



A Novel Data-mining Model for Automated Prediction of Low Birth Weight.

By

UZAPI HANGE

Reg. No: 14101345

BSc (Applied Business Computing) (University of Sunderland)

Department of Computer Science and Information Systems

College of Science

Botswana International University of Science and Technology

uzapi.hange@studentmail.biust.ac.bw, (+267) 765 715 21

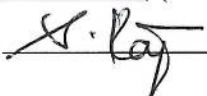
A Dissertation Submitted to the College of Science in Partial Fulfilment of the Requirements for the Award of the Degree of Master of Science in Information Systems of BIUST

Supervisor: Dr. Rajalakshmi Selvaraj

Department of Computer Science and Information Systems

College of Science, BIUST

selvaraj@biust.ac.bw, (+267) 724 454 26

Signature  Date: 3/4/2017

Co-Supervisor: Dr. Malatsi Galani

Department of Computer Science and Information Systems

College of Science, BIUST

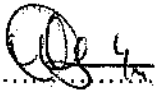
galanijm@biust.ac.bw, (+267) 726 828 55

Signature  Date: 3/4/2017

January, 2017

Declaration and copyright

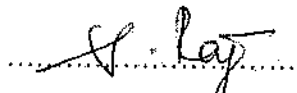
I, **Uzapi Hange**, declare that this dissertation is my own original work and that it has not been presented and will not be presented to any other university for a similar or any other degree award.

Signature 

This dissertation/thesis is copyright material protected under the Berne Convention, the Copyright Act of 1999 and other international and national enactments, in that behalf, on intellectual property. It must not be reproduced by any means, in full or in part, except for short extracts in fair dealing; for researcher private study, critical scholarly review or discourse with an acknowledgement, without the written permission of the office of the Provost, on behalf of both the author and the BIUST.

Certification

The undersigned certifies that he/she has read and hereby recommends for acceptance by the College of Science a dissertation/thesis titled: *A Novel Data-mining Model for Automated Prediction of Low Birth Weight*, in fulfilment of the requirements for the degree of Master of Science/Doctor of Philosophy in (*Information Systems*) of the Botswana International University of Science and Technology.



Dr. Rajalakshmi Selvaraj
(Supervisor)

Date: 3/4/2017



Dr. Malatsi Galani
(Co-Supervisor)

Date: 3/4/2017

Acknowledgements

First and foremost I would like to direct my sincere gratitude to the God Almighty for carrying me throughout this journey and for placing people in my life to help me achieve my goal. I would especially like to thank Dr. Rajalakshmi Selvaraj and Dr. Malatsi Galani for their guidance and support throughout the process of pursuing my MSc that led to the completion of this thesis.

I am indebted to the staff at the Botswana International University of Science and Technology, specifically the computer science and information systems department for their generous support. My research wouldn't have taken its form if it wasn't for Dr. Dimane Mpoeleng, Dr. Hlomani Hlomani, Dr. Keletso Letsholo, Professor Heiko Schroöder, and Professor Irina Zlotnikova to mention a few.

The financial support and overall facilitation of my study by the Botswana International University of Science and Technology cannot go unnoticed. Finally, my strong appreciation goes to my classmates for their comments and constructive feedback. To my best friends Rachel Kapeko, Kushata Ntwaetsile, Sasha Meroro, Thotobolo Morapedi, and Umpenyu Bernice Dema, thank you for moral support during the time I devoted to my study.

Dedication

I would like to dedicate this dissertation to my family; Agnes Hange, Petrus Hange, Utjiua Timon, Vaavi Hange, Tumbee Hange, Cisca Kefas, Joseph Kefas, Uerihorora Hange Kefas, Tapita Nauta, Shadreck Kefas, Patricia Kemumuine, Kaumbu Katjipaha, Uaendondonja Daniel, and Uamunikua Kangunde. I would not have made it this far if it was not for their full support.

Table of Contents

Declaration and copyright.....	i
Certification.....	ii
Acknowledgements.....	iii
Dedication.....	iv
List of tables.....	viii
List of figures.....	ix
Abstract.....	x
1. Introduction.....	1
1.1 Introduction to the Problem.....	1
1.2 Background.....	2
1.3 Problem Statement.....	4
1.4 General Objective.....	4
1.5 Specific Objectives.....	4
1.6 Research Questions.....	5
1.7 Significance of the Study.....	5
1.8 Dissertation Organization.....	5
2 Literature Review.....	7
2.1 Data-mining Process.....	7
2.2 Data-mining Tasks.....	8
2.2.1 Predictive data-mining (classification) methods.....	8
2.3 Evaluating Data-mining Models Developed for Healthcare Settings.....	12
2.3.1 Confusion Matrix.....	12
2.4 Prior Application of Decision Trees in Healthcare.....	15
2.5 Challenges of Mining Healthcare Data.....	17
2.5.1 Imbalanced Data.....	17
2.5.2 High Dimensionality.....	18
2.6 Low Birth Weight and Its Determinants.....	19
2.7 Low Birth Weight Prediction (Related Works).....	20
2.7.1 Low Birth Weight Prediction Using Statistical Models.....	20
2.7.2 Low Birth Weight Prediction Using Data-mining Models.....	23
2.8 Literature Review Summary.....	23
Chapter 3.....	25
3 Methodology.....	25
3.1 An Architecture of the Research Approach.....	25
3.2 Dataset.....	26
3.3 Data Selection.....	26

3.4	Data Pre-processing	26
3.5	Model Building.....	32
3.5.1	Model Building Using J48 Algorithm	32
3.5.2	Model Building Using REPTree Algorithm	33
3.5.3	Model Building Using Random Tree Algorithm	33
3.6	Model Evaluation.....	33
3.7	Predicting Birth Weight Based On an Independent Synthetic Dataset.....	34
3.8	Methodology Summary	34
Chapter 4		35
4	Presentation of Results and Discussion	35
4.1	Experimental Design.....	35
4.1.1	Experimentation One (Model Building with Imbalanced Data).....	36
4.1.2	Experimentation Two (Model Building with SMOTE).....	37
4.1.3	Experimentation Three (Model Building with the Best Selected Attributes (CFS).....	38
4.1.4	Selection of the best model.....	41
4.1.5	Predicting Birth Weight Based On an Independent Synthetic Dataset.....	41
4.2	Results and Discussion Summary	52
Chapter 5		54
5	Conclusions and Recommendations	54
5.1	Study Overview	54
5.2	Conclusions.....	54
5.3	Recommendations.....	55
6	References.....	56
7	Appendices	62
7.1	Appendix A: North Carolina State Centre for Health Statistics-2006 NC State Birth File Description.....	62
7.2	Appendix B: The J48 result of run information of experiment one using the actual data.....	71
7.3	Appendix C: The REPTree result of run information of experiment one using the actual data	71
7.4	Appendix D: The Random tree result of run information of experiment one using the actual data	72
7.5	Appendix E: The J48 result of run information of experiment two using SMOTE.....	72
7.6	Appendix F: The REPTree result of run information of experiment two using SMOTE	73
7.7	Appendix G: The Random tree result of run information of experiment two using SMOTE	73
7.8	Appendix H: The J48 result of run information of experiment three using feature selection	74
7.9	Appendix I: The REPTree result of run information of experiment three using feature selection	74

7.10	Appendix J: The Random tree result of run information of experiment three using feature selection	75
7.11	Appendix K: The J48 result of run information based on an independent dataset	75

List of tables

Table 1 Missing values together with their respective handling mechanisms 27

Table 2 Number of synthetic case generate by SMOTE.....30

Table 3 default parameter settings for building a J48 model 32

Table 4 default parameter settings for building a REPTree model 33

Table 5 default parameter settings for building a Random Tree model 33

Table 6 Performance output of models when applied to an imbalanced dataset 36

Table 7 performance output of models built with SMOTE 37

Table 8 performance output of models when built with the best selected features 40

Table 9 A comparison of experimental results based on the AUC of models 41

Table 10 Summary output of the J48 model when applied to new unseen data 42

List of figures

Figure 1 An overview of the steps in the KDD process [35]	7
Figure 2 Data-mining tasks	8
Figure 3 Confusion matrix for a binary classification problem [24]	12
Figure 4 A depiction of an AUC for a particular classifier [47]	15
Figure 5 Research approach architecture	25
Figure 6 Birth weight frequencies per class	29
Figure 7 Implementation of SMOTE in WEKA	30
Figure 8 Final distribution per class, after SMOTE 600%	31
Figure 9 Rankings by CFS algorithm and best search strategy	39
Figure 10 The independent dataset distribution per class	42
Figure 11 Confusion matrix for J48 model when applied to new unseen data	43
Figure 12 ROC Curve for class 0	44
Figure 13 ROC curve for class 1	45
Figure 14 ROC curve for class 2	46
Figure 15 ROC curve for class 3	47
Figure 16 ROC curve for class 4	48
Figure 17 ROC curve for class 5	48
Figure 18 ROC curve for class 6	49
Figure 19 ROC curve for class 7	50
Figure 20 ROC curve for class 8	51
Figure 21 ROC curve for class 9	52

Abstract

Birth weight is one of the major factors that determine the overall future health outcomes. Predicting birth weight can enable medical practitioners to make early obstetric interventions, thus minimising complications associated with low birth weight. Data-mining models are receiving a great deal of attention for making predictions based on a vast amount of low birth data. Low birth weight research has especially focused on identifying the risk factors of low birth weight. However, the prediction of actual birth weight values based on the identified low birth weight risk factors, which can play a significant role in the identification of mothers at the risk of delivering low birth weight infants, remains unsolved. Since the performance of data-mining techniques is dependent on the underlying problem, it is vital to analyse their relative performances for any given task. Therefore the goal of this thesis was to develop a data-mining model that predicts the actual birth weight with a relatively higher Area Under the receiver operating Characteristic (AUC).

The prediction was based on low birth weight risk factors and birth data from the North Carolina State Centre for Health Statistics of 2006. In order to extract interesting patterns from data the knowledge discovery in databases process model was utilized. The steps followed were data selection, data pre-processing, model building, and model evaluation/interpretation. Decision trees were used for classifying birth weight and tested on the actual imbalanced datasets, the balanced dataset using Synthetic Minority Oversampling Technique (SMOTE), as well as with all the features and reduced feature using correlation based feature selection algorithm.

The results highlighted that models built with balanced datasets using the SMOTE algorithm produce a relatively higher AUC comparative to models built with imbalanced datasets. It was also discovered that building models with reduced features through correlation based feature selection algorithm, give a comparatively higher AUC as opposed to models built with all features. The J48 decision tree built with reduced features outperformed REPTree and Random tree with an AUC of 90.3%, and thus it was selected as the best model. When applying the selected J48 model to new unseen data, and comparing it to the testing set, we reached a conclusion that the feasibility of using J48 in birth weight prediction would offer the possibility to reduce obstetric related complications and thus improving the overall obstetric healthcare.

Chapter 1

1. Introduction

This chapter sets the scene and outlines why there is a need to conduct this research. The chapter is divided into several sub-sections, the introduction to the problem, background, problem statement, the research objectives and questions that the study seeks to answer, the significance of the study and finally the organisation of the dissertation.

1.1 Introduction to the Problem

Continuous attempts to improve healthcare delivery has led to rapid adoption of Electronic Health Records (EHR) [1], [2], [3], [4]. This adoption has resulted in the dramatic increase of data commonly referred to as “big data”, [5], [6], [7] hence substantially widening the gap between unprocessed and useful information available for decision making. For this reason, the healthcare industry is viewed as ‘data-rich’ but ‘knowledge-poor’ [8]. While EHR have been effective in enabling medical data acquisition, they have been less fruitful in helping medical practitioners transform this data into useful knowledge for important decision-making processes, such as estimating clinical outcomes.

Obstetrics is a field of study focused on pregnancy, childbirth, and the post-delivery period. Obstetric cases have been reported to constitute a considerable fraction of medical data [9], since childbirth is the most common reason for hospital admissions. Information collected during childbirth relates to the infant, mother and the father. This data is accumulated in medical databases and when analysed, trends and patterns are unveiled. Thus medical practitioners can advance their knowledge and hence making insightful diagnosis and treatments [5], [7].

Birth weight is a significant determinant of a new-born’s probability of survival, long term growth and psychological development [10], [11]. Thus, in medical research, birth weight has been under intense scrutiny. Particularly, the casual factors of low birth weight (LBW), have been the major focus of previous investigations. Evidence shows that low birth weight new-borns are at a higher risk of perinatal and infant mortality and other health complications such as hypertensive adults, mental retardation, and respiratory distress syndrome (RDS) [12], [13], [14], [15], [16], [17], [18], [19], [20]. RDS can result in repeated respiratory infections, delay in growth, learning disabilities and asthma.

Low birth weight is often classified as extremely low (less than 1000 grams), very low (less than 1 500 grams), moderately low (less than 2,500 grams), or normal (at least 2,500 grams) weight at birth [13], regardless of the gestational age [11]. According to the World Health

Organization (WHO), about 20 million or approximately 15.5% of births worldwide are LBW and 95.6% of these are from developing countries [12]. Reducing cases of LBW by at least one third is amongst the seven aims for the current decade of the “A world fit for Children” programme of the United Nations Children’s fund [12]. As a result, there is an intensifying demand for robust and non-invasive methods of estimating birth weight and predicting new-borns at a higher risk of being born with LBW [18]. The development of simple and inexpensive tools are critical, especially in developing countries, where high-cost equipment and trained specialists are scarce.[21] In addition, accurate assessment and prediction of LBW permits early obstetric interventions and overall improvement of maternal outcomes [18], [20], [21], [22], [23].

1.2 Background

Low birth weight prediction dates back to many years ago. Analysts have been striving to analyse low birth data using statistical methods, some of them more accurate than others. However, statistical methods have limited capacity to unearth hidden relationships in large datasets [18], [24]. Thus, of recent, data-mining has emerged as one of the most useful solutions to handle enormous volumes of LBW data.

Data-mining is defined as a field that unearths new and useful trends from large historical datasets, by applying diverse methods borrowed from multiple disciplines, including; machine learning, statistics, and soft computing. [25], [26]. It achieves its goal by mining patterns from huge datasets and consequently builds models that could later perform predictive or descriptive tasks. The ultimate goal of predictive data-mining is to apply methods that make future predictions for new unknown cases. In contrast, descriptive data-mining unveil patterns from a vast amount of data [18]. Unveiling patterns can be interesting and useful, however, it is more difficult to gain direct benefits from descriptive data-mining than predictive data mining. Descriptive data-mining only communicates patterns found in the data, while predictive data-mining can be used in real life settings to make predictions on unseen cases [26].

There exists a few studies focused on data-mining models for LBW prediction and these studies are largely based on descriptive data-mining. Thus, previous studies have paid considerable attention to capturing important genetic, demographic, obstetric factors associated with LBW prevalence. The identified factors can be profiled and used to make predictions on birth weight outcomes [27]. However, the applicability of data-mining to estimate the actual birth weight based on the identified risk factors has not been studied.

Another noteworthy gap is that these models' performance assessment is often carried out on training and testing sets, which can lead to overly optimistic results. However, research by [28] states that an unbiased evaluation of model performance is realized when an entirely new third test is used for final scoring of the model. Performance on training and testing datasets only offers a baseline for how the model will perform, and thus further assessment of the actual performance of the model is fundamental.

As such, it is reasonable to state that a novel solution is necessary to develop a model that predicts future birth weight outcomes based on the identified factors rather than exhaustively listing risk factors for LBW. The study also concludes that model performance should also be performed on a separate independent test set. Therefore the study aims at developing a data-mining model that estimates birth weight values for unborn babies, with a relatively higher AUC. The model makes use of the available socio-demographic and obstetric risk factors.

The prediction model's ultimate goal is to improve the identification of mothers at a risk of delivering LBW babies. The birth weight variable is measured in grams and can belong to either of the following classes; 0 (500 or less), 1 (501-1000), 2 (1001-1500), 3 (1501-2000), 4 (2001-2500), 5 (2501-3000), 6 (3001-3500), 7 (3501-4000), and 8 (4001-4500) and 9 (4501 or more). These classes were derived from the birth weight group from the North Carolina State Centre for Health Statistics of 2006 in **Appendix A**.

One factor that determines the success of a data-mining project depends on the selection of the right algorithm for the problem at hand. Classification is one of the most popular functions in data-mining, therefore this study compares different classification algorithms based on the AUC. Several classification algorithms have been proposed, and for the aim of this study, we have selected decision tree algorithms namely C4.5, REPTree and Random Tree. Decision trees are among the top 10 most commonly used algorithms. Experiments were conducted and the decision tree model with a higher AUC was used for classifying future birth weight values based on an independent synthetic data set. A further performance assessment of the model when applied to new unseen data was conducted.

Two special techniques were applied to advance the predictive capabilities of decision trees, with particular emphasis on increasing the AUC since it has greater value than the classifier accuracy in clinical settings [24], [27].

- Increasing the cases of the underrepresented birth weight classes using SMOTE to handle the issue of imbalanced clinical data.

- Employing feature selection techniques to extract the most relevant features for prediction.

1.3 Problem Statement

Predicting the likelihood of a baby being born with LBW is challenging and overwhelming to medical practitioners. It requires evaluating several different contributing factors from large datasets which reside in multiple data sources. Analytics tools are necessary to help users analyse these large volumes of data to a manageable size, so that relevant sections can be focused on [29].

Numerous studies in the literature have attempted to analyse LBW data using data-mining models [15], [18], [19], [30], [31], [32], however, several problems exist with these studies. Firstly, they have largely focused on exploring the risk factors for LBW, while the prediction of the actual birth weight based on the identified risk factors has received considerably less attention. Secondly, prior work is only limited to two binary outcomes, where the risk is presented as either LBW/not or high risk/low risk. This practice does not convey the degree of the LBW, since the binary classes are too broad. Moreover, previous studies are centred around measuring LBW outcomes using categorical values such as extremely low, very low, low, moderately low and normal birth weight, instead of actual continuous values [11]. Thus, the problem this study seeks to address is the paucity of data on data-mining research that estimates the actual birth weight based on risk factors of LBW.

1.4 General Objective

To develop a decision tree model that predicts birth weight values for new-borns with a relatively higher AUC, in an attempt to identify mothers at a high risk of delivering LBW babies.

1.5 Specific Objectives

1. To increase the class distribution of the minority class using SMOTE technique.
2. To select the most significant features for prediction of birth weight using correlation based (CFS) feature selection algorithm.
3. To select a decision tree model with the highest AUC, by comparing experimental results of C4.5, REPTree and Random tree classifiers.
4. To assess the performance of the selected algorithm based on an independent synthetic dataset.

1.6 Research Questions

1. Can decision trees built from a dataset with artificially increased cases of the minority class possibly yield a higher AUC than the ones built from raw imbalanced data?
2. What attributes are significant for prediction of birth weight values?
3. What is the effect of the Correlation-based Feature Selection (CFS) algorithm on the AUC of decision tree models?
4. Which decision tree algorithm predicts birth weight values with a higher AUC?
5. How does the selected model perform when applied to new unseen patient cases?

1.7 Significance of the Study

This research will be a significant endeavour in automating the prediction of LBW through a novel data mining model. The contribution is two-fold, from the medical perspective and information systems perspective;

1. From the medical perspective, this study develops a data mining model that will predict the birth weight of a new born, in an attempt to identify mothers at a risk of delivering low birth weight babies. The early prediction of low birth weight is of paramount importance, since the number of pregnancies ending in LBW is currently increasing. Early identification of mothers at risk of delivering LBW babies could aid medical practitioners to modify pregnancy management strategies, in an attempt to avert low birth weight delivery. In addition, by being prevised of their increased risk, mothers would be in a position to familiarize themselves with LBW prevention measures without medical intervention, for instance by managing their diet. The prediction model will aid medical practitioners in personalized decision-making with regard to low birth weight prediction. Predicting LBW envisions improved overall quality of maternal healthcare, as preventative interventions will be put in place well in advance.
2. From an information systems perspective, this research offers an approach for modelling medical practitioner's decision making process. The approach offers a means of transforming medical data into useful knowledge to alert medical practitioners of potential outcomes before they transpire.

1.8 Dissertation Organization

The dissertation is organised into five chapters;

Chapter 1 Introduction, this chapter offers the rationale for conducting this research. The research objectives and questions are also stated.

Chapter 2: Literature Review, this chapter provides a background on data mining and information visualisation techniques, as well as the state of the art of low birth prediction research. This chapter also highlights the gaps in the current literature and how this study attempts to curb the identified gaps.

Chapter 3: Methodology, this chapter discusses all the research methods, research design, data collection and data analysis methods and tools employed in this study.

Chapter 4: Presentation of Results and Discussion, presents the results and a discussion of these results on the basis of performance measures and experimentations conducted.

Chapter 5: Conclusions and Recommendations, this chapter provides the conclusion and recommendations based on the outcomes of the experiments.

Chapter 2

2 Literature Review

This chapter brings to light elements related to the data-mining process, its application in healthcare, and the issues that impede the successful implementation of data mining in medical settings. Previous work related to low birth weight prediction are also reviewed.

2.1 Data-mining Process

Data-mining has received a great deal of attention for bridging the gap of interpreting data into knowledge to support decision-making. Thus, data-mining's main role is to unveil hidden knowledge and patterns from large datasets [30], [31], [32]. These patterns have to be comprehensible, significant, and useful, so as to contribute new knowledge. Data-mining is commonly referred to as the Knowledge Discovery in Databases (KDD), whereas it is a phase in the KDD process [33], [34], as demonstrated in *Figure 1*.

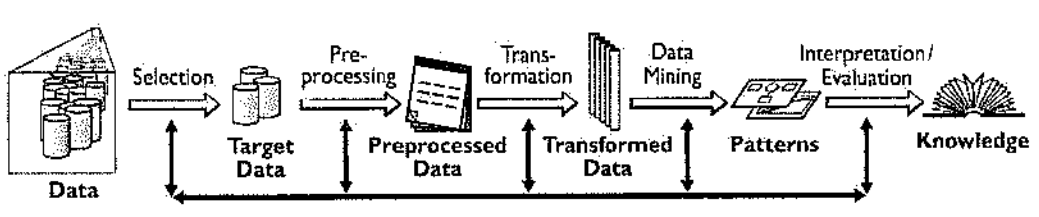


Figure 1 An overview of the steps in the KDD process [35]

KDD is defined by Fayyad [35] as the process of utilizing data-mining techniques to extract hidden knowledge from a dataset. KDD employs supplementary techniques such as data pre-processing, data sampling and data transformation if required. KDD consists of five steps:

Selection: In this phase the focus is on building a target dataset or concentrating on a subsection of attributes on which knowledge discovery will be executed.

Pre-processing: This phase involves data cleaning where by missing values, outliers, noise and other anomalies are detected and corrected to obtain consistent data.

Transformation: Data transformation makes use of dimensionality reduction and data is converted into the correct format required for knowledge discovery.

Data-mining: The data-mining step implements learning algorithms also known as classifiers, on training data so that they can learn patterns that can be used for future predictions. It is commonly known that more training dataset leads to a more reliable model. The success of data mining lies in the right choice and integration of learning algorithms to effectively solve given problems [36].

Interpretation/Evaluation: This stage puts special emphasis on evaluation and interpretation of the discovered knowledge. It also involves deployment of models to real life settings.

2.2 Data-mining Tasks

Data-mining tasks can be grouped into two main categories, namely: predictive data mining tasks and descriptive tasks. Descriptive data-mining seeks to build models to uncover patterns that are interpretable to humans, whereas predictive data mining aims at forecasting some answers of interest. In predictive data-mining attributes can either be predictor (independent) or response (dependent) attributes. Predictor attributes serve as inputs to the predictive model. The response attribute holds the outcome we obtain from the prediction model. The response attribute in predictive data-mining is predefined, however, on the other hand, in descriptive data-mining there is no need to predefine the response attribute.

The response can either be numerical or categorical, thus bringing us to subcategories of predictive data mining, that is, classification (categorical) and regression (numerical) [36]. An overview of the data mining tasks is represented in *Figure 2*.

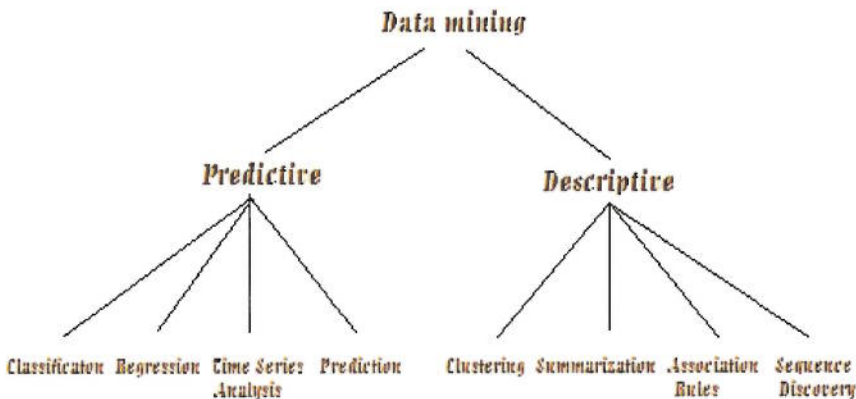


Figure 2 Data-mining tasks

2.2.1 Predictive data-mining (classification) methods

Classification is a supervised learning method whose principal objective is to build models based on known data and use the model to predict the category for new data [6], [30], [32]. In classification, dividing a supplied dataset into training and test sets develops models. The training set is run through one or numerous classification algorithms and the classifier models are subsequently built. The test set is then used to evaluate the performance of the models. Lastly, the classifier model together with various evaluation metrics are presented, demonstrating its capability to correctly classify new cases [6]. Classification methods differ

in the way they employ modelling approaches. Their main basis of comparison is listed below:

- The way in which they handle noise and missing data.
- The kind of attributes they take as input (categorical, ordinal, continuous).
- Presentation of models.
- The models' capability to explain the conclusions made when models are used for decision-making.
- And finally how models generalize or perform with new unknown cases.

We discuss some of the classification methods that frequently appear in the literature.

2.2.1.1 Decision Tree Classification

According to several researchers such as Senthilkumar and Paulraj [15], Jeyarani et al. [30], Gupta et al. [32], Adeyemo and Adeyeye [33], Podgorelec et al. [37], Yahia and El-taher [38] decision trees are flow chart like trees where each internal node depicts a test on an attribute, each branch represents an outcome of the test and leaf nodes represent classes, whereas the top node is known as the root node.

In this study, decision trees were used for classification and prediction due to the following reasons; they have been reported to perform relatively better than other classification algorithms, not only in their construction but accuracy as well. Furthermore, they can deal with high dimensional data without compromising their performance. Another notable benefit of decision trees is their capability to display model results in comprehensible formats [38].

2.2.1.2 Construction of Decision Trees

The prime aim of decision tree classification is to map cases with the same behaviour to the same class. Decision tree building employs two main methods, the top-down approach (general to specific) and the bottom-up (specific to general).

The top-down approach begins with a broad partition and recursively divide cells until certain local stopping criterion is satisfied [15], [30], [32], [33], [37], [38]. In contrast, the bottom-up approach, begins with the largest possible number of leaves, each entailing either a single case from the database or all the cases with the same values for each predictor. At every level, cells are merged into one broad cell, in the end, all cells form the universal partition at the root.

Top-down approaches are the most prominent even though they are often coupled with bottom-up approaches. This is known as over-fitting and pruning a tree. A tree is purposely grown into a very large tree and recursively cut down into a smaller size using a pruning strategy that varies from the growing strategy [39].

2.2.1.3 Decision tree learning algorithms

There exists a number of decision tree learning algorithms, also known as classifiers [15]. The top widely adopted decision tree learning algorithms are Iterative Dichotomiser 3 (ID3), Classification and Regression Tree (CART), Chi-squared Automatic Interaction detection (CHAID), C4.5, REPTree and Random trees [40]. In this study, C4.5, Reduced error pruning trees (REPTree) and Random trees are compared so as to select the best performing decision tree model. The ability of these algorithms to handle multi-classification problems, missing data, and interpretability of models [15], [30], [32], [33], [37], [38], are among the reasons for selecting these algorithms. This section provides a discussion of the selected algorithms.

2.2.1.3.1 C4.5 Decision Tree

The C4.5 decision tree has been extensively applied to medical problems. C4.5 [41], [42] is a successor of ID3, developed by Quinlan Ross (1993). Comparative to earlier versions of decision trees, C4.5 algorithm takes both continuous and categorical attributes as training input. It divides continuous data into discrete intervals, this is widely known as discretization. The data is arranged at every node in order to determine the best splitting feature using gain ratio impurity technique. Another strong point of C4.5 algorithm is its ability to handle missing values. It accomplishes this by using probability values for missing values. These values are computed from the identified frequencies in that instance. For instance, let X be a Boolean variable. If variable X has six values with $X=1$, and four with $X=0$, then according to the probability theory, the probability of $X=1$ is 0.6 and the probability of $X=0$ is 0.4. At this stage, the attribute is partitioned into two fractions: the 0.6 fraction of the instances is allocated to the branch for $X=1$ and the 0.4 fraction is allocated to the other branch of tree [43].

2.2.1.3.2 REPTree

REPTree algorithm is a fast classifier that builds a classification/regression tree via information gain/variance. It prunes its trees by means of reduced-error pruning (with back-fitting). REPTree algorithm only sorts the instances for numeric variables once. Missing values are handled in the same manner as C4.5 Decision Tree [44].

2.2.1.3.3 Random Tree

In recent years, the random tree learning algorithm has received immense attention in medical research. With k random variables at each node, a random tree is a tree constructed at random from a wide range of possible trees. Each tree has an equal probability of being selected. Put differently, the distribution of trees is uniform. Random trees are very efficient and the fusion of large random trees normally leads to accurate models [44]. Other classification methods are discussed below;

2.2.1.4 Other classification algorithms

Other alternative classification algorithms are explained in this section together with their merits and demerits, and why there were not found fit to be used in this research.

2.2.1.4.1 Artificial Neural Networks classification

It is stipulated in many works such as Kaur and Wasan [6], Gupta et al [32], Catley et al [27], Sana et al [45], that Artificial Neural Networks (ANNs) simulate human brain and consist of a set of nodes called neurons. ANNs provide a learning mechanism to perform a task given the initial data. The merits of using ANN are that they are fault tolerant and can make their own representation of information. However, ANNs are difficult to visualize and suffer from slow convergence time during model training.

2.2.1.4.2 Naïve Bayes classification

Naïve Bayes is described by Jeyarani et al [30], as a supervised learning and statistical method that does classification based on probability. Naïve Bayes reads data from the training sets and uses Bayes theorem to forecast probabilities of all classifications. For every instance, the classification that has the highest probability is selected as the prediction class. Naïve Bayes are suitable for solving multi-classification problems [46], however, comprehending their results still remains a challenge.

2.2.1.4.3 Support Vector Machines classification

Support Vector Machines (SVMs) are non-probabilistic binary linear, supervised learning models that perform classification by marking each instance in a training set as belonging to one of two classes. An SVM then builds a model that maps the new cases into one class or the other. They perform well with interactions among features and repeating features [32]. Similar to Naïve Bayes classification, SVMs are suitable for solving multi-classification problems [46], however, interpreting their results is challenging.

The next section highlights the methods that are used for evaluating data mining models developed for healthcare settings.

2.3 Evaluating Data-mining Models Developed for Healthcare Settings

Data-mining research has made advancements in developing models aimed at supporting decision making. As novel models are being developed, performance measures are essential to guide research in promising directions. Evaluation has emerged as one of the components that play a major role in moulding research and therefore data-mining models need to be assessed so as to determine if they indeed serve the purpose they have been developed for. The selection of a performance measure is largely dependent on the specifics of the problem at hand. There is no globally suitable performance measure, one performance measure may be suitable in one context and fail to give a correct estimation on the expected performance in a different scenario. Therefore, choosing the suitable performance measure for a specific problem is necessary to avoid the possibilities of a failed data-mining operation.

Data-mining classifiers are typically evaluated by a confusion matrix [24], as illustrated in *Figure 3* (for binary classification problem) and it is discussed below, together with other metrics derived from the matrix.

2.3.1 Confusion Matrix

Tefera [24] explains the confusion matrix is a useful measure for assessing how well classifiers can identify instances of diverse classes. The columns in a confusion matrix represent the Predicted class and the rows represent the Actual class. The TN (True Negatives) is the number of negative instances that were correctly predicted as negative instances. FP (False Positives) is the number of negative instances that were incorrectly predicted as positive. FN (False Negatives) is the number of positive instances that were incorrectly predicted as negative and TP (True Positives) is the number of positive instances that were indeed correctly predicted as positive.

		Predicted Class	
		P+	P-
Actual Class	C ₁ +	True Positive(TP)	False Negative(FN)
	C ₂ -	False Positive(FP)	True Negative(TN)

Figure 3 Confusion matrix for a binary classification problem [24]

Using the confusion matrix a four extreme cases can be deduced [47]:

1. **A Perfect Classifier** – This classifier correctly classifies all instances. All the positive instances are correctly classified as positive and all the negative instances are correctly classified as negative.

2. **The Worst Classifier** – This classifier is incapable of correctly predicting any positives or negatives.

3. **An Ultra-Liberal Classifier** – This classifier predicts all instances as positive.

4. **An Ultra-Conservative Classifier** – This classifier predicts all instances as negative.

Several evaluation measures that can be computed from the confusion matrix include the accuracy, precision, recall, F-measure and ROC curve.

2.3.1.1 Accuracy

In the initial stages of data-mining, accuracy (Acc) was the most extensively used measure of performance. Accuracy measures the overall effectiveness of the algorithm. It is computed by measuring the percentage of correctly classified instances out of all the instances; $Acc = (TP+TN) / (TP+TN+FP+FN)$. Even today accuracy is being used to evaluate novel classifiers. However, it has been criticized for providing over optimistic results, especially when the distribution of data is highly imbalanced [48], [49]. Data-mining algorithms would often forecast the most frequent class in these problems, thus, a great identification of the majority class can translate to a great accuracy, regardless of how the model recognises the minority class. Such distributions are common in medical data [50], [51], therefore the most appropriate evaluation measures need to be used to evaluate the effectiveness of models before they can be applied to real life clinical practices.

Frequently recommended metrics for evaluating medical data-mining models are the ROC and AUC, sensitivity (recall), specificity, F-Measure, and precision [52]. These metrics are discussed in the subsequent section.

2.3.1.2 F-Measure

F-measure is a metric that uses two other metrics (precision and recall) to measure performance.

Precision means the number of positive instances classified divided by all the instances classified. Thus, precision is denoted as follows;

$$Precision = \frac{TP}{TP + FP}$$

Sensitivity is also known as the True Positive Rate (TPR) or recall, this is the ratio of the number of positive instances classified over all the positive instances [52]. Sensitivity is computed as follows;

$$Sensitivity = \frac{TP}{TP + FN}$$

Based on these definitions F-measure is defined as demonstrated below;

$$F - measure = \frac{2 \times precision \times recall}{precision + recall}$$

2.3.1.3 Specificity

Specificity is used for the purpose of measure the proportion of negative cases that were correctly classified as negative, which is $1 - FP$ (False positive) or can be denoted as follows, [24], [52];

$$Specificity = \frac{TN}{TN + FP}$$

2.3.1.4 ROC curve (AUC)

The ROC curve is a two-dimensional graph representing the ratio of false positive and true positive rate. On an ROC curve the X-axis represents the percentage of the FP ($1 - specificity$) = $FP / (TN + FP)$ and the Y-axis represent the TP (sensitivity) = $TP / (TP + FN)$. It has been recommended as an appropriate measure of diagnostic accuracy by clinical epidemiologists [48], [52].

Two ROC curves for two models are shown in **Figure 4**. The curve depicts a diagonal line where for each true positive of a certain model, it is likely to encounter a false positive [24]. Thus, if the ROC curve is closer to the diagonal line, then the model is less accurate. If the performance of the model is outstanding, it is more likely to encounter true positives. Therefore, the ideal point on the ROC curve would be (0, 1) that is all positive instances are classified correctly and no negative examples are incorrectly classified as positive. The AUC is a standard performance metric for a ROC curve. It takes any value between [0, 1]. A random classifier yields an AUC of ~0.5. A larger AUC represents a better model [47], [48], [49], [53], [54], **Figure 4** shows a visual example of an AUC.

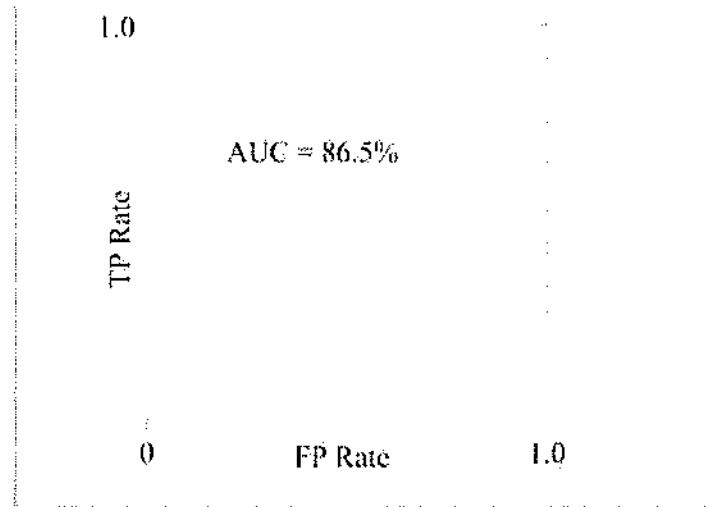


Figure 4 A depiction of an AUC for a particular classifier [47]

2.4 Prior Application of Decision Trees in Healthcare.

The healthcare industry is amongst the industries where the quality of decision making, directly affects the quality of life. Therefore, accurate diagnosis and prognosis is crucial. Literature regarding applications of data mining in healthcare indicates that different data mining techniques and technologies have been widely adopted to solve healthcare problems ranging from; informing diagnostics, disease impact forecasting, scaling down healthcare costs, healthcare insurance claims identification, and medical malpractice prevention [54], [55], [56].

In Pakistan research by Sana et al [45], demonstrated the success of data-mining by developing a Birth Prediction Classification model (BPC) based on a high dimensional dataset with 50 factors that influence birth. Two data mining algorithms, J48 (C4.5) and multilayer perceptron neural network were utilized for model building. Chi square method was used to extract relevant factors and the models were evaluated using cross validation. The results of the study indicated that C4.5 decision trees can be successfully applied to predict birth delivery mode with a recall (sensitivity) of 80.2 % and an AUC of 81.6 %. The Multilayer Perceptron neural network, obtained a recall of 82.3 %, and an AUC of 90 %. While the Multilayer Perceptron neural network outperformed the decision trees, neural networks are generally slow during training and their results are somewhat difficult to interpret. This study advocates for the use of data-mining techniques in the medical domain.

Summarizing the results of work by Geregiorgis [57], we conclude that decision trees are suitable for the healthcare domain. This work applied J48 decision tree and Naïve Bayes to

predict the possibility of the use of contraceptive methods among women aged between 15 and 49 years. The study employed the 2005 Ethiopia Demographic Health Survey data. The J48 decision tree was selected as the best model with an overall accuracy of 82.85%, a sensitivity of 87.3% and specificity of 74.7%. The Naïve Bayes produced an accuracy of 85.87%, a sensitivity of 89.7%, and a specificity of 53.8%. Although the naïve Bayes algorithm had a comparative larger accuracy and sensitivity, there was a discrepancy in classifying true positive values and true negative values. The true negative values were fairly low.

In their research, Hailu [28], conducted a comparative study on the prediction capability of decision trees, Neural network, and , logistic regression in building HIV test prediction models. The models predict HIV test outcomes among adults in Ethiopia based on the 2011 Ethiopia Demographic Health Survey (EDHS) data. The results revealed that the decision trees performed better than other classifiers, with random tree algorithm producing an accuracy of 96%, and C4.5 (J48) decision tree obtaining 79% classification accuracy. Other algorithms' classification accuracy was as follows; neural network (78%) and Logistic regression had the least accuracy of 74%.

A review of a study by Soleimani et al [58], affirmed the applicability of decision trees in healthcare settings. This work designed a C4.5 decision tree using a dataset from the Tabriz healthcare centre with 80 pregnant women and 5 attributes (age, number of pregnant, heart status, blood pressure and delivery time). The decision tree's goal was to classify caesarean section into two classes, yes or no. The study used WEKA software and the C4.5 decision tree was able to successfully classify caesarean section with an ROC of 94.6 %, recall of 86.3% and specificity of 89.1%.

A study by Taft et al [59] demonstrated an improved performance of C4.5 decision tree and Naïve Bayes by developing synthetic cases with SMOTE and applying 10-fold cross validation method. This research developed prediction models for finding Adverse Drug Event (ADE) in women admitted for Labour and Delivery. The predictor variables were patient risk factors and comorbidities. The data came from the Enterprise Data Warehouse (EDW) of Intermountain Healthcare in Salt Lake City, Utah. 135,000 individual patients from years 2002-2005 were extracted. The study concluded that models developed with the use SMOTE can aid the development of evidence based ADE monitoring systems.

2.5 Challenges of Mining Healthcare Data

Numerous difficulties are encountered when data-mining healthcare data. Healthcare datasets are generally complex, they are large, imbalanced and possess high dimensionality. If the complexity is not dealt with, the data-mining process can become daunting and produce poor results. This section consists of a description of the challenges of mining medical data, alongside methods that are used to counteract the challenges.

2.5.1 Imbalanced Data

Medical datasets predominantly suffer from data imbalance. The data imbalance problem is realized when the number of instances in one class is much smaller (minority class) than the instances another class (majority class). Learning in imbalanced dataset is problematic, especially when the outcome of interest is the minority class. Classifiers end up incorrectly classifying the minority instances as belonging to the majority class. From a medical perspective, misclassification of such cases is expensive, since medical practitioners are often interested in correctly classifying rare outcomes [48], [50], [51].

A series of methods for enhancing performance of algorithms when dealing with imbalanced data have been proposed. These methods can be grouped into three wide categories, data centred (sampling methods), algorithm centred and hybrid solutions [48]. This study employed data centred methods and therefore this section will only be limited to the discussion of data centred methods.

Data centred methods are concerned with modifying the distribution of data. This task is accomplished either randomly or deciding which instances to remove from the majority class or add to the minority class, or by artificially generating data. Techniques under this category include random over-sampling, under-sampling, or more advanced techniques, such as SMOTE [47].

SMOTE is one of the most popular oversampling techniques. It has been successfully applied in number of studies such as [24], [48], [51] and serves as a base for other techniques [60]. SMOTE performs its balancing by creating new synthetic cases of the minority class by considering its neighbours. The resulting oversampled cases are not simple duplication of the data but an expression of the variable values of the complete dataset [59]. For this reason SMOTE was employed in this study.

2.5.2 High Dimensionality

Medical data has databases with millions of records and large numbers of variables. These datasets increase the likelihood that a data-mining classifier will obtain spurious patterns that are invalid. Possible solutions include very efficient algorithms, massive parallel processing, dimensionality reduction techniques, and incorporation of prior knowledge [61]. Dimensionality reduction through the use of feature selection was utilized and is thoroughly discussed below;

2.5.2.1 Dimensionality Reduction through Feature Selection

While some datasets require no or little preparation prior to the data mining exercise, others, especially high dimensional datasets require thorough preparation. This is due to the following problems that exist with high dimensional datasets [61];

1. Computational costs increase as features (dimensions) increase. This is known to as the curse of dimensionality.
2. Some features within the dataset may be irrelevant to the task at hand and hence impacting the reliability of the results.
3. There may be association between features in the dataset that may affect the classification performance.

Feature selection also known as relevant analysis [61], [47] reduces dimensionality (number of features) by selecting a subset of top features while eliminating irrelevant or redundant features. Irrelevant features are the ones that provide no useful information for solving the problem at hand. Feature selection optimizes model performance by avoiding overfitting, it also reduces training time and increases comprehensibility of the mined results [47], [62].

Feature selection algorithms can be categorized into two wide groups, **wrappers and filters**. Wrappers employ the learning algorithm to examine the relevance of features. Filters examine features according to some heuristics based on the overall characteristics of the data. This is done independently of the data-mining algorithm to be used. Filters have been renowned for this independency. Studies have reported that filters are relatively faster than wrappers [61], [63]. Hence filters are used in this research. The following section provides a discussion of low birth weight and its determinants.

2.6 Low Birth Weight and Its Determinants

Earlier studies have established premature delivery (short gestation < 37 weeks) and foetal growth retardation as the major two causes of low birth weight. These two conditions are affected to a great extent by numerous factors that are closely related to the infant, the mother or the physical environment. The factors play an important role in determining the new-born's birthweight and future health outcomes [10], [12].

It is well known to medical practitioners that factors that largely influence foetal growth and development are infant gender, parity, maternal height, ethnic origin, pre-pregnancy weight, paternal weight and height, maternal birth weight, history of low-birth-weight infants, gestational weight gain and caloric intake, the mother's gestational nutrition, diet, general morbidity and episodic illness [10], [22].

Further investigation into the determinants of low birth weight revealed that firstborns weigh less than subsequent infants, girls are lighter boys, and twins are fairly lighter than singletons. Young women, women of short stature and low pre-pregnancy weight women frequently deliver low birthweight babies [12], [22]. Furthermore, intake of toxic substances such as alcohol, tobacco and other drugs is significantly correlated with slow foetal growth. Mothers with poor socio-economic backgrounds are also prone to delivering low birth weight infants. In those contexts, the baby's low birthweight stems mainly from poor gestational nutrition. Physically demanding work during pregnancy is also a contributing factor to poor foetal growth [12], [14].

On the other hand, gestational duration's well established causal factors include pre-pregnancy weight, history of premature births or spontaneous abortion, and cigarette smoking. Other factors are maternal complications such as hypertension, and high prevalence of infections such as HIV, syphilis, urinary tract infection and other infections [10], [12].

In conclusion, low birthweight defines a varied cluster of infants: some are delivered prematurely (preterm birth), some with restricted growth, and others are delivered both early and with restricted growth. Preterm birth is recognized as a leading cause of death, morbidity and disability. Low birthweight due to restricted foetal growth affects the person throughout their entire life and is linked with a higher incidence of adult diseases, such as type 2 diabetes, hypertension and cardiovascular disease. An auxiliary risk for girls is giving birth to low birth weight infants when they become mothers [12], [13], [14], [15], [16], [17], [18], [19], [20]. Thus, low birthweight is an important indicator of health; and efforts must

therefore be channelled into measuring birth weight as accurately as possible and planning infant care accordingly.

2.7 Low Birth Weight Prediction (Related Works)

In the past years, researchers have made efforts to predict LBW babies and LBW risk factors. The aim of this section is to evaluate the current techniques used in medical research to estimate low birth weight. These techniques include both statistical and data mining techniques.

2.7.1 Low Birth Weight Prediction Using Statistical Models

In their study, Ben-haroush et al [22] presented an evaluation of the accuracy of sonographic measures in estimating foetal weight at 28-34 gestational age to predict small and large for-gestational-age babies (SGA, LGA) at term. This study was carried out on 259 low risk births whose measurements were taken at 28-34 weeks of gestation. The participants were recruited at term delivery. The birth weight were transformed into percentiles based on locally developed growth charts and compared. A statistical method called multivariate linear stepwise regression analysis was employed to predict the birth weight and birth weight percentile. The Bland and Altman plot and Passing and Bablok regression were utilised to measure the calculated birth weight against the actual birth weight. Both the sonographic estimated foetal weight and the computed birth weight bore low positive predictive values in predicting SGA or LGA infants. However, the computed birth weight had high negative predictive values of 99.5% and 100%, for SGA and LGA infants respectively. These values suggest that the data was highly imbalanced and in conclusion, this study produced poor results and it was only concentrated on low risk pregnancies. Therefore, further studies targeting high risk pregnancies was proposed.

Research by Moraes et al [64] employed classical logistic regression and multilevel logistic regression models to discover risk factors for low birth weight for singleton live born babies in Rio Grande do Sul State, Brazil, in 2003. The research used the live births data from an Information System. The study successfully identified a correlation between low birth weight and frequency of prenatal visits, prematurity, maternal age, inherited anomalies, childbirth delivery mode, sex, marital status, maternal occupation, place of delivery, parity, and schooling. The multilevel models established a correlation between urbanization of the micro-region. Less urbanized micro-regions, present higher risks of single mothers giving birth to LBW babies.

Similar to the abovementioned study, Yadav and Lee [14] applied statistical analysis to low birth data to build a LBW predictive model. A cross sectional study was conducted at Seremban General Hospital in Malaysia between May 2007 and March 2008. The number of deliveries throughout the time of the study and associated maternal factors were documented. The total number of records was 666, 573 normal and 93 LBW births. The Chi-square statistics and odds ratios examined the correlation between LBW and maternal factors. Logistic regression was used to develop a LBW predictive model. The model was used to forecast the factors associated with LBW in terms of order. The ROC curve estimated the sensitivity and specificity of the LBW model. Another model, step wise regression model, was developed and independent variables were included in a hierarchical form and the likelihood ratio test was used to do a comparative analysis between the initial and the nested models which emanated from the first model. Among the maternal factors, older women (>35 years) were noticed to likely have LBW babies, mothers from poor socio-economic background also were bound to have LBW babies. Other significant risk factors of LBW are high and low blood pressure. The LBW model yielded a sensitivity rate of 80% and specificity of 75 % using ROC curve.

Das et al [16] studied mother's lifestyle characteristics that have a significant effect on her baby birth weight using Joint generalized linear log-normal statistical modelling of mean and variance. A dataset that was originally collected in 1986 at Baystate Medical Centre, Springfield, Massachusetts was utilized. The data had maternal information about 189 women and 11 attributes. The analysis was able to detect that the mother's age, race, history of premature labour, her weight at last menstrual period, smoking status, the presence of uterine irritability and hypertension history, are statistically influential to the baby's birth weight. Contrary to the previous study, this work concluded that joint log-normal models are more effective than conventional simple, multiple, logistic regression.

In South America, a study by Marshall et al [65] developed and validated a model for very low birth weight neonatal mortality prediction, using historical birth data in 16 neonatal intensive care units from Argentina, Chile, Paraguay, Peru and Uruguay (NEOCOSUR Network.). All infants weighing 500-1500 grams from the 1st of October 2000 to the 30th May 2003 were included in this study. The model was developed using 1801 instances, while 450 instances served as a test sample. This model was compared with two existing scores by using the AUC, and it was highly predictive for in-hospital mortality with 85%

AUC. The new model outperformed the Clinical Risk Index for Babies (CRIB) and the NICHD risk models when applied to the NEOCOSUR Network.

In Bangladesh, research was conducted by Nasreen et al [66], based on 720 expectant women from two rural sub-districts to examine the effect of maternal antepartum depressive and anxiety symptoms on infant low birth weight. Antepartum depression symptoms were examined using the Edinburgh Postpartum Depression Scale (EPDS), and antepartum anxiety, using the State Trait Anxiety Inventory (STAI). Birth weight of 583 singleton live babies born at 37 gestational age (weeks) and above, was recorded within 48 hours of delivery. Structured interviews were also carried out and descriptive statistics were performed using Chi-square, Fisher's exact, and independent-sample t tests. A multiple logistic regression model was employed to detect predictors of low birth weight. The research revealed that maternal antepartum depressive and anxiety symptoms are significantly related to low birth weight.

Vanka et al [67] carried out a study to scrutinize the relationship between maternal periodontitis and the risk of low birth weight using logistic regression analysis. Two hundred and thirty two (232) expectant women who had reached the second trimester and with at least 4 teeth with periodontal pockets were studied. Structured interviews were carried out after the delivery to gather data concerning the weight of the baby. The periodontal status was examined based on Community Periodontal Index (CPI) and Loss of Attachment (LOA) as per the WHO recommendations. A large proportion of the population (134) with deep pockets delivered babies with low birth weight, as compared to 98 who had normal weight babies.

In Cleveland, Hack et al [68] conducted a longitudinal study of 242 very-low-birth-weight young adults and 233 normal weight young adults born between 1977 and 1979 at Rainbow Babies and Children's Hospital. The main aim of the study was to assess the level of education, cognitive and academic performance and levels of chronic illness and risk taking behaviour of these young adults at the age of 20 years. Univariate comparisons between the very-low-birth-weight and normal-birth-weight groups were made with the use of Student's t-test for continuous variables and with the chi-square test or Fisher's exact test for discrete variables. Logistic regression was used for dichotomous outcomes, and multiple linear regression was used for continuous outcomes. It was reported that 74% very-low-birth-weight young adults graduated from high school compared to normal-birth-weight young adults (83%). Very-low-birth-weight subjects had a lower average IQ (87 compared

to. 92). They also had greater levels of neurosensory deficiencies (10 percent vs. <1 percent, $P<0.001$). Thus low birth weight has a significant effect on the level of education, cognitive and academic performance of low birth weight adults. The next section discusses work related to the utilization of data-mining to predict low birth weight.

2.7.2 Low Birth Weight Prediction Using Data-mining Models

Senthilkumar and Paulraj [15] applied a wide range of data-mining algorithms to predict LBW babies. The algorithms were Naïve Bayes', neural network, random forest, classification tree, SVM, and logistic regression. The study also used the dataset from Baystate Medical Centre, Springfield, Massachusetts. The classification tree yielded a higher prediction accuracy of 89.95% when using validation method on training data. Random forest gave a lower overall prediction accuracy 70.9% than other algorithms. This work identified mother's last weight (pounds) prior to becoming pregnant, mother's age, the number of physician visits during the first trimester, and the number of previous premature labours as outstanding influencing factors of LBW. Smoke, hypertension, urine irritability, and race were said to be weak influencing factors of LBW. The results of the study were compared with that of logistic regression in other works and logistic regression's prediction accuracy was outperformed by the data-mining techniques employed in this study.

Desalegn [18] introduced J48 and PART rule induction algorithm (with pruning) in predicting low birth weight using Ethiopia demographic and health survey data. Experiments were conducted on a dataset containing 9, 861 records dataset. J48 decision tree classifiers using pruned technique achieved an accuracy rate of 94.7%. PART rule induction algorithm gave an accuracy rate of 94.35%. In spite of the high accuracy rates for both algorithms the research had some hiccups. The results of his study could be biased because evaluation was on the basis of 10 -fold cross validation and thus the actual performance of the model based on unseen data was ignored. Moreover, the class attribute was only limited to binary outcomes (low/normal birth weight). A summary of the statistical and data-mining related studies is discussed below. The next section also sums up the literature review chapter.

2.8 Literature Review Summary

It was observed in the obstetrical research that, statistical methods are predominantly used for estimating low birth weight. Despite the substantial results yielded by classical statistical methods, they have been largely criticized for having limited capacity to discover hidden

relationships in large datasets [18] [69]. Large datasets are comprised of a wide range of potential variables which make the possible combinations of variables highly complex for statistical methods to analyse these data in a reasonable amount of time. Therefore, statistical methods alone, are insufficient for handling complex medical datasets. In contrast, data-mining can handle large and complex datasets.

Thus, the unsuitability of statistical models to handle complex datasets can be solved by the use of data-mining techniques [24]. However, what remains unresolved about the examined studies is the use of data-mining techniques for predicting birth weight based on LBW risk factors. In addition, these models may have not offered means for evaluating model performance on a distinct independent dataset.

Thus, after thrashing out the literature the study concludes that decision trees have the potential predict birth weight based on LBW risk factors. To the best of our knowledge, a study of this kind is novel in LBW prediction. The methodology followed by this study is outlined on the next chapter.

Chapter 3

3 Methodology

This chapter discusses the architecture of our approach, the data-mining techniques, and research methods used in this research. All the data preparation and experiments were performed in the Waikato environment for data analysis (WEKA) workbench. WEKA data-mining workbench is an open source and free software that was developed in New Zealand at the University of Waikato using JAVA programming language. This workbench has been widely accepted in both academia and the corporate world. It offers researchers a platform to conduct experiments by applying various to datasets of their choice [70]. Therefore, making it suitable for the purpose of this research.

3.1 An Architecture of the Research Approach

The research approach architecture mainly consists the data-mining process and an investigative method to assess the predictive power of the best selected model on unknown data. The workflow entails acquisition of the dataset, data selection, data pre-processing, model building based on three classifiers (C4.5, REPTree and Random tree), model evaluation and the actual usage of the selected model on an independent dataset. The architecture is represented in *Figure 5*.

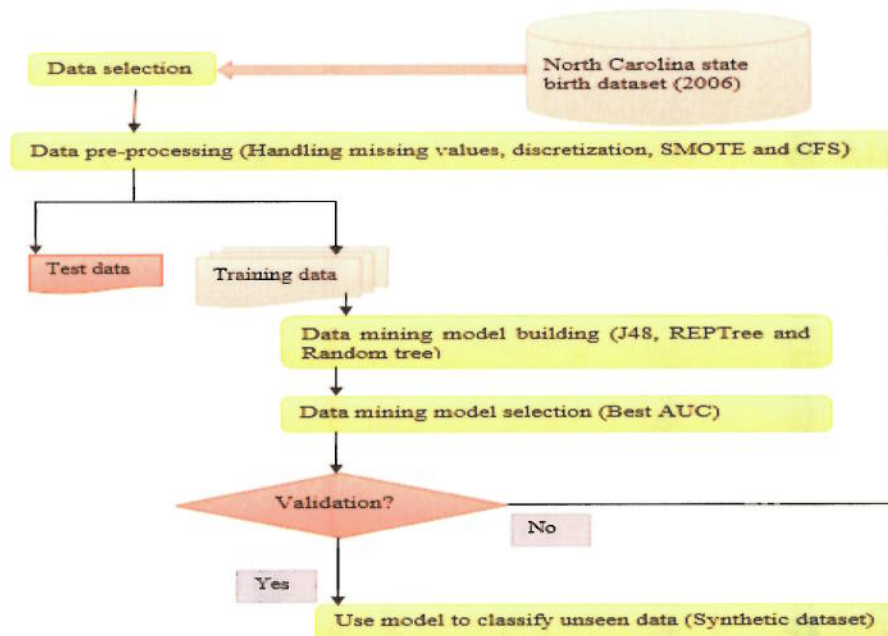


Figure 5 Research approach architecture

3.2 Dataset

The research utilized a publicly available dataset containing births recorded at the North Carolina State Centre for Health Statistics in 2006. The target variable was the birth weight group. The target variable was measured in grams and it had 10 classes; 0 (500 or less), 1 (501-1000), 2 (1001-1500), 3 (1501-2000), 4 (2001-2500), 5 (2501-3000), 6 (3001 -3500), 7 (3501-4000), and 8 (4001-4500) and 9 (4501 or more). Using the target variable, the mothers who are likely to deliver birth weight babies can be identified. Initially, the dataset contained a total population of 10, 000 births and 131 continuous and categorical predictor variables as indicated in the data file description in **Appendix A**. The next step was to select relevant data for model building.

3.3 Data Selection

The data file description in **Appendix A** was used to gain understanding of the variables and structure of the dataset. Anomalies and redundant variables and other variables that were not relevant to this study were eliminated using Microsoft Excel software. These included; the place of residence, date of birth, month of birth, day of birth, year of birth, Date of Last Live Birth (Date Lbirth), Month of Last Live Birth (month Lbirth), Year of Last Live Birth (year Lbirth), Date of Last Termination (Date term), Month of Last Termination (month term), Year of Last Termination (year term), and race. Prenatal, trimester prenatal, visits, Kessner index, Birth weight (lbs), birth weight remainder, birth weight (g), low birth, race mom, race dad, father minority, mother minority, and smoker were all redundant and were deleted. Features that were related to the conditions of the newborn such as gest age (BC- Weeks gestation as reported on birth certificate), Appearance; Pulse; Grimace; Activity; and Respiration (APGAR) score 1 min, APGAR score 5 min , Infant transfer status (infant tran), Mother transfer status (mom tran) were eliminated. The final number of attributes for initial modelling were 90 and were exported to WEKA for pre-processing.

3.4 Data Pre-processing

The most important but tedious task in data-mining is data pre-processing. Real world medical datasets are often noisy, incomplete or inconsistent which can distort useful patterns and make the learning process overwhelming. The possible reasons for inconsistencies are human errors and values that are designed to take no values [71]. Several data pre-processing techniques exist to deal with the inconsistencies in datasets and thus improving the quality of data for effective data mining.

3.4.1.1 Handling Missing Values

Missing values are data with no records. Missing data can significantly impact models performance. Different options are available for handling missing values. One option is to ignore missing values, this option has been reported to lead to biasness in the data. Other options that are preferred to the abovementioned option include [72];

- Filling out the missing data manually, this can be tedious and almost impossible to achieve when dealing with enormous datasets.
- Replace missing values with techniques like mean, mode, and random imputation values.

Our dataset had a considerable amount of missing values and thus the study employed a feature named *“ReplaceMissingValues”* in WEKA to automatically replace missing values. This feature replaces continuous missing values with the mean and nominal missing values with the mode. Attributes with a large number of missing values discarded from analysis. *Age* and *education of the father* attributes accounted for approximately 18% of missing values, and therefore they were deleted. *Table 1*, shows the attributes which had missing values together with how they were dealt with.

Table 1 Missing values together with their respective handling mechanisms

Attribute name	Number of missing values	Percentage of missing (%)	Handling mechanism (WEKA)
<i>County of residence of mother</i>	179	1.79	Replace with mode
<i>Age of father</i>	1,753	17.53	Delete attribute
<i>Education of father (years)</i>	1,778	17.78	Delete attribute
<i>Education of mother</i>	37	0.37	Replace with mean
<i>Total Preg (Total Pregnancies (Including This One))</i>	13	0.13	Replace with mean
<i>Bdead (Number of Children Born Alive Now Dead)</i>	9	0.09	Replace with mean
<i>Terms (Number of Other Terminations)</i>	12	0.12	Replace with mean
<i>Gest Age (Completed Weeks of Gestation (calculated))</i>	8	0.08	Replace with mean

Birth weight group	2	0.02	Replace with mode
NumChild (Number of Living Children)	7	0.07	Replace with mean
AveCigs (Average number of Cigarettes Used Daily)	41	0.41	Replace with mean
AveDrink (Average number of Drinks Consumed Weekly)	49	0.49	Replace with mean
WT Gain (Weight Gained)	305	3.05	Replace with mean

Some attributes related to the *events of labour and/or delivery* such as *Febrile* ($> 100^{\circ}\text{F}$ or 38°C), *Meconium* (Meconium, moderate/heavy), *Rupture* (Premature rupture of membrane), *Abruptio* (Abruptio placenta), *Previa* (Placenta previa), *Bleeding* (Other excessive bleeding), *Preclab* (Precipitous labour (< 3 hours)), *Prolab* (Prolonged labour (> 20 hours)), *Dyslab* (Dysfunctional labour), *Breech* (Breech/Malpresentation), *Cephalo* (Cephalopelvic disproportion), *Prolapse* (Cord prolapse), *Anesth* (Anesthetic complications), *Distress* (Fetal distress), and *Othlabor* (other) had 4 missing values and were replaced with modes.

Medical History related attributes such as *Anemia*, *Cardiac* (Cardiac disease), *Aclung* (Acute or chronic lung disease), *Diabetes*, *Herpes* (Genital Herpes), *Hydran* (Hydramnios/Oligohydramnios), *Hemoglob* (Hemoglobinopathy), *Hyperch* (Hypertension, Chronic), *Hyperpr* (Hypertension, pregnancy), *Eclamp* (Eclampsia), *Cervix* (Incompetent Cervix), *Pinfant* (Previous infant 4000+ grams), *Preterm* (Previous preterm /small infant), *Renal* (Renal disease), *Rhsen* (Rh sensitization), *Uterine* (Uterine Bleeding), and *Othmed* (other) all had 9 missing values and were replaced with modes.

Attributes related to **Congenital Anomalies** such as *Anenceph* (Anencephalus), *Spinabif* (Spina bifida/Meningocele), *Hydrocep* (Hydrocephalus), *Microcap* (Microcephalus), *Othcns* (Other central nervous system), *Heartmal* (Heart malformations), *Othcirc* (Other circulatory/respiratory), *Rectalat* (Rectal atresia/stenosis), *Fistula* (Tracheo-esophageal/fistula), *Omphalac* (Omphalocele/Gastroschisis), *Othgast* (Other gastro anomalies), *Malgen* (Malformed genitalia), *Rangene* (Renal agenesis), *Othuuro* (Other urogenital anomalies), *Cleft* (Cleft lip/palate), *Polydact* (Polydactyly/Syndactyly/Adactyly), *Clubfoot*, *Dhernia* (Diaphragmatic hernia), *Othms* (Other musculoskeletal/integumental),

Downs (Down’s Syndrome), *Othchrom* (Other chromosomal anomalies), and other all had 1 missing value and were replace with modes. The next step was to handle data imbalance.

3.4.1.2 Handling Imbalanced Data

Building predictive models based on imbalanced datasets and multiple classes can be overwhelming and countering imbalance is necessary [24]. WEKA visualizations revealed that the dataset was imbalanced. Class 6 (3001 -3500 grams) was the most frequent with 3, 769 instances. It was followed by class 7 (3501-4000 grams) with 2, 660 instances, then class 5 (2501-3000) with 1, 843 instances, and then Classes 8 (4001-4500 grams) and 4 (2001-2500) with 720 and 531 instances respectively. This can be imputed to the fact that birth weights of 3500-4000 g are the most frequent in obstetrics [10], [11]. The least frequent classes were classes 0 (500 or less grams), 1 (501-1000 grams), 2 (1001-1500 grams), 9 (4501 or more), and 3 (1501-2000) with 32, 80, 75, 112 and 178 instances respectively. The frequency per class is shown in *Figure 6*.

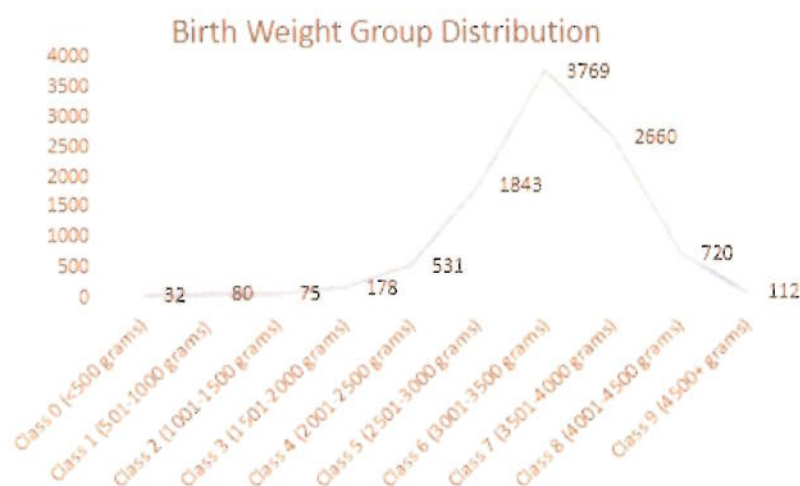


Figure 6 Birth weight frequencies per class

Our objective was to bridge the gap between the minority and majority class distributions by altering the distribution of the minority classes, while maintaining the distribution of the majority classes. SMOTE was found fit for accomplishing this objective. This technique was applied to the dataset to create synthetic instances for the minority classes (0, 1, 2, 3, 4, 5, 8, and 9). The parameters for implementing SMOTE are presented in *Figure 7*.

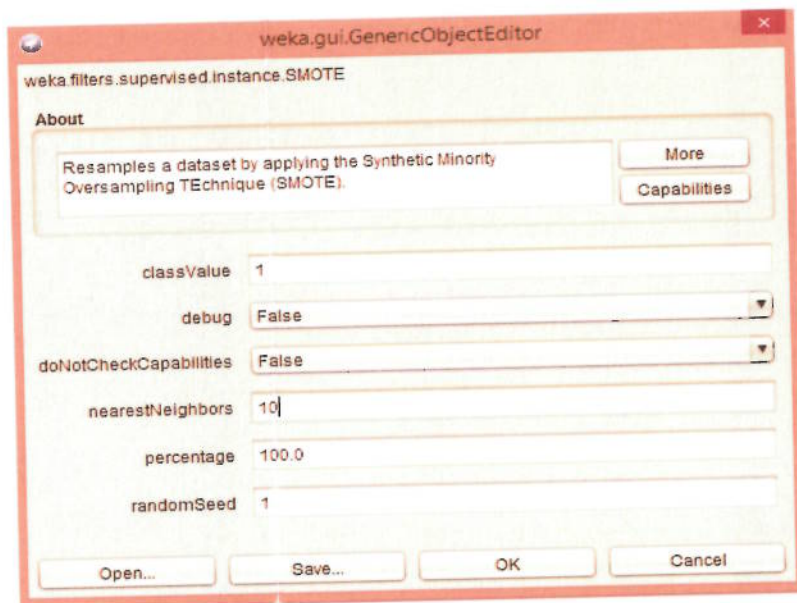


Figure 7 Implementation of SMOTE in WEKA

Depending upon the amount of over-sampling required, neighbours from the k nearest neighbours are randomly chosen [48]. For instance, if the over-sampling needed is 200%, only two neighbours from the five nearest neighbours are chosen and one sample is generated in the direction of each [51]. Our approach used ten nearest neighbours. The number of synthetic cases generated in each iteration is shown in *Table 2*.

Table 2 Number of synthetic case generate by SMOTE

Class Label	Initial count	SMOTE 100%	SMOTE 200%	SMOTE 300%	SMOTE 400%	SMOTE 500%	SMOTE 600%
0	32	64	128	256	512	1,024	2,048 (Stop)
1	80	160	320	640	1280	2,560 (Stop)	2,560
2	75	150	300	600	1,200	2,400 (Stop)	2,400
3	178	356	712	1,424	2,848 (Stop)	2,848	2,848
4	531	1,062	2,124 (Stop)	2,124	2,124	2,124	2,124
5	1,843	3,686 (Stop)	3,686	3,686	3,686	3,686	3,686
6	3769	No SMOTE (3769)	3769	3769	3769	3769	3769

7	2660	No SMOTE (2660)	2660	2660	2660	2660	2660
8	720	1440	2880 (Stop)	2880	2880	2880	2880
9	112	224	448	896	1792	3584 (Stop)	3584

Initially *Class 0* had 32 cases and after SMOTE 600% it had a total of 2, 048 cases. *Class 1* was comprised of 80 cases before SMOTE and after SMOTE 500%, it had 2, 560 cases. *Class 2* entailed 75 cases and after SMOTE 500% it had 2, 400 cases. *Class 3* contained 178 cases prior to SMOTE and held 2, 848 after SMOTE 400%. *Class 4* originally had 531 cases and 2, 124 after SMOTE 200%. The initial class distribution for *Class 5* was 1, 843 cases and 3, 686 after SMOTE 100%. The distributions for *Class 6* and *Class 7* were maintained since no SMOTE was implemented on the two classes. Class 6 had 3, 769 while Class 9 had 2660 cases. *Class 8* had 720 cases and a final class distribution of 2, 880 after SMOTE 200%. Finally *Class 9* had 112 cases and a total of 3, 584 cases after SMOTE 500%. *Figure 8* shows the final distribution in each class after SMOTE 600%.

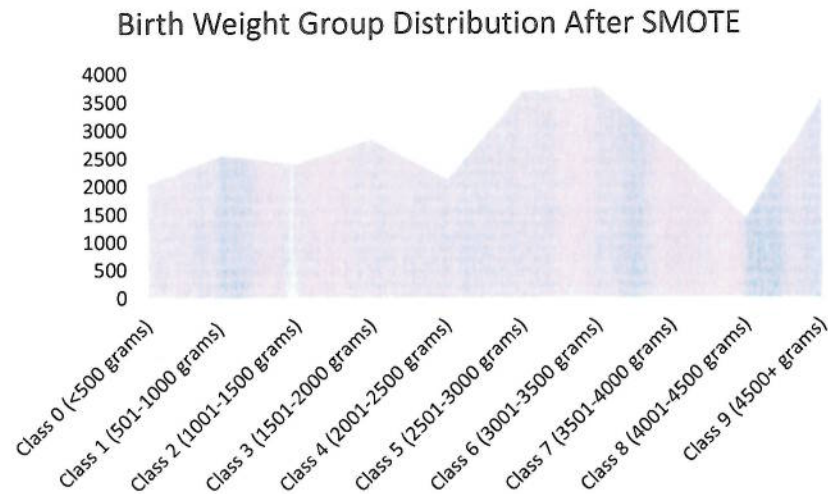


Figure 8 Final distribution per class, after SMOTE 600%

After handling data imbalance, feature selection was employed to reduce data dimensionality since our dataset had 90 features.

3.4.1.3 Feature Selection

By eliminating features that do not add any value to the accuracy and efficiency of the algorithm, performance is enhanced [63], [73]. A filter feature selection method named the

Correlation-based Feature Selection Algorithm (CFS), coupled with the Best fit search method were used to reduce dimensionality. CFS searches for features that are highly correlated with the target classes yet have minimal inter-correlation amongst the features themselves. CFS can be denoted as follows (*Error! Reference source not found.*);

$$r_{zc} = \frac{k\overline{r_{zi}}}{\sqrt{(k + k(k - 1)\overline{r_{ii}})}}$$

Where r_{zc} is the correlation between the added up features, the class attribute, k is the total number of features, $\overline{r_{zi}}$ is the mean of the correlations amongst the features and the class attribute, and $\overline{r_{ii}}$ is the average inter-correlation amongst features. Inter-correlation is defined as the capacity of a feature to predict another feature. Therefore redundant features would be vastly correlated [47]. The next step dealt with the building of models.

3.5 Model Building

Models are built by training classifiers before they can be used to perform prediction on new data. Predictive model building experiments were conducted on both imbalanced data and balanced data using meta-classifiers. Meta-classifiers enabled the discretization of numeric attributes before model building. Discretization converts continuous values into nominal values. This method has proven to improve accuracy and reduce algorithm learning time. In this study it was necessary to discretize data since some learning algorithms such as C4.5 do not support multi-way split on numeric attributes. The discretized dataset was applied to three different decision trees learning algorithms, C4.5 (J48 in WEKA), REPTree and Random tree. The models were built on 80% of the total instances and 20 % was held out for testing in order to estimate the performance of the models. The parameter settings for all the three algorithms are presented as follows;

3.5.1 Model Building Using J48 Algorithm

The construction of a J48 classifier in predicting the birth weight group was done using default parameter settings in the Weka explorer as exhibited in *Table 3*.

Table 3 default parameter settings for building a J48 model

Parameter	Description	Default Value
-M	The minimum number of instances per leaf	2
-C	ConfidenceFactor used in post pruning	0.25

A standard pruning confidence factor of 0.25 was used to prune branches where the algorithm was 25% or more confident so as to avoid having specific branches that would not be representative for generalisation. Pruning avoids over-fitting of the data [74].

3.5.2 Model Building Using REPTree Algorithm

The other classification technique applied in this study is the REPTree algorithm. The algorithm also utilized the the default values as shown in *Table 4*.

Table 4 default parameter settings for building a REPTree model

Parameter	Description	Default Value
-M	The minimum number of instances per leaf	2
-V	The minimum numeric class variance proportion of train variance for split.	0.001
-N	The number of folds for reduced error pruning.	3
-S	The Seed for random data shuffling.	1
-L	The maximum depth of a tree.	-1

3.5.3 Model Building Using Random Tree Algorithm

The random tree algorithm was also applied for building the classification model based on the default parameter settings and they are presented in *Table 5*.

Table 5 default parameter settings for building a Random Tree model

Parameter	Description	Default Value
-K	The number of attributes to randomly investigate.	0
-M	The minimum number of instances per leaf	1
-V	The minimum numeric class variance proportion of train variance for split.	0.001
-S	The seed for random number generator.	1

3.6 Model Evaluation

Once the model is trained, it is important to test its performance. Experimental results based on the test set were evaluated in terms of the details provided by the ROC curve and confusion matrix. Sensitivity, Specificity, and Accuracy were derived from the confusion

matrix. A further assessment of the model performance when run through unknown data was conducted based on a separate synthetic dataset. The construction of the dataset is detailed in the next section.

3.7 Predicting Birth Weight Based On an Independent Synthetic Dataset

The challenges of acquiring medical data are acute due to privacy related issues. Medical practitioners are generally reluctant to share patient data with researchers for fear of possible lawsuits that may arise, should anomalies be detected in patient medical histories [54], [75]. Therefore, this study opted to develop a synthetic dataset for model testing and classification.

An independent synthetic dataset with 10, 000 cases and the best selected attributes was randomly generated using the data analysis tool in Microsoft Excel. The distribution for discrete variables was a normal distribution. The cases were generated based on mean and standard deviation of the discrete variables. For categorical variables, a uniform distribution was used. Each value had an equal chance of being selected. A random seed of 20 was used for ease of retrieval of the same set of distribution. The synthetic dataset was loaded in WEKA for prediction, predictions were done based on the selected decision tree model. The results were saved in comma separated value (CSV) format and an auxiliary performance analysis of the model was performed. This task was carried out to assess how the model performs when dealing with unseen data.

3.8 Methodology Summary

This chapter has provided an outline of how data was collected and how experimentations were conducted. The original dataset as well as the modified datasets using SMOTE and Correlation-based Feature Selection algorithms were assessed for performance and lastly a synthetic dataset was produced and utilized for further performance evaluation of the selected model when applied to new patient cases.

Chapter 4

4 Presentation of Results and Discussion

The main goal of the study was to predict birth weight in order to identify mothers at the risk of giving birth to LBW babies. Therefore, data-mining classification methods were selected to develop predictive models. Predictive model building is an iterative procedure and therefore it is crucial to perform multiple experiments with different classifiers in order to select the best model for solving the problem at hand [24]. The experiments were conducted by using a meta-classifier in conjunction with J48, REPTree and Random tree and discretization as mentioned in the previous chapter. The performance of the algorithms were evaluated on both the unbalanced and balanced datasets.

4.1 Experimental Design

Various experimentations were carried out for each classifier and percentage split validation was used to estimate the performance of each classifier. Out of all the instances, 80% was utilized for training and the remaining 20% made up the test set. Overall classification accuracy, sensitivity, specificity and AUC were used to evaluate and compare the performance of the models. The AUC serves as an indicator of the overall performance of the algorithm [76]. Therefore models with the highest AUC were deemed as the best.

The experiments were generally carried out in three phases;

1. The first phase was performed on the *actual imbalanced data*.
2. The second phase made use of instances generated by *SMOTE* algorithm (balanced data)
3. The third experiment utilized the *best selected features*.

The above scenarios were considered to answer the following research questions;

1. Can decision trees built from a dataset with artificially increased cases of the minority class possibly yield a higher AUC than the ones built from raw imbalanced data?
2. What attributes are significant for prediction of birth weight values?
3. What is the effect of the CFS algorithm on the AUC of the decision tree models?
4. Which decision tree algorithm predicts birth weight values with a higher AUC?
5. How does the selected model perform when applied to new unseen patient cases?

After setting all the necessary parameters for the learning algorithms, the three phases of the experiments described above were conducted. A thorough description and outcome of the experiments is provided below;

4.1.1 Experimentation One (Model Building with Imbalanced Data)

This experimentation made use of the original dataset (imbalanced data) with 10, 000 instances and 90 features. The original dataset was analysed to determine whether decision trees built from a dataset with artificially increased cases of the minority class could yield a higher AUC than the ones built from raw imbalanced data. Models were built through training all the three classifiers based on 8, 000 instances. Performance evaluation was done on the remaining 2, 000 instances. A summary of the obtained output from the three different experiments based on the test data is represented in Table 6.

Table 6 Performance output of models when applied to an imbalanced dataset

Performance measure	J48 algorithm	REPTree	Random Tree
<i>Correctly classified</i>	735	783	571
<i>Incorrectly classified</i>	1265	1217	1429
<i>Accuracy (%)</i>	36.8	39.2	28.6
<i>Sensitivity/Recall (%)</i>	36.8	39.2	28.6
<i>FP rate</i>	0.251	0.266	0.253
<i>Specificity (1-FP) * 100</i>	74.9	73.4	74.7
<i>AUC (%)</i>	59.9	62.9	51.8
<i>Mean absolute error</i>	0.136	0.139	0.1434

As illustrated in Table 6 the REPTree model has the most correctly classified instances and the least incorrectly classified instances. Thus, REPTree has the highest accuracy of 39.2%, followed by J48 algorithm with an overall accuracy of 36.8%. Random tree gives the worst accuracy of 28.6%. The overall accuracy for this experiment is poor, however, a close investigation into other performance measures such as recall/sensitivity, specificity and AUC can produce results of interest to medical practitioners.

An evaluation based on the sensitivity highlights that the models are almost incapable of classifying the positive instances as positive, when applied to the imbalanced dataset. REPTree gives the highest sensitivity of 39.2%, J48 comes after REPTree with 36.8% and Random tree is the least sensitive to positive instances with a sensitivity rate of 28.6%. Interestingly the sensitivity rates are observed to be similar to the accuracy rates of the models. In contrast, the specificity rates for all the models are relatively higher than their sensitivity rates. Thus, the models can better recognize the non-LBW classes compared to the LBW classes. The specificity for the models varies with a relatively small margin. J48 has the best specificity of 74.9%, followed Random tree with 74.7%, whereas REPTree holds 73.4%.

For this experiment, the REPTree model predicts birth weight with a comparatively higher AUC of 62.9%, followed by J48 model with 59.9%. The lowest AUC is attained by Random tree (51.8%). REPTree also gives the highest sensitivity of 39.2%, J48 comes after REPTree with 36.8%, whereas Random tree has the worst sensitivity (28.6%). Thus, the models' ability to distinguish all the classes was inadequate.

To sum up, the results show that REPTree produced the best results. However it is observed that all models are better at predicting the non-LBW classes comparative to the LBW classes. Thus, our investigation reveals that class imbalance degrades performance of decision trees in predicting the birth weight values. Therefore, the results of this experimentation would not be applicable in predicting birth weight in clinical settings, thereby heightening the demand for further experimentations. This experimentation was compared to the next experimentation that was performed with the SMOTE algorithm.

4.1.2 Experimentation Two (Model Building with SMOTE)

This experiment made use of 28, 559 instances generated by SMOTE algorithm. Out of these instances 80% (22, 848) were used for training and the remaining 20 % (5, 711) for performance evaluation. In this section we present the results of the algorithms when implemented on data that was modified with SMOTE algorithm, and with all the 90 features.

Table 7 performance output of models built with SMOTE

Performance measure	J48 algorithm	REPTree	Random Tree
<i>Correctly classified</i>	3788	3524	3427
<i>Incorrectly classified</i>	1924	2188	2285
<i>Accuracy (%)</i>	66.3	61.7	60
<i>Sensitivity/Recall (%)</i>	66.3	61.7	60
<i>FP rate</i>	0.044	0.050	0.051
<i>Specificity (1-FP) * 100</i>	95.6	95.0	94.9
<i>AUC (%)</i>	89.0	88.3	82.7
<i>Mean absolute error</i>	0.0752	0.0871	0.0824

Comparative to experimentation one, **Table 7** depicts an increase in accuracy rates of models when optimized with SMOTE. J48 model has the highest accuracy of 66.3%; thus, an increment of 29.5% is realized. REPTree has the next highest accuracy of 61.7%. Unlike in experimentation one, REPTree's accuracy rate is slightly lower than that of J48. Similar to experimentation one, Random tree yields the lowest accuracy rate (60%), however it is almost at par with the other two models.

A substantial increment in the sensitivity of all models is also observed in this experimentation. Thus, the ability of the models to classify the LBW classes is strengthened. This could be more interesting to medical staff, since the aim of the models is to correctly classify LBW classes. The J48 model scores the highest sensitivity of 66.3%, followed by REPTree with 61.7% and lastly Random tree with 60%. These results are fairly good, however, further improvements are necessary.

The specificity rates are still observed to be higher than the sensitivity rates, however the margin between the two performance measures is significantly lower compared to experimentation one. Based on this observation, it can be inferred that the models' recognition of the LBW classes as opposed to the non-LBW classes is intensified. The best specificity is obtained by J48 with a rate of 95.6%. Unlike in experimentation one, where Random tree came second, the REPTree model outperformed the Random tree in terms of recognizing the non-LBW classes as non-LBW by only 0.1%. Thus REPTree had a specificity of 95.0% and Random tree came last with 94.9%.

Similarly, we realized a significant increase in the AUC of all the three models. J48 had 59.9% in experimentation one and relative to experimentation two, the AUC for J48 is 89.0%. This is also the highest AUC obtained in this experimentation. REPTree had the highest AUC of 62.9%, however in experimentation two comes second with 88.3%. Likewise, Random tree realises an increase from 51.8% to 82.7%. This is the least obtained AUC in experimentation two. Another performance measure under study is the sensitivity and it is substantially enhanced in all the models.

In conclusion a model with the highest AUC is deemed better than a model with the lowest AUC. Thus the results suggest that J48's output would be more of value to birth weight prediction. In addition, following the above discussion of results, it is evident that birth weight prediction models optimized by artificially increasing the minority class instances using SMOTE yield significantly better results relative to models built from actual raw imbalanced data. The optimized birth weight prediction models improved the accuracy, sensitivity, specificity and the AUC.

The next experimentation was done based on the best selected features after SMOTE.

4.1.3 Experimentation Three (Model Building with the Best Selected Attributes (CFS)).

The third experimentation was conducted using reduced attributes. Prior to the experimentation, attributes selection was necessary. Correlation based feature selection

algorithm together with best search strategy were applied to determine the most significant attributes for birth weight prediction. In addition, the dataset with reduced features was used to test the effect the CFS algorithm on the AUC of the decision tree models. The rankings of the most 16 predictive attributes out of the 90 supplied attributes are summarized in *Figure 9*.

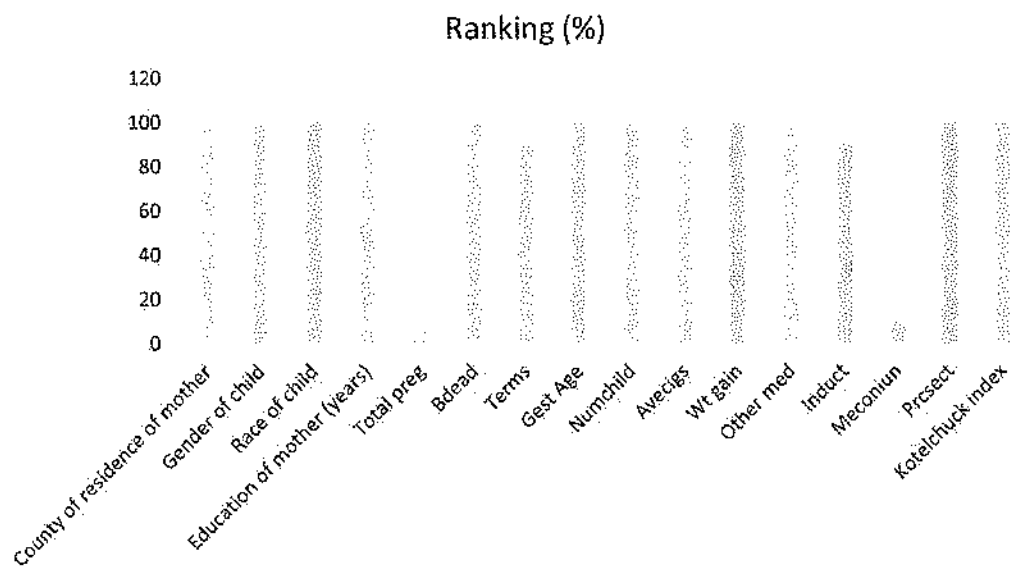


Figure 9 Rankings by CFS algorithm and best search strategy

Figure 9 indicates that *Total preg* (number of total pregnancies) and *Meconium* (Meconium, moderate/heavy) attributes are least correlated to the birth weight group with a ranking of 10% each and thus were manually eliminated.

County of residence of mother, *Gender of child*, *Race of child*, *Education of mother (years)*, *Bdead* (the number of children born alive now Dead), *Terms* (number of other terminations), *Gest Age* (completed weeks of gestation (calculated)), *Numchild* (number of living children), *Avecigs* (average number of cigarettes used daily), *Wt gain* (weight gained), *Other med* (other medical conditions), *induct* (Induction of labour), *Presect* (repeat caesarean section), and *Kotelchuck index* are all highly correlated to the birth weight group with a score of 100%, and thus 14 attributes were retained and used in further experimentations.

It can be inferred that the experimentation has confirmed some of the determinants of birth weight mentioned in the literature such as *gender of the baby*, *ethnic origin*, *gestational weight gain*, *gestational age*, and *intake of toxic substances* such as cigarettes [10].

The performance of models was assessed using reduced features and the results are presented below;

Table 8 performance output of models when built with the best selected features

Performance measure	J48 algorithm	REPTree	Random Tree
<i>Correctly classified</i>	3789	3527	3712
<i>Incorrectly classified</i>	1923	2185	2000
<i>Accuracy (%)</i>	66.3	61.7	65.0
<i>Sensitivity/Recall (%)</i>	66.3	61.7	65.0
<i>FP rate</i>	0.046	0.051	0.045
<i>Specificity (1-FP) * 100</i>	95.4	94.9	95.5
<i>AUC (%)</i>	90.3	88.4	85.3
<i>Mean absolute error</i>	0.0786	0.0872	0.0724

Table 8 shows that the number of correctly classified instances increased in contrast to experiment two. The number of correctly classified instances for J48 increased by 1, REPTree by 3 and Random tree realized the highest increment of 285 instances. However, J48 has the highest number of correctly classified instance. The accuracy for both J48 and REPTree is constant. Random tree earned an improvement of 5% in terms of accuracy, nevertheless, it maintained the lowest accuracy rate and J48 achieved the highest accuracy rate.

Additional comparison based on the sensitivity of models, reveals that the sensitivity is constant for both J48 and REPTree, while Random tree has an improvement of 5 %. J48. Thus feature selection has minimal impact on the sensitivity of birth weight prediction models. A further assessment shows a slight decline in the specificity rates of J48 and REPTree, 0.2% for J48 and 0.1% for REPTree. The specificity for Random tree is enhanced by 0.6% and thus Random tree outperforms J48 and REPTree in terms of in categorizing the non-LBW classes as indeed non-LBW. However, it can be argued that medical practitioners are often interested in classifying the positive class (LBW) as positive. Thus, Random tree cannot be chosen as the best model based on its sensitivity.

In terms of the AUC all the models have an improvement. J48 improved by 1.3%, REPTree 0.1% and Random tree 2.6%.

Given the observed trends in the results, the increasing values of AUC confirm the positive impact of feature selection through CFS algorithm, on birth weight prediction models. However it is highlighted that CFS slightly increases the mean absolute error for J48 and REPTree models and there was a slight decline in the mean absolute error for the Random

tree model. After feature selection, a comparison of all the experiments was done to determine the best model among others.

4.1.4 Selection of the best model

The goal of this section was to select a decision tree model with the highest AUC, for future prediction of birth weight values. To fulfil this purpose, the above experimental results were compared based on the AUC. This answered the following research question; which decision tree algorithm predicts birth weight values with a higher AUC. The AUC for all the experiments are shown in *Table 9*.

Table 9 A comparison of experimental results based on the AUC of models

Performance measure	J48 algorithm	REPTree	Random Tree
<i>Experimentation one (model building with imbalanced data)</i>	59.9	62.9	51.8
<i>Experimentation two (model building with SMOTE)</i>	89.0	88.3	82.7
<i>Experimentation three (model building with the best selected attributes)</i>	90.3	88.4	85.3

As indicated in *Table 9*, in experimentation one, the AUC of the REPTree model is better than others and therefore REPTree is the best model to deal with imbalanced data. However, further experimentation reveal that J48 is generally better at predicting birth weight, with the highest AUC in both experimentation two and three.

By implementing the SMOTE on the original imbalanced dataset, the AUC is significantly improved for all the models. The utilization of feature selection through CFS further enhances the AUC of the models.

Therefore we conclude that the dataset that produces the best AUC is the one with balanced data and reduced features (*Experimentation three*). This combination produces AUC rates of 90.3% (J48), 88.4% (REPTree) and 85.3%. Looking at this experimentation, **J48** attained the highest *AUC (90.3%)* and therefore was selected as the best model for future prediction of birth weight values.

4.1.5 Predicting Birth Weight Based On an Independent Synthetic Dataset

In order to assess how the selected model performs when applied to new unseen patient cases. The J48 model was used to predict birth weight based on an independent data set was. This task was performed to assess how the selected model performs when applied to new

unseen patient cases. The dataset had 10, 000 instances and the best selected features from the feature selection experiment. The distribution per class is represented in *Figure 10*.

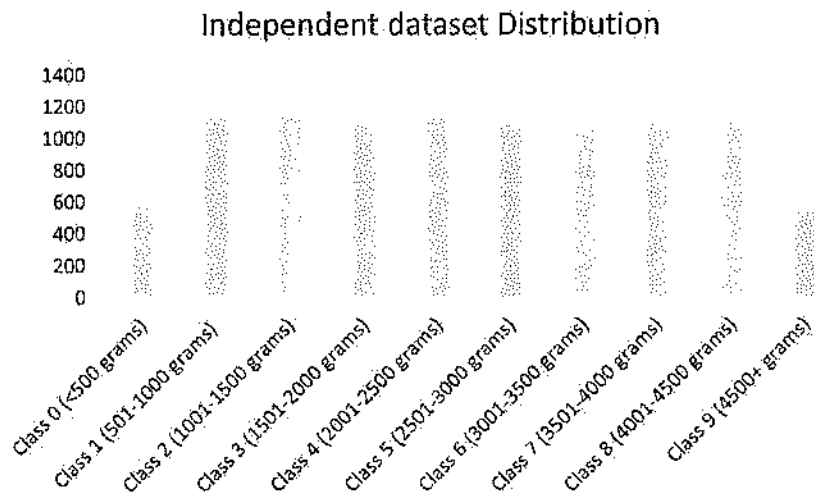


Figure 10 The independent dataset distribution per class

The distribution for the classes was almost at par except for classes 0 and 1 which had 597 and 536 instances respectively, and therefore is reasonable to conclude that the data is imbalanced. As mentioned in the literature review (Imbalanced Data) section, data imbalance degrades performance. [24], [48] Performance assessment was on the basis of the, AUC, sensitivity and specificity as constructed from the confusion matrix. The accuracy and mean absolute error are also mentioned. First we present a summary output of the model and then explain the confusion matrix.

Table 10 Summary output of the J48 model when applied to new unseen data

Correctly classified	5, 433
Incorrectly classified	4, 567
Accuracy (%)	54.3
Sensitivity/Recall (%)	54.3
FP rate	0.056
Specificity (1-FP) * 100	94.4
AUC (%)	84.6
Mean absolute error	0.1316

As illustrated in *Table 10*, the overall predictive accuracy of the model on an independent dataset was 54.3%. Although the model performance regressed in terms of the accuracy and

AUC, it maintained a substantial AUC of 84.6%. A specificity of 94.4% and sensitivity of 54.3% were scored by the model.

The results of this model can be used in real clinical settings for mapping new patient data into multiple birth weight groups, and thereby assisting in early recognition of mothers at the risk of low birth weight infants. In addition, there is a likelihood of improvement in interventions to improve care during pregnancy, childbirth and the postnatal stage.

A detailed performance assessment by class as represented in the Confusion Matrix is discussed in the next section.

The Confusion Matrix in *Figure 11* represents the number of instances per class that are designated to all possible classes as per the classifier’s predictions. The columns symbolise the predictions, and the rows the actual class.

```

=== Confusion Matrix ===

```

	a	b	c	d	e	f	g	h	i	j	<-- classified as
273	324	0	0	0	0	0	0	0	0	0	a = 0
0	613	520	0	0	0	0	0	0	0	0	b = 1
0	0	347	802	0	0	0	0	0	0	0	c = 2
0	0	0	799	294	0	0	0	0	0	0	d = 3
0	0	0	0	1114	0	17	0	0	0	0	e = 4
0	0	0	0	0	0	672	416	0	0	0	f = 5
0	0	0	0	0	0	0	1015	0	39	0	g = 6
0	0	0	0	0	0	0	507	600	0	0	h = 7
0	0	0	0	0	0	82	322	708	0	0	i = 8
0	0	0	0	0	0	120	54	276	86	0	j = 9

Figure 11 Confusion matrix for J48 model when applied to new unseen data

It can be observed from the above confusion matrix that out of 10, 000 instances 5, 433 (sum of the values in the diagonal column) were correctly classified and 4, 567 were misclassified, thus the overall accuracy rate obtained by J48 when applied to new data was 54.3% $(5433/10\ 000) \times 100$. The accuracy declined by 12% compared to the one achieved in Experiment three. A justification for the declined could be related to overfitting, thus the model was too specific to training data. Up next we discuss the sensitivity, specificity and AUC obtained by each class.

Class 0 (<500 grams)

The Confusion Matrix shows that 273 out of 597 infants that were less than 500 grams were correctly predicted as less weighing 500 grams. Thus, the sensitivity for Class 0 was $(273/597) \times 100 = 45.7\%$. This means only 45.7% of the instances that belonged to Class 0 were indeed classified as belonging to Class 0. On the other hand 324 infants that weighed 500 grams were incorrectly classified as weighing to 501-1000 grams. Predicting infants that are likely to belong to Class 0 would be of paramount importance to medical staff,

however, the model performed poorly when classifying Class 0 as positive. On the contrary this misclassification would be less costly since both the weight groups 500 grams and 501-100 grams are low birth weight.

The specificity measures the true negative rate, therefore the sensitivity for class 0 was $(9403/9403+0) * 100 = 100\%$. Thus all the values that did not belong to class 0 were correctly classified as not belonging to class 0. The AUC was also considered and it is shown in *figure 12*.



Figure 12 ROC Curve for class 0

Figure 12 reveals that the model was exceptionally good in terms of the AUC. It was capable of distinguishing between the True Positives and False Positives with a score of 91.1%. Looking at the AUC, the model can be applicable used for predicting birth weight values of less than 500 grams.

Class 1 (501-1000 grams)

When the J48 model was used to predict the group for infants that were 501-1000 grams, a fair sensitivity of 54.1 % was produced. Thus out of all the 1, 133 infants that weighed 501-1000 grams, the model was able to recognize 613 of these infants as indeed weighing 501-1000 grams. The sensitivity was calculated as follows, $(613/1133)*100= 54.1\%$. On the contrary 520 instances were incorrectly predicted as class 2. The sensitivity signifies that the model was slightly above average in correctly classifying class 1, and thus it can be used for future predictions, though further improvements would be necessary.

The specificity/true negative rate for Class 1 was $8543 / (8543+324) * 100=96.3\%$. Thus the model was able to correctly classify 96.3% of infants as not belonging weighing 501-1000 grams. *Figure 13* depicts the AUC for class 1.

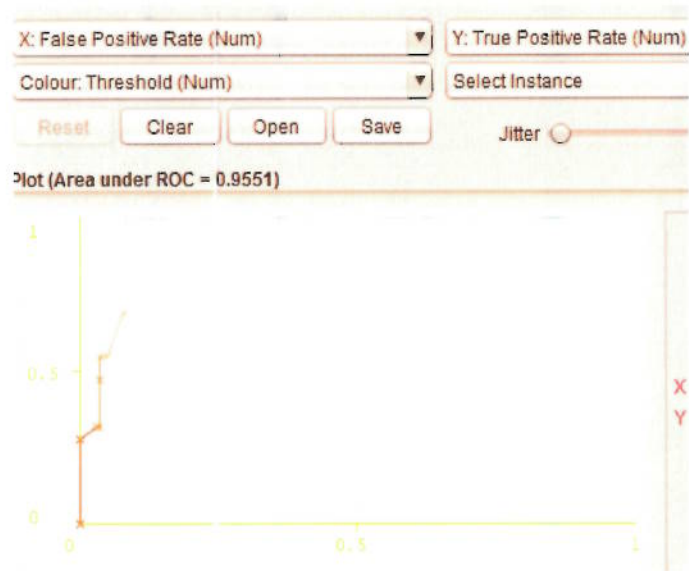


Figure 13 ROC curve for class 1

The AUC was 95.5%, this value denotes that the model performed tremendously in differentiating between the True Positive and False Positive instances. Therefore it would be viable to use the model in the application domain.

Class 2 (1001-1500 grams)

The confusion matrix shows that for Class 2 (1001-1500g), the sensitivity was $(347/1149) * 100 = 30.2\%$. Thus a large fraction of instances weighing 1001-1500 grams, were incorrectly predicted. This is supported by the number of the incorrectly classified instances which was 802 out 1, 149. These instances were instead classified as weighing 1501-2000 grams. J48 model produced a specificity of $94.1\% = 8331 / (8331+520)*100$.

There was a discrepancy between the sensitivity and specificity rates, thus the model was biased towards identifying 1501-2000 grams birth weight group rather than 1001-1500 grams. The last performance measure was the AUC (*Figure 14*).



Figure 14 ROC curve for class 2

Figure 14 shows that the model scored an AUC of 89.7% when classifying instances that were 1001-1500 grams. Hence it can be concluded that the model was not random and can be utilized for future predictions.

Class 3 (1501-2000 grams)

The Confusion Matrix demonstrates that the sensitivity for Class 3 was $(799/1093) * 100 = 73.1\%$. It was observed that the sensitivity for Class 3 had an improvement comparative to Classes 0, 1 and 2. Out of the 1, 093 instances 294 were misclassified as Class 4. These values highlight the robustness and possibility of further usage of the model in correctly predicting Class 3.

The instances that did not belong to Class 3 (true negatives), were correctly classified with a specificity of $8105 / (8105+802) * 100 = 91.0\%$. The specificity was outstanding, however, the sensitivity was more desirable, as it is better to correctly classify a high risk event, than to incorrectly classify a low risk event. The AUC for Class 3 is displayed in **Figure 15**.

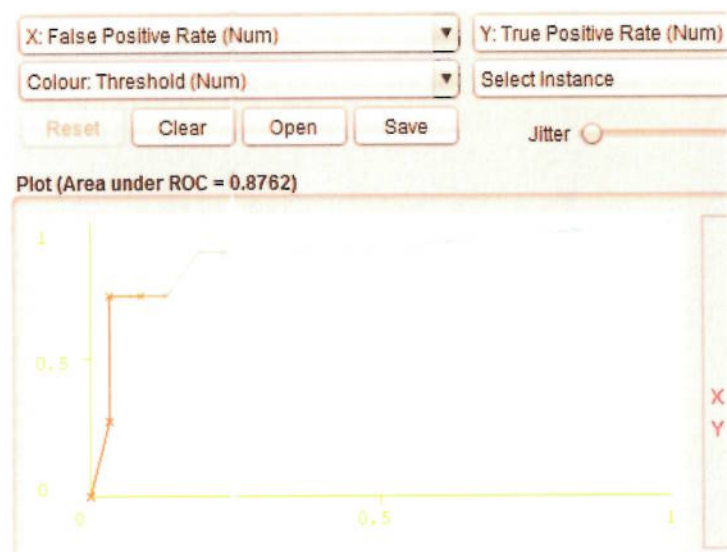


Figure 15 ROC curve for class 3

The AUC for predicting Class 3 was 87.6%. This depicts that the model would be beneficial to medical staff for predicting infants weighing 1501-2000 grams.

Class 4 (2001-2500 grams)

From the confusion matrix, we discover that the sensitivity for Class 4 is $(1114/1131) \times 100 = 98.5\%$. The specificity for this class was $8575 / (8575 + 294) \times 100 = 96.7\%$. The model was able to clearly differentiate instances that belonged to Class 4 and the ones that belonged to other classes.

Only 17 instances out of 1, 131 were incorrectly classified as Class 6 (3000-3500). This implies 17 infants who were 2000-2500 grams (low birth weight), were incorrectly classified as normal. The misclassification would be costly as the aim is to correctly classify low birth weight as positive. However, these results can be applied in practical settings, since the misclassification rate is minute (0.015%). The AUC earned was 97.1 % as exhibited in *Figure 16*.

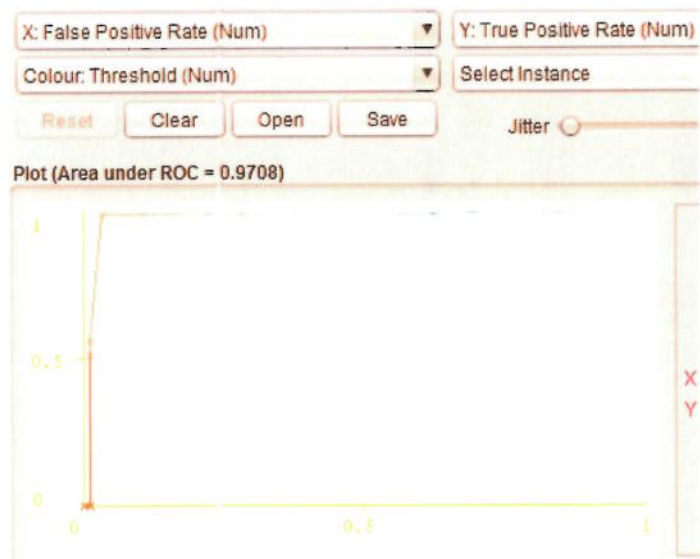


Figure 16 ROC curve for class 4

When predicting Class 4, the model had the capability to clearly separate instances belonging to class 4 from those that did not belong to class 4, with a score of AUC 87.6%. Thus, the model would be significant in predicting future instances belonging to Class 4.

Class 5 (2501-3000 grams)

The number of correctly classified instances for Class 5 out of a total of 1, 088, was 672 and thus the sensitivity was $(672/1088) * 100 = 61.8\%$. The model incorrectly classified 416 instances as Class 6, however, both classes were normal births and therefore the misclassification would be to a lesser extent costly. The specificity was found to be 97.7%, calculated as $8710 / (8710+202)*100$. The AUC is showcased in **Figure 17**.

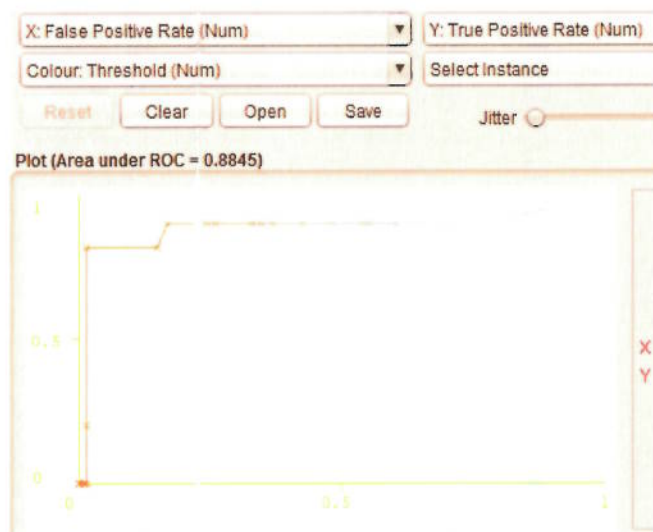


Figure 17 ROC curve for class 5

An AUC of 87.6% was yielded by the model when predicting Class 5, thus the model had the ability to distinguish Class 5 from other classes. The results show that the model would be suitable for forecasting future birth weight values.

Class 6 (3001-3500 grams)

The Confusion Matrix also indicates that the sensitivity for Class 6 was 96.3% $((1015/1054) * 100)$. The specificity on the other hand was 85.3%, $(7630 / (7630+1316) * 100)$. Out of the total instances that belonged to Class 6, 39 instances were misclassified as class 8. As indicated in **Figure 18**, the AUC was 88.3%.



Figure 18 ROC curve for class 6

Figure 18 represents that the model would be desirable for future predictions in clinical operations, as the AUC was significantly high.

Class 7 (3501-4000 grams)

The model was able to correctly classify 600 out of 1107 instances as Class 7. Consequently, the sensitivity for was 54.2%., whereas 507 instances were incorrectly classified as class 6. The attained specificity was 89.0% $(7989 / (7989+984)* 100)$. The obtained AUC is revealed in **Figure 19**.



Figure 19 ROC curve for class 7

A great AUC score of 86.2% was obtained by the model when classifying instances belonging to 3501-4000 grams. Therefore the model would be feasible in categorizing future birth weight values.

Class 8 (4001-4500 grams)

It is depicted in the Confusion Matrix that Class 8 had a total of 1,112 instances. However all of them were incorrectly classified, thus the sensitivity for Class 8 was 0%. The model was not able to recognize instances that belonged to Class 8 as belonging to Class 8.

A total of 82 instances were incorrectly classified as Class 5 (2500-3000 grams), 322 as Class 6 (3001-3500 grams) and 708 as Class 7 (3501-4000 grams). Although there was a misclassification, the model was still able to successfully classify these instances as normal births (not low birth weight).

The specificity for this class was discovered to be 98.6% $(8763 / (8763 + 125)) * 100$. The AUC of the model can be observed from **Figure 20**.

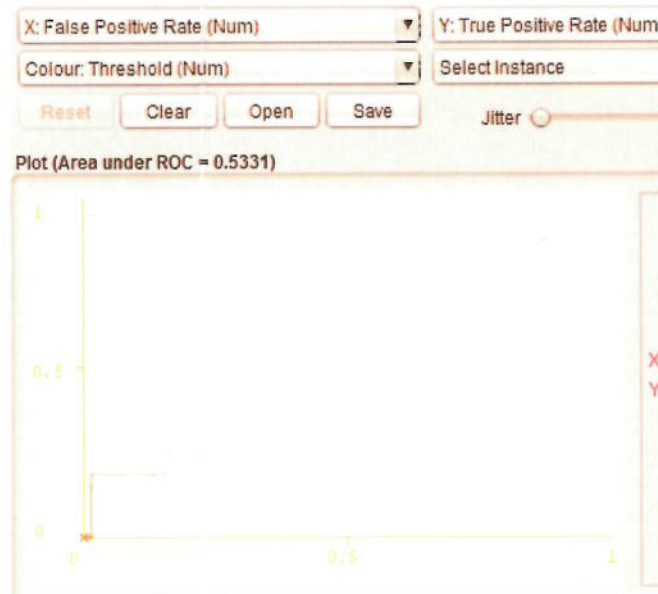


Figure 20 ROC curve for class 8

From *Figure 20*, we conclude that the AUC was 53.3.0%. The model was almost random as the AUC was, slightly greater than 50% with 3.3%. Therefore the model can be slightly trusted when predicting 4001-4500 grams. The poor performance can be attributed to the fact that in normal clinical settings, a birth weight group of 4001-4500 grams is rare.

Class 9 (4500+ grams)

It can be deduced from the Confusion Matrix that the sensitivity for Class 9 was 0%. None of the instances that belonged to Class 9 were correctly predicted as Class 9. Out of 536 instances, 120 were incorrectly classified as Class 5 (2501-3000 grams), 54 as Class 6, 276 as Class 7 (3501-4000 grams), and 86 as Class 8 (4001-4500 grams). Similar to Class 9, the model was able to identify the instances as normal, despite the misclassification.

The specificity for class 9 was 100% (9464/9464+0). Thus the model was utterly incapable of recognizing class 9. Similar to class 8, the poor results can be imputed to the fact that the weight group of 4500+ grams is rare. An AUC for this class is illustrated in *Figure 21*.

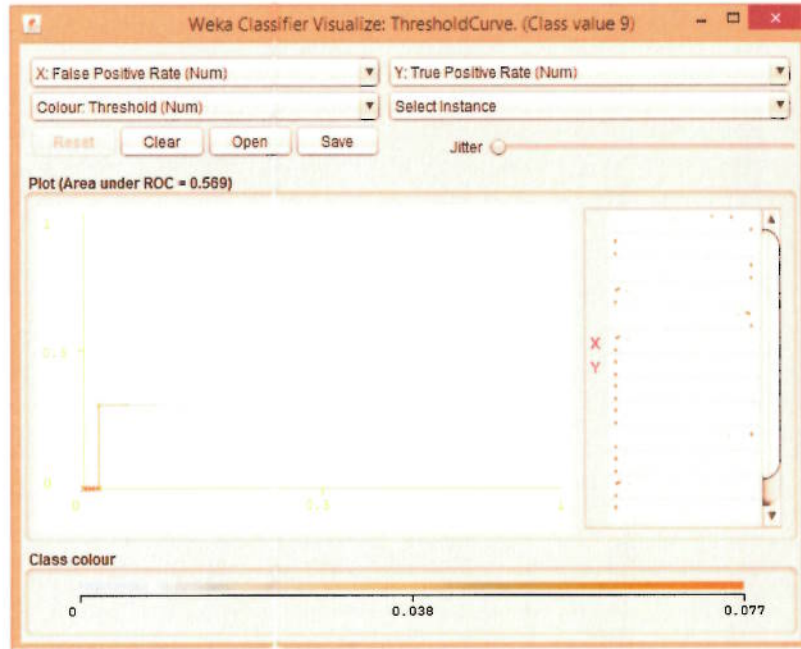


Figure 21 ROC curve for class 9

The obtained AUC for class 9 was 56.9%, this was slightly higher than the AUC for class 8. However, the model was above the threshold (50%), by only 6.9%, thus it can be deduced that the model was almost random and further improvement would be necessary before it can be utilized to predict future instances weighing 4 500+ grams.

4.2 Results and Discussion Summary

In summary, several decision tree models (J48, REPTree and Random Tree) that predict birth weight groups were built. Three broad experiments were performed based on the original imbalanced data, with balanced data through SMOTE and with 14 selected 8 attributes of the dataset.

The experiments were designed to assess the effects of balanced and imbalanced data, and feature selection on the AUC of birth weight prediction models. The models were evaluated based on percentage split (20% test set). A comparison of the AUC aided the selection of the best model for future birth weight predictions.

In Experimentation one (with imbalanced data), all the models performed poorly with REPTree attaining the highest AUC of 62.9%, followed by J48 (59.9%) and lastly Random Tree (51.8%). The implementation SMOTE for balancing data (Experimentation two) led to an enhancement in the AUC rates of all the models. The J48 model performed the best with an AUC of 89.0%, followed by REPTree (88.3%) and Random Tree (85.3%). This

indicates that models built with balanced data achieve better results comparative to models built with imbalanced data.

Finally in Experimentation three (with 14 selected attributes), J48 performed relatively better in predicting birth weight with the highest AUC of .90.3%, followed by REPTree (88.4%), and lastly the Random Tree (85.3%). This was a further improvement from the last experiment. Experimentation three was the best and J48 model from this experiment was selected as the best model.

A further assessment of the J48 model was carried out on an independent synthetic dataset and the model performed tremendously well with an AUC of 84.6%, Sensitivity of 54.3%, Specificity of 94.4% and 54.3% Accuracy rate. The AUC indicates that the model would be able to correctly classify birth weight with a degree of 84.6%.

Comparative to related work in the literature, the SMOTE optimized J48 model achieved a slightly lower that the model by Marshall et al [65]. The J48 model produced an AUC of 84.6% when applied to unseen data, while their model predicted hospital mortality with 85% AUC.

In terms of the sensitivity, a model by Yadav and Lee [14] yielded a sensitivity of 80%, whereas the J48 model produced a relatively lower sensitivity of 54.3.3%. The poor sensitivity rate can be partially imputed to varying approaches. The J48 model predicts the actual birth weight values, which is novel. On the other hand, the study by Yadav and Lee [14], exhaustively lists down LBW risk factors. In addition, the two models employed different datasets and therefore it would be rational to produce different results.

The results of the model were also comparable to Naïve Bayes', neural network, random forest, classification tree, SVM, and logistic regression models in work by Senthilkumar and Paulraj [15]. The SMOTE optimized J48 model had a lower accuracy than all the models by a margin of 4.6% -23.7%. The rationale for poor performance can be attributed to the use of varying evaluation methods. In their study, Senthilkumar and Paulraj [15], used cross validation method, while our study utilized the percentage split method of evaluation.

In conclusion, the obtained levels of AUC indicate that, the use of the SMOTE optimized J48 model in practical settings is highly feasible. Prediction of birth weight outcomes is of great importance for identifying mothers at a high risk of delivering LBW infants. Identifying these mothers can enable the medical staff to tailor special care to try and minimize complications related to LBW. The outcomes of the predictive models could also be employed in post-delivery consultations with mothers.

Chapter 5

5 Conclusions and Recommendations

5.1 Study Overview

Birth weight is one of the main factors determining neonatal and prenatal survival, therefore birth weight has been studied in the past years so as to determine low birth weight risk factors. The use of the identified risk factors for prediction of the actual birth weight values using data mining techniques has been neglected. Predicting the birth weight can assist in the early identification of expectant mothers at the risk of delivering low birth weight babies. Low birth weight is often grouped into low (<1 500 grams), low (<2,500 grams), or normal ($\leq 2,500$ grams) weight at birth. Low birth weight is associated with several obstetric complications such as infant mortality rate, respiratory infections, and hypertensive adults.

The aim of the study was to develop a decision tree model which predicts birth weight with a relatively higher AUC based on the identified low birth weight risk factors. The study made use of the North Carolina State Centre for Health Statistics (2006). In order to extract interesting trends from data and the steps followed were data selection, data pre-processing, model building, and model evaluation/interpretation.

Decision trees were used for classification purposes. The various decision trees under study predict the birth weight in to either of the following ten classes; 0 (500 or less), 1 (501-1000), 2 (1001-1500), 3 (1501-2000), 4 (2001-2500), 5 (2501-3000), 6 (3001 -3500), 7 (3501-4000), and 8 (4001-4500) and 9 (4501 or more). The experimentations were done on the actual imbalanced datasets, the balanced dataset using SMOTE algorithm, together with reduced feature using correlation based feature selection algorithm.

5.2 Conclusions

In this study, an approach based on decision tree modelling in order to predict the birth weight of baby was presented. The necessary steps required to achieve this goal were fully described and experiments with different decision trees were conducted and compared. An evaluation based on the hold out method and independent synthetic dataset was conducted. The results indicated that J48 model yielded the best results in terms of both the AUC and accuracy. As the medical field demands prediction with a higher AUC, it is suggested that the J48 algorithm from experiment three may be practically used in birth weight prediction. The study found that bridging the gap between the minority and majority classes in the training set using SMOTE can significantly improve the AUC of decision trees. Furthermore

models built used reduced features through feature selection also increases decision tree models' AUC. From the feature selection experimentation we observed that attributes such County of residence of mother, Gender of child, Race of child, Education of mother (years), Bdead (the number of children born alive now Dead), Terms (number of other terminations), Gest Age (completed weeks of gestation (calculated)), Numchild (number of living children), Avecigs (average number of cigarettes used daily), Wt gain (weight gained), Other med (other medical conditions), induct (Induction of labour), Presect (repeat caesarean section), and Kotelchuck index are the most determinant factors of birth weight. In general the results indicate our approach as a promising avenue of research.

5.3 Recommendations

The study illustrates the utility of predictive data-mining in making early clinical interventions. Although encouraging results were obtained, it was noted that further research efforts need to be conducted to fully utilize the potential of data-mining technology in improving the obstetric healthcare. In particular, the following areas were identified as deserving further research work;

It has been observed that the performance of classification algorithms differs with datasets. Therefore, additional data sets, particularly in the context of Botswana must be considered to assess the applicability of data mining in predicting birth weight based on other LBW risk factors. Furthermore, it is essential to conduct new experiments based on other data mining methods such as artificial neural networks and support vector machines since they have also been proved to be important techniques in the healthcare.

Since this study has used an independent synthetic dataset to assess performance of the selected model, it is suggested that more evaluation should be conducted using actual data to obtain more accurate results. In addition, medical data is constantly updated and models built from flat files often become obsolete since the models ignore new data. Therefore, implementation using real-time information to provide quality obstetric care should be considered. Real-time data-mining can be accomplished by utilizing big data frameworks such as Hadoop and MongoDB. The possibility of linking WEKA to these frameworks should also be tested.

Finally, the possibility of incorporating information visualization techniques to effectively explain the discovered complex patterns ought to be explored.

6 References

- [1] F. Allan, L. M. Ahumada, J. A. Galvez, and M. A. Rehman, "A review of analytics and clinical informatics in health care," *J Med Syst*, vol. 38, no. 45, 2014.
- [2] D. W. Bates, S. Saria, L. Ohno-machado, and A. Shah, "Big data in health care: Using analytics to identify and manage high-risk and high-cost patients," vol. 33, no. 7. pp. 1123–1131, 2014.
- [3] C. Bennett and T. W. Doub, "Data mining and electronic health records: Selecting optimal clinical treatments in practice," in *Proceedings of the 6th International Conference on Data Mining*, 2010, pp. 313–318.
- [4] M. Huang, M. Chen, and S. Lee, "Integrating data mining with case-based reasoning for chronic diseases prognosis and diagnosis," *Expert Syst. Appl.*, vol. 32, pp. 856–867, 2007.
- [5] W. Raghupathi and V. Raghupathi, "Big data analytics in healthcare: promise and potential," *Heal. Inf. Sci. Syst.*, vol. 2, no. 1, pp. 1–10, 2014.
- [6] H. Kaur and S. K. Wasan, "Empirical study on applications of data mining techniques in healthcare," *J. Comput. Sci.*, vol. 2, no. 2, pp. 194–200, 2006.
- [7] L. Li, S. Bagheri, H. Goote, A. Hasan, and G. Hazard, "Risk adjustment of patient expenditures: A big data analytics approach," in *2013 IEEE International Conference on Big Data*, 2013, pp. 12–14.
- [8] C. A Catley, "An integrated hybrid data mining system for preterm birth risk assessment based on a 'Semantic web services for healthcare' framework," 2007.
- [9] E. S. Hall and S. N. Thornton, "Extracting nursing practice patterns from structured labor and delivery data sets," in *AMIA 2007 Symposium Proceedings*, 2007, pp. 309–313.
- [10] M. S. Kramer, "Determinants of low birth weight: methodological assessment and meta-analysis," *Bull. World Health Organ.*, vol. 65, no. 5, pp. 663–737, 1987.
- [11] A. K. Blanc and T. Wardlaw, "Monitoring low birth weight: an evaluation of international estimates and an updated estimation procedure," *Bull. World Health Organ.*, vol. 83, no. 3, pp. 178–185, 2005.
- [12] United Nations Children's Fund and World Health Organization, "Low birth weight, country regional and global estimates," 2004.
- [13] N. E. Reichman, "Low birth weight and school readiness," *The Future of children*, vol. 15, no. 1. pp. 91–116, 2005.

- [14] H. Yadav and N. Lee, "Maternal factors in predicting low birth weight babies," *Med J Malaysia*, vol. 68, no. 1, pp. 44–47, 2013.
- [15] D. Senthilkumar and S. Paulraj, "Prediction of low birth weight infants and its risk factors using data mining techniques," in *Proceedings of the 2015 International Conference on Industrial Engineering and Operations Management*, 2015, pp. 186–194.
- [16] R. N. Das, R. S. Devi, and J. Kim, "Mothers' lifestyle characteristics impact on her neonates' low birth weight," *Int. J. Women's Heal. Reprod. Sci.*, vol. 2, no. 4, pp. 229–235, 2014.
- [17] M. P. Debbink and M. D. M. Bader, "Racial Residential Segregation and Low Birth Weight in Michigan's Metropolitan Areas," *Am. J. Public Health*, vol. 101, no. 9, pp. 1714–1720, 2011.
- [18] B. Desalegn, "Predicting low birth weight using data mining techniques on Ethiopia demographic and health survey data sets," Addis Ababa University, 2011.
- [19] M. P. Harris, "Predicting low birth weight through data mining." 2009.
- [20] A. S. Shittu, O. Kuti, E. O. Orji, N. O. Makinde, S. O. Ogunniyi, O. O. Ayoola, and S. S. Sule, "Clinical versus sonographic estimation of foetal weight in Southwest Nigeria," *J Heal. Popul NUTR*, vol. 25, no. 1, pp. 14–23, 2007.
- [21] M. R. Torloni, N. Sass, J. L. Sato, A. C. P. Renzi, M. Fukuyama, and P. R. de Lucca, "Clinical formulas, mother's opinion and ultrasound in predicting birth weight," *Sao Paulo Med J*, vol. 126, no. 3, pp. 145–149, 2008.
- [22] A. Ben-haroush, Y. Yogev, M. Hod, and J. Bar, "Predictive value of a single early fetal weight estimate in normal pregnancies," *Eur. J. Obstet. Gynecol. Reprod. Biol.*, vol. 130, pp. 187–192, 2007.
- [23] L. J. Salomon, J. P. Bernard, and Y. Ville, "Estimation of fetal weight: reference range at 20–36 weeks' gestation and comparison with actual birth-weight reference range," *Ultrasound Obs. Gynecol*, vol. 29, pp. 550–555, 2007.
- [24] M. Tefera, "Application of data mining to predict urinary fistula surgical repair outcome," Addis Ababa University, 2012.
- [25] S. K. Wasan, V. Bhatnagar, and H. Kaur, "The impact of data mining techniques on medical," *Data Sci. J.*, vol. 5, no. 1, pp. 119–126, 2006.
- [26] J. Soni, U. Ansari, D. Sharma, and S. Soni, "Predictive data mining for medical diagnosis: an overview of heart disease prediction," *Int. J. Comput. Appl.*, vol. 17, no. 8, pp. 43–48, 2011.

- [27] C. Catley, M. Frize, C. R. Walker, and D. C. Petriu, "Predicting high-risk preterm birth using artificial neural networks," *IEEE Trans. Inf. Technol. Biomed.*, vol. 10, no. 3, pp. 540–549, 2006.
- [28] T. G. Hailu, "Comparing data mining techniques in HIV testing prediction," *Intell. Inf. Manag.*, vol. 7, no. May, pp. 153–180, 2015.
- [29] M. Gagnon, J. Payne-gagnon, J. Fortin, G. Paré, J. Côté, and F. Courcy, "A learning organization in the service of knowledge management among nurses : A case study," *Int. J. Inf. Manage.*, vol. 35, pp. 636–642, 2015.
- [30] D. S. Jeyarani, G. Anushya, R. R. Rajeswari, and A. Pethalakshmi, "A Comparative study of decision tree and naive bayesian classifiers on medical datasets," *Int. J. Comput. Appl.*, pp. 5–7, 2013.
- [31] T. A. Hilage and R. V Kulkarni, "Review of literature on data mining," *IJRRAS*, vol. 10, no. 1, pp. 107–114, 2012.
- [32] S. Gupta, D. Kumar, and A. Sharma, "Data mining classification techniques applied for breast cancer diagnosis and prognosis," *Indian J. Comput. Sci. Eng.*, vol. 2, no. 2, pp. 188–195, 2011.
- [33] O. O. Adeyemo and T. O. Adeyeye, "Comparative study of ID3 / C4.5 decision tree and multilayer perceptron algorithms for the prediction of typhoid fever," *African Journa Comput. ICT*, vol. 8, no. 1, 2015.
- [34] R. Redpath, "A Comparative study of visualization techniques for data mining," Monash University, 2000.
- [35] U. Fayyad, "The KDD process for extracting useful knowledge from volumes of data," *Commun. ACM*, vol. 39, no. 11, pp. 27–34, 1996.
- [36] R. Bellazzi and B. Zupan, "Predictive data mining in clinical medicine : Current issues and guidelines," *Int. J. Med. Inform.*, vol. 77, pp. 81–97, 2008.
- [37] V. Podgorelec, P. Kokol, B. Stiglic, and I. Rozman, "Decision trees : an overview and their use in medicine," *J. Med. Syst.*, vol. 26, no. 5, pp. 445–463, 2002.
- [38] M. E. Yahia and M. E. El-taher, "A new approach for evaluation of data mining techniques," *Int. J. Comput. Sci. Inf. Issues*, vol. 7, no. 5, pp. 181–186, 2010.
- [39] B. D. Pitt, "Applications of data mining techniques to electric load profiling," University of Manchester Institute of Science, 2000.
- [40] X. Wu, V. Kumar, J. R. Quinlan, J. Ghosh, Q. Yang, H. Motoda, G. J. Mcalachlan, A. Ng, B. Liu, P. S. Yu, Z. Zhou, S. Michael, J. H. David J, and Dan Steinberg, "Top 10 algorithms in data mining," *Knowl Inf Syst*, vol. 14, pp. 1–37, 2008.

- [41] K. R. Lakshmi and S. P. Kumar, "Utilization of data mining techniques for prediction of diabetes disease survivability," *Int. J. Sci. Engineering Res.*, vol. 4, no. 6, pp. 933–942, 2013.
- [42] M. M. Mazid, A. B. M. S. Ali, and K. S. Tickle, *Improved C4.5 Algorithm for Rule Based Classification*.
- [43] K. Polat and S. Gunes, "A novel hybrid intelligent method based on C4.5 decision tree classifier and one-against-all approach for multi-class classification problems," *Expert Syst. Appl.*, vol. 36, pp. 1587–1592, 2009.
- [44] S. Ravichandran, V. B. Srinivasan, and C. Ramasamy, "Comparative study on decision tree techniques for mobile call detail record," *J. Commun. Comput.*, vol. 9, pp. 1331–1335, 2012.
- [45] A. Sana, S. Razzaq, and J. Ferzund, "Automated diagnosis and cause analysis of cesarean section using machine learning techniques," *Int. J. Mach. Learn. Comput.*, vol. 2, no. 5, pp. 677–680, 2012.
- [46] A. G. Eapen, "Application of data mining in medical applications," University of Waterloo, 2004.
- [47] O. Villacampa, "Feature selection and classification methods for decision making : A comparative analysis," Nova Southeastern University, 2015.
- [48] N. V Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE : synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, 2002.
- [49] M. A. Mazurowski, P. A. Habas, J. M. Zurada, J. Y. Lo, J. A. Baker, and G. D. Tourassi, "Training neural network classifiers for medical decision making: The effects of imbalanced datasets on classification performance \$," *Neural Networks*, vol. 21, pp. 427–436, 2008.
- [50] M. Khalilia, S. Chakraborty, and M. Popescu, "Predicting disease risks from highly imbalanced data using random forest," *BMC Med. Informatics Decis. Mak.*, vol. 11, no. 51, pp. 1–13, 2011.
- [51] L. M. Taft, R. S. Evans, C. R. Shyu, M. J. Egger, N. Chawla, J. A. Mitchell, S. N. Thornton, B. Bray, and M. Varner, "Countering imbalanced datasets to improve adverse drug event predictive models in labor and delivery," *J. Biomed. Inform.*, vol. 42, pp. 356–364, 2009.
- [52] K. J. Cios and G. W. Moore, "Uniqueness of medical data mining," *Artif. Intell. Med.*, vol. 26, pp. 1–24, 2002.
- [53] J. Huang and C. X. Ling, "Using AUC and accuracy in evaluating learning

- algorithms," *IEEE Trans. Knowl. Data Eng.*, vol. 17, no. 3, pp. 299–310, 2005.
- [54] T. Debela, "Developing a predictive model for fertility preference of women of reproductive age using data mining techniques," Addis Ababa University, 2013.
- [55] H. C. Koh and G. Tan, "Data Mining Applications in Healthcare," *J. Healthc. Inf. Manag.*, vol. 19, no. 2, pp. 64–72, 2005.
- [56] D. P. Shukla, S. B. Patel, and A. K. Sen, "A Literature review in health informatics using data mining techniques," *Int. J. Softw. Hardw. Res. Eng.*, vol. 2, no. 2, pp. 123–129, 2014.
- [57] A. Gebregiorgis, "Application of data mining to predict the likelihood of contraceptive method use among women aged 15-49," Addis Ababa University, 2012.
- [58] F. Soleimani, P. Mohammadi, and P. Hakimi, "Application of decision tree algorithm for data mining in healthcare operations : A Case Study," *Int. J. Comput. Appl.*, vol. 52, no. 6, p. 8887, 2012.
- [59] L. M. Taft, E. RS, S. CR, M. JA, T. SN, B. BE, and V. M, "Application of knowledge discovery in databases methodologies for predictive models for pregnancy adverse events," The University of Utah, 2010.
- [60] M. Beckmann, N. F. F. Ebecken, B. S. L. Lima, and P. De Lima, "A KNN undersampling approach for data balancing," *J. Intell. Learn. Syst. Appl.*, vol. 7, pp. 104–116, 2015.
- [61] M. Ramaswami and R. Bhaskaran, "A Study on feature selection techniques in educational data mining," *J. Comput.*, vol. 1, no. 1, pp. 7–11, 2009.
- [62] S. Shojaee, A. Mustapha, F. Sidi, and M. A. Jabar, "A study on classification learning algorithms to predict crime status," *Int. J. Digit. Content Technol. its Appl.*, vol. 7, no. October 2016, pp. 361–369, 2013.
- [63] V. Kumar and S. Minz, "Feature selection : A literature review," *Smart Comput. Rev.*, vol. 4, no. 3, pp. 211–229, 2014.
- [64] A. B. de Moraes, R. R. Zanini, J. Riboldi, and Elsa Regina Justo Giuglini, "Risk factors for low birth weight in Rio Grande do Sul State , Brazil," vol. 28, no. 12, pp. 2293–2305, 2012.
- [65] G. Marshall, J. L. Tapia, D. Ivonne, C. Grandi, C. Barros, A. Alegria, J. Standen, R. Panizza, A. Bancalari, J. Lacarruba, and J. Fabres, "A new score for predicting neonatal very low birth weight mortality risk in the NEOFOSUR south American network," *J. Perinatol.*, vol. 25, pp. 577–582, 2005.

- [66] H. E. Nasreen, Z. N. Kabir, Y. Forsell, and M. Edhborg, "Low birth weight in offspring of women with depressive and anxiety symptoms during pregnancy : results from a population based study in Bangladesh," *BMC Public Health*, vol. 10, no. 515, 2010.
- [67] S. Vanka, V. Saxena, A. Vanka, and K. S. Ravi, "Maternal periodontitis and low birth weight babies," *J Orofac Res*, vol. 3, no. 4, pp. 253–255, 2013.
- [68] M. Hack, D. J. Flannery, M. Schluchter, L. Cartar, E. Borawski, and N. Klein, "Outcomes in young adulthood for very-low-birth-weight infants," *N. Engl. J. Med.*, vol. 346, no. 3, pp. 149–157, 2002.
- [69] B. Shneiderman, "Inventing discovery tools : Combining information visualization with data mining," in *Keynote for Discovery Science 2001 Conference*, 2001, pp. 1–12.
- [70] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software : An update," *SIGKDD Explor.*, vol. 11, no. 1, pp. 10–18, 2009.
- [71] S. Singhal and M. Jena, "A Study on WEKA tool for data preprocessing , classification and clustering," *Int. J. Innov. Technol. Explor. Eng.*, vol. 2, no. 6, pp. 250–253, 2013.
- [72] K. S. Joseph and T. Ravichandran, "A comparative evaluation of Software effort estimation using REPTree and K* in handling with missing Values," *Aust. J. Basic Appl. Sci.*, vol. 6, no. 7, pp. 312–317, 2012.
- [73] R. Setiono, "Feature selection : an Ever evolving frontier in data mining," in *JMLR: Workshop and Conference Proceedings*, 2010, pp. 4–13.
- [74] V. Podgorelec, P. Kokol, B. Stiglic, and I. Rozman, "Decision Trees : An overview and their use in medicine," *J. Med. Syst.*, vol. 26, no. 5, pp. 445–463, 2002.
- [75] B. Mounia and C. Habiba, "Big data privacy in healthcare Moroccan context," in *Procedia Computer Science*, 2015, vol. 63, pp. 575–580.
- [76] L. Tanner, M. Schreiber, J. G. H. Low, A. Ong, T. Tolfvenstam, Y. L. Lai, L. C. Ng, Y. S. Leo, L. T. Puong, S. G. Vasudevan, C. P. Simmons, L. Martin, and E. E. Ooi, "Decision tree algorithms predict the diagnosis and outcome of dengue fever in the early phase of illness," *PLoS Negl Trop Dis*, vol. 2, no. 3, 2008.

7 Appendices

7.1 Appendix A: North Carolina State Centre for Health Statistics-2006 NC State Birth File Description

<i>Position</i>	<i>Variable Description</i>	<i>Variable Name</i>	<i>Coding</i>
1	Hospital Type Code	Institution type	See Table D
2 - 6	Place of Residence	Place of residence mother	NOT USED
8 - 10	County of Residence	County of residence mother	001 – 100 See Table B 999 Out of State
13	Plurality	Plurality of birth	1 Singleton 2 Twins 3 Triplets 4 Quadruplets 5 Quintuplets or Higher
	Birth Date	Date of birth	MM/DD/YY
	Month of Birth	Month of birth	1 - 12
	Day of Birth	Day of birth	1 - 31
	Year of Birth	Year of birth	2006
22	Sex	Gender of child	1 Male 2 Female
23	Race of Child	Race of Child	0 Other Nonwhite 1 White 2 Black 3 American Indian 4 Chinese 5 Japanese 6 Hawaiian 7 Filipino 8 Other Asian or Pacific Islander
24 - 25	Age of Father	FAGE	14 – 95 Years of age
26 - 27	Age of Mother	MAGE	10 – 55 Years of age

28 - 29	Education of Father	Education of father (years)	Years of education
30 - 31	Education of Mother	Education of mother (years)	Years of education
32 - 33	Total Pregnancies (Including This One)	Total Preg	01 – 25 Total Pregnancies
34 - 35	Number of Children Born Alive Now Dead	BDcad	00 – 25 Children
36 - 37	Number of Other Terminations	Terms	00 – 25 Abortions and/or Fetal Deaths
38 - 43	Date of Last Live Birth	Date LBirth, Month LBirth, Year LBirth	MMYYYY Format Year must be less than or equal to current year 00 in first 2 positions denotes unknown month/not applicable; 0000 in last four positions denotes unknown year/not applicable 000000 denotes No Prior Live Births or unknown Can be equal to month/year of birth for 2nd and higher order births in multiple birth events
44	Outcome of Last Delivery	LOutcome	1 Live Birth 2 Fetal Death 9 Not Applicable/Unknown
45 - 46	Completed Weeks of Gestation (calculated)	Gest Age	18 – 45 Weeks
47 - 48	Month Prenatal Care Began	Prenatal	0 No Prenatal Care 01 – 09 First – Ninth month of Pregnancy
49 - 50	Number of Prenatal Visits	Visits	00 – 49 Visits 99 Unknown
51 - 52	Birth weight group	Birth weight group	00 500 grams or less 01 501 – 1000 grams 02 1001 – 1500 grams 03 1501 – 2000 grams 04 2001 – 2500 grams

			05 2501 – 3000 grams 06 3001 – 3500 grams 07 3501 – 4000 grams 08 4001 – 4500 grams 09 4501 grams or more 99 Unknown
53	Marital Status	Marital	1 Married 2 Not Married
54	Attendant	Birth Attendant	1 MD/Physician 2 DO Physician 3 Certified Midwife 4 Other Midwife 5 Other
56 - 57	Number of Living Children	Numchild	00 – 25 Children
58 - 63	Date of Last Termination	Date Term (mmyyyy), Month Term, Year Term	MMYYYY Format Year must be less than or equal to current year 00 in first 2 positions denotes unknown month/not applicable; 0000 in last four positions denotes unknown year/not applicable 000000 denotes No Prior Terminations or unknown
64 - 65	Pounds of Birthweight	Birth weight (lbs)	00 – 14 Pounds 15 15 Pounds or Over (Ounces is 00)
66 - 67	Ounces of Birthweight	Birth weight (remainder oz.)	00 – 15 Ounces
	Birth weight (grams)	Birth weight (g)	Birth weight in grams
	Low Birth Weight?	Low Birth	Birth weight $\leq$ 2500 grams = Y Birth weight > 2500 grams = N
68	Race of Mother	RaceMom	Same coding as Race of Child

69	Race of Father	RaceDad	Same coding as Race of Child
	Minority Status of Mother	Mother Minority	White or Nonwhite
	Minority Status of Father	Father Minority	White or Nonwhite
70	Hispanic Origin of Mother	HISPMOM	C Cuban M Mexican N Non-Hispanic O Other Hispanic P Puerto Rican S Central/South American U Unknown
71	Hispanic Origin of Father	HISPDAD	See values for position 70, above
77 - 78	Average # Cigarettes Used Daily	AveCig	0 No Tobacco Use 01 - 97 Cigarettes per day
79 - 80	Average # Drinks Consumed Weekly	AveDrink	0 No Alcohol Use 01 - 97 Drinks per week
81 - 82	Weight Gained	Wt Gain	0 Weight Loss or No Gain 01 - 97 Pounds Gained
83 - 99	Medical History		
83	Anemia	ANEMIA	0 Condition not present 1 Condition present
84	Cardiac disease	CARDIAC	
85	Acute or chronic lung disease	ACLUNG	
86	Diabetes	DIABETES	
87	Genital Herpes	HERPES	
88	Hydramnios/Oligohydramnios	HYDRAM	
89	Hemoglobinopathy	HEMOGLOB	
90	Hypertension, Chronic	HYPERCH	
91	Hypertension, Pregnancy	HYPERPR	
92	Eclampsia	ECLAMP	
93	Incompetent Cervix	CERVIX	
94	Previous infant 4000+ grams	PINFANT	
		PRETERM	

95	Previous preterm /small infant	RENAL	
96	Renal disease	RHSEN	
97	Rh sensitization	UTERINE	
98	Uterine Bleeding	OTHMED	
99	Other		
100 - 106	Obstetric Procedures		
	Amniocentesis	AMNIO	0 Condition not present
100	Electronic fetal monitoring	MONITOR	1 Condition present
101	Induction of labour	INDUCT	
102	Stimulation of labour	STIMULA	
103	Tocolysis	TOCOLY	
104	Ultrasound	ULTRA	
105	Other	OTHPROC	
106			
107 - 122	Events of Labour and/or Delivery		
	Febrile (> 100°F or 38°C)	FEBRILE	0 Condition not present
107	Meconium, moderate/heavy	MECONIUM	1 Condition present
108	Premature rupture of membrane	RUPTURE	
109	Abruptio placenta	ABRUPTIO	
110	Placenta previa	PREVIA	
111	Other excessive bleeding	BLEEDING	
112	Seizures during labor	SEIZURES	
113	Precipitous labor (<3 hours)	PRECLAB	
114	Prolonged labor (> 20 hours)	PROLAB	
115	Dysfunctional labor	DYSLAB	
116	Breech/Malpresentation	BREECH	
117	Cephalopelvic disproportion	CEPHALO	
118	Cord prolapse	PROLAPSE	
119	Anesthetic complications	ANESTH	
	Fetal distress		

120	Other	DISTRESS	
121		OTHLABOR	
122			
123 - 128	Method of Delivery		
123	Vaginal	VAGINAL	0 Condition not present
124	Vaginal after C-section Primary C-section	VAGCSECT	1 Condition present
125	Repeat C-section	PRCSECT	
126	Forceps	RECSECT	
127	Vacuum	FORCEPS	
128		VACUUM	
129	Kotelchuck Index of the Adequacy of Prenatal Care	Kotelchuck Index	0 Missing Information 1 Inadequate 2 Intermediate 3 Adequate 4 Adequate Plus
130 - 151	Congenital Anomalies		0 Condition not present
130	Anencephalus	ANENCEPH	1 Condition present
131	Spina bifida/Meningocele	SPINABIF	9 Unknown
131	Hydrocephalus	HYDROCEP	
132	Microcephalus	MICROCEP	
133	Other central nervous system	OTHCNS	
134	Heart malformations	HEARTMAL	
135	Other circulatory/respiratory	OTHCIRC	
136	Rectal atresia/stenosis	RECTALAT	
137	Tracheo-esophageal/fistula	FISTULA	
138	Omphalocele/Gastroschisis	OMPHALAC	
139	Other gastro anomalies	OTHGAST	
140	Malformed genitalia	MALGEN	
141	Renal agenesis	RANGENE	
142	Other urogenital anomalies	OTHURO	
143	Cleft lip/palate	CLEFT	
144	Polydactyly/Syndactyly/Adactyly	POLYDACT	

145	Club foot	CLUBFOOT	
146	Diaphragmatic hernia	DHERNIA	
147	Other musculoskeletal/integumental	OTHMS	
148	Down's Syndrome	DOWNS	
149	Other chromosomal anomalies	OTHCHROM	
150	Other	OTHER	
151			
152 - 153	Weeks gestation as reported on birth certificate	Gest Age (BC)	Gestational age reported on birth certificate
162 - 163	Apgar Score 1 Minute	APGAR Score (1 minute)	00 – 10 Apgar Points
164 - 165	Apgar Score 5 Minutes	APGAR Score (5 minute)	00 – 10 Apgar Points
166	Kessner Index	Kessner Index	1 Adequate 2 Intermediate 3 Inadequate
167	Infant Transfer Status	InfantTran	1 Infant Transferred to other facility After delivery 2 No Transfer
168	Mother Transfer Status	MomTran	1 Mother Transferred to other facility BEFORE delivery 2 No Transfer

169 - 177	Conditions of the Newborn		
169	Anemia	IANEMIA	0 Condition not present
170	Birth injury	BINJURY	1 Condition present
171	Fetal alcohol syndrome	FAS	
172	Hyaline membrane disease/RDS	HYALINE	
173	Meconium aspiration syndrome	ASPIRATE	
174	Assisted ventilation < 30 min	VENTLESS	
175	Assisted ventilation ≥ 30 min	VENTMORE	
176	Seizures	ISEIZURE	
177	Other	OTHINF	

Table B
County of Occurrence
County of Residence

001	Alamance	021	Chowan	041	Guilford	061	Mitchell	081	Rutherford
002	Alexander	022	Clay	042	Halifax	062	Montgomery	082	Sampson
003	Alleghany	023	Cleveland	043	Harnett	063	Moore	083	Scotland
004	Anson	024	Columbus	044	Haywood	064	Nash	084	Stanly
005	Ashe	025	Craven	045	Henderson	065	New Hanover	085	Stokes
006	Avery	026	Cumberland	046	Hertford	066	Northampton	086	Surry
007	Beaufort	027	Currituck	047	Hoke	067	Onslow	087	Swain
008	Bertie	028	Dare	048	Hyde	068	Orange	088	Transylvania
009	Bladen	029	Davidson	049	Iredell	069	Pamlico	089	Tyrrell
010	Brunswick	030	Davie	050	Jackson	070	Pasquotank	090	Union
011	Buncombe	031	Duplin	051	Johnston	071	Pender	091	Vance
012	Burke	032	Durham	052	Jones	072	Perquimans	092	Wake
013	Cabarrus	033	Edgecombe	053	Lee	073	Person	093	Warren
014	Caldwell	034	Forsyth	054	Lenoir	074	Pitt	094	Washington
015	Camden	035	Franklin	055	Lincoln	075	Polk	095	Watauga
016	Carteret	036	Gaston	056	McDowell	076	Randolph	096	Wayne
017	Caswell	037	Gates	057	Macon	077	Richmond	097	Wilkes

018	Catawba	038	Graham	058	Madison	078	Robeson	098	Wilson
019	Chatham	039	Granville	059	Martin	079	Rockingham	099	Yadkin
020	Cherokee	040	Greene	060	Mecklenburg	080	Rowan	100	Yancey

Table D
Hospital Type Code

–	En Route to Facility
0	Home or Non-Institution
1	General Hospital
2	TB Hospital
3	Mental Hospital
4	Chronic Hospital
5	Penal Hospital
6	Veterans Hospital
7	Nursing & Rest Home
8	Clinic & Doctor’s Office
9	Other Institution
~	Other Institution

Table F

Race

0	Other non-White
1	White
2	Black
3	American Indian
4	Chinese
5	Japanese
6	Hawaiian
7	Filipino
8	Other Asian
9	Unknown

7.2 Appendix B: The J48 result of run information of experiment one using the actual data

Correctly Classified Instances	735	36.75	%							
Incorrectly Classified Instances	1265	63.25	%							
Kappa statistic	0.123									
Mean absolute error	0.136									
Root mean squared error	0.3027									
Relative absolute error	90.9343	%								
Root relative squared error	110.3427	%								
Total Number of Instances	2000									
=== Detailed Accuracy By Class ===										
	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class	
	0.125	0.001	0.333	0.125	0.182	0.202	0.747	0.131	0	
	0.389	0.004	0.467	0.389	0.424	0.421	0.852	0.285	1	
	0.500	0.004	0.529	0.500	0.514	0.510	0.936	0.381	2	
	0.368	0.015	0.326	0.368	0.346	0.333	0.833	0.281	3	
	0.217	0.031	0.306	0.217	0.254	0.218	0.683	0.189	4	
	0.258	0.143	0.278	0.258	0.268	0.118	0.606	0.235	5	
	0.587	0.461	0.424	0.587	0.492	0.121	0.571	0.410	6	
	0.280	0.195	0.346	0.280	0.309	0.091	0.584	0.313	7	
	0.040	0.028	0.103	0.040	0.057	0.018	0.572	0.100	8	
	0.000	0.001	0.000	0.000	0.000	-0.002	0.576	0.015	9	
Weighted Avg.	0.368	0.251	0.340	0.368	0.346	0.120	0.599	0.307		
=== Confusion Matrix ===										
a	b	c	d	e	f	g	h	i	j	<-- classified as
1	5	1	0	1	0	0	0	0	0	a = 0
2	7	3	4	1	1	0	0	0	0	b = 1
0	3	9	5	0	1	0	0	0	0	c = 2
0	0	3	14	7	5	7	1	1	0	d = 3
0	0	1	9	26	42	34	7	1	0	e = 4
0	0	0	5	26	91	170	52	9	0	f = 5
0	0	0	6	16	98	430	167	15	0	g = 6
0	0	0	0	5	72	287	151	24	1	h = 7
0	0	0	0	2	14	80	49	6	0	i = 8
0	0	0	0	1	3	7	9	2	0	j = 9

7.3 Appendix C: The REPTree result of run information of experiment one using the actual data

Correctly Classified Instances	783	39.15	%							
Incorrectly Classified Instances	1217	60.85	%							
Kappa statistic	0.1329									
Mean absolute error	0.139									
Root mean squared error	0.2762									
Relative absolute error	92.9551	%								
Root relative squared error	100.6905	%								
Total Number of Instances	2000									
=== Detailed Accuracy By Class ===										
	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class	
	0.625	0.002	0.625	0.625	0.623	0.623	0.998	0.554	0	
	0.389	0.003	0.538	0.389	0.452	0.453	0.956	0.365	1	
	0.722	0.012	0.361	0.722	0.481	0.505	0.987	0.326	2	
	0.289	0.010	0.355	0.289	0.319	0.309	0.869	0.195	3	
	0.008	0.003	0.167	0.008	0.016	0.025	0.779	0.210	4	
	0.289	0.151	0.291	0.289	0.290	0.139	0.632	0.257	5	
	0.690	0.341	0.424	0.690	0.525	0.146	0.582	0.422	6	
	0.256	0.151	0.384	0.256	0.307	0.121	0.623	0.346	7	
	0.007	0.003	0.167	0.007	0.013	0.019	0.602	0.103	8	
	0.000	0.000	0.000	0.000	0.000	0.000	0.547	0.013	9	
Weighted Avg.	0.392	0.266	0.350	0.392	0.345	0.130	0.629	0.326		
=== Confusion Matrix ===										
a	b	c	d	e	f	g	h	i	j	<-- classified as
5	3	0	0	0	0	0	0	0	0	a = 0
2	7	6	1	0	2	0	0	0	0	b = 1
0	3	13	1	0	1	0	0	0	0	c = 2
0	0	7	11	0	15	5	0	0	0	d = 3
0	0	4	11	1	64	34	6	0	0	e = 4
0	0	3	2	1	102	190	54	1	0	f = 5
1	0	3	4	1	102	505	115	1	0	g = 6
0	0	0	1	2	48	343	138	3	0	h = 7
0	0	0	0	1	14	93	42	1	0	i = 8
0	0	0	0	0	2	16	4	0	0	j = 9

7.4 Appendix D: The Random tree result of run information of experiment one using the actual data

```

Correctly Classified Instances      571          28.55 %
Incorrectly Classified Instances  1429          71.45 %
Kappa statistic                    0.0341
Mean absolute error                0.1434
Root mean squared error           0.3671
Relative absolute error            95.8988 %
Root relative squared error       138.8138 %
Total Number of Instances        2000

```

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.000	0.003	0.000	0.000	0.000	-0.003	0.559	0.021	0
	0.000	0.003	0.000	0.000	0.000	-0.003	0.550	0.023	1
	0.000	0.008	0.000	0.000	0.000	-0.008	0.493	0.009	2
	0.000	0.017	0.000	0.000	0.000	-0.018	0.498	0.019	3
	0.100	0.049	0.118	0.100	0.107	0.055	0.544	0.071	4
	0.201	0.164	0.208	0.201	0.205	0.038	0.528	0.139	5
	0.417	0.402	0.374	0.417	0.394	0.014	0.509	0.370	6
	0.309	0.256	0.309	0.309	0.309	0.058	0.521	0.287	7
	0.099	0.054	0.130	0.099	0.113	0.051	0.511	0.082	8
	0.045	0.012	0.040	0.045	0.043	0.031	0.509	0.012	9
Weighted Avg.	0.286	0.253	0.274	0.286	0.279	0.033	0.518	0.257	

=== Confusion Matrix ===

```

a b c d e f g h i j <-- classified as
0 0 0 1 0 3 4 0 0 0 | a = 0
0 0 2 2 1 4 7 2 0 0 | b = 1
0 1 0 3 2 3 5 3 1 0 | c = 2
0 0 1 0 4 10 12 9 2 0 | d = 3
0 3 2 4 12 25 41 26 6 1 | e = 4
1 1 2 8 19 71 151 84 15 1 | f = 5
2 0 4 11 36 131 305 196 37 10 | g = 6
2 0 4 3 21 79 220 167 38 6 | h = 7
0 0 0 2 7 15 59 47 15 6 | i = 8
0 0 0 0 2 0 11 7 1 1 | j = 9

```

7.5 Appendix E: The J48 result of run information of experiment two using SMOTE

```

Correctly Classified Instances      3788          66.3165 %
Incorrectly Classified Instances  1924          33.6835 %
Kappa statistic                    0.6229
Mean absolute error                0.0752
Root mean squared error           0.2197
Relative absolute error            41.9751 %
Root relative squared error       73.4533 %
Total Number of Instances        5712

```

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.978	0.003	0.964	0.978	0.971	0.968	0.996	0.971	0
	0.962	0.003	0.968	0.962	0.965	0.962	0.992	0.964	1
	0.939	0.007	0.928	0.939	0.934	0.928	0.988	0.946	2
	0.874	0.026	0.782	0.874	0.825	0.807	0.957	0.840	3
	0.435	0.036	0.468	0.435	0.451	0.413	0.835	0.402	4
	0.380	0.074	0.449	0.380	0.412	0.329	0.781	0.377	5
	0.528	0.123	0.409	0.528	0.461	0.365	0.808	0.359	6
	0.255	0.059	0.304	0.255	0.277	0.213	0.786	0.252	7
	0.545	0.034	0.659	0.545	0.596	0.556	0.869	0.581	8
	0.922	0.017	0.880	0.922	0.900	0.887	0.979	0.916	9
Weighted Avg.	0.663	0.044	0.660	0.663	0.659	0.617	0.890	0.637	

=== Confusion Matrix ===

```

a b c d e f g h i j <-- classified as
396 2 0 2 3 2 0 0 0 0 | a = 0
5 484 10 4 0 0 0 0 0 0 | b = 1
1 11 450 12 4 0 1 0 0 0 | c = 2
3 2 12 484 28 19 3 0 2 1 | d = 3
2 0 6 60 167 89 31 8 17 4 | e = 4
3 1 6 39 103 297 193 60 70 9 | f = 5
1 0 0 8 29 130 420 170 34 3 | g = 6
0 0 1 5 19 62 281 133 17 4 | h = 7
0 0 0 5 4 57 88 61 334 64 | i = 8
0 0 0 0 0 8 10 5 33 623 | j = 9

```

7.6 Appendix F: The REPTree result of run information of experiment two using SMOTE

Correctly Classified Instances	3524	61.6947 %								
Incorrectly Classified Instances	2188	38.3053 %								
Kappa statistic	0.5707									
Mean absolute error	0.0871									
Root mean squared error	0.2285									
Relative absolute error	48.6698 %									
Root relative squared error	76.3784 %									
Total Number of Instances	5712									
=== Detailed Accuracy By Class ===										
	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class	
	0.978	0.002	0.971	0.978	0.974	0.972	0.995	0.982	0	
	0.936	0.004	0.935	0.936	0.946	0.941	0.989	0.959	1	
	0.904	0.011	0.885	0.904	0.895	0.885	0.980	0.938	2	
	0.819	0.036	0.707	0.819	0.759	0.734	0.935	0.739	3	
	0.341	0.037	0.398	0.341	0.367	0.327	0.791	0.294	4	
	0.296	0.069	0.404	0.296	0.341	0.259	0.754	0.321	5	
	0.620	0.140	0.417	0.620	0.498	0.410	0.853	0.401	6	
	0.192	0.053	0.266	0.192	0.223	0.161	0.818	0.254	7	
	0.387	0.034	0.575	0.387	0.462	0.422	0.834	0.458	8	
	0.855	0.046	0.715	0.855	0.779	0.750	0.955	0.793	9	
Weighted Avg.	0.617	0.050	0.606	0.617	0.604	0.560	0.883	0.590		
=== Confusion Matrix ===										
a	b	c	d	e	f	g	h	i	j	<-- classified as
396	1	1	5	1	1	0	0	0	0	a = 0
5	471	14	12	1	0	0	0	0	0	b = 1
0	11	433	26	7	1	1	0	0	0	c = 2
1	5	23	454	40	14	9	0	1	7	d = 3
1	4	13	79	131	79	32	6	21	18	e = 4
3	1	3	44	88	231	208	57	80	66	f = 5
0	0	2	6	23	105	493	147	14	5	g = 6
0	0	0	4	9	64	329	100	10	6	h = 7
2	0	0	9	21	53	99	64	237	128	i = 8
0	0	0	3	8	24	12	2	49	578	j = 9

7.7 Appendix G: The Random tree result of run information of experiment two using SMOTE

Correctly Classified Instances	3427	59.9965 %
Incorrectly Classified Instances	2285	40.0035 %
Kappa statistic	0.5524	
Mean absolute error	0.0824	
Root mean squared error	0.255	
Relative absolute error	46.0083 %	
Root relative squared error	85.2355 %	
Total Number of Instances	5712	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.958	0.004	0.946	0.958	0.952	0.948	0.981	0.945	0
	0.891	0.013	0.870	0.891	0.880	0.869	0.955	0.857	1
	0.864	0.016	0.830	0.864	0.847	0.832	0.949	0.824	2
	0.798	0.032	0.731	0.798	0.763	0.737	0.908	0.704	3
	0.375	0.041	0.400	0.375	0.387	0.345	0.733	0.258	4
	0.297	0.082	0.364	0.297	0.327	0.234	0.697	0.267	5
	0.438	0.121	0.368	0.438	0.400	0.295	0.725	0.295	6
	0.238	0.062	0.277	0.238	0.256	0.188	0.697	0.201	7
	0.444	0.040	0.574	0.444	0.500	0.453	0.783	0.402	8
	0.910	0.040	0.752	0.910	0.823	0.802	0.951	0.821	9
Weighted Avg.	0.600	0.051	0.588	0.600	0.591	0.542	0.827	0.533	

=== Confusion Matrix ===

a	b	c	d	e	f	g	h	i	j	<-- classified as
388	2	2	3	2	3	0	2	1	2	a = 0
4	448	15	8	3	6	10	1	5	3	b = 1
2	10	414	21	7	4	10	2	6	3	c = 2
1	9	25	442	30	13	15	5	8	6	d = 3
1	13	14	48	144	49	58	14	21	22	e = 4
6	14	12	37	85	232	206	74	66	49	f = 5
4	4	7	12	29	153	348	189	44	5	g = 6
1	4	3	5	30	97	224	124	26	8	h = 7
3	5	6	23	28	67	69	35	272	105	i = 8
0	6	1	6	2	14	5	2	25	615	j = 9

7.8 Appendix H: The J48 result of run information of experiment three using feature selection

Correctly Classified Instances	3789	66.334 %								
Incorrectly Classified Instances	1923	33.666 %								
Kappa statistic	0.6222									
Mean absolute error	0.0786									
Root mean squared error	0.2139									
Relative absolute error	43.9007 %									
Root relative squared error	71.8211 %									
Total Number of Instances	5712									
=== Detailed Accuracy By Class ===										
	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class	
	0.985	0.002	0.978	0.985	0.982	0.980	0.998	0.975	0	
	0.968	0.004	0.962	0.968	0.965	0.962	0.993	0.969	1	
	0.933	0.007	0.927	0.933	0.930	0.924	0.981	0.934	2	
	0.866	0.023	0.800	0.866	0.832	0.814	0.964	0.863	3	
	0.383	0.031	0.473	0.383	0.423	0.389	0.850	0.408	4	
	0.373	0.071	0.454	0.373	0.409	0.328	0.792	0.391	5	
	0.677	0.153	0.418	0.677	0.517	0.434	0.857	0.402	6	
	0.176	0.040	0.305	0.176	0.223	0.175	0.837	0.275	7	
	0.450	0.027	0.665	0.450	0.537	0.504	0.855	0.532	8	
	0.935	0.025	0.833	0.935	0.881	0.866	0.979	0.884	9	
Weighted Avg.	0.663	0.046	0.659	0.663	0.652	0.615	0.903	0.640		
=== Confusion Matrix ===										
a	b	c	d	e	f	g	h	i	j	<-- classified as
399	1	0	0	3	2	0	0	0	0	a = 0
3	487	9	4	0	0	0	0	0	0	b = 1
2	13	447	10	5	1	1	0	0	0	c = 2
0	2	10	480	30	23	5	0	2	2	d = 3
1	1	11	54	147	106	34	8	17	5	e = 4
1	2	5	34	86	291	219	45	83	15	f = 5
1	0	0	9	23	104	538	106	11	3	g = 6
0	0	0	2	8	56	356	92	5	3	h = 7
1	0	0	5	7	54	123	48	276	99	i = 8
0	0	0	2	2	4	12	3	21	632	j = 9

7.9 Appendix I: The REPTree result of run information of experiment three using feature selection

Correctly Classified Instances	3527	61.7472 %								
Incorrectly Classified Instances	2185	38.2528 %								
Kappa statistic	0.8712									
Mean absolute error	0.0872									
Root mean squared error	0.2275									
Relative absolute error	45.6854 %									
Root relative squared error	76.0525 %									
Total Number of Instances	5712									
=== Detailed Accuracy By Class ===										
	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class	
	0.980	0.002	0.971	0.980	0.975	0.974	0.995	0.982	0	
	0.940	0.004	0.954	0.940	0.947	0.942	0.991	0.962	1	
	0.898	0.011	0.881	0.898	0.889	0.879	0.982	0.939	2	
	0.816	0.035	0.717	0.816	0.764	0.738	0.935	0.743	3	
	0.336	0.037	0.396	0.336	0.363	0.323	0.783	0.284	4	
	0.288	0.073	0.384	0.288	0.329	0.243	0.752	0.313	5	
	0.619	0.142	0.414	0.619	0.496	0.407	0.852	0.395	6	
	0.190	0.051	0.273	0.190	0.224	0.164	0.827	0.262	7	
	0.405	0.033	0.593	0.405	0.481	0.441	0.837	0.480	8	
	0.861	0.045	0.720	0.861	0.784	0.756	0.960	0.808	9	
Weighted Avg.	0.617	0.051	0.607	0.617	0.605	0.561	0.884	0.593		
=== Confusion Matrix ===										
a	b	c	d	e	f	g	h	i	j	<-- classified as
397	1	1	4	1	1	0	0	0	0	a = 0
5	473	14	10	1	0	0	0	0	0	b = 1
0	11	430	27	8	2	1	0	0	0	c = 2
1	6	23	452	40	15	9	0	2	6	d = 3
2	4	15	76	129	79	32	5	20	22	e = 4
3	1	3	47	91	225	215	53	76	67	f = 5
0	0	2	6	22	107	492	145	14	7	g = 6
0	0	0	4	11	65	327	99	9	7	h = 7
1	0	0	3	19	66	101	58	248	117	i = 8
0	0	0	1	4	26	12	2	49	582	j = 9

7.10 Appendix J: The Random tree result of run information of experiment three using feature selection

Correctly Classified Instances	3712	64.986 %
Incorrectly Classified Instances	2000	35.014 %
Kappa statistic	0.6084	
Mean absolute error	0.0724	
Root mean squared error	0.2382	
Relative absolute error	40.4388 %	
Root relative squared error	79.6301 %	
Total Number of Instances	5712	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.983	0.003	0.966	0.983	0.974	0.972	0.992	0.971	0
	0.938	0.006	0.944	0.938	0.941	0.936	0.980	0.943	1
	0.929	0.006	0.931	0.929	0.930	0.924	0.981	0.933	2
	0.877	0.024	0.798	0.877	0.836	0.818	0.950	0.820	3
	0.484	0.041	0.462	0.484	0.473	0.434	0.798	0.371	4
	0.347	0.084	0.396	0.347	0.370	0.279	0.709	0.298	5
	0.459	0.108	0.408	0.459	0.432	0.335	0.748	0.311	6
	0.255	0.064	0.288	0.255	0.269	0.200	0.697	0.192	7
	0.543	0.033	0.667	0.543	0.599	0.560	0.836	0.519	8
	0.922	0.028	0.813	0.922	0.864	0.847	0.965	0.872	9
Weighted Avg.	0.650	0.045	0.643	0.650	0.645	0.601	0.853	0.595	

=== Confusion Matrix ===

a	b	c	d	e	f	g	h	i	j	<-- classified as
398	3	1	2	0	1	0	0	0	0	a = 0
9	472	13	5	4	0	0	0	0	0	b = 1
0	17	445	11	5	1	0	0	0	0	c = 2
0	4	8	486	26	17	3	2	1	7	d = 3
0	1	9	58	186	78	28	10	8	6	e = 4
4	0	1	29	108	271	174	78	69	47	f = 5
0	1	1	5	38	155	363	184	40	6	g = 6
0	2	0	6	21	84	248	133	21	7	h = 7
1	0	0	4	13	67	69	56	333	70	i = 8
0	0	0	3	2	10	7	4	27	623	j = 9

7.11 Appendix K: The J48 result of run information based on an independent dataset

Correctly Classified Instances	5433	54.33 %
Incorrectly Classified Instances	4567	45.67 %
Kappa statistic	0.4876	
Mean absolute error	0.1316	
Root mean squared error	0.2732	
Total Number of Instances	10000	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.457	0.000	1.000	0.457	0.628	0.665	0.910	0.637	0
	0.541	0.037	0.654	0.541	0.592	0.549	0.955	0.704	1
	0.302	0.089	0.400	0.302	0.344	0.276	0.897	0.483	2
	0.731	0.090	0.499	0.731	0.593	0.545	0.876	0.538	3
	0.985	0.033	0.791	0.985	0.878	0.867	0.971	0.804	4
	0.618	0.023	0.769	0.618	0.685	0.656	0.884	0.706	5
	0.963	0.147	0.435	0.963	0.600	0.593	0.893	0.415	6
	0.542	0.111	0.379	0.542	0.446	0.371	0.862	0.363	7
	0.000	0.014	0.000	0.000	0.000	-0.040	0.533	0.196	8
	0.000	0.000	0.000	0.000	0.000	0.000	0.569	0.142	9
Weighted Avg.	0.543	0.056	0.495	0.543	0.495	0.462	0.846	0.513	

=== Confusion Matrix ===

a	b	c	d	e	f	g	h	i	j	<-- classified as
273	324	0	0	0	0	0	0	0	0	a = 0
0	613	520	0	0	0	0	0	0	0	b = 1
0	0	347	802	0	0	0	0	0	0	c = 2
0	0	0	799	294	0	0	0	0	0	d = 3
0	0	0	0	1114	0	17	0	0	0	e = 4
0	0	0	0	0	672	416	0	0	0	f = 5
0	0	0	0	0	0	1015	0	39	0	g = 6
0	0	0	0	0	0	507	600	0	0	h = 7
0	0	0	0	0	82	322	708	0	0	i = 8
0	0	0	0	0	120	34	276	86	0	j = 9