**UNIVERSITY OF GHANA**


# MISCLASSIFICATION COST SENSITIVE LEARNING FOR PREDICTING GONORRHEA INFECTION STATUS IN GHANA

BY

**BEHENE ERIC**

**(10244343)**

THIS THESIS IS SUBMITTED TO THE UNIVERSITY OF GHANA, LEGON IN PARTIAL FULFILMENT OF THE REQUIREMENT FOR THE AWARD OF MPHIL STATISTICS DEGREE

JUNE, 2017

## DECLARATION

I hereby declare that with the exception of cited references to other people's work which has been acknowledged, this work is as the result of my own research work done under supervision and has neither been presented elsewhere either in part or whole for another degree.

Student: Behene Eric (10244343)

Signature .....

Date .....

Principal Supervisor: Dr. Isaac Baidoo

Signature: .....

Date: .....

Co-Supervisor: Dr. F.O.Mettle

Signature: .....

Date: .....

## DEDICATION

Every challenging work needs sacrifice and dedication. This work is dedicated to my God Almighty and my family for their support and care.

**ACKNOWLEDGEMENT**

I thank my academic supervisor Dr. Isaac Baidoo for his advice, encourage and motivation. I also express my gratitude to Dr. F.O. Mettle for his support to make this thesis a reality. I also appreciate the support of the Staff of Noguchi Memorial Institute for Medical Research and staff of Naval Medical Research Unit 3 Ghana Detachment (NAMRU-3GD) especially Mrs. Naiki Pupulampu Attram.

Lastly, to my mother Theresa Nyamah and brother Samuel Behene for their prayers and sacrifice.

## ABSTRACT

Gonorrhea, which is one of the most frequently reported sexually transmitted infection is caused by a bacterium called *Neisseria gonorrhoeae*. This disease can causes a serious public health problem worldwide, with about 88 million new infections occurring each year. Failure to treat this disease can result into pelvic inflammatory disease (PID), chronic pain and also damage the female reproductive organ. In males it can lead to reduced fertility and sterility. In developing countries, the unavailability of diagnostic capacity due to cost, lack of equipment and trained personnel has led to the syndrome based management of sexually transmitted infection (STI). Due to these challenges, there is the need for statistical models for gonorrhoea diagnosis which can easily be obtained and implemented with the appropriate expertise. In diagnosing sexually transmitted infection, a false positive has different impact than vice versa. Assuming equal misclassification cost in such models can lead to incorrect decisions and also incur financial cost and harm to the patient. Many classifiers do not allow integration of cost into model development process but rather are designed to improve prediction accuracy assuming equal misclassification cost. The aim of the study is to develop cost sensitive statistical models for predicting gonorrhoea infection. For the data used for the study, 80% was used for training and 20% for testing. The results indicated that, the cost sensitive classifiers had a reduced total classification cost than the cost insensitive classifiers. Also, the classification cost of all laboratory diagnostic method except culture was lower than the cost sensitive and insensitive model. The class distribution weakly affected the cost sensitive classifiers but not the cost insensitive classifiers.

**TABLE OF CONTENT**

DECLARATION ..... ii
DEDICATION ..... iii
ACKNOWLEDGEMENT ..... iv
ABSTRACT ..... v
TABLE OF CONTENT ..... vi
LIST OF TABLES ..... ix
LIST OF FIGURES ..... x
LIST OF ABBREVIATIONS ..... xi
CHAPTER ONE ..... 1
Introduction ..... 1
1.1 Diagnostic test ..... 1
1.2 Cost Sensitive Learning ..... 3
1.3 Problem Statement ..... 3
1.4 Objectives ..... 4
1.5 Significance ..... 4
1.6 Limitation ..... 5
CHAPTER TWO ..... 6
LITERATURE REVIEW ..... 6
2.1 Brief History ..... 6
2.2 Statistical models application to sexually transmitted infection ..... 8
2.3 Application of classification trees to other medical diagnosis ..... 9
2.4 Research work on Cost sensitive Methods ..... 13
2.5 Conclusion ..... 15
CHAPTER THREE ..... 16
METHODOLOGY ..... 16
3.1 Source of Data ..... 16
3.2 Description of Data ..... 17
3.3 Definition of Some Medical and Statistical Terminology ..... 18
3.4 Logistic Regression ..... 18
3.4.1 Assumption of logistic regression ..... 18
3.4.2 Model Specification ..... 18
3.4.3 Model Evaluation and Diagnostic ..... 19

3.4.3.1 Likelihood Ratio test ..... 19
3.4.3.2 Hosmer-Lemeshow test ..... 20
3.4.3.3 Pearson Residual ..... 21
3.4.3.4 Deviance Residual ..... 21
3.4.3.5 Wald Statistics ..... 22
3.4.3.6 Linearity test ..... 22
3.4.4 Parameter estimation of logistic regression ..... 23
3.4.4.1 Maximum likelihood estimation of parameters of logistic regression ..... 23
3.4.4.2 Variable Selection ..... 24
3.4.4.3 Bayesian Estimation ..... 24
3.4.4.4 Prior Distribution ..... 25
3.4.4.5 Posterior Distribution ..... 25
3.4.4.6 Markov Chain Monte Carlo (MCMC) ..... 26
3.4.4.7 Gibbs sampling ..... 26
3.4.4.8 Burn in ..... 27
3.5 Classification tress ..... 27
3.5.1 Basic Definitions used in classification tree ..... 27
3.5.2 Tree Construction ..... 28
3.5.3 ID3 ..... 28
3.5.4 C4.5 ..... 29
3.5.5 CART ..... 29
3.5.6 Splitting Criterion ..... 29
3.5.7 Information gain ..... 30
3.5.7.1 Gain Ratio ..... 30
3.5.7.2 Entropy ..... 30
3.5.7.3 Gini index ..... 31
3.5.8 Pruning ..... 31
3.5.8.1 Reduced error pruning ..... 31
3.5.8.2 Cost complexity pruning ..... 32
3.6 Random forest ..... 32
3.6.1Random forest construction ..... 34
3.6.2 Bagging ..... 35
3.7 Cost sensitive modelling ..... 35
3.7.1 Modifying the classification of the predicted probability scores obtained from logistic regression to include unequal misclassification cost ..... 37

3.8 Sampling ..... 38
3.9 Performance Measure ..... 39
F-measure: ..... 41
3.9.1 Receiving Operating Characteristics (ROC) ..... 42
CHAPTER FOUR ..... 43
DATA PRESENTATION AND ANALYSIS ..... 43
4.1 Data and Preliminary Analysis ..... 44
4.2 Training Data and Model fitting ..... 47
4.2.1 Logistic Regression ..... 49
4.2.2 Classification tree ..... 51
4.2.3 Random Forest ..... 52
4.3 Cost-Sensitive Models ..... 53
4.3.1 Classification of logistic regression predicted probability score to include unequal misclassification cost ..... 54
4.3.2 Classification tree with unequal misclassification cost ..... 55
4.4 Comparing the performance of the models using training Data ..... 56
4.5 Effect of Total classification cost on cost sensitive and insensitive method ..... 56
4.6 Model Validation ..... 57
4.6.1 Comparing laboratory diagnostic methods with cost sensitive and insensitive models on the testing data ..... 58
4.6.2 Effect of class distribution and cost sensitive method on classification cost ..... 59
4.7 Summary of results ..... 60
CHAPTER FIVE ..... 62
CONCLUSIONS AND RECOMMENDATIONS ..... 62
5.1 Discussion ..... 62
5.1.1 Comparing Logistic regression, classification tree and Random forest ..... 62
5.1.2 Effect of Classification cost on Laboratory diagnostic method and skewed class distribution of Cost sensitive and insensitive classifiers ..... 64
5.2 Conclusion ..... 65
5.3 Recommendation ..... 66
REFERENCE ..... 67
APPENDIX ..... 73

# LIST OF TABLES

Table 4.1a:Background information of the respondent .....46

Table 4.1b:Background information of the respondent .....47

Table 4.2a:Description of Training Data .....48

Table 4.2b:Description of Training Data continuation .....49

Table 4.3: Logistic Regression model using maximum likelihood estimation.....50

Table 4.4: Confusion Matrix.....54

Table 4. 5: Comparing Performance of the various model using training data .....56

Table 4.6: Comparing Performance of the various model using test data .....57

Table 1a: Bayesian logistic regression.....73

## LIST OF FIGURES

Figure 4.1: Distribution of the predicted probability score of the training data .....51

Figure 4.2: Tree structure for gonorrhea data .....52

Figure 4.3: Variable importance for Random forest .....53

Figure 4.4: Determination of optimal cut off of predicted scores using unequal classification cost .....54

Figure 4.5 : Tree structure for gonorrhea data using a cost ratio of 1:4 .....55

Figure 4.6: Effect of classification cost on cost sensitive and cost insensitive classifiers.....57

Figure 4.7: Total cost of classification of Laboratory method, Cost sensitive and insensitive classifiers.....58

Figure 4.8: Laboratory diagnostic methods and Cost sensitive models.....59

Figure 4.9: Effect of class distribution on classification cost of the classifiers .....60

Figure 1a: Pearson Residuals plotted against predictor one by one.....73

Figure 1b: Posterior distribution of the model parameters .....74

Figure 1b: Posterior distribution of the model parameters (Cont.).....74

Figure 1c: Posterior distribution of the model parameters (Cont) .....75

Figure 1d: Posterior distribution of the model parameters (Cont.).....75

Figure 1e : Error rate for the number of trees .....75

## LIST OF ABBREVIATIONS

| | |
|---|---|
| WHO | World Health Organization |
| CDC | Centre for Disease Control and Prevention |
| NAAT | Nucleic Acid Amplification Test |
| AUC | Area Under the Curve |

# CHAPTER ONE

## Introduction

Gonorrhea, is one of the most frequently reported sexually transmitted infection caused by a bacterium called *Neisseria gonorrhoeae*. It causes serious public health problem worldwide, with about 88 million new infections occurring each year (Smith, 2016). It is the third most prevalent sexually transmitted infection (STI) worldwide (WHO, 2005).

Center for Disease Control and Prevention (CDC) report in 2010 stated that about 700000 new cases of *Neisseria gonorrhoeae* are diagnosed yearly in the United States which makes it the second most frequent reported STI after *Chlamydia trachomatis.*

This disease can be acquired through having unprotected vaginal, oral or anal sex with an infected person just like any STI. It can also be transmitted from an infected mother to child through birth. Gonorrhoea can be avoided through abstinence or use of condom during sexual intercourse as both partners can reduce the chances of acquiring it. Failure to treat this disease can result into pelvic inflammatory disease (PID), chronic pain and also damage the female reproductive organ. In males it can lead to reduced fertility and sterility (Handsfield *et al*., 1974).

Men are more symptomatic to the disease as compared to women. They mostly present with symptoms such as penile discharge, painful and frequent urination. Women normally present with increased vaginal discharge, painful urination, lower back pain and spotting between menstrual periods which may occur alone or in combination and may range from hardly visible to severe spotting (Smith, 2016).

### 1.1 Diagnostic test

In developing countries, the unavailability of diagnostic capacity due to cost, lack of equipment and trained personnel has led to the syndromic management of STI. Even when there are

equipment and trained personnel to help in diagnosing these diseases, they are usually found in the urban centres (Meade & Cornelius, 2012).

The main diagnostic techniques for gonorrhoea include, culture, direct microscopic testing and Nucleic Acid Amplification Testing (NAAT) which is currently recommended by CDC (Papp *et al*., 2014).

Culture, which is the gold standard for bacterium identification has a high specificity and sensitivity and is also optimal for antimicrobial susceptibility testing (Murray *et al*., 2003)

Another method which is also useful is the direct microscopy and is more preferable for diagnosing symptomatic gonococcal urethritis in men. This test is not appropriate for the diagnosis of extra-genital infections since non-pathogenic Gram-negative diplococci may be present and may result in false positives (Bignell *et al.*, 2006). This method requires highly trained personnel and also specimen required for the testing needs to be stored and transported under appropriate conditions to maintain organism viability (Whiley *et al*., 2006). This is however not often the case in developing countries. The challenges associated with specimen collection and transportation required for culture based diagnosis has resulted in the development and application of nucleic acid detection methods such as Nucleic Acid Amplification test (NAAT) which utilises urine samples (Whiley *et al*., 2006). Test results from these methods of diagnosing are obtained from urine and specimen which can be obtained in minimal invasive ways. This diagnostic method is often not available in developing countries since it requires specialized equipment which are very expensive and require experienced personnel.

In recent years, molecular diagnostics like nucleic acid amplification technique (NAAT) has captured attention and been recommended as the optimal test for diagnosing gonorrhoea. As compared to culture it is more sensitive and provides faster results (Cosentino *et al*., 2012). In some developing countries, NAAT may not be available in public hospitals hence patients are

requested to go to private laboratories where costs are much higher. These laboratories may provide better results but may not be reliable due to lack of national quality assurance programs to certify them (Ndongmo, 2005).

Due to these challenges, there is the need for statistical models for gonorrhoea diagnosis which can easily be obtained and implemented with the appropriate expertise.

### 1.2 Cost Sensitive Learning

This method is a type of learning in data mining which considers the misclassification cost (and other possible types of cost such as test cost) with the aim to minimize the total cost. Total cost (cost of classification) refers to the total number of people who has been misclassified. In diagnosing sexually transmitted infection, a false positive has different impact than vice versa. To be able to solve this problem, a cost-sensitive classification is obtained which considered the varying misclassification cost (false positive and false negative) using cost matrix (Zhou & Liu, 2006). The cost matrix is used during the model building processes and it is quite subjective (Zadrony & Elkan, 2001).

### 1.3 Problem Statement

There are several classifiers designed for medical diagnosis (Chou & Shapiro, 2003). However, many of these do not allow for integration of misclassification cost into their model development process. Instead, they are designed to improve prediction accuracy assuming equal misclassification cost (Jiang & Cukic, 2009). The dataset used for gonorrhoea prediction is usually unbalanced with regards to the proportion of gonorrhoea positives to gonorrhoea negatives. The class distribution is usually skewed in favour of the gonorrhoea negatives and may cause poor performance when detecting gonorrhoea positive instances. . Assuming equal misclassification cost in such models can lead to incorrect conclusions regarding their diagnosis which can result in significant harm to the patient Therefore, in development of predictive models for gonorrhoea diagnosis, there is the need to reduce the number of people

who could be misclassified by including misclassification cost in the model since this usually varies in real life. From literature, very few studies as reported in Ling *et al*., (2004) and Domingo (1999) have considered the essence of misclassification cost when developing predictive models.

### 1.4 Objectives

The aim of the study is to develop cost sensitive statistical model for predicting gonorrhoea infection.

**Specific Objectives**

- To fit a cost insensitive classifier for predicting gonorrhoea status using logistic regression , classification tress ,and random forest and also to induce cost-sensitive criterion into these classifiers
- To determine the effect of classification cost on cost sensitive and insensitive method
- To determine whether classification cost is affected by both the skewed distribution of class and cost sensitivity
- To compare traditional laboratory diagnostic methods with cost-sensitive and insensitive classifiers

### 1.5 Significance

According to World Health Organisation's (WHO) Western Pacific region manual of tests, an ideal diagnostic tool for reproductive tract infection is one in which results are easily made available to patients, inexpensive, highly sensitive and specific, requires no specialised equipment and also samples are obtained by non-invasive procedure (Verma *et al*.,2009). This study seeks to achieve this goal and also propose an alternative tool for diagnosing gonorrhoea in resource constrained environment which will be beneficial to the clinicians in order to limit the symptomatic management of the disease.

### 1.6 Limitation

The limitation for the study included;

- The data obtained considered only symptomatic patients, hence results obtained cannot be generalised to asymptomatic patients who may or may not have gonorrhoea
- Data for the study was obtained from only three health facilities, it would have produce better results if other health facilities across the country were included. This was not achieved due to scarcity of epidemiological data on gonorrhoea.

### Future Research

The study focused on only misclassification cost which was used to determine the total classification and hence, it would be important to consider the other forms of cost in future research

# CHAPTER TWO

# LITERATURE REVIEW

This chapter of the thesis deals with reviews of related work of previous authors regarding statistical models of medical diagnosis, sexually transmitted disease and cost sensitive models. This review will be based on the methods, data finding and conclusions.

## 2.1 Brief History

Logistic regression evolved in the 19th Century and was used to describe the growth of population and chain reaction (Cramer, 2002). It is a popular model used in biomedical informatics to study the relationship between response and predictors that included physiological data. One of the medical areas it has been applied is cancer prediction (Yusulf *et al*., 2012).

Another method which is also used for classification is the Naïve Bayes classifiers which has been studied since 1950's.This is a probabilistic classifier which uses Bayes theorem. Similarities between logistic regression and Naïve Bayes is that they are both linear classification. Logistic regression estimates the conditional probability of response variable given predictors from the data by minimizing the error. Hence it is termed as a discriminating model whiles Naïve Bayes estimate a joint probability of the response variable and the predictors from the data hence it is a generative model (Kolluru, 2014).

In some instances, researchers would resort to use nonparametric models like decision trees and Random forest for classification. Regarding decision tress, there are two types; Classification trees which have a discrete response variable and Regression tree which uses a continuous response variable.

Decision trees was initially introduced in the 1960 as one of the proactive methods in data mining which has widely been used in several fields such as agriculture, astronomy, image

processing, medicine, software development, financial, manufacturing and production ( Hastie *et al*., 2009).The most used decision tree algorithm are classification and regression tree(CART) (Breiman *et al*., 1984), Interactive Dichotomizer 3(ID3) which was developed by Quinlan (1979, 1983, 1986) and C4.5(Quinlan,1993) which is an improvement of the ID3.These classification methods tend to identify classes in which the object belongs from a descriptive trait.

In 2001, Leo Breiman developed Random forest in other to improve the performance of CART The name came from random decision forest which was first proposed by Tin Kam

Ho. The method is a **combination** of Breiman "bagging" idea and Ho's random subspace method which a collation of decision trees with a regulated variation.

In most data mining techniques, varying misclassification cost produces different cost hence traditional data mining techniques which aim at minimizing error and assume equal misclassification tend to perform poorly in this area. Cost-sensitive learning (CSL) is an extension of traditional inductive learning for tackling the imbalance in misclassification errors, i.e., minimizing classification costs. This issue is practical but challenging, because different classification errors often have distinct costs in real-world applications. As a result of this, Turney in 1995 developed cost sensitive learning for addressing classification with non-uniform cost. In his published paper in 2000, he identified nine major type of cost which are;

- Misclassification
- Test
- Teacher
- Computational
- Interventional
- Unwanted achievement
- Human computer interaction

- Cost of cases
- Cost of instability

**2.2 Statistical models application to sexually transmitted infection**

Researchers have adopted various statistical technique to model sexually transmitted disease. In epidemiological research, the focus is to identify risk factors of a disease in which standard analytical technique (e.g Logistic regression) are used for the analysis. Below are some literature in the field of sexually transmitted disease.

Gardella *et al*. (2005) conducted a study to determine the risk factors of herpes simplex virus (HSV) acquisition among pregnant women at risk. The women and their partners were enrolled and tested for HSV. The risk factors for HSV susceptibility, exposure and acquisition were determined using logistic regression.

Hupert *et al*. (2006) conducted a study which was to determine the association between urinary tract infection and sexually transmitted infection using history, clinical and laboratory findings of symptomatic women. A cross sectional of 296 sexually active women between age 14-22 years were recruited. Logistic regression and CART was used for the analysis and both methods identified virtually the same risk factors. The misclassification rate for CART was obtained to be 38% but that of logistic regression was not stated in the manuscript since the focus of the study was not on the evaluation of the performance of the two models.

Kershaw *et al*. (2007) aimed to use individual, family and community level characteristics to construct a clinical classification tree to help identify women who are at risk of acquiring sexually transmitted infection during pregnancy. This result of the study was to assist clinician who normally uses informal decision trees in making clinical decisions.

### 2.3 Application of classification trees to other medical diagnosis

Long *et al*. (1993) compared the performance of logistic regression to decision trees induction in classifying patient as having acute cardiac ischemia. The data for the study was obtained from six different new England hospital ranging from urban teaching and rural nonteaching hospital. The training data had a total of 3453 patients with 59 clinical features whiles the test data set compose of 2320 patients. The results indicate that, comparing the error rate of the logistic regression with ID3, it perform better on the test set than the training dataset. Comparing the area under the receiving operating characteristic (ROC) curve for the test data, LR(0.89) was still better than ID3(0.82). The ID3 was pruned to improve the performance but the results indicated that LR still outperformed it even though it showed improvement over the default ID3.

Ture *et al*. (2005), compared the performance of classification techniques which is used for the prediction of essential hypertension. Among the classifiers were three decision trees (Chi-square Automatic interaction Detector, CART and Quick unbiased efficient statistical tree), three statistical algorithm (Logistic regression, Flexible discriminant analysis and Multivariate Adaptive Regression spline and two neural networks (Multilayer perceptron and Radial Bias function). For the study, a respective analysis was done on a total of 694 who were obtained from Cardiology Clinic of Trakya University Medical Faculty in Turkey, 2002–2003.The dataset was split into 75% for training and 25% for testing. Findings from the study indicated that the two neural network methods performed better than the other classifiers but the decision trees remained advantageous than the statistical algorithms and neural network since the probabilities available for each terminal node remain dependent on the tree structure and its interpretation may not be the same as the other classifiers.

In the work of de Queiroz Mello *et al*., (2006) which was at aimed to develop a predictive model using logistic regression and classification trees for smear negative pulmonary

tuberculosis (SNPT) for outpatient in areas with scare resources. The study enrolled 551 patients with clinical symptoms of SNPT in which data was divided into training and validation set. Model performance were evaluated using sensitivity, specificity and area under the ROC curve. Classification tree models performed better than logistic regression for the training data but on the validation data logistic regression performed slightly better.

Kurt *et al*. (2008) compared the performance of logistic regression, classification and regression tree and neural network for predicting coronary artery disease (CAD). A retrospective dataset of size 1225 was obtained in Cardiology Clinic of Trakya University Medical Faculty in Turkey between January 2002 and February 2003 was used for the study. Findings indicated that neural network outperformed the other classifiers using the area under the ROC curve. The difference between the value of the area under ROC curve of logistic regression and CART was statistically insignificant.

Lavanya and Rani (2011) evaluated the performance of ID3, C4.5 and CART classifiers on some medical dataset. These were Diabetes, Heart stat log, Thyroid, Breast cancer and Arrhythmia dataset which were obtained from the UCI machine learning repository. Performance measure such as accuracy and time of complexity was access using the 10 fold cross validation of the various datasets. The results on the experimental data indicated that CART performed better than the other two algorithm and also had an improved classification for the medical data set.

The study of Abdullah and Rajalaxmi (2012) was to use Random forest to improve the prediction accuracy and to investigate the various event related to Coronary heart disease (CHD). Data were obtained from the UCI machine learning repository. The classifiers were evaluated using Kappa statistics, classification error and root mean square error. Results

indicated that Random forest performed better base on the evaluation measures than the decision trees.

In the research of Adeyemo and Adeyeye (2015), the performance of ID3, C4.5 and Multilayer perceptron (MLP) Artificial Neural Network in the prediction of typhoid fever were compared. The data was obtained from a Nigerian hospital and divided into a training and testing set. Classifiers were evaluated based on accuracy, root mean square error, F-measure, area under the ROC curve, mean absolute error, relative absolute error, the mean relative square error and Kappa statistics. The result indicated that MLP had a high accuracy and performed better on the other evaluating measures than the other two classifiers. In comparing the two decision tree classifiers, C4.5 outperformed ID3 in terms of area under the ROC curve, misclassification rate root mean square error and the other evaluation measures.

Mohammed (2016) analysed and compared the performance of various classification methods used to diagnose Parkison disease. These classification methods were Naïve Bayes, Support vector machine (SVM) and decision tress. The dataset used contained 22 features obtained from 31 people of which 23 had the Parkison disease (PD). In the study two data set were used which are; Actual PD and discretised PD dataset. For the discretised dataset, continuous variables are discretised. The PD disease was diagnosed using numerous features obtained from the human voice. The dataset was divided into training (70%) and testing (30%).Also cross validation method was used without splitting the dataset into training and testing. Method used to compare the performance of the various models were accuracy. Naive Bayes performed better on the discretised data set with cross validation yielding 84.6% accuracy than compared to using the actual data set. SVM and decision tress also obtained a high accuracy of 96.5% and 89.6% respectively on the discretised data set.

Heish *et al*. (2010) evaluated the performance of Random forest, Support vector machine and Artificial Neural Network compared with logistic regression to diagnose acute appendicitis. Data for the study was from January 2006 to December 2008 in which patients who were suspected of acute appendicitis were enrolled. Sixteen input variables which are commonly used in diagnosing acute appendicitis were used. Operation note and pathology report was used to confirm the diagnosis of acute appendicitis. Those who did not obtain any operation note were followed to make sure they were not false negative. The data set was divided into two, seventy-five percent for training (i.e used for the development of the various models) and twenty five percent was for testing. The area under the Receiving operation characteristic curve (AUC) ,accuracy(AC),sensitivity(SN),specificity(SP),positive predictive value(PPV) and negative predictive value(NPV) were used to evaluate model performance. A total of 180 patients were enrolled in which 135 patients were used for training and 45 patients for testing. The AUC for testing data set for the various models are Random forest (0.98), Support vector machine (0.96), artificial neural network (0.91) and logistic regression (0.77). Random forest had a high AC, SN, SP, PPV and NPV than logistic regression also the SN (0.94) and SP (1.0) values of Random Forest was the same for Artificial Neural Network and Support Vector machine respectively. This is an indication that random forest can predict acute appendicitis more accurately and can be an effective tool for clinician decision making.

Jin *et al*. (2014) compared the prediction test of some data mining algorithm using a data set containing liver disease patients. The data set used was collected from Andhra Pradesh, India which is made up of 414 confirmed liver disease patient and 165 people suspected of liver disease. There were 11 predictor variables which were also used. The classification algorithm which were used are Naïve Bayes, Decision Tree, Multi-layer perceptron, Random forest, logistic regression and K-Nearest Neighbour. These algorithm performance measure were evaluated base on precision, sensitivity, specificity, accuracy, AUC, and root mean square error

(RMSE).The results indicate that, in terms of precision and specificity, Naïve Bayes is superior to the other classification algorithm. Also, Logistic regression together with Random forest showed the highest AUC value. The RMSE value of logistic regression was the lowest (0.42) which mean the difference between the actual and actual and expected value is small which indicates of a relatively low error rate as compared to the other models.

Danjuma & Osofisan (2015) study was aimed to identify the most performing predictive data mining algorithms used in the diagnosis of Erythemato-squamous disease. Predictive models developed were Naïve Bayes, Multilayer Perceptron and J48 decision tree. A 10-fold cross validation and a set of performance measure were used to evaluate the predictive performance of the models. The results indicated that, Naïve Bayes had the best accuracy (97.4%) as compared with the other classifiers.

**2.4 Research work on Cost sensitive Methods**

Cost sensitive learning is a machine learning approach which considers cost of misclassification. Basically, they can be grouped into two; Direct method and Meta learning. Turney (1995) made a contribution in the direct method of cost sensitive learning by developing an algorithm such as ICET which incorporated misclassification cost in the fitness function of genetic algorithm.

Ling *et al*. (2004) also made a contribution to direct method of cost sensitive learning by considering classification cost in the tree generating process which selected attributes with reduced expected total cost instead of attributes with minimum entropy.

Meta learning method converts cost insensitive to cost sensitive classifiers. This method can be grouped into two which are thresholding and sampling.

Domingo (1999) developed a method called MetaCost which used cost insensitive bagging on decision trees to produce estimated probabilities using training data and then applied thresholding to obtain the predicted class.

Witten & Frank (2005) also used a cost insensitive algorithm to obtain the probability estimates and then applied thresholding to predict the class labels.

Regarding to Sampling method, it modify the class distribution of training data and then classifiers using cost insensitive classifiers directly on the sampled data.

In the work of Weisis (2003), the effective of class distribution on decision tree was investigated using under sampling and oversampling method to obtain various measuring performance using accuracy and Area under the curve(AUC).The conclusion that both the under sampling and over sampling method for dealing with class imbalance problem were both effective.

Chawla *et al*. (2002), proposed an approached termed Synthetic Minority Over sampling technique which tend to reduce the overfitting problem in over sampling technique. The method create synthetic data base on the minority class and has proven to be effective

Pazzani *et al*. (1994) compared various cost sensitive decision trees to cost insensitive CART and C4.5.The method which was used in selecting decision split in the tree were GINI criterion with altered prior and also used the misclassification cost as the test selection metric. The study found that the original CART and C4.5 performed better than the cost sensitive trees in terms of minimizing misclassification cost.

Sahin *et al*. (2013) developed a new cost sensitive decision tree which minimize the cost of misclassification while selecting the splitting attribute at each terminal node. This model performance was compared with known traditional classification models such as CART, C5 etc. on credit dataset. Performance measure used were accuracy, true positive rate and save loss

rate. The findings of the study indicated that, the cost sensitive tree algorithm outperformed the existing well known methods.

## 2.5 Conclusion

The review of literature from the various authors indicate that most studies in medical diagnosis use cost insensitive models to predict infection status of various diseases. This thesis would seek to introduce cost of misclassification in the various models since in real life there is variation in the cost of misclassification. Review of previous study has shown that few studies have accounted for the different cost of misclassification resulting from type I and type II errors in developing classifiers especially for sexually transmitted disease. But most of the classifiers assume equal cost even though that is not the case in real life.

# CHAPTER THREE

# METHODOLOGY

This chapter of the thesis describe the various cost insensitive statistical models such as logistic regression, classification tree random forest which was used to fit the data. Also, how cost of misclassification was included into these models were described together with the various performance measure which were used to evaluate these models.

## 3.1 Source of Data

Secondary data from the U.S Naval Medical Research Unit 3, Ghana detachment (NAMRU 3 GD) was used for the thesis. The data are Four year data set which span from 2012 to 2016. Patients who were enrolled into the main study titled "**Sexually Transmitted Disease (STD) Surveillance Characterizing Gonorrhea and Chlamydia Prevalence and Gonorrhea Resistance Profile in Ghana"** were from 37 Military Hospital, Adabraka polyclinic and three Garrison clinics in Takoradi ( Naval Sick bay, Airforce Medical center and 2 Medical Reception station ).For patients to participate in the study, they needed to eligible to be enrolled.

**Inclusion Criteria**

- Aged ≥ 18 may provide independent (autonomous) consent
- Aged > 11 and <18 may participate but will require parental consent and child's assent
- Patients presenting with an STI syndrome.(i.e., urethritis in men and cervicitis in females)
- Pregnant women may be included. This is a group which may particularly benefit from information regarding STI transmission in order to protect themselves and their foetus

If the Patient fulfil the above criteria then a questionnaire is administered, urine sample (for Nucleic Acid Amplification (NAAT) testing) is obtained and two swabs of discharge from penis or the vagina/cervix is obtained (for culture and gram stain testing).

**Exclusion Criteria**

Patients presenting without an STI syndrome or suspicion of Gonorrhea

### 3.2 Description of Data

The variables which were used in the study are described below.

**Dependent Variable**

The dependent variable is gonorrhoea status of patients. This infection status was obtained using Nucleic acid amplification test which is a molecular diagnostic method.

**Independent Variable**

The table below provide a description of the various independent variables used;

Table 3.1 : Independent variables used in modelling

| Variable | Description |
|---|---|
| **Demographic** | |
| Gender | Binary |
| Age | Count |
| Marital status | Discrete |
| Educational level | Discrete |
| **Clinical Presentation** | |
| Painful urination | Binary |
| Discharge | Binary |
| Pain in penis or vagina | Binary |
| Foul smell | Binary |
| Painful sex | Binary |
| Bleeding from penis or vagina | Binary |
| Itching of Genital | Binary |
| **Sexual Behaviour** | |
| Alcohol intake | Binary |
| Use of Condom | Discrete |
| Having more than one sexual partner in the partner month | Discrete |

### 3.3 Definition of Some Medical and Statistical Terminology

**Culture:** It is a laboratory diagnostic method use for diagnosing gonorrhea infection.

**Majority voting:** Majority voting occurs when majority of a particular class is predicted than the other classes.

**Machine learning**: Machine learning is a field of computer science that learn from data without relying on rules base programing**.**

### 3.4 Logistic Regression

Logistic regression is a most common model used in medical diagnosis to fit a binary or dichotomous response variable (Hilbe, 2011). It was first used in the 19$^{th}$ century to describe population growth and it now been adopted in biomedical research to model the log odds of the response variable using the logistic function (Cramer, 2002). The response variable(Y) takes two values which is 0 and 1.The event Y=1 is the success of the event and Y=0 is the failure.

$$Y_i = \begin{cases} 1 & if\ the\ outcome\ is\ observed \\ 0 & otherwise \end{cases}$$

#### 3.4.1 Assumption of logistic regression

- Normal distribution is not necessary or assumed for the response variable
- Normally distributed description of errors are not assumed
- Equal variance is not assumed for each level of the independent variables
- Linearity is assumed for the log odds of the response variable and the covariate.

#### 3.4.2 Model Specification

Let us consider a response variable $Y$ which is binary and having a Bernoulli distribution

$$y \sim B(1, \pi)$$

$$log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \sum_{i=1}^{K} \beta_i X_i \qquad (3.0)$$

$\pi$ is the probability of success

where,

$\underset{\sim}{\boldsymbol{X}} = (x_0, \ldots . x_k)'$ are independent variables

$\underset{\sim}{\boldsymbol{\beta}} = (\beta_0, \ldots . \beta_k)$ are the unknown parameter

### 3.4.3 Model Evaluation and Diagnostic

The goodness of fit of logistic regression is assessed using Likelihood ratio test (LRT) and Hosmer-Lemeshow test (HL).The purpose of LRT is used to determine the overall significance of the predictors in the model (Bewick, Cheek, & Ball, 2005).Hosmer-Lomeshow test is used to determine if there exit some form of interaction or linearity in the model.

Other diagnostic measures used to evaluate the model is deviance, Wald statistics and linearity test. The Wald statistic is used to determine the significance of each predictor in the model (Hosmer Jr, Lemeshow, & Sturdivant, 2013).Linearity test is used to determine if the linerarity between the logs odd of the dependent and the covariate is assumed correctly. Below is the detailed description of the methods.

#### 3.4.3.1 Likelihood Ratio test

Likelihood ratio test compares the likelihood of the data under the full model against the likelihood of a model with fewer predictors. The likelihood ratio statistic is asymptotically distributed as chi-square with one degree of freedom. It is use to determine the overall significance of the predictors

The LRT test statistics is given by;

$$LRT = 2 * (\log L_F - \log L_R) \qquad (3.2)$$

where

Log likelihood of full model ($L_F$) and Log likelihood of reduced model ($L_R$)

$$\log L_F = \sum_{i=1}^{n}\left[y_i\left(\beta_o + \sum_{k=1}^{K} X_i\beta_i\right) - 2y_i \log\left(\frac{1}{1+e^{\beta_o+\sum_{k=1}^{K}\beta_i X_i}}\right) + \ln\left(\frac{1}{1+e^{\beta_o+\sum_{k=1}^{K}\beta_i X_i}}\right)\right]$$

$$\log L_R = \sum_{i=1}^{n}\left[y_i\beta_o - 2y_i \log\left(\frac{1}{1+e^{\beta_o}}\right) + \ln\left(\frac{1}{1+e^{\beta_o}}\right)\right]$$

When the p-value obtained less than 0.05 then the model is described not to fit the data.

### 3.4.3.2 Hosmer-Lemeshow test

Hosmer-Lemeshow test is another approach of determining the goodness of fit of the data been divided into various subgroups with similar predicted probabilities. The test seeks to find out if the proportion of events observed in the subgroup is the same as the predicted probabilities using Pearson chi square.

The Hosmer-Lemeshow goodness of fit is calculated as follows;

$$\chi^2_{HL} = \sum_{i=1}^{G}\frac{(O_i - N_i\bar{\pi}_i)}{\left(N_i\bar{\pi}_i(1-\bar{\pi}_i)\right)} \tag{3.3}$$

where:

$G$ is the number of subgroups;

$O_i$ is the number of responses in the $i^{th}$ group;

$N_i$ is the number of observation in the $i^{th}$ group;

$\bar{\pi}_i$ is the average predicted probabilities in the $i^{th}$ group.

The test statistics approaches a chi-square distribution with $G-2$ degree of freedom.

Small p-values indicate a poor fit to the data whiles large p-values which is 0.05 or more indicate otherwise.

### 3.4.3.3 Pearson Residual

Person Residual is the difference between the observed and the predicted probabilities from the model divided by binomial standard deviation from the predicted probabilities which is use to correct the uneven variation in the actual residuals .Mathematically it can be expressed as;

$$PR_i = \frac{Y_i - \hat{\pi}_i}{\left(\hat{\pi}_i(1-\hat{\pi}_i)\right)^{\frac{1}{2}}} \tag{3.4}$$

The Pearson residual test statistic follow a chi-square distribution with $N-k$ and it is given by;

$$\chi^2 = \sum_i \frac{(y_i - \hat{\pi}_i)^2}{\hat{\pi}_i(1-\hat{\pi}_i)} \tag{3.4.1}$$

### 3.4.3.4 Deviance Residual

Deviance residual is used to determine if specific observations properly fit the model. The residual deviance is positive when the specified observation is greater than the predicted probability and negative otherwise. It is expressed mathematical as below;

$$RD_i = \pm\left\{2\left[y_i log\left(\frac{y_i}{\hat{\pi}_i}\right) + (1-y_i) log\frac{(1-y_i)}{1-\hat{\pi}_i}\right]\right\}^{\frac{1}{2}} \tag{3.5}$$

The test statistics of the Deviance is expressed as;

$$D = -2\sum_{i=1}^{n}[\hat{\pi}_i log(\hat{\pi}_i) + (1-\hat{\pi}_i) log(1-\hat{\pi}_i)]$$

### 3.4.3.5 Wald Statistics

Wald statistic is the ratio of the parameter estimated square over the square of the estimated standard error of the estimated parameter. It is asymptotically distributed as a chi-square with one degree of freedom. Normally, it assumes that the individual predictor variable have a significant influence on the response.

$$W_i = \frac{\beta_i^2}{SE_{\beta i}^2} \tag{3.6}$$

where,

$SE_{\beta_i}^2$ is the standard error of the $i^{th}$ estimated parameter which is the square root of the $i^{th}$ diagonal element of the estimated covariance matrix.

### 3.4.3.6 Linearity test

One of the ways of assessing the linearity assumption of the log odds of the response variable and the continuous covariate is the use of locally weighted scatter plot smoothing (LOESS). It combines the simple form of the least square regression with flexible form of non-linear regression. It fits simple models to localised subsets of the data to build up a function that would describe the deterministic component of changes in the data points. A functional form of the data is not required to be specified to be able to fit a model and hence, it is able to show complex relationship in a data that could be ignored.

The model can be expressed as;

$$y_i = m(x_i) + \varepsilon_i \tag{3.7}$$

Where $m$ is a regression function which is not specified and $\varepsilon_i$ is the random error. The LOESS method is used to estimate the function $m$

### 3.4.4 Parameter estimation of logistic regression

The parameters of the logistic regression model were estimated using maximum likelihood and Bayesian estimation method. Below is a detailed description of the methods.

#### 3.4.4.1 Maximum likelihood estimation of parameters of logistic regression

Maximum likelihood estimation is the method used to find the parameters which maximize the likelihood function. It is use to find the smallest possible deviance between the observed and the predicted values which is stated as;

$$\boldsymbol{\pi_i} = \frac{\mathbf{1}}{\mathbf{1} + \boldsymbol{e}^{-(\sum_{k=0}^{K} \boldsymbol{\beta_i x_{ik}})}} \qquad i = 1,2, \ldots . n \tag{3.8}$$

The probability distribution of the response is represented as;

$$f(y_i) = \pi^{y_i}(1-\pi)^{1-y_i} \tag{3.9}$$

Likelihood function can be expressed as

$$L(y_i, \beta_o, \beta_1, \ldots, \beta_k) = \prod_{i=1}^{n} \pi^{y_i}(1-\pi)^{1-y_i} \tag{3.10}$$

Taking $log$ of both side

$$l(y_i, \beta_o, \beta_1, \ldots, \beta_k) = \log\left(\prod_{i=1}^{n} \pi^{y_i}(1-\pi)^{1-y_i}\right)$$

$$= \sum_{i=1}^{n} y_i \log \pi + \sum_{i=1}^{n} (1-y_i)\log(1-\pi)$$

$$= \sum_{i=1}^{n} y_i \log \frac{\pi}{1-\pi} + \sum_{i=1}^{n} \log(1-\pi)$$

Differentiating the above equation with respect to $\beta_k$ yields and using $\pi$ as in equation (3.8):

$$\frac{\partial l(\beta)}{\partial \beta_k} = \sum_{i=1}^{n} y_i x_{ik} - n_i \cdot \frac{1}{1 + e^{\sum_{k=0}^{K} x_{ik}\beta_k}} \cdot \frac{\partial}{\partial \beta_k}\left(1 + e^{\sum_{k=0}^{K} x_{ik}\beta_k}\right)$$

$$= \sum_{i=1}^{n} y_i x_{ik} - n_i \cdot \frac{1}{1 + e^{\sum_{k=0}^{K} x_{ik}\beta_k}} \cdot e^{\sum_{k=0}^{K} x_{ik}\beta_k} \cdot \frac{\partial}{\partial \beta_k} \sum_{k=0}^{K} x_{ik}\beta_k$$

$$= \sum_{i=1}^{n} y_i x_{ik} - n_i \cdot \frac{1}{1 + e^{\sum_{k=0}^{K} x_{ik}\beta_k}} \cdot e^{\sum_{k=0}^{K} x_{ik}\beta_k \cdot x_{ik}}$$

The maximum likelihood can be obtained by setting each $k+1$ equations to zero in the equation above and solving for each $\beta_k$.Where $k$ is the number of independent variables specified in the model.

### 3.4.4.2 Variable Selection

The Akaike information criterion (AIC) is used to select variables for the model. It is a measure of relative quality of statistical models hence given a set of models, AIC estimates the quality of the model relative to the other models hence it is use to select the best model. Mathematically it is express as;

$$AIC = 2k - 2\log L$$

$$= 2k + Deviance$$

where $L$ is the maximum value of the likelihood function.

### 3.4.4.3 Bayesian Estimation

Bayesian method of estimation is flexible which does not demand compliance with assumption as that of the case of maximum likelihood approach used by classical techniques. The flexibility of the Bayesian method is enhanced by the use of Markov chain Monte carlo (MCMC) method which is used as a base sampling (Acquah , 2013). It has become a method for fitting various

non-linear regression models. The method is not often used because very minimal is understood about its concept in Bayesian analysis and its application in logistic regression.

In the Bayesian analysis framework, there are three component associated with parameter estimation, these are; prior distribution, likelihood function and posterior distribution.

The likelihood function can be expressed as,

$$L(y_i, \beta_o, \beta_1, \dots, \beta_k) = \prod_{i=1}^{n} \pi^{y_i}(1-\pi)^{1-y_i}$$

#### 3.4.4.4 Prior Distribution

There are two types of prior distribution which are often used. They are informative prior and Non-informative prior. Informative prior distribution are used if the likelihood values of the unknown parameters are known. For non-informative prior, they are used if little is known about the coefficient values of the parameter. In the study, the prior used is the multivariate normal prior on $\beta$ expressed as;

$$\beta_j \sim N(\mu_j, \sigma_j^2) \tag{3.11}$$

The common choice of μ is zero and $\sigma$ is set to 1000 which is chosen large enough to be considered as a non-informative prior (Acquah, 2013).

#### 3.4.4.5 Posterior Distribution

The posterior is obtained by multiplying the prior and the likelihood function as given by;

$$P(\beta/y, x) = \prod_{i=1}^{n} \pi^{y_i}(1-\pi)^{1-y_i} \prod_{j=1}^{K} \frac{1}{\sqrt{2\pi\sigma_j}} exp\left\{\frac{-1}{2}\left(\frac{\beta_j - \mu_j}{\sigma_j}\right)^2\right\} \tag{3.12}$$

The above equation has no closed form hence Gibbs sampler was used to solve and approximate the properties of the marginal posterior distribution of each parameter.

### 3.4.4.6 Markov Chain Monte Carlo (MCMC)

It is a computational method in Bayesian estimation which is used to obtain sequence of random samples from a probability distribution. In this method, values of the parameters $\beta$ are obtained from an approximate distribution and then correcting those drawn to better approximate the desired posterior distribution. For a Markov chain, a sample at $t+1$ depends on the sample at $t$. There are two MCMC algorithm that are used; Gibbs sampler and Metropolis Hastings algorithm (Medova , 2008). In this thesis, the Gibbs sampler was used.

### 3.4.4.7 Gibbs sampling

The Gibbs sampler produces a sequence of samples from the joint distributing of two or more random variable. It can be used to sample from the joint distribution if the full conditional distribution for each parameter is known. It is uses an iterative procedure to sample from the posterior

The algorithm is as follows;

0. Set an arbitrary starting value $\{\beta_0^0, \ldots . \beta_k^0\}$
1. Draw $\beta_0^1$ from the $P(\beta_0/\beta_1^0, \ldots \beta_k^0, y, x)$
2. Draw $\beta_1^1$ from the $P(\beta_1/\beta_0^1, \beta_2^0 \ldots \beta_k^0, y, x)$
3. Draw $\beta_2^1$ from the $P(\beta_2/\beta_0^1, \beta_1^1, \beta_3^0 \ldots \beta_k^0, y, x)$

.
.
.

k. Draw $\beta_k^1$ from the $P(\beta_k/\beta_0^1, \beta_1^1, \beta_3^1 \ldots \beta_{k-1}^1, y, x)$

This complete iteration 1 of the Gibbs sampler, hence $\{\beta_0^1, \ldots . \beta_k^1\}$ is obtained.

For the second iteration,

1. Draw $\beta_0^2$ from the $P(\beta_0/\beta_1^1, \ldots \beta_k^1, y, x)$
2. Draw $\beta_1^2$ from the $P(\beta_0/\beta_0^2, \beta_2^1 \ldots \beta_k^1, y, x)$
3. Draw $\beta_2^2$ from the $P(\beta_2/\beta_0^2, \beta_1^2, \beta_3^1 \ldots \beta_k^1, y, x)$

.
.
.
k. Draw $\beta_k^2$ from the $P(\beta_k/\beta_0^2, \beta_1^2, \beta_3^2 \ldots \beta_{k-1}^2, y, x)$

This complete iteration 1 of the Gibbs sampler, hence $\{\beta_0^2, \ldots . \beta_k^2\}$ is obtained

The procedure continues until $T$ iterations are obtained and we have $\{\beta_0^T, \ldots . \beta_k^T\}$

#### 3.4.4.8 Burn in

The time it takes for a markov chain to converge depends on the starting point, however in practice, certain number of draws are thrown out which is known as burn in. This makes the draw closer to the stationary distribution and less dependent on the starting point.

### 3.5 Classification tress

This method is a form of decision tress which is based on the repeated partitioning of data to become homogenous in other to estimate the conditional probabilities of the outcome given predictor variables. In situations where the response variable is discrete, a classification tree is used and when the response variable is continuous, a regression tree is used (De'ath & Fabricius ,2000).

The main components of a classification tree are; nodes and branches. The important process for constructing a classification tree are splitting, stopping and pruning.

### 3.5.1 Basic Definitions used in classification tree

**Nodes**: The classification tree has three types of nodes which are root, internal and leave node. Root nodes result in the subdivision of observation into two or more mutually exclusive subsets. The internal node have top part of it connected to the parent node whiles the bottom

part is connected to the child or leave node. The leave node is the representation of the final results of the events of the classification three.

**Branches:** Branches represent chance outcomes that originate from root nodes and internal nodes. A classification tree model is formed using an ordered branches in which each path from the root node through internal nodes to a leaf node represents a classification decision rule.

### 3.5.2 Tree Construction

The classification tree is constructed using the steps below;

1. In the development of classification tree, the feature with the highest information gain is selected to be root node.

2. For each value of the feature at the root node create new descendant of node

3. Sort training examples to the leaf nodes

4. When the training examples are perfectly classified, then **stop,** else iterate over the new leave nodes

There are various algorithm for constructing classification trees, the commonly known ones are ID3, C4.5 and CART.

### 3.5.3 ID3

This is a tree algorithm which is constructed in two phases: tree building and pruning. It uses information gain to choose the splitting attribute and accept only categorical attribute in building a tree model. It has bad performance when there are noise (Peng, Chen, & Zhou, 2009).

To build the tree, information gain measured by entropy is calculated for all the attributes and the attributes with the highest information gain is assigned the root node.

Continuous attributes are handled by discretizing or directly considering the values to find the best split point by taking the threshold value on the attribute. ID3 does not support pruning

### 3.5.4 C4.5

This algorithm is base also base on that of Hunt's (Kohavi & Quinlan, 2002). It handles both continues and categorical attributes in building the tree. For it to be able to handle continuous attributes,C4.5 splits the attributes values into two partition based on the selected threshold as one child and the other as another. It uses Gain ratio as defined in (3.8.6.1) as an attribute selection measure to build the tree. It removes the biasness of information gain where there are many outcome values of an attribute. It also handles missing attribute values.

### 3.5.5 CART

This algorithm is also based on Hunt's which also handles both continuous and categorical attributes to construct the decision tree. CART uses Gini index for selecting attributes for the tree .This method does not use probability assumptions unlike the other methods. CART produces binary split and also use cost complexity pruning to eliminate unreliable branches from the decision tree to improve accuracy. It can handle missing values (Lewis, 2007).

### 3.5.6 Splitting Criterion

In construction of a decision tree, it is important to choose a split that provide the important information concerning a class label or predominantly of one class. This is to obtain nodes that are pure.

Impurity of nodes is the measure in which a node is completely homogenous. There are many impurity measure (splitting criteria) used for classification tree, these are information gain, Gini index, Gain ratio etc. For the regression tree, the splitting criteria used is the sum of square about the group mean or the sum of the absolute deviation about the median.

### 3.5.7 Information gain

It is the change of entropy. It is use to determine which attribute in a given training feature is useful for discriminating between the classes and also provide information on how important a given attribute is to enhance ordering in the nodes of a decision tree.

Information gain calculates the entropy of examples after a split on an attribute and subtract it from the entropy of the examples at the node of splitting.

Mathematically, information gain is defined as ;

$$Gain(E, Attribute) = Entropy(X) - \sum_{v \in Value(X)} \frac{|X_v|}{|X|} Entropy(X_v) \quad (3.13)$$

Or

$$Gain(G, Attribute) = Gini(X) - Gini(X)_{split}$$

where $Gini(X)$ is expressed in (3.5.3.7)

#### 3.5.7.1 Gain Ratio

The Information gain ratio is the ratio between the information gain and the intrinsic value (IV) .It biases the decision tree by considering attributes with large number of diverse values which solve the setback in information gain

$$IV(E, Attribute) = -\sum_{v \in Value(A)} \frac{|X_v|}{|X|} \log_2 \frac{|X_v|}{|X|}$$

$$GR(E, Attribute) = \frac{Gain(E, Attribute)}{IV(E, Attribute)}$$

#### 3.5.7.2 Entropy

Entropy is a measure of uncertainty (impurity) associated with the attribute. The entropy E of a discrete random variable $X = x_1, \ldots x_n$ and probability $p(x_i)$ is defined as;

$$Entropy(X) = -\sum_{i=1}^{C} p(x_i) \log_2 p(x_i) \tag{3.14}$$

Where c is the number of classes and $p(x_i) \log_2 p(x_i)$ is assumed to be zero if $p(x_i) = 0$ or $p(x_i) = 1$

Note that if Entropy is low then information gain is high

#### 3.5.7.3 Gini index

Gini index is the probability that a random sample is been classified correctly. The aim of Gini index is to reduce misclassification rate. Mathematically it is expressed as;

$$Gini(X) = 1 - \sum_{i=1}^{C} (p(x_i))^2 \tag{3.15}$$

When the $node\ \ p$ is partition into $k$, the quality of the split is stated as;

$$Gini(X)_{split} = \sum_{v \in Value(X)} \frac{|X_v|}{|X|} Gini(X)$$

### 3.5.8 Pruning

It is a method of removing certain portion of the decision tree which have no significant influence and provide little power in classifying labels which at the end improve prediction accuracy and reduces overfitting. There are several methods for pruning, the commonly used ones are reduced error and cost complexity pruning.

#### 3.5.8.1 Reduced error pruning

Reduced error pruning commence at the leave where each node is assigned the majority class. When the prediction accuracy does not change then the tree is maintained. This method of pruning is simple and fast (Patel & Upadhyay, 2012).

### 3.5.8.2 Cost complexity pruning

This method generate series of trees $T_o, \dots T_{max}$ where $T_o$ is the initial tree and $T_{max}$ the maximum number of tree. For any subtree, the complexity ($\lceil \check{T} \rceil$) is defined as the number of terminal nodes in $T$. Then the cost complexity ($R_\alpha(T)$ ) measures the penalised on the resubstitution error rate. The resubstitution error rate is not an appropriate measure for obtaining subtrees because it always favour the bigger ones. Including a complexity penalty to resubstitution error turns to favour small trees hence reducing the cost complexity when pruning the tree. Mathematically, it can be expressed as;

$$R_\alpha(T) = R(T) + \alpha\lceil \check{T} \rceil$$

where $\alpha \geq 0$ is the complexity parameter and $R(T)$ is the resubstitution error rate. Note that as $\alpha$ approaches infinity, the tree of size 1 (i.e single root node) becomes the biggest tree.

For a pre-selected $\alpha$ then an identified subtree $T(\alpha)$ which minimize the cost complexity would be obtained such that

$$R_\alpha\big(T(\alpha)\big) = \max_{T \leq T_{max}} R_\alpha(T)$$

## 3.6 Random forest

Random forest are ensemble methods which means they are made up of other small models in which predictions are obtained by combining the output of these smaller models which are the classification or regression tress. It is normally useful for exploratory analysis, detection of interaction and non-linearity without initially specifying them in the model.

The training algorithm for random forest uses the techniques of bootstrap aggregating or bagging. For this method, observations for each trees are selected randomly likewise the variables

$$H(x) = \frac{\underset{y}{\operatorname{argmax}} \sum_{i=1}^{K} I(h_i(x_i) = y)}{K} \tag{3.16}$$

where;

$H(x)$ is the class label through majority voting,

$K$ is the number of decision tress,

and $h_i$ is the $i^{th}$ tree of the random forest.

In other to fine tune the forest, two parameters must be considered;

- Number of trees that would correspond to a stable classifier
- Number of random variables used in each tree

Variable importance are estimated based on margin of cases which is defined as proportion of votes for true class minus maximum proportion of votes. It measures the association between a given variable and its classification results.

$$VI_j^t = \sum_{i=1}^{N} \frac{\beta_t I(y_i = y_i^t)}{|\beta_t|} - \sum_{i=1}^{N} \frac{\beta_t I(y_i = y_{i,a}^t)}{|\beta_t|} \tag{3.17}$$

where;

$\beta_t$ is the out-of-Bag(OOB) sample for tree $t \in (1,2, \ldots ntree)$,

$y_i^t$ is the $i^{th}$ predicted class before permutation in $t$,

$y_{i,a}^t$ is the $i^{th}$ predicted class after permutation in $t$.

The variable importance for variable $j$ in the Random Forest (RF) is;

$$VI_j = \sum_{t=1}^{ntree} \frac{VI_m}{ntree}$$

The merit of RF over logistic regression is that essential variables can automatically be selected no matter the number used initially. It does not use stepwise regression to select variables. It uses 63% of cases to construct each tree while the remaining 37% of case which is out-of-bag (OOB) is used to evaluate the performance of the tree (Steinberg, Golovnya, & Cardell, 2004).

### 3.6.1Random forest construction

1. Obtain $n\,tree$ which are independent and identically distributed bootstrap samples from the original data with $N$ available case. The samples are obtained with replacement.
2. In growing each tree, $N-1$ bootstrap samples are obtained and the remaining cases called out of bag (OOB) are used to validate the tree.

   **Out of Bag Estimate**

   Out of bag estimate is the average prediction error on each training sample which use only trees which did not have those observation in their bootstrapped sample. It is a method which is used to measure the error of prediction of the Random forest and other machine learning models which makes cross validation not necessary in these models.

3. If there are $M$ variables, $m < M$ random specified variables at each node are chosen such that the best split of $m$ is use to split the node.The number of $m$ variables remain the same during the forest growing process. But the variables might vary when growing the forest.
4. Prediction are obtained by averaging the majority vote obtained from the $n\,tree$
5. Error rate are obtained by averaging the prediction obtained by the OOB data which were used during the bootstrap iteration process

### 3.6.2 Bagging

Bagging is obtaining repeated samples from a dataset in other to produce $C$ different bootstrapped training dataset where the $C^{th}$ bootstrapped training set is trained to obtain $\hat{f}^c$. Each separate prediction model are averaged out to obtain a low variance statistical learning model.

$$\hat{f}(x) = \frac{1}{C}\sum_{c=1}^{C} \hat{f}^c(x) \tag{3.18}$$

### 3.7 Cost sensitive modelling

In this method, the loss matrix is incorporated into the classification tree and random forest. These are used to weight misclassification differently. In medical diagnosis, false positive (type I error) and false negative (type II error) have different cost. The classification takes into consideration how much to penalise each incorrect classification in a given choice split. The cost matrix $L$ is stated as;

$$L = \begin{pmatrix} C_{TP} & C_{FP} \\ C_{FN} & C_{TN} \end{pmatrix}$$

where;

$C_{FP}$ is the cost of false positive,

$C_{FN}$ is the cost of false negative,

$C_{TP}$ is the cost of true positive,

and $C_{TN}$ is the cost of true negative.

The cost matrix is use to adjust the way split are chosen. The tree is constructed in such a way that it uses the splitting criteria which minimize the misclassification cost rather than

minimizing the entropy. It can also be used to tune the threshold on the probability of classification.

The theory of cost sensitive learning as reported in Elkan (2001), Zadrozny and Elkan (2001) describe how misclassification cost plays important role in cost sensitive learning algorithm. Since the study is focused on binary classification, let $i, j$ represent the two classes. For an instance to be classified to minimise the expected cost, then expected cost of predicting an instance $x$ to belong to class $i$ can represented as;

$$R(i/x) = \sum_j P(j/x)C(i,j) \ , i = 0{,}1 \ and \ j = 0{,}1 \tag{3.19}$$

where;

$P(j/x)$ is the probability that an instance $x$ belong to $j$,

$C_{TP}$ is equal to $C(1{,}1)$,

$C_{TN}$ is equal to $C(0{,}0)$,

$C_{FN}$ is equal to $C(0{,}1)$,

and $C_{Fp}$ is equal to $C(1{,}0)$.

The classifier would classify an instance into a positive class if and only if ;

$$P(0/x)C(1{,}0) + P(1/x)C(1{,}1) \leq P(0/x)C(0{,}0) + P(1/x)C(0{,}1)$$

Or

$$P(0/x)C_{FP} + P(1/x)P(0/x)C_{TP} \leq P(0/x)C_{TN} + P(1/x)C_{FN}.$$

In this thesis, we assumed that $C_{TP}$=$C_{TN}$=0

Therefore,

$$P(0/x)C_{FP} \leq P(1/x)C_{FN}$$

note that $P(0/x) = 1 - P(1/x)$, hence;

$$\big(1 - P(1/x)\big)C_{FP} \leq P(1/x)C_{FN}$$

$$P(1/x) \geq \frac{C_{FP}}{C_{FN} + C_{FP}} \tag{3.20}$$

From above, it can be observed that there is a threshold $\frac{C_{FP}}{C_{FN}+C_{FP}}$ for the classifier to classify an instance into a positive class. When the cost of misclassification is equal, that is $C_{FP} = C_{FN} = 1$ the threshold is 0.5

### 3.7.1 Modifying the classification of the predicted probability scores obtained from logistic regression to include unequal misclassification cost

In other to determine the optimal decision threshold for the predicted probability scores of logistic regression the Receiving Operation characteristic (ROC) expressed in (3.8.1) analysis was used. Also the relative cost of the false negative and false positive was considered.

$$Total\ expected\ cost(C_{Total}) = \pi_o(1 - SP)C_{FP} + \pi_1(1 - SN)C_{FN}$$

where;

$SP$ is the Specificity,

$SN$ is the sensitivity,

$\pi_o$ is prior probability of negative cases,

and $\pi_1$is the prior probability of positive cases.

Since Sensitivity (SN) is a function of false positive rate of the ROC curve. The total expected cost is equivalent;

$$C_{Total} = \pi_o FPRC_{FP} + \pi_1[1 - ROC(FPR)]C_{FN}$$

For a minimal cost, the optimal cut off is obtained by differentiating with respect to $FPR$ and setting it to zero;

$$\frac{dROC(FPR)}{dFPR} = \frac{\pi_o}{\pi_1}\frac{C_{FP}}{C_{FN}}$$

Alternately, the optimal cut off could be obtained by evaluating the various cut off corresponding to $(1 - SP)$ in which total expected cost is minimum.

### 3.8 Sampling

Let $(X, Y)$ be the original unbalanced training sample and $(\aleph, \gamma)$ be the balanced sample, which means$(\aleph, \gamma) \subseteq (X, Y)$. Suppose $s$ is a binary variable for selection which takes values 1 if the point is in $(\aleph, \gamma)$ and 0 if otherwise, then, the posterior distribution of the training data for the balanced data and that on the original data can be obtained.

For $\beta = p(s = 1/y = 0)$ is the probability of selecting negative instances with under sampling, $p = p(y = 1/x)$ is the posterior probability of the positive class on the original dataset and $p_s = p(y = 1/x, s = 1)$ is the posterior after sampling. Then

$$p_s \frac{p}{p + \beta(1 - p)} \qquad 3.21$$

which can also be expressed as;

$$p = \frac{\beta p_s}{\beta p_s - p_s + 1}$$

To balance an unbalanced class distribution corresponding to the cases then, $\beta = \frac{p(y=1)}{p(y=0)} \approx \frac{N^+}{N^-}$

where $N^+$the number of positive instances, $N^-$ is the number of negative instances in the dataset and $\frac{N^+}{N^-}$ is the minimum value for $\beta$.

When $\beta = 1$ , all the negative instances are used for training whiles for $\beta < 1$ a subset of the negative instances are used for training. For $\beta < \frac{N^+}{N^-}$ , we would have more positive than negative cases.

In under sampling, the number of negative is $N_s^- = \beta N^-$ while the number of positive is $N_s^+ = N^+$

In this thesis, the under sampling method was used to vary the ratio of the gonorrhea negative cases to that of the gonorrhea positive (i.e 40:60, 50:50 and 60:40).

### 3.9 Performance Measure

Assessing model's performance is an essential part in machine learning in which it becomes impossible to compare models without an evaluation. The performance measure which were used are; confusion matrix, which identifies the errors in the classifiers; g-mean, which combines the performance of two classifiers; ROC curve; F-measure and Kappa. The ROC is used to determine the discriminative power of the classifiers. Kappa is used to determine the agreement between the classifiers used. F-measure is used to determine the classifiers performance to predict the positive class whiles the g-mean is used to determine if the negative class has been over fitted or the positive class has been under fitted.

#### Confusion Matrix

The Confusion matrix basically show the type of classification error a classifier make as shown below:

| Predicted class | | |
|---|---|---|
| + | - | Actual Class |
| TP | FN | + |
| FP | TN | - |

TP is true positive,

TN is true negative,

FP represent the false positive,

and FN represent the false negative.

**Accuracy:**

It is usually defined over all the classification error and it is calculated as;

$$Accuracy = \frac{TP + TN}{TP + TN + FN + FP}$$

**Recall:**

It a measure of proportion of true positives verses examples which are classified by a system as positive (Borges, 2016). This is also known as sensitivity

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

**Specificity:**

It is a measure of proportion of truly negative verse examples which are classified by a system as negative.

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

**Precision:**

It is a measure of proportion of positive which are classify by a system as truly positive

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

**Negative predictive value:**

It is a measure of proportion of negative which are classify by a system as truly negative.

$$Negative\ Predictive\ value = \frac{TN}{TN + FN}$$

**F-measure:**

It is a measure which combines the trade-offs of recall and precision and produces a single outcome which is a measure the goodness of a classifier in the presence of uncommon event (Sokolova, Japkowicz, & Szpakowicz, 2006).

$$F - measure = \frac{2 * Recall * Precesion}{Recall + Precision}$$

**Geometric-mean:**

It measure of the product of the prediction accuracy of sensitivity and specificity. It indicate the classification performance on the majority and minority class (Bekkar, Djemaa, & Alitouche , 2013).

$$G - mean = \sqrt{sensitivity * specificity}$$

**Kappa:**

Kappa statistics is a measure that compares the observed agreement with the expected agreement. It adjust the accuracy by accounting for the correct prediction by chance only. The maximum value is one which indicate perfect agreement between the predictions of the models. Kappa values less than one indicate imperfect agreement (Biswas, 2006).

$$Kappa = \frac{P_o - P_e}{1 - P_e}$$

where;

$P_o = \frac{TP+TN}{N}$,

$P_e = \frac{1}{(N)^2}[(TP + FN)(TP + FP) + ((FP + TN)(FN + TN)]$,

and $N = TP + TN + FP + FN$.

### 3.9.1 Receiving Operating Characteristics (ROC)

The ROC curve is a plot that classify the trade-off between the True Positive rate (Sensitivity) and False Positive rate (1-Speficificity) across a series of cut-off points. The area under the ROC curve is considered as an effective measure of intrinsic validity of the diagnostic test. The curve is useful in; evaluating the discriminatory ability of the test, compare efficiency of two or more medical test for assessing the same disease and finding optimal cut-off point to at least misclassify disease and non-disease.

There are parametric and non-parametric methods for obtaining the ROC curve. Non-parametric methods uses the trapezoidal rule to obtain the area under the ROC curve .Other methods use Man-Whitney statistic also known as Wilcoxon rank-sum to calculate the area under the ROC curve.

Parametric methods are used when the statistical distribution of the test values of disease and non-disease is known. Often, binomial distribution is use which gives a smooth ROC curve.

# CHAPTER FOUR

## DATA PRESENTATION AND ANALYSIS

This chapter presents the analysis of cost insensitive learning and sensitive learning for predicting Gonorrhea infection status and discussion of results obtained.

Percentage and frequency was used to describe both the training and testing data. Logistic regression was used to fit a model for the training data in which the parameters were estimated using the maximum likelihood method. The variables were selected into the model using Akaike information criterion (AIC).Diagnostic measures such as log likelihood ratio, residual deviance, linearity test etc. were accessed. In other to use the model as a classifier, the probability distribution of the predicted score was accessed. An optimal cut off 0.5 which assumed equal cost of misclassification was initially used as a threshold for classification. Another optimal cut off 0.26 was obtained by considering unequal misclassification cost and this threshold was used for classification.

The parameters of the logistic regression were also estimated using the Bayesian method and also the model diagnostic were also accessed.

A classification tree model which assumed equal misclassification cost was developed. The variables were selected using information gain. In other to include unequal misclassification cost in the model, a loss matrix was used in the classification tree.

Finally, the random forest model with equal classification cost was use to fit the training data. Also, to include unequal misclassification cost, lost matrix was used to classify the predicted probabilities of the random forest.

The various cost insensitive and sensitive models performance were mainly evaluated using total classification cost and also measures such as accuracy, receiving operating characteristic Area Under Curve(AUC), F-measure, G-mean, type I and type II error were also used .

The model was validated using a testing data using the optimal cut off obtained from the trained data. These threshold were used to classify the probability scores. Cost sensitive and insensitive classification tree and random forest models were evaluated on the testing data. This classifiers were also compared to the laboratory diagnostic methods such gram stain and culture using total classification cost (benefit).They were also evaluated on the other performance measure.

Under sampling method was used to adjust the class distribution, cost sensitive and insensitive method was train on the adjusted dataset and then tested on the hold out data. Total classification cost was used to compare the effect of the class adjustment on the cost sensitive and insensitive method. Data was analysed using R version 3.3.3.

### 4.1 Data and Preliminary Analysis

In the Table 4.1a and 4.1b below, 906 patients, were enrolled from the clinical facilities during the study period, of which 28% were diagnosed of having gonorrhea. The study enrolled more females than males but yet high proportion of the males were diagnosed of the infections than females. Majority (70%) of the patients who participated in the study were 31 years and below and high proportion (31%) of them were diagnosed of gonorrhea.

The study enrolled 581(64%) of patients who were not married as at the time of enrolment. Five percent (5%) of these patients were previously married and are now single either by divorce or deceased of spouse. This is an indication that majority of the patient who were enrolled had never been married.

Majority of clinical presentation were discharge (88%) and pain during urination (49%).Other symptoms which were reported were itching of the genital, foul smell, painful sex, ulcers, bleeding from penis or vagina and pain in penis or vagina.

Regarding sexual behaviour, only 12% of the patients have had more than one sexual partner in the past month of which 51% of them were diagnosed with gonorrhoea. Forty one percent

(41%) of the enrolled patients never used condom during sexual intercourse whiles 59% of the patients at least used condom once during sexual intercourse. Among the patient who always use condom during sexual intercourse 21% of them were diagnose of gonorrhoea.

Only 289(32%) of the patients enrolled had a behaviour of drinking alcohol of which 38% them were diagnosed gonorrhea.

Table 4.1a:Background information of the respondent

| Variable | Total N(%) | Gonorrhea +ve N(%) | Gonorrhea -ve N(%) |
|---|---|---|---|
| **Demographic** | | | |
| Gender | | | |
| Male | 390(43.0) | 169(43.3) | 221(56.7) |
| Female | 516(57.0) | 85(16.4) | 431(83.6) |
| Age | | | |
| 18-24years | 247(27.3) | 70(28.3) | 177(71.7) |
| 25-31years | 385(42.5) | 123(31.9) | 262(68.1) |
| 32-38years | 171(18.9) | 45(26.3) | 126(73.7) |
| 39years and above | 103(11.4) | 16(15.5) | 87(84.5) |
| Marital status | | | |
| Single | 551(60.8) | 161(29.2) | 390(70.8) |
| Previously Married | 30(3.3) | 10(33.3) | 20(66.7) |
| Married | 325(35.9) | 83(25.5) | 242(74.5) |
| **Clinical Presentation** | | | |
| Painful urination | | | |
| Yes | 445(49.1) | 150(33.7) | 295(66.3) |
| No | 461((50.9) | 104(22.6) | 357(77.4) |
| Discharge | | | |
| Yes | 800(88.3) | 228(28.5) | 572(71.5) |
| No | 106(11.7) | 26(24.5) | 80(75.5) |
| Pain in penis or vagina | | | |
| Yes | 256(28.3) | 70(27.3) | 186(72.7) |
| No | 650(71.7) | 184(28.3) | 466(71.7) |
| Foul smell | | | |
| Yes | 214(23.6) | 50((23.4) | 164((76.6) |
| No | 692(76.4) | 204(29.5) | 488(70.5) |
| Painful sex | | | |
| Yes | 129(14.2) | 30(23.3) | 99(76.7) |
| No | 777(85.8) | 224(28.8) | 553(71.2) |
| Bleeding from penis or vagina | | | |
| Yes | 69(7.6) | 21(30.4) | 48(69.6) |
| No | 837(92.4) | 233(25.7) | 604(74.3) |
| Itching of Genital | | | |
| Yes | 162(17.9) | 32(19.8) | 130(80.2) |
| No | 744(82.1) | 222(29.8) | 522(70.2) |
| Ulcers | | | |
| Yes | 53(5.8) | 15(28.3) | 38(71.7) |
| No | 853(94.2) | 239(28.0) | 6149(72.0) |

Table 4.1b:Background information of the respondent

| Variable | Total N(%) | Gonorrhea +ve N(%) | Gonorrhea -ve N(%) |
|---|---|---|---|
| **Sexual Behaviour** | | | |
| Alcohol intake | | | |
| Yes | 289(31.9) | 109(37.7) | 180(62.3) |
| No | 617(68.1) | 145(68.1) | 472(31.9) |
| Use of Condom | | | |
| Never | 375(41.4) | 107(28.5) | 268(71.5) |
| Rarely | 244(26.9) | 74(30.3) | 170(69.7) |
| Most occasion | 147(16.2) | 43(29.3) | 104(70.7) |
| Always | 140(15.5) | 30(21.4) | 110(78.6) |
| Having more than one sexual partner in the pastner month | | | |
| Yes | 112(12.4) | 57(50.9) | 55(49.1) |
| No | 794(87.6) | 197(24.8) | 597(75.2) |

## 4.2 Training Data and Model fitting

The result presented in Table 4.2a and 4.2b is a description of the patient in the training data consisting of 80% of the original data. The description of the characteristic is approximately the same as that of Table 4.1a and 4.1b.

Table 4.2a:Description of Training Data

| Variable | Total N(%) | Gonorrhea +ve N(%) | Gonorrhea -ve N(%) |
|---|---|---|---|
| **Demographic** | | | |
| Gender | | | |
| Male | 318(43.7) | 142(44.7) | 176(55.3) |
| Female | 410(56.3) | 63(15.4) | 347(84.6) |
| Age | | | |
| 18-24years | 199(27.3) | 58(29.1) | 141(70.9) |
| 25-31years | 313(43.0) | 103(32.9) | 210(67.1) |
| 32-38years | 133(18.3) | 31(23.3) | 102(76.7) |
| 39years and above | 83(1.4) | 13(15.7) | 70(84.3) |
| Marital status | | | |
| Single | 446(61.3) | 133(26.8) | 313(73.2) |
| Previously Married | 26(3.6) | 8(30.8) | 18(69.2) |
| Married | 256(35.2) | 64(25.0) | 192(75.0) |
| **Clinical Presentation** | | | |
| Painful urination | | | |
| Yes | 367(50.4) | 128(34.9) | 239(65.1) |
| No | 361(49.6) | 77(21.3) | 284(78.7) |
| Discharge | | | |
| Yes | 646(88.7) | 181(28.1) | 465(71.9) |
| No | 82(11.3) | 24(29.3) | 58(70.7) |
| Pain in penis or vagina | | | |
| Yes | 210(28.8) | 58(27.6) | 152(72.4) |
| No | 518(71.2) | 147(28.4) | 371(71.6) |
| Foul smell | | | |
| Yes | 170(23.4) | 41(54.1) | 129(45.9) |
| No | 558(76.6) | 164(29.4) | 394(70.6) |
| Painful sex | | | |
| Yes | 105(14.2) | 26(24.8) | 79(75.2) |
| No | 623(85.6) | 179(28.7) | 504(71.3) |
| Bleeding from penis or vagina | | | |
| Yes | 55(7.6) | 13(23.6) | 42(76.4) |
| No | 673(92.4) | 192(28.5) | 481(71.5) |
| Itching of Genital | | | |
| Yes | 136(18.7) | 26(19.1) | 110(80.9) |
| No | 592(81.3) | 179(30.2) | 413(69.8 |
| Ulcers | | | |
| Yes | 44(6.0) | 12(27.3) | 32(72.7) |
| No | 684(94.0) | 193(28.2) | 481(71.8) |

Table 4.2b:Description of Training Data continuation

| Variable | Total N(%) | Gonorrhea +ve N(%) | Gonorrhea -ve N(%) |
|---|---|---|---|
| Wartsong | | | |
| Yes | 26(3.6) | 7(26.9) | 19(73.1) |
| No | 702(96.4) | 198(28.2) | 504(71.8) |
| **Sexual Behaviour** | | | |
| Alcohol intake | | | |
| Yes | 239(32.8) | 88(36.8) | 151(63.2) |
| No | 489(67.2) | 117(23.9) | 372(76.1) |
| Use of Condom | | | |
| Never | 301(41.3) | 85(28.2) | 216(71.8) |
| Rarely | 200(27.5) | 63(31.5) | 137(68.5) |
| Most occasion | 114(15.7) | 35(30.7) | 79(69.3) |
| Always | 113(15.5) | 22(19.5) | 91(80.5) |
| Having more than one sexual partner in the partner month | | | |
| Yes | 93(12.8) | 48(51.6) | 45(48.4) |
| No | 635(87.2) | 157(24.7) | 478(75.3) |

Models which were fitted with the training data were logistic regression, classification tree and Random forest.

### 4.2.1 Logistic Regression

The LR model was developed using 14 variables available in Table 4.2a and 4.2b .The reduced final model gave 5 significant variables using stepwise selection procedure and selecting model with the lowest AIC. The variables which remained in the model were Age, Gender, Pain during urination, Condom usage and having more than one sexual partner in the past month. Almost all thes variables were significantly associated with gonorrhea infection except Pain during urination. Similar results were obtained using Markov chain Monte Carlo Bayesian regression as in Table 1a in (Appendix).

Table 4.3: Logistic Regression model using maximum likelihood estimation

| Variable | Estimate | SE | P-value |
|---|---|---|---|
| Intercept | -2.58 | 0.33 | <0.001** |
| Age | | | |
| 25-31years | 0.06 | 0.22 | 0.77919 |
| 32-38years | -0.47 | 0.28 | 0.09632 |
| 39years and above | -1.05 | 0.36 | 0.003858* |
| Male | 1.37 | 0.2 | <0.001** |
| Burning | 0.35 | 0.19 | 0.06252 |
| Condom usage | | | |
| Never | 0.88 | 0.3 | 0.00324* |
| Rarely | 0.81 | 0.31 | 0.00834* |
| Most occasion | 0.7 | 0.34 | 0.03744* |
| More than one sexual partner in past month | 0.67 | 0.25 | 0.00788* |

AIC=770.93 Residual deviance=748.93

In Figure 4.1, both distributions of the predicted probabilities of the positive and negative cases are slightly skewed to the left. The reason for this is because the dataset used consists of majority of the negative class which makes the predicted scores pulled towards a lower number.

When developing models for prediction, it is aimed to make it as accurate as possible. The above density distribution clearly shows that accuracy is not a suitable measurement for the model. Since the prediction of logistic regression uses probabilities, it is necessary to include unequal misclassification which would determine the optimal cut off which would have low total classification cost. This makes it cost effective.

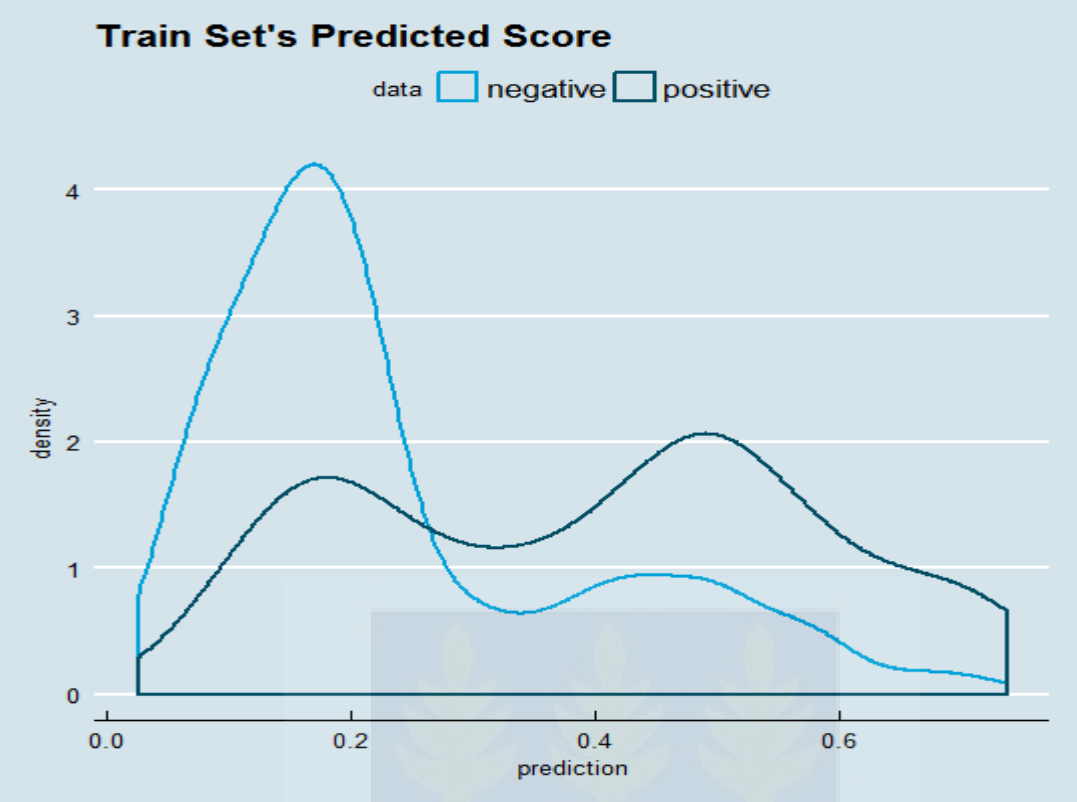


Figure 4.1: Distribution of the predicted probability score of the training data

### 4.2.2 Classification tree

In this classification tree, Gini splitting criterion and assuming equal misclassification cost was used which produced tree in Figure 4.2. Prior probabilities was also assumed to be proportional to the number of positive-Negative gonorrhea infection status in the training data (0.28 and 0.72 respectively). The classification tree had a resubstitution estimate of misclassification of 0.23 and CV classification error rate of 0.25. The tree structure defines five decision profile of gonorrhoea infection in which two predicted positive while three predicted negative infection status of gonorrhea. The 56% of females which ended up in the leave node were predicted to be negative to the disease. Males who had Pain during urination and less than 30 years were predicted to be positive to the infection whiles those who did not show clinical symptom of Pain during urination but had multiple sexual partner in past months were also predicted to be positive.

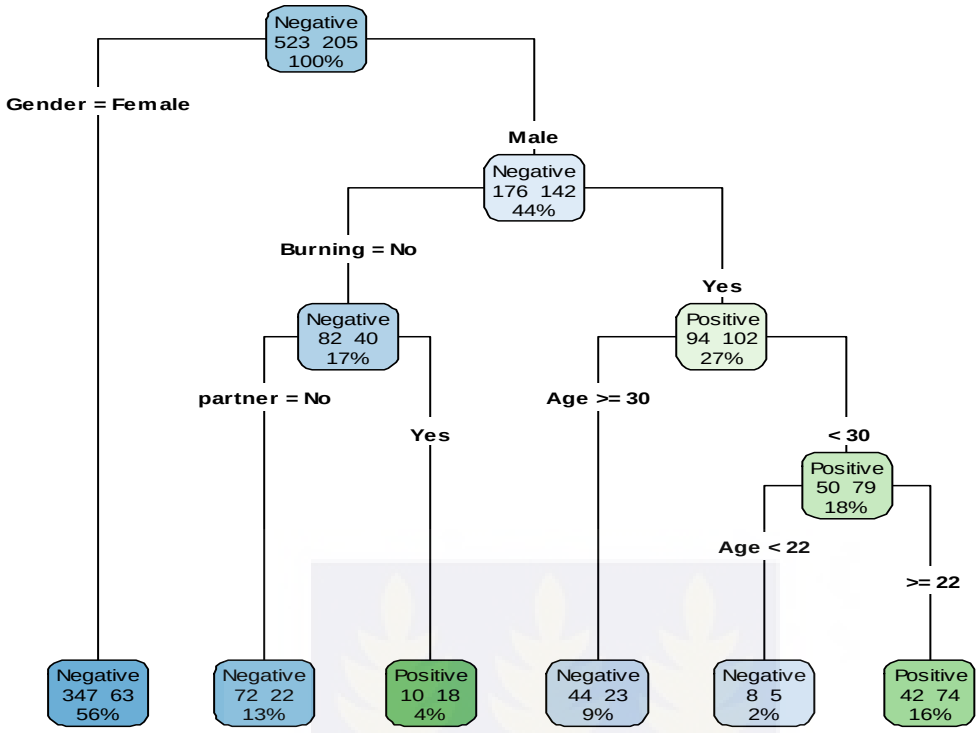


Figure 4.2: Tree structure for gonorrhea data

### 4.2.3 Random Forest

The variable importance plot given in Figure 4.3 shows how important each variable is when classifying the data. These variable importance were measured using the mean decreasing gini which is a measure of how each variable contribute to the purity of each node. The most important five variables were Gender, Age, condom usage, marital status and Pain during urinating. The model gave an out-of-bag (OOB) error rate of 26.4%

Regarding the number of tree to have in the model, use of Figure 1a in Appendix gives an indication. The results indicate that increasing the number of trees decreases the OOB error.

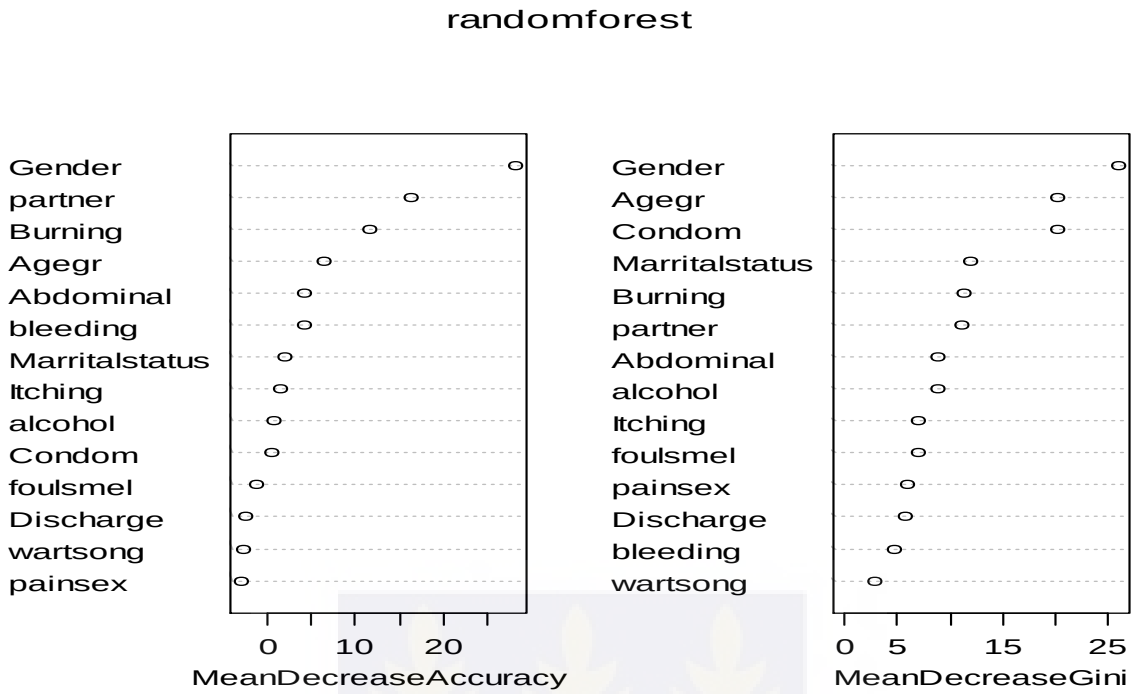


Figure 4.3: Variable importance for Random forest

### 4.3 Cost-Sensitive Models

The cost of false positive and false negative differ in medical diagnosis, hence, in this study misclassification cost was assumed to vary. Therefore, cost of false positive was set at 1 whiles the cost of false negative was adjusted from 1 to 25 shown in Table 4.4. The effect of misclassification error is sated below;

*False negative*

- Spread of disease
- Loss of fertility

*False positive*

- Drug abuse
- Unnecessary financial burden

Table 4.4: Confusion Matrix

| | Confirmed Test(NAAT) | | |
|---|---|---|---|
| | | Positive | Negative |
| Statistical Model | Positive | TP=0 | FP=1 |
| | Negative | FN | TN=0 |

### 4.3.1 Classification of logistic regression predicted probability score to include unequal misclassification cost

From Figure 4.1, the predicted probabilities were skewed to the left hence there is the need to obtain optimal threshold which could be used to classify the model. In Figure 4.4, the false negative was given a high cost than the false positive which yielded a 0.26 probability cut off (threshold).

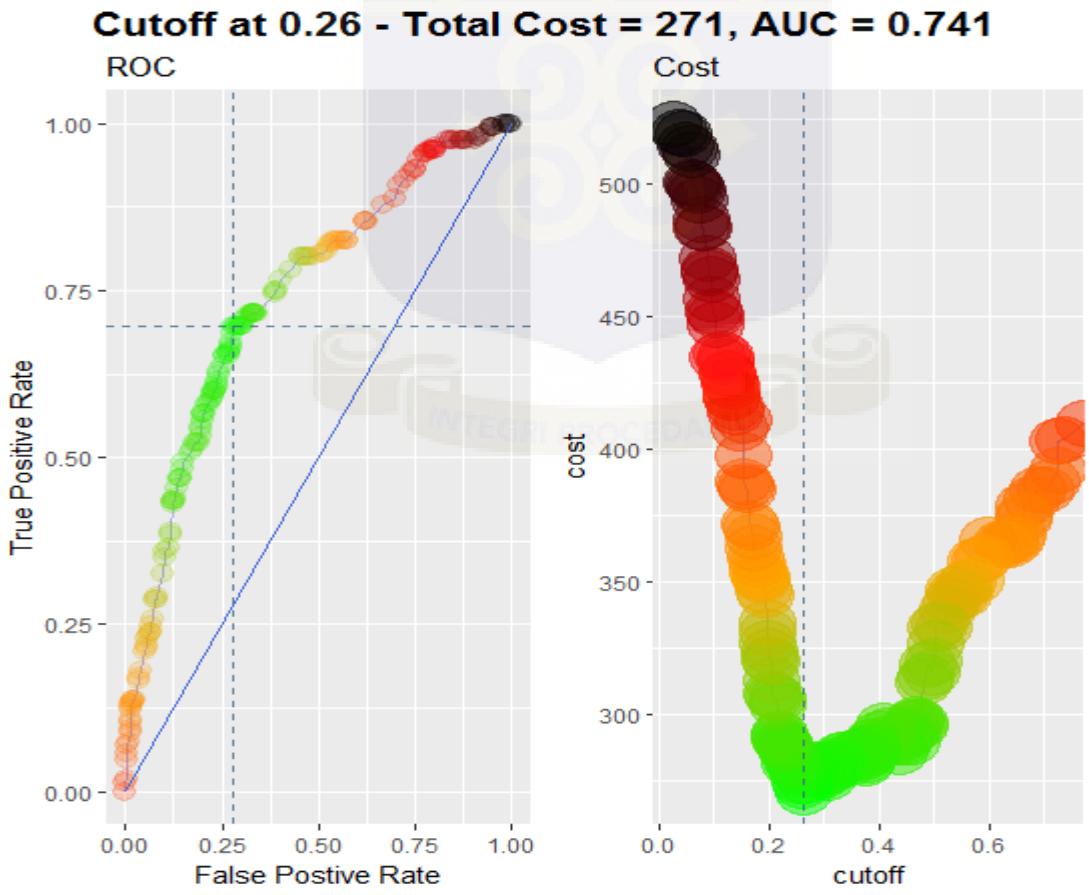


Figure 4.4: Determination of optimal cut off of predicted scores using unequal classification cost

### 4.3.2 Classification tree with unequal misclassification cost

Assumption for the study was that misclassification of a minority class incur a higher cost hence the cost ratio between false negative and false positive was adjusted until decision trees with a reduced type II error was obtained. In Figure 4.5, cost ratio of 1:4 gave seven decision profile. Forty four (44%) of the males who ended up in the leave node were predicted to be negative to gonorrhea whiles for females, 6% of them who ended up in the leave node reported of never or not been consistent of condom usage and also age equal or more than 28years were predicted to be positive to the infection. Also, 4% of the females who were predicted positive in one of the leave nodes were married and partners or they themselves never use condom. This result obtained differs from instances of assuming equal cost of misclassification.

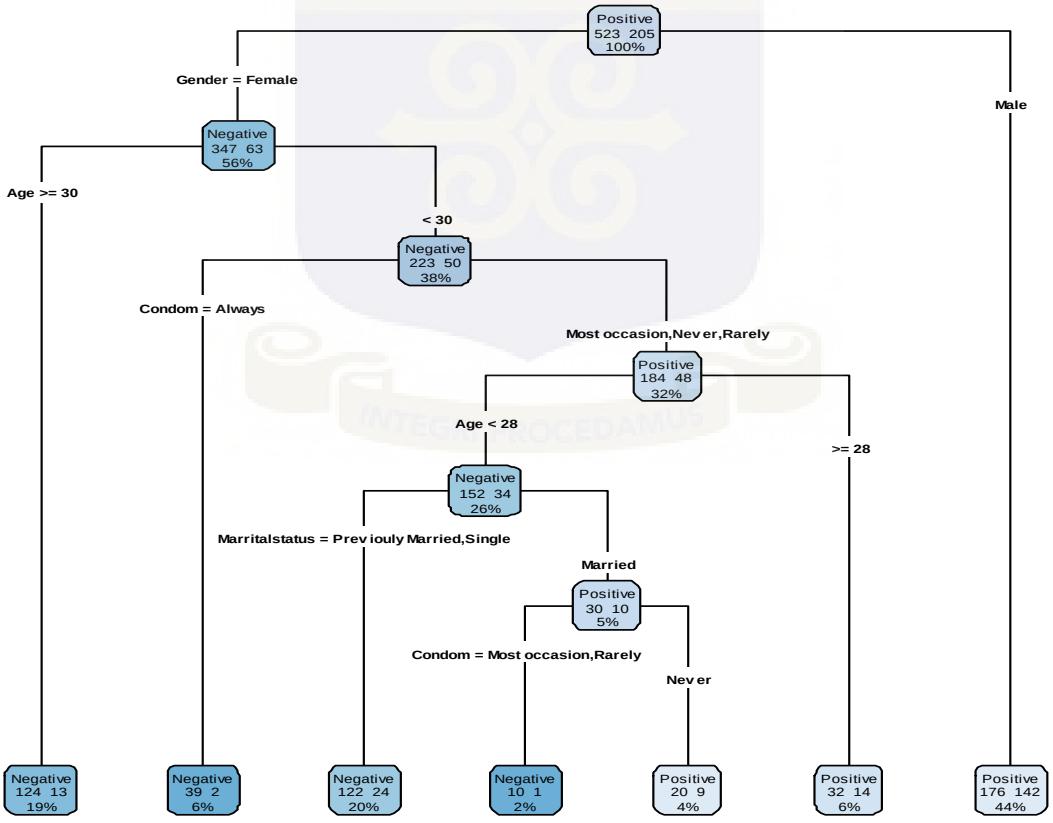


Figure 4.5 : Tree structure for gonorrhea data using a cost ratio of 1:4

### 4.4 Comparing the performance of the models using training Data

Assessing the performance of the various models on the training data indicate that the Random forest had the highest F-measure, AUC and Accuracy. Adding cost to the tree models makes it also better than logistic regression on the training data.

Table 4. 5: Comparing Performance of the various model using training data

| | Accuracy | AUC | F-measure | G-mean |
|---|---|---|---|---|
| LR | 0.74 | 0.74 | 0.43 | 0.56 |
| LR with optimal cut off 0.26 | 0.71 | 0.74 | 0.58 | 0.71 |
| CART | 0.77 | 0.72 | 0.53 | 0.64 |
| CART(Cost Ratio=1:4) | 0.72 | 0.72 | 0.55 | 0.67 |
| RF | 0.86 | 0.91 | 0.71 | 0.76 |
| RF(Cost Ratio=1:4) | 0.84 | 0.91 | 0.74 | 0.84 |

Adjusting the cost of misclassification between the false negative and false positive to obtain an optimal cut off of 0.26 for logistic regression model had a reduction in accuracy than logistic regression with 0.5 threshold which considers equal cost of misclassification. But the F-measure and G-mean was higher. Also, the other classifiers which considered unequal cost of misclassification also had a reduction in accuracy but a high F-measure and G-mean.

### 4.5 Effect of Total classification cost on cost sensitive and insensitive method

In other to calculate the total classification cost, the cost matrix defined in Table 4.4 was used. The results in Figure 4.6 indicated that, the cost sensitive classifiers produced low total classification cost than the cost insensitive classifiers. Adjusting the misclassification cost made the cost sensitive classifiers fall below 1000 cost unit whiles that of the cost insensitive classifiers was above 2000 cost unit. This is an indication that the cost sensitive method have a reduced total number of people who misclassified than the cost insensitive method.

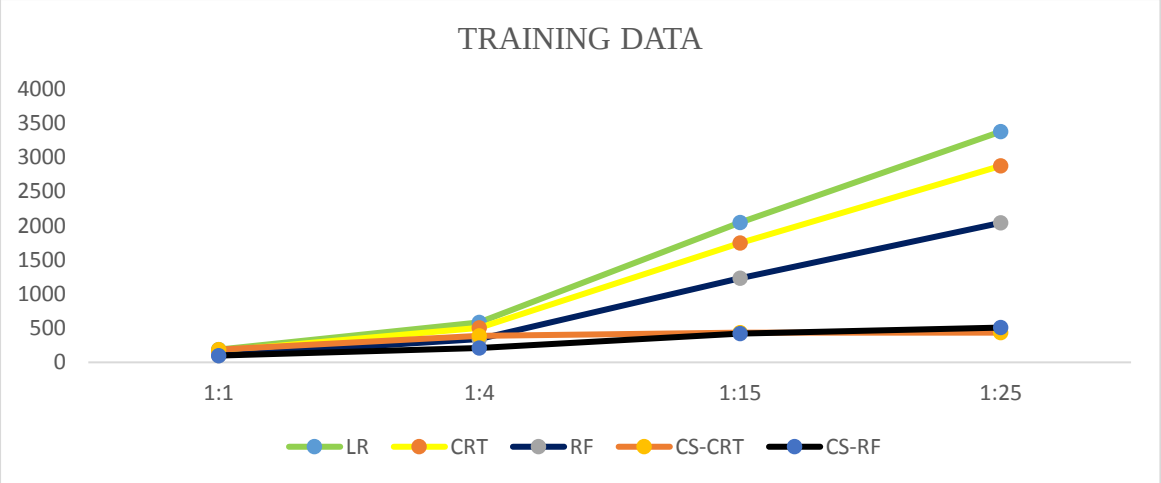


Figure 4.6: Effect of classification cost on cost sensitive and cost insensitive classifiers

### 4.6 Model Validation

The results from the cost insensitive and cost sensitive models which were obtained using the training data were tested using a holdout data. This data was 20% (178) of the entire observation in the dataset.

From table 4.6 below, logistic regression with an optimal threshold of 0.5 performed better in terms of Accuracy than the other classifiers. A change of the optimal threshold of the logistic regression classifier to 0.26 improves the F-measure and G-mean.

Making the tree base method cost-sensitive, also reduces the accuracy of the model. Cost sensitive trees had a better F-measure and G-mean than logistic regression with optimal threshold of 0.5.

Table 4.6: Comparing Performance of the various model using test data

| | Accuracy | AUC | F-measure | G-mean |
|---|---|---|---|---|
| LR | 0.74 | 0.64 | 0.37 | 0.51 |
| LR with Optimal cut off 0.26 | 0.64 | 0.64 | 0.40 | 0.60 |
| CART | 0.70 | 0.61 | 0.27 | 0.43 |
| CART(Cost Ratio=1:4) | 0.57 | **0.61** | **0.46** | 0.59 |
| RF | 0.74 | 0.66 | 0.36 | 0.49 |
| RF(Cost Ratio=1:4) | 0.65 | 0.66 | 0.45 | 0.56 |

### 4.6.1 Comparing laboratory diagnostic methods with cost sensitive and insensitive models on the testing data

The laboratory diagnostic method had a lower total classification cost (higher benefit) than the cost sensitive and insensitive method except culture as seen in Figure 4.7. The cost sensitive classifiers had their classification cost less than 300 cost unit whiles the cost insensitive classifiers had their classification cost more than 1000 cost unit.

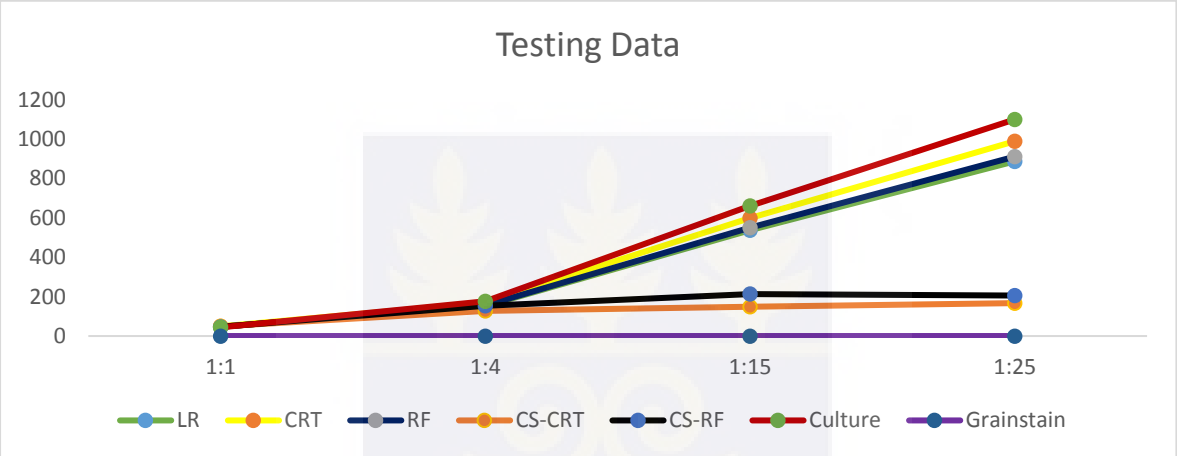


Figure 4.7: Total cost of classification of Laboratory method, Cost sensitive and insensitive classifiers

The results in Figure 4.8 indicate that Grain stain test was perfect. Cost sensitive tress a reduced type II error than Culture. This model also outperformed culture in terms of F-measure, Geometric mean and Kappa which a measure of agreement between classifiers. The reference test used was the results from Nucleic acid amplification test.

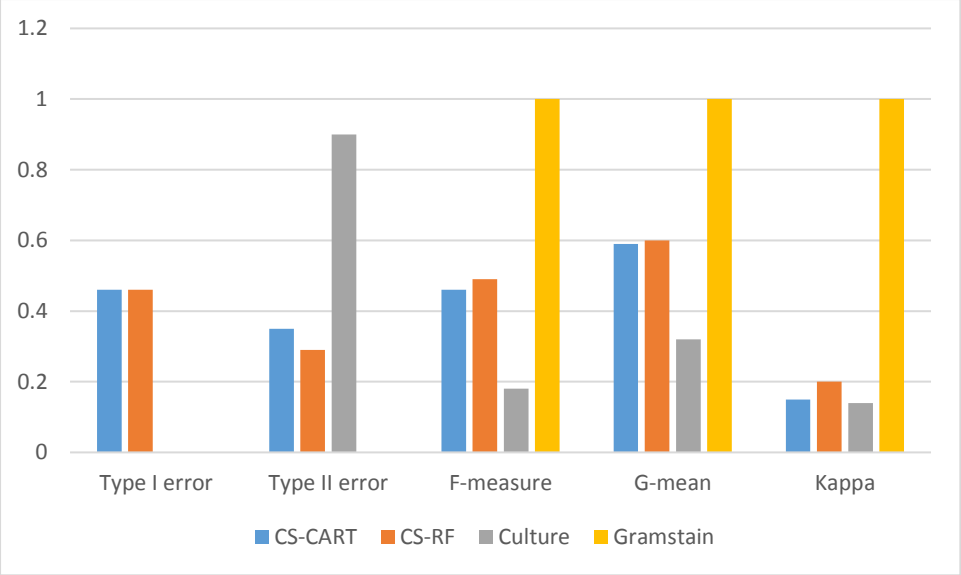


Figure 4.8: Laboratory diagnostic methods and Cost sensitive models

### 4.6.2 Effect of class distribution and cost sensitive method on classification cost

To evaluate the effect of the class distribution on classification cost, the total cost for the cost insensitive and sensitive classifiers were calculated for each dataset where class distribution was adjusted (i.e the class distribution of gonorrhea negative to positive was adjusted in a ratio of 40:60, 50:50 and 60:40 using the under sampling method). The results in Figure 9.0 indicate that when the ratio between the two classes was 60:40 a less classification cost was obtained using the cost insensitive classifiers. This is an indication that less classification cost is obtained when the data contains more gonorrhea positive cases than the negative cases. Regarding the cost sensitive classifiers, adjustment in the ratio of the class distribution weakly affected the classification cost.

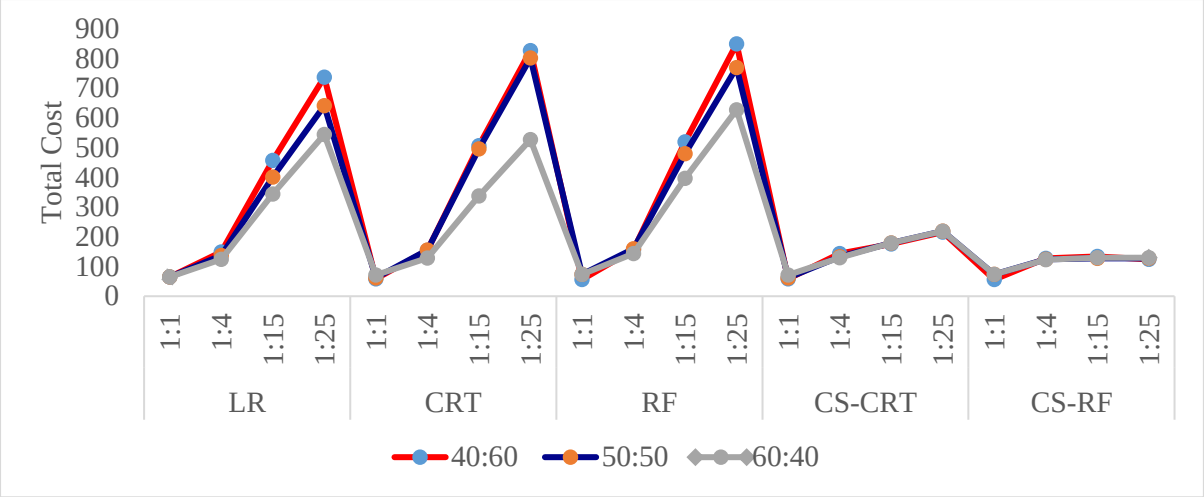


Figure 4.9: Effect of class distribution on classification cost of the classifiers

**4.7 Summary of results**

The models were fitted with logistic regression, classification and Random forest using equal cost of misclassification. For logistic regression model, the goodness of fit was tested using Hosmer-Lemoshow, log likelihood and deviance in which p-value greater than 0.05 obtained was an indication of a good fit to the training data. The misclassification rate of the model was 26%, F-measure was 43% and G-mean was 56%. Regarding the classification tree, the important variables selected were Gender, age, Pain during urination and more than one partner in the past month which was similar to that obtained in the logistic regression model. The model misclassification rate was 23%, F measure was 53% and G-mean was 64%. For Random forest, the misclassification rate was 14%, F-measure was 71% and G-mean was 76%.The cost of misclassification was varied (Cost of False negative was considered higher than cost of false positive) for the various models which yielded a reduction in the total classification cost and accuracy of the classifiers. For logistic regression an optimal threshold of 0.26 was obtained when the ROC curved which included the cost of misclassification was used to obtain a cut off. This increased the misclassification rate but improved the F-measure and G-mean. Similar results were obtained with the inclusion of the cost matrix in the classification tree and Random forest. The cost sensitive models performed well on the training data than the cost insensitive

models in terms of reduction on the total classification cost. Also, the class distribution had a weak effect on cost insensitive classifiers but weakly affected the cost sensitive classifiers. When the models where evaluated on a testing data, it had a poor performance than the training data but the cost sensitive models still out performed the cost insensitive models using F-measure and G-mean. Misclassification rate was not a good measure to evaluate this models since the data was unbalanced.

# CHAPTER FIVE

# CONCLUSIONS AND RECOMMENDATIONS

The final chapter of this thesis discussed the results and also deals with the conclusion and recommendation from the study.

## 5.1 Discussion

The results obtained from the study are discussed below;

### 5.1.1 Comparing Logistic regression, classification tree and Random forest

In this study, three traditional models were developed for the prediction gonorrhea infection. These are logistic regression, classification trees and Random forest. Furthermore, the models were evaluated using Classification cost. In addition, Accuracy, sensitivity, specificity, Area under the ROC curve, F-measure, and G-mean were used to determine the performance of the models.

The logistic regression identified four features which are age, gender, and Painful urination and Condom usage as significant for predicting gonorrhoea infection status which were similar for the classification trees and Random forest. The choice of these features for the classification tree and Random forest was based on the information gain and Mean decreasing Gini respectively.

The LR model was transformed into classifier by considering a 0.5 probability threshold for classification of gonorrhea infection whiles for Random forest, the most important variables were selected based on the mean decreasing Gini.

The tree base models had reduced misclassification error in the learning phase than the logistic regression (Table 4.5) but had a high misclassification error for the hold out data (Table 4.6).

The area under ROC curve for LR on the test data set is 0.64 while that of Classification tree and Random forest are 0.61 and 0.63 respectively. The difference between the AUC of LR and tree classifiers were not statistically significant likewise comparing the tree classifiers using

Delong's method as reported in Table 2a. This indicate that the classifiers virtually have the same probability rank of randomly choosing positive cases higher than a randomly choosing a negative case.

LR makes an assumption of the log odds of the response being a linear combination of the predictors. Since all the variables used in the model construction are discrete, the linearity assumption is satisfied hence minimizing the specification error which reduce the chance of overfitting in LR and making it more robust in the testing data than the tree classifiers. The tree classifiers poor performance for the testing data may be due to the fact that the pattern of features in the learning data set may not have been similar to the testing data set hence resulting in overfitting. One way of curbing this situation is pruning the classification tree which also did not yield any much difference when it was implemented. For Random forest, a portion of the data called out of bag is use to validate the model. This undergo bootstrap processes to obtain a prediction for the model which is an average of the bootstrap tree. In terms of accuracy, AUC F-measure and g-mean, Random Forest performed better than Classification tree on the test data (Table 4.6). Its performance compared to logistic regression was slightly high in terms accuracy and area under the ROC curve which is similar to the findings of Jin *et al*., (2007).

The type I error(False positive) for the classifiers were very less than the type II error(False Negative) which is as a result of the imbalance class distribution in the data set which made the models more likely to predict the negative case than the positive. These classification algorithms assume equal class distribution hence without any adjustment leads to classification bias toward the majority class.

In medical diagnosis, type II error are more severe than type I error (Feitas *et al*.,2009), this means for type II error ,individuals who have the disease have been misclassified as not having the disease. To address class imbalance in data set, some researchers choose to use random sampling method such as under-sampling and Over-sampling which does not have any

theoretical basis .This method tend to adjust the prior distribution of the learning data to be able to obtain a balance class distribution. Other method to use is the cost sensitive learning which can accept information cost and also assign different cost to the various misclassification error. This is very difficult to implement since in most cases the misclassification cost are unknown hence need to be assumed.

### 5.1.2 Effect of Classification cost on Laboratory diagnostic method and skewed class distribution of Cost sensitive and insensitive classifiers

Since the impact of each type of error has its own financial cost and harm to the individual, in medical diagnosis there is much focus to reduce the misclassification of type II error. From the study, cost sensitive trees helped to reduce the classification cost when the cost of misclassification for the errors are adjusted.

In comparing cost-sensitive classification tree and random forest to traditional classifiers which assume cost of misclassification, the cost sensitive classifiers performed better in terms of reduction in total classification cost. It also performed better in terms of F-measure and g-mean than the cost insensitive models. Even though traditional classifiers had a high accuracy than the cost-sensitive tree base methods, this measure is not appropriate for evaluating performance of classifiers since the data is imbalance and biases towards the majority class (Weng & Pong, 2006). The use of F-measure and G-mean are appropriate instances when the there is an imbalance data. F-measure is a combination of sensitivity and positive predictive value which is effective to use when there is an imbalance class distribution in the data. G-mean also has a combination of sensitivity and specificity.

Gram stain test had a perfect test which was extremely better than the cost sensitive models.

The other laboratory method which is culture had a high classification cost than the cost sensitive and insensitive classifiers. The poor performance of this laboratory diagnostic method

might be due to lack of adherence to standard operating procedures which may not have been followed when collecting samples, transporting and testing.

In other to determine the effect of class distribution of the classification cost, under sampling method in which varying ratio of gonorrhea negative to positive were used. The results in Figure 4.9 clear show that the class distribution weakly affect the classification cost of the cost sensitive classifiers but for the cost insensitive classifiers, the class distribution affect the classification cost. The reason why the class distribution did not affect the classification cost of the cost sensitive classifiers is due to the fact that the method uses a cost matrix which encodes the penalty of misclassifying a data sample.

## 5.2 Conclusion

In this work, a number of cost sensitive and cost insensitive algorithm were proposed for gonorrhea prediction. The study investigated whether the cost-sensitive approach for gonorrhoea prediction result in less total classification cost than the cost insensitive and laboratory diagnostic method. Also, it was investigated if the class distribution of the data had an effect on the total classification cost of the cost sensitive and insensitive classifier. The results obtained indicated that, cost sensitive methods outperform cost insensitive methods in terms of reduction of total classification cost (higher benefit). But it did not outperform the laboratory diagnostic methods except culture which even had a poor performance than the cost insensitive method. The class distribution of the data weakly affected the cost sensitive method but affected that of the cost insensitive method.

## 5.3 Recommendation

The following are some recommendation to consider;

- Misclassification cost, rather than accuracy of the classifiers should be considered when selecting statistical diagnostic model. This will help clinicians to make more effective decision in diagnosing by minimizing the number of False negative which has serious impact on patients and society at large
- Cost sensitive classifiers can be used if data for statistical prediction is skewed (i.e imbalanced problem in the dataset).
- Finding from classification tree can serve as a guide for clinicians when diagnosing gonorrhea. In most cases they rely on informal decision which enables them provide treatment for patients.

## REFERENCE

Abdullah, A. S., & Rajalaxmi, R. (2012, April). A data mining model for predicting the coronary heart disease using random forest classifier. In *International Conference in Recent Trends in Computational Methods, Communication and Controls*.

Acquah, H.D. (2013). Bayesian Logistic Regression Modelling via Markov Chain Monte Carlo Algorithm. Journal of Social and Development Sciences, *4(4)*, 193-197.

Adeyemo, O., Adeyeye, T., & Ogunbiyi, D. (2015). Comparative Study of ID3/C4. 5 Decision tree and Multilayer Perceptron Algorithms for the Prediction of Typhoid Fever. *African Joumal of Computing & ICT, IEEE,8(1)*,103-112.

Archer, K. J., & Lemeshow, S. (2006). Goodness-of-fit test for a logistic regression model fitted using survey sample data. *Stata Journal*, *6*(1), 97-105.

Austin, P.C. (2006). A comparison of regression trees, logistic regression, generalized additive models, and multivariate adaptive regression splines for predicting AMI mortality. Statistics in Medicine, *26(15)*, 2937-2957.

Bekkar, M., Djemaa, H. K., & Alitouche, T. A. (2013). Evaluation measures for models assessment over imbalanced data sets. *Iournal Of Information Engineering and Applications, 3*(10),27-38.

Biswas, B. (2006). Assessing agreement for diagnostic devices. In *FDA/Industry Statistics Workshop*. FDA.

Borges, L. S. R. (2016). Diagnostic Accuracy Measures in Cardiovascular Research. *Int J Cardiovasc Sci*, *29*(3), 218-222.

Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5-32.

Centres for Disease Control and Prevention. (2017). 10 ways STDs impact women differently from men. *CDC Fact Sheet*, (April), 1. Retrieved from http://www.cdc.gov/nchhstp/newsroom/docs/STDs-Women-042011.pdf

Centres for Disease Control and Prevention . (2010). National Center for HIV/AIDS, Viral Hepatitis, STD, and TB Prevention Division of STD Prevention.

Chawla, N.V., Bowyer, K.W., Hall, L.O., & Kegelmeyer, W.P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence and Research 16,* 321-357.

Chen, J.J., Tsai, C. A., Moon, H., A, H., Y, J. J., & Chen, C.H. (2006).The use of decision threshold adjustment in classification for cancer prediction.

Chou, Y.Y., & Shapiro, L. G. (2003). A hierarchical multiple classifier learning algorithm. *Pattern Analysis & Applications, 6*(2), 150-168.

Cosentino , L.A., Campbell, T., Jett , A., Macio , I., Zamborsky, T., Cranston, R. D., & Hiller,S.L. (2012). Use of Nucleic Amplification Testing for Diagnosis of Anorectal Sexually Transmitted infection. *Journal of Clinical Microbiology, 80(6),* 2005-2008.

Cramer, J. S. (2002). The origins of logistic regression.

Dal Pozzolo, A., Caelen, O., Johnson, R. A., & Bontempi, G. (2015, December). Calibrating probability with undersampling for unbalanced classification. In *Computational Intelligence, 2015 IEEE Symposium Series on* (pp. 159-166). IEEE.

Danjuma, K., & Osofisan,A.O.(2015).Evaluation of Predictive Data Mining Algorithms in Erythemato-Squamous Disease Diagnosis. *International Journal of Computer Science, 11(6),* 85-94.

De'ath, G., & Fabricius, K. E. (2000). Classification and regression trees: a powerful yet simple technique for ecological data analysis. *Ecology, 81*(11), 3178-3192.

De Queiroz Mello, F. C., do Valle Bastos, L. G., Soares, S. L. M., Rezende, V. M., Conde, M. B., Chaisson, R. E., & Werneck, G. L. (2006). Predicting smear negative pulmonary tuberculosis with classification trees and logistic regression: a cross-sectional study. *BMC Public Health*, *6*(1), 43.

Domingos, P. (1999, August). Metacost: A general method for making classifiers cost-sensitive. In *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 155-164). ACM.

Effects of STIs on Pregnancy | SexInfo Online. (n.d.). Retrieved November 25, 2016, from http://www.soc.ucsb.edu/sexinfo/article/effects-stis-pregnancy

Gardella, C., Brown, Z., Wald, A., Selke, S., Zeh, J., Morrow, R. A., & Corey, L. (2005). Risk factors for herpes simplex virus transmission to pregnant women: a couples study. *American journal of obstetrics and gynecology, 193*(6), 1891-1899.

Handsfield, H. H., Lipman, T. O., Harnisch, J. P., Tronca, E., & Holmes, K. K. (1974). Asymptomatic gonorrhea in men: diagnosis, natural course, prevalence and significance. *New England Journal of Medicine, 290*(3), 117-123.

Hastie, T. J., and Tibshirani, R. J. (1990). Generalized Additive Models. New York: Chapman & Hall.

Hsieh,C.H.,Lu,R.H.,Lee,N.H.,Chiu,W.T.,Hsu,M.H.,&Li,Y.C.(2010).Novel solution for an old disease: diagnosis of acute appendicitis with random forest, support vector machines, and artificial neural networks.Surgery,*149(1),*87-93.

Hssina, B., Merbouha, A., Ezzikouri, H., & Erritali, M. (2014). A comparative study of decision tree ID3 and C4. 5. *International Journal of Advanced Computer Science and Applications*, *4*(2), 13-19.

Huppert, J. S., Biro, F., Lan, D., Mortensen, J. E., Reed, J., & Slap, G. B. (2007). Urinary symptoms in adolescent females: STI or UTI?. *Journal of adolescent health*, *40*(5), 418-424.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning* (Vol. 112). New York: springer.
.

Jiang, Y., & Cukic, B. (2009, May). Misclassification cost-sensitive fault prediction models. In *Proceedings of the 5th international conference on predictor models in software engineering* (p. 20). ACM.

Jin, H., Kim, S.,& Kim ,J.(2014).Decision Factors on effective Liver Patient Data Prediction. International Journal of Bio-Science-Technology, *6(4)*, 167-178.

Kazemnejad,A.,Zayeri,F.,Aishah,H.,Gharaaghaji,R.,& Salehi,M.(2010).A Bayesian analysis of bivariate ordered categorical response using a latent variable regression model: Application to diabetic retinopathy data. Scientific Research and Essays,*5(11)*,1264-1273.

Kershaw, T. S., Lewis, J., Westdahl, C., Wang, Y. F., Rising, S. S., Massey, Z., & Ickovics, J. (2007). Using clinical classification trees to identify individuals at risk of STDs during pregnancy. *Perspectives on sexual and reproductive health*, *39*(3), 141-148.

King, G., & Zeng, L. (2001). Logistic regression in rare events data. *Political analysis*, 137-163.

Kohavi, R., & Quinlan, J. R. (2002). *Data mining tasks and methods: Classification: decision-tree discovery*. Paper presented at the Handbook of data mining and knowledge discovery.

Kolluru,M. (n.d.). What is the difference between logistic regression and Naive Bayes? - Quora. Retrieved September 13, 2016, from https://www.quora.com/What-is-the-difference-between-logistic-regression-and-Naive-Bayes

Kurt, I., Ture, M., & Kurum, A. T. (2008). Comparing performances of logistic regression, classification and regression tree, and neural networks for predicting coronary artery disease. *Expert Systems with Applications*, *34*(1), 366-374.

Lavanya, D., & Rani, K. U. (2011). Performance evaluation of decision tree classifiers on medical datasets. *International Journal of Computer Applications*, *26*(4).

Lecture 14 Diagnostics and model checking for logistic regression. (2004). Retrieved from https://courses.washington.edu/b515/l14.pdf

Leung, K. M. (2007). Naive bayesian classifier. *Polytechnic University Department of Computer Science/Finance and Risk Engineering*.

Lewis, R. J. (2007). *An introduction to classification and regression tree (CART) analysis.* Paper presented at the 2000 annual meeting of the society for academic emergency medicine. San Francisco. Disponível em: www. saem. org/download/lewis1. pdf. Acesso em: mar.

Ling, C.X., Yang, Q., Wang, J., & Zhang, S. (2004). Decision Trees with Minimal Costs. *In Proceedings of 2004 International Conference on Machine Learning (ICML'2004)*.

Liu, Y. (2007). *On goodness-of-fit of logistic regression model.* (PhD Thesis), Kansas State University.

Long, W. J., Griffith, J. L., Selker, H. P., & D'agostino, R. B. (1993). A comparison of logistic regression to decision-tree induction in a medical domain. *Computers and Biomedical Research*, *26*(1), 74-97.

Mohammed, G.S. (2016). Parkison's disease diagnosis: Detecting the effect of attribute selection and discretization of Parkinson's disease dataset on the performance of classifier algorithm.Open access library Journal, *3*(11), 1-11.

Meade, J. C., & Cornelius, D. C. (2012). Sexually transmitted infections in the tropics *Current Topics in Tropical Medicine*, Dr. Alfonso Rodriguez-Morales (Ed.), ISBN: 978-953-51-0274-8, InTech, Available from:http://www.intechopen.com/books /current-topics-in-tropical-medicine/sexually -transmitted-infections-in-the-tropics
.

Medova, E. (2008). *Bayesian Analysis and Markov Chain Monte Carlo Simulation*: Wiley Online Library.

Murray P.R., Baron E.J., Pfaller M.A., Jorgensen J.H., & Yolken R.H. (2003).Manual of Clinical Microbiology. 8th ed. Washington DC:American Society for Microbiology.

Ndongmo, B.C. (2005). Clinical laboratory diagnostics in Africa. *African Technology Development Forum Journal, 2(3),* 21-22.

Papp, J. R., Schachter, J., Gaydos, C. A., & Van Der Pol, B. (2014). Recommendations for the laboratory-based detection of Chlamydia trachomatis and Neisseria gonorrhoeae—2014. *MMWR. Recommendations and reports: Morbidity and mortality weekly report. Recommendations and reports/Centers for Disease Control*, *63*(0), 1-19.

Parker, C. (2011, December). An analysis of performance measures for binary classifiers. In *Data Mining (ICDM), 2011 IEEE 11th International Conference on* (pp. 517-526). IEEE.

Patel, N., & Upadhyay, S. (2012). Study of various decision tree pruning methods with their empirical comparison in WEKA. *International journal of computer applications*, *60*(12).

Pazzani, M., Merz, C., Murphy, P., Ali, K., Hume, T., & Brunk, C. (1994). Reducing misclassification costs. In *Proceedings of the Eleventh International Conference on Machine Learning* (pp. 217-225).

Peng, C. Y. J., Lee, K. L., & Ingersoll, G. M. (2002). An introduction to logistic regression analysis and reporting. *The journal of educational research*, *96*(1), 3-14.

Peng, W., Chen, J., & Zhou, H. (2009). An implementation of ID3-decision tree learning algorithm. *From web. arch. usyd. edu. au/wpeng/DecisionTree2. pdf Retrieved date: May*, *13*.

Sahin, Y., Bulkan, S., & Duman, E. (2013). A cost-sensitive decision tree approach for fraud detection. *Expert Systems with Applications*, *40*(15), 5916-5923.

Salameh,P.,Waked,M.,Khayat,G.,& Dramaix, M.(2014).Bayesian and Frequentist Comparison for Epidemiologists: A Non Mathematical Application on Logistic Regressions. *The open Epidemiology*, *7*(1), 17-26.

Schachter J., Moncada J., & Liska S. (2008). Nucleic acid amplification tests in the diagnosis of chlamydial and gonococcal infections of the oropharynx and rectum in men who has sex with men. *Sexually transmitted disease*, *35*(7), 637-42.

Smith,L.(2016). *Gonorrhea: Causes, Systems and Treatments*. Retrieved December 13, 2016 from http://www.medicalnewstoday.com/articles/155653.php.

Smith, A. F., & Roberts, G. O. (1993). Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society. Series B (Methodological)*, *55*(1), 3-23.

Sokolova, M., Japkowicz, N., & Szpakowicz, S. (2006). *Beyond accuracy, F-score and ROC: a family of discriminant measures for performance evaluation.* Paper presented at the Australian conference on artificial intelligence.

Steinberg, D., Golovnya, M., & Cardell, N. S. (2004). Data Mining with Random Forests™.

Therneau, T. M., & Atkinson, E. J. (1997). *An introduction to recursive partitioning using the RPART routines* (Vol. 61, p. 452). Mayo Foundation: Technical report.

Ture, M., Kurt, I., Kurum, A. T., & Ozdamar, K. (2005). Comparing classification techniques for predicting essential hypertension. *Expert Systems with Applications*, *29*(3), 583-588.

Turney, P. D. (1995). Cost-sensitive classification: Empirical evaluation of a hybrid genetic decision tree induction algorithm. *Journal of artificial intelligence research*, *2*, 369-409.

Verma, R., Sood, S., Kapil, A., & Sharma, V. K. (2009). Diagnostic approach to gonorrhoea: Limitations. *Indian Journal of Sexually Transmitted Diseases and AIDS*, *30*(1), 61.

Weiss, G. (2003). The Effect of Small Disjuncts and Class Distribution on Decision Tree Learning, *Ph.D. Dissertation, Department of Computer Science, Rutgers University*, New Brunswick, New Jersey.

Whiley, D. M., Tapsall, J. W., & Sloots, T. P. (2006). Nucleic acid amplification testing for Neisseria gonorrhoeae: an ongoing challenge. *The Journal of Molecular Diagnostics, 8*(1), 3-15.

Yusuff, H., Mohamad, N., Ngah, U., & Yahaya, A. (2012). Breast cancer analysis using logistic regression. *International Journal of Research and Reviews in Applied Sciences, 10*(1),14-22.

Zadrozny, B.,& Elkan, C.(2001).Learning and making decisions when costs and probabilities are both unknown. *Proceedings of the Seventh International Conference on Knowledge Discovery and Data Mining*. pp. 204-2013.ACM Press.

Zhou, Z. H., & Liu, X. Y. (2006). ON MULTI-CLASS COST-SENSITIVE LEARNING. *Proceedings of the 21st National conference on Artificial intelligence*, pp. 567-572. Boston, MA.

# APPENDIX

## Logistic Regression Diagnostic measure

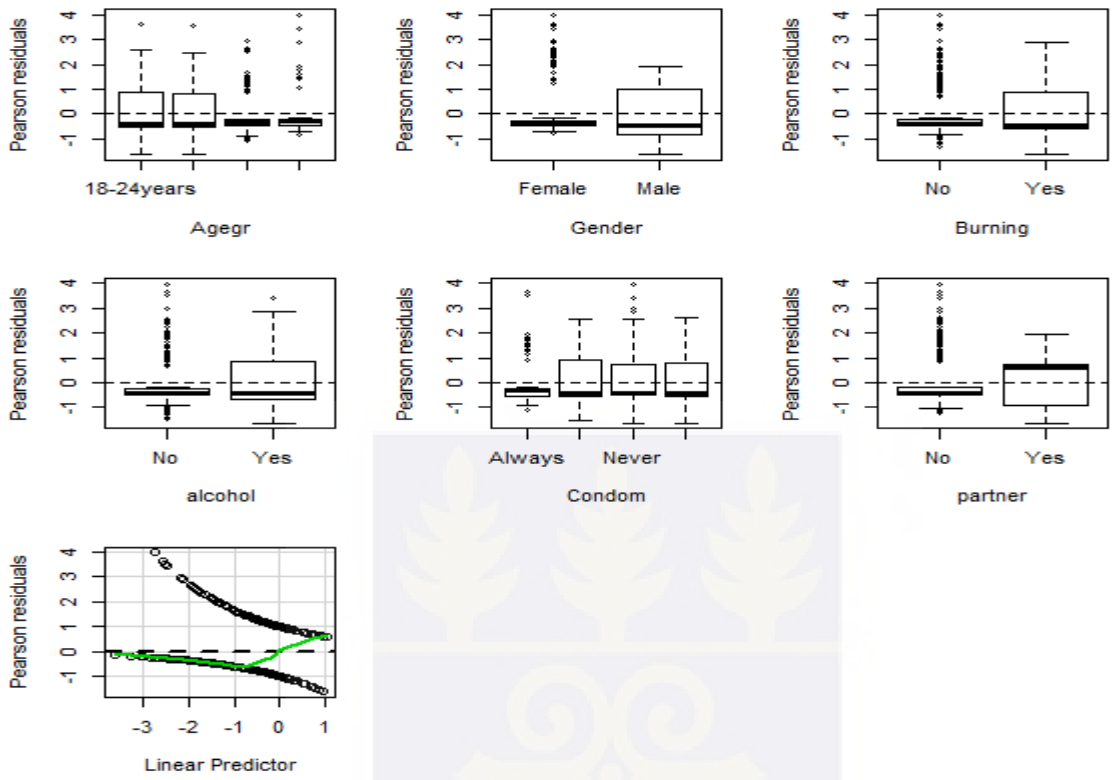


Figure 1a: Pearson Residuals plotted against predictor one by one.

Table 1a: Bayesian logistic regression

| Variable | Mean | Std. Dev. | Lower | Upper |
|---|---|---|---|---|
| Intercept | -2.58 | 0.30 | -3.19 | -2.0 |
| Age | | | | |
| 25-31years | 0.06 | 0.21 | -0.34 | 0.49 |
| 32-38years | -0.46 | 0.28 | -1.01 | 0.10 |
| 39years and above | -1.08 | 0.37 | -1.85 | -0.41 |
| Male | 1.38 | 0.19 | 1.02 | 1.76 |
| Pain during Urination | 0.36 | 0.19 | -0.03 | 0.73 |
| Condom usage | | | | |
| Never | 0.85 | 0.28 | 0.31 | 1.33 |
| Rarely | 0.78 | 0.29 | 0.31 | 1.37 |
| Most occasion | 0.66 | 0.32 | 0.22 | 1.37 |
| More than one sexual partner in past month | 0.65 | 0.27 | 0.13 | 1.16 |

**Bayesian regression diagnostic measures**

On the left is Time series of the parameter in the model as MCMC iterates and on the right is the probability density estimate of the parameters which likely to occur at the peak of the distribution(posterior mode)

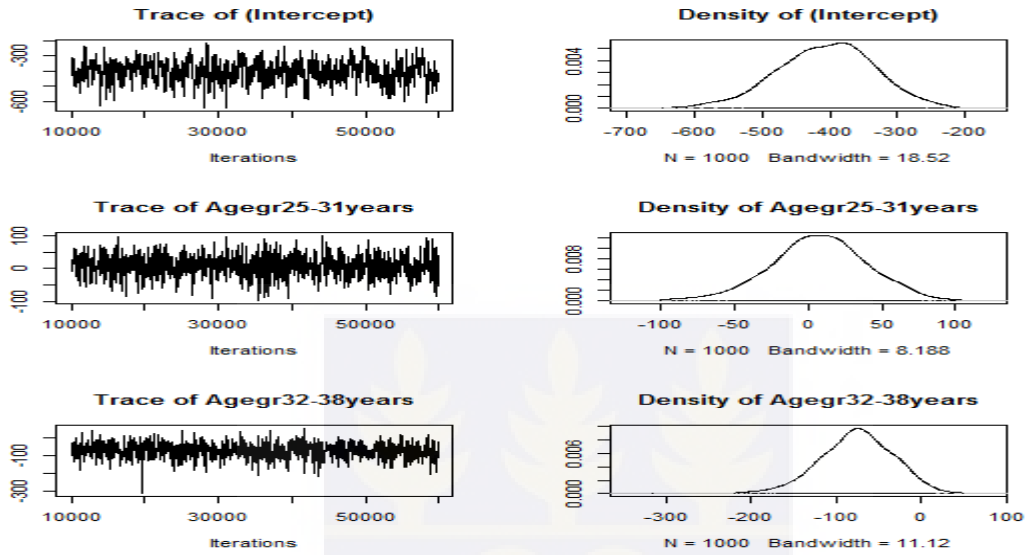


Figure 1b: Posterior distribution of the model parameters

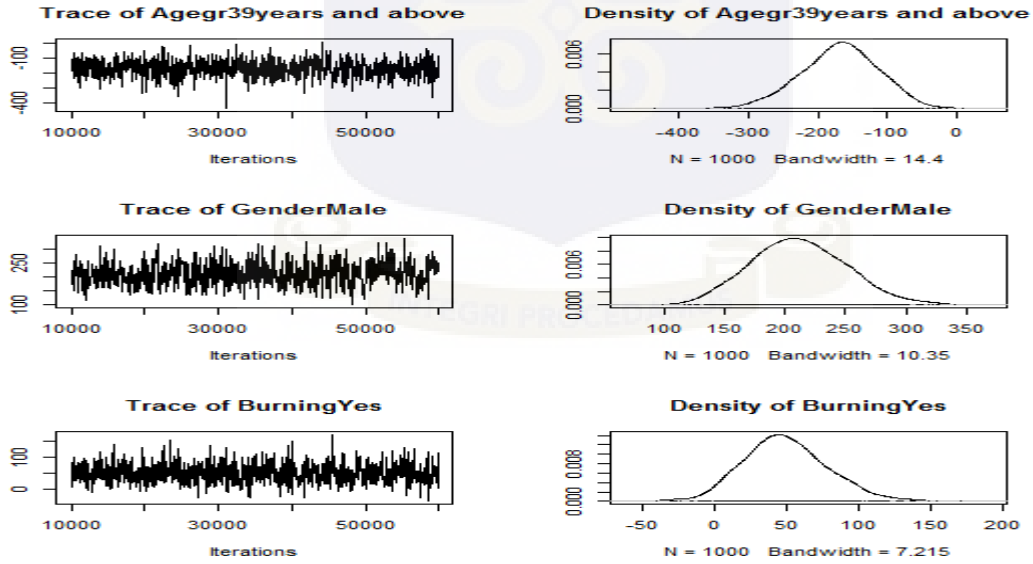


Figure 1b: Posterior distribution of the model parameters (Cont.)

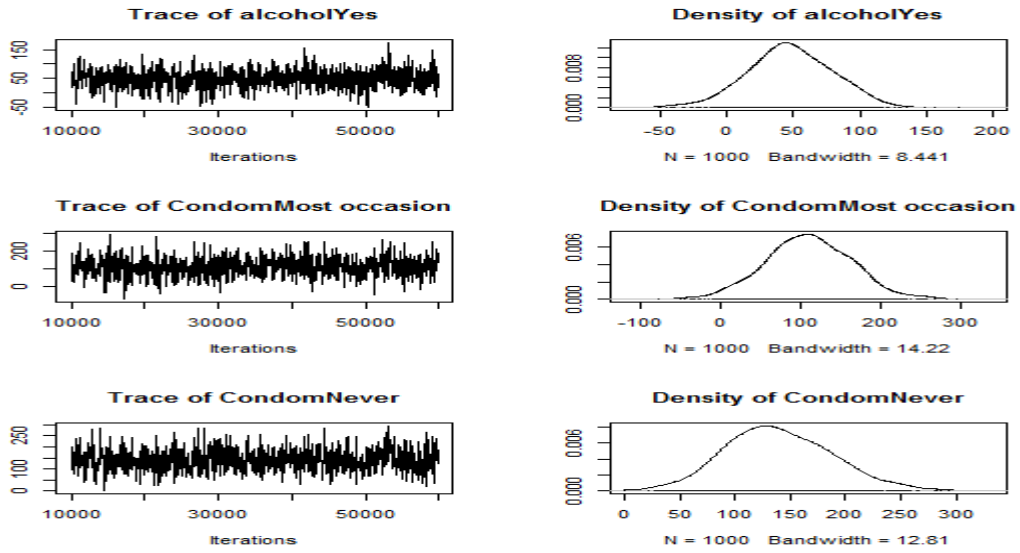


Figure 1c: Posterior distribution of the model parameters (Cont)

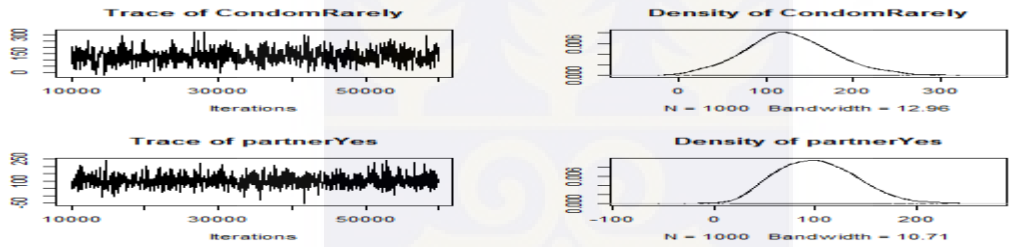


Figure 1d: Posterior distribution of the model parameters (Cont.)

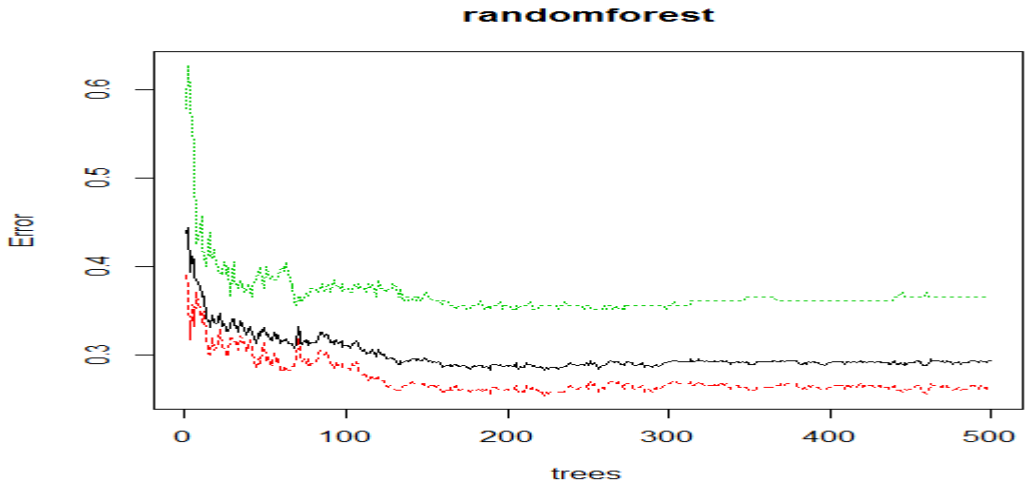


Figure 1e : Error rate for the number of trees