COLLEGE OF BASIC AND APPLIED SCIENCES

SCHOOL OF PHYSICAL AND MATHEMATICAL SCIENCES

DEPARTMENT OF STATISTICS

A PROPOSED METHOD FOR CLASSIFICATION; AN APPLICATION TO FORECASTING STUDENTS' QUALIFICATION FOR ADMISSION INTO FIRST DEGREE PROGRAMS IN GHANAIAN UNIVERSITIES

THIS THESIS IS SUBMITTED TO THE UNIVERSITY OF GHANA, LEGON

IN PARTIAL FULFILLMENT OF THE REQUIREMENT FOR THE AWARD OF

MPHIL STATISTICS DEGREE

BY

DELALINAM BESA AMEGAGO

(10507134)

JULY, 2016

DECLARATION

I declare that this thesis is a write-up of my own work towards the award of Master of Philosophy degree in Statistics, and that no part of it has been presented for any other degree in the university of Ghana or elsewhere except where due acknowledgement had been made in the text.

Delalinam Besa Amegago Signature ……. Date

(10507134)

Supervisors' Declaration

Dr. F.O. Mettle Signature……. Date

(Principal Supervisor)

Dr. K. Doku-Amponsah Signature……. Date

(Co-Supervisor)

ABSTRACT

This study proposes a method for classification based on a Pattern Trace Network and Principal Component Analysis. Using data from a public Senior High School, the proposed method was applied to forecast whether or not a Senior High School (SHS) student would qualify for admission into the university. The method's performance and the performance of Logistic Regression were compared. The independent variables used in the study are gender, boarding status, program of study, and continuous assessment. 79% of the data was used as a training set to estimate the parameters of the models and the remainder was used as a test set to evaluate the performances of the methods. The stages of the modelling procedure are: grouping the data into smaller subsets, preprocessing, Principal Component Analysis, computation of mean vectors of new descriptors of the pre-defined classes in the data. Judging from the methods' performances on the test set, both methods are equally good, with 78% overall predictive accuracy each. In terms of specificity and sensitivity, the Proposed Method had 87% and 63% respectively while Logistic Regression had 84% and 68%. The study concludes that the proposed method is an efficient alternative to solving a classification problem whose implementation does not necessarily require a sufficiently large sample. The study also reveals that, as far as continuous assessment is concerned, marks obtained by a Senior High School Student in the third term of S.H.S. 1 is enough to predict whether or not he or she would qualify to gain admission into the university. The study recommends among other things that the proposed pattern trace network approach is an effective method for classification.

DEDICATION

This work is dedicated to the following people:

1. My parents, Prof. Modesto M.K. Amegago and Mrs. Georgina M. Amegago.
2. All my siblings, including Mawulorm Amegago (of blessed memory).
3. Mrs. Eyram Amegago.
4. All friends and well-wishers.

ACKNOWLEDGEMENTS

I thank the almighty God for granting me good health and intellect to enable me to undertake this project. While I am very grateful to Gertrude Akuffo for encouraging me to enroll for the MPHIL. Statistics program in the University of Ghana, I would like to register my appreciation to my biological parents: Prof. and Mrs. Modesto Amegago for the financial and moral support that they offered me throughout my period of study.

I am very grateful to my supervisors: Dr. F.O. Mettle (Principal Supervisor) and Dr. K. Doku-Amponsah for the patience and tolerance they exhibited, both in teaching and supervising me; all through this work. I also thank Mr. Louis Asiedu for his guidance and mentorship.

I am particularly grateful to my brother, Mawuenam Amegago for allowing me to use his laptop when my laptop got spoilt in the middle of the project, not forgetting the rest of my family members: Venunye, Senyo, Xoenam, and Eyram. Most importantly, I convey my profound gratitude to Mrs. Eyram Amegago for the spiritual, emotional, and financial support offered me throughout my period of study. I am also grateful to Mr. D.K. Sedanu-Kwawu (headmaster, Keta Senior High Technical School), Mr. and Mrs. Tay (for assisting me to get data for this work), Mr. Henry Wosor and all staff members of Keta Senior High Technical School for the support given me.

Table of Contents

LIST OF FIGURES

LIST OF TABLES

1. INTRODUCTION

1.1 Background of Study

A classification problem arises when objects of two or more categories or populations or classes have certain common characteristics/features such that the class to which a given object belongs cannot be immediately known. Classification is the task of assigning a class label to a given object based upon a rule of assignment (Ripley, 1997; Chandola *et al.*, 2009). The rule of assignment is generally derived from an example set of known objects and their corresponding class labels (Chandola *et al.*, 2009). This rule of assignment is arrived at by exploring the features that distinguish members of one class (population) from the other. The technique (process) of describing the differential features of objects (observations) from the several known classes is referred to as discrimination (Johnson and Wichern, 2007). An example of a classification problem is the task of finding the species to which an organism belongs. Another example is determining whether a currency is a fake one or not. There are several other classification problems such as determining whether or not an out-patient is suffering from a certain illness, determining whether or not an email message is spam. The system of processes/procedures designed to assign a class label to an object is called a classifier. In the language of statistics, it suffices to say that a classifier is a model that predicts the class of an object or an observation. The special cases where objects are to be assigned exactly one of two class labels are referred to as binary classification problems or 2-class classification problems. Consequently, a binary classifier is a model that assigns to an observation exactly one of two class labels. This study focuses on the binary classification problem and proposes a method for solving it; a method that can as well be applied to solving classification problems involving more than two classes (multi-class problems).

Usually, binary classification problems arise when there is uncertainty about the occurrence of an event **of interest**. In such cases, one often needs to have an idea of the likelihood of the occurrence of the event in order to make a prudent decision or to take a wise step. On one occasion, the interest may be to know if a patient is suffering from a particular disease or not; on another occasion, the interest may be in knowing if or not a patient would expire (die) as a result of a particular illness, other times, we may want to know whether or not a credit applicant would default on the facility when granted, whether or not it would rain on a particular date in the year. A politician may be eager to know whether or not a certain group of electorates would vote for him, the management/sponsors/supporters of a certain football team may also wish to know whether or not their football team would win a match. These are all further examples of classification problems.

In probability theory, a binomial/binary random variable can be thought of as the number of times, X, that an **event of interest** occurs in a number, n, of Bernoulli trials. Where a Bernoulli trial refers to a random experiment with exactly two possible outcomes; referred to as success and failure (Syayib, 2013). The outcome is referred to as "success" when the event of interest occurs and "failure" when the event of interest does not occur. In short, a binomial variable is a variable that has only two possible outcomes, namely, success and failure. The output of a binary classifier can be categorized into four categories. These are: True Negatives (TN), False Positives, False Negatives, and True Positives. A True Negative is an outcome that is known to be a failure and at the same time predicted by the classifier as failure. A False Positive is an outcome that is known to be a failure but predicted by the classifier as success. A False Negative is an outcome that is known to be a success but predicted by the classifier as failure. Finally, a True Positive is an outcome that is actually success and also predicted by the classifier as success.

There exists several methods or techniques for forecasting the class label of a binary outcome. Two popular statistical methods used for solving a binary classification problem are linear discriminant analysis (LDA) and logistic regression (LR). The linear discriminant analysis approach finds a linear function of the predictor variable(s) (known as the discriminant), that best discriminates between the two classes of the binary response variable. A threshold is then set on the function to classify instances as successes or failures. The linear discriminant is best used only when all the predictor variables are normally distributed (Johnson and Wichern, 2007). There are a couple of other assumptions that must be satisfied for the implementation of the LDA, these assumptions are discussed later in Chapter 2. The logistic regression on the other hand makes fewer assumptions about the data and so it is preferred to discriminant analysis in many situations when the assumptions of the later are not met. Logistic regression is best used when all the predictor variables are categorical and also, when there is a sufficiently large sample. The forecast of linear discriminant analysis can be described as a point forecast while that of the logistic regression can be described as a probability forecast (Lahiri *et al.*, 2012). This is because whereas the linear discriminant assigns exactly one of the two classes (success/failure) to an observation, based upon a set threshold, the logistic regression computes the conditional probability of the observation being a success. In this way, the end-user(s) of the forecast are given an idea of the degree of uncertainty associated with the forecast. Nevertheless, ultimately, the logistic regression also assigns the observation to one of the two classes based upon a threshold set on the value of the probability. Also, in the case of the logistic regression, there is a more formalized way of assessing the influence of each predictor variable on the occurrence of the event of interest. These attributes of the logistic regression approach make it more popular in binary classification problems.

Over the years, with the increasing ability of computer software programs to perform complex computational tasks in an incredibly short period of time, the development of pattern recognition techniques of solving classification problems has been greatly inspired.

Pattern Recognition refers to the discovery of patterns and regularities in raw data through the use of computer algorithms (Bishop, 2006).

In the literature, the use of the term "pattern recognition" is somewhat varied and confusing. Some writers simply used the term as the computer science terminology for classification. Wasserman (2011) is one such example. Undeniably, classification appears to be the most common application of pattern recognition efforts. This notwithstanding, pattern recognition also lends it techniques to other modelling domains, these include:

- regression
- parsing

Singh (2014) describes a pattern as an object or an event that can be given a name, Singh also describes a pattern class as a set of patterns that have similar characteristics. In this study, a pattern is seen as an arrangement/organization of features that is reasonably similar for objects/members/individuals of the same class/population, and for that matter can be used to distinguish members of one population from members of another population. A pattern can also be thought of as the design or shape or form of an object. Based on this idea, pattern recognition can be viewed as any attempt, statistical or otherwise, aimed at finding out (identifying) characteristics or features that distinguish among members of different populations. When this attempt is made based on some underlying statistical model, the result is in the domain of statistical pattern recognition. The term "**discrimination"** is mostly used in statistics to represent this process. Statistical pattern recognition is basically about finding a linear function that discriminates

one population from the other, examples include the linear discriminant analysis and logistic regression mentioned earlier. Over the years, the capacity of computers in undertaking complex computing tasks has increased greatly, inspiring the development of non-statistical pattern recognition methods. Basically, classification by non-statistical pattern recognition is achieved by estimating a non-linear composite function. The estimated non-linear composite function is usually an algorithm (a set of rules) that is used to assign a pre-defined class to new observations/objects. It can be recalled that in Mathematics, a function is a rule that assigns to every member of a set, called the domain, one and only one corresponding member of another set called the range; and a composite function is a function of another function. Getting the optimal set of rules that best fixes the classification task is achieved by iterating over several possibilities and selecting the best result. This is where the computing power of computers come to play. The merit of the so called non-statistical pattern recognition methods is that, they are able to dig into the data structure and bring out hidden patterns that otherwise would have gone unrecognized. These are very helpful in this era when industries generate very complex high dimensional data sets. From this point onwards, the term pattern recognition is used in the sense of non-statistical pattern recognition. On a whole, pattern recognition essentially aims at formalizing, explaining, and visualizing the pattern in data. In the strict sense of the phrase: "pattern recognition", it suffices to say that the job of pattern recognition ends at identifying the features that distinguish one population from the other and classification is the task that pattern recognition can be used to achieve. It is however, conventional to refer to the union of both of them as pattern recognition.

There are many pattern recognition techniques/algorithms that are employed in achieving classification, the following are some examples:

- Linear discriminant analysis

- Quadratic discriminant analysis
- Decision trees, decision lists
- Naive Bayes classifier
- Kernel estimation and K-nearest-neighbor algorithms
- Neural networks (multi-layer perceptrons)
- Support vector machines
- Gene expression programming

This study proposes an algorithmic method of binary classification based on a Pattern Tracing Network concept. The method is applied on a data set comprising of students' variables and qualification for admission into a bachelor's degree program. Here, a student's sitting for the examinations is the Bernoulli trial, and the two possible outcomes of the trial are:

1. The student obtains grades that qualify him or her for admission into a first degree program
2. The student does not obtain grades that qualify him or her for admission into a first degree program

In this study, the first outcome is the event of interest and so is referred to as "success". The second outcome is referred to as "failure".

In Ghana, the criteria for admission to pursue degree programs in the universities are set by the National Accreditation Board (NAB) of Ghana. At the end of a barely two and half year Senior High School (S.H.S.) program, students sit for the West Africa Senior School Certificate examination, (hereafter, WASSCE). The WASSCE result is one of the main qualifications for

entry into bachelor's degree programs in Ghanaian universities. The WASSCE is conducted by the West African Examinations Council (WAEC)

WASSCE is now written twice in each year, one written by the regular students in May/June each year and the other by private candidates in September/October each year. The private candidates are usually those who have completed S.H.S. but have not had the grades they desire to proceed to the tertiary institutions. Nowadays, however, Senior High School students who wish to evaluate their academic preparedness or feel that they are ready to sit for the examinations or for any other reason also sit for the examination, at the end of their second year in the S.H.S. Before the introduction of the WASSCE, the WAEC conducted the Senior High School Certificate Examination (SSSCE) which succeeded the Ordinary Level and the Advance level examinations. The WASSCE was introduced in May/June 2006. The SSSCE spanned the period 1993 to 2005. The WASSCE uses a nine-point scale to grade candidates in each subject. The WASSCE grading structure is shown in Table 1.1

Table 1.1: WASSCE Grading Structure

Grade	Range of Marks (in percentage)	Interpretation
A1	80-100	Excellent
B2	70-79	Very Good
B3	60-69	Good
C4	55-59	Credit
C5	50-54	Credit
C6	45-49	Credit
D7	40-44	Pass
E8	35-39	Pass
F9	0-34	Fail

Source: **Ghana Education Service (GES) Teaching Syllabus for Core Mathematics**

1.2 Problem Statement

The assumptions of the traditional statistical methods (Logistic Regression and Linear Discriminant Analysis) of classification impose certain limitations on the extent of their applicability. For instance, Logistic Regression is unable to give meaningful and reliable measures of model fit when the data set is not sufficiently large. Additionally, Logistic Regression estimates coefficients by the method of maximum likelihood, which relies heavily on large sample sizes to produce unbiased coefficients. Consequently, in cases where sample sizes are small the maximum likelihood estimator has a very high tendency of being biased. Another notable shortfall of the method is a phenomenon referred to as, sparseness. This is when some values of the independent variable have 0 or 1 probability of occurrence in the data. That is to say that any time the independent variable takes on those values, the outcome is always either a success or a failure. Meanwhile, this is a common phenomenon in real life and should not have been a problem. It is very common to observe sparseness in the binary classification data when there are more predictor variables in the model especially when there are several continuous predictor variables. This is because when there are several predictor variables, the likelihood of observing a particular combination of all the predictor variables only once or rarely in the data set is very high. When it so happens that such a combination occurs only once in the data, then obviously the computed conditional probability of such an observation is clearly either 0 or 1. This is problematic in logistic regression because the method works by fitting the curve

$$\ln\left[\frac{\Pr\left(Y=1 \mid X_i = x_{i1},...,x_{ip}\right)}{\Pr\left(Y=0 \mid X_i = x_{i1},...,x_{ip}\right)}\right] = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip}, \qquad (1.1)$$

where $X_1,...,X_p$ are predictor variables of some binary response Y; $\beta_0, \beta_1,..., \beta_p$ regression coefficients, and $\ln$ is the natural logarithm function. This implies that, whenever

$\Pr\left(Y=1 \mid X_i = x_{i1},...,x_{ip}\right)=1$, implying that $\Pr\left(Y=0 \mid X_i = x_{i1},...,x_{ip}\right)=0$, the LHS of (1.1) is undefined. At the same time, whenever $\Pr\left(Y=1 \mid X_i = x_{i1},...,x_{ip}\right)=0$ implying that $\Pr\left(Y=0 \mid X_i = x_{i1},...,x_{ip}\right)=1$, the LHS of (1.1) is undefined. Although there exists some techniques of fixing this problem, it (the problem) makes the method more useful for situations where all predictor variables are categorical. The linear discriminant approach is a very good method for classification when all predictor variables are continuous but it assumes that each of these variables has a univariate normal probability density and so any departure from this assumption yields unreliable results. This makes the method less useful in cases where some independent variables are not continuous. It has several other assumption that, in practice, the data seldom satisfies. The fact that all these traditional methods of classification have their respective shortfalls is what has motivated the development of new computer-intensive non-linear models of classification. Same is the motivation for this research. This study also seeks to address the challenges with the criteria that the Senior High Schools adopt in selecting students for the examinations. The schools usually, do not measure the students' performance based on the requirements of the higher institutions before registering them for the final external exams.

1.3 Objectives of Study

1.3.1 General Objective

To develop and propose an alternative method of classification that does not make much assumptions about the data (as in the case of the traditional methods thus addressing the drawbacks of the existing methods) and apply it to solve a classification problem on an educational data set; a method that would be efficient for a data with both categorical and continuous predictor variables.

1.3.2 Specific Objectives

The specific objectives of the study are:

1. To determine whether or not the proposed method is a good alternative to the logistic regression method by comparing the predictive performances of the two methods on several metrics.
2. To find out which variables/factors are statistically significant in predicting students' success in the WASSCE (qualification for entry into a bachelors program in Ghanaian universities) and the extent to which they do so.

1.4 Significance of Study

- The main significance of this study is that it seeks to provide an alternative method of classification that makes almost no assumption about the data structure.
- As far as the data used in the study is concerned, the relevance of the study can also be seen in the ways that it would:
    - Shed more light on the various factors that influence a student's success in the WASSCE and how they (the factors) interact with the existing clusters/sub-groups of the students
    - Inform school authorities on what to look for in presenting a candidate for the WASSCE
    - Inform the parents and guardians of the student on steps they should take to ensure that their wards perform well in the examination
    - Serve as a comparative study and dichotomize the predictive performances of two binary classification models

- Induce a paradigm shift among researchers who have been investigating issues about students' performance in examinations and serve as a reference point for future works

1.5 Limitations of the Study

- Due to time constraint and data limitations, the proposed method could not be applied on several data sets
- The proposed method alone is not able to explain the relative significance of the predictor variables to the determination of the response variable.

1.6 Organization of Study

This thesis is in five chapters. The first chapter introduces the subject matter of the study; stating the problem that the study attempts to address, significance of the study, the goals that the study seeks to achieve, and the limitations of the study. The second chapter reviews some existing methods of classification and previous works related to the study, the third chapter describes in detail, the two methods that are employed in addressing the problem with illustrations of how the methods are applied on a hypothetical data set; relevant mathematical formulae, theorems, and ideas are also presented in the third chapter. The fourth chapter begins by performing some exploratory data analyses with the view of getting a fair idea of the general trends in the data. In the fourth chapter, results of the implementation of the methods described in chapter 3 are presented, the performances of the models used by the methods are analyzed and compared on various metrics. The fifth chapter makes inferences/conclusions based on the findings in the fourth chapter and suggests directions for future research.

2. LITERATURE REVIEW

2.1 Introduction

In this study, classification of the WASSCE candidates' examination outcome as success or failure is viewed as an example of a binary classification problem situated within the broader context of classification and as such, reviews the literature available and relevant to the study on the following sub-themes:

i. Overview of some classification methods,
ii. Pattern Recognition methods of classification, and
iii. Predicting students' performance in examinations based on their school-based data
iv. Evaluation of Classifiers

2.2 Overview of Some Classification methods

This section discusses the salient features of five classification models briefly.

2.2.1 Linear Discriminant Analysis

Linear Discriminant Analysis (LDA) is a generalization of the Fisher linear discriminant (Fisher, 1936) used when the following assumptions are met:

- Within-group variances are homogeneous
- All independent variables are normally distributed
- All pairs of independent variables are linearly related
- Prior probabilities of the classes are known.

It models the difference between the classes of the data. It attempts to reduce dimensional as well as maintain as much discriminatory information as possible simultaneously. Given a data Y, consisting of two classes Y_0 and Y_1 , LDA finds a linear function, $K(\beta)$ that best discriminates between the two classes. The function K is a linear combination of a vector of weights/coefficients β^T and the independent variable(s) X , that is

$$K = K(\beta, X) \tag{2.1}$$

Since the data is fixed, (2.1) can just be written as

$$K = K(\beta). \tag{2.2}$$

The linear function K is sought to maximize the ratio, R, of the variance between the classes to the variance within the classes. A threshold c is then set and a new observation y_i is classified into either of the classes depending on the value of K relative to c. The variation between the two classes Σ_b is defined as:

$$\Sigma_b = (\mu_0 - \mu_1)(\mu_0 - \mu_1)^T, \tag{2.3}$$

where μ_0 and μ_1 are the respective mean vectors of the classes labelled 0 and 1. Similarly, the variation within the classes Σ_w is defined as:

$$\Sigma_w = \Sigma_0 + \Sigma_1, \tag{2.4}$$

where Σ_0 and Σ_1 are the respective variance-covariance matrices of class 0 and class 1. Consequently, the ratio to maximize is

$$R = \frac{\beta^T \Sigma_b \beta}{\beta^T \Sigma_w \beta}. \tag{2.5}$$

It turns out (from differentiating and equating to zero) that the value of β that maximizes this ratio is

$$\beta^T = \Sigma_w^{-1} \mu_0 - \mu_1 \tag{2.6}$$

A new observation y_i is assigned the class labelled 0, if $\beta^T \left[y_i - \frac{1}{2}(\mu_0 + \mu_1) \right] \geq \log \frac{p(Y_0)}{p(Y_1)}$, otherwise, it is classified as belonging to the class labelled 1. Here, $p(Y_0)$ is the prior probability of the class Y_0 and $p(Y_1)$ is the prior probability of the class Y_1. Most classification problems, however, seldom satisfy the assumptions of LDA. For instance, when some or all the independent variables are categorical, then the linear discriminant cannot be applied.

2.2.2 The Linear Probability Model

Traditionally, ordinary least squares (OLS) regression was used to fit the linear probability model,

$$Y = X\beta + \varepsilon \tag{2.7}$$

and the resulting conditional expectation $E(Y \mid X = x_i) = x_i \beta$ is interpreted as a conditional probability that the outcome or attribute of interest will occur; i.e. $\Pr(y_i = 1 | x_i) = x_i \beta$. The basic idea of the linear probability model is to look up for a linear combination of the predictor variables. The assumption here is that x (the explanatory variable(s)) is fixed and $E(\varepsilon) = 0$. several problems arise from this however, these are:

- There is no guarantee that it will produce conditional probabilities that lie in the unit interval $[0,1]$. Although a lot of effort has been made to constrain the predictions of the linear probability model to the unit interval [0, 1], they are considered ungainly and estimating an alternative model by the method of Maximum Likelihood is considered preferable. Moreover, these techniques (of constraining the predictions of the linear probability model to the unit interval [0, 1]) are usually data dependent and so the resulting estimates have no sampling properties. Under such circumstances no meaningful inference is made.
- The residuals have no known distribution, at least in finite samples although, asymptotically, they are normally distributed.
- The spread of the errors is not uniform/constant, it is data-dependent (heteroskedasticity). This makes the results of tests of significant unreliable.

2.2.3 Probability Models for Binary Outcome

Consequent to the problem(s) of the linear probability models, probability models have evolved to solve the non-probability of predictions problem of the linear probability models, unlike the linear probability models, the probability models estimate coefficients by the method of Maximum Likelihood. This amounts to building models satisfying the following conditions:

i. $\lim_{x_i, b \to +\infty} p(y_i = 1) = 1.$

ii. $\lim_{x_i, b \to -\infty} p(y_i = 1) = 0.$

The probability models find a function F that maps values of the explanatory variables from the real line to the unit interval $[0,1]$ i.e. $F(X): \mathbb{R} \to [0,1]$

Two cumulative distribution functions that probability models use for this purpose are the normal probability distribution function and the logistic function. A model that uses the normal probability distribution function is referred to as a probit model.

$$The\ probit: F(x) = P(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} exp\left[-\left(\frac{t^2}{2}\right)\right] dt \tag{2.8}$$

$$\text{The } logit: F(x) = L(x) = \frac{1}{1+e^{-(\beta_0+\beta_1 x)}} \tag{2.9}$$

Thus, in general, the model for the binary response is written as $P(Y=1|\boldsymbol{X}=\boldsymbol{x}) = F(\boldsymbol{x})$. Here, $\boldsymbol{x}$ is either a single predictor variable or a vector of several predictor variables.

Both the logit and the probit calculate the probability of the event of interest occurring and set a threshold to discriminate between the two classes. A new observation is then classified as being the event or not the event based upon the threshold. The good thing here is that the threshold can be adjusted based upon the level of accuracy desired by the researcher. The common decision boundary that people choose is: classify a new observation, O_{new} as the event if $P(Y=1|X=O_{new}) \geq 0.5$.

Whereas the logit is more popular in the health sciences, probably due to the fact that its coefficients can be interpreted in terms of odds ratios, the probit is used by some economists and political scientists due to the fact that it can handle the problem of heteroskedasticity (Albright, n.d.). One issue with logistic regression is that it relies on sufficiently large samples to give a reliable measure of the goodness-of-fit of the model. In conclusion, even though the linear probability models do not assume any linear relationship between the independent variables and the response, they attempt to fit a linear relationship between a link function of the response and

the independent variables. In recent times, the large and complex nature of the data sets (in which classification is desired) coupled with advancement in computer software power, fitting a linear function between the independent variables and link function of a sort is no longer desired. A non-linear function that best discriminates between the classes preferable. A prudent decision is to model the variation in the dependent variable with respect to the existing clusters in the data, where a cluster can best be defined as a set of observations that has an appreciable number of similar characteristics.

2.2.4 Decision Trees

Decision trees are hierarchical arrangements of stages used to classify objects/items/ data points into pre-defined populations/classes/categories. At each of these stages, a decision is taken to either assign an object to one population or proceed to the next stage of decision making. In a decision tree, each stage of decision making is referred to as a node. The initial node of the tree is referred to as a root whilst the last node is referred to as a leaf. All other intermediary nodes are referred to as branches. At each node, a decision is taken on the object based on a set, say t, of value of one independent variable. The threshold t partitions the set of the predictor into two disjoint sets $(-\infty, t)$ and $[t, \infty)$. More complex trees can be built by partitioning the set into more than two disjoint sets. The decision is either to assign the object to one of the pre-defined populations or refer it to the next node where a decision would be taken based on another predictor variable, the process continues until a suitable stopping criteria is satisfied.

Consider a scenario in which objects $O_i(x_1, x_2, x_3)$ are to be classified into two pre-defined populations π_1 and π_2 based on three predictors x_1, x_2, and x_3, Figure 2.1 illustrates the procedure.

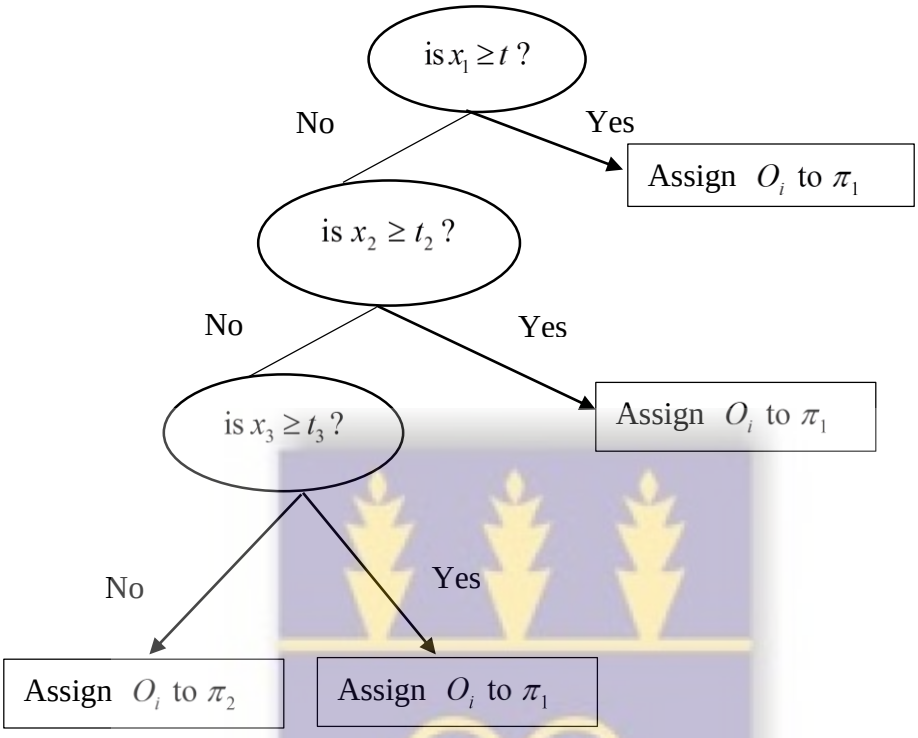

Figure 2.1: A Decision Tree

This is only an example of a decision tree. This decision tree makes it look as if decision trees can be applied only when dealing with continuous variables but this is not the case. In general, a decision tree can accept categorical predictors or continuous predictors or both. Various decision tree algorithms use various measures to determine which variables are involved at the root node. Some of these measures include:

- Information Gain
- Gini Index
- Gain ratio

In their paper entitled "Data Mining Applications: A comparative Study for Predicting Student's performance", Yadav *et al.*, (2012) listed 12 different decision tree algorithms, they are:

- CLS

- ID3
- IDE3+
- C4.5
- C5.0
- CART
- Random Tree
- Random Forest
- SLIQ
- Public
- OCI
- Clouds

Continuous splitting of the data into several subsets to have homogeneous collection of objects in each set can result in over fitting. To prevent this, a process called pruning is carried out after the tree has been built. Pruning is the removal of unnecessary intermediary nodes in order to remove noise and outliers. This generalizes the model and increases its predictive accuracy on new queries (data points).

2.2.5 Artificial Neural Networks (ANN)

An Artificial Neural Network is a system of interconnected neurons that exchange messages among themselves. An Artificial Neural Network is a learning model designed to mimic the process by which biological nervous systems communicate to effect an action. It is a very computer-intensive algorithmic procedure used to approximate non-linear functions of the input variables (Johnson and Wichern, 2007). An ANN is made of three types of layers of neurons; an

input layer, hidden layer(s), and output layer. A layer is made up of smaller units called neurons. Each neuron is a node/site in the network that receives and transmits input signals. A relation links neurons of successive layers. The strength of this links is determined by synaptic weights. Mathematically speaking, the neural network is a composite function; the input layer admits the input of the function and the output layer produces the output. The output of the input layer is the input of the hidden layer, and as well, the output of the hidden layer is the input of the output layer. When there are several hidden layers, the output of one hidden layer is the input of the next hidden layer. The hidden layer or layers are the levels at which the computations are done. The synaptic weights are the parameters of the relations that link an input from one layer to the next layer in the network. These parameters are estimated by means of a training data set. The job of a network algorithm is to learn the training set and produce a function that best approximates the unknown function that maps inputs to outputs in the training set. In training, for given inputs of the training set, the network's outputs are compared to the actual outputs and the error (the difference between the two) is computed. The error is communicated to the input layer from the output layer through the hidden layers by a process called backward propagation. This communication prompts the network to adjust the weights (between the layers) in a direction that would minimize the error. This adjustment of weights continue until such a time when no further adjustment improves the result. A network architecture is the organization of the nodes (in the layers) and the types of connections permitted between them (Johnson and Wichern, 2007). There are two types of network architecture, feed-forward neural networks and feed-back neural networks. In feed-forward networks, neurons in one layer are only connected to neurons in the succeeding layer, connections to neurons in the same layer or preceding layer is not permitted. A feed-back network, however, permits connections in all directions.

Figure 2.2 illustrates a feed-forward neural network that has n neurons in its input layer, 2 neurons in its hidden layer and 2 neurons in the output layer. This design can be used to solve a binary classification problem, where only one of the two neurons in the output layer will be "excited" based upon a predefined rule

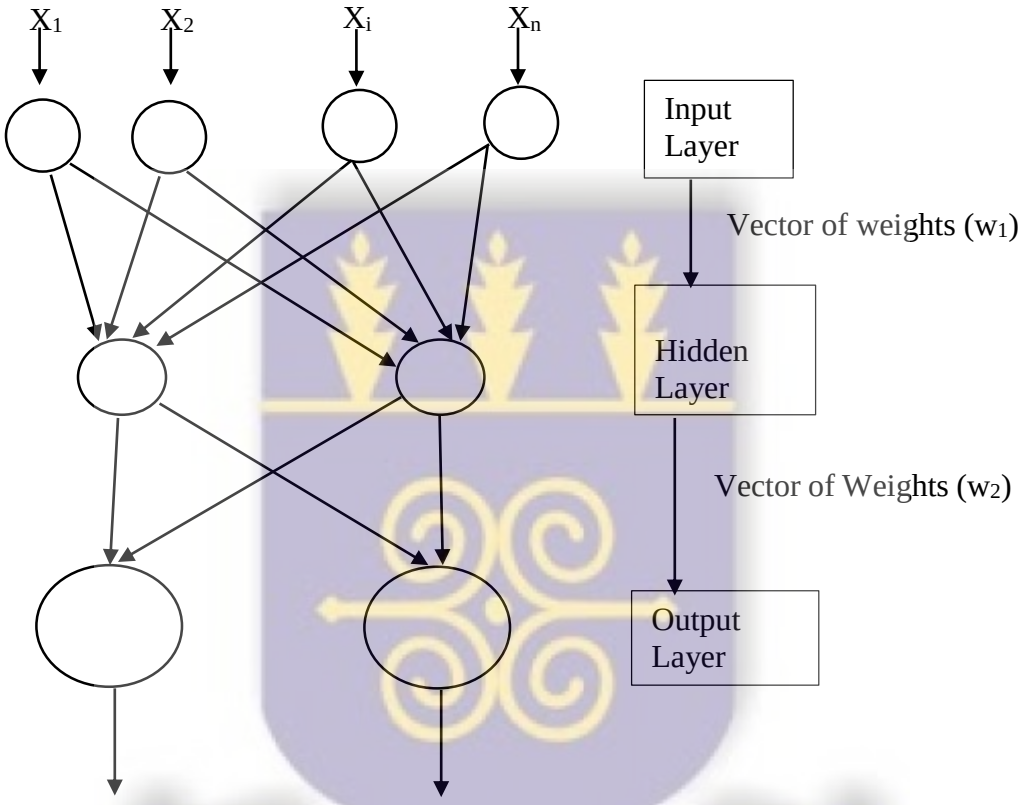

Figure 2.2: A Simple feed-forward Neural Network

2.3 Review of Related Works

2.3.1 Non-Linear Classification Models

The focus of this subsection of the literature is to highlight a cross-section of classification works achieved via pattern recognition/data mining techniques.

Odeh *et al*., (2006) compared the performances of logistic regression, artificial neural networks and adaptive neuro-fuzzy inference system in predicting credit default using data from Farm Credit System. They examined the performances of the three models based on their respective sensitivity and specificity measures. Defining sensitivity as a percentage measure of the proportion of default cases correctly identified by the model to total number of default cases in the entire sample and Specificity as the measure of the proportion of non-events correctly identified by the model to total number of non-events in the sample, they obtained the results displayed in Table 2.1.

Table 2.1: Summary of model performances by Odeh *et al*., (2006)

	LReg	ANFIS	ANN
In-Sample			
Sensitivity	44.42	69.87	61.51
Specificity	53.95	64.71	60.42
Out-of-Sample			
Sensitivity	90.77	51.97	43.78
Specificity	5.32	49.62	76.85

They concluded that prediction performances vary with model used but could not tell which model was the best. They also mentioned that the performance of these models may change given a new

data set. This conclusion of theirs opens up the discusion on finding a model that would perform best (if possible) in predicting the credit default.

Lahiri (2006) also examined the performances of three statistical and data mining techniques namely: logistic regression, decision tree, and Neural Networks. Each of the techniques were used to build models for 8 different sample sizes with two different sampling methods. It was found that for a given data set, none of the three models outperformed the others. He stated however that the absolute value of prediction accuracy varied among the three data sets indicating that the data distribution and data characteristics play a role in the actual prediction accuracy. This finding of Lahiri consolidates what has been suspected by Odeh *et al*., This further suggests that for any type (domain) of data, it would be prudent to examine the performance of several classifiers in order to know which classifier performs best on it. The effect of sample size and sampling method were also examined against two sets of test data and it was found that neither of them had significant influence on the accuracy of the predictions.

Martin (2009) developed an Artificial Neural Network to predict maize yield based on large-scale meteorological phenomena. He used Genetic algorithms to find the optimal subset of inputs for the model. The data set of 385 patterns were divided into three, a development set, selection set, and evaluation set. Martin used the development set to train the ANN models and determine the fitness of individuals for the Genetic Algorithm and used the selection set to determine when training should be terminated to avoid overtraining and to select ANN parameters. The evaluation set was used to test the performance of the model. The study found that indicators of large-scale meteorological phenomena can be successfully used for modeling crop yields. Martin further concluded that "the improvement in model accuracy when using GA-selected subsets of inputs indicates that GA/ANN hybrids are useful tools for meteorological and agricultural applications". While Martin (2009) made this observation for meteorological and agricultural applications, the

challenge still remains in research to investigate how this hybrid would to in other areas of application. His work ended by recommending that future research should apply these methods to predicting crop (not only maize) yield in other geographical locations.

Ghatge and Halkarnikar (2012) in a paper entitled "Estimation of Credit Risk for Business Firms of Nationalized Bank by Neural Network Approach" used Back Propagation neural network and expert evaluation to predict the credit risk category of business firms of Nationalized Bank. Using a data set of 23 instances, they trained the network with 14 and kept nine (9) for testing. Their data consisted of three (3) categorical variables and 10 continuous variables. Having found that the Neural Network showed more accurate results than the bank's own manual calculation, they concluded that the Artificial Neural Network is the best method for estimating credit risks in banks.

A very common application of pattern recognition is human face recognition. In this case the training set is a collection of several images of an individual, and a new face is recognized based on the training set. In such applications the test set is made up of different images of the same individual that was not included in training. It is worth reiterating that, pattern recognition is the bigger domain that encompasses the various sub- domains such as detection, classification, and face recognition.

Alkalin (2003) developed a face authentication neural network based on Principal Component Analysis. For any given individual, several images are taken and the principal components of the person's face are extracted. These components of each individual are fed into the input layer of the neural network. To authenticate a new test image after some pre-processing has been carried out, it is simulated with all saved principal components of the network. The image with maximum output is selected, when it satisfies some pre-defined threshold, it is reported as the host of the input face, otherwise it is reported as unknown and added to the face library.

2.3.2 Predicting Students' Performance in Examinations

Minaei-Bidgoli *et al*., (2003) employed a number of classifiers to predict the classes of the final grades of students in a an online course, namely, "introductory physics course for scientists and engineers". These are Quadratic Bayesian classifier, 1-nearest neighbor (1-NN), k-nearest neighbor (k-NN), Parzen-window, multilayer perceptron (MLP), and Decision Tree. They hold the view that classification is a task that finds applications in a variety of fields and as such, it is impossible to find a single classifier that can give good results on all data sets. They considered the following independent variables in predicting the classes of the final grades of students in the course:

1. Total number of correct answers
2. Getting the problem right on the first try, vs. those with high number of tries. (Success at the first try)
3. Total number of tries for doing homework. (Number of attempts before correct answer is derived)
4. Time spent on the problem until solved (more specifically, the number of hours until correct. The difference between time of the last successful submission and the first time the problem was examined). Also, the time at which the student got the problem correct relative to the due date. Usually better students get the homework completed earlier.
5. Total time spent on the problem regardless of whether they got the correct answer or not. (Difference between time of the last submission and the first time the problem was examined)
6. Participating in the communicating mechanisms vs. those working alone LON-CAPA provides online interaction both with other students and with the instructor. Were these used?

These variables are mentioned in this work since, this thesis, even though it is mainly about developing a classifier, it is also about investigating students' variables that are significant in

achieving academic success. Minaei-Bidgoli *et al*., conducted this study with the aim of finding answers to the following two major research questions as qouted directly from their own paper:

1) Can we find classes of students? In other words, do there exist groups of students who use these online resources in a similar way? If so, can we identify that class for any individual student? With this information, can we help a student use the resources better, based on the usage of the resource by other students in their groups?
2) Can we classify the problems that have been used by students? If so, can we show how different types of problems impact students' achievement? Can we help instructors to develop the homework more effectively and efficiently?

They classified students according to their grades in three different ways. In the first instance, the response variable was modelled as a nine-outcome variable, in the second instance, it was modelled as a three-outcome variable, and in the third case, and the response was treated as a binary variable. In the third case, a student either passes or fails. The classifiers were applied in all these three segmentations of the response variable. All their classification procedures involved some pre-processing (normalization). The error rates of all classifiers were calculated in each case after which the combination of multiple classifiers (CMC) method was used to improve classification performance. Their study elucidated two methods of classifier combinations, namely, offline CMC and online CMC. They modified CMC and recorded significant improvement in classification accuracy in all three segmentations of the response variable. In a lesser-known application, they employed genetic algorithm in finding optimal combinations of classifiers that maximize classification performance in all three instances. Their results show that using the genetic algorithm to combine classifiers gives better results compared to not using it.

Regarding the performances of the individual classifiers in the cases where the response has been categorized into 2, 3, and 9 their results are recorded in Table 2.2

Table 2.2 : Results of Minaei-Bidgoli *et al*., (2003) on Performance of Classifiers

Case	Best Classifier	Percentage Accuracy
2-classes	K-NN	82.3%
3-classes	CART	60%
9-classes	CART	43%

The key lessons learnt from the work of Minaei-Bidgoli *et al*., are summarised as follows:

i. Combination of multiple classifiers improves classification accuracy.
ii. The use of genetic algorithm in classifier combination improves predictive accuracy.
iii. In cases where number of features is low, feature weighting is better than feature selection.
iv. Data mining approach to predicting students' performance is a good means of improving academic performance amongst student, seeing that it reveals latent variables that otherwise, would go unnoticed.

They finally suggested that more students should enter the online learning environment so that data mining would serve a useful purpose in studying factors affecting students' performance.

Yadav *et al*., evaluated the performances of three (3) decision tree algorithms in predicting the end of semester performance of university students in a course. The explanatory variables they used include: Attendance, Class test, lab work, Seminar and Assignment marks. Their response variable has four outcomes. Marks ≥ 60% were labeled "First", marks ≥ 45 &<60% were labeled "Second

marks ≥ 36 &<45% were labeled "Third", and marks < 36% were labeled "Fail". The algorithms used for classification is ID3, C4.5 and CART. They used an evaluation approach called the 10-fold cross-validation to evaluate the performances of the algorithms since they did not keep any portion of the data set for validation. Their study found that the CART (Classification and Regression Trees) algorithm has the highest predictive accuracy with 56% of correctly classified instances. The performance of the ID3 was next to CART 52% of correctly classified instances. The C4.5 algorithm classified 45.8% of the instances correctly.

In 2013, Wushishi and Usman used Pearson product moment correlation coefficient to find the relationship between Senior Secondary School Certificate Examination Mathematics grades and final Nigeria Certificate of Education (NCE) Mathematics students' results of Niger State College of Education Minna. They collected data from examination records of 67 students (45 males and 22 females). Their hypothesis test result rejected the null hypothesis that "There is no significant relationship between WASSCE entry grades in mathematics and academic achievement of final NCE mathematics students' result" at 0.5 level of significance. Their findings also indicate that gender is significant in predicting the students' academic achievement in the final NCE mathematics result with male students doing better than female students. This focused only on studying the relationship between entry Mathematics grades and final Mathematics grades.

In 2014, Agbaje and Alake used questionnaires and past academic performance records in three science subjects (Biology, Chemistry and Physics) to study the predictive strengths of students' variables (students' gender, study habit, attitude to and interest of students in Science subjects) in predicting students' academic performance with a sample of 125 students in Ikere Local Government Area of Ekiti state, Nigeria. Their analysis was done using Pearson's Moment Correlation and Multiple regression analysis. Their finding was that study habit, attitude to and

interest of students in Science subjects are better predictors of students' performance in science subjects, while student gender has no influence on students' academic performance. It is surprising how Agbaje and Alake in studying the performance of science students did not include Elective Mathematics, seeing that the response variable (academic performance) is measure of a combination of students' performance in three subjects (Biology, Chemistry and Physics) and the fact that Elective Mathematics in most Senior High Schools is compulsory for science students. The fact is that the constitution of science subjects such as Chemistry and Physics is such that Mathematics as an academic subjects is an indispensable tool. The paper of Agbaje and Alake also did not mention how the variables were measured. Their study would have been more comprehensive if they outlined the range of values that the various levels of students' interest, students' study habit and students' attitude could assume.

In 2008, Omirin and Ale tested the hypotheses:

1. There is no significant relationship between the performances of students in mock and WASSCE English.
2. There is no significant relationship between the performances of students in mock and WASSCE Mathematics.

They analyzed the relationships between the results of the mock examinations and the WASSCE by finding the correlations between the two. A simple linear regression model of the WASSCE results on the mock results was also fitted. Data was collected on 306 students from 12 public secondary schools. Their study finds that the mock examinations results in English and

Mathematics are positively correlated with the WASSCE results in English and Mathematics respectively.

2.4 Evaluation of Classifiers

The most simplistic way to evaluate the performance of a binary classification algorithm is to divide the data set into two; a training set and a validation set ; train the classifier on the training set and then use the validation set to estimate the performance of the classifier. This evaluation method is common with binary classifiers. The discussion here would for the sake of convenience restrict attention to dichotomous classifiers. The most basic statistic used in evaluating classifier performance is predictive accuracy. The predictive Accuracy of a classifier is defined as the number of observations correctly predicted divided by the total number of observations in the sample. It is often expressed as a percentage. This statistic gives a general idea of the performance of the classifier but can often times mask relevant information from the end-user of the classifier's forecast. If, for example, a data set contains 100 failures and 900 successes and the classifier predicts 2 failures correctly and 850 success correctly, then the classifiers accuracy would be $\frac{2+850}{1000}\times 100\%$. This is equal to 85.2%. This result looks good but does not tell anything about the classifier's specific ability in predicting success or failure. An inspection of the outputs however reveals that the classifier is considerably bad at predicting failures. It may turn out that, in the application domain, the cost incurred in wrongly predicting a true failure as success is much higher than that of predicting success as failure. If the relative proportions of success and failure in the data were close, a high value of accuracy would have meant that the classifier is good, both in correctly identifying successes and failures. In the language of classifier evaluation, the proportion of the outcome of interest that are in the data is referred to as **prevalence** and so statistic:

predictive accuracy is said be biased because it is prevalence dependent. Consequently, the need arises to evaluate a classifier based on its respective abilities in predicting success and failure as well. The ratio of number of successes that are correctly predicted (by the classifier) to the number of successes in the sample is referred to as the **sensitivity** of the classifier. Analogously, the ratio of number of failures that are correctly predicted (by the classifier) to the number of failures in the sample is referred to as the **specificity** of the classifier. The importance of each of these measures is highly a function of the problem/application domain. For example, in a credit lending decision where denying a credit-worthy applicant a loan does not make the lending institution lose considerably, sensitivity of the classifier would be highly desired (this statement is made, based on the assumption that "credit default" is the event of interest in the classification problem). Here, the interest is in detecting non-credit-worthy applicants. A good judgment of the classifier's predictive performance is made when the following four statistic are computed: the number of failures correctly predicted as failures (TN), the number of successes wrongly predicted as failures (FP), the number of failures wrongly predicted as successes (FN), the number of successes correctly predicted as successes (TP). A table that presents these four statistics is referred to as confusion matrix (Elkan, 2102).

Table 2.3: Structure of the Confusion Matrix

Actual Class	**Predicted Class**		**Totals**
	0	***1***	
0	TN	FP	TN+FP
1	FN	TP	FN+TP
Totals	TN+FN	FP+TP	TN+FP+FN+TP

0 represents failure and 1 represents to success.

Elkan (2012) maintains that no single statistic or pair of statistics can completely indicate the performance of a classifier. He therefore suggested that, when presenting a report, it the full confusion matrix should be explicitly presented to allow readers to compute their own statistics.

Much as sensitivity and specificity give specific information on the performance of the classifier, there are two outstanding questions that neither of them answers. These are: "if the classifier predicts success, how sure are we that the outcome is truly a success?" and "if the classifier predicts failure, how sure are we that the outcome is truly a failure?" These questions are respectively answered by the Positive Predictive Value (PPV) and the Negative Predictive Value (NPV). The positive predictive value is the fraction of the observations predicted as success that are actually success, similarly, the negative predictive value is the fraction of all observation predicted as failures that are actually failures. PPV is also referred to as **precision**, according to Powers (2012), it is also called confidence, in the language of data mining. A single metric used to measures a classifier's performance is the F-score. F-score employs precision and recall (the number of correctly predicted positive successes divided by the number of success in the sample) to compute a fraction. An F-score of 1 indicates that the classifier is flawless while an F-score of 0 indicates that the classifier is completely flawed. For a real positive number β , F_β- score is defined as:

$F_\beta\text{- score} = (\beta^2+1)\dfrac{\text{precision}\times\text{recall}}{\beta^2(\text{precision}+\text{recall})}$. Van Rijsbergen (1979) wrote that F_β "measures the effectiveness of retrieval with respect to a user who attaches β times as much importance to recall as precision". Thus, the balanced F_1- score for a receiver who attaches equal importance to both recall and precision is $2\dfrac{\text{precision}\times\text{recall}}{\text{precision}+\text{recall}}$. This is interpreted as the harmonic mean of precision

and recall (Lozano *et al*., 2010; Powers, 2012). The problem with F_1- score is that, it the not take TN, which according to Powers (2012), can vary freely with time, into consideration.

While partitioning a sample into a training set and a validation set comes across as a simple means of evaluating classifier performance, it can lead to loss of information about the model's parameters. This danger is high when dealing with relatively small data sets. This is because, usually, a sufficiently large data set is needed to produce reliable parameter estimates. A more sophisticated means of evaluating a classifier's performance is to partition the data, D into smaller number, n of subsets $S_1, S_2, ..., S_i, ..., S_n$ ensuring that each of these subsets is of some constant size, k. When this is done, then at any instance i, the classifier is trained on the complement of S_i in D and the performance metrics discussed earlier are computed. Finally, the average of each metric is computed over all the n times that the classifier has been trained and quoted as the metric of the classifier. The question arises as to what should be the size of each of the subsets S_i to produce optimal results. Intuitively, an optimal result would be produced when the size of each of these subsets is 1. This means that at any time of training the classifier, all but one of the data points should be used. This, however, is computationally infeasible (Elkan, 2012). Elkan (2012) wrote: "In recent research the most common choice for k is 10".

With this methodology of evaluation, a more reliable means of analyzing classifier performance is by means of R.O.C. curve. R.O.C. stands for Receiver Operating Characteristic. The ROC curve of a classifier is the graph of the classifier's True Positive rate (plotted on the vertical axis) against its False Positive rate (plotted on the horizontal axis) for the various n training samples. The coordinates of the point of intersection of the True Positive rate and the False Positive rate is (0,0). The entire plane in which the ROC curve lies is the rectangle, R, whose vertices are (0, 0), (1, 0),

(1, 1), and (0, 1). The coordinates of a perfect classifier would be (0, 1), that is False Positive rate equals 0 and True Positive rate equals 1. Analogously, the coordinates of a worst classifier would be (1, 0). The leading diagonal of the rectangle R divides the ROC space into two; points above it indicate that the classifier is better than chance and points below it indicate that the classifier is worse than chance. A point on the diagonal indicates that the classifier classifies purely by chance.

Powers suggests that classifier performance should be evaluated base on unbiased measures such as the ROC curve analysis, informedness, markedness, and correlation.

In this thesis, due to the limitations of time and space, only specificity sensitivity and accuracy were used to evaluate classifier performance but the study continues beyond this thesis.

2.5 Summary of Revelations of the Literature Review

- Research has admitted that traditional linear models of classification such as discriminant analysis and logistic regression have limitations.
- The predictive performance of each method of classification depends on the type of data used, no method is universally superior.
- Non-linear pattern recognition methods have being used more in other areas such image recognition and loan default prediction compared to their use on educational data.
- Some classification methods produce a probability forecast (which is consequently used to predict the class) of the likelihood of an observation belonging to one of the predefined classes of a categorical variable while other methods produce a point forecast.
- Educational data mining is still an evolving domain.

- Studies on students' performance only consider modelling performance in individual subject areas.
- Most of the published works on students' performance may not have been done in Ghana

2.6 Gaps in Literature that the Study Seeks to Address

Consequent to the revelations of the literature review, the study seeks, amongst other things to:

1. Develop a point classifier that also outputs a probability value that would signify the strength/reliability of the forecast.
2. Contribute to the knowledge that educational data mining is bringing to bear in the educational industry.
3. Measure/ model students' academic performance holistically, as the totality of students' excellence in all subjects that he/she studies.
4. Measure students' performance by his/her chances of progressing to the next level on the academic ladder and not just by his/her success in some subject.

3. METHODOLOGY

3.1 Chapter overview

This chapter is in 3 sections, section 1 describes the data used for the study; it entails the method of data collection, description of variables in the data, and the rational for choice of explanatory variables. Section 2 describes the two methods of analysis employed in the study; the proposed method and the logistic regression method. Section 3 defines the measures on which the classification performances of the methods shall be compared.

3.2 Data Collection and Description

3.2.1 Data Acquisition

The data for this study was obtained from the records of Keta Senior High Technical School, situated in the Volta Region of Ghana. It comprises 11 variables of students who wrote the West Africa Senior School Certificate Examination in the year 2014. These are:

1. Student's qualification for admission into the university (it is shortened as "Admit")
2. Program of Study of student
3. Residence of student
4. Gender of student
5. Student's B.E.C.E aggregate (Entry grade)
6. End of third term marks in English Language in the SHS 1
7. End of third term marks in Core Mathematics in the SHS 1
8. Average of marks obtained by a student in all other subjects apart from English Language and Core Mathematics at the end of the third term of SHS1

9. End of third term marks in English Language in the SHS 2
10. End of third term marks in Core Mathematics in the SHS 2
11. Average of marks obtained by a student in all other subjects apart from English Language and Core Mathematics at the end of the third term of SHS1

The students' qualification for admission into the university was calculated from the grades obtained by the student in the WASSCE. A WAEC master sheet of grades of all students who wrote the examination was obtained from the school's administrative office. Based upon the NAB's admission criteria, the student's qualification for admission was determined (calculated). The variables: program of study, residence, gender, were obtained from files of the students. The remaining 6 variables were also obtained from the carbon copies of the terminal reports of the students. The full data set is presented at the appendix in Table 7.1

In the data presented in Table 7.1, the variables have been coded, Table 3.1 gives the actual variables and the codes associated with them for easy comprehension.

Table 3.1: Meanings of Variables

Variable number	Variable code	Meaning of code
1	Std	Row number of student in the data set
2	Admit	Whether or not the student qualifies to be admitted to pursue a bachelor's degree program in the university
3	prog_code	A number representing the program that the student had studied
4	resid_code	a number indicating the type of residence 1 if student is a boarding student and 0 if day student
5	gen_code	The gender of the student, 0 if female and 1 if male
6	Entry Grade	A number indicating the aggregate that the student brought to Senior High School
7	Eng1	The marks(in percentage), in English language, that the student has gotten for first year third term
8.	Math1	The marks(in percentage), in Mathematics, that the student has gotten for first year third term
9.	f_av	Arithmetic mean of the marks in the remaining 5 or 6 subjects that the student took in first year third term
10.	Eng2	The marks(in percentage), in English language, that the student has gotten for second year third term
11.	Math2	The marks(in percentage), in Mathematics, that the student has gotten for second year third term
12.	s_av	Arithmetic mean of the marks in the remaining 5 or 6 subjects that the student took in second year third term

Source: Own research

Table 3.2: Computation of Response Variable

Student	Math	English	Social	Int. Science	Chemistry	Physics	Biology	E-Math	Admit
1	B3	D7	A1	B2	C5	B2	A1	B2	0
2	C5	C6	B3	B2	C6	C5	C6	C5	1
3	B2	C4	B2	A1	B2	D7	D7	B3	0

1. The **prog_code** is one of the numbers 1,2,3,4 indicating which program the student pursued in the Senior High School. In the school for which the data was taken, there are seven (7) programs of study namely,
    a. Agricultural Science
    b. Business
    c. General Arts
    d. Home Economics
    e. Science
    f. Technical
    g. Visual Arts

These programs have been re-grouped as four new programs. The Business program is maintained as Business, the General Arts program is maintained as General Arts, the Science program is maintained as Science and the Agricultural Science, Home Economics, Technical and Visual Arts programs are now grouped as Vocational. The rational for this re-categorization is researcher's knowledge of the similarity that exist in the values of (the variables that influence) the response variable of interest (Admit). Additionally, it is known that the programs Home Economics and Visual Arts offer some elective subjects in common namely, General Knowledge in Arts,

Chemistry, and Biology. The Agricultural Science and the Technical students also read Elective Mathematics, Chemistry and Physics in common. One other reason for putting these four programs in one common group is the relatively smaller number of students (compared to the Business, General Arts and Science programs) who wrote the WASSCE from these programs taking cognizance of the fact that the method of data analysis deals with the existing clusters in the sample and smaller number of cluster sizes would not produce very meaningful and generalizable results/conclusions. Table 3.3 illustrates the re-grouping of the programs and their respective codes

Table 3.3: Re-grouping of the Programs of Study

No.	Program Name	Program group	code
1	Agricultural Science	Vocational	4
2	Business	Business	1
3	General Arts	General Arts	2
4	Home Economics	Vocational	4
5	Science	Science	3
6	Technical	Vocational	4
7	Visual Arts	Vocational	4

2. **Entry Grade** is the aggregate of the grades of the students in English, Mathematics, Integrated, and Social studies and best two grades inn two other subjects in the Basic Education certificate examination. The B.E.C.E uses a 9 point scale in grading the candidates in each subject with 1 being the highest grade and 9 being the lowest grade, thus the best aggregate a candidate can get is 6 and the worst grade is 54. Practically, B.E.C.E candidates enter into the Senior High Schools with aggregates as worse as 40- 42.
3. Eng1 is the total of the marks obtained by the student in English Language in the third term of the first year expressed as a percentage. All class works, home works, group works etc. constitute 30% percent of this mark while the end of third term examination score constitute 70%.

4. Math1 is the total of the marks obtained by the student in Mathematics in the third term of the first year expressed as a percentage. All class works, home works, group works etc. constitute 30% percent of this mark while the end of third term examination score constitute 70%.
5. f_av is the average of marks obtained by the student in the third term of the first year in the remaining subjects (after taking out English Language and Mathematics), that Social Studies, Integrated Science and the elective some subjects. Some students read 3 elective subjects whilst most students read 4 elective subjects

3.2.2 Rational for Choice of Predictor Variables

Basically, the researcher sets out to find the association between the variables chosen and the response. There are also some factors motivating this choice. The choice of the continuous variables is because they are the direct analogous of the variables that are used in the determination of the response variable since the one goal of the study is to investigate the extent of correlation between the student performance in the school and the performance in external examination. The two variables: English Language marks and Mathematics marks have been chosen as independent variables because they are two variables that a student must perform well in so as to enable him/her to qualify for admission into the university. From the admission requirements, a student who obtains a grade worse than C6 in either English Language or Mathematics automatically does not qualify for admissions. The study therefore chooses these variables in order to investigate the relationship between students' performance in these subjects in the internal examinations and the external examinations. The main reason for choosing the end of third term marks is that it is more representative of the students' academic performance since it is a promotion examination and the students give their bests in the examinations in order to get promoted to the next level. Also in the

first year, most students do not report to school in the first term. The same reasoning goes into choosing the second year third term marks also. Now, the reason for including both the end of first year and the end of second year result for the same subject is to investigate the existence and possibly the strength of any association between the two, and to track the students' progress in the subject over the academic terms.

3.3 Data Analysis

Summary statistics were calculated for all variables. Of the 241 observations, 191 (representing approximately 79% of the observations) were systematically sampled to constitute the training sample whilst the remaining 50 observations (representing approximately 21% of the observations) are removed to test the models. The test set is constituted to, as much as possible, include observations in all the 16 clusters in the data and also to represent approximately 21% of the observations in each cluster. The definition of a cluster as applied to this study is given in section 3.3.1.1. In each cluster the ratio of the size, N of the cluster to the pre-determined sample size, n was calculated and rounded to an integer k. A number j, lying between 1 and k was randomly selected to give the sample $\{j, j+k, j+2k, ..., j+(n-1)k\}$. This systematic sampling method is described in "Applied Statistics" (Shayib, 2013).

Two methods of analysis have been used to investigate the response variable; the proposed pattern recognition method and binary logistic regression. In the remainder of the subsection, each of the methods is described; the major mathematical/statistical techniques involved are also described and illustrated.

3.3.1 The Pattern Recognition Approach

The Pattern Recognition Approach to the binary classification task in this study employs the concept of Principal Components Analysis (PCA). PCA is used to extract the important attributes of the observations, the Euclidean distance of the feature vector of an observation from the mean vectors of the pre-defined classes is used in classifying the observation.

The stages involved are clustering/subgrouping (where all observations with the same value of all the categorical variables are grouped as one), pre-processing, feature extraction, discrimination and classification. The major statistical tools used at each of these stages are described subsequently.

3.3.1.1 Clustering/Grouping

In this study, a cluster refers to one of all the possible combinations of the levels of the categorical variables in the data. Observations that have the same values for all the categorical variables constitute one cluster. Given a data matrix of n number of categorical variables where the first categorical variable has n_1 number of levels, the second categorical variable has n_2 number of levels, the k^{th} categorical variable has n_k number of levels, and the n^{th} categorical variable has n_n number of levels, then by the basic principle of counting, there exists $n_1 \times n_2 \ldots \times n_k \times \ldots n_n$ combinations of the levels of the categorical variables (referred to in this study as clusters). In this study, there are three categorical variables. They are: program of study, residential status, and gender. Program of study has 4 levels, residential status has 2 levels, and gender has two levels. This results in a total of 16 clusters of the observations in the study. Table 3.4 gives the definitions and labels the clusters.

Table 3.4: Definition and Labelling of the Clusters

Program	Residence	Gender	Cluster number
1	1	1	1
1	0	1	2
1	1	0	3
1	0	0	4
2	1	1	5
2	0	1	6
2	1	0	7
2	0	0	8
3	1	1	9
3	0	1	10
3	1	0	11
3	0	0	12
4	1	1	13
4	0	1	14
4	1	0	15
4	0	0	16

The codes in columns 1, 2, and 3 of Table 3.4 are as defined for the categorical variables in Section 3.1.1.

3.3.1.2 Preprocessing

In general, preprocessing refers to any of several modifications that can be performed on the input data with the intent of simplifying subsequent operations that are to be carried out on the data without losing relevant information about the data (Corso, 2013). There are several preprocessing techniques, the choice of a particular preprocessing technique is usually dependent on the characteristics of the data. In this study, the preprocessing technique employed on the data is **Mean Centering.** To mean center a column of a continuous variable simply means to compute a new column whose entries are given by the original entry minus the mean of the column. Given a data set $X = (X_1, X_2, ..., X_p)$ of predictor variables, the mean centered data denoted W is given by

$$W = X - \bar{X} \tag{3.1}$$

Where $\bar{X} = \begin{bmatrix} \bar{X}_1 \\ \bar{X}_2 \\ \vdots \\ \bar{X}_P \end{bmatrix}$. For the purpose of illustration, consider the column $C_1 = \begin{bmatrix} 14 \\ 18 \\ 25 \\ 17 \\ 11 \\ 13 \end{bmatrix}$ then, the mean of C_1 is calculated as $\bar{C}_1 = \frac{14+18+17+25+11+13}{6} = 16.3$ and the mean centered column is

$$W1 = \begin{bmatrix} 14-16.3 \\ 18-16.3 \\ 25-16.3 \\ 17-16.3 \\ 11-16.3 \\ 13-16.3 \end{bmatrix} = \begin{bmatrix} -2.3 \\ 1.7 \\ 8.7 \\ 0.7 \\ -5.3 \\ -3.3 \end{bmatrix}$$

The purpose of mean centering is to remove noise from the data.

3.3.1.3 Feature extraction

Significant features (predictor variables) of the data points were identified by means of Principal Component Analysis (PCA). PCA is a statistical technique used to summarize the overall variation in the data. It measures the extent of variation that respective features/variables/attributes of the data constitute to the overall variation in the data.

Consider the data matrix $X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}$ of p continuous variables $X_1, X_2, \ldots, X_p$ observed for n observations $O_1, O_2, \ldots, O_n$. Then:

$x_{11}, x_{12}, \ldots, x_{1p}$ are the components of O_1, that is, $\mathbf{O_1}$ is the vector $(x_{11}, x_{12}, \ldots x_{1p})$,

$x_{21}, x_{22}, \ldots, x_{2p}$ are the components of the vector $\mathbf{O_2}$, that is, O_2 is the vector $(x_{21}, x_{22}, \ldots x_{2p})$

$x_{k1}, x_{k2}, \ldots, x_{kp}$ are the components of the vector $\boldsymbol{O_k}$, that is, O_k is the vector $(x_{k1}, x_{k2}, \ldots x_{kp})$

$x_{n1}, x_{n2}, \ldots x_{np}$ are the components of the vector $\boldsymbol{O_n}$, that is, O_n is the vector $(x_{n1}, x_{n2}, \ldots x_{np})$

While it is common to find correlations between pairs of the columns of X, a set of new axes which is then used to transform data X to a new data, say $\mathbf{X^*}$ such that:

1. The columns of $\mathbf{X^*}$ are uncorrelated
2. The columns of $\mathbf{X^*}$ account for decreasing proportions of the total variation in $\mathbf{X^*}$

The components of the new data, $\mathbf{X^*}$, satisfying the above conditions are referred to as the principal components of the original data X. Generally such a matrix can be obtained from the matrix multiplication

$$\mathbf{X^* = X \times L} \tag{3.2}$$

Where $\boldsymbol{L}$ is the matrix of the new axes; $\boldsymbol{L}$ is called the rotation matrix $\mathbf{L}$. $\mathbf{L}$ is the matrix of principal component coefficients. Each column of X^* is referred to as a vector of principal components of $\boldsymbol{X}$. Each row of $\mathbf{X^*}$ represents the principal components of some k^{th} observation, O_k. Every entry of $\mathbf{X^*}$ is a principal component with $\mathbf{X^*}(i, j)$ representing the j^{th} principal component of i^{th} observation in $\boldsymbol{X}$. In general, $\mathbf{L}$ can be obtained from an Eigen Value Decomposition (EVD) of the covariance matrix, $\boldsymbol{\Sigma}$ of $\boldsymbol{X}$ or a Singular Value

Decomposition (SVD) of the data matrix, $\boldsymbol{X}$. In this study, $\mathbf{L}$ is obtained by Singular Value Decomposition. In this case, $\mathbf{L}$ must satisfies the following:

1. The columns of $\mathbf{L}$ must be orthonormal singular vectors of $\boldsymbol{X}$, i.e. each column of L has unit length and is orthogonal to any other column of $\mathbf{L}$
2. The columns of $\mathbf{L}$ are ordered, with the first column accounting for the highest proportion of the overall variance and the last column accounting for the lowest proportion of the overall variance i.e. successive columns account for decreasing proportions of the total variation in the data.
3. Number of rows of $\mathbf{L}$ must be equal to the number of columns of $\boldsymbol{X}$ (number of variables in the data)

The following result from linear algebra supports this technique.

Result 3.1 (Singular Value Decomposition)

For any matrix $X(n \times p)$ $\mathbf{X}(n \times p)$ whose entries come from a field K, which is either the field of real numbers or the field of complex numbers, there exists matrices:

i. $\boldsymbol{U}(n \times r)$ whose columns are the left-singular vectors of $\boldsymbol{X}$

ii. $\boldsymbol{D}$, an $r \times r$ diagonal matrix of singular values of X where $r \leq \min(n, p)$ is the rank of $\boldsymbol{X}$

iii. $\boldsymbol{V}(p \times r)$ whose columns are the right-singular vectors of $\boldsymbol{X}$ Such that

$$\boldsymbol{X} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^T$$

The ordered triple $(\mathbf{U},\mathbf{D},\mathbf{V})$ is referred to as a Singular Value Decomposition of $\boldsymbol{X}$.

In essence, the Singular Value Decomposition theorem guarantees that for any a matrix $\boldsymbol{X}$, a matrix, V of orthonormal right-singular vectors of $\boldsymbol{X}$ can be found, from the decomposition

$\boldsymbol{X} = \boldsymbol{UDV}^T$, where $\mathbf{D}$ is the diagonal matrix of singular values of $\boldsymbol{X}$

Since in (3.2), $\mathbf{L}$ is required to be the matrix of orthonormal singular vectors of $\boldsymbol{X}$, then $\mathbf{V}$ coincides with $\mathbf{L}$.

Assume that the data, $\boldsymbol{X}$ is full rank, then:

i. When $n < p$, $\mathbf{V}$ is a $p \times n$ matrix and $\mathbf{X}^*$ is an $n \times n$ $(n \times p \times p \times n)$ matrix.

ii. When $n > p$, $\mathbf{V}$ is a $p \times p$ matrix and $\mathbf{X}^*$ is an $n \times p$ $(n \times p \times p \times p)$ matrix.

Usually, the structure of a multivariate data is such that $n >> p$.In this study, however, since the analyses are done on the subsets (referred to in the study as clusters) of the data set and some of the subsets have sufficiently small number of observations, it has been the case that $n < p$ in some clusters.

To illustrate the SVD technique, consider a data matrix

$$\boldsymbol{Y}(5 \times 2) = \begin{bmatrix} 65 & 40 \\ 80 & 35 \\ 75 & 30 \\ 90 & 50 \\ 65 & 70 \end{bmatrix}$$

The mean vector of $\boldsymbol{Y}$ is $(75, 45)$

When the columns of $\boldsymbol{Y}$ are mean centered, the resulting matrix is

$$\tilde{\boldsymbol{Y}} = \begin{bmatrix} -10 & -5 \\ 5 & -10 \\ 0 & -15 \\ 15 & 5 \\ -10 & 25 \end{bmatrix}$$

It can be verified that the decomposition

$$\begin{bmatrix} -10 & -5 \\ 5 & -10 \\ 0 & -15 \\ 15 & 5 \\ -10 & 25 \end{bmatrix} = \begin{bmatrix} -0.06183887 & 0.5506182 \\ -0.33928574 & -0.1003567 \\ -0.44424621 & 0.2099429 \\ 0.01871727 & -0.7909369 \\ 0.82665356 & 0.1307325 \end{bmatrix} \begin{bmatrix} 32.41852 & 0 \\ 0 & 19.97598 \end{bmatrix} \begin{bmatrix} -0.2795877 & -0.9601202 \\ 0.9601202 & -0.2795877 \end{bmatrix}^T$$

is unique. The matrix of principal components Y as described earlier, is given by

$$\mathbf{Y}^* = \mathbf{Y} \begin{bmatrix} -0.2795877 & -0.9601202 \\ 0.9601202 & -0.2795877 \end{bmatrix}.$$

This gives $\mathbf{Y}^* = \begin{bmatrix} 65 & 40 \\ 80 & 35 \\ 75 & 30 \\ 90 & 50 \\ 65 & 70 \end{bmatrix} \begin{bmatrix} -0.2795877 & -0.9601202 \\ 0.9601202 & -0.2795877 \end{bmatrix} = \begin{bmatrix} 20.231609 & -73.59132 \\ 11.237194 & -86.59518 \\ 7.834531 & -80.39664 \\ 22.843120 & -100.39020 \\ 49.035215 & -81.97895 \end{bmatrix}$

3.3.1.4 Discrimination

The mean of principal components of all observations with class label 0 is computed, and also the mean of all observations with class label 1 is computed. These two mean vectors are estimates of the centers of the two classes.

Illustration

Consider a training sample T displayed in Table 3.5 in which $\mathbf{X_1}$ and $\mathbf{X_2}$ are predictors of $\mathbf{Y}$

Table 3.5: A Training Sample for Illustration

X_1	X_2	Y
65	40	0
80	35	1
75	30	1
90	50	1
65	70	0

It has been shown already that the principal components matrix of the predictors $\mathbf{X_1}$ and $\mathbf{X_2}$ is

$$\begin{bmatrix} 20.231609 & -73.59132 \\ 11.237194 & -86.59518 \\ 7.834531 & -80.39664 \\ 22.843120 & -100.39020 \\ 49.035215 & -81.97895 \end{bmatrix}.$$

Denoting the collection of principal components of data points labelled 0 and 1 respectively by $\mathbf{T_0^*}$ and $\mathbf{T_1^*}$ gives

$$\mathbf{T_0^*} = \begin{bmatrix} 20.231609 & -73.59132 \\ 49.035215 & -81.97895 \end{bmatrix}, \quad \mathbf{T_1^*} = \begin{bmatrix} 11.237194 & -86.59518 \\ 7.834531 & -80.39664 \\ 22.843120 & -100.39020 \end{bmatrix}$$

Let the vector of means of the columns of $\mathbf{T_0^*}$ be $\mathbf{M_0}$. Similarly, let the vector of means of the columns of $\mathbf{T_1^*}$ be $\mathbf{M_1}$. Then $\mathbf{M_0} = (34.63, -77.76)$ and $\mathbf{M_1} = (13.97, -89.13)$

3.3.1.5 Classification/ Class Prediction of new observations

Given a new query (observation/data point) $\boldsymbol{O}_{new}$, let the vector if its principal components be $\boldsymbol{S}_{new}$, then

$$\boldsymbol{S}_{new} = \boldsymbol{O}_{new} \times \boldsymbol{L}_c \tag{3.3}$$

Here, $\boldsymbol{L}_c$ is the rotation matrix (of the cluster to which $\boldsymbol{O}_{new}$ belongs). Denoting as d_0, the Euclidean distance between $\boldsymbol{S}_{new}$ and $\mathbf{M}_0$ (the center of the class labelled 0); and denoting as d_1, the Euclidean distance between $\boldsymbol{S}_{new}$ and $\mathbf{M}_1$ (the center of the class labelled 1) gives the following equations:

$$|\boldsymbol{S}_{new} - \mathbf{M}_0| = d_0 \tag{3.4}$$

$$|\boldsymbol{S}_{new} - \mathbf{M}_1| = d_1 \tag{3.5}$$

$\boldsymbol{O}_{new}$ is classified as:

i. failure if $d0 \leq d1$

ii. success if $d0 > d1$

The Euclidean distance, d_e between two arbitrary n-dimensional vectors u and v where $u = (u_1, u_2, \ldots u_k, \ldots, u_n)$ and $v = (v_1, v_2, \ldots v_k, \ldots, v_n)$ is given by

$$d_e = \sqrt{(u_1 - v_1)^2 + (u_2 - v_2)^2 + \cdots + (u_k - v_k)^2 + \cdots + (u_n - v_n)^2} \tag{3.6}$$

The reason for including the equality in the first decision is that, there may be other predictors of the event that are not included in the model and also, the cost of misclassifying success as failure

is lesser than that of misclassifying a failure as success. Figure 3.1 illustrates the classification process. Additionally, the classifier also computes and indicates the following indices:

1. The Euclidean distance between the centers of the two classes, denoted D.
2. d_0
3. d_1
4. $d = |d_0 - d_1|$
5. $\frac{d}{D}$

The ratio $\frac{d}{D}$, indicates, to some extent the confidence in the predicted class of the query/observation. It is however, not directly used in the determination of the class label of the observation.

By the definitions of d and D, the following facts hold:

i. $d \geq 0$

ii. $d = D$ when $\boldsymbol{O}_{new}$ does not lie within the space between the two centers

iii. $d < D$ when d lies within the space between the two centers

i, ii, and iii thus imply that $0 \leq \frac{d}{D} \leq 1$

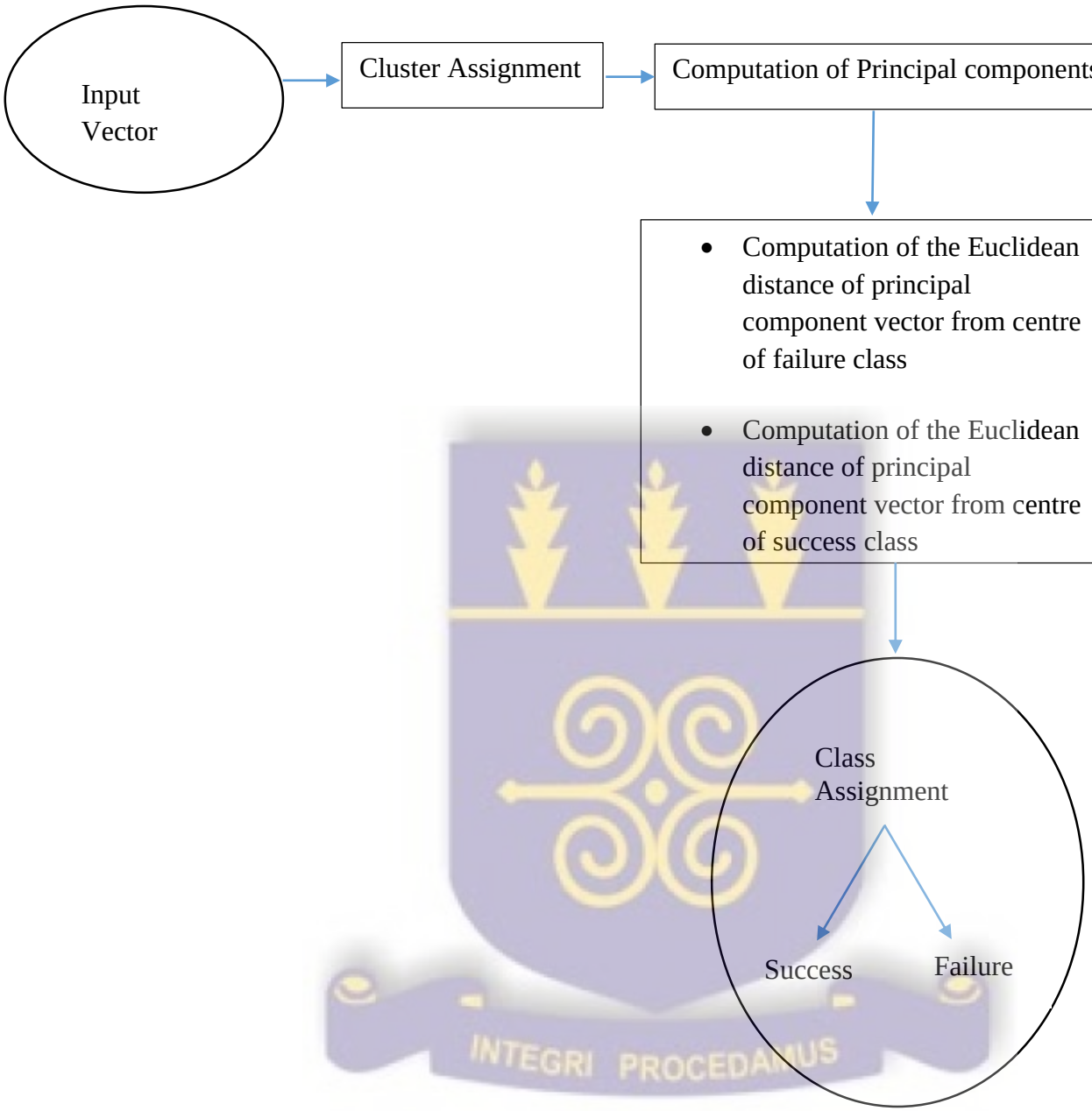

Figure 3.1: Flow Diagram of the Classification Process

The following illustration is based on the parameters of the illustration training set T introduced earlier.

Illustration

Consider a test data point $\boldsymbol{O_i} = (40, 30)$.

I. Vector of Principal components of $\boldsymbol{O}_i$ is

$$[40,30]\begin{bmatrix} -0.2795877 & -0.9601202 \\ 0.9601202 & -0.2795877 \end{bmatrix} = [-3.031474, -46.79242]$$

II. $$D=\sqrt{(34.63-13.97)^2+(-77.76--89.13)^2} = 23.58$$

III.

a) The Euclidean distance between the principal components vector and the estimated center of the class labelled 0 is

$$\begin{aligned} d_0 &= \sqrt{(34.63--3.03)^2+(-77.76--46.80)^2} \\ &= \sqrt{1418.276+958.5216} \\ &= 48.75 \end{aligned}$$

b) The Euclidean distance between the principal components vector and the estimated center of the class labelled 1 is

$$\begin{aligned} d_1 &= \sqrt{(13.97-(-3.03))^2+(-89.13-(-46.79))^2} \\ &= \sqrt{289+1792.676} \\ &= 45.63 \end{aligned}$$

IV. $d = |d_1 - d_0| = 3.12$

V. Clearly, $d_1 < d_0$

VI. The observation is classified as success by assigning to it the value 1

The classifier thus returns the following row of outputs (shown in Table 3.7) for $O_i = (40,30)$.

Table 3.6: Illustration of the output of the classifier

Cluster	D	d0	d1	d	d/D	P. Admit
	23.58	48.75	45.63	3.12	0.132	1

The P. Admit in Table 3.6 stands for the predicted qualification for admission into the university.

3.3.2 The Logistic Regression Approach

3.3.2.1 Introduction

Logistic regression is a statistical method used to measure/model the relationship between a categorical outcome variable and one or more independent variables through a logistic function. The independent variable(s) may be categorical, continuous, or a mixture of both categorical and continuous variables. When the categorical outcome has two levels, the logistic regression is referred to as binary logistic regression. Cases where the outcome has more than two values are referred to as multinomial logistic regression. In this project, the outcome can assume only two values; either a student has met the minimum requirement for admission into a Ghanaian university to pursue a first degree program or not. This makes modelling it as a binary logistic regression problem an appropriate choice. The binary logistic regression model predicts the probability that the outcome variable will be the event of interest. This is the conditional probability of occurrence of the event for the given the value(s) of the predictor variable(s). A researcher would usually assign the value 1 to the outcome variable when the desired event occurs and a value 0 the event does not occur. The implementation of the logistic regression hinges on the fact that an observation must have occurred several times in the data. The following are assumptions of the method:

1. The observations $X_1, X_2, \ldots, X_k, \ldots, X_n$ in the data are independently distributed
2. The outcome/response variable is binomially distributed or coming from an exponential family (Binomial, Poisson, Multinomial, Normal…)

3. The relationship between the independent variable(s) and the response variable need not be linear but there is a linear relationship between the natural logarithm of the probability of occurrence of the response variable and the explanatory variables.
4. Homogeneity of variance (homoscedasticity) need not be satisfied.
5. Errors are independent but do not need to be normally distributed
6. Method of Maximum Likelihood is used to estimate parameters
7. Goodness-of-fit measures rely on sufficiently large samples.

3.3.2.2 Modelling Procedure

Given training data set D, the logistic regression model is fit by the following procedure. Firstly, the data is aggregated to simplify calculations. The aggregated data indicates the frequencies the unique observations, number of successes, and number of failures. The odds of success for each observation are the calculated. Logits (natural logarithms) of the odds are computed, this is a transformation of binary response variable to a real-valued variable. The next stage involves fitting a straight curve of the logit in terms of the predictor variables. Finally, the logistic function assigns success probabilities to future observations.

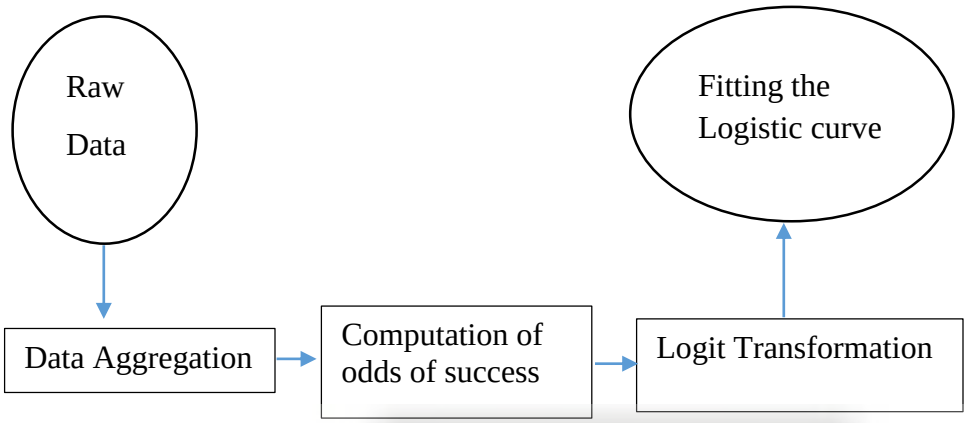

Figure 3.2: Flow Diagram of procedure for fitting the Logistic Curve

Classification of future observations (as either success or failure) is based upon a set probability threshold. It is common practice to classify observations whose predicted probability is less than 0.5 as failure and otherwise, classify the observation a success.

The major Mathematical/Statistical tools involved in each of the stages of the modelling process are presented in the sequel using the data in Table 3.7 for illustrations where necessary.

Table 3.7: A Sample Data Set for illustrating the Logistic Regression Procedure

observation	X_1	X_2	Y	observation	X_1	X_2	Y
1	1	0	1	**16**	2	0	1
2	1	1	0	**17**	2	0	1
3	2	0	0	**18**	2	1	1
4	2	1	0	**19**	2	1	1
5	3	0	0	**20**	2	1	1
6	3	1	0	**21**	2	1	1
7	1	0	0	**22**	3	0	0
8	1	0	1	**23**	3	0	1
9	1	0	1	**24**	3	0	1
10	1	0	0	**25**	3	0	1
11	1	1	1	**26**	3	0	1

observation	X_1	X_2	Y	observation	X_1	X_2	Y
12	1	1	1	**27**	3	1	0
13	1	1	1	**28**	3	1	1
14	2	0	0	**29**	3	1	1
15	2	0	0	**30**	3	1	1

3.3.2.2.1 Data Aggregation

This is a representation of the distinct observations in the data, together with their success and failure counts. The number of distinct observations given by the product of the number of levels of all the categorical variables.

Illustration

The data in Table 3.8 has two explanatory variables X_1 and X_2. The X_1 takes on 3 possible values while the X_2 takes on 2 possible values and so the composite variable $X = (X_1, X_2)$ takes on 6 different 2-dimensional vectors. It also means that a maximum of 6 distinct observations exist in the data set. The aggregated data is presented in Table 3.8

Table 3.8: Aggregated Data for Table 3.7

Observation(i)	Variables		Frequency	Frequency of successes	Frequency of Failures
	X_1	X_2		n(Y=1)	n(Y=0)
1	1	0	5	3	2
2	1	1	4	3	1
3	2	0	5	2	3
4	2	1	5	4	1
5	3	0	6	4	2
6	3	1	5	3	2
Totals			30	19	11

3.3.2.2.2 Odds of Success

The Odds of success of an observation is a measure of the ratio of the likelihood of the observation being the event of interest to the likelihood of it not being the event of interest. The Odds is a function of the predictor variable(s). Let x_i be the i^{th} element of $\boldsymbol{X}$, then the Odds of success of the outcome, given that $X = x_i$ is given by

$$Odds\ (x_i) = \frac{\mathrm{p}(Y = 1 \mid X = x_i)}{1 - \mathrm{p}(Y = 1 \mid X = x_i)} \tag{3.7}$$

It is convenient to label $\mathrm{Prob}(Y = 1 \mid X = x_i)$ as $\mathrm{p_i}$

So that

$$\mathrm{Odds}(x_i) = \frac{p_i}{1 - p_i} \tag{3.8}$$

Simply put, $Odds = \frac{P(Success)}{P(Failure)}$

From the data, the Odds of $X = (1,0) = \frac{3/5}{2/5} = \frac{3}{2} = 1.5$

Let

$$\mathrm{Odds}(\mathrm{x_i}) = \pi_i \tag{3.9}$$

When the Odds is less than 1, it means that the chances of the outcome being a success is less compared to the chances of it being a failure.

When Odds is greater than 1, it means that there is greater likelihood of the outcome being a success compared to it being a failure. Finally, an Odds of 1 means that there are equal chances of success and failure. Table 3.10 gives the Odds of the observations of the composite variable X

Table 3.9: Odds associated with the various values of the composite variable

x_i	π_i
(1,0)	1.50
(1,1)	3.00
(2,0)	0.67
(2,1)	4.00
(3,0)	2.03
(3,1)	1.50

3.3.2.2.3 Odds Ratio

The Odds Ratio of a particular predictor variable is a measure of association between the variable and the outcome (the dependent variable). This does not apply to the composite variable, it does to the distinct independent variables. It measures the ratio by which the odds increase/decrease for a 1-unit change in the independent variable. Mathematically,

$$Odds\ Ratio\ (of\ X_j) = \frac{Odds\left(X_j\right)}{Odds\left(X_j - 1\right)} \tag{3.10}$$

When the Odds Ratio equals 1, it means that the independent variable does not affect the outcome

When Odds Ratio is greater than 1, it means that a unit change in the variable leads to a corresponding change in the outcome (in the same direction), that is, the variable affects the outcome positively.

Finally, when the Odds Ratio is less than 1, it means that a unit change in the variable leads to a corresponding change in the opposite direction, that is, the variable affects the outcome negatively.

3.3.2.2.4 The Logit Transform

Since a straight curve of the binary response cannot be fit in terms of the independent variables, it is converted to a continuous variable by means of a transformation. In logistic regression, this transformation is the logit transform. The input of the logit is the odds (of success) of the outcome while the output is a real number.

$$\text{logit}(\pi_i) : (0,\infty) \rightarrow (-\infty,\infty) \tag{3.11}$$

If Y is a binary response and y_i is the i^{th} realization of Y, the logit of y_i is defined as:

$$logit(y_i) = \ln\left(\frac{p_i}{1-p_i}\right) \tag{3.12}$$

(3.12) is the same as

$$logit(y_i) = \ln(\pi_i) \tag{3.13}$$

Illustration

From the illustration data, logit of the third observation, $(2,0)$ is $\ln(0.67) = -0.40048$

The logits of all the observations are computed and shown in the right-most column of Table 3.11

Table 3.10: Illustration of Results of Logit Transformation

x_i	π_i	$\ln(\pi_i)$
(1,0)	1.50	0.4055
(1,1)	3.00	1.0986
(2,0)	0.67	-0.4055
(2,1)	4.00	1.3863
(3,0)	2.03	0.7082
(3,1)	1.50	0.4055

The definition of the logit transform implies clearly that $logit(p_i)$ is linear in p_i

But $p_i = p(Y=1|X=x_i)$ and so $logit(y_i)$ is linear in $X=(X_1, X_2, \ldots X_k, \ldots X_p)$

And so the relation

$$logit(y_i) = \tilde{X}\beta$$

Holds

Where $\tilde{\boldsymbol{X}} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ 1 & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}$ and $\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \\ \vdots \\ \beta_p \end{bmatrix}$ (3.14)

For the composite variable $\boldsymbol{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}$ and regression coefficients $\beta_0, \beta_1, \ldots \beta_k, \ldots, \beta_p$ which are estimated by method of Maximum Likelihood.

Now let

$$y_i^* = \text{logit}(y_i) \quad (3.15)$$

Then $\mathbf{Y}^* = \begin{bmatrix} y_1^* \\ y_2^* \\ \vdots \\ y_i^* \\ \vdots \\ y_n^* \end{bmatrix}$ and

$$\mathbf{Y}^* = \tilde{\mathbf{X}}\boldsymbol{\beta} \tag{3.16}$$

Which is the system

$$\begin{aligned} y_1^* &= \beta_0 + \beta_1 x_{11} + \ldots + \beta_k x_{1k} + \ldots + \beta_p x_{1p} \\ y_2^* &= \beta_0 + \beta_1 x_{21} + \ldots + \beta_k x_{2k} + \ldots + \beta_p x_{2p} \\ &\vdots \\ y_i^* &= \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik} + \ldots + \beta_p x_{ip} \\ &\vdots \\ y_n^* &= \beta_0 + \beta_1 x_{n1} + \ldots + \beta_k x_{nk} + \ldots + \beta_p x_{np} \end{aligned} \tag{3.17}$$

For the Right Hand Side of (3.13) to be defined, $\pi_i \notin (-\infty, 0]$

This implies that $p_i \neq 0,1 \left(\text{since by the definition of } p_i, p_i \notin (-\infty, 0)\right)$

Thus, the transform does not permit subsets

$$X_{0,1} := \{x_i | p_i = 0,1\} \tag{3.18}$$

To be in the data (on which the model's use is to be applied). The phenomenon of having a large proportion of such subsets in the data is referred to as *sparseness*. Sparseness is more of a problem to the model when it is associated with categorical variables.

This subsection is concluded by reiterating the fact that the output of the logit transform lies in the interval $(-\infty, \infty)$.

3.3.2.2.5 The Logistic Function

The Logistic is a function that converts the **real-valued** output of the logit transform to a probability. It is the inverse of the logit transformation.

Substituting (3.12) into (3.15) gives

$y^* = \ln\left(\frac{p_i}{1-p_i}\right)$, but

$$y^* = \ln\left(\frac{p_i}{1-p_i}\right) \Rightarrow e^{y^*} = \frac{p_i}{1-p_i}$$

$$e^{y^*}(1-p_i) = p_i$$

$$e^{y^*} - p_i.e^{y^*} = p_i$$

$$p_i + p_i.e^{y^*} = e^{y^*}$$

$$p_i\left(1+e^{y^*}\right) = e^{y^*}$$

$$p_i = \frac{e^{y^*}}{\left(1+e^{y^*}\right)}$$

$$Pr(Y=1|X=x_i) = \frac{e^{y^*}}{\left(1+e^{y^*}\right)}$$

$$\Rightarrow Pr(Y=1|X=x_i) = \frac{1}{1+e^{-y^*}}. \quad (3.19)$$

Substituting (3.16) and (3.17) into (3.19) gives

$$F(\boldsymbol{X}) = Pr(\boldsymbol{Y}=1|\boldsymbol{X}=\boldsymbol{x}) = \frac{1}{1+e^{-(\boldsymbol{x}\boldsymbol{\beta})}} \quad (3.20)$$

for the vector of outcomes Y and

$$\mathrm{f}\left(x_i\right) = Pr\left(y_i = 1 | X = x_i = x_{i1}, \ldots, x_{ip}\right) = \frac{1}{1+e^{-\left(\beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}+\ldots+\beta_k x_{ik}+\ldots+\beta_p x_{ip}\right)}} \tag{3.21}$$

for a particular outcome y_i Figure 3.3 shows the graph of a logistic function.

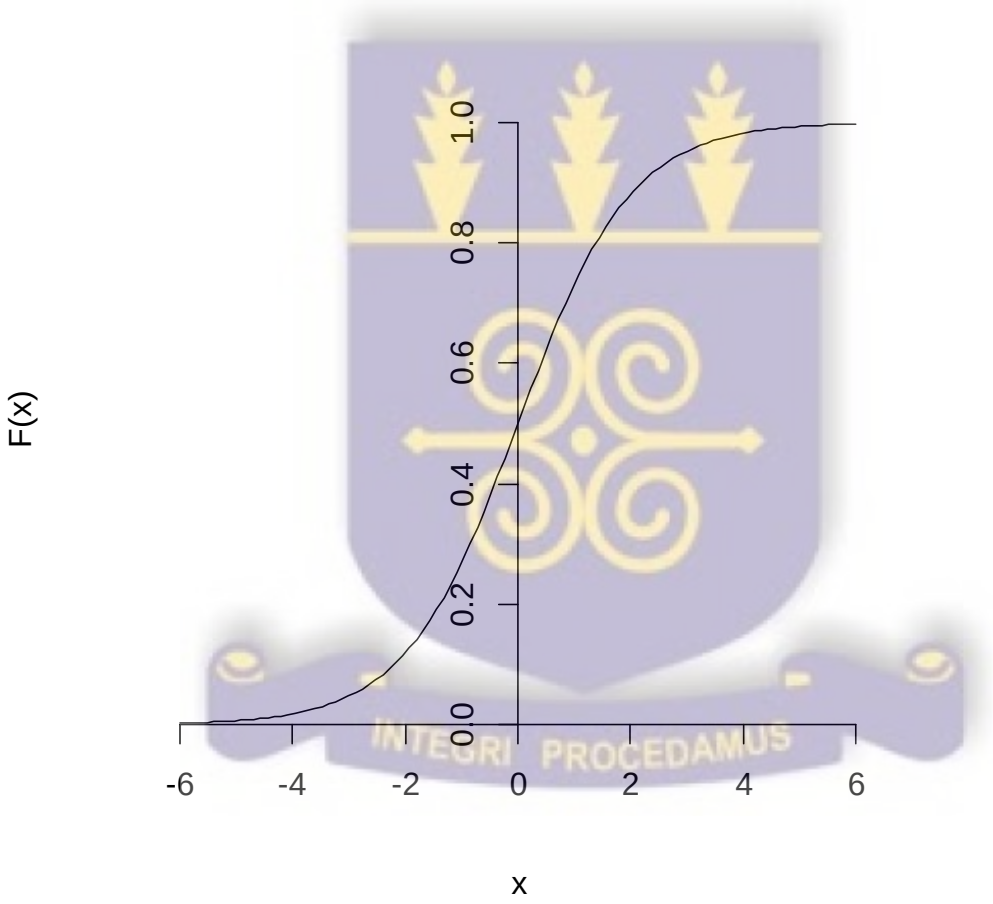

Figure 3.3: Graph of a Logistic function

The graph of the logistic function shows that the output of the logistic function lies between and including 0 and 1.

3.3.2.2.6 Estimation of Parameters

The parameters of (3.21) are estimated by the method of Maximum Likelihood. For the logistic regression, the least squares method cannot produce minimum variance unbiased estimators for the parameters (Czepiel, 2012). This follows from the one of assumptions of the methodology, that is: the variance of the response variable is not homogeneous across all values of the predictor variables. The method of Maximum Likelihood finds the set of parameters that have the greatest probability of producing the observed data. This is done by maximizing the Likelihood function of the parameter (the joint probability mass function (p.m.f.) of Y) and finding the vector of parameters for which it is maximum. Now, since the observations are independent, this joint p.m.f. is simply the product of all the probability mass functions, and also since the observations are identically distributed, each has the same probability mass function. Now, each observation, y_i is binomially distributed with number of trials n_i and success probability p_i i.e.

$$y_i \sim Binom(n_i, p_i)$$

and so the Likelihood function of β is

$$L(\beta) = f(\mathbf{Y};\boldsymbol{\beta}) = \prod_{i=1}^{n} \frac{n_i!}{y_i!(n_i - y_i)!} p_i^{y_i} (1-p_i)^{n_i - p_i} \quad . \qquad (3.22)$$

From calculus, the optimum (minimum or maximum) point occurs where the first derivative equals 0. This means that the Maximum Likelihood estimates of $\boldsymbol{\beta}$ can be found by differentiating (3.22) and equating the result to zero, provided that the second derivative is less than zero. Since the factorial terms of (3.22) do not contain p_i, they are essentially constants that can be ignored. Maximizing (3.22) without the factorial terms will produce the same result as if they were included

(Czepiel, 2012). This means that the Maximum Likelihood Estimate (MLE) can be found by maximizing

$$\prod_{i=1}^{n} p_i^{y_i} (1-p_i)^{n_i-p_i}.$$

This equals

$$\prod_{i=1}^{n} p_i^{y_i} \frac{(1-p_i)^{n_i}}{(1-p_i)^{y_i}}$$

Which implies that

$$L(\beta) = \prod_{i=1}^{n} \left(\frac{p_i}{1-p_i} \right)^{y_i} (1-p_i)^{n_i} \tag{3.23}$$

Substituting (3.17) into (3.15) gives

$$logit(y_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik} + \ldots + \beta_p x_{ip} = \sum_{k=0}^{p} x_{ik}\beta_k. \tag{3.24}$$

This implies that

$$\ln\left(\frac{p_i}{1-p_i} \right) = \sum_{k=0}^{p} x_{ik}\beta_k \quad . \tag{3.25}$$

Taking the exponent of both sides of (3.25) gives

$$\left(\frac{p_i}{1-p_i} \right) = e^{\sum_{k=1}^{p} x_{ik}\beta_k} \tag{3.26}$$

Making p_i the subject gives

$$p_i = \frac{e^{\sum_{k=1}^{p} x_{ik}\beta_k}}{1+e^{\sum_{k=1}^{p} x_{ik}\beta_k}} \tag{3.27}$$

Substituting (3.26) for the first term and (3.27) for the second term of (3.23) gives

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{p} \left(e^{\sum_{k=1}^{p} x_{ik}\beta_k} \right)^{y_i} \left(1 - \frac{e^{\sum_{k=1}^{p} x_{ik}\beta_k}}{1+e^{\sum_{k=1}^{p} x_{ik}\beta_k}} \right)^{n_i} \tag{3.28}$$

This implies that $L(\boldsymbol{\beta}) = \prod_{i=1}^{p} \left(e^{y_i \sum_{k=0}^{p} x_{ik}\beta_k} \right) \left(\frac{1}{1+e^{\sum_{k=0}^{p} x_{ik}\beta_k}} \right)^{n_i}$ which is equal to

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{p} \left(e^{y_i \sum_{k=0}^{p} x_{ik}\beta_k} \right) \left(1+e^{\sum_{k=0}^{p} x_{ik}\beta_k} \right)^{-n_i} \tag{3.29}$$

The natural logarithm of both sides of (3.29) shall be taken to make the differentiation easier. This transformation is justified, since the natural logarithm function is monotonic and so the maximum of $L(\boldsymbol{\beta})$, if it exists, will as well be the maximum of the logarithm of the likelihood function.

$$l(\boldsymbol{\beta}) = \ln L(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left[e^{y_i \left(\sum_{k=0}^{p} x_{ik}\beta_k \right)} - n_i . \ln \left(1+e^{\sum_{k=0}^{p} x_{ik}\beta_k} \right) \right] \tag{3.30}$$

Taking, partial derivative of $l(\beta)$with respect to β_k implies

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

$$= \sum_{i=1}^{n} y_i x_{ik} - n_i \cdot \frac{1}{1+e^{\sum_{k=0}^{p} x_{ik}\beta_k}} \cdot e^{\sum_{k=0}^{p} x_{ik}\beta_k} \cdot \frac{\partial}{\partial \beta_k} \sum_{k=0}^{p} x_{ik}\beta_k \qquad (3.31)$$

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

The maximum likelihood estimates for β can be found by setting each of the $p+1$ equations in (3.31) to zero and solving for each β_k. This is not algebraically possible (Czepiel, 2012) and so the solution has to be numerically estimated by an iterative process.

3.3.2.2.7 Interpretation of Parameter Estimates

Model parameters are more meaningful when interpreted in terms of the odds Table 3.11 gives the interpretations of the parameter estimates.

Table 3.11: Interpretation of Model Parameters

Parameter Estimate	Sign	Interpretation
β_0	Negative	Odds of occurrence of the event decrease by a

Parameter Estimate	Sign	Interpretation
		multiplicative factor of e^{β_1} in the absence of all explanatory variables in the model.
β_0	Positive	Odds of occurrence of the event increase by a multiplicative factor of e^{β_0} in the absence of all explanatory variables in the model
$\beta_k, k = 1, ..., p$	Negative	On average, odds of occurrence of the event decrease by a multiplicative factor of e^{β_k} for a 1-unit increase in the explanatory variable x_k ; given that all other explanatory variables in the model remain unchanged.
$\beta_k, k = 1, ..., p$	Positive	On average, odds of occurrence of the event increase by a multiplicative factor of e^{β_k} for a 1-unit increase in the explanatory variable x_k ; given that all other explanatory variables in the model remain unchanged

3.3.2.2.8 Test for Significance of Parameter Estimates

For each β_k the null hypothesis $H_0 : \beta_k = 0$ is tested against the alternative $H_A : \beta_k \neq 0$ to check the significance of the estimate at various levels of significance. The test statistic is the Wald statistic, $W = \dfrac{\beta_k}{s.e\left(\beta_k\right)}$ is the ratio of the estimate to its standard error which is the positive square root of the element s_{kk} of covariance matrix of the vector β. $s.e\left(\beta_k\right)$ means standard error of β_k. It can be shown that, for large values of n, β is approximately, normally distributed.

If an estimate β_k is found to be significant at some α level of significant, it means that when β_k is used in place of β_k 100, times it will make inaccurate predictions $\alpha(100)$ or less times, on average.

3.3.2.2.9 Goodness of Fit Measure

Deviance and Likelihood Ratio Chi-Square Statistic are two good measures used to test the goodness of fit of a logistic regression model. The measure used to assess the goodness of the model fit in this study is the **deviance.** It measures the variation in the data Y that has not been captured by the model. If there exists a saturated model, the deviance, denoted D, is defined as

$$D(y) = -2\ln\frac{\text{likelihood of the fitted model}}{\text{likelihood of the saturated model}}. \qquad (3.32)$$

Here, a **saturated model** can be defined as a model with a theoretically perfect fit. Two types of deviance shall be calculated, the null deviance and the fitted deviance. The null deviance represents the difference between the null model and the saturated model, the fitted model

represents the difference between the fitted model and the saturated model. A null model is one with only the intercept.

$$D_{\text{null}} = -2\ln\frac{\text{likelihood of null model}}{\text{likelihood of saturated model}}. \tag{3.33}$$

Mostly, a saturated model does not exist, in such cases the deviance is defined as

$$D(y) = -2\ln(\text{likelihood of the fitted(null) model}). \tag{3.34}$$

A very good fitted model is one whose deviance is significantly smaller than the null model's deviance, this indicates that the set of explanatory variables improve the model fit.

3.4 Measures of Classifier Performance

The performances of two classifiers are measured based on specificity, sensitivity, and accuracy. These metrics have been defined in chapter 2 but are defined again mathematically in the following three equations ((3.35), (3.36), (3.37)):

$$Specificity = \frac{TN}{TN + FP} \tag{3.35}$$

$$Sensitivity = \frac{TP}{TP + FN} \tag{3.36}$$

$$\text{Accuracy} = \frac{\text{TN+TP}}{\text{TN+FN+TP+FP}} \tag{3.37}.$$

Sensitivity is also referred to as, recall (Powers, 2007; Elkan, 2012). In all cases, the confusion matrices from which the metrics are calculated are also explicitly presented as suggested by Elkan (2012).

4. ANALYSIS AND RESULTS

4.1 Chapter Overview

This chapter is in two main sections, the first section entails some exploratory analysis of the data. The focus here is on identifying the variables that may be indicative of the difference between the two classes of the response variable. This section is not about the models used in the main analysis. The second section provides the results obtained from the two main method of the analysis. In all cases brief discussion of the trends detected are made.

4.2 Exploratory data Analysis/ Summary Statistics of the data

This section gives some summary statistics of both the categorical and continuous variables in the data set and discusses the possible dimension along which the classes could be maximally discriminated.

4.2.1 Summary Statistics of the Categories

This subsection gives summary statistics of the categorical variables in the data.

4.2.1.1 Summary statistics for qualification for admission

An appreciably higher proportion of students in the data failed to meet the minimum requirement for admission. Out of 241 students who wrote the examinations, it was only 100 whose grade combinations meet the minimum requirements for admission into first degree programs. This is an observation that steps must be taken to address. Table 4.1 gives summary statistics for the variable: qualification for admission into the university.

Table 4.1: Frequency Table for Qualification for admission

Admit	Count	Percentage in Category
0	141	58.51
1	100	41.49
Total	241	100.00

4.2.1.2 Frequency Table for Program of Study

The General Arts programs is the program with the highest number of candidates, 32. 37% of the students in the data set have read General Arts. It may be that the more students simply love the program or the school reserves a higher quota for the program. The combination of the Agricultural Science, Home Economics, Visual Arts, and Technical Programs records the lowest frequency. This is the same as the proportion of candidates that are from the business department. It means that each of these four programs presents fewer number of candidates compared to the other three programs. Table 4.2 gives summary statistics for program of study.

Table 4.2: Frequency Table for program of study

Program of Study	Count	Percentage in Category
1	50	20.75
2	78	32.37
3	69	28.63
4	50	20.75
Total	241	100.00

4.2.1.3 Summary Statistics for Residential Status of the Students

The majority of the students were in the boarding house, Table 4.3 gives summary statistics for the residence status of the students.

Table 4.3: Frequency Table for residential status

Residence	Count	Percentage in Category
0	172	71.36

Residence	Count	Percentage in Category
1	69	28.63
Total	241	100.00

4.2.1.4 Summary Statistics for Gender

52. 28 % of the students were males while the remaining 47.72% were females. This tells that the proportion of girls in schools is still lower than that of boys. Table 4.4 displays the gender distribution of students in the data.

Table 4.4: Frequency Table for Gender

Gender	Count	Percentage in Category
0	115	47.72
1	126	52.28

4.2.2 Summary Statistics for the continuous variables across the various categories

This section gives summary statistics for the continuous variables across the various categorical variables. The short forms of the variable names are used in the tables, they have been defined in chapter 3. Table 4.5 gives some summary statistics of the continuous variables. It can be seen that the end of first year mathematics marks has the highest variation in the data, followed by the second year mathematics marks. The results give an initial suspicion that these variables would be better discriminators of the two classes in the data. Entry Grade has the least variance among all the continuous variables (18.64) but this variance cannot be compared to the variances of the other variables since its range of value differs from the range of the other variables. The continuous assessment variables are in percentages while the entry grades range from 6 to 28. The variance in the average marks obtained by students in the first year is considerably higher than the variance in the marks obtained by the students in the second year. It is likely that, moving on to second

year, students who were not performing well during the first year have improved that is why the variation has reduced. The drop in the variance (from 319.90 to 210.30) of the marks scored by the students in mathematics from the first year to the second year also deepens this suspicion. The same observation is made for the marks obtained by students in English Language in the first and second years. It suggests that that in the long run, the performances of most students would improve.

Table 4.5: Summary Statistics of the continuous variables

Variable	Min	1st Qu.	Median	Mean	3rd Qu.	Max	Variance
Entry Grade	6.00	12.00	14.00	14.56	17.00	28.00	18.64
Eng1	26.00	53.00	61.00	60.68	68.00	88.00	122.08
Math1	10.00	42.00	57.00	57.40	72.00	99.00	319.90
f_av	33.00	51.00	58.00	59.22	67.00	89.00	116.48
Eng2	29.00	52.00	58.00	57.95	64.00	86.00	82.30
Math2	12.00	46.00	57.00	56.14	65.00	94.00	210.30
s_av	31.00	54.00	61.00	60.78	67.00	84.00	90.39

4.2.2.1 Distribution of Entry Grade across Categories

Comparing the entry grades of students in the two classes reveals that there exists slight variations in the grades of students who eventually qualifies for admission and those who did not. For instance, the mean mark of students in the failure class is 15.95 whilst that of the success class is 12.61. One would expected much larger difference. It gives a hint that entry grade might not be a significant discriminator between the two classes. Considering gender as a variable, the descriptive statistics reveal that very slight difference exists between the average performance of males and females; with the mean grade of girls being 15.05 and that of boys being 14.12. In terms of variance, girls record a quite higher value, this indicates that there is greater competition among boys than among girls at the Junior School Level.

Table 4.6: Descriptive statistics for Entry Grade across the various categories

Category	Levels	Min	Max	Mean	Variance	Fraction in Category
Admit	Fail	7	28	15.95	17	0.59
	Pass	6	26	12.61	14.54	0.41
Program	Business	9	26	14.66	9.86	0.21
	G. Arts	6	27	15.05	18.52	0.32
	Science	6	26	11.87	14.53	0.26
	Vocational	7	28	17.1	17.68	0.21
Residence	Day	6	28	15.51	24.52	0.29
	Boarder	6	27	14.19	15.91	0.71
Gender	Female	6	28	15.05	21.54	0.48
	Male	7	27	14.12	15.72	0.52

4.2.2.2 Exploratory analysis of students' performance in English in SHS 1

The average mark of students who eventually qualify for admission in English Language in SHS 1 is 67.35 that of students who do not qualify for admission is 53.37, the difference is 13.98. This is quite high. The minimum mark of students who eventually did not qualify is 26 and the maximum is 27, this in a negligible difference. It suggests that first year English Language marks would be a good discriminator between the two classes. It is also interesting to note that the least mark of students in the success class is greater than the highest mark of students in failure class. The variances also show that students in the success class are closer to one another in performance than students in the failure class. The general arts class records the least variation the first year English Language marks, implying that there is keener competition there. This results is also not surprising since most of the general arts electives are reading subjects, some even read Literature in English as an elective subject and so it is in not surprising to find such closeness in the marks. Their average mark is also higher. It suggests that English language marks may not be a good discriminator between the two classes of the response variable among general arts students. Table

4.7 presents the descriptive statistics of first year English Language marks across the various categorical variables.

Table 4.7: Statistics of marks of Students in English Language in SHS 1

Category	Levels	Min	Max	Mean	Variance	Fraction in Category
Admit	Fail	26.00	27.00	53.37	81.04	0.59
	Pass	43.00	88.00	67.35	66.42	0.41
Program	Business	42.00	83.00	61.48	99.4	0.21
	G. Arts	43.00	83.00	63.63	72.29	0.32
	Science	34.00	88.00	65.57	90.86	0.26
	Vocational	26.00	70.00	49.10	87.89	0.21
Residence	Day	26.00	88.00	58.62	158.18	0.29
	Boarder	31.00	83.00	61.50	106.05	0.71
Gender	Female	31.00	83.00	60.30	124.81	0.48
	Male	26.00	88.00	61.02	120.32	0.52

4.2.2.3 Exploratory analysis of students' mark in Mathematics in SHS 1

The average mark of students in the success class is more than that of the student in the failure class by a margin of a margin of 22.98. The variance of the success class is also smaller than that of the failure class, also indicating that there is keener competition among students in the success class. In terms of program of study, Science students record the highest mean mark among all other programs of study while students in the vocational program record the least mean mark. Girls have a lesser mean mark of 51.61 compared to boys. First year Mathematics marks have a higher variation among girls than boys; 285.6 for girls and 294.81 for boys. This means that much as the girls have lower average mark than boys, their marks are closer to one another, indicating that in general, boys perform better than girls in Mathematics in SHS 1.

Table 4.8: Statistics of SHS 1 Mathematics marks across the various categories

Category	Levels	Min	Max	Mean	Variance	Fraction in Category
Admit	Fail	10.00	78.00	47.87	185.79	0.59
	Pass	35.00	99.00	70.85	200.55	0.41
Program	Business	27.00	86.00	58.34	242.35	0.21
	G. Arts	22.00	89.00	51.08	285.11	0.32
	Science	35.00	99.00	72.02	244.16	0.26
	Vocational	10.00	90.00	47.92	239.79	0.21
Residence	Day	10.00	99.00	55.42	331.04	0.29
	Boarder	22.00	91.00	58.20	315.11	0.71
Gender	Female	10.00	90.00	51.61	285.60	0.48
	Male	23.00	99.00	62.68	294.81	0.52

4.2.2.4 Exploratory Analysis of Average Performance of students in Other Remaining Subjects in SHS 1

The statistics reveal that on average, students who would eventually qualify for admission have higher marks than those who would not. The minimum mark of the success class is 10 more than that of the failure class. The maximum mark of the success class is also 15 more than that of the failure class. The average mark of the success class is 15 more than that of the failure class. Lower variance in the success class indicates that there is keener competition among those students who would eventually qualify for admission.

Table 4.9: Descriptive statistics for First Average across the various categories

Category	Levels	Min	Max	Mean	Variance	Fraction in Category
Admit	Fail	33.00	74.00	53.00	58.00	0.59
	Pass	43.00	89.00	68.00	76.00	0.41
Program	Business	43.00	79.00	60.00	113.00	0.21
	G. Arts	33.00	82.00	56.00	96.00	0.32

Category	Levels	Min	Max	Mean	Variance	Fraction in Category
	Science	40.00	89.00	67.00	85.00	0.26
	Vocational	33.00	74.00	55.00	83.00	0.21
Residence	Day	33.00	89.00	58.00	144.00	0.29
	Boarder	33.00	82.00	60.00	106.00	0.71
Gender	Female	33.00	80.00	57.00	91.00	0.48
	Male	33.00	89.00	61.00	133.00	0.52

The observations made for the statistics of the marks of students in SHS 1 across the three variable: first year Mathematics marks, first year English Language marks, and first year average mark in the remaining five or six subjects indicate that first year continuous assessment records of students can serve as a good basis for discriminating between academically good students and academically bad students, at least to some extent. This is because on all the three variables, the average performance of the "good" students is higher than the average performance of the "bad" students.

4.2.2.5 Exploratory Analysis of Students' Performance in English in SHS 2

In SHS 2 the average performance of students in the success class is greater than that of students in the failure class and there is a remarkably greater variation among those students in the failure class compared to those in the success class. The average performance of science students is the highest and that of the vocational program is the lowest. In the case of the first year, the general arts students did best in English Language, on average but here, the science students lead marginally. Boys have a slightly higher average mark than girls in this variable.

Table 4.10: Descriptive statistics for SHS 2 English Language across the various categories

Category	Levels	Min	Max	Mean	Variance	Fraction in Category
Admit	Fail	29.00	77.00	54.32	11.96	0.59

Category	Levels	Min	Max	Mean	Variance	Fraction in Category
	Pass	48.00	86.00	63.06	52.60	0.41
Program	Business	36.00	70.00	57.06	48.06	0.21
	G. Arts	37.00	77.00	59.91	69.67	0.32
	Science	43.00	75.00	61.22	63.72	0.26
	Vocational	29.00	86.00	51.64	103.62	0.21
Residence	Day	29.00	75.00	56.68	82.63	0.29
	Boarder	31.00	86.00	58.45	81.75	0.71
Gender	Female	31.00	77.00	57.51	78.67	0.48
	Male	29.00	86.00	58.34	85.94	0.52

4.2.2.6 Exploratory Analysis of students Performance in Mathematics in SHS 2

The maximum mark of students in the failure class is 75% while the minimum mark of students in the success class is 38. This implies that students in the two classes cannot be completely separated in terms of SHS 2 mathematics marks. The average mark of students in the failure class is 49.50 while that of students in the success class is 65.50, implying that on the average, students in the success class have performed better than students in the failure class in SHS 2 Mathematics. There is a greater variation in the marks of students in the failure class compared to those in the success class (153.55 for failure class and 141.32 for success class) suggesting that there is keener competition among students in the success, it tells also that the mean mark is more representative of them. The higher variance in the failure class tells how less the mean mark represents the class.

Table 4.11: Descriptive statistics for SHS2 Core Mathematics marks across the categories

Category	Levels	Min	Max	Mean	Variance	Fraction in Category
Admit	Fail	12.00	75.00	49.50	153.55	0.59
	Pass	38.00	94.00	65.50	141.32	0.41

Category	Levels	Min	Max	Mean	Variance	Fraction in Category
Program	Business	40.00	86.00	62.48	118.83	0.21
	G. Arts	20.00	84.00	50.47	170.93	0.32
	Science	36.00	94.00	62.08	118.78	0.26
	Vocational	12.00	77.00	51.14	249.63	0.21
Residence	Day	18.00	89.00	56.04	214.76	0.29
	Boarder	12.00	94.00	56.17	209.95	0.71
Gender	Female	18.00	86.00	52.61	193.33	0.48
	Male	12.00	94.00	59.36	205.56	0.52

Across the programs of study business class records the highest mean performance of 62.48 with a variance of 118.83, science students follow with a mean mark of 62.08 and a variance of 118.78. The mean mark of boys is greater than the mean mark of girls, also indicating that in SHS 2 boys still perform better than girls in Mathematics, just as in the case of SHS 1.

4.2.2.7 Exploratory Analysis of Average Performance Students in other Remaining Subjects in SHS 2

Minimum mark of students in the failure class is 39 while that of students in the success class is 31. The maximum of the failure class is 83 while that of the success class is 84. Both classes have the same mean mark. The closeness in these statistics suggests that this variable would be a good discriminator between the two classes. It is interesting to know that with the same mean mark, the variance of the failure class is smaller than that of the success class. This suggests that in the second year, the average marks of the students does not speak much about the difference between the two classes. The mean marks of the students in the students across all the programs of study are also almost the same. In terms of residential status, means of the two classes are exactly the same. Gender wise too, there is not great variation in students' performance on this variable. The closeness between the two classes of students in second year marks is surprising and needs to be investigated.

Table 4.12: Descriptive statistics for Second Average across the various categories

Category	Levels	Min	Max	Mean	Variance	Fraction in Category
Admit	Fail	39.00	83.00	61.00	54.00	0.59
	Pass	31.00	84.00	61.00	89.00	0.41
Program	Business	37.00	84.00	61.00	89.00	0.21
	G. Arts	45.00	81.00	60.00	92.00	0.32
	Science	44.00	81.00	62.00	78.00	0.26
	Vocational	31.00	83.00	60.00	105.00	0.21
Residence	Day	37.00	83.00	61.00	85.00	0.29
	Boarder	31.00	84.00	61.00	93.00	0.71
Gender	Female	39.00	81.00	61.00	93.00	0.48
	Male	31.00	84.00	60.00	89.00	0.52

4.2.3 Analysis of the differences in the mean values of the variables between the two classes of the response variable

Table presents the differences in the mean values of the seven continuous variables between the two classes. The values are the means of the success class minus the means of the failure class. For example, the mean entry grade of students in the success class is 12.61 while that of students in the failure class is 15.95, subtracting the later from the former gives -3.34 as shown in the first column of the second row of Table 4.13. Inspection of this table suggests that the first year Mathematics marks would best discriminate between the two classes, followed by the first year average marks of students in the remaining subjects. It is also seen that the SHS 2 average marks does not discriminate between the two classes at all. All the differences between the two classes in SHS 1 marks are higher than those of SHS 2. This suggests strongly that SHS 1 continuous assessment record is a better predictor of students' eventual success than the second year continuous assessment record.

Table 4.13: Differences between the two classes in terms of the Mean Marks

Entry grade	Eng.1	Math.1	f_av	Eng. 2	Math. 2	s_av
-3.34	13.98	22.98	15	8.74	16	0

The exploratory analysis of the data suggests directions of maximum variance in the data, however they remain suggestions, the confirmatory analysis would throw more light on this suspicions.

4.3 Results

This section presents the classification results from using both the Proposed Method (Pattern Recognition) and Logistic Regression.

4.3.1 Results of Proposed Method (Pattern Recognition) Using all Variables

Table 4.14 gives a structure of the result of the Pattern Recognition methodology for the training set using all variables. The full result is in appendix 7.2

Table 4.14: Results of the proposed method for Training Set

std	Admit	Cluster	D	d0	d1	d	d/D	P. Admit	Accuracy
1	1	7	18.45	21.24	9.92	11.32	0.61	1	Accurate
2	0	3	44.83	15.05	50.64	35.59	0.79	0	Accurate
3	0	4	118.99	16.9	124.8	107.9	0.91	0	Accurate
…	…	…	…	…	…	…	…	…	…
…	…	…	…	…	…	…	…	…	…
…	…	…	…	…	…	…	…	…	…
239	0	8	54.51	24.87	41.81	16.94	0.31	0	Accurate
240	0	13	32.8	25	29.82	4.82	0.15	0	Accurate
241	0	5	34.47	20.44	45.79	25.35	0.74	0	Accurate

The corresponding confusion matrix is Table 4.15

Table 4.15: Confusion Matrix of the Result of the Proposed Method on Training Set

	Predicted Class		
Actual Class	0	1	Total
0	102	8	110
1	12	69	81
Total	114	77	191

- Percentage Specificity $= \frac{102}{110} \times 100\% = 93\%$
- Percentage Sensitivity $= \frac{69}{81} \times 100\% = 85\%$
- Percentage Accuracy $= \frac{102 + 69}{191} \times 100\% \simeq 90\%$

Table 4.16 gives a structure of the result of the Proposed (Pattern Rec.) Method for the test set using all variables, the full result is in appendix 7.

Table 4.16: Structure of Result of Proposed Method for Test Set.

std	Admit	Cluster	D	d0	d1	d	d/D	P. Admit	Accuracy
11	0	9	26.3	16.96	35.91	18.95	0.72	0	Acc.
13	0	14	122.31	19.72	123.56	103.84	0.85	0	Acc.
21	1	11	18.29	32.8	17.9	14.9	0.81	1	Acc.
…	…	…	…	…	…	…	…	…	…
180	0	11	18.29	13.63	26.2	12.57	0.69	0	Acc.
192	0	1	23.47	30.44	34	3.56	0.15	0	Acc.
214	1	1	23.47	28.04	29.42	1.38	0.06	0	Inacc.

Table 4.17: Confusion Matrix of Result of Proposed Method on Test set

	Predicted Class		
Actual Class	0	1	Total
0	27	4	31
1	7	12	19
Total	34	16	50

From the table,

- Percentage Specificity $= \frac{27}{31} \times 100\% = 87\%$,
- Percentage Sensitivity $= \frac{12}{19} \times 100\% = 63\%$,
- Percentage Accuracy $= \frac{27+12}{50} \times 100\% = 78\%$.

The results shows that this model it is generally better at predicting the class of observations in the training set than predicting the class of new observations. It can be seen that, for new observations, the model is better at identifying observations whose actual class label is 0 i.e. specificity. The result also shows that the model has a relatively high chance of making error when it comes to identifying observations whose actual class is 1. It is therefore best used in cases where the identification of observations whose actual class is 0 is the priority.

4.3.2 Results of Logistic Regression

An initial model of all 9 explanatory variables was fitted, Table 4.18 gives the estimates of the regression parameters, their corresponding standard errors, their z values (test statistic) and their p-values. The p-values measure the significance of the estimates. Specifically, the p-value is the probability of making an error in using the estimate in the place of the unknown parameter.

Table 4.18: Estimates of Regression Parameters

Coefficients:	Estimate	Std.Error	z value	Pr(>|z|)
(Intercept)	-19.149	4.06693	-4.708	2.50E-06
prog_code2	-0.04951	0.78044	-0.063	0.9494
prog_code3	-0.48141	0.83184	-0.579	0.5628
prog_code4	-1.60281	1.12411	-1.426	0.1539
resid_code1	0.43684	0.62967	0.694	0.4878

Coefficients:	Estimate	Std.Error	z value	Pr(>|z|)
gen_code1	-0.01181	0.58726	-0.02	0.9839
Entry.Grade	-0.03784	0.06808	-0.556	0.5783
Eng1	0.04913	0.0346	1.42	0.1557
Math1	0.05205	0.0264	1.972	0.0486
f_av	0.20183	0.04873	4.142	3.44E-05
Eng2	0.05788	0.04381	1.321	0.1864
Math2	0.01695	0.03116	0.544	0.5864
s_av	-0.05517	0.03272	-1.686	0.0918

Null deviance: 260.362 on 190 degrees of freedom
Fitted deviance: 98.938 on 178 degrees of freedom
AIC: 124.94

Selection of the final model was based on the following three factors in descending order of importance:

1. Maximizing predictive accuracy.
2. Minimizing the number of predictor variables.
3. Significance of variables.

In terms of significance of variables, a p-value of less than 0.05 was considered as ideal. In this regard, the variables: third term core mathematics marks of SHS 1, and the mean of marks of other remaining subjects of SHS1 (Math1 and f_av) qualify to be in the final model. It turned out that the model with these two variables also satisfy the first two requirements. However, the model with all explanatory variables has a smaller residual deviance compared to the model with only

the two variables but since the focus of the study is on evaluating the predictive accuracy, the model with the lesser number of explanatory variables has been chosen as the final model.

Table 4.19 gives the summary of the final model fitted with Math1 and f_av.

Table 4.19: Parameters of Model Fitted with Significant Variables Only

Variables	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-15.6401	2.26054	-6.919	4.56E-12
Math1	0.05611	0.01646	3.409	0.000652
f_av	0.19753	0.03814	5.179	2.23E-07

Null deviance: 260.36 on 190 degrees of freedom
Fitted deviance: 121.52 on 188 degrees of freedom
AIC: 127.52

The final model can therefore be written as:

$$P(Y=1|\text{Math1},\text{f\_av}) = \frac{1}{1+e^{15.64-0.056*\text{Math1}-0.197*\text{f\_av}}}. \quad (4.1)$$

(4.1) gives the probability that a student qualifies for admission given his or her first year third term mark in core mathematics and average of the marks obtained in the remaining subjects with the exception of English language. The intercept being -15.6401 implies that the probability of success of a student is $\frac{1}{1+e^{15.64}} = 1.6\times10^{-7} \simeq 0$ in the absence the two significant independent variables. This simply means that a student who obtains a score of zero in the two variables has a zero chance of qualifying for admission to the university. The coefficient f_av being 0.19753 implies that the odds of success increase by the multiplicative factor $e^{0.19753} = 1.22$ for every unit change in f_av, when Math1 remains constant. The coefficient of Math1 is 0.05611. This implies

that the odds of success increase by the multiplicative factor $e^{0.05611} = 1.06$ for every unit change in Math1, when f_av remains constant. The structure of the results of the model is presented in Table 4.20 , the full result is in appendix 7.4

Table 4.20: Results of Logistic Reg. on Training Set

std	A. Admit	P(Y=1)	P. Admit	Accuracy
1	1	0.3313	0	Inaccurate
2	0	0.0287	0	Accurate
3	0	0.0714	0	Accurate
…	…	…	…	…
…	…	…	…	…
…	…	…	…	…
239	0	0.1670	0	Accurate
240	0	0.5975	1	Inaccurate
241	0	0.0537	0	Accurate

The confusion matrix of this result is Table 4.20

Table 4.21: Confusion Matrix of Results of Logistic Regression on Training set

	Predicted Class		
Actual Class	**0**	**1**	**Total**
0	99	11	110
1	13	68	81
Total	112	79	191

- Percentage Specitivity $= \frac{99}{110} \times 100\% = 90$
- Percentage Sensitivity $= \frac{68}{81} \times 100\% = 84\%$

- Percentage Accuracy $= \frac{167}{191} \times 100\% = 87\%$

The full results of the Logistic Regression on the test set is in appendix 7.5, the confusion matrix is displayed in Table 4.21

Table 4.22: Confusion Matrix of Results of Logistic Regression on Test Set

	Predicted Class		
Actual Class	0	1	Total
0	26	5	31
1	6	13	19
Total	32	18	50

From the table,

- Percentage Specificity $= \frac{26}{31} \times 100\% = 84\%$,
- Percentage Sensitivity $= \frac{13}{19} \times 100\% = 68\%$,
- Percentage Accuracy $= \frac{26+13}{50} \times 100\% = 78\%$.

Comparing the training test set results reveals clearly that this model performs better on the training data than the test data, this is because it has a considerably higher sensitivity ratio in the training data set than the test data set while maintaining the same ratio of specificity in both the training and test sets. This makes the overall predictive accuracy higher in the training set. It can be inferred that the model does best in out-sample specificity i.e. identifying new observations that are actually failures.

4.3.3 Results of Proposed Method with only Significant Variables

In section 4.2.1, the results of employing the Pattern Recognition method on the data were presented and in section 4.3.2, the results of employing the Logistic Regression method were presented. The final logistic regression model contains only two of the explanatory variables because the model containing only those two variables was found to be the model containing the significant variables and at the same time producing the highest accuracy. The Pattern Recognition method is employed on the data using only the variables contained in the final logistic regression model in order to compare the performances of the two methods based on the same set of variables. Both of these significant explanatory variables happen to be continuous, simplifying the entire data set into a single cluster of observations since no categorical variables have been found to be significant. The matrix of principal components coefficients computed for the entire data (in this case) using the two variables is

```
            PC1         PC2
Math1 0.9052410 -0.4248984
f_av  0.4248984  0.90524
```

This means that the principal components of each observations $O_i = (\text{Math1}, \text{f\_av})$ is given by

$$S_i = (0.91\text{Math1} + 0.42\,\text{f\_av}, -0.42\text{Math1} + 0.91\text{f\_av}) \quad (4.2)$$

Table 4.23: Summary of P.C.A of the Data Using the Two Significant Predictors

Component	Standard deviation	Proportion of Variance	Cumulative Proportion
PC1	20.1408	0.8834	0.8834
PC2	7.3170	0.1166	1.0000

Table 4.23 gives the correlations between the original variables and the new Principal Components.

Table 4.24: Correlation between Original Components and Principal Components

	Math1	f_av	PC1	PC2
Math1	1.00000	0.67666	0.98577	-0.16809
f_av	0.67666	1.00000	0.79080	0.61207
PC1	0.98577	0.79080	1.00000	0.00000
PC2	-0.16809	0.61207	0.00000	1.00000

The matrix reveals Math1 contributes significantly to the determination of the first principal component even more than how f_av contributes to the determination of the second principal component. Meanwhile the first principal component alone contributes significantly more to the

variation in the entire data compared to the second component, this, essentially, is highlighting the significance of Math1 in determining the variation in the data as revealed by variance of the variables calculated in section 1 of this chapter.

Now, the principal components of the observations are segregated into two classes, based on the class labels of the observations (as discussed earlier) and means/estimated centers have been computed for each class. Denoting the mean vectors of the columns of the classes labeled 0 and 1 by $\mathbf{M_0}$ and $\mathbf{M_1}$ respectively gives

$$\mathbf{M_0} = (66.31934, 27.81090)$$

$$\mathbf{M_1} = (94.84, 31.45)$$

The Euclidean distance between these two vectors is 28.75. The classification of new observations is carried out as described already, using these two vectors as the estimated centers of the two classes and the results are presented subsequently.

The result of the training set is in Appendix 7.6, the structure is displayed in Table 4.25

Table 4.25: Result of Proposed Method with Significant Variables on Training Set

std	D	d0	d1	d	d/D	P. Admit	Admit	Accuracy
1	28.75	9.5	19.57	10.07	0.35	0	1	Inacc.
2	28.75	5.49	32.88	27.39	0.95	0	0	Acc.
3	28.75	9.4	24.34	14.94	0.52	0	0	Acc.
…	…	…	…	…	…	…	…	…
…	…	…	…	…	…	…	…	…
239	28.75	11.8	19.34	7.54	0.26	0	0	Acc.
240	28.75	18.5	10.27	8.23	0.29	1	0	Inacc.
241	28.75	14.31	40.88	26.57	0.92	0	0	Acc.

The confusion matrix of the results for the training set is Table 4.26

Table 4.26: Confusion Matrix of results of Proposed Method with only Significant Variables on Training Set

	Predicted Class		
Actual Class	0	1	Total
0	91	19	110
1	14	67	81
Total	105	86	191

- Percentage Specificity $= \frac{91}{110} \times 100\% = 83\%$
- Percentage Sensitivity $= \frac{67}{81} \times 100\% = 83\%$
- Percentage Accuracy $= \frac{91+67}{191} \times 100\% = 83\%$

The full result of test set is in appendix 7.7, the associated confusion matrix is Table 4.28

Table 4.27: Result of Proposed Method with Significant Variables on Test Set

Std.	**D**	**d0**	**d1**	**d**	**d/D**	**P. Admit**	**Admit**	**Accuracy**
11	28.75	17.17	13.95	3.22	0.11	1	0	Inacc.
13	28.75	11.42	36.96	25.54	0.89	0	0	Acc.
21	28.75	27.91	7.59	20.32	0.71	1	1	Acc.
…	…	…	…	…	…	…	…	…
…	…	…	…	…	…	…	…	…
…	…	…	…	…	…	…	…	…
180	28.75	14.89	14.33	0.56	0.02	1	0	Inacc.
192	28.75	7.25	34.78	27.53	0.96	0	0	Acc.
214	28.75	7.95	29.22	21.27	0.74	0	1	Inacc.

Table 4.28: Confusion Matrix of Proposed Meth. With only Sig. Variables on Test Set

	Predicted Class

Actual Class	0	1	Total
0	27	4	31
1	7	12	19
Total	34	16	50

- Percentage Specificity $= \frac{27}{31} \times 100\% = 87\%$
- Percentage Sensitivity $= \frac{12}{19} \times 100\% = 63\%$
- Percentage Accuracy $= \frac{27+12}{50} \times 100\% = 78\%$

The model performs better in the training set compared to the test set with an accuracy of 83% in the training data and 78% in the test set. Considering the performance of this model on the training data and new observations, it can be seen that the strength of the model lies in identifying failures (in this case students whose performances will not help them to gain admission), this is because the model has a high percentage of specificity, 83% both in the training data and in the test data set. In terms of identifying the event of interest (Sensitivity), it does significantly better on the training data compared to the test data.

4.3.4 Concluding Notes on the Results of the Two Methods

So far, the proposed method was applied on the testing/validation data set twice; in the first instance, all predictor variables were included in the model, in the second instance, only variables found to be significant from the logistic regression analysis were included in the model. In both instances, the method's predictive accuracy remains 78%. This result coincides with result of the logistic regression also gave a 78% overall accuracy. On the measure of specificity, the proposed method maintains a higher value of 87% in both instances while the logistic regression method

records a specificity of 84%. The logistic regression approach records a marginal higher sensitivity of 68% while the proposed method maintains 63% specificity in both cases. On a whole, this results suggests that both methods are equally good.

5. CONCLUSIONS AND RECOMMENDATIONS

5.1 Conclusions

The conclusions follow from the research objectives.

5.1.1 Conclusions on General Objective

A good method of classification has been developed, based on the concepts Principal Components Analysis and Pattern Trace Networks. It does not require a sufficiently large sample for its implementation. The proposed method is applicable to solving a classification problem in the education industry.

5.1.2 Conclusions on Specific Objective 1

Research objective 1 states: "to determine whether or not the proposed pattern recognition method is a good alternative to the logistic regression method by comparing the predictive performances of the two methods on several metrics". Judging from the predictive performances of the two methods on the test set, the study concludes that the predictive accuracies of the two methods are identical. The second conclusion is that: The strength of both methods in predicting a particular class label varies directly as the relative proportion of observations with that class label in the training data (the prior probability of the class). The next conclusion is that: Both methods are better at predicting failure than success. The final conclusion is that: The proposed method is only slightly better at predicting failure (specificity) compared to Logistic Regression while the Logistic Regression is also slightly better at predicting success (sensitivity) than the proposed method. There remain the suspicion that the proposed method can perform better than logistic regression.

5.1.3 Conclusions on Specific Objective 2

Research objective 2 states: "Find out which variables/factors are significant in predicting students' success in the WASSCE (qualification for entry into a bachelors program in Ghanaian universities) and the extent to which they do so". The study concludes that, marks obtained by students in Core-Mathematics are important indicators of their qualification for admission into the universities. The study also reveals that average of the marks obtained by a student in Social Studies, Integrated Science, and the elective subjects is very significant in forecasting students' success in the WASSCE, measured by qualification for admission into the university. The study further maintain that that marks obtained by students in the third term of the first year are the most significant indicators of success (qualification for admission into the university) in the WASSCE as far as the variables used in this study are concerned.

5.2 Recommendations

Consequent to the findings of this study, the proposed method is hereby recommended as an appropriate algorithm for classification problems; a very good alternative to the Logistic Regression. It is also recommended that: for a data set with more explanatory variables, especially, categorical variables, with relatively small sample size, a Pattern Recognition approach should be considered for solving the classification problem. This is because the implementation of the Logistic Regression depends on sufficiently large sample size. For relatively smaller sample sizes, where there are more categorical variables, the problem of sparsing (described in chapter 3) is most likely to emerge and when this occurs, the logistic regression is not the best approach. The study also suggests that, in training a classifier, the training set should be constituted by equal proportions of objects of the pre-defined classes/categories/populations to mitigate bias associated with prevalence. Finally, it is recommended that the first year end of third term examination marks of

students should be taken seriously by parents, school authorities and all other stakeholders of education. In this regard, a drastic move should be initiated to pay special attention to students who would perform abysmally in this examinations, it will be very good to make such students repeat the class to improve upon their performance.

5.3 Concluding Remarks

This study has successfully formulated a classification algorithm based on the concept of pattern tracing/recognition and was successful at applying it on an educational data. It should have been able to apply it to several data to confirm the capability of the proposed method. The study has quite extensively but not exhaustively reviewed issues relating to classification and classifier evaluation, only that it has used simplistic and biased measures to evaluate the performance of the classifiers. It may be true that this is not the first study that has designed a classifier based on the concepts of pattern recognition and Principal Components Analysis but it may not be true that this study has not designed a non-probability classifier and has at the same time associated a probability measure for indicating the reliability of the forecast. Additionally, this study has contributed to exemplifying the fact that a pattern recognition methodology can be successfully applied to a class prediction problem in the educational industry. In the near future, the distances of the observations from the mean vectors of the classes would be studied, their asymptotic distributions would be investigated to provide a broader perspective of the classifier's predictive strength.

6. REFERENCES

Agbaje, R.O., & Alake, E.M. (2014). Students' Variables as Predictor of Secondary School Students' Academic Achievement in Science Subjects. International Journal of Scientific and Research Publications, 4(9), 1–5.

Albright, J.(n.d.). *What is the Difference Between Logit and Probit Models?*. Retrieved from http://www.methodsconsultants.com/tutorial/what-is-the-difference-between-logit-and-probit-models/

Alkalin, V. (2003). Face Recognition Using Eigenfaces and Neural Networks. The Middle East Technical University.

Asiedu, L. (2016). Statistical Assessment of PCA / SVD and FFT- PCA / SVD on Variable Facial Expressions Original Research Article, (January). http://doi.org/10.9734/BJMCS/2016/22141

Asiedu, L., O. Adebanji, A., Oduro, F., & O. Mettle, F. (2015). Statistical Evaluation of Face Recognition Techniques under Variable Environmental Constraints. International Journal of Statistics and Probability, 4(4). http://doi.org/10.5539/ijsp.v4n4p93

Bishop, C. M. (2006). Pattern recognition. *Machine Learning*, *128*.

Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection: A survey. *ACM computing surveys (CSUR)*, *41*(3), 15.

Corso, J. (2013). Introduction to Pattern Recognition

Czepiel, S. A. (2012). Maximum Likelihood Estimation of Logistic Regression Models: Theory and Implementation. Class Notes, 1–23. Retrieved from papers3://publication/uuid/4E1E1B7E-9CAC-4570-8949-E96B51D9C91D

Elkan, C. (2012). Evaluating classifiers. University of San Diego, California, retrieved [01-11-2012] from http://cseweb. ucsd. edu/~ elkan B, 250.

Everitt, B., & Hothorn, T. (2011). An introduction to applied multivariate analysis with R. Media. http://doi.org/10.1007/978-1-4419-9650-3

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of eugenics*, *7*(2), 179-188.

Ghatge, M. A. R., & Halkarnikar, P. P. (2012). Estimation of Credit Risk for Business Firms of Nationalized Bank by Neural Network Approach. International Journal of Electronics and Computer Science Engineering, 2, 828–834.

Gutierrez-Osuna, R. (2005). Pattern Recognition - Lecture 10: Linear discriminants analysis, 1–738. Retrieved from http://research.cs.tamu.edu/prism/lectures/pr/pr_l10.pdf

Johnson, R. A., & Wichern, D. W. (2007). Applied multivariate statistical analysis (6th edition). Upper Saddle River, NJ: Prentice hall.

Lahiri, K., & Yang, L. (2012). Forecasting binary outcomes. *Handbook of economic forecasting*, *2*.

Lahiri, R. (2006). Comparison of Data Mining and Statistical Techniques for Classification Model.

Lozano, J.A., Santafé, G. Inz, I. (2010). Classifier performance evaluation and Comparison. International Conference on Machine Learning and Applications (ICMLA 2010).

Martin, M.C. (2009). Crop Yield Prediction Using Artificial Neural Networks and Genetic Algorithms.

Minaei-Bidgoli, B., Kashy, D. a., Kortemeyer, G., & Punch, W. F. (2003). Predicting student performance: an application of data mining methods with an educational web-based system. 33rd Annual Frontiers in Education, 2003. FIE 2003., 1, T2A_13–T2A_18. http://doi.org/10.1109/FIE.2003.1263284

Odeh, O.O., Featherstone, A.M. & Sanjoy, D. (2006). Predicting Credit Default in an Agricultural Bank: Methods and Issues. Selected Paper prepared for presentation at the Orlando, Florida, February 5-8, 2006Southern Agricultural Economics Association Annual Meetings

Omirin, M.S., & Ale V.M. (2008). Predictive Validity of English and Mathematics Mock Examination Results of Senior Secondary School Students Performance in WASCE in Ekiti-State Nigeria. Pakistan Journal of Social Sciences 5(2): 139-141, 2008

Powers, D. M. W. (2007). Evaluation: From Precision, Recall and F-Factor to ROC, Informedness, Markedness & Correlation, 2(1), 24. http://doi.org/10.1.1.214.9232

Ripley, B. D. (1997) Classification (update). In *Encyclopedia of Statistical Sciences*, eds S. Kotz, C. B. Read and D. L. Banks, volume Update 1. New York: John Wiley and Sons. [331]

Shayib, M.A., (2013). Applied Statistics (1st edition).

Singh,S.(2014). Introduction to pattern recognition. Retrieved from https://www.youtube.com/watch?v=mWYUx_HJeSM

Van Rijsbergen, C. J. (1979). Information Retrieval (2nd edition.). Butterworth.

Wasserman, L. (2011). All of statistics: A Concise Course in Statistical Inference. http://doi.org/10.1017/CBO9781107415324.004

Wushishi, D. I., & Usman, H. (2013). Relationship Between Senior Secondary School Certificate Examination (SSCE) Mathematics Grades And Final Nigeria Certificate Of Education (NCE) Mathematics Students Results Of Niger State College Of Education Minna. International Journal of Humanities and Social Science Invention, 2(2), 16–21.

Yadav, S., Bharadwaj, B., & Pal, S. (2012). Data mining applications: A comparative study for predicting student's performance. International Journal of Innovative Technology & Creative Engineering, 1(12), 13–19. Retrieved from http://arxiv.org/abs/1202.4815

Ghana Education Service Teaching Syllabus for Core Mathematics for Senior High Schools (2010)

7. APPENDIX

Table 7.1: The Study Data Set

std	Admit	prog_ code	resid_ code	gen_c ode	Entry Grade	Eng1	Math 1	f_av	Eng2	Math 2	s_av
1	1	2	1	0	7	64	55	60	60	55	65
2	0	1	1	0	17	50	47	48	46	49	43
3	0	1	0	0	26	59	57	50	60	60	53
4	0	4	1	1	15	50	31	46	51	28	46
5	0	2	1	0	27	54	26	47	58	36	47
6	0	3	1	1	14	81	77	58	56	44	58
7	1	2	1	1	14	71	84	62	61	53	53
8	1	3	1	0	13	81	65	76	69	69	73
9	0	4	1	0	18	44	52	55	49	66	61
10	0	4	0	1	15	70	29	55	55	45	64
11	0	3	1	1	14	66	65	57	50	41	57
12	0	1	1	1	19	46	53	47	61	71	52
13	0	4	0	1	18	38	47	42	45	45	45
14	0	4	1	1	12	57	51	61	47	12	63
15	0	2	0	0	12	68	33	54	55	48	62
16	1	1	1	1	15	52	73	66	58	78	59
17	1	3	0	1	8	88	99	84	75	89	84
18	0	3	1	0	15	75	76	62	70	46	61
19	0	4	0	1	16	43	55	49	42	64	54
20	0	4	0	0	14	65	58	52	74	62	61
21	1	3	1	0	11	81	67	74	62	68	64
22	0	4	1	1	19	48	69	58	52	49	55
23	1	1	0	1	14	60	64	49	49	50	48
24	0	4	1	1	15	51	64	60	46	69	59
25	0	4	0	0	28	40	50	55	45	60	62
26	1	1	0	1	14	66	62	49	52	63	58
27	1	3	1	0	19	57	62	63	63	57	63
28	1	3	1	1	10	68	87	77	59	72	68
29	1	3	1	1	10	62	72	74	73	94	76
30	1	2	1	1	8	73	89	77	77	75	72
31	1	2	1	1	21	67	74	61	65	63	64
32	0	4	0	1	22	60	49	56	61	50	55
33	1	3	1	1	15	73	70	60	71	64	71
34	0	4	1	0	14	42	45	56	46	53	57
35	0	4	0	1	17	45	38	40	48	41	37
36	1	3	1	1	10	72	81	71	67	76	75

std	Admit	prog_ code	resid_ code	gen_c ode	Entry Grade	Eng1	Math 1	f_av	Eng2	Math 2	s_av
37	0	2	1	1	9	67	26	49	55	31	49
38	1	1	1	1	10	58	79	74	61	78	69
39	1	1	1	0	13	59	85	69	66	86	73
40	0	2	1	0	17	70	40	53	64	35	55
41	1	3	0	0	6	73	82	72	68	81	69
42	1	1	1	1	15	54	47	43	53	62	54
43	0	2	1	1	12	78	57	62	56	40	58
44	1	2	1	1	15	57	59	50	51	69	59
45	1	3	1	1	11	65	77	73	68	63	63
46	0	2	1	1	15	60	63	53	57	71	58
47	1	3	0	1	12	64	71	69	59	68	65
48	1	1	1	0	15	71	48	60	51	56	53
49	0	2	1	0	17	70	44	62	54	45	63
50	1	4	1	1	16	58	68	71	86	65	78
51	0	2	1	0	15	83	56	65	77	34	63
52	1	2	1	0	7	69	40	61	70	59	63
53	1	1	1	1	17	66	67	62	54	59	59
54	0	2	1	0	15	47	40	52	58	42	47
55	1	3	1	1	11	51	85	80	65	82	73
56	1	3	1	1	11	58	83	82	61	64	77
57	0	2	1	1	13	62	59	53	51	47	58
58	0	2	1	1	23	54	69	50	57	58	59
59	1	1	1	0	15	75	78	74	56	71	65
60	1	2	1	1	12	65	74	64	59	69	69
61	1	3	1	1	26	61	77	67	67	59	73
62	0	2	0	0	10	67	51	58	73	44	50
63	1	2	0	0	6	72	87	76	67	84	81
64	1	1	1	1	10	64	57	58	62	70	60
65	1	1	1	0	9	77	82	75	62	77	77
66	0	3	0	1	11	59	62	62	47	53	62
67	1	2	1	0	19	56	39	54	54	38	66
68	1	1	1	1	15	67	45	65	59	52	61
69	1	3	0	1	16	52	60	63	59	46	64
70	0	1	1	0	13	57	53	44	56	63	45
71	1	1	1	1	15	75	74	76	68	62	69
72	1	3	1	0	11	65	0	69	51	63	64
73	1	3	1	1	9	74	77	71	69	59	75
74	0	1	1	0	15	53	47	47	53	61	54
75	0	1	0	0	13	53	27	55	52	41	55
76	1	1	1	1	16	74	76	71	62	68	71
77	0	4	1	0	15	43	36	60	58	25	50
78	1	3	0	1	8	63	64	51	56	54	145

std	Admit	prog_ code	resid_ code	gen_c ode	Entry Grade	Eng1	Math 1	f_av	Eng2	Math 2	s_av
79	0	1	1	0	15	53	49	53	51	54	48
80	1	2	1	0	14	43	66	60	70	46	53
81	0	2	1	0	17	61	35	46	59	45	49
82	1	4	0	0	14	65	90	72	60	77	76
83	0	3	1	1	12	63	47	65	73	55	61
84	0	1	1	0	16	48	35	52	57	53	46
85	0	1	1	1	15	48	36	49	45	45	52
86	1	2	0	1	12	60	75	74	66	78	78
87	0	2	1	0	13	58	59	47	70	46	49
88	0	2	1	0	14	65	46	62	70	38	58
89	0	2	0	0	19	59	41	50	52	37	60
90	0	2	0	0	23	51	48	38	48	23	46
91	0	4	1	0	18	47	40	62	49	52	63
92	1	3	1	0	6	62	65	58	56	60	64
93	0	2	0	0	21	60	35	52	54	35	49
94	0	2	0	1	13	66	67	52	69	52	57
95	0	3	1	1	10	66	35	42	51	36	54
96	0	4	0	0	16	47	51	46	53	46	52
97	0	4	0	0	27	32	42	46	46	56	46
98	0	1	1	0	20	61	44	57	53	55	52
99	0	3	1	0	18	57	46	53	52	46	56
100	0	4	1	0	21	48	35	54	50	42	53
101	0	4	1	1	9	61	23	47	55	14	45
102	0	2	1	0	13	69	39	51	57	45	57
103	0	4	1	1	20	48	39	58	67	50	71
104	0	3	0	1	12	54	64	72	54	55	65
105	1	3	1	1	11	77	67	75	61	63	77
106	0	2	1	0	14	58	44	54	51	44	52
107	1	3	0	1	11	78	65	63	62	46	68
108	1	3	0	1	7	68	74	68	51	64	75
109	0	1	1	1	14	42	76	57	53	73	59
110	1	4	1	1	23	54	54	57	58	64	62
111	0	2	1	1	14	56	49	50	56	43	52
112	0	1	1	0	9	74	35	61	63	56	64
113	0	2	1	1	16	73	48	61	64	41	58
114	1	3	1	1	10	57	88	67	62	78	67
115	0	4	1	0	14	62	69	74	60	54	75
116	0	2	1	0	15	68	45	49	46	34	48
117	1	2	0	1	11	66	62	68	57	72	70
118	0	1	1	0	10	50	61	45	36	50	50
119	0	4	0	0	20	52	10	68	44	36	65
120	0	4	1	0	7	58	49	57	64	43	69

std	Admit	prog_ code	resid_ code	gen_c ode	Entry Grade	Eng1	Math 1	f_av	Eng2	Math 2	s_av
121	1	3	1	0	13	65	90	68	71	83	75
122	0	2	1	0	19	59	48	42	62	57	50
123	0	4	1	0	16	45	55	48	45	66	51
124	0	2	0	0	15	64	29	58	65	43	57
125	0	2	1	0	16	72	33	57	72	58	61
126	0	4	0	1	17	46	36	53	56	50	63
127	0	1	1	1	14	67	48	44	46	40	49
128	1	3	0	0	12	72	79	72	64	67	70
129	0	4	1	0	11	55	72	66	59	71	76
130	0	1	0	1	14	48	61	58	60	62	55
131	0	2	1	1	12	77	27	51	71	37	54
132	0	3	1	1	13	59	66	67	47	60	65
133	1	3	1	0	6	57	70	69	59	38	67
134	1	1	0	1	15	69	77	79	68	76	74
135	1	3	1	0	13	58	81	76	73	80	69
136	0	2	1	1	20	45	40	48	54	61	59
137	0	2	0	1	13	45	42	49	58	43	65
138	0	4	0	0	19	48	54	71	55	73	78
139	0	4	1	1	15	53	36	39	38	45	48
140	1	1	1	1	14	66	86	78	70	80	73
141	0	2	1	1	13	61	45	48	57	42	57
142	0	3	0	1	21	34	60	40	60	47	47
143	0	1	1	0	18	77	40	52	56	45	54
144	1	3	1	1	9	56	79	74	53	54	73
145	0	3	0	1	16	61	51	61	43	54	65
146	1	3	1	1	17	64	81	77	50	63	75
147	0	2	1	0	21	53	57	56	55	48	59
148	1	2	0	0	16	67	48	57	52	60	58
149	1	3	1	1	10	76	86	72	65	57	69
150	0	2	1	0	15	66	60	58	62	50	60
151	1	3	0	1	13	68	75	67	69	69	69
152	0	4	1	0	13	49	71	49	53	70	64
153	1	4	1	0	16	58	80	69	60	77	67
154	1	2	0	1	11	70	71	61	67	63	66
155	1	1	1	1	9	74	69	64	69	78	63
156	1	1	0	1	13	50	55	78	64	53	67
157	1	2	1	1	17	80	83	73	76	71	65
158	1	2	1	1	13	68	88	82	75	71	81
159	1	1	0	1	13	58	62	62	55	65	63
160	0	1	0	0	16	67	32	62	57	50	55
161	1	3	1	1	15	73	80	69	59	60	64
162	0	4	1	1	16	48	64	60	34	63	54

std	Admit	prog_ code	resid_ code	gen_c ode	Entry Grade	Eng1	Math 1	f_av	Eng2	Math 2	s_av
163	1	2	1	0	13	62	39	59	55	40	61
164	0	2	1	0	18	67	33	60	52	48	63
165	0	4	0	1	17	56	44	47	54	67	46
166	1	2	1	0	11	72	53	57	70	47	61
167	1	3	1	0	7	72	52	51	66	53	62
168	1	2	1	0	19	56	35	64	48	40	64
169	0	3	1	1	11	65	57	63	51	43	67
170	0	1	0	0	15	65	44	48	55	52	59
171	0	2	1	0	12	58	41	33	51	45	51
172	1	3	1	1	10	65	86	72	61	84	81
173	0	4	1	0	15	52	41	48	48	49	53
174	0	4	1	0	18	44	31	59	61	19	44
175	0	2	0	1	16	59	51	41	53	75	66
176	1	3	1	1	15	66	69	69	63	61	68
177	1	1	0	1	17	67	73	76	54	81	75
178	0	1	1	1	19	47	75	64	53	68	64
179	1	1	1	1	18	61	68	69	59	62	61
180	0	3	1	0	9	63	62	59	51	51	60
181	0	1	1	0	13	61	50	48	50	59	49
182	0	4	1	0	15	36	26	49	31	41	47
183	0	2	1	0	18	66	45	53	56	39	51
184	0	4	0	1	17	55	51	46	43	51	56
185	0	3	1	1	17	57	53	56	52	47	52
186	1	1	1	0	15	83	42	56	66	67	63
187	1	3	1	0	7	71	70	68	61	57	63
188	1	3	0	1	13	50	76	61	67	40	64
189	1	3	1	1	12	66	91	70	69	80	79
190	0	1	1	1	11	50	70	67	64	62	57
191	0	4	0	1	18	26	40	46	54	57	56
192	0	1	1	1	17	66	41	54	53	63	50
193	0	2	1	0	22	66	27	51	42	57	62
194	1	3	1	1	8	67	80	74	55	64	74
195	0	2	1	0	13	67	52	57	60	58	66
196	1	2	1	0	10	77	86	80	74	60	74
197	0	2	1	0	11	69	78	57	68	67	58
198	0	2	1	0	14	50	36	41	58	29	50
199	1	2	1	1	9	73	68	60	64	63	54
200	1	2	1	0	13	76	66	64	68	59	66
201	0	4	1	0	17	54	64	53	40	49	47
202	0	1	0	0	17	66	47	61	53	63	62
203	1	1	1	1	13	71	57	66	62	75	61
204	0	4	0	1	18	39	31	61	29	45	55

std	Admit	prog_ code	resid_ code	gen_c ode	Entry Grade	Eng1	Math 1	f_av	Eng2	Math 2	s_av
205	1	2	0	0	9	77	70	76	73	54	76
206	0	3	1	1	8	68	68	54	59	66	60
207	1	3	1	0	13	83	61	72	74	65	69
208	0	4	0	1	21	30	38	33	44	39	39
209	1	3	0	1	11	86	98	89	75	83	83
210	0	2	1	1	14	59	53	42	37	43	45
211	0	4	1	0	18	49	42	60	53	44	59
212	0	2	0	0	18	68	52	54	53	48	51
213	0	4	0	0	10	57	42	47	45	66	57
214	1	1	1	1	14	73	53	47	64	63	62
215	0	4	0	1	27	58	41	48	63	52	56
216	0	2	1	0	23	65	48	48	57	52	49
217	0	2	1	0	15	61	42	40	56	53	55
218	0	2	1	1	12	63	73	49	55	61	59
219	1	1	1	1	13	66	75	65	59	63	72
220	1	3	1	1	10	61	81	67	68	73	75
221	0	3	1	0	10	51	78	58	57	64	54
222	1	3	1	1	12	62	86	73	64	64	70
223	0	2	1	0	18	61	34	45	61	43	52
224	1	1	1	0	19	60	82	74	65	71	72
225	0	4	0	0	23	46	34	55	54	18	31
226	1	3	1	1	14	69	73	65	69	75	69
227	0	4	1	0	15	31	45	57	45	54	56
228	0	1	1	1	11	50	53	58	52	63	51
229	0	2	1	0	17	53	33	54	50	41	59
230	0	4	1	0	19	42	50	48	49	57	59
231	0	3	0	1	9	68	61	60	51	48	54
232	0	2	0	0	22	63	31	55	62	41	62
233	0	2	0	0	20	46	48	45	62	54	58
234	0	2	1	1	15	63	42	52	64	62	55
235	1	3	0	1	19	63	91	76	58	83	78
236	1	3	1	0	7	64	72	63	66	63	64
237	0	2	1	0	15	55	22	47	52	20	55
238	1	2	1	1	11	58	37	74	62	57	73
239	0	2	0	0	17	70	60	54	58	50	57
240	0	4	1	1	21	50	64	63	58	66	65
241	0	2	1	1	25	69	34	55	58	52	56

Table 7.2: Full Results of Proposed Method on Training Set

std	Admit	Cluster	D	d0	d1	d	d/D	P.Admit	Accuracy
1	1	7	18.45	21.24	9.92	11.32	0.61	1	Acc.
2	0	3	44.83	15.05	50.64	35.59	0.79	0	Acc.
3	0	4	118.99	16.9	124.8	107.9	0.91	0	Acc.
4	0	13	32.8	31.97	57.45	25.48	0.78	0	Acc.
5	0	7	18.45	23.57	38.69	15.12	0.82	0	Acc.
6	0	9	26.3	23.81	30.78	6.97	0.27	0	Acc.
7	1	5	34.47	40.01	22.08	17.93	0.52	1	Acc.
8	1	11	18.29	33.74	19.21	14.53	0.79	1	Acc.
9	0	15	49.16	20.02	37.83	17.81	0.36	0	Acc.
10	0	14	122.31	29.48	130.22	100.74	0.82	0	Acc.
12	0	1	23.47	19	35.91	16.91	0.72	0	Acc.
14	0	13	32.8	37.56	58.14	20.58	0.63	0	Acc.
15	0	8	54.51	13.11	57.3	44.19	0.81	0	Acc.
16	1	1	23.47	21.47	18.22	3.25	0.14	1	Acc.
17	1	10	30.91	69.1	40.35	28.75	0.93	1	Acc.
18	0	11	18.29	21.23	23.25	2.02	0.11	0	Acc.
19	0	14	122.31	30.9	142.41	111.51	0.91	0	Acc.
20	0	16	60.71	33.75	42.97	9.22	0.15	0	Acc.
22	0	13	32.8	17.57	34.86	17.29	0.53	0	Acc.
23	1	2	21.88	28.07	29.19	1.12	0.05	0	Inacc.
24	0	13	32.8	25.11	29.66	4.55	0.14	0	Acc.
27	1	11	18.29	12.5	16.51	4.01	0.22	0	Inacc.
28	1	9	26.3	36.92	13.79	23.13	0.88	1	Acc.
30	1	5	34.47	59.06	25.75	33.31	0.97	1	Acc.
33	1	9	26.3	23.67	21.19	2.48	0.09	1	Acc.
34	0	15	49.16	11.24	52.11	40.87	0.83	0	Acc.
35	0	14	122.31	13.19	114.83	101.64	0.83	0	Acc.
37	0	5	34.47	33.81	62.77	28.96	0.84	0	Acc.
38	1	1	23.47	30.99	20.64	10.35	0.44	1	Acc.
39	1	3	44.83	56.19	20.37	35.82	0.8	1	Acc.
41	1	12	6	6	0	6	1	1	Acc.
44	1	5	34.47	23.27	31.56	8.29	0.24	0	Inacc.
45	1	9	26.3	26.13	10.02	16.11	0.61	1	Acc.
46	0	5	34.47	23.67	22.14	1.53	0.04	1	Inacc.
49	0	7	18.45	19.57	22.7	3.13	0.17	0	Acc.
53	1	1	23.47	18.42	14.46	3.96	0.17	1	Acc.
54	0	7	18.45	21.39	28.37	6.98	0.38	0	Acc.
56	1	9	26.3	34.68	16.83	17.85	0.68	1	Acc.
58	0	5	34.47	23.1	28.18	5.08	0.15	0	Acc.
59	1	3	44.83	50.22	12.23	37.99	0.85	1	Acc.
60	1	5	34.47	35.41	14.32	21.09	0.61	1	Acc.
61	1	9	26.3	33.19	26.61	6.58	0.25	1	Acc.
63	1	8	54.51	67.92	18.3	49.62	0.91	1	Acc.

std	Admit	Cluster	D	d0	d1	d	d/D	P.Admit	Accuracy
64	1	1	23.47	20.78	16.34	4.44	0.19	1	Acc.
65	1	3	44.83	59.03	16.92	42.11	0.94	1	Acc.
66	0	10	30.91	5.14	31.54	26.4	0.85	0	Acc.
67	1	7	18.45	16.28	24.93	8.65	0.47	0	Inacc.
68	1	1	23.47	31.22	30.31	0.91	0.04	1	Acc.
69	1	10	30.91	15.54	34.42	18.88	0.61	0	Inacc.
71	1	1	23.47	37.62	16.52	21.1	0.9	1	Acc.
73	1	9	26.3	25.39	13.02	12.37	0.47	1	Acc.
74	0	3	44.83	9.77	43.51	33.74	0.75	0	Acc.
76	1	1	23.47	32.29	13.23	19.06	0.81	1	Acc.
79	0	3	44.83	9.76	47.09	37.33	0.83	0	Acc.
82	1	16	60.71	60.71	0	60.71	1	1	Acc.
83	0	9	26.3	25.54	39.09	13.55	0.52	0	Acc.
84	0	3	44.83	15.22	55.08	39.86	0.89	0	Acc.
86	1	6	27.12	35.51	11.36	24.15	0.89	1	Acc.
87	0	7	18.45	23.11	23.96	0.85	0.05	0	Acc.
89	0	8	54.51	13.71	63.13	49.42	0.91	0	Acc.
91	0	15	49.16	12.46	50.21	37.75	0.77	0	Acc.
92	1	11	18.29	16.27	15.57	0.7	0.04	1	Acc.
93	0	8	54.51	20.29	66.37	46.08	0.85	0	Acc.
94	0	6	27.12	22.38	27.78	5.4	0.2	0	Acc.
96	0	16	60.71	15.83	60.86	45.03	0.74	0	Acc.
98	0	3	44.83	13.56	46.32	32.76	0.73	0	Acc.
99	0	11	18.29	24.17	37.87	13.7	0.75	0	Acc.
100	0	15	49.16	16.3	62.42	46.12	0.94	0	Acc.
102	0	7	18.45	12.51	19.75	7.24	0.39	0	Acc.
103	0	13	32.8	23.94	40.46	16.52	0.5	0	Acc.
104	0	10	30.91	13.11	26.22	13.11	0.42	0	Acc.
105	1	9	26.3	22.85	18.35	4.5	0.17	1	Acc.
107	1	10	30.91	24.15	28.31	4.16	0.13	0	Inacc.
108	1	10	30.91	21.14	16.32	4.82	0.16	1	Acc.
109	0	1	23.47	18.96	29.74	10.78	0.46	0	Acc.
110	1	13	32.8	32.8	0	32.8	1	1	Acc.
112	0	3	44.83	27.03	47.82	20.79	0.46	0	Acc.
113	0	5	34.47	22.34	37.41	15.07	0.44	0	Acc.
114	1	9	26.3	44.38	26.74	17.64	0.67	1	Acc.
116	0	7	18.45	20.72	31.09	10.37	0.56	0	Acc.
118	0	3	44.83	24.82	50.81	25.99	0.58	0	Acc.
119	0	16	60.71	37.98	93.16	55.18	0.91	0	Acc.
120	0	15	49.16	27.73	53.73	26	0.53	0	Acc.
121	1	11	18.29	42.16	32.36	9.8	0.54	1	Acc.
123	0	15	49.16	23.49	44.97	21.48	0.44	0	Acc.
124	0	8	54.51	15.6	60.51	44.91	0.82	0	Acc.

std	Admit	Cluster	D	d0	d1	d	d/D	P.Admit	Accuracy
125	0	7	18.45	30.05	28.79	1.26	0.07	1	Inacc.
127	0	1	23.47	35.61	47.15	11.54	0.49	0	Acc.
130	0	2	21.88	0	21.88	21.88	1	0	Acc.
132	0	9	26.3	17.03	24.03	7	0.27	0	Acc.
133	1	11	18.29	21.91	27.82	5.91	0.32	0	Inacc.
134	1	2	21.88	38.99	20.8	18.19	0.83	1	Acc.
135	1	11	18.29	46.07	31.4	14.67	0.8	1	Acc.
136	0	5	34.47	22.51	47.48	24.97	0.72	0	Acc.
138	0	16	60.71	30.45	44.06	13.61	0.22	0	Acc.
139	0	13	32.8	26.38	48.72	22.34	0.68	0	Acc.
140	1	1	23.47	42.43	25.23	17.2	0.73	1	Acc.
141	0	5	34.47	14.34	44.03	29.69	0.86	0	Acc.
144	1	9	26.3	25.59	20.8	4.79	0.18	1	Acc.
145	0	10	30.91	13.3	40.16	26.86	0.87	0	Acc.
146	1	9	26.3	29.17	15.63	13.54	0.51	1	Acc.
147	0	7	18.45	20.88	18.8	2.08	0.11	1	Inacc.
149	1	9	26.3	30.89	15.09	15.8	0.6	1	Acc.
150	0	7	18.45	21.38	8.54	12.84	0.7	1	Inacc.
151	1	10	30.91	31.69	8.97	22.72	0.74	1	Acc.
152	0	15	49.16	32.84	26.17	6.67	0.14	1	Inacc.
153	1	15	49.16	49.16	0	49.16	1	1	Acc.
154	1	6	27.12	21.63	11.36	10.27	0.38	1	Acc.
155	1	1	23.47	37.33	18.38	18.95	0.81	1	Acc.
156	1	2	21.88	24.88	22.66	2.22	0.1	1	Acc.
157	1	5	34.47	52.82	21.6	31.22	0.91	1	Acc.
158	1	5	34.47	57.6	25.01	32.59	0.95	1	Acc.
160	0	4	118.99	18.01	113.04	95.03	0.8	0	Acc.
161	1	9	26.3	23.78	10.95	12.83	0.49	1	Acc.
162	0	13	32.8	27.75	31.1	3.35	0.1	0	Acc.
163	1	7	18.45	14.32	18.93	4.61	0.25	0	Inacc.
164	0	7	18.45	15.69	23.11	7.42	0.4	0	Acc.
165	0	14	122.31	21.92	132.19	110.27	0.9	0	Acc.
166	1	7	18.45	21.96	12.17	9.79	0.53	1	Acc.
167	1	11	18.29	32.18	30.2	1.98	0.11	1	Acc.
168	1	7	18.45	21.84	27.26	5.42	0.29	0	Inacc.
169	0	9	26.3	13.22	35.89	22.67	0.86	0	Acc.
170	0	4	118.99	9.91	110.83	100.92	0.85	0	Acc.
171	0	7	18.45	24.86	40.28	15.42	0.84	0	Acc.
172	1	9	26.3	42.19	19.38	22.81	0.87	1	Acc.
173	0	15	49.16	25.74	56.42	30.68	0.62	0	Acc.
174	0	15	49.16	37.5	78.66	41.16	0.84	0	Acc.
175	0	6	27.12	22.38	41.24	18.86	0.7	0	Acc.
176	1	9	26.3	16.52	12.9	3.62	0.14	1	Acc.

std	Admit	Cluster	D	d0	d1	d	d/D	P.Admit	Accuracy
177	1	2	21.88	35.16	21.2	13.96	0.64	1	Acc.
178	0	1	23.47	19.6	23.35	3.75	0.16	0	Acc.
179	1	1	23.47	25.27	15.02	10.25	0.44	1	Acc.
181	0	3	44.83	9.19	40.27	31.08	0.69	0	Acc.
182	0	15	49.16	34.28	77.47	43.19	0.88	0	Acc.
183	0	7	18.45	8.86	17.65	8.79	0.48	0	Acc.
184	0	14	122.31	14.24	127.77	113.53	0.93	0	Acc.
185	0	9	26.3	15.91	39.79	23.88	0.91	0	Acc.
186	1	3	44.83	33.67	38.9	5.23	0.12	0	Inacc.
187	1	11	18.29	17.43	13.38	4.05	0.22	1	Acc.
188	1	10	30.91	30.93	33.97	3.04	0.1	0	Inacc.
189	1	9	26.3	43.57	18.94	24.63	0.94	1	Acc.
190	0	1	23.47	16.89	20.43	3.54	0.15	0	Acc.
191	0	14	122.31	23.09	117.92	94.83	0.78	0	Acc.
193	0	7	18.45	27.02	39.41	12.39	0.67	0	Acc.
194	1	9	26.3	27.51	11.93	15.58	0.59	1	Acc.
195	0	7	18.45	22.59	21.47	1.12	0.06	1	Inacc.
196	1	7	18.45	59.55	42.4	17.15	0.93	1	Acc.
197	0	7	18.45	45.38	32.66	12.72	0.69	1	Inacc.
198	0	7	18.45	26.09	40.97	14.88	0.81	0	Acc.
199	1	5	34.47	29.25	12.98	16.27	0.47	1	Acc.
200	1	7	18.45	36.28	24.38	11.9	0.64	1	Acc.
201	0	15	49.16	21.63	41.98	20.35	0.41	0	Acc.
202	0	4	118.99	13.74	129.98	116.24	0.98	0	Acc.
203	1	1	23.47	28.01	14.39	13.62	0.58	1	Acc.
204	0	14	122.31	28.79	109.52	80.73	0.66	0	Acc.
205	1	8	54.51	44.72	18.3	26.42	0.48	1	Acc.
206	0	9	26.3	18.19	22.19	4	0.15	0	Acc.
207	1	11	18.29	34.01	20.7	13.31	0.73	1	Acc.
208	0	14	122.31	29.38	113.17	83.79	0.69	0	Acc.
209	1	10	30.91	68.86	40.2	28.66	0.93	1	Acc.
210	0	5	34.47	23.96	50.15	26.19	0.76	0	Acc.
211	0	15	49.16	16.71	56.11	39.4	0.8	0	Acc.
212	0	8	54.51	13.69	45.62	31.93	0.59	0	Acc.
213	0	16	60.71	22.25	59.59	37.34	0.62	0	Acc.
215	0	14	122.31	23.32	130.97	107.65	0.88	0	Acc.
216	0	7	18.45	11.78	19.37	7.59	0.41	0	Acc.
217	0	7	18.45	16.13	30.33	14.2	0.77	0	Acc.
218	0	5	34.47	26.96	24.34	2.62	0.08	1	Inacc.
219	1	1	23.47	22.36	11.06	11.3	0.48	1	Acc.
220	1	9	26.3	34.24	17.93	16.31	0.62	1	Acc.
221	0	11	18.29	20.75	22.22	1.47	0.08	0	Acc.
222	1	9	26.3	31.12	8.26	22.86	0.87	1	Acc.

std	Admit	Cluster	D	d0	d1	d	d/D	P.Admit	Accuracy
223	0	7	18.45	11.69	28.52	16.83	0.91	0	Acc.
224	1	3	44.83	48.75	16.11	32.64	0.73	1	Acc.
225	0	16	60.71	35.61	88.96	53.35	0.88	0	Acc.
226	1	9	26.3	29.03	14.64	14.39	0.55	1	Acc.
227	0	15	49.16	17.38	53.42	36.04	0.73	0	Acc.
228	0	1	23.47	11.99	27.36	15.37	0.65	0	Acc.
229	0	7	18.45	16.39	29.24	12.85	0.7	0	Acc.
230	0	15	49.16	13.1	48.5	35.4	0.72	0	Acc.
231	0	10	30.91	12.08	32.36	20.28	0.66	0	Acc.
232	0	8	54.51	20.29	63.4	43.11	0.79	0	Acc.
233	0	8	54.51	26.64	59.41	32.77	0.6	0	Acc.
234	0	5	34.47	15.04	35.4	20.36	0.59	0	Acc.
235	1	10	30.91	47.13	24.44	22.69	0.73	1	Acc.
236	1	11	18.29	17.31	13.4	3.91	0.21	1	Acc.
237	0	7	18.45	36.02	47.84	11.82	0.64	0	Acc.
238	1	5	34.47	28.81	39.51	10.7	0.31	0	Inacc.
239	0	8	54.51	24.87	41.81	16.94	0.31	0	Acc.
240	0	13	32.8	25	29.82	4.82	0.15	0	Acc.
241	0	5	34.47	20.44	45.79	25.35	0.74	0	Acc.

Table 7.3: Full Results of Proposed Method on Test Set

std	Admit	Cluster	D	d0	d1	d	d/D	P.Admit	Accuracy
11	0	9	26.3	16.96	35.91	18.95	0.72	0	Acc
13	0	14	122.31	19.72	123.56	103.84	0.85	0	Acc
21	1	11	18.29	32.8	17.9	14.9	0.81	1	Acc
25	0	16	60.71	24.83	56.92	32.09	0.53	0	Acc
26	1	2	21.88	25.53	22.96	2.57	0.12	1	Acc
29	1	9	26.3	48.88	32.73	16.15	0.61	1	Acc
31	1	5	34.47	31.57	12.26	19.31	0.56	1	Acc
32	0	14	122.31	22.38	139.01	116.63	0.95	0	Acc
36	1	9	26.3	36.41	17.28	19.13	0.73	1	Acc
40	0	7	18.45	17.4	22.2	4.8	0.26	0	Acc
42	1	1	23.47	32.76	41.66	8.9	0.38	0	Inacc
43	0	5	34.47	32.71	39.46	6.75	0.2	0	Acc
47	1	10	30.91	23.73	12.76	10.97	0.35	1	Acc
48	1	3	44.83	18.83	35.61	16.78	0.37	0	Inacc
50	1	13	32.8	47.55	48.97	1.42	0.04	0	Inacc
51	0	7	18.45	37.65	29.72	7.93	0.43	1	Inacc
52	1	7	18.45	26.57	20.85	5.72	0.31	1	Acc
55	1	9	26.3	48.41	29.67	18.74	0.71	1	Acc
57	0	5	34.47	12.69	31.93	19.24	0.56	0	Acc
62	0	8	54.51	23.86	44.91	21.05	0.39	0	Acc
70	0	3	44.83	18.22	45.22	27	0.6	0	Acc
72	1	11	18.29	69.27	70.68	1.41	0.08	0	Inacc
75	0	4	118.99	26.3	99.72	73.42	0.62	0	Acc
77	0	15	49.16	30.57	70.36	39.79	0.81	0	Acc
78	1	10	30.91	16.67	31.03	14.36	0.46	0	Inacc
80	1	7	18.45	37.98	30.39	7.59	0.41	1	Acc
81	0	7	18.45	9.95	25.86	15.91	0.86	0	Acc
85	0	1	23.47	34.3	51.73	17.43	0.74	0	Acc
88	0	7	18.45	25.17	18.84	6.33	0.34	1	Inacc
90	0	8	54.51	31.23	75.98	44.75	0.82	0	Acc
95	0	9	26.3	37.37	62.95	25.58	0.97	0	Acc
97	0	16	60.71	27.67	72.36	44.69	0.74	0	Acc
101	0	13	32.8	47.65	67.41	19.76	0.6	0	Acc
106	0	7	18.45	8.41	19.25	10.84	0.59	0	Acc
111	0	5	34.47	14.15	41.45	27.3	0.79	0	Acc
115	0	15	49.16	37.28	26.36	10.92	0.22	1	Inacc
117	1	6	27.12	22.44	9.29	13.15	0.48	1	Acc
122	0	7	18.45	25.42	34.64	9.22	0.5	0	Acc
126	0	14	122.31	11.21	125.01	113.8	0.93	0	Acc
128	1	12	6	12	6	6	1	1	Acc
129	0	15	49.16	38.56	12.17	26.39	0.54	1	Inacc
131	0	5	34.47	39.3	60.17	20.87	0.61	0	Acc

std	Admit	Cluster	D	d0	d1	d	d/D	P.Admit	Accuracy
137	0	6	27.12	23.92	45.35	21.43	0.79	0	Acc
142	0	10	30.91	41.86	56.52	14.66	0.47	0	Acc
143	0	3	44.83	25.27	49.57	24.3	0.54	0	Acc
148	1	8	54.51	19.18	42.79	23.61	0.43	0	Inacc
159	1	2	21.88	29.46	20.55	8.91	0.41	1	Acc
180	0	11	18.29	13.63	26.2	12.57	0.69	0	Acc
192	0	1	23.47	30.44	34	3.56	0.15	0	Acc
214	1	1	23.47	28.04	29.42	1.38	0.06	0	Inacc

Table 7.4: Logistic Regression Result for Training Set

std	A. Admit	P(Y=1)	P. Admit	Accuracy
1	1	0.3313	0	Inaccurate
2	0	0.0287	0	Accurate
3	0	0.0714	0	Accurate
4	0	0.0080	0	Accurate
5	0	0.0074	0	Accurate
6	0	0.5342	1	Inaccurate
7	1	0.7892	1	Accurate
8	1	0.9534	1	Accurate
9	0	0.1349	0	Accurate
10	0	0.0411	0	Accurate
12	0	0.0328	0	Accurate
14	0	0.3254	0	Accurate
15	0	0.0422	0	Accurate
16	1	0.8165	1	Accurate
17	1	0.9985	1	Accurate
18	0	0.7050	1	Inaccurate
19	0	0.0534	0	Accurate
20	0	0.1077	0	Accurate
22	0	0.4227	0	Accurate
23	1	0.0855	0	Inaccurate
24	0	0.4508	0	Accurate
27	1	0.5703	1	Accurate
28	1	0.9885	1	Accurate
30	1	0.9897	1	Accurate
33	1	0.5348	1	Accurate
34	0	0.1137	0	Accurate
35	0	0.0037	0	Accurate
37	0	0.0110	0	Accurate

std	A. Admit	P(Y=1)	P. Admit	Accuracy
38	1	0.9680	1	Accurate
39	1	0.9404	1	Accurate
41	1	0.9602	1	Accurate
44	1	0.0792	0	Inaccurate
45	1	0.9569	1	Accurate
46	0	0.1630	0	Accurate
49	0	0.2840	0	Accurate
53	1	0.5905	1	Accurate
54	0	0.0421	0	Accurate
56	1	0.9946	1	Accurate
58	0	0.1310	0	Accurate
59	1	0.9662	1	Accurate
60	1	0.7602	1	Accurate
61	1	0.8716	1	Accurate
63	1	0.9860	1	Accurate
64	1	0.2719	0	Inaccurate
65	1	0.9776	1	Accurate
66	0	0.5214	1	Inaccurate
67	1	0.0581	0	Inaccurate
68	1	0.4315	0	Inaccurate
69	1	0.5426	1	Accurate
71	1	0.9714	1	Accurate
73	1	0.9373	1	Accurate
74	0	0.0237	0	Accurate
76	1	0.9339	1	Accurate
79	0	0.0815	0	Accurate
82	1	0.9742	1	Accurate
83	0	0.4592	0	Accurate
84	0	0.0321	0	Accurate
86	1	0.9603	1	Accurate
87	0	0.0454	0	Accurate
89	0	0.0304	0	Accurate
91	0	0.2407	0	Accurate
92	1	0.3691	0	Inaccurate
93	0	0.0321	0	Accurate
94	0	0.1667	0	Accurate
96	0	0.0243	0	Accurate
98	0	0.1287	0	Accurate
99	0	0.0698	0	Accurate
100	0	0.0470	0	Accurate
102	0	0.0330	0	Accurate
103	0	0.1197	0	Accurate
104	0	0.8978	1	Inaccurate

std	A. Admit	P(Y=1)	P. Admit	Accuracy
105	1	0.9495	1	Accurate
107	1	0.6110	1	Accurate
108	1	0.8748	1	Accurate
109	0	0.4709	0	Accurate
110	1	0.2057	0	Inaccurate
112	0	0.1642	0	Accurate
113	0	0.2895	0	Accurate
114	1	0.9264	1	Accurate
116	0	0.0312	0	Accurate
118	0	0.0346	0	Accurate
119	0	0.1615	0	Accurate
120	0	0.1636	0	Accurate
121	1	0.9449	1	Accurate
123	0	0.0442	0	Accurate
124	0	0.0720	0	Accurate
125	0	0.0738	0	Accurate
127	0	0.0140	0	Accurate
130	0	0.3185	0	Accurate
132	0	0.7854	1	Inaccurate
133	1	0.8718	1	Accurate
134	1	0.9864	1	Accurate
135	1	0.9805	1	Accurate
136	0	0.0196	0	Accurate
138	0	0.8044	1	Inaccurate
139	0	0.0027	0	Accurate
140	1	0.9900	1	Accurate
141	0	0.0257	0	Accurate
144	1	0.9680	1	Accurate
145	0	0.3254	0	Accurate
146	1	0.9839	1	Accurate
147	0	0.2010	0	Accurate
149	1	0.9679	1	Accurate
150	0	0.3064	0	Accurate
151	1	0.8585	1	Accurate
152	0	0.1216	0	Accurate
153	1	0.9226	1	Accurate
154	1	0.5970	1	Accurate
155	1	0.7054	1	Accurate
156	1	0.9455	1	Accurate
157	1	0.9688	1	Accurate
158	1	0.9959	1	Accurate
160	0	0.1683	0	Accurate
161	1	0.9226	1	Accurate

std	A. Admit	P(Y=1)	P. Admit	Accuracy
162	0	0.4508	0	Accurate
163	1	0.1421	0	Inaccurate
164	0	0.1260	0	Accurate
165	0	0.0201	0	Accurate
166	1	0.1967	0	Inaccurate
167	1	0.0661	0	Inaccurate
168	1	0.2622	0	Inaccurate
169	0	0.5006	1	Inaccurate
170	0	0.0244	0	Accurate
171	0	0.0011	0	Accurate
172	1	0.9679	1	Accurate
173	0	0.0207	0	Accurate
174	0	0.0956	0	Accurate
175	0	0.0092	0	Accurate
176	1	0.8654	1	Accurate
177	1	0.9698	1	Accurate
178	0	0.7703	1	Inaccurate
179	1	0.8587	1	Accurate
181	0	0.0338	0	Accurate
182	0	0.0110	0	Accurate
183	0	0.0662	0	Accurate
184	0	0.0243	0	Accurate
185	0	0.1673	0	Accurate
186	1	0.0978	0	Inaccurate
187	1	0.8481	1	Accurate
188	1	0.6623	1	Accurate
189	1	0.9642	1	Accurate
190	0	0.8208	1	Inaccurate
191	0	0.0133	0	Accurate
193	0	0.0171	0	Accurate
194	1	0.9697	1	Accurate
195	0	0.1880	0	Accurate
196	1	0.9932	1	Accurate
197	0	0.4989	0	Accurate
198	0	0.0040	0	Accurate
199	1	0.5068	1	Accurate
200	1	0.6693	1	Accurate
201	0	0.1708	0	Accurate
202	0	0.2781	0	Accurate
203	1	0.6445	1	Accurate
204	0	0.1357	0	Accurate
205	1	0.9644	1	Accurate
206	0	0.2390	0	Accurate

std	A. Admit	P(Y=1)	P. Admit	Accuracy
207	1	0.8813	1	Accurate
208	0	0.0009	0	Accurate
209	1	0.9994	1	Accurate
210	0	0.0125	0	Accurate
211	0	0.1928	0	Accurate
212	0	0.1135	0	Accurate
213	0	0.0180	0	Accurate
215	0	0.0207	0	Accurate
216	0	0.0303	0	Accurate
217	0	0.0046	0	Accurate
218	0	0.1341	0	Accurate
219	1	0.8034	1	Accurate
220	1	0.8947	1	Accurate
221	0	0.5481	1	Inaccurate
222	1	0.9735	1	Accurate
223	0	0.0078	0	Accurate
224	1	0.9728	1	Accurate
225	0	0.0537	0	Accurate
226	1	0.7850	1	Accurate
227	0	0.1352	0	Accurate
228	0	0.2298	0	Accurate
229	0	0.0422	0	Accurate
230	0	0.0338	0	Accurate
231	0	0.4096	0	Accurate
232	0	0.0458	0	Accurate
233	0	0.0170	0	Accurate
234	0	0.0469	0	Accurate
235	1	0.9888	1	Accurate
236	1	0.6993	1	Accurate
237	0	0.0059	0	Accurate
238	1	0.7414	1	Accurate
239	0	0.1670	0	Accurate
240	0	0.5975	1	Inaccurate
241	0	0.0537	0	Accurate

Table 7.5: Results of Logistic Regression on Test Set

std	Admit	Prob	P. Admit	Accuracy	std	Admit	Prob	P. Admit	Accuracy
11	0	0.1391	0	Accurate	80	1	0.6715	1	Accurate
13	0	0.0001	0	Accurate	81	0	0.0131	0	Accurate
21	1	0.9863	1	Accurate	85	0	0.0071	0	Accurate
25	0	0.0018	0	Accurate	88	0	0.7247	1	Inaccurat
26	1	0.1299	0	Inaccurat	90	0	0.0005	0	Accurate
29	1	0.9942	1	Accurate	95	0	0.0021	0	Accurate
31	1	0.7852	1	Accurate	97	0	0.0003	0	Accurate
32	0	0.0248	0	Accurate	101	0	0.0012	0	Accurate
36	1	0.9931	1	Accurate	106	0	0.0511	0	Accurate
40	0	0.1337	0	Accurate	111	0	0.0271	0	Accurate
42	1	0.0024	0	Inaccurat	115	0	0.8891	1	Inaccurat
43	0	0.4258	0	Accurate	117	1	0.7210	1	Accurate
47	1	0.8236	1	Accurate	122	0	0.0040	0	Accurate
48	1	0.3312	0	Inaccurat	126	0	0.0032	0	Accurate
50	1	0.9072	1	Accurate	128	1	0.9623	1	Accurate
51	0	0.8873	1	Inaccurat	129	0	0.6255	1	Inaccurat
52	1	0.7191	1	Accurate	131	0	0.0412	0	Accurate
55	1	0.9979	1	Accurate	137	0	0.0065	0	Accurate
57	0	0.1298	0	Accurate	142	0	0.0006	0	Accurate
62	0	0.5715	1	Inaccurat	143	0	0.0836	0	Accurate
70	0	0.0558	0	Accurate	148	1	0.1236	0	Inaccurat
72	1	0.7416	1	Accurate	159	1	0.7652	1	Accurate
75	0	0.0149	0	Accurate	180	0	0.1430	0	Accurate
77	0	0.0173	0	Accurate	192	0	0.0482	0	Accurate
78	1	0.1036	0	Inaccurat	214	1	0.1565	0	Inaccurat

Table 7.6: Results of Pattern Recognition Method with Significant Variable On Training Set

std	D	d0	d1	d	d/D	P.admit	Admit	Accuracy
1	28.75	9.5	19.57	10.07	0.35	0	1	Inacc.
2	28.75	5.49	32.88	27.39	0.95	0	0	Acc.
3	28.75	9.4	24.34	14.94	0.52	0	0	Acc.
4	28.75	18.72	47.33	28.61	1	0	0	Acc.
5	28.75	23.11	51.34	28.23	0.98	0	0	Acc.
6	28.75	29.15	11.67	17.48	0.61	1	0	Inacc.

std	D	d0	d1	d	d/D	P.admit	Admit	Accuracy
7	28.75	36.81	13.35	23.46	0.82	1	1	Acc.
8	28.75	28.19	10.42	17.77	0.62	1	1	Acc.
9	28.75	4.12	24.69	20.57	0.72	0	0	Acc.
10	28.75	19.29	45.62	26.33	0.92	0	0	Acc.
12	28.75	7.95	29.22	21.27	0.74	0	0	Acc.
14	28.75	8.14	22.85	14.71	0.51	0	0	Acc.
15	28.75	15.23	42.16	26.93	0.94	0	0	Acc.
16	28.75	27.82	2.81	25.01	0.87	1	1	Acc.
17	28.75	59.31	30.57	28.74	1	1	1	Acc.
18	28.75	29.1	7.62	21.48	0.75	1	0	Inacc.
19	28.75	8.06	26.4	18.34	0.64	0	0	Acc.
20	28.75	9.88	22.16	12.28	0.43	0	0	Acc.
22	28.75	21.29	11.32	9.97	0.35	1	0	Inacc.
23	28.75	16.37	21.51	5.14	0.18	0	1	Inacc.
24	28.75	17.12	12.21	4.91	0.17	1	0	Inacc.
27	28.75	16.82	11.97	4.85	0.17	1	1	Acc.
28	28.75	45.42	16.68	28.74	1	1	1	Acc.
30	28.75	47.14	18.45	28.69	1	1	1	Acc.
33	28.75	22.77	9.11	13.66	0.48	1	1	Acc.
34	28.75	4.17	30.31	26.14	0.91	0	0	Acc.
35	28.75	16.82	44.91	28.09	0.98	0	0	Acc.
37	28.75	22.64	50.52	27.88	0.97	0	0	Acc.
38	28.75	37.06	8.35	28.71	1	1	1	Acc.
39	28.75	39.97	12.51	27.46	0.96	1	1	Acc.
41	28.75	38.59	10.04	28.55	0.99	1	1	Acc.
44	28.75	11.29	23.11	11.82	0.41	0	1	Inacc.
45	28.75	34.85	6.18	28.67	1	1	1	Acc.
46	28.75	14.79	18.4	3.61	0.13	0	0	Acc.
49	28.75	9.62	29.29	19.67	0.68	0	0	Acc.
53	28.75	20.68	8.72	11.96	0.42	1	1	Acc.
54	28.75	8.33	36.56	28.23	0.98	0	0	Acc.
56	28.75	45.06	16.9	28.16	0.98	1	1	Acc.
58	28.75	21.05	19.09	1.96	0.07	1	0	Inacc.
59	28.75	36.24	7.6	28.64	1	1	1	Acc.
60	28.75	27.89	5	22.89	0.8	1	1	Acc.
61	28.75	31.85	4.84	27.01	0.94	1	1	Acc.
63	28.75	44.91	16.21	28.7	1	1	1	Acc.
64	28.75	9.93	18.87	8.94	0.31	0	1	Inacc.
65	28.75	40.12	11.37	28.75	1	1	1	Acc.
66	28.75	16.27	12.49	3.78	0.13	1	0	Inacc.
67	28.75	9.24	36.6	27.36	0.95	0	1	Inacc.
68	28.75	12.08	27.75	15.67	0.55	0	1	Inacc.
69	28.75	15.23	13.76	1.47	0.05	1	1	Acc.

std	D	d0	d1	d	d/D	P.admit	Admit	Accuracy
71	28.75	34.31	7.39	26.92	0.94	1	1	Acc.
73	28.75	33.76	5.03	28.73	1	1	1	Acc.
74	28.75	6.47	33.52	27.05	0.94	0	0	Acc.
76	28.75	32.91	4.16	28.75	1	1	1	Acc.
79	28.75	0.86	28.29	27.43	0.95	0	0	Acc.
82	28.75	45.75	17.8	27.95	0.97	1	1	Acc.
83	28.75	11.71	25.77	14.06	0.49	0	0	Acc.
84	28.75	13.29	41.07	27.78	0.97	0	0	Acc.
86	28.75	33.82	5.8	28.02	0.97	1	1	Acc.
87	28.75	12.52	25.61	13.09	0.46	0	0	Acc.
89	28.75	7.96	36.66	28.7	1	0	0	Acc.
91	28.75	11.93	33.19	21.26	0.74	0	0	Acc.
92	28.75	17.41	13.12	4.29	0.15	1	1	Acc.
93	28.75	13.29	41.07	27.78	0.97	0	0	Acc.
94	28.75	18.83	17.64	1.19	0.04	1	0	Inacc.
96	28.75	7.86	31.31	23.45	0.82	0	0	Acc.
98	28.75	5.58	30.83	25.25	0.88	0	0	Acc.
99	28.75	2.25	30.83	28.58	0.99	0	0	Acc.
100	28.75	13.23	40.3	27.07	0.94	0	0	Acc.
102	28.75	9.51	37.91	28.4	0.99	0	0	Acc.
103	28.75	10.32	35.18	24.86	0.86	0	0	Acc.
104	28.75	24.43	9.09	15.34	0.53	1	0	Inacc.
105	28.75	28.66	8.31	20.35	0.71	1	1	Acc.
107	28.75	19.36	9.46	9.9	0.34	1	1	Acc.
108	28.75	29.65	1.69	27.96	0.97	1	1	Acc.
109	28.75	28.02	12.28	15.74	0.55	1	0	Inacc.
110	28.75	6.84	21.92	15.08	0.52	0	1	Inacc.
112	28.75	15.27	38.29	23.02	0.8	0	0	Acc.
113	28.75	7.65	25.7	18.05	0.63	0	0	Acc.
114	28.75	42.06	15.61	26.45	0.92	1	1	Acc.
116	28.75	5.41	33.86	28.45	0.99	0	0	Acc.
118	28.75	15.27	26.4	11.13	0.39	0	0	Acc.
119	28.75	40.93	62.5	21.57	0.75	0	0	Acc.
120	28.75	3.73	26.28	22.55	0.78	0	0	Acc.
121	28.75	44.27	17.52	26.75	0.93	1	1	Acc.
123	28.75	8.64	27.15	18.51	0.64	0	0	Acc.
124	28.75	19.77	44.81	25.04	0.87	0	0	Acc.
125	28.75	15.65	41.21	25.56	0.89	0	0	Acc.
127	28.75	9.36	34.83	25.47	0.89	0	0	Acc.
130	28.75	13.6	15.75	2.15	0.07	0	0	Acc.
132	28.75	22.41	6.73	15.68	0.55	1	0	Inacc.
133	28.75	26.82	2.5	24.32	0.85	1	1	Acc.
134	28.75	38.55	11.18	27.37	0.95	1	1	Acc.

std	D	d0	d1	d	d/D	P.admit	Admit	Accuracy
135	28.75	39.84	11.17	28.67	1	1	1	Acc.
136	28.75	9.81	38.56	28.75	1	0	0	Acc.
138	28.75	18.57	18.63	0.06	0	0	0	Acc.
139	28.75	18.85	47.09	28.24	0.98	0	0	Acc.
140	28.75	45.11	16.36	28.75	1	1	1	Acc.
141	28.75	6.25	34.45	28.2	0.98	0	0	Acc.
144	28.75	37.06	8.35	28.71	1	1	1	Acc.
145	28.75	8.14	22.85	14.71	0.51	0	0	Acc.
146	28.75	40.42	11.84	28.58	0.99	1	1	Acc.
147	28.75	9.17	20.08	10.91	0.38	0	0	Acc.
149	28.75	42.13	13.89	28.24	0.98	1	1	Acc.
150	28.75	12.66	16.49	3.83	0.13	0	0	Acc.
151	28.75	30.06	3.07	26.99	0.94	1	1	Acc.
152	28.75	23.19	19.82	3.37	0.12	1	0	Inacc.
153	28.75	35.42	7.51	27.91	0.97	1	1	Acc.
154	28.75	24.03	7.91	16.12	0.56	1	1	Acc.
155	28.75	23.35	5.91	17.44	0.61	1	1	Acc.
156	28.75	25.56	19.78	5.78	0.2	1	1	Acc.
157	28.75	39.95	11.33	28.62	1	1	1	Acc.
158	28.75	49.02	20.39	28.63	1	1	1	Acc.
160	28.75	18.38	41.06	22.68	0.79	0	0	Acc.
161	28.75	35.42	7.51	27.91	0.97	1	1	Acc.
162	28.75	17.12	12.21	4.91	0.17	1	0	Inacc.
163	28.75	10.81	34.89	24.08	0.84	0	1	Inacc.
164	28.75	16.61	40.45	23.84	0.83	0	0	Acc.
165	28.75	7.63	35.86	28.23	0.98	0	0	Acc.
166	28.75	6.01	22.77	16.76	0.58	0	1	Inacc.
167	28.75	4.45	27.12	22.67	0.79	0	1	Inacc.
168	28.75	16.97	37.8	20.83	0.72	0	1	Inacc.
169	28.75	13.04	16.53	3.49	0.12	0	0	Acc.
170	28.75	6.82	35.26	28.44	0.99	0	0	Acc.
171	28.75	21.6	47.66	26.06	0.91	0	0	Acc.
172	28.75	42.13	13.89	28.24	0.98	1	1	Acc.
173	28.75	8.99	37.72	28.73	1	0	0	Acc.
174	28.75	18.12	42.63	24.51	0.85	0	0	Acc.
175	28.75	12.66	35.11	22.45	0.78	0	0	Acc.
176	28.75	26.01	3.5	22.51	0.78	1	1	Acc.
177	28.75	33.57	7.25	26.32	0.92	1	1	Acc.
178	28.75	28.82	5.38	23.44	0.82	1	0	Inacc.
179	28.75	25.22	4.5	20.72	0.72	1	1	Acc.
181	28.75	5.64	30.61	24.97	0.87	0	0	Acc.
182	28.75	22.64	50.52	27.88	0.97	0	0	Acc.
183	28.75	3.24	31.69	28.45	0.99	0	0	Acc.

std	D	d0	d1	d	d/D	P.admit	Admit	Accuracy
184	28.75	7.86	31.31	23.45	0.82	0	0	Acc.
185	28.75	5.46	23.3	17.84	0.62	0	0	Acc.
186	28.75	6.76	33.06	26.3	0.91	0	1	Inacc.
187	28.75	26.25	2.61	23.64	0.82	1	1	Acc.
188	28.75	28.81	8.52	20.29	0.71	1	1	Acc.
189	28.75	45.91	18.55	27.36	0.95	1	1	Acc.
190	28.75	25.7	3.06	22.64	0.79	1	0	Inacc.
191	28.75	11.03	39.68	28.65	1	0	0	Acc.
193	28.75	21.35	48.84	27.49	0.96	0	0	Acc.
194	28.75	37.9	9.15	28.75	1	1	1	Acc.
195	28.75	5.25	23.63	18.38	0.64	0	0	Acc.
196	28.75	46.23	17.57	28.66	1	1	1	Acc.
197	28.75	30	12.99	17.01	0.59	1	0	Inacc.
198	28.75	17.38	45.86	28.48	0.99	0	0	Acc.
199	28.75	20.87	9.85	11.02	0.38	1	1	Acc.
200	28.75	20.72	8.05	12.67	0.44	1	1	Acc.
201	28.75	15.79	17.91	2.12	0.07	0	0	Acc.
202	28.75	7.74	26.65	18.91	0.66	0	0	Acc.
203	28.75	15.4	15.74	0.34	0.01	0	1	Inacc.
204	28.75	18.84	42.21	23.37	0.81	0	0	Acc.
205	28.75	31.42	7.65	23.77	0.83	1	1	Acc.
206	28.75	19.79	15.43	4.36	0.15	1	0	Inacc.
207	28.75	22.61	11.94	10.67	0.37	1	1	Acc.
208	28.75	22.78	49.69	26.91	0.94	0	0	Acc.
209	28.75	61.23	32.56	28.67	1	1	1	Acc.
210	28.75	12.32	33.11	20.79	0.72	0	0	Acc.
211	28.75	9.1	31.73	22.63	0.79	0	0	Acc.
212	28.75	3.84	25.26	21.42	0.75	0	0	Acc.
213	28.75	8.89	37.46	28.57	0.99	0	0	Acc.
215	28.75	8.99	37.72	28.73	1	0	0	Acc.
216	28.75	5.36	32.11	26.75	0.93	0	0	Acc.
217	28.75	14.73	41.92	27.19	0.95	0	0	Acc.
218	28.75	25.16	19.77	5.39	0.19	1	0	Inacc.
219	28.75	29.2	4.52	24.68	0.86	1	1	Acc.
220	28.75	35.51	8.69	26.82	0.93	1	1	Acc.
221	28.75	30.14	12.09	18.05	0.63	1	0	Inacc.
222	28.75	42.58	14.15	28.43	0.99	1	1	Acc.
223	28.75	16.49	45.24	28.75	1	0	0	Acc.
224	28.75	39.59	10.85	28.74	1	1	1	Acc.
225	28.75	14.31	40.88	26.57	0.92	0	0	Acc.
226	28.75	27.38	3.8	23.58	0.82	1	1	Acc.
227	28.75	4.86	29.91	25.05	0.87	0	0	Acc.
228	28.75	6.67	22.27	15.6	0.54	0	0	Acc.

std	D	d0	d1	d	d/D	P.admit	Admit	Accuracy
229	28.75	15.23	42.16	26.93	0.94	0	0	Acc.
230	28.75	5.64	30.61	24.97	0.87	0	0	Acc.
231	28.75	14.41	14.45	0.04	0	0	0	Acc.
232	28.75	17.3	43.72	26.42	0.92	0	0	Acc.
233	28.75	8.36	34.13	25.77	0.9	0	0	Acc.
234	28.75	6.36	34.8	28.44	0.99	0	0	Acc.
235	28.75	48.41	19.87	28.54	0.99	1	1	Acc.
236	28.75	25.66	5.79	19.87	0.69	1	1	Acc.
237	28.75	26.98	54.99	28.01	0.97	0	0	Acc.
238	28.75	23.5	35.88	12.38	0.43	0	1	Inacc.
239	28.75	11.8	19.34	7.54	0.26	0	0	Acc.
240	28.75	18.5	10.27	8.23	0.29	1	0	Inacc.
241	28.75	14.31	40.88	26.57	0.92	0	0	Acc.

Table 7.7: Full Results of Proposed Method with Significant Variables for Test Set

std	D	d0	d1	d	d/D	P.Admit	Admit	Accuracy
11	28.75	17.17	13.95	3.22	0.11	1	0	Inacc.
13	28.75	11.42	36.96	25.54	0.89	0	0	Acc.
21	28.75	27.91	7.59	20.32	0.71	1	1	Acc.
25	28.75	2.43	26.37	23.94	0.83	0	0	Acc.
26	28.75	14.45	22.38	7.93	0.28	0	1	Inacc.
29	28.75	31.49	5.26	26.23	0.91	1	1	Acc.
31	28.75	26.89	7.91	18.98	0.66	1	1	Acc.
32	28.75	2.76	26.74	23.98	0.83	0	0	Acc.
36	28.75	37.23	8.79	28.44	0.99	1	1	Acc.
40	28.75	8.23	36.12	27.89	0.97	0	0	Acc.
42	28.75	10.43	36.25	25.82	0.9	0	1	Inacc.
43	28.75	12.32	16.91	4.59	0.16	0	0	Acc.
47	28.75	27.64	1.51	26.13	0.91	1	1	Acc.
48	28.75	6.65	26.02	19.37	0.67	0	1	Inacc.
50	28.75	26.51	5.02	21.49	0.75	1	1	Acc.
51	28.75	14.01	16.92	2.91	0.1	0	0	Acc.
52	28.75	11.22	33.41	22.19	0.77	0	1	Inacc.
55	28.75	45.42	16.81	28.61	1	1	1	Acc.
57	28.75	10.79	20.75	9.96	0.35	0	0	Acc.
62	28.75	5.41	24.04	18.63	0.65	0	0	Acc.
70	28.75	10.51	31.52	21.01	0.73	0	0	Acc.
72	28.75	22.22	8.5	13.72	0.48	1	1	Acc.
75	28.75	21.28	47.53	26.25	0.91	0	0	Acc.
77	28.75	13.91	37.53	23.62	0.82	0	0	Acc.
78	28.75	15.96	19.69	3.73	0.13	0	1	Inacc.

std	D	d0	d1	d	d/D	P.Admit	Admit	Accuracy
80	28.75	18.98	10.91	8.07	0.28	1	1	Acc.
81	28.75	15.13	43.86	28.73	1	0	0	Acc.
85	28.75	12.97	41.5	28.53	0.99	0	0	Acc.
88	28.75	8.93	27.34	18.41	0.64	0	0	Acc.
90	28.75	15.36	39.33	23.97	0.83	0	0	Acc.
95	28.75	17.43	46.07	28.64	1	0	0	Acc.
97	28.75	9.63	38.05	28.42	0.99	0	0	Acc.
101	28.75	26.01	54.07	28.06	0.98	0	0	Acc.
106	28.75	4.27	32.09	27.82	0.97	0	0	Acc.
111	28.75	3.44	30.07	26.63	0.93	0	0	Acc.
115	28.75	29.29	6.29	23	0.8	1	0	Inacc.
117	28.75	20.11	10.52	9.59	0.33	1	1	Acc.
122	28.75	11.36	36.28	24.92	0.87	0	0	Acc.
126	28.75	12.22	39.75	27.53	0.96	0	0	Acc.
128	28.75	35.99	7.27	28.72	1	1	1	Acc.
129	28.75	26.93	2.81	24.12	0.84	1	0	Inacc.
131	28.75	21.35	48.84	27.49	0.96	0	0	Acc.
137	28.75	7.59	36.34	28.75	1	0	0	Acc.
142	28.75	17.81	31.36	13.55	0.47	0	0	Acc.
143	28.75	8.33	36.56	28.23	0.98	0	0	Acc.
148	28.75	3.65	27.17	23.52	0.82	0	1	Inacc.
159	28.75	16.27	12.49	3.78	0.13	1	1	Acc.
180	28.75	14.89	14.33	0.56	0.02	1	0	Inacc.
192	28.75	7.25	34.78	27.53	0.96	0	0	Acc.
214	28.75	7.95	29.22	21.27	0.74	0	1	Inacc.

std	Admit	Cluster	D	d0	d1	d	d/D	P.Admit	Accuracy
11	0	9	26.3	16.96	35.91	18.95	0.72	0	Acc
13	0	14	122.31	19.72	123.56	103.84	0.85	0	Acc
21	1	11	18.29	32.8	17.9	14.9	0.81	1	Acc
25	0	16	60.71	24.83	56.92	32.09	0.53	0	Acc
26	1	2	21.88	25.53	22.96	2.57	0.12	1	Acc
29	1	9	26.3	48.88	32.73	16.15	0.61	1	Acc
31	1	5	34.47	31.57	12.26	19.31	0.56	1	Acc
32	0	14	122.31	22.38	139.01	116.63	0.95	0	Acc
36	1	9	26.3	36.41	17.28	19.13	0.73	1	Acc
40	0	7	18.45	17.4	22.2	4.8	0.26	0	Acc
42	1	1	23.47	32.76	41.66	8.9	0.38	0	Inacc
43	0	5	34.47	32.71	39.46	6.75	0.2	0	Acc
47	1	10	30.91	23.73	12.76	10.97	0.35	1	Acc
48	1	3	44.83	18.83	35.61	16.78	0.37	0	Inacc
50	1	13	32.8	47.55	48.97	1.42	0.04	0	Inacc
51	0	7	18.45	37.65	29.72	7.93	0.43	1	Inacc
52	1	7	18.45	26.57	20.85	5.72	0.31	1	Acc
55	1	9	26.3	48.41	29.67	18.74	0.71	1	Acc
57	0	5	34.47	12.69	31.93	19.24	0.56	0	Acc
62	0	8	54.51	23.86	44.91	21.05	0.39	0	Acc
70	0	3	44.83	18.22	45.22	27	0.6	0	Acc
72	1	11	18.29	69.27	70.68	1.41	0.08	0	Inacc
75	0	4	118.99	26.3	99.72	73.42	0.62	0	Acc
77	0	15	49.16	30.57	70.36	39.79	0.81	0	Acc
78	1	10	30.91	16.67	31.03	14.36	0.46	0	Inacc
80	1	7	18.45	37.98	30.39	7.59	0.41	1	Acc
81	0	7	18.45	9.95	25.86	15.91	0.86	0	Acc
85	0	1	23.47	34.3	51.73	17.43	0.74	0	Acc
88	0	7	18.45	25.17	18.84	6.33	0.34	1	Inacc
90	0	8	54.51	31.23	75.98	44.75	0.82	0	Acc
95	0	9	26.3	37.37	62.95	25.58	0.97	0	Acc
97	0	16	60.71	27.67	72.36	44.69	0.74	0	Acc
101	0	13	32.8	47.65	67.41	19.76	0.6	0	Acc
106	0	7	18.45	8.41	19.25	10.84	0.59	0	Acc
111	0	5	34.47	14.15	41.45	27.3	0.79	0	Acc
115	0	15	49.16	37.28	26.36	10.92	0.22	1	Inacc
117	1	6	27.12	22.44	9.29	13.15	0.48	1	Acc
122	0	7	18.45	25.42	34.64	9.22	0.5	0	Acc
126	0	14	122.31	11.21	125.01	113.8	0.93	0	Acc
128	1	12	6	12	6	6	1	1	Acc
129	0	15	49.16	38.56	12.17	26.39	0.54	1	Inacc
131	0	5	34.47	39.3	60.17	20.87	0.61	0	Acc
137	0	6	27.12	23.92	45.35	21.43	0.79	0	Acc

std	Admit	Cluster	D	d0	d1	d	d/D	P.Admit	Accuracy
142	0	10	30.91	41.86	56.52	14.66	0.47	0	Acc
143	0	3	44.83	25.27	49.57	24.3	0.54	0	Acc
148	1	8	54.51	19.18	42.79	23.61	0.43	0	Inacc
159	1	2	21.88	29.46	20.55	8.91	0.41	1	Acc
180	0	11	18.29	13.63	26.2	12.57	0.69	0	Acc
192	0	1	23.47	30.44	34	3.56	0.15	0	Acc
214	1	1	23.47	28.04	29.42	1.38	0.06	0	Inacc