0.016 (mikro: 0.912) (positive class: true)
AUC (pessimistic): 0.912 +/- 0.016 (mikro: 0.912) (positive class: true)
precision: 89.36% +/- 3.63% (mikro: 89.40%) (positive class: true)
recall: 88.32% +/- 3.15% (mikro: 88.29%) (positive class: true)
f_measure: 88.78% +/- 2.64% (mikro: 88.84%) (positive class: true)
sensitivity: 88.32% +/- 3.15% (mikro: 88.29%) (positive class: true)
positive_predictive_value: 89.36% +/- 3.63% (mikro: 89.40%) (positive class:
true)
negative_predictive_value: 88.43% +/- 3.10% (mikro: 88.43%) (positive class:
true)

alpha = 0.7, lambda = 1.0E-4

```

Figure 5. 2B: Results from running a 10-fold cross-validation and grid-search for parameter optimisation of Logistic Regression

The recall and precision are helpful when diagnosing the outputs produced by the classifier. The last classifier comparison is the SVM against the results produced by the decision tree classifier. Table 5.2C and Figure 5.2C give the results of the decision tree, whose classification criterion was also executed via the grid-search for optimum values.

Table 5. 2C Confusion matrix for a decision tree classifier

	true false	true true	class precision
pred. false	914	83	91.68%
pred. true	51	882	94.53%
class recall	94.72%	91.40%	

For the decision tree, there were no parameters that were selected, however, a minimal gain and a classification criterion was selected between four possible criterions: namely information gain, Gini index, gain ratio, and accuracy. The decision tree had a confidence value of 0.25, and the maximum depth of the tree was set to 10, and the tree

was also set up to apply pruning. From Figure 5.2C, a Gini index classification criterion was selected for the decision tree with a minimal gain of 0.01. The decision tree outperformed the SVM classifier on the given training data. An accuracy of 93.05% was produced by the decision tree classifier. This is not a rare performance, where other studies such as those of (Kirkos, Spathis, & Manolopoulos, 2008) have found the C4.5 decision tree outperforming SVM and neural networks. Accuracy has been the preferred method to record the performance of the results. A classic study by (Ling, Huang, & Zhang, 2003) has proven that ROC's AUC is a better and more reliable measure than accuracy. Considering that SVM has a higher AUC value than that of the decision tree, then this is a sign that more assessment needs to be made on the training data. Figure 5.3 shows the ROC (receiver operating characteristic) curve for all three classifiers. The ROC curve was used to compare the performance of each classifier. The ROC output is used for determining an effective threshold so that values that are above the threshold represent a specific classification event.

```

PerformanceVector:
accuracy: 93.05% +/- 1.33% (mikro: 93.06%)
ConfusionMatrix:
True:   false  true
false:  914    83
true:   51     882
classification_error: 6.95% +/- 1.33% (mikro: 6.94%)
kappa: 0.861 +/- 0.026 (mikro: 0.861)
AUC (optimistic): 0.905 +/- 0.018 (mikro: 0.905) (positive class: true)
AUC: 0.915 +/- 0.014 (mikro: 0.915) (positive class: true)
AUC (pessimistic): 0.926 +/- 0.015 (mikro: 0.926) (positive class: true)
precision: 94.53% +/- 1.61% (mikro: 94.53%) (positive class: true)
recall: 91.45% +/- 1.42% (mikro: 91.40%) (positive class: true)
f_measure: 92.95% +/- 1.14% (mikro: 92.94%) (positive class: true)
sensitivity: 91.45% +/- 1.42% (mikro: 91.40%) (positive class: true)
positive_predictive_value: 94.53% +/- 1.61% (mikro: 94.53%) (positive class:
true)
negative predictive value: 91.59% +/- 2.02% (mikro: 91.68%) (positive class:
true)

Decision Tree.criterion      = gini_index
Decision Tree.minimal_gain   = 0.01

```

Figure 5. 2C: Results from running a 10-fold cross-validation and grid-search for parameter optimisation of decision tree

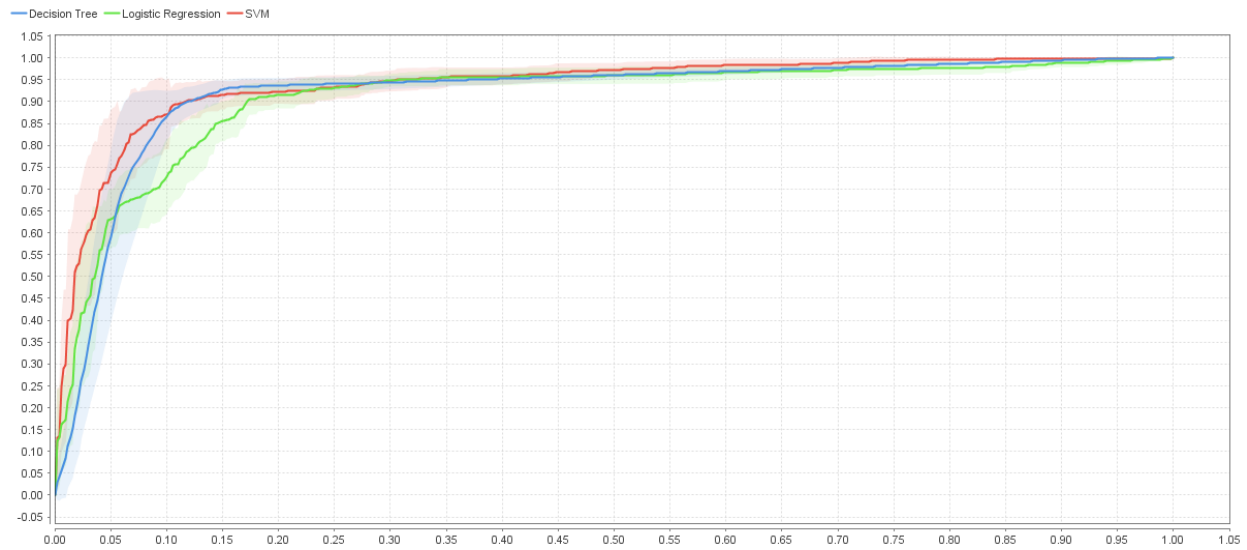


Figure 5. 3: ROC curve for decision tree, Logistic Regression and SVM

A good model yields a curve that lies on the North-Western part of the plot, while a bad model yields a curve that is far from the north-western position. The x-axis of the curve represents the rate of false positives, while the y-axis is the rate of true positives. The ROC is meant to address details that were missed by the accuracy measure and the classification error.

5.4 SECOND EXPERIMENT: UNSTRUCTURED DATA

In this section of the study, the researcher aims to show the results obtained from the application of classification rules and machine learning algorithms on unstructured data using Conditional Random Fields sequence classifier. The task was to extract meaning from unstructured data, then standardize it in order to enable searchability, comparability and exchangeability. The task involved the extraction of smoking information and determining if the patient is a current smoker; is a past smoker; or is a non-smoker. Therefore, there are three classes from which each document should be classified, and as thus the classifier is evaluated on its ability to correctly assign an appropriate tag or class on the correct document based on the gold standard. This means that if the gold standard matches with the predictions made by the classifier,

then that is regarded as a correct prediction. The CLAMP software uses the CRFSuite library (Okazaki, 2007) and this library outputs the precision, recall and the F1-measure score. These evaluation measures were produced for all the five folds that were executed for selecting the best model and for optimizing the parameters. In each fold, CRFSuite uses the Limited-memory Broyden-Fletcher-Goldfarb-Shanno (L-BFGS) algorithm for estimating the CRF parameters, and default settings were used for the CRF parameter. A paper by (Okazaki, 2007) has more details on CRFSuite implementation. Table 5.3 represents the summarized results for the 5-folds of cross-validation which executed for a minimum of three hours for each model. It must also be noted that the results in this section are presented differently than on the previous section, since more than one class is predicted. Instead of representing results through a confusion matrix, the researcher will thus present the results through a micro and macro-averaging for the precision, recall and F-measure scores.

Table 5. 3 Results for smoking status detection produced by a CRF sequence classifier

	Output from CLAMP prewritten rules (A)						Output from customized rules (B)					
	P	R	F1	TP	Prd	G	P	R	F1	TP	Prd	G
Past Smoker	0.788	0.882	0.832	82	104	93	0.724	0.750	0.737	84	116	112
Current Smoker	0.833	0.784	0.808	80	96	102	0.574	0.684	0.624	39	68	57
Non-Smoker	0.783	0.722	0.751	65	83	90	0.714	0.652	0.682	60	84	92
Macro Avg.	0.801	0.796	0.797				0.670	0.695	0.681			
Micro Avg.	0.8021	0.8438	0.822				0.6829	0.7011	0.692			

Precision, Recall and F-Measure

The researcher started training the model with annotations that were produced from executing both prewritten rules and customized rules. Prewritten rules are represented as output “A”, and annotations from customised rules are represented as output “B” (see Table 5.3). Therefore, these results will be referred to as specified throughout this section. Table 5.3 shows an abbreviated version of the evaluation measure, the items below give an expanded version of these abbreviations:

- P: Precision
- R: Recall
- F1: F-measure
- TP: True positives
- Prd: Prediction count
- G: Gold standard

The findings show that output “A” has outperformed output “B” for all the predictions. There were 112 “pastsmoker” annotations that met the gold standard for output “B”, however, according to output “A”, there are only 104 annotations for the “pastsmoker” class. Now based on the gold standard, it can be said that none of the annotations agree to have the same gold standard, it is only the “nonsmoker” class that has approximately the same quantity of annotations between output “A” and output “B”. Furthermore, it can be observed that the number of true positives are 5 annotations apart, and the number of predictions differ by 1 between the two outputs. Another observation is that the “pastsmoker” has the highest F-measure for both output “A” (0.832) and output “B” (0.737), meaning that there is a balance between precision and recall for this class, however there is a need for improvement. The researcher has further computed the micro and macro-averaging since there are multiple classes to be predicted. (Wang & Domeniconi, 2008) say that the micro-average is used for computing the average precision at a document level, or at an annotated corpus level, unlike macro, which computes the average precisions from all the classes that are used. One of the key takeaways is that macro-averaging has more influence on smaller classes, and on the other hand micro-averaging has a high measure of effectiveness on larger classes (Manning et al., 2009).

Furthermore, the precision and recall results give more information about the class distribution and the correctness of the methods used for identifying correct classes. The class with both the highest precision and recall is a sign that the rules that were used were able to detect the smoking statuses in the given corpus. In addition, the test data had enough tests for the calculation of predictions for the same class, meaning there was a good class coverage. A true reflection of such a case was the precision of the “currentsmoker” class for output “A”, which was 83%, while the recall was 78%, thus indicating a roughly balanced trade-off between precision and recall. Therefore, the annotation improvements will be based on enhancing class coverage on the test set, and also improving the rules for class detection.

5.5 FIRST EXPERIMENT DISCUSSIONS

The researcher has demonstrated that laboratory data that is standardized in LOINC could be used to formulate a predictive model, which could be used to predict the LOINC codes that should be assigned to unstandardized data. Firstly, the researcher discusses the results obtained through the performance indicators.

ROC results

The ROC curves are less biased by the class distribution; these curves are used together with AUC, which measures how good the area under the curve is; where the larger the curve, the better the model. The results presented in Figure 5.3 were made based on the following setup, a 10-fold cross-validation, with a split ratio of 0.9 and the sampling was set to “shuffled” for the ROC function. From the ROC curves, it can be observed that SVM outperforms the rest of the models, where the SVM has a higher number of true positives while incurring a small percentage of false positives, however this was based on small sample. According to (Witten & Frank, 2011), the results of the ROC are interpreted based on the shaded area, or convex hull. The study by (Witten & Frank, 2011) argues that one should always operate at the upper boundary of the

convex hull. Other measures were also used to determine the correctness of the generated models.

Recall and Precision

Recall was used to determine records that were mistakenly predicted to match whilst they actually not matching. Precision then looks at all the predicted records to be matching, determining the fraction of them that are actually matching. Recall was used to check the prediction coverage of the classifier, because prior to predictions, it is already known how many records match and those that do not match. If the recall gives an output with a lesser number than the actual number, it is a sign of an incorrect classifier. A high recall is an indication that there are less chances of misclassifying a non-matching record as a match. A high precision means that the chances of misclassifying a matching record as non-matching was small. A recall of 91.5% for the decision tree and 91.11% for SVM is a good result, although, on the other hand, the precision was also high. These two measures were balanced by the high F-measure score, because having a high precision trades off recall and vice versa.

Error analysis

The researcher has found that matching and mapping laboratory data to a standard is a laborious exercise that needs time and skill to perfect, as noted by (Abhyankar et al., 2012; Lee, Groß, Hartung, Liou, & Rahm, 2013). Hence the researcher has attempted to solve the problem through a machine learning process. As the results have been shown on the previous sections, it was not enough to just accept these results without questioning their accuracy. Error analysis was applied for checking the correctness of the produced models. During error analysis, the researcher checked for high bias and high variance. There are various methods for testing high bias and high variance. An expert in machine learning (Ng, 2011) suggests that one of the methods for detecting high bias and high variance is through the observation of the training error and the testing error. Having said that, a random sample of 100 sets from the training data was then used to train the classifier. At that time, the researcher used the SVM classifier

because (Singh, 2010; Xu, Caramanis, & Mannor, 2008) said SVM has proven to be a robust classifier. The result of the error analysis is shown in Figure 5.4 where it shows that when more training sets are added, the classifier's test error decreases, resulting in a logarithmic curve, and it was the same with the train error, however, in an opposite direction. The objective of this was to get a test error value that was close to the training error value, provided there was low test error and training error. The problem with the tests produced, was a high recall and high precision, which was a positive attribute of a good classifier. However, the researcher went on to manually assess the predicted results. It was discovered that the classifier could not distinguish between common tests such as cholesterol ldl and cholesterol hdl. This was caused by the fact that the similarity weight between the two observation names was 0.9522, and that the rest of the weights for other features gave a high score, which meant that the two records were the same. If adding more features would fix the problem, then (Ng, 2011) says that is a sign that the model is highly biased.

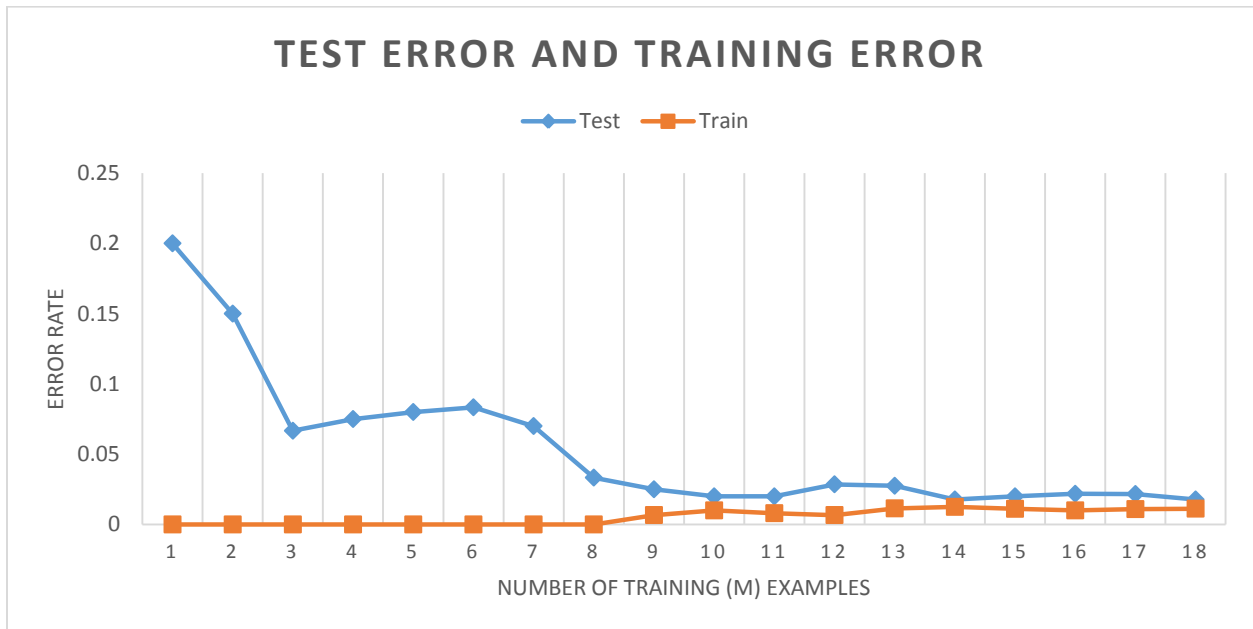


Figure 5. 4: SVM model classification error

Therefore, the researcher revisited the features and identified more features to be used for the classifier.

Feature refinery for distinct matching

The researcher has discovered that adding more distinct features improves the performance of the classifier. This ultimately removes false matches, and these findings are similar to that of (Kum, Krishnamurthy, Machanavajjhala, Reiter, & Ahalt, 2014). The features that were added include the following: *valuerange*, *hastimeaspect*, *sameunits*, *patientgender*, *testtime*, and the *testrank*.

- *valuerange*: checks whether the recorded source observation value is in the same range as the target observation.
- *hastimeaspect*: checks if the target observation is measured at certain intervals, for instance, whether blood pressure is being measured every hour, and there is specific LOINC code used for identifying such tests.
- *sameunits*: removes the metric value on a unit of measure, e.g. mg/L would be g/L. Another test may use a different metric such as kg/L, and the “SameUnits” feature would register the two units as similar.
- *patientgender*: captures the patient’s gender, where some tests differ by gender.
- *testtime*: refers to the time when the source tests was taken. The value of this feature was calculated for a patient by specifying a test day, and then checking how often the same test occurs. If the test is done on every specific interval, then variable “TestTime” will store the value of the interval. If, for instance, a test is done every hour, then the “TestTime” features will store a value of “hourly”.
- *testrank*: The LOINC top 2000 document includes 98% of the tests from three large institutions and the document ranks tests based on the observed usage.

Other researchers have also seen the need to extend observation names. (Kim et al., 2012) have discovered that extending the observation name improved LOINC mapping. The researcher has also extended the observation names, unit of measure, and test category. In fact, when the decision tree model was run, the *expUomWeight* and

UomWeight were the most dominant feature at the root of the tree (see Appendix C for the decision tree model). Existing literature in laboratory data standardization has shown that the *unit of measure* feature is important and necessary for identifying tests (Abhyankar et al., 2012; Fidahusseini & Vreeman, 2014; Lin, Vreeman, & Huff, 2011). (Vreeman, Hook, & Dixon, 2015) discovered that LOINC-mappers find it informative when they learn how other mappers map from other organizations.

They therefore introduced a ranking attribute, which was used to see how often other LOINC-mappers have mapped to the test in question. As a result, these features were used to disambiguate common tests that have the same observation name, which caused confusion about which one should be used for mapping. The tests were ran again for both SVM and decision tree, however this time for a small sample of 200 tests. The decision tree achieved an accuracy of 91%, while the SVM had achieved an accuracy of 92.50% (see Table 5.4A and Table 5.4B, respectively).

Table 5. 4A Confusion matrix for a decision tree classifier with accuracy: 91.00% +/- 5.83% (mikro: 91.00%)

	true false	true true	class precision
pred. false	132	12	91.67%
pred. true	6	50	89.29%
class recall	95.65%	80.65%	

Table 5. 4B Confusion matrix for the SVM classifier with accuracy: 92.50% +/- 5.12% (mikro: 92.50%)

	true false	true true	class precision
pred. false	135	12	91.84%
pred. true	3	50	94.34%
class recall	97.83%	80.65%	

It is realised that the sample used in these figures is small, due to an extensive process of identifying extra features for disambiguating common tests. The idea of a small dataset was meant to remove complexity in the mapping process, so that one can easily identify what causes bias and variance or errors in the training data. The researcher tested the performance of the SVM classifier by adding 22 unlabelled records to be predicted by the classifier, where the classifier was able to correctly predict 62% of the unlabelled records. This was a good indication that the extra step of feature engineering was necessary, and this was a sign that on a bigger dataset the prediction accuracy would increase, since a greater amount of data has proven to reduce the classification error while improving the accuracy.

5.6 SECOND EXPERIMENT DISCUSSIONS

The results of the second experiment have shown that the annotations that were produced through the predefined rules have outperformed the rules that the researcher has written. The rules that were written by the researcher were able to differentiate between the “currentsmoker” and the “pastsmoker” based on the time frame. Meaning that if the date of quitting smoking for the patient is less than a year, then such a record is classified as a “currentsmoker”, otherwise it is classified as a “pastsmoker”. Although the predefined rules yielded the best classification performance for all the classes, these rules only looked for the occurrence of keywords such as “former”, “quit”, “no longer” for assigning the “pastsmoker” class. Also, word features that represent the past (such as history or used to) were used together with the smoking-based words (such as smoking, smokes, tobacco, or cigarette) for identifying various types of smokers. For instance, the phrase “Tobacco: 40 year history of smoking” would be classified as a “pastsmoker” because of the keyword “history” and “smoking” appearing together.

The results from output “A” in Table 5.3 does not seem to have followed the classification rules that were defined by (Uzuner, Goldstein, Luo, & Kohane, 2008) This observation emanates from records that were classified incorrectly as shown in Figure

5.5A and Figure 5.5B. Therefore, it can be said that the annotation on Figure 5.5B is correct because someone that has quit smoking less than a week ago should be classified as a “currentsmoker” instead of as a “pastsmoker”.

1 Social History: Patient states he quit **smoking** **PastSmoker** 1 week ago, although he smells of cigarettes. He states that he does not currently drink or usedrugs. He lives at home, is on disability and takes care of himself.

Figure 5. 5A: Annotations from CLAMP’s predefined rules

1 Social History: Patient states he quit **smoking** **CurrentSmoker** **temporal** 1 week ago, although he smells of cigarettes. He states that he does not currently drink or usedrugs. He lives at home, is on disability and takes care of himself.

Figure 5. 5B: Annotations from the custom developed rules

However, the classifier from output B has performed poorly for the classification tasks, and the researcher also identified that within a single annotated corpus, there were sometimes more than one class. For illustration purposes see Figure 5.6.

1 Social History: **CurrentSmoker** +current tobacco use but has not **NonSmoker** smoked since surgery 1 mo **temporal** ago.[**12-30**] drinks ETOH per week Lives with husband, no children. Currently works as a secretary

Figure 5. 6: A double class annotation where a record is classified as “currentsmoker” and “nonsmoker” at the same time.

Since there were many classes that were identified on a single record as shown in Figure 5.6, the accuracy of the produced model was reduced when a single annotated corpus was classified to more than a single class because the number of gold standard records would increase, which would result in more false negatives and ultimately make the recall value lower than it is supposed to be. Therefore, as part of data pre-processing, the researcher had to rewrite the rules, and also improve the training time so that it becomes easier and more efficient to train the models. Also, one of the challenges that the researcher had experienced regarding the writing of rules was that he was not adept at UIMA RUTA for multi-class detection rules. Therefore, instead of writing rules for annotating the three classes at one go, he resorted to writing rules for identifying two classes at a time. Thus, the first rules were between “nonsmoker” class and “smoker” class, the “smoker” class includes both “currentsmoker” and “pastsmoker”. Then the second set was based on the “nonsmoker” and the “pastsmoker” class. The reason for these rules was to cover phrases such as “He has a history of tobacco use, but does not smoke currently.” Initially this was classified as both “currentsmoker” and “nonsmoker”. The classifier predicted that it is a “currentsmoker” because of the use of words such “history of smoking” without specifying the time frame when the person quit smoking, and it was classified also as “nonsmoker” because of the phrase “does not smoke”. Then the third and last set of rules were between the “currentsmoker” and the “pastsmoker” classes.

Efficiency and error analysis

The three hours of training the model was undertaken because a full corpus was loaded instead of only loading text that contained smoking information. Each annotated corpus contained approximately a minimum of 3500 words, and within that corpora, information that was of interest to the researcher was about 20 to 100 words. However, in order to only load the relevant information onto a file for training, one had to open each document and search for the required information, then copy and save the extracted text into a new file. Initially, the researcher used this method which was lengthy and cumbersome. However, because of this inefficient method, he then decided to write a

simple C#.net windows program that could read the clinical corpora and extract the relevant information and save it with a proper name that eases the process of annotation preparation and data training. (See Appendix D for the screenshots, and the code which has been shared on Google drive as shown in Appendix E). Extracting the relevant information did not only help with the inefficient processes, but it also gave the researcher the opportunity to train with more and relevant data, since it is known from a classic study by (Banko & Brill, 2001) that the algorithm's performance improves as more relevant data is added. When the corpora were shortened into relevant text, the annotation process took less than 2 minutes for 195 records, and training and testing as well took less than 5 minutes which was a massive improvement. Therefore, the researcher took advantage of this and added more training and test data.

Re-evaluating the model

Following the CRISP-DM framework allowed one the flexibility to revisit the data collection and preparation process frequently, even after the model was tested and evaluated. One should remember that the model was tested on a gold standard that the researcher was satisfied with, which meant that when new training and test data was added, the rules that were used to generate the gold standard were not changed, however the ones that generated double classes (see Figure 5.6) were updated so that only one class was selected. Therefore, the researcher collected more training and testing data which amounted to a total of 1242 annotated corpus, and for training and testing the researcher continued with the k-fold cross validation method. The same word representation features were selected as discussed in section 4.6.

The classifier was trained and tested and the summarized results for the F-measure score was 94.4% for "nonsmoker", 54.1% for "pastsmokers" and 80.2% for "currentsmokers", see Table 5.5. These results showed a sharp increase for both "nonsmoker" and "currentsmoker" records, while the "pastsmoker" performance decreased. The performance of the "nonsmoker" class has surpassed that of (Liu et al.,

2012) for document-level classification. (Liu et al., 2012) did a similar study where they focused on transferability of the smoking status detection module at different institutions, and their results for “nonsmoker” detection have shown an F-measure of 97% for sentence-level classification, 93% for document-level classification and 87% for patient-level classification. However, the current study was not specific on the type of clinical notes, as all training and testing was done at a document level and each document represented a unique patient from the MIMIC-III database.

Table 5. 5 An earlier test results for the Named Entity Extraction for the patient’s smoking status and other relevant information through a CRF sequence classifier

Output from customized rules						
	P	R	F1	TP	Prd	G
CurrentSmoker	0.807	0.797	0.802	467	529	518
NonSmoker	0.969	0.919	0.944	569	587	619
PastSmoker	0.521	0.562	0.541	223	428	397
Macro Avg.	0.7656	0.7593	0.7623			
Micro Avg.	0.815	0.821	0.818			

An earlier study by (Sohn & Savova, 2009) had obtained a much higher F-measure of 97% for the “nonsmoker” detection class at a document-level, while a recent study by (Liu et al., 2012) obtained an F-measure of 93%. Getting more relevant training data has proven to have more influence on the performance of the algorithm. As the researcher added more training data, the F-measure of the “pastsmoker” increased from 0.54 to 0.657, while recall also increased to 0.669. Increasing precision means that false positives are reduced, and also increasing recall means that false negatives are also reduced, and the balance between precision and recall is important for improving the performance of the algorithm.

Table 5. 6 Later test results for the Named Entity Extraction for the patient’s smoking status and other relevant information through a CRF sequence classifier

Output from customized rules						
	P	R	F1	TP	Prd	G
CurrentSmoker	0.839	0.821	0.830	439	523	535
NonSmoker	0.973	0.925	0.948	613	630	663
PastSmoker	0.646	0.669	0.657	410	635	613
Macro Avg.	0.8193	0.805	0.8116			
Micro Avg.	0.838	0.807	0.822			

It should also be noted that these results were generated by the same rules that were used to build the gold standard, and when more new data was added for training and testing, the performance increased rather than decreased which might be a sign that the rules that the researcher had defined were robust to the change in data, which implies that they could be implemented for extracting smoking status from other institutions. However, the classifier produced poor performance for the “pastsmoker” even when more data was added, so one can see that from a total of 613 gold standard records for the “pastsmoker” class, only 410 were correctly predicted which resulted in a precision of 64.6% which was still poor. Apart from the “pastsmoker” results, it is worth mentioning that these results were as good as the data that was used, in this case the MIMIC-III data. Therefore, the results might be influenced or biased to the manner in which the health clinician captured the data. Furthermore, the performance of these rules could further be tested on different data from different institutions. Part of the output produced by CLAMP includes a “.jar” file, which could be reused for annotation purposes on other projects. In the following chapter, the researcher discusses the meaning of these results in terms of interoperability in healthcare.

Mapping to a coding standard

The goal for extracting smoking status information was to organize this information so that it would be easier to search for and standardize how this information is represented and shared across different institutions and health systems. Also, once the information is extracted, more analysis could be made on the same information. In Figure 5.7 the researcher shows the predictions that were made by the classifier, and additional information was also extracted such as the smoking frequency and temporal information. Although the classifier was able to extract time-based information, it did not know that “since” could be used to identify the length of time that the patient has been smoking, therefore more research could be done for such cases.

1	Social History:
2	He lives alone but has a girlfriend and is in contact with his
	predict predict predict CurrentSmoker Frequency temporal
3	mother/sister. [**Name (NI) **] smokes 1.5 ppd since age 15. He has largely
	predict temporal
4	quit EtOH for the last 3 yrs but occasionally relapses and has a
5	pint of whiskey. He used to drink 0.5l hard alcohol. He uses
6	marijuana on occasion but denies IVDU.

Figure 5. 7: Predicted Named Entities from the CRF classifier

Figure 5.7 represents the resulting performance of the model that was produced from training the CRF classifier. Now on a production application, the produced model could be used to annotate data that has not been annotated without the need to go through the training and testing process again. Table 5.7 represents results from the annotated text that could further be mapped into a coding standard such as SNOMED-CT and LOINC, as mapping to a coding standard ensures that the data is exchangeable across different institutions and health systems. One can also observe from Table 5.7 that each entity that was extracted has been automatically mapped to CUI codes by the UMLS encoding algorithm. Now mapping to SNOMED-CT and LOINC would need one to write

a simple computer program that takes as input the predicted class name, e.g. “currentsmoker” and maps to coding standard code, see Table 5.8.

Table 5. 7 Results from executing rules on clinical text data

Start	End	Class	CUI	Entity Extracted
112	118	CurrentSmoker	C0037366	smokes
123	126	Frequency	C0032739	1.5 <u>ppd</u>
133	139	Temporal	C1850825	age 15
179	184	Temporal	C2302314	3 yrs

Mapping to a coding standard helps during data exchange and data sharing, and the researcher suggests that Fast Healthcare Interoperability Resources (FHIR) can be used for exchanging a patient’s coded information between health care institutions. FHIR represents clinical data as resources and each resource contains data that is represented by coding standards. It uses RESTFul API to exchange messages between two parties, and the messages could be represented in JSON, XML (Mandel, Kreda, Mandl, Kohane, & Ramoni, 2016) and now also includes a Turtle format. FHIR also uses profiles to group common use cases that are defined together, and the profiles contain data constraints, Value Sets and examples. Box 1 shows the use of FHIR profiles where a coding standard is used together with the identified class name (current some day smoker).

Table 5. 8 SNOMED-CT coding information according to the UMLS metathesaurus

Concept code	Concept name	Coding system
428071000124103	Heavy tobacco smoker	Current Heavy tobacco smoker
428061000124105	Light tobacco smoker	Current Light tobacco smoker
428041000124106	Current some day smoker	Current some day smoker
8517006	Former smoker	Former smoker
266919005	Never smoked tobacco	Never smoked tobacco
77176002	Current smoker	Current smoker
449868002	Smokes tobacco daily	Smokes tobacco daily
266927001	Tobacco smoking consumption unknown	Tobacco smoking consumption unknown

Box 1- Smoking status profile in a JSON file format

```

{
  "resourceType": "Observation",
  "id": "5-smokingstatus",
  "meta": {
    "versionId": "1",
    "lastUpdated": "2018-01-31T19:48:22Z"
  },
  "text": {
    "status": "generated",
    "div": "<div xmlns=\\"http://www.w3.org/1999/xhtml\\">Tobacco smoking status:
Current some day smoker</div>"
  },
  "status": "final",
  "category": {
    "coding": [
      {
        "system": "http://hl7.org/fhir/observation-category",
        "code": "social-history",
        "display": "Social History"
      }
    ],
    "text": "Social History"
  },
  "subject": {
    "reference": "Patient/1032702"
  },
  "issued": "2016-03-18T05:27:04Z",
  "valueCodeableConcept": {
    "coding": [
      {
        "system": "http://snomed.info/sct",
        "code": "428041000124106",
        "display": "Current some day smoker"
      }
    ],
    "text": "She is a past smoker, quit five years ago. She has a 50 pack year
history of tobacco usage."
  }
}

```

5.7 CONCLUSION

In this chapter, the researcher reports the results of the two experiments that were conducted throughout the duration of this study. This chapter references the evaluation stage of the CRISP-DM model. All the sections in this chapter are meant to cover as much detail as possible about the results of the experiments. Multiple similarity measure functions were used to evaluate whether the source string matches with the target string. In section 5.2, the researcher evaluated the performance of Jaro-Winkler against Edit distance for laboratory data, and Jaro-Winkler outperformed Edit distance. Then in section 5.3 the researcher went on to compare the classifiers that were used namely SVM, Decision Trees and Logistic Regression. It was discovered that the Decision Trees classifier outperformed SVM and Logistic Regression, while SVM performed second-best. However, when more distinct features were added then the SVM performed better than the Decision Trees classifier. In section 5.4, the researcher covered experiments that involved the testing of annotations that have been identified from clinical text data (corpora). Results were reported, and in section 5.5 the researcher discussed results that were obtained from training the structured data. Section 5.6 discussed the results from the unstructured data, and also suggested the use of FHIR resources for data exchange and clinical message representation.

CHAPTER 6

Conclusions and Recommendations

6. IMPLICATIONS OF THE FINDINGS

6.1 INTRODUCTION

In this chapter, the researcher focuses on attempts made to address the interoperability problem, where in section 6.2, the researcher discusses the framework used. Section 6.3 addresses the attempts made to solve the interoperability problem for both structured and unstructured data. Section 6.4 talks about the study limitations, outlook, and lessons learned.

6.2 IMPLICATION OF THE FINDINGS BASED ON CRISP-DM PROCESS

How can the process of data compliance across health care providers be automated through machine learning concepts?

This question detailed the step-by-step process that the researcher used in order to address the research problem. The researcher has used the CRISP-DM framework (see Figure 6.1) as guideline for conducting this study. CRISP-DM was used because the researcher has sought to solve this problem through data science concepts. Firstly, the researcher has identified the actual problem that this study aims to address. Namely, the lack of interoperability between health care providers. If health care systems are operated in silos, then there is a high chance that the data will not be semantically and syntactically interoperable. In Chapter One, the researcher addressed the causes of the lack of interoperability, and the effects of this problem were also addressed in the same chapter. Chapter Two addressed the properties of the data, in terms of what prevents this data from being interoperable. Relevant data was collected from different sources in order to simulate the problem that currently exists in health care. It was indeed true that data from different sources is stored differently, and this causes the data to be loosely structured, where ultimately, the data becomes non-interoperable. In Chapter Three, the researcher identified methods that could be used to normalize the data in such a way that it is easier to process on a computer. These

methods are data normalization, data cleaning and data preparation. In Chapter Four, the researcher identified the predictive models to be built for the data that has been identified for this study.

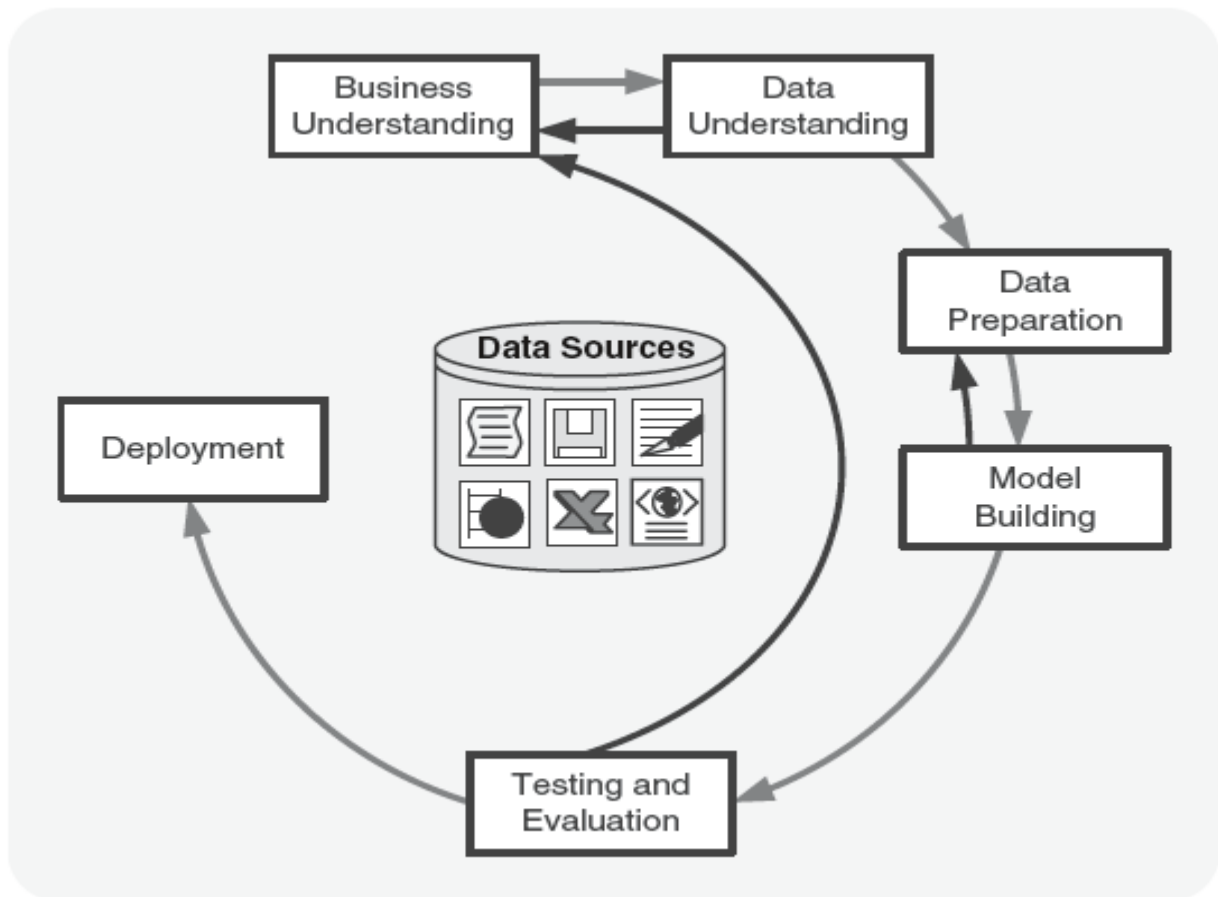


Figure 6. 1: CRISP-DM process flow (Source: (Olson & Delen, 2008))

Before predictive models could be applied, feature engineering was applied on the data, where features were extracted, and for unstructured data (corpora), the researcher had built rules in order to formulate an annotated corpus for the purpose of training. A gold standard was established for both structured and unstructured data, and this establishment was achieved through testing and evaluating the coverage of the gold standard on the training data that was identified. Then in Chapter Five, models were built and evaluated through the test data using a v-fold cross validation. The whole execution of this study was guided by the CRISP-DM framework, although different types of data, tools and different feature extraction and selection methods were also

used. This study has followed the process as shown in Figure 6.1 with an exception of the last process depicted, which is not covered in this study because there is no system that will be implemented. However, the output from the produced models could further be used as input to other systems such as a Clinical Decision Support System or Analytical Systems, and it could also be used for research purposes because it would have simplified making the data comparable. The CRISP-DM framework has proven to be a useful guideline for performing all the data mining processes for health-based data, whether it is structured or unstructured. In addition, this study only covered the variety property of big data, meaning it has partially captured the use of big data in health care through the CRISP-DM guideline. However, (Li, Thomas, & Osei-Bryson, 2016) have proposed a new framework called a snail shell process model. This framework is said to be suitable for the challenges that come with big data, also it was built to improve problem formulation, monitor and update models, and move between phases in the Knowledge Discovery and Data Mining (KDDM) process. Therefore, in future one would like to explore this framework further when addressing a problem that fully captures all the properties (volume, velocity, veracity, variety) of big data.

6.3 IMPLICATION OF THE FINDINGS

The researcher has set out a goal to use machine learning for addressing the problem of syntactic and semantic interoperability in health care. This study was focused on clinical observation data that could be mapped to a standard. It has been mentioned previously that health coding or terminology standards could be used to achieve data interoperability. The researcher therefore learnt how to apply a coding standard from data that had already been standardized. For structured data, the researcher applied a machine learning algorithm to learn from the patterns of the already standardized data. While for unstructured data, due to the lack of clinical data annotation knowledge, and the lack of previously and freely available smoking status annotations, the researcher thus opted to write rules for creating the annotations. The annotated documents (corpus) were used as input to the sequence-based machine learning algorithm (CRF)

and the corpora were used for supervising the algorithm. In the following subsections, the researcher talks about the implications of his findings.

6.3.1 INTEROPERABILITY FOR STRUCTURED DATA

In South Africa, (Adebesin, Kotzé, et al., 2013) reflected on the lack of well-skilled standards developers as one of the reasons it is difficult to implement health standards. Standards change over time, they are expensive to implement, and as mentioned above, there are many to choose from. Therefore, the researcher has identified common health standards that are prescribed by the Meaning Use program which is aimed at fixing the lack of interoperability in healthcare. The researcher has come up with an approach that uses machine learning in order to address the standards implementation problem. What the researcher proposed is an automated method for estimating the similarity between two potentially similar data objects. Data matching concepts were used as defined by (Bonifadi et al., 2011; Christen, 2008, 2012; Jahns & Veit, 2012) in order to identify similarities between related records. The objective was not to integrate one dataset to the next as it is done with record linkage and record matching but, it was about learning how one dataset (target) structures its data so that its patterns could be applied to one (source) whose data should be transformed. To the knowledge of the researcher, the approach that the researcher had used is unique, because it used record linkage and data matching concepts to compare data standardized data and unstandardized data so that the unstandardized data could be mapped to the one which is standardized. Also in this study, it is shown that standardized data implements the LOINC coding standard, while the unstandardized is the data to be transformed to LOINC. The researcher has experimented with the LOINC coding standard, because it is free and easy to use. Other researchers including (Abhyankar et al., 2012; Fidahusseini & Vreeman, 2014; Kim et al., 2012; Lee et al., 2013; Vreeman et al., 2015) have achieved a high accuracy while mapping to LOINC through the RELMA mapping tool. These researchers' method loaded the data to be mapped into the RELMA tool, then the RELMA tool predicts the potential matching observation to which the data should be mapped. The core difference between the RELMA tool and the current study (for structured data), is that this study although it

used laboratory observation names and LOINC to conduct experiments, the approach that the researcher proposes could be used to standardize data of any form. It should be noted that this study was not aimed at creating another clinical observation mapping system or tool, but it was testing whether a standard could be learned, irrespective of whether it is SNOMED-CT, LOINC, CPT, ICD-10, RxNorm or any other coding standard.

From mapping to LOINC, the researcher has learnt that other observations could not be mapped because the starting word of the observation name was completely different to the one it should be mapped to. For instance, “Blood Urea Nitrogen” from the MIMIC-III database could not be mapped to LOINC, because LOINC uses “Urea Nitrogen” instead. Therefore, n-gram could have been used to achieve such mapping. Mapping to LOINC also provides an educational platform that allows clinicians to learn new methods of referring to observation names, for instance there is no observation called “Lactic acid” in the LOINC database, however it is called as such in the MIMIC-III database, where LOINC has “lactate”. According to (Cormont et al., 2011), all acids should be written in the form of salts, and hence such information is vital when mapping to LOINC. It was also proven from this study that different databases use different naming to record the same information, the CareVue and the MetaVision HISs are an example of this.

6.3.2 INTEROPERABILITY FOR UNSTRUCTURED DATA

The goal for the classification of unstructured data was to address the standardization of behavioural or environmental data. This was required because patients are often affected by the environment. (Wild, 2012) mentions that chemical exposure such as arsenic or benzene result into epigenetic changes, and even the patient’s smoking status has a certain pattern in microRNA expression. Since environmental data is often recorded in an unstructured textual form, the aim was to extract meaningful concepts from this unstructured data, then standardize it so that the recording of it is not affected by location, time, institution or the person recording it. By standardizing this information,

it would then be structured in way that makes it easily comparable, searchable and exchangeable across disparate healthcare institutions. Firstly, information had to be extracted from unstructured text then mapped to a corresponding coding or terminology standard. However the process of extracting the data was different from previously related studies by (Liu et al., 2012; Sohn & Savova, 2009). These researchers attempted to address this problem through a customized cTakes program, which was applied at three different levels namely: sentence, document and patient. These researchers achieved a high performing model at both sentence and document-level. However, the annotations that were created in this study were also able to produce a high classification performance, especially for the classification of the “nonsmoker” class. This study had some similarities with the study by (Liu et al., 2012), since both have used a rule-based method and a machine learning method for training. However, (Liu et al., 2012; Sohn & Savova, 2009) had used SVM for learning how to annotate the given corpora, and in this study the researcher has used a CRF classifier. CRFs have previously outperformed non-structured SVM (Li, Kipper-Schuler, & Savova, 2008), and they mainly focused on predicting a large number of variables that depend on one another such as English phrases and the parts of speech tags. This study has gone beyond the classification of a smoking status and has used word shape, random indexing and word embedding features for understanding the meaning in text data. A highly cited paper by (Kenter & de Rijke, 2015) has established that word embedding features allow one to find semantic similarities between words, since words that are syntactically or semantically similar appear close to one another in a semantic space. All the features that the researcher used were for achieving high performance, even though the results of the smoking status prediction were lower than expected especially for the prediction of the “pastsmoker” class. However, it was observed that as more data was added, the performance of the classifier was constantly improving even for the “pastsmoker” class. Additionally, the rules would need to be updated so that they are able to cover complex conditions on a given text data.

Now a high performing classifier for all the classes would imply that the generated annotated corpora would be reusable for other research projects for detecting smoking status in clinical text. In addition, (Albright et al., 2013) have also seen the potential that distributable clinical annotated corpora have in the improvement of clinical decision support systems; clinical research combining phenotype and genotype data, quality control, comparative effectiveness and medication reconciliation, just to mention a few useful clinical applications. In this study, the researcher had to start from scratch building rules for annotating clinical documents for the purpose of identifying patient's smoking status, then mapping it to a suitable coding standard. However, had these annotations been freely available for research purposes, it would have catalysed the annotation of other documents and the mapping process. Ultimately, the researcher was able to use NLP and machine learning methods to get the patient's smoking information, such as the quantity and frequency of cigarettes smoked, and the dates associated with the usages. The extracted data was automatically mapped to UMLS CUI codes. Mapping to CUI codes helped to make the data interoperable because of common methods to identify and represent the data. The researcher also had suggestions on how to further map the predicted classes into a coding standard. However, despite what was achieved in this study, there were still limitations that were identified.

6.4 LIMITATIONS, FUTURE AND ADVICE

The selected databases were heterogeneous in structure because they were collected from two unrelated sources. MIMIC-III database contains both structured and unstructured data. Additionally, the data was not collected on a real-time basis, and it can be easily stored on a traditional database without needing a distributed processing framework such as Hadoop. This showed that the selected databases do not qualify to be labelled as big data. However, the data standardization technique that the researcher proposed can be applied on big data. This could be done in real-time where system A wants to exchange patient's data with system B. The data to be exchanged would be formatted and standardized so that the receiving end is able to interpret it.

Furthermore, applying what the researcher proposes in a real-time database is still to be explored. This study has only explored the *variety* characteristic of big data, *volume*, *velocity* and *veracity* is still to be explored.

The data to be mapped to the target dataset often comes from heterogeneous data sources, and dataset-based matching systems implement a wrapper (Fengguang, Xie, & Liquan, 2009). A wrapper helps compose the data so that it can be integrated to the target data source, where the data format could be “XML”, “CSV”, “HTML”, “RDF” and more. In this study, the researcher assumes that the structured data has already been composed in a readable format that can be queried through SQL. Therefore, the researcher has not used any wrappers for this data, even though it came from multiple sources. Further it was identified that LOINC has around 84868 observation names, and only a total of 1070 unique observation names were used in this study. The researcher has adapted the guidelines for mapping to LOINC from studies by (Abhyankar et al., 2012; Kim et al., 2012). These researchers have advised that mapping ought to be supervised by an expert. Since mapping to coding standards in South Africa is still a research task, therefore the researcher did not consult an expert for the LOINC mapping tasks. An annotation expert is also required for the annotation of clinical data, in the case of this study this was smoking status. However, in future, the researcher will compile a detailed guideline as per the advice of (Pustejovsky & Stubbs, 2013) for the purposes of annotation, and then consult an expert for the annotation task.

As for the unstructured data, the researcher used the predefined dictionaries that came with the CLAMP tool, this was done because of the lack of detailed documentation on how the dictionaries were created. It would have been more advantageous if the researcher had been able to use his own dictionaries and n-grams that were suitable for smoke status classification. The knowledge and the application of the UIMA RUTA rule language was an important component for the automatic annotations task. However, the researcher had spent a lot of time learning the scripting language which has a steep learning curve, and this was also the view of (Pablo, 2014). Furthermore, the storage of

the annotated data has not been thoroughly covered, but experts suggest the use of a NoSQL database such as CouchDB (Rea et al., 2012).

As for the mapping tasks, this study suggested the use of UMLS CUI for each of the extracted concepts. However, for the purpose of mapping the smoking status classes, the researcher suggested the use of FHIR profiles which emphasizes the use of coding standards for clinical data. However, other studies (Oniki et al., 2016; Pathak et al., 2013; Wu et al., 2013) have also used the same coding standard, but instead of implementing them through FHIR profiles, they have used Clinical Element Models (CEM). According to (Oniki et al., 2016), CEMs were developed by Intermountain for the SHARPn project. The SHARPn project is also called Strategic Health IT Advanced Research Project. Through this project open-source tools were developed for the purpose of standardizing EHR data for secondary use. During the initial stages of this study, the researcher had tried to use one of tools (cTakes) for the standardization of unstructured data. However, he could not install the tool and then he resorted to finding other tools such as CLAMP which was useful for the purpose of this study. However, CLAMP is not open-source, and when the researcher used it, it was still in its infancy Version 1.3. Therefore in future, the researcher would like to explore more of the SHARPn tools for the problem of standardizing the timeline for long-lived patient's data across multiple data sources, and scaling technologies such as Hadoop would come handy in addressing this problem. Key lessons were that a project of this nature needs a proper project plan, therefore project management skills are a necessity, hence the use of CRISP-DM provided a valuable guideline for conducting this study. The researcher has also discovered the importance of being agile and experimenting early in the project, while focusing on small data and less complex algorithms. Therefore, the use of RapidMiner was advantageous and beneficial for this study, because conducting experiments is quick and much clearer since it uses a visual representation of the classification processes, also it allowed one to extend the functionality of the algorithms by writing Python code. The first experiment of this study yielded results that show that the method that the researcher used could be used for finding errors in data.

Considering that one set of data is correct and standardized, comparing that set with another data set would show where these two sets match and where they do not match. The obvious application is the data mapping tool between disparate datasets.

6.5 CONCLUSION

This chapter presented a summarized version of the work that was done in this study. Firstly, on the introduction in section 6.1 the researcher identified what each section was meant to cover. In section 6.1 the researcher presented the implications that CRISP-DM framework had on this study, then in section 6.3 the researcher reflected on the findings of both experiments that were conducted, thereafter in section 6.4 the researcher mentioned limitations that were experienced while conducting experiments and he explores potential future studies and advises on lessons learned while carrying out the study.

REFERENCES

- Abdullah, M. F., & Ahmad, K. (2013). The mapping process of unstructured data to structured data. In *International Conference on Research and Innovation in Information Systems, ICRIIS* (Vol. 2013, pp. 151–155).
<http://doi.org/10.1109/ICRIIS.2013.6716700>
- Abhyankar, S., Demner-Fushman, D., & McDonald, C. J. (2012). Standardizing clinical laboratory data for secondary use. *Journal of Biomedical Informatics*, *45*(4), 642–650. <http://doi.org/10.1016/j.jbi.2012.04.012>
- Abiteboul, S. (1997). Querying semi-structured data. In *International Conference on Database Theory* (pp. 1–18). http://doi.org/10.1007/3-540-62222-5_33
- Adebesin, F., Foster, R., Kotzé, P., & Van Greunen, D. (2013). A Review of Interoperability Standards in E-health and Imperatives for their Adoption in Africa. *Applied Computing*, *50*(50), 55–72.
- Adebesin, F., Kotzé, P., Greunen, D. van, & Foster, R. (2013). Barriers & challenges to the adoption of E-Health standards in Africa. *Proceedings of Health Informatics South Africa 2013 (HISA 2013) Conference*. Retrieved from <http://researchspace.csir.co.za/dspace/handle/10204/6910>
- Adrián, G., Francisco, G. E., Marcela, M., Baum, A., Daniel, L., & Fernán, G. B. de Q. (2013). MongoDB: An open source alternative for HL7-CDA clinical documents management. *Open Source International Conference - CISL 2013*, 0–5.
<http://doi.org/10.13140/RG.2.1.3033.7128>
- Aggarwal, C. C., & Reddy, C. K. (2013). *Data Clustering: Algorithms and Applications* (1st ed.). Chapman & Hall/CRC.
- Albright, D., Lanfranchi, A., Fredriksen, A., Styler, W. F., Warner, C., Hwang, J. D., ... Savova, G. K. (2013). Towards comprehensive syntactic and semantic annotations of the clinical narrative. *Journal of the American Medical Informatics Association*, *20*(5), 922–930. <http://doi.org/10.1136/amiajnl-2012-001317>
- Aouicha, M. Ben, Ali, M., & Taieb, H. (2016). Computing semantic similarity between biomedical concepts using new information content approach. *JOURNAL OF BIOMEDICAL INFORMATICS*, *59*, 258–275.
<http://doi.org/10.1016/j.jbi.2015.12.007>

- Arlot, S., & Celisse, A. (2010). A survey of cross-validation procedures for model selection *. *Statistics Surveys*, 4, 40–79. <http://doi.org/10.1214/09-SS054>
- Asghari, E., & Keyvanpour, M. (2015). XML document clustering: techniques and challenges. *Artif Intell Rev*, 43, 417–436. <http://doi.org/10.1007/s10462-012-9379-2>
- Avati, A., Jung, K., Harman, S., Downing, L., Ng, A., & Shah, N. H. (2017). Improving Palliative Care with Deep Learning. Retrieved from <http://arxiv.org/abs/1711.06402>
- Banko, M., & Brill, E. (2001). Scaling to very very large corpora for natural language disambiguation. *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics - ACL '01*, 26–33. <http://doi.org/10.3115/1073012.1073017>
- Barbarito, F., Pincioli, F., Barone, A., Pizzo, F., Ranza, R., Mason, J., ... Marceglia, S. (2015). Implementing the lifelong personal health record in a regionalised health information system : The case of Lombardy , Italy. *Computers in Biology and Medicine*, 59, 164–174. <http://doi.org/10.1016/j.combiomed.2013.10.021>
- Barbulescu, M., Grigoriu, R., Halcu, I., Neculoiu, G., Sandulescu, V. C., Marinescu, M., & Marinescu, V. (2013). Integrating of structured, semi-structured and unstructured data in natural and build environmental engineering. In *2013 11th RoEduNet International Conference* (pp. 1–4). IEEE. <http://doi.org/10.1109/RoEduNet.2013.6511738>
- Barrett, G., Levell, P., & Milligan, K. (2013). A Comparison of Micro and Macro Expenditure Measures Across Countries Using Differing Survey Methods, 1–27. <http://doi.org/10.3386/w19544>
- Belle, A., Thiagarajan, R., Soroushmehr, S. M. R., Navidi, F., Beard, D. A., & Najarian, K. (2015). Big Data Analytics in Healthcare. *BioMed Research International*, 2015, 1–16. <http://doi.org/10.1155/2015/370194>
- Ben-Hur, A., & Weston, J. (2010). A user's guide to support vector machines. *Methods in Molecular Biology (Clifton, N.J.)*, 609, 223–239. http://doi.org/10.1007/978-1-60327-241-4_13
- Bennett, C. C. (2012). Utilizing RxNorm to support practical computing applications: Capturing medication history in live electronic health records. *Journal of Biomedical*

- Informatics*, 45(4), 634–641. <http://doi.org/10.1016/j.jbi.2012.02.011>
- Biba, M., & Xhafa, F. (2011). *Learning Structure and Schemas from Documents* (Vol. 375). <http://doi.org/10.1007/978-3-642-22913-8>
- Bilbao-Osorio, B., Dutta, S., & Lanvin, B. (2013). The global information technology report 2013. In *World Economic Forum* (pp. 1–383).
- Bilenko, M. (2006). Learnable Similarity Functions and Their Application to Record Linkage and Clustering. *Citeseer*, 2003(3), 449–467. <http://doi.org/10.14778/2733004.2733024>
- Bonifadi, A., Mecca, G., Papotti, P., & Velegarakis, Y. (2011). *Schema Matching and Mapping*. (Z. Bellahsene, A. Bonifati, & E. Rahm, Eds.). Berlin, Heidelberg: Springer Berlin Heidelberg. <http://doi.org/10.1007/978-3-642-16518-4>
- Bousquet, O. (2003). New approaches to statistical learning theory. *Annals of the Institute of Statistical Mathematics*, 55(2), 371–389. <http://doi.org/10.1023/A:1026303510251>
- Bousquet, O. (2004). Introduction to Statistical Learning Theory. *Biological Cybernetics*, 3176(1), 169–207. http://doi.org/10.1007/978-3-540-28650-9_8
- Brown, P. F., Della Pietra, V. J., deSouza, P. V, Lai, J. C., & Mercer, R. L. (1992). Class-based n-gram models of natural language. *Computational Linguistics*, 18(4), 467–479. Retrieved from <http://www.cs.cmu.edu/~roni/11761/PreviousYearsHandouts/classlm.pdf>
- Chang, C. C.-C. C. C.-C. C., & Lin, C. C. C.-J. (2011). A Library for Support Vector Machines. *ACM Transactions on Interlligent Systems and Technology (TIST)*, 2(3), 39. <http://doi.org/10.1145/1961189.1961199>
- Chawla, N. V., & Davis, D. A. (2013). Bringing Big Data to Personalized Healthcare: A Patient-Centered Framework. *Journal of General Internal Medicine*, 28(S3), 660–665. <http://doi.org/10.1007/s11606-013-2455-8>
- Chen, M., Mao, S., Zhang, Y., & Leung, V. C. M. (2014). *Big Data: Related Technologies, Challenges and Future Prospects*. Retrieved from <https://books.google.com.br/books?id=0wwqBAAAQBAJ>
- Cheng, Y., Zhang, K., Xie, Y., Agrawal, A., & Choudhary, A. (2012). On Active Learning in Hierarchical Classification, 2468–2471.

- Chih-Wei Hsu, Chih-Chung Chang, and C.-J. L. (2008). A Practical Guide to Support Vector Classification. *BJU International*, 101(1), 1396–400.
<http://doi.org/10.1177/02632760022050997>
- Christakis, N. A., & Fowler, J. H. (2009). *Connected: The surprising power of our social networks and how they shape our lives*. Little, Brown.
- Christen, P. (2008). Febrl – A Freely Available Record Linkage System with a Graphical User Interface. *Second Australasian Workshop on Health Data and Knowledge Management HDKM 2008*, 80, 17–25. Retrieved from
<http://datamining.anu.edu.au/linkage.html>
- Christen, P. (2012a). *Data Matching. Web Services: Concepts, Architectures and Applications*. Berlin, Heidelberg: Springer Berlin Heidelberg.
<http://doi.org/10.1007/978-3-642-31164-2>
- Christen, P. (2012b). *Data matching: Concepts and techniques for record linkage, entity resolution, and duplicate detection*. *Change*. <http://doi.org/10.1007/978-3-642-31164-2>
- Christopher D. Manning, Prabhakar Raghavan, & Hinrich Schütze. (2009). *An Introduction to Information Retrieval*. Retrieved from <https://nlp.stanford.edu/IR-book/pdf/irbookonlinereading.pdf>
- Clifton, C. (2004). Introduction to Data Mining What Is Data Mining ? What is Data Mining ? Real Example from the NBA Why Data Mining ?— Potential Applications, 1–26. Retrieved from
<https://pdfs.semanticscholar.org/c126/349ec999f6f0c93ea6d75c273c9e9abdc6ea.pdf>
- Coleman, A., Herselman, M. E., & Potass, D. (2012). E-health readiness assessment for e-health framework for Africa: A case study of hospitals in South Africa. In *Lecture Notes of the Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering* (Vol. 91 LNICST, pp. 162–169). Springer Berlin Heidelberg. http://doi.org/10.1007/978-3-642-29262-0_24
- Collins, M. (2011). Lecture 10 : Discriminative Training for MT / the Brown et al . Word Clustering Algorithm Discriminative Training for MT. *Word Journal Of The International Linguistic Association*. Retrieved from

- <http://www.cs.columbia.edu/~mcollins/courses/6998-2011/lectures/lec11.pdf>
- Cormont, S., Vandenbussche, P.-Y., Buemi, A., Delahousse, J., Lepage, E., & Charlet, J. (2011). Implementation of a platform dedicated to the biomedical analysis terminologies management. *AMIA ... Annual Symposium Proceedings. AMIA Symposium, 2011*, 1418–27. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/22195205>
- Creswell, J. W. (2014). *Research design: Qualitative, quantitative, and mixed methods approaches*. Sage publications.
- Cronje, J. C. (2014). What Is This Thing Called “ Design ” in Instructional Design Research ?— The ABC Instant Research Question Generator. In A. Moreira, O. Benavides, & A. J. Mendes (Eds.), *Media in Education: Results from the 2011 ICEM and SIIE joint Conference* (pp. 15–29). New York, NY: Springer New York. <http://doi.org/10.1007/978-1-4614-3175-6>
- CSIR, & NDoH. (2014). National Health Normative Standards Framework for Interoperability in eHealth in South Africa: Version 2.0, (March). Retrieved from <http://hufee.meraka.org.za/Hufeesite/staff/the-hufee-group/paula-kotze-1/hnsf-complete-version>
- Dai, C., Lin, D., Bertino, E., & Kantarcioglu, M. (2008). An Approach to Evaluate Data Trustworthiness Based on Data Provenance, 82–98.
- Demchenko, Y., De Laat, C., & Membrey, P. (2014). Defining architecture components of the Big Data Ecosystem. In *2014 International Conference on Collaboration Technologies and Systems, CTS 2014* (pp. 104–112). IEEE. <http://doi.org/10.1109/CTS.2014.6867550>
- Ding, Z., Yang, Q., & Wu, H. (2011). Massive Heterogeneous Sensor Data Management in the Internet of Things. *2011 International Conference on Internet of Things and 4th International Conference on Cyber, Physical and Social Computing*, (5), 100–108. <http://doi.org/10.1109/iThings/CPSCoM.2011.6>
- Do, H.-H. (2009). Data Conflicts. In L. LIU & M. T. ÖZSU (Eds.), *Encyclopedia of Database Systems* (pp. 565–569). Boston, MA: Springer US. http://doi.org/10.1007/978-0-387-39940-9_97
- Doan, A., Halevy, A., Ives, Z., Doan, A., Halevy, A., & Ives, Z. (2012a). 1 – Introduction.

- In *Principles of Data Integration* (pp. 1–18). <http://doi.org/10.1016/B978-0-12-416044-6.00001-6>
- Doan, A., Halevy, A., Ives, Z., Doan, A., Halevy, A., & Ives, Z. (2012b). 4 – String Matching. In *Principles of Data Integration* (pp. 95–119). <http://doi.org/10.1016/B978-0-12-416044-6.00004-1>
- Doan, A., Halevy, A., Ives, Z., Doan, A., Halevy, A., & Ives, Z. (2012c). Schema Matching and Mapping. In Z. Bellahsene, A. Bonifati, & E. Rahm (Eds.), *Principles of Data Integration* (pp. 121–160). Berlin, Heidelberg: Springer Berlin Heidelberg. <http://doi.org/10.1016/B978-0-12-416044-6.00005-3>
- Doan, A., Halevy, A., & Ives, Z. G. (2012). *Principles of data integration*. Morgan Kaufmann. Retrieved from <http://www.sciencedirect.com/science/book/9780124160446>
- Dong, X. L., & Srivastava, D. (2013). Big data integration. *2013 IEEE 29th International Conference on Data Engineering (ICDE)*, 1245–1248. <http://doi.org/10.1109/ICDE.2013.6544914>
- Ebadollahi, S., Ebadollahi, S., Coden, A. R., Coden, A. R., Tanenblatt, M. a., Tanenblatt, M. a., ... Amir, A. (2006). Concept-based electronic health records. *Proceedings of the 14th Annual ACM International Conference on Multimedia - MULTIMEDIA '06*, 997. <http://doi.org/10.1145/1180639.1180859>
- Eichelberg, M., Aden, T., & Riesmeier, O. (2005). A Survey and Analysis of Electronic Healthcare Record Standards. *ACM Computing Surveys*, 37(4), 277–315. <http://doi.org/10.1145/1118890.1118891>
- Fagin, R., Haas, L. M., Hernández, M., Miller, R. J., Popa, L., & Velegrakis, Y. (2009). Clio: Schema Mapping Creation and Data Exchange. Retrieved from http://0-download.springer.com.oasis.unisa.ac.za/static/pdf/753/chp%25253A10.1007%25252F978-3-642-02463-4_12.pdf?originUrl=http%253A%252F%252F0-link.springer.com.oasis.unisa.ac.za%252Fchapter%252F10.1007%252F978-3-642-02463-4_12&token2=exp=1493634895~acl=
- Feinleib, D. (2014). Big Data. In *Big Data Bootcamp: What Managers Need to Know to Profit from the Big Data Revolution* (pp. 1–14). Berkeley, CA: Apress. http://doi.org/10.1007/978-1-4842-0040-7_1

- Feldman, B., Martin, E. M., & Skotnes, T. (2012). *Big Data in Healthcare - Hype and Hope. Dr. Bonnie 360 degree (Business Development for Digital Health)* (Vol. 2013). Retrieved from <http://www.riss.kr/link?id=A99883549>
- Fengguang, X., Xie, H., & Liqun, K. (2009). Research and implementation of heterogeneous data integration based on XML. *2009 9th International Conference on Electronic Measurement & Instruments*, (Ameii), 4–715. <http://doi.org/10.1109/ICEMI.2009.5274686>
- FHIR. (2011). Observation-example-respiratory-rate.xml - FHIR v3.0.1. Retrieved November 10, 2017, from <https://www.hl7.org/fhir/observation-example-respiratory-rate.xml.html>
- Fidahusseini, M., & Vreeman, D. J. (2014). A corpus-based approach for automated LOINC mapping. *Journal of the American Medical Informatics Association*, 21(1), 64–72. <http://doi.org/10.1136/amiajnl-2012-001159>
- Friess, P., & Vermesan, O. (2011). *Internet of Things - Global Technological and Societal Trends From Smart Environments and Spaces to Green ICT*. River publishers. Retrieved from <http://books.google.com/books?hl=en&lr=&id=Eug-RvslW30C&pgis=1>
- Fu, Z., Christen, P., & Boot, M. (2010). A supervised learning and group linking method for historical census household linkage. *Conferences in Research and Practice in Information Technology Series*, 121, 153–162. Retrieved from http://0-delivery.acm.org.oasis.unisa.ac.za/10.1145/2490000/2483646/p153-fu.pdf?ip=163.200.81.46&id=2483646&acc=PUBLIC&key=646D7B17E601A2A5%252E24AFF711EFAADD7C%252E4D4702B0C3E38B35%252E4D4702B0C3E38B35&CFID=767729360&CFTOKEN=14363728&__acm__=1496604093_6
- Gao, J., & Koronios, A. (2015). Unlock the Value of Unstructured Data in EAM. In W. B. Lee, B. Choi, L. Ma, & J. Mathew (Eds.), *Proceedings of the 7th World Congress on Engineering Asset Management (WCEAM 2012)* (pp. 265–275). Cham: Springer International Publishing. http://doi.org/10.1007/978-3-319-06966-1_25
- Gao, J., Xie, C., & Chuanqi, T. (2016). Big Data Validation and Quality Assurance -- Issues, Challenges, and Needs. In *2016 IEEE Symposium on Service-Oriented System Engineering (SOSE)* (pp. 433–441). IEEE.

<http://doi.org/10.1109/SOSE.2016.63>

Garla, V., Re, V. Lo, Dorey-Stein, Z., Kidwai, F., Scotch, M., Womack, J., ... Brandt, C. (2011). The Yale cTAKES extensions for document classification: architecture and application. *Journal of the American Medical Informatics Association*, 18(5), 614–620. <http://doi.org/10.1136/amiajnl-2011-000093>

Gharehchopogh, F. S., & Khalifelu, Z. a. (2011). Analysis and evaluation of unstructured data: text mining versus natural language processing. In *2011 5th International Conference on Application of Information and Communication Technologies (AICT)* (pp. 1–4). IEEE. <http://doi.org/10.1109/ICAICT.2011.6111017>

Google. (2013). Google Ngram Viewer. Retrieved January 17, 2018, from https://books.google.com/ngrams/graph?content=does+not+smoke+*%2Cnever+smoked+*&year_start=1800&year_end=2008&corpus=15&smoothing=3&share=&direct_url=t2%3B%2Cdoes+not+smoke+%2A%3B%2Cc0%3B%2Cs0%3B%3Bdoes+not+smoke+or%3B%2Cc0%3B%3Bdoes+not+s

Gorunescu, F. (2011). Data Mining: Concepts, Models and Techniques. *Data Mining - Concepts, Models and Technique*, 1–357. <http://doi.org/10.1007/978-3-642-19721-5>

Groves, P., Kayyali, B., Knott, D., & Van Kuiken, S. (2013). The “big data” revolution in healthcare: accelerating value and innovation. *McKinsey Global Institute*, (January), 1–22. Retrieved from http://www.images-et-reseaux.com/sites/default/files/medias/blog/2013/12/mckinsey_131204_-_the_big_data_revolution_in_healthcare.pdf

Gruenheid, A., Dong, X. L., & Srivastava, D. (2014). Incremental record linkage. *Proceedings of the VLDB Endowment*, 7(9), 697–708. <http://doi.org/10.14778/2732939.2732943>

Hamel, L. (2009). Elements of Statistical Learning Theory. *Knowledge Discovery with Support Vector Machines*, 171–181.

Han, J., Kamber, M., & Pei, J. (Computer scientist). (2012). *Data mining : concepts and techniques*. Elsevier/Morgan Kaufmann. Retrieved from <http://0-www.sciencedirect.com.oasis.unisa.ac.za/science/book/9780123814791>

Harrington, P. (2012). *Machine learning in action* (Vol. 5). Manning Greenwich, CT.

- Hassanpour, S., & Langlotz, C. P. (2016). Information extraction from multi-institutional radiology reports HHS Public Access. *Artif Intell Med*, 66, 29–39.
<https://doi.org/10.1016/j.artmed.2015.09.007>
- Hassanzadeh, O., Pu, K. Q., Miller, R. J., Popa, L., Hernandez, M. A., & Ho, H. (2013). Discovering Linkage Points over Web Data. *VLDB*, 6(6), 445–456.
<http://doi.org/10.14778/2536336.2536345>
- Haux, R. (2006). Health information systems - past, present, future. *International Journal of Medical Informatics*, 75(3–4), 268–281.
<http://doi.org/10.1016/j.ijmedinf.2005.08.002>
- Higgins, D., & Burstein, J. (2007). Sentence similarity measures for essay coherence. *Proceedings of the 7th International Workshop on Computational Semantics IWCS*, (January), 1–12. Retrieved from
http://scholar.google.com/scholar?hl=en&q=Sentence+similarity+measure+for+essay+coherence&btnG=Search&as_sdt=2000&as_ylo=&as_vis=0#2
- Hill, E., Fry, Z. P., Boyd, H., Sridhara, G., Novikova, Y., Pollock, L., & Vijay-Shanker, K. (2008). AMAP: Automatically mining abbreviation expansions in programs to enhance software maintenance tools. *Proceedings - International Conference on Software Engineering*, 79–88.
<http://doi.org/http://doi.acm.org/10.1145/1370750.1370771>
- Hinssen, P. (2012). *The Age of Data-driven Medicine*. Retrieved from www.datascienceseries.com
- Hofmann, M. (2006). Support Vector Machines - Kernels and the Kernel Trick. Retrieved from http://www.cogsys.wiai.uni-bamberg.de/teaching/ss06/hs_svm/slides/SVM_Seminarbericht_Hofmann.pdf
- Holzinger, A. (2016). Machine Learning for Health Informatics: State-of-the-Art and Future Challenges, 211. <https://doi.org/10.1007/978-3-319-50478-0>
- Huang, T., Lan, L., Fang, X., An, P., Min, J., & Wang, F. (2015). Promises and Challenges of Big Data Computing in Health Sciences. *Big Data Research*, 2(1), 2–11. <http://doi.org/10.1016/j.bdr.2015.02.002>
- IBM. (2013). *Data-driven healthcare organizations use big data analytics for big gains*. Retrieved from <http://www03.ibm.com/industries/ca/en/healthcare/>

documents/Data_driven_healthcare_organizations_use_big_data_analytics_for_big_gains.pdf

- Islam, S. M. R., Kwak, D., Kabir, M. H., Hossain, M., & Kwak, K. S. (2015). The internet of things for health care: A comprehensive survey. *IEEE Access*, 3, 678–708. <http://doi.org/10.1109/ACCESS.2015.2437951>
- Jahns, V., & Veit. (2012). Principles of data integration by Anhai Doan, Alon Halevy, Zachary Ives. *ACM SIGSOFT Software Engineering Notes*, 37(5), 43. <http://doi.org/10.1145/2347696.2347721>
- Jian, W. S., Hsu, C. Y., Hao, T. H., Wen, H. C., Hsu, M. H., Lee, Y. L., ... Chang, P. (2007). Building a portable data and information interoperability infrastructure-framework for a standard Taiwan Electronic Medical Record Template. *Computer Methods and Programs in Biomedicine*, 88(2), 102–111. <http://doi.org/10.1016/j.cmpb.2007.07.014>
- Johnson, A. E. W., Pollard, T. J., Shen, L., Lehman, L. H., Feng, M., Ghassemi, M., ... Mark, R. G. (2016). MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3, 160035. <http://doi.org/10.1038/sdata.2016.35>
- Kaisler, S., Armour, F., Espinosa, J. A., & Money, W. (2013). Big Data: Issues and Challenges Moving Forward. In *2013 46th Hawaii International Conference on System Sciences* (pp. 995–1004). IEEE. <http://doi.org/10.1109/HICSS.2013.645>
- Kanerva, P. (2009). Hyperdimensional computing: An introduction to computing in distributed representation with high-dimensional random vectors. *Cognitive Computation*, 1(2), 139–159. <http://doi.org/10.1007/s12559-009-9009-8>
- Kenter, T., & de Rijke, M. (2015). Short Text Similarity with Word Embeddings. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management - CIKM '15* (pp. 1411–1420). <http://doi.org/10.1145/2806416.2806475>
- Kim, H., El-Kareh, R., Goel, A., Vineet, F. N. U., & Chapman, W. W. (2012). An approach to improve LOINC mapping through augmentation of local test names. *Journal of Biomedical Informatics*, 45(4), 651–657. <http://doi.org/10.1016/j.jbi.2011.12.004>
- Kim, M. H., XuYu, J., & Unland, R. (2011). *Database Systems for Advanced*

- Applications*. (J. X. Yu, M. H. Kim, & R. Unland, Eds.), *Lecture Notes in Computer Science* (Vol. 6588). Berlin, Heidelberg: Springer Berlin Heidelberg.
<http://doi.org/10.1007/978-3-642-20152-3>
- Kirkos, E., Spathis, C., & Manolopoulos, Y. (2008). Support vector machines, Decision Trees and Neural Networks for auditor selection. *Journal of Computational Methods in Sciences and Engineering*, 8, 213–224. Retrieved from
<http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.217.4046&rep=rep1&type=pdf>
- Kleynhans, A. (2011). Is South Africa ready for a national Electronic Health Record (EHR)?, (70992657), 1–100. Retrieved from
http://uir.unisa.ac.za/bitstream/handle/10500/6128/2011_MBL3_Research_Report_A-M_Kleynhans.pdf?sequence=1
- Kluegl, P., Toepfer, M., Beck, P. D., Fette, G., & Puppe, F. (2016). UIMA Ruta: Rapid development of rule-based information extraction applications. *Natural Language Engineering*, 22(1), 1–40. <http://doi.org/10.1017/S1351324914000114>
- Koller, D., & Friedman, N. (2009). *Probabilistic Graphical Models- Principles and Techniques*. *Journal of Chemical Information and Modeling* (Vol. 53).
<http://doi.org/10.1017/CBO9781107415324.004>
- Krzysztof Cios, Witold Pedrycz, Roman Swiniarski, & Lukasz Kurgan. (2007). *Data Mining A Knowledge Discovery Approach*. <http://doi.org/13:978-0-387-33333-5>
- Kuhn, M., & Johnson, K. (2013). *Applied Predictive Modeling*. New York, NY: Springer New York. <http://doi.org/10.1007/978-1-4614-6849-3>
- Kum, H.-C., Krishnamurthy, A., Machanavajjhala, A., Reiter, M. K., & Ahalt, S. (2014). Privacy preserving interactive record linkage (PPIRL). *Journal of the American Medical Informatics Association*, 21(2), 212–220. <http://doi.org/10.1136/amiajnl-2013-002165>
- Kumar, A. (2015). Machine Learning – When to Use Logistic Regression vs. SVM. Retrieved April 26, 2017, from <http://vitalflux.com/machine-learning-use-logistic-regression-vs-svm/>
- Kurgan, L. A., & Musilek, P. (2006). A survey of Knowledge Discovery and Data Mining process models. *Knowledge Engineering Review*, 21(1), 1–24.

<http://doi.org/10.1017/S0269888906000737>

- Lan, M., Tan, C. L., Su, J., & Lu, Y. (2009). Supervised and Traditional Term Weighting Methods for Automatic Text Categorization. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 31(4), 721–735.
<http://doi.org/10.1109/TPAMI.2008.110>
- Lapão, L. V., da Silva, M. M., & Gregório, J. (2017). Implementing an online pharmaceutical service using design science research. *BMC Medical Informatics and Decision Making*, 17(1), 31. <http://doi.org/10.1186/s12911-017-0428-2>
- Leavitt, N. (2010). Will NoSQL Databases Live Up to Their Promise? *Computer*, 43(2), 12–14. <http://doi.org/10.1109/MC.2010.58>
- Lee, L.-H., Groß, A., Hartung, M., Liou, D.-M., & Rahm, E. (2013). A multi-part matching strategy for mapping LOINC with laboratory terminologies. *Journal of the American Medical Informatics Association*, 21(5), 1–9. <http://doi.org/10.1136/amiajnl-2013-002139>
- Li, D., Kipper-Schuler, K., & Savova, G. (2008). Conditional Random Fields and Support Vector Machines for Disorder Named Entity Recognition in Clinical Texts. *BioNLP*, 94–95. Retrieved from <http://www.aclweb.org/anthology/W08-0615>
- Li, Y., Thomas, M. A., & Osei-Bryson, K. M. (2016). A snail shell process model for knowledge discovery via data analytics. *Decision Support Systems*, 91, 1–12.
<https://doi.org/10.1016/j.dss.2016.07.003>
- Lin, M.-C., Vreeman, D. J., & Huff, S. M. (2011). Investigating the semantic interoperability of laboratory data exchanged using LOINC codes in three large institutions. *AMIA ... Annual Symposium Proceedings / AMIA Symposium. AMIA Symposium, 2011*, 805–14. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3243154/pdf/0805_amia_2011_proc.pdf
- Ling, C. X., Huang, J., & Zhang, H. (2003). AUC: A Better Measure than Accuracy in Comparing Learning Algorithms (pp. 329–341). Springer, Berlin, Heidelberg.
http://doi.org/10.1007/3-540-44886-1_25
- Liu, H., Kumar, T. K. A., & Thomas, J. P. (2015). Cleaning Framework for Big Data - Object Identification and Linkage. In *2015 IEEE International Congress on Big Data*

- (pp. 215–221). IEEE. <http://doi.org/10.1109/BigDataCongress.2015.38>
- Liu, M., Shah, A., Jiang, M., Peterson, N. B., Dai, Q., Aldrich, M. C., ... Xu, H. (2012). A study of transportability of an existing smoking status detection module across institutions. *AMIA ... Annual Symposium Proceedings. AMIA Symposium, 2012*, 577–86. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/23304330>
- Liu, X., Lang, B., Yu, W., Luo, J., & Huang, L. (2011). AUDR: An Advanced Unstructured Data Repository. In *Proceedings - 2011 6th International Conference on Pervasive Computing and Applications, ICPCA 2011* (pp. 462–469). <http://doi.org/10.1109/ICPCA.2011.6106548>
- Longadge, R., & Dongre, S. (2013). Class Imbalance Problem in Data Mining Review. Retrieved from <http://arxiv.org/abs/1305.1707>
- Lopez, D., & Blobel, B. (2009). A development framework for semantically interoperable health information systems. *International Journal of Medical Informatics*, 78(2), 83–103. <http://doi.org/10.1016/j.ijmedinf.2008.05.009>
- Luo, G., Tang, C., & Thomas, S. B. (2012). Intelligent personal health record: experience and open issues. *Journal of Medical Systems*, 36(4), 2111–2128. <http://doi.org/10.1007/s10916-011-9674-5>
- Malley, B., Ramazzotti, D., & Wu, J. T. (2016). Data Pre-processing. In *Secondary Analysis of Electronic Health Records* (pp. 115–141). Cham: Springer International Publishing. http://doi.org/10.1007/978-3-319-43742-2_12
- Malvey, D., & Slovensky, D. J. (2014). Overview. In *mHealth* (pp. 1–17). Boston, MA: Springer US. http://doi.org/10.1007/978-1-4899-7457-0_1
- Mandel, J. C., Kreda, D. A., Mandl, K. D., Kohane, I. S., & Ramoni, R. B. (2016). SMART on FHIR: A standards-based, interoperable apps platform for electronic health records. *Journal of the American Medical Informatics Association*, 23(5), 899–908. <http://doi.org/10.1093/jamia/ocv189>
- Mansingh, G., Osei-Bryson, K.-M., & Asnani, M. (2016). Exploring the antecedents of the quality of life of patients with sickle cell disease: using a knowledge discovery and data mining process model-based framework. *Health Systems*, 5(1), 52–65. <http://doi.org/10.1057/hs.2015.3>
- Marszalek, M. (2009). A Tutorial on Hidden Markov Models. Retrieved from

- <http://www.cogsci.ucsd.edu/~ajyu/Teaching/Tutorials/hmm.pdf>
- Masilela, T. C., Foster, R., & Chetty, M. (2013). *The eHealth Strategy for South Africa 2012-2016: how far are we? South African Health Review Review*.
- Matshidze, P., & Hanmer, L. (2007). Health Information Systems in the Private Health Sector. *South African Health Review*, 89–102. Retrieved from citeulike-article-id:13234454
- Mayer-Schönberger, V., & Cukier, K. (2013). *Big Data: A Revolution that Will Transform how We Live, Work, and Think*. Houghton Mifflin Harcourt.
- Mayo Clinic. (2015). Mayo Medical Laboratories. Retrieved October 24, 2017, from <https://www.mayomedicallaboratories.com/>
- Mayosi, B. M., Lawn, J. E., Niekerk, A. Van, Bradshaw, D., Karim, S. S. A., Coovadia, H. M., & South, L. (2012). Review Health in South Africa : changes and challenges since 2009. *Www.TheLancet.Com*, 380(9858), 5–19. [http://doi.org/doi.org/10.1016/S0140-6736\(12\)61814-5](http://doi.org/doi.org/10.1016/S0140-6736(12)61814-5)
- Mcdonald, C., Huff, S., Deckard, J., Armson, S., Abhyankar, S., & Vreeman, D. J. (2017). Logical Observation Identifiers Names and Codes (LOINC ®) Users' Guide. Retrieved from <https://loinc.org/download/loinc>
- Mehrabi, S., Krishnan, A., Sohn, S., Roch, A. M., Schmidt, H., Kesterson, J., ... Palakal, M. (2015). DEEPEN: A negation detection system for clinical text incorporating dependency relation into NegEx. *JOURNAL OF BIOMEDICAL INFORMATICS*, 54, 213–219. <https://doi.org/10.1016/j.jbi.2015.02.010>
- Melton, G. B., Parsons, S., Morrison, F. P., Rothschild, A. S., Markatou, M., & Hripcsak, G. (2006). Inter-patient distance metrics using SNOMED CT defining relationships. *Journal of Biomedical Informatics*, 39(6), 697–705. <http://doi.org/10.1016/j.jbi.2006.01.004>
- Messina, A., Storniolo, P., & Urso, A. (2016). Keep It Simple, Fast and Scalable: A Multi-model NoSQL DBMS as an (eb) XML-over-SOAP Service. In *2016 30th International Conference on Advanced Information Networking and Applications Workshops (WAINA)* (pp. 220–225). IEEE. <http://doi.org/10.1109/WAINA.2016.71>
- Meyer, D. (2009). Support vector machines. *Engineering*, 1(December), 1–8. <http://doi.org/10.1002/wics.049>

- Michel, J.-B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., Google, T., ...
Lieberman, E. (2011). Quantitative analysis of culture using millions of digitized books. *Science January*, 14(3316014), 176–182.
<http://doi.org/10.1126/science.1199644>
- Mitchell, T. M. (1997). Machine learning. *Burr Ridge, IL: McGraw Hill*, 45(37), 870–877.
- Moffat, A., Zobel, J., Skeppstedt, M., Daudaravičius, V., Duneld, M., Browne, A., ...
Zeng-Treitler, Q. (2008). Synonym extraction and abbreviation expansion with ensembles of semantic spaces. *ACM Transactions on Information Systems*, 27(1), 1–27. <http://doi.org/10.1186/2041-1480-5-6>
- Moniruzzaman, A. B. M., & Hossain, S. A. (2013). Nosql database: New era of databases for big data analytics-classification, characteristics and comparison. *Nosql Database: New Era of Databases for Big Data Analytics-Classification, Characteristics and Comparison*, 6(4), 1–14. Retrieved from [http://scholar.google.com/scholar?q=Nosql database: New era of databases for big data analytics-classification, characteristics and comparison&btnG=&hl=en&num=20&as_sdt=0%2C22 VN - readcube.com](http://scholar.google.com/scholar?q=Nosql+database:+New+era+of+databases+for+big+data+analytics-classification,+characteristics+and+comparison&btnG=&hl=en&num=20&as_sdt=0%2C22+VN+-+readcube.com)
- Moraes, R., Valiati, J. F., Gavião Neto, W. P., & Neto, W. P. G. (2013). Document-level sentiment classification: An empirical comparison between SVM and ANN. *Expert Systems with Applications*, 40(2), 621–633.
<http://doi.org/http://dx.doi.org/10.1016/j.eswa.2012.07.059>
- Mxoli, A., Mostert-Phipps, N., & Gerber, M. (2014). Personal Health Records: Design considerations for the South African context. In *Design, Development and Research* (Vol. 16, pp. 124–245).
<http://doi.org/http://researchspace.csir.co.za/dspace/handle/10204/7712>
- Mxoli, A., Mostert-Phipps, N., & Gerber, M. (2015). Personal Health Records in the South African Healthcare Landscape: A SWOT Analysis. In *ICTs for Inclusive Communities in Developing Societies* (pp. 345–357).
- Nagy, M., Preckova, P., Seidl, L., & Zvarova, J. (2010). Challenges of interoperability using HL7 v3 in Czech healthcare. *Studies in Health Technology and Informatics*, 155, 122–128. <http://doi.org/10.3233/978-1-60750-563-1-122>
- Nasien, D., Yuhaniz, S. S., & Haron, H. (2010). Statistical Learning Theory and Support

- Vector Machines. In *2010 Second International Conference on Computer Research and Development* (pp. 760–764). IEEE. <http://doi.org/10.1109/ICCRD.2010.183>
- Natarajan, K., Li, J., & Koronios, A. (2009). Data mining techniques for data cleaning. *Engineering Asset Lifecycle Management*, 796–804. http://doi.org/10.1007/978-0-85729-320-6_91
- National Department of Health (South Africa). (2015). *National Health Insurance for South Africa: Towards Universal Health Coverage. White Paper*. Retrieved from <http://www.doh.gov.za/list.php?type=National Health Insurance>
- Ng, A. (2000). CS229 Lecture notes. Retrieved from <http://cs229.stanford.edu/notes/cs229-notes1.pdf>
- Ng, A. (2011). *Advice for applying Machine Learning*. Stanford University.
- Ng, A. (2016). Machine Learning Yearning, 1–23.
- Ng, A., & Jordan, M. I. (2002). On generative vs. discriminative classifiers: A comparison of logistic regression and naive bayes. *Proceedings of Advances in Neural Information Processing*, 28(3), 169–187. <http://doi.org/10.1007/s11063-008-9088-7>
- Nickerson, D., Atalag, K., de Bono, B., Geiger, J., Goble, C., Hollmann, S., ... Hunter, P. (2016). The Human Physiome: how standards, software and innovative service infrastructures are providing the building blocks to make it achievable. *Interface Focus*, 6(2), 20150103. <http://doi.org/10.1098/rsfs.2015.0103>
- Noumeir, R. (2011). Sharing medical records: The XDS architecture and communication infrastructure. *IT Professional*, 13(4), 46–51. <http://doi.org/10.1109/MITP.2010.123>
- Okazaki, N. (2007). CRFsuite: a fast implementation of Conditional Random Fields (CRFs). Retrieved from <http://www.chokkan.org/software/crfsuite/>
- Olson, D., & Delen, D. (2008). *Advanced Data Mining Techniques*. <http://doi.org/10.1007/978-3-540-76917-0>
- Omholt, S. W., & Hunter, P. J. (2016). The Human Physiome: a necessary key for the creative destruction of medicine. *Interface Focus*, 6(2), 20160003. <http://doi.org/10.1098/rsfs.2016.0003>
- Oniki, T. A., Zhuo, N., Beebe, C. E., Liu, H., Coyle, J. F., Parker, C. G., ... Huff, S. M. (2016). Clinical element models in the SHARPN consortium. *Journal of the*

American Medical Informatics Association, 23(2), 248–256.

<http://doi.org/10.1093/jamia/ocv134>

- Orgun, B., & Vu, J. (2006). HL7 ontology and mobile agents for interoperability in heterogeneous medical information systems. *Computers in Biology and Medicine*, 36(7–8), 817–836. <http://doi.org/10.1016/j.compbimed.2005.04.010>
- Pablo, A. D. (2014). Apache UIMA and the Watson Jeopardy! TM System Big Data Montreal Meetup. Retrieved from <http://duboue.net/papers/20140204uima-bigdata.pdf>
- Paschou, M., Sakkopoulos, E., Sourla, E., & Tsakalidis, A. (2012). Health Internet of Things: Metrics and methods for efficient data transfer. *Simulation Modelling Practice and Theory*, pp. 186–199. <http://doi.org/10.1016/j.simpat.2012.08.002>
- Patel, C. J., Pho, N., McDuffie, M., Easton-Marks, J., Kothari, C., Kohane, I. S., & Avillach, P. (2016). A database of human exposomes and phenomes from the US National Health and Nutrition Examination Survey. *Scientific Data*, 3, 160096. <https://doi.org/10.1038/sdata.2016.96>
- Pathak, J., Bailey, K. R., Beebe, C. E., Bethard, S., Carrell, D. S., Chen, P. J., ... Chute, C. G. (2013). Normalization and standardization of electronic health records for high-throughput phenotyping: the SHARPn consortium. *Journal of the American Medical Informatics Association*, 20(e2), e341–e348. <http://doi.org/10.1136/amiajnl-2013-001939>
- Perkins, L. S., Andrews, P., Panda, D., Morton, D., Bonica, R., Werstiuk, N., & Kreiser, R. (2011). A Survey of Load Balancing Techniques for Data Intensive Computing. *The 2009 International Symposium on Collaborative Technologies and Systems CTS 2009*, 41(4), c1–c1. <http://doi.org/10.1007/978-1-4614-1415-5>
- Philip Chen, C. L., & Zhang, C. Y. (2014). Data-intensive applications, challenges, techniques and technologies: A survey on Big Data. *Information Sciences*, 275, 314–347. <http://doi.org/10.1016/j.ins.2014.01.015>
- Ponomareva, N., Rosso, P., Pla, F., & Molina, A. (2007). Conditional Random Fields vs. Hidden Markov Models in a biomedical Named Entity Recognition task. *Proc. of Int. Conf. Recent Advances in Natural Language Processing, RANLP*, 479–483. Retrieved from

<https://pdfs.semanticscholar.org/1bbe/6b9e2310fbae3560dcb5ef2961272684d5aa.pdf>

Poon, C. C. Y., Zhang, Y.-T., & Bao, S.-D. (2006). A novel biometrics method to secure wireless body area sensor networks for telemedicine and m-health. *IEEE Communications Magazine*, 44(4), 73–81.

Porter, M. E., & Lee, T. H. (2013). *The Strategy That Will Fix Health Care*. Boston, MA: Harvard Business Review Webinar.

Prestasi kecekapan pengurusan kewangan dan agihan zakat: perbandingan antara majlis agama islam negeri di Malaysia. (n.d.).

<http://doi.org/10.1017/CBO9781107415324.004>

Pustejovsky, J., & Stubbs, A. (2013). *Natural language annotation for machine learning: A guide to corpus-building for applications*. O'Reilly Media. Retrieved from <http://www.amazon.com/Natural-Language-Annotation-Machine-Learning/dp/1449306667>

Pustejovsky, J., & Stubbs, a. (2013). *Natural language annotation for machine learning*. Vasa. <http://doi.org/1332788036>

Raghupathi, W., & Raghupathi, V. (2014). Big data analytics in healthcare: promise and potential. *Health Information Science and Systems*, 2(1), 3.

<http://doi.org/10.1186/2047-2501-2-3>

Rea, S., Pathak, J., Savova, G., Oniki, T. A., Westberg, L., Beebe, C. E., ... Chute, C. G. (2012). Building a robust, scalable and standards-driven infrastructure for secondary use of EHR data: The SHARPN project. *Journal of Biomedical Informatics*, 45(4), 763–771. <http://doi.org/10.1016/j.jbi.2012.01.009>

Regenstrief Institute. (2016). LOINC (Logical Observation Identifiers Names and Codes). Retrieved October 24, 2017, from <https://loinc.org/>

Rivo, E., De La Fuente, J., Rivo, Á., García-Fontán, E., Cañizares, M. Á., & Gil, P. (2012). Cross-Industry Standard Process for data mining is applicable to the lung cancer surgery domain, improving decision making as well as knowledge and quality management. *Clinical and Translational Oncology*, 14(1), 73–79.

<http://doi.org/10.1007/s12094-012-0764-8>

Saliccioli, J. D., Crutain, Y., Komorowski, M., & Marshall, D. C. (2016). Sensitivity

- Analysis and Model Validation. In *Secondary Analysis of Electronic Health Records* (pp. 263–271). Cham: Springer International Publishing. http://doi.org/10.1007/978-3-319-43742-2_17
- Sarawagi, S. (2007). Information Extraction, 1(3), 261–377. <http://doi.org/10.1561/15000000003>
- Savova, G. K., Masanz, J. J., Ogren, P. V, Zheng, J., Sohn, S., Kipper-Schuler, K. C., & Chute, C. G. (2010). Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. *Journal of the American Medical Informatics Association*, 17(5), 507–513. <http://doi.org/10.1136/jamia.2009.001560>
- Seebregts, C., Barron, P., Tanna, G., Benjamin, P., & Fogwill, T. (2016). MomConnect: an exemplar implementation of the Health Normative Standards Framework in South Africa. *South African Health Review*, Jan 2016(1), 125–135. Retrieved from http://journals.co.za/docserver/fulltext/healthr/2016/1/healthr_2016_a13.pdf?expires=1515022420&id=id&accname=guest&checksum=551FDB2B9055D1E1A5CB3DCDCCA22A4D
- Setiono, R., Liu, H., Motoda, H., Setiono, R., & Zhao, Z. (2010). Feature Selection : An Ever Evolving Frontier in Data Mining. *Journal of Machine Learning Research: Workshop and Conference Proceedings 10: The Fourth Workshop on Feature Selection in Data Mining*, 4–13.
- Shah, N. H., & Tenenbaum, J. D. (2012). The coming age of data-driven medicine: translational bioinformatics' next frontier. *Journal of the American Medical Informatics Association*, 19(e1), e2–e4. <http://doi.org/10.1136/amiajnl-2012-000969>
- Sharma, S., & Osei-Bryson, K.-M. (2010). Toward an integrated knowledge discovery and data mining process model. *The Knowledge Engineering Review*, 25(1), 49. <http://doi.org/10.1017/S0269888909990361>
- Siddiqi, M. H., Alam, M. G. R., Hong, C. S., Khan, A. M., & Choo, H. (2016). A novel maximum entropy markov model for human facial expression recognition. *PLoS ONE*, 11(9). <http://doi.org/10.1371/journal.pone.0162702>
- Singh, A. (2010). Support Vector Machines. Retrieved from <http://www.cs.cmu.edu/~aarti/Class/10701/slides/Lecture12.pdf>

- Smith, J., Fridsma, D., & Johns, M. (2014). Igniting an Interoperable Healthcare System.
- Smits, M., & Cornet, R. (2014). A comparison of two Detailed Clinical Model representations: FHIR and CDA. Retrieved from <http://dare.uva.nl/cgi/arno/show.cgi?fid=573070>
- Sohn, S., & Savova, G. K. (2009). Mayo clinic smoking status classification system: extensions and improvements. *AMIA ... Annual Symposium Proceedings. AMIA Symposium, 2009*, 619–23. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/20351929><http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2815365>
- Soysal, E., Wang, J., Jiang, M., Wu, Y., Pakhomov, S., Liu, H., & Xu, H. (2017). CLAMP – a toolkit for efficiently building customized clinical natural language processing pipelines. *Journal of the American Medical Informatics Association*. <http://doi.org/10.1093/jamia/ocx132>
- Sutton, C., & Mccallum, A. (2011). An Introduction to Conditional Random Fields. *Machine Learning*, 4(4), 267–373. <http://doi.org/10.1561/22000000013>
- Swan, M. (2009). Emerging patient-driven health care models: An examination of health social networks, consumer personalized medicine and quantified self-tracking. *International Journal of Environmental Research and Public Health*, 492–525. <http://doi.org/10.3390/ijerph6020492>
- Swan, M. (2012a). Health 2050: The Realization of Personalized Medicine through Crowdsourcing, the Quantified Self, and the Participatory Biocitizen. *Journal of Personalized Medicine*, 2(3), 93–118. <http://doi.org/10.3390/jpm2030093>
- Swan, M. (2012b). Sensor Mania! The Internet of Things, Wearable Computing, Objective Metrics, and the Quantified Self 2.0. *Journal of Sensor and Actuator Networks*, 1(3), 217–253. <http://doi.org/10.3390/jsan1030217>
- Talbur, J. R. (2011). *Entity Resolution and Information Quality*. *Entity Resolution and Information Quality*. Elsevier. <http://doi.org/10.1016/C2009-0-63396-1>
- Taleb, I., Dssouli, R., & Serhani, M. A. (2015). Big Data Pre-processing: A Quality Framework. In *2015 IEEE International Congress on Big Data* (pp. 191–198). <http://doi.org/10.1109/BigDataCongress.2015.35>
- Tang, N. (2014). Big Data Cleaning. In *Web Technologies and Applications* (Vol. 8709,

- pp. 13–24). http://doi.org/10.1007/978-3-319-11116-2_2
- Tene, O., & Polonetsky, J. (2012). Privacy in the age of big data: a time for big decisions. *Stanford Law Review Online*, 64, 63.
- The Competition Commission South Africa. (2016). Private Healthcare Inquiry Public Hearings Day 3, 18 February Live at CSIR #HMI. Retrieved April 17, 2018, from https://youtu.be/-cn_eIW48X0
- Tian, Y., Shi, Y., & Liu, X. (2012). Recent advances on support vector machines research. *Technological and Economic Development of Economy*, 18(1), 5–33. <http://doi.org/10.3846/20294913.2012.661205>
- Topol, E. (2015). *The Patient Will See You Now: The Future of Medicine is in Your Hands*. New York, NY: Basic Books. <http://doi.org/10.4258/hir.2015.21.4.321>
- Turvey, C., Klein, D., Fix, G., Hogan, T. P., Woods, S., Simon, S. R., ... Nazi, K. (2014). Blue Button use by patients to access and share health record information using the Department of Veterans Affairs' online patient portal. *Journal of the American Medical Informatics Association*, 21(4), 657–663. <http://doi.org/10.1136/amiajnl-2014-002723>
- Uzuner, O., Goldstein, I., Luo, Y., & Kohane, I. (2008). Identifying patient smoking status from medical discharge records. *Journal of the American Medical Informatics Association : JAMIA*, 15(1), 14–24. <http://doi.org/10.1197/jamia.M2408>
- Vapnik, V. (2013). *The nature of statistical learning theory*. Springer science & business media.
- Viana-Ferreira, C., Ribeiro, L. S., & Costa, C. (2014). A Framework for Integration of Heterogeneous Medical Imaging Networks. *The Open Medical Informatics Journal*, 8(1), 20–32. <http://doi.org/10.2174/1874431101408010020>
- Viangteeravat, T., Anyanwu, M. N., Nagisetty, V., Kuscu, E., Sakauye, M., & Wu, D. (2011). Clinical data integration of distributed data sources using Health Level Seven (HL7) v3-RIM mapping. *Journal of Clinical Bioinformatics*, 1(1), 32. <http://doi.org/10.1186/2043-9113-1-32>
- Villars, R. L., & Olofson, C. W. (2011). *Big Data : What It Is and Why You Should Care*. Retrieved from www.idc.com
- Vreeland, A., Persons, K. R., Primo, H., Bishop, M., Garriott, K. M., Doyle, M. K., ...

- Bashall, C. (2016). Considerations for Exchanging and Sharing Medical Images for Improved Collaboration and Patient Care: HIMSS-SIIM Collaborative White Paper. *Journal of Digital Imaging*. <http://doi.org/10.1007/s10278-016-9885-x>
- Vreeman, D. J., Hook, J., & Dixon, B. E. (2015). Learning from the crowd while mapping to LOINC. *Journal of the American Medical Informatics Association*, 22(6), 1205–1211. <http://doi.org/10.1093/jamia/ocv098>
- Vreeman, D. J., & McDonald, C. J. (2005). Automated Mapping of Local Radiology Terms to LOINC. *American Medical Informatics Association Annual Symposium Proceedings*, 769–773. <http://doi.org/57553> [pii]
- Wang, P., & Domeniconi, C. (2008). Building semantic kernels for text classification using wikipedia. In *Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD 08* (p. 713). New York, New York, USA: ACM Press. <http://doi.org/10.1145/1401890.1401976>
- Weber, S. (2010). Design Science Research : Paradigm or Approach? *Proceedings of the 16th Americas Conference on Information Systems*, Paper 214. Retrieved from <http://aisel.aisnet.org/amcis2010>
- Wenfei Fan, Jianzhong Li, Nan Tang, & Wenyan Yu. (2014). Incremental Detection of Inconsistencies in Distributed Data. *IEEE Transactions on Knowledge and Data Engineering*, 26(6), 1367–1383. <http://doi.org/10.1109/TKDE.2012.138>
- Wild, C. P. (2012). The exposome: from concept to utility. *International Journal of Epidemiology*, 41(1), 24–32. <http://doi.org/10.1093/ije/dyr236>
- Winter, A., Haux, R., Ammenwerth, E., Brigl, B., Hellrung, N., & Jahn, F. (2011). *Health Information Systems*. London: Springer London. <http://doi.org/10.1007/978-1-84996-441-8>
- Witten, I. H., & Frank, E. (2011). *Data mining: practical machine learning tools and techniques*. *Complementary literature None*. <http://doi.org/0120884070>, 9780120884070
- World Health Organization. (2008). Toolkit on Monitoring Health Systems Strengthening: Health Information Systems, (June).
- Wu, S. T., Kaggal, V. C., Dligach, D., Masanz, J. J., Chen, P., Becker, L., ... Chute, C. G. (2013). A common type system for clinical natural language processing. *Journal*

- of Biomedical Semantics*, 4(1), 1. <http://doi.org/10.1186/2041-1480-4-1>
- Xing, E. (2007). Hidden Markov Model and Conditional Random Fields. *Graphical Models*, 1–22. Retrieved from <http://www.cs.cmu.edu/~epxing/Class/10708-07/Slides/lecture12-CRF-HMM-annotation.pdf>
- Xu, H., Caramanis, C., & Mannor, S. (2008). Robustness and Regularization of Support Vector Machines. *Journal of Machine Learning Research*, 10. Retrieved from <http://arxiv.org/abs/0803.3490>
- Xu, Y., Zhen, L., Yang, L., & Wang, L. (2009). Classification Algorithm Based on Feature Selection and Samples Selection, 631–638.
- Zamani, H., & Croft, W. B. (2016). Embedding-based Query Language Models. *Proceedings of the 2016 ACM International Conference on the Theory of Information Retrieval - ICTIR '16*, 147–156. <http://doi.org/10.1145/2970398.2970405>
- Zhao, X., Wang, G., Bi, X., Gong, P., & Zhao, Y. (2011). XML document classification based on ELM. *Neurocomputing*, 74(16), 2444–2451. <http://doi.org/10.1016/j.neucom.2010.12.038>

APPENDIX A-1: ETHICAL CONSENT LETTER FROM UNISA



**UNISA COLLEGE OF SCIENCE, ENGINEERING AND TECHNOLOGY'S
(CSET) RESEARCH AND ETHICS COMMITTEE**

9 June 2017

Ref #: 040/MN/2017/CSET_SOC
Name: Mandlenkosi Absalom Walter
Student #: 43615554

Dear Mandlenkosi Absalom Walter

Decision: Ethics Approval for three years (No humans involved)

Researcher: Mandlenkosi Absalom Walter
43615554@mylife.unisa.ac.za, +27 84 493 4536

Supervisor (s): Prof F. Bankole
bankofo@unisa.ac.za, +27 11 670 9476

Proposal: Health Systems Data Management

Qualification: MsC in Computing

Thank you for the application for research ethics clearance by the Unisa College of Science, Engineering and Technology's (CSET) Research and Ethics Committee for the above mentioned research. Ethics approval is granted for a period of three years from 9 June 2017 to 9 June 2020.

1. The researcher will ensure that the research project adheres to the values and principles expressed in the UNISA Policy on Research Ethics.
2. Any adverse circumstance arising in the undertaking of the research project that is relevant to the ethicality of the study, as well as changes in the methodology, should be communicated in writing to the Unisa College of Science, Engineering and Technology's (CSET) Research and Ethics Committee. An amended application could



RECEIVED
2017-06-12
Office of the Deputy Executive Dean
College of Science, Engineering & Technology

University of South Africa
Preller Street, Muckleneuk Ridge, City of Tshwane
PO Box 392 UNISA 0003 South Africa
Telephone: +27 12 429 3111 Facsimile: +27 12 429 4150
www.unisa.ac.za

be requested if there are substantial changes from the existing proposal, especially if those changes affect any of the study-related risks for the research participants.

3. The researcher will ensure that the research project adheres to any applicable national legislation, professional codes of conduct, institutional guidelines and scientific standards relevant to the specific field of study.
4. Only de-identified research data may be used for secondary research purposes in future on condition that the research objectives are similar to those of the original research. Secondary use of identifiable human research data require additional ethics clearance.

Note:

The reference number 040/MN/2017/CSET_SOC should be clearly indicated on all forms of communication with the intended research participants, as well as with the Unisa College of Science, Engineering and Technology's (CSET) Research and Ethics Committee

Yours sincerely

Ade de Veiga

Dr. A Da Veiga

Chair: Ethics Sub-Committee School of Computing, CSET

I. Osunmakinde

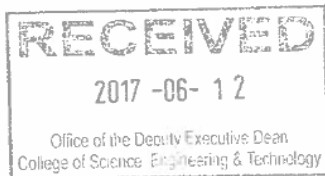
Prof I. Osunmakinde

Director: School of Computing, CSET

B. Mamba

Prof B. Mamba

Executive Dean: College of Science, Engineering and Technology (CSET)



Approved - decision template – updated Aug 2016

University of South Africa
Pretter Street, Muckleneuk Ridge, City of Tshwane
PO Box 392 UNISA 0003 South Africa
Telephone: +27 12 429 3111 Facsimile: +27 12 429 4150
www.unisa.ac.za

APPENDIX A-2: REPORT TO AUTHORIZE THE USE OF MIMIC-III DATABASE FOR RESEARCH

COLLABORATIVE INSTITUTIONAL TRAINING INITIATIVE (CITI PROGRAM) COMPLETION REPORT - PART 1 OF 2 COURSEWORK REQUIREMENTS*

* NOTE: Scores on this [Requirements Report](#) reflect quiz completions at the time all requirements for the course were met. See list below for details. See separate Transcript Report for more recent quiz scores, including those on optional (supplemental) course elements.

- **Name:** Mandlenkosi Ngwenya (ID: 6308971)
- **Institution Affiliation:** Massachusetts Institute of Technology Affiliates (ID: 1912)
- **Institution Email:** 43615554@mylife.unisa.ac.za
- **Institution Unit:** University Of South Africa Computing

- **Curriculum Group:** Human Research
- **Course Learner Group:** Data or Specimens Only Research
- **Stage:** Stage 2 - Refresher Course

- **Record ID:** 23026778
- **Completion Date:** 01-May-2017
- **Expiration Date:** 30-Apr-2020
- **Minimum Passing:** 90
- **Reported Score*:** 97

REQUIRED AND ELECTIVE MODULES ONLY	DATE COMPLETED	SCORE
SBE Refresher 1 – Defining Research with Human Subjects (ID: 15029)	01-May-2017	2/2 (100%)
SBE Refresher 1 – Privacy and Confidentiality (ID: 15035)	01-May-2017	2/2 (100%)
SBE Refresher 1 – Assessing Risk (ID: 15034)	01-May-2017	2/2 (100%)
SBE Refresher 1 – Research with Children (ID: 15036)	01-May-2017	2/2 (100%)
SBE Refresher 1 – International Research (ID: 15028)	01-May-2017	2/2 (100%)
Biomed Refresher 2 - Instructions (ID: 764)	01-May-2017	No Quiz
Biomed Refresher 2 – History and Ethical Principles (ID: 511)	01-May-2017	3/3 (100%)
Biomed Refresher 2 – Regulations and Process (ID: 512)	01-May-2017	2/2 (100%)
Biomed Refresher 2 – SBR Methodologies in Biomedical Research (ID: 515)	01-May-2017	3/4 (75%)
Biomed Refresher 2 – Genetics Research (ID: 518)	01-May-2017	2/2 (100%)
Biomed Refresher 2 – Records-Based Research (ID: 516)	01-May-2017	3/3 (100%)
Biomed Refresher 2 - Populations in Research Requiring Additional Considerations and/or Protections (ID: 519)	01-May-2017	1/1 (100%)
Biomed Refresher 2 – HIPAA and Human Subjects Research (ID: 526)	01-May-2017	5/5 (100%)
Biomed Refresher 2 – Conflicts of Interest in Research Involving Human Subjects (ID: 681)	01-May-2017	3/3 (100%)
How to Complete the CITI Refresher Course and Receive a Completion Report (ID: 922)	01-May-2017	No Quiz

For this Report to be valid, the learner identified above must have had a valid affiliation with the CITI Program subscribing institution identified above or have been a paid Independent Learner.

Verify at: www.citiprogram.org/verify/2kfccd9d56-7664-48eb-b6df-ad95c754ea8b-23026778

Collaborative Institutional Training Initiative (CITI Program)
Email: support@citiprogram.org
Phone: 888-529-5929
Web: <https://www.citiprogram.org>

COLLABORATIVE INSTITUTIONAL TRAINING INITIATIVE (CITI PROGRAM)

COMPLETION REPORT - PART 2 OF 2 COURSEWORK TRANSCRIPT**

** NOTE: Scores on this Transcript Report reflect the most current quiz completions, including quizzes on optional (supplemental) elements of the course. See list below for details. See separate Requirements Report for the reported scores at the time all requirements for the course were met.

- **Name:** Mandlenkosi Ngwenya (ID: 6308971)
- **Institution Affiliation:** Massachusetts Institute of Technology Affiliates (ID: 1912)
- **Institution Email:** 43615554@mylife.unisa.ac.za
- **Institution Unit:** University Of South Africa Computing

- **Curriculum Group:** Human Research
- **Course Learner Group:** Data or Specimens Only Research
- **Stage:** Stage 2 - Refresher Course

- **Record ID:** 23026778
- **Report Date:** 04-Feb-2018
- **Current Score**:** 97

REQUIRED, ELECTIVE, AND SUPPLEMENTAL MODULES	MOST RECENT	SCORE
Biomed Refresher 2 - Instructions (ID: 764)	01-May-2017	No Quiz
Biomed Refresher 2 – History and Ethical Principles (ID: 511)	01-May-2017	3/3 (100%)
Biomed Refresher 2 – Regulations and Process (ID: 512)	01-May-2017	2/2 (100%)
Biomed Refresher 2 – SBR Methodologies in Biomedical Research (ID: 515)	01-May-2017	3/4 (75%)
Biomed Refresher 2 – Records-Based Research (ID: 516)	01-May-2017	3/3 (100%)
Biomed Refresher 2 – Genetics Research (ID: 518)	01-May-2017	2/2 (100%)
SBE Refresher 1 – International Research (ID: 15028)	01-May-2017	2/2 (100%)
SBE Refresher 1 – Defining Research with Human Subjects (ID: 15029)	01-May-2017	2/2 (100%)
Biomed Refresher 2 - Populations in Research Requiring Additional Considerations and/or Protections (ID: 519)	01-May-2017	1/1 (100%)
SBE Refresher 1 – Assessing Risk (ID: 15034)	01-May-2017	2/2 (100%)
SBE Refresher 1 – Privacy and Confidentiality (ID: 15035)	01-May-2017	2/2 (100%)
SBE Refresher 1 – Research with Children (ID: 15036)	01-May-2017	2/2 (100%)
Biomed Refresher 2 – HIPAA and Human Subjects Research (ID: 526)	01-May-2017	5/5 (100%)
Biomed Refresher 2 – Conflicts of Interest in Research Involving Human Subjects (ID: 681)	01-May-2017	3/3 (100%)
How to Complete the CITI Refresher Course and Receive a Completion Report (ID: 922)	01-May-2017	No Quiz

For this Report to be valid, the learner identified above must have had a valid affiliation with the CITI Program subscribing institution identified above or have been a paid Independent Learner.

Verify at: www.citiprogram.org/verify/2kfcdd9d56-7664-48eb-b6df-ad95c754ea8b-23026778

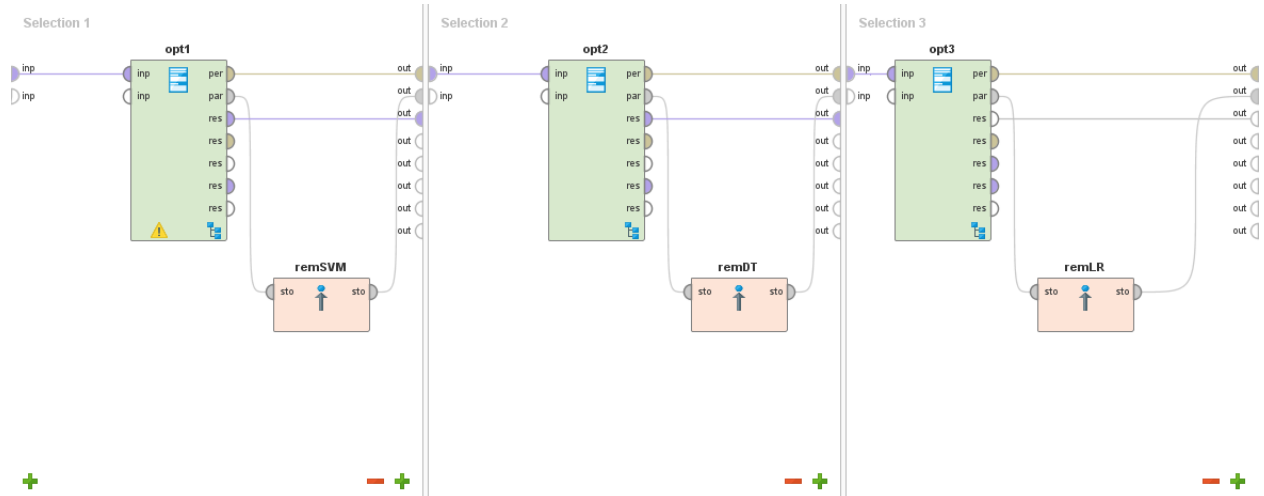
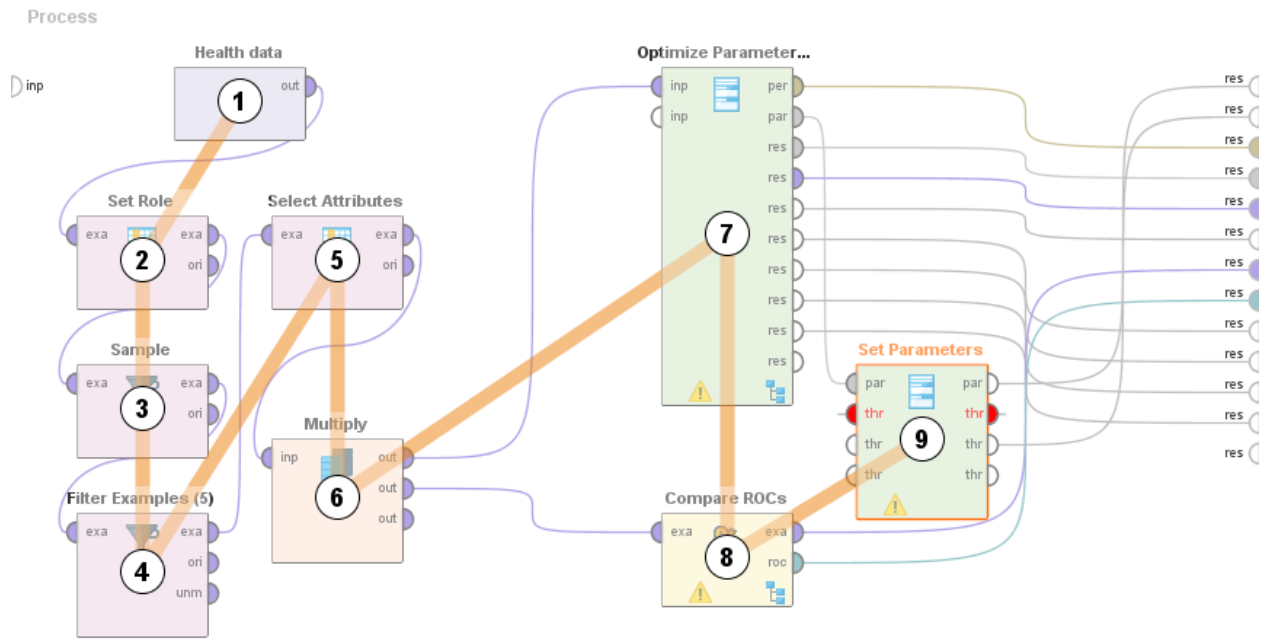
Collaborative Institutional Training Initiative (CITI Program)

Email: support@citiprogram.org

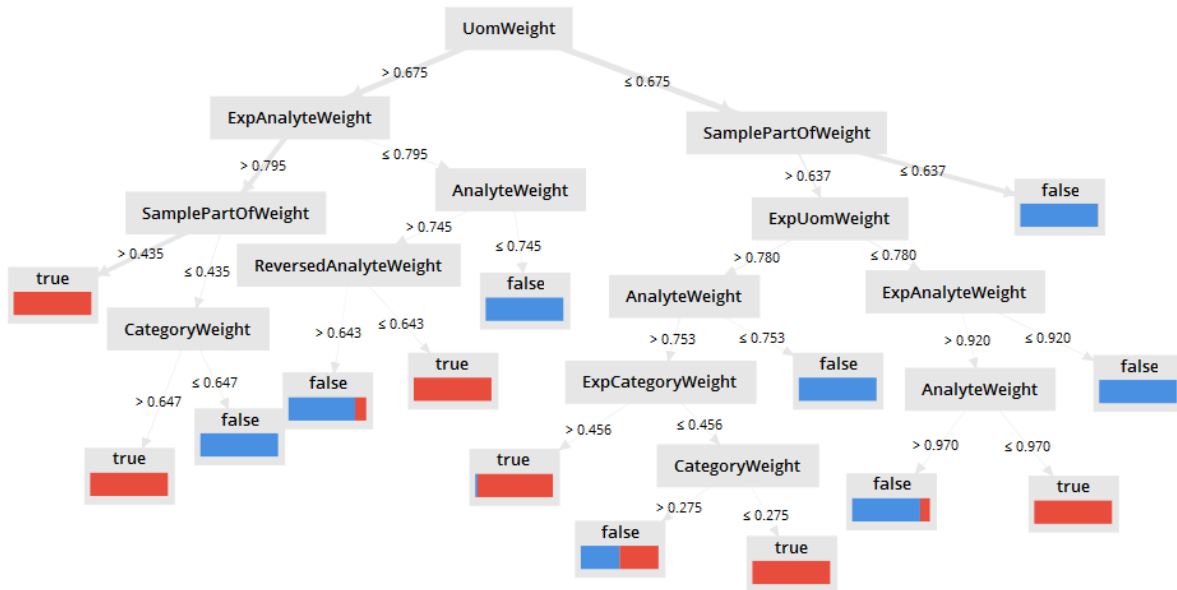
Phone: 888-529-5929

Web: <https://www.citiprogram.org>

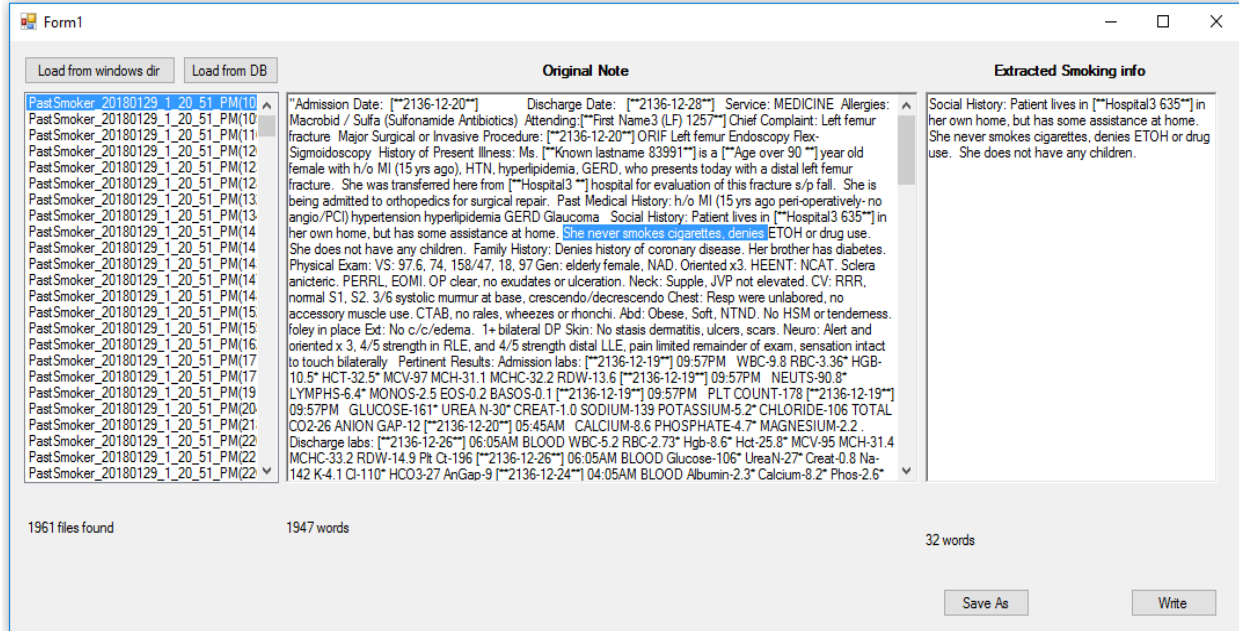
APPENDIX B: PROCESS FLOW FOR ROC RESULTS COMPARISON BETWEEN SVM, DECISION TREES AND LOGISTIC REGRESSION



APPENDIX C: DECISION TREE, SPLITTING CRITERION EVALUATION



APPENDIX D: SCREENSHOT OF THE PROGRAM THE RESEARCHER WROTE FOR THE PURPOSE OF EXTRACTING SMOKING INFORMATION FROM A LARGE TEXT FILE



Code for this program is accessible as shown in Appendix E.

APPENDIX E: SETUP FILES AND RESULTS FROM EXPERIMENT 1 AND EXPERIMENT 2

Accessible via google drive: <https://drive.google.com/open?id=1iSXX-CAJaSXbFbhmdpmmvFzaAEG8Nupg>

- Experiment 1
 - o Setup files and executable files include, results are in a .txt file
 - Code for simialrity weight calculation (requires Octave or Matlab)
 - SVM files (requires RapidMiner)
 - Decision tree files (requires RapidMiner)
 - Logistic Regression (requires RapidMiner)

- Experiment 2:
 - o Setup files include (requires CLAMP software):
 - NegationDictionary
 - NamedEntityRecognizerLooku
 - PartOfSpeechTagger
 - UIMA Ruta rule scripts
 - Section Identifier
 - SentenceDetector
 - TemporalRecognizer
 - TemporalRelation
 - Tokenizer
 - UserDefinedRelations
 - All Word representation features
 - o Program for extracting relevant content from text big files
 - o Results and annotations (contains .txt and .xmi files)