

Mining Students' Academic Performance using Decision Tree Algorithms

Adesean Barnabas Adeyemo¹ Gbemisola Kuye²
University of Ibadan
Nigeria

Abstract

In this paper we present an evaluation of the factors that contribute to the academic performance of students admitted into the university. The variables of interest are the entry qualification and admission mode and how these factors affect the academic performance of the students. The evaluation was carried out using computer software that implements the decision tree algorithm. Student admission data obtained from a case study department in one of the university's in Nigeria was used. The results indicate that the observed performance of student whose admission into the case study department is through the University Matriculation Examinations (UME) depends more on their respective Senior School Certificate Examination (SSCE) performance than their entry scores in the UME examination used as the basis for their admission.

Keywords: Data Mining, Decision Tree Algorithms, Students' Academic Performance.

Introduction

Data mining is the exploration and analysis, by automatic or semi automatic means of large quantities of data in order to discover meaningful patterns and rules (Berry & Linoff, 1997). It is also called data or knowledge discovery in databases. Data mining tools predict future trends and behaviors, allowing decision makers make proactive, knowledge-driven decisions. Data mining tools can answer questions that traditionally were too time-consuming to resolve.

In the recent past, it has been observed that the performance of students in the first two years of undergraduate studies do not match their performance in the Joint Admissions and Matriculation Boards (JAMB) University Matriculation Examination (UME) which is used as the basis for the admission of the these students into the university. The observed trend indicates that some of the students, who scored very high marks at UME selection exams, end up with very poor results in their first or second year degree exams. Furthermore, students who inadvertently gain admission into the university with average UME scores excel in their studies. It is a primary objective of this work to examine how data mining can help education managers. Data mining can assist in the on-going school reform agenda. If the recommended post JAMB tests proposed and already adopted by Nigerian Universities for the 2005/2006 admission exercise will be successful and sustainable then educational administrators, counselors, and teachers need to incorporate the predictive potential of data mining into the administration of the schools. The

need to score a candidates SSCE result according to a chosen scale and then combine it with the standard test (UME) scores to determine the candidates eligibility for admission into the university can be justified using data mining techniques.

In this study the knowledge discovery potential of data mining has been used to evaluate the academic performance of undergraduate students admitted into the first year (100 level) class of the case study department. In this case, the goal of our knowledge discovery process is to identify the patterns that influence and or affect the performance of undergraduate students after their first year degree examinations. The case data for the analysis was obtained from one of the academic departments of a university in the southwestern part of Nigeria. A data-mining model was developed to test some hypotheses that form the basis for the analysis of students' performance after their first year degree examinations.

Data Mining with Decision Tree Algorithms

Data mining takes advantage of advances in the fields of artificial intelligence (AI) and statistics. Regression and classification, the oldest and most well known statistical techniques, are utilized in data mining. New techniques and new approaches to older algorithms, such as artificial neural networks, decision trees and discriminate analysis, are applied to data mining as well.

Decision trees are tree-shaped structures that represent a series of rules that lead to sets of decisions. These decisions generate rules for the classification of a dataset. Decision trees present a clear, logical model that can be understood easily by people who are not mathematically inclined. They can handle both continuous and categorical variables. Categorical variables such as gender, race, religion, marital status and geographic region are difficult to model using numerically oriented techniques such as regression. In contrast, decision trees handle categorical variables easily.

Decision trees can perform classification as well as regression. The predicted value from a decision tree is not simply a numerical value but can be a predicted category such as male/female, malignant/benign, frequent buyer/occasional buyer, first class, second class upper, second class lower and third class. Decision trees automatically handle interactions between variables and are highly accurate. There are basically two decision tree methods; these are Classification and Regression Trees (CART) and Chi-Square Automatic Interaction Detection (Statsoft, 2003).

Methodology

The process of data mining consists of three steps thus: Data preparation, Model Building and Deployment. During data collection, the relevant data is gathered. Once the data has been assembled, its quality must be verified. Incomplete (lacking certain attributes of interest, or containing only aggregate data), noisy (containing errors, or outlier values that deviate from expected), and inconsistent (for example, discrepancies in the department codes used to categorize items) data are common. Data cleaning routines attempt to clean the data by filling in missing values; smoothing noisy data, identifying or removing outliers, and resolving inconsistencies. Finally, the cleaned data are transformed into a format suitable for data mining. The model building stage involves considering various models, evaluating their predictive

performance and choosing the one that is most useful in solving the problem. In the deployment stage new data sets are applied to the model selected in the model building stage to generate predictions or estimates of the expected outcome (Statsoft, 2003).

Data Preparation

The data preparation processes is under the gathering/collection of data, and data cleaning through data transformation. The data gathering process involves the collection of the raw data about students, which include the UME score, SSCE results which were then scored (both of which constitute the entry requirements into the University) and other information like age, gender and mode of entry to the degree program.

Data can be inconsistency, incomplete or noisy. This may be as a result of a number of factors ranging from data entry or transmission problem, discrepancy in the naming convention, duplicated records or incomplete data. All these affect the analysis. The data used for the analysis was entered into an Excel spreadsheet file. The actual matriculation number of the students were not used: The data format is shown in table 1.

Table 1. Data format

S/no	Field Name	Field Type
1	Serial number	Continuous
2	Matriculation number	Continuous
3	Age	Continuous
4	Sex	Categorical
5	No of sittings	Continuous
6	SSCE result	Continuous
7	Total SSCE score	Continuous
8	UME score	Continuous
9	100 level CGPA	Continuous/Categorical
10	Manner of admission	Categorical

The software program chosen for the analysis is DTREG. DTREG provides state-of-the-art modeling methods including Single Tree, Tree Boost, Decision Tree Forests, Support Vector Machines (SVM) and Logistic Regression.

Building and Pruning Decision Trees using DTREG

The steps taken by DTREG (Sherrod, 2006) to build and prune a tree can be summarized as follows:

1. To build the tree
 - a. i. Examine each node and find the best possible split
 - ii. Examine each predictor variable.
 - iii. Examine each possible split on each predictor
 - b. Create two child nodes
 - c. Determine which child node each row goes into. This may involve using surrogate splitters

- d. Continue the process until a stopping criterion (for example, minimum node size) is reached.
2. To prune the tree
 - a. Build a set of cross-validation trees
 - b. Compute the cross validated misclassification cost for each possible tree size
 - c. Prune the primary tree to the optimal size

The process used to split a node is the same whether the node is the root node with all of the rows or a child node many levels deep in the tree. The only difference is the set of rows in the node being split. DTREG tries each predictor variable to see how well it can divide the node into two groups.

If the predictor is continuous, a trial split is made between each discrete value (category) of the variable. The number of splits evaluated is equal to the number of discrete values of the predictor variable less one. If there are more actual discrete values than this parameter setting, values are grouped together into value ranges. This process is repeated by moving the split point across all possible division points. The best improvement found from any split point is saved as the best possible split for that predictor variable in this node. The process is then repeated for each other predictor variable. The best split found for any predictor variable is used to perform the actual split on the node. The next best five splits are saved as “competitor splits” for the node.

If the predictor variable is categorical and the target variable is continuous, the categories of the predictor variable are sorted so that the mean value of the target variable for the rows having each category of the predictor are increasing. The splitting process then tries each split point between each category of the predictor. This is very similar to the process used for continuous predictor variables except the categories are arranged by values of the target variable rather than by values of the predictor variable. The number of splits evaluated is equal to the number of categories of the predictor variable less one.

If both the target variable and the predictor variable are categorical, to perform an exhaustive search DTREG must evaluate a potential split for every possible combination of categories of the predictor variable. The number of splits is equal to $2^k - 1$ where k is the number of categories of the predictor variable. Because of the exponential growth that results, the computation time to do an exhaustive search becomes prohibitive when there are more than about 12 predictor categories. In this case, DTREG uses a clustering technique to group the target categories.

Various criteria have been proposed for evaluating splits, but they all have the same basic goal which is to favor homogeneity within each child node and heterogeneity between the child nodes. The heterogeneity (or dispersion) of target categories within a node is called the “node impurity”. The goal of splitting is to produce child nodes with minimum impurity. The impurity of every node is calculated by examining the distribution of categories of the target variable for the rows in the group. A “pure” node, where all rows have the same value of the target variable, has an impurity value of zero. When a potential split is evaluated, the probability-weighted average of the impurities of the two child nodes is subtracted from the impurity of the parent node. This reduction in impurity is called the improvement of the split. The split with the greatest improvement is the one used.

DTREG provides two methods for evaluating the quality of splits when building classification trees. These are Gini method and Entropy method. The minimum variance within nodes is the only method provided when building regression trees. The minimum variance/least squares criteria is essentially the same criteria used by traditional, numeric regression analysis (that is, line and function fitting). Experience has shown that the splitting criterion is not very important, and Gini and Entropy yield trees that are very similar although the Gini method is considered slightly better than the Entropy method.

When a decision tree is used to predict values of the target variable, rows are run through the tree down to the point where they reach a terminal node. The category assigned to the terminal node is the predicted value for the row being evaluated. For regression trees built with a continuous target variable, the value assigned to a node is simply the average value of the target variable for all rows that end up in the node weighted by the row weights. For classification trees built with a categorical target variable, the determination of what category to assign to a node is the category that minimizes the misclassification cost for the rows in the node.

Several criteria are used to limit how large the tree constructed by DTREG is. Once a tree is built, a pruning method is used to reduce its size to the optimal number of nodes. There are two reasons why it is desirable to generate trees of optimal size. First, if a situation can be described and explained equally well by two descriptions, the description that is simpler and more concise is generally preferred. The same is true with decision trees: if two trees provide equivalent predictive accuracy, the simpler tree is preferred because it is easier to understand and faster to use for making predictions. Second, and more importantly, smaller trees may provide greater predictive accuracy for unseen data than larger trees. Another way to understand why large trees can be inferior to smaller trees is that the large trees fit and model minor "noise" in the data, whereas smaller trees model only the significant data factors. The primary goal of the pruning process is to generate the optimal size tree that can be generalized to other data beyond the learning dataset.

The V-fold cross validation method is used by DTREG to determine the optimal tree size. V-fold cross validation is a technique for performing independent tree size tests without requiring separate test datasets and without reducing the data used to build the tree (Sherrod, 2006).

Building a Decision Tree Model

A set of data values called the learning or training dataset is needed to build a decision tree model. The training data set is used by the decision tree software to learn how the value of a target variable is related to the values of predictor variables. A decision tree can be used to predict the values of the target variable based on values of the predictor variables. The decision tree software will generate a regression model or a classification model depending on whether the target variable is continuous or categorical. If the target variable is continuous, a regression model is generated. If the target variable is categorical, then a classification model is generated. To predict the value (category) of the target variable using a classification tree, the values of the predictor variables are used to move through the tree until a terminal (leaf) node is reached. Then the category that is shown for that node is used for the prediction. The decision tree software can be used to run a set of data rows through a generated decision tree in order to generate a new data

file showing the predicted value of the target variable and other information for each row.

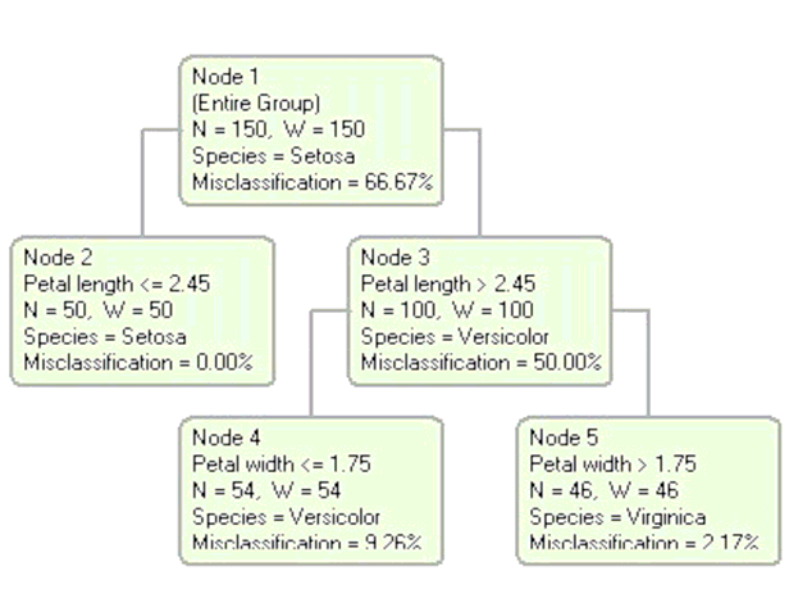


Figure 1. Decision tree nodes (Sherrod, 2006)

In the decision tree shown in figure 1, the rectangular boxes shown in the tree are called “nodes”. Each node represents a set of records (rows) from the original dataset. Nodes that have child nodes (nodes 1 and 3 in figure 1) are called “interior” nodes. Nodes that do not have child nodes (nodes 2, 4 and 5 in figure 1) are called “terminal” or “leaf” nodes. The topmost node (node 1 in figure 1) is called the “root” node. The root node represents all the rows in the dataset. In the top of the node box is the node number. The “N = nn” line shows how many rows (cases) fall in the node. The “W = nn” line shows the sum of the weights of the rows in the node (Sherrod, 2006).

Results and Discussion

The data mining technique used was a simple classification technique. It is a well-trained predictive, classifier model that can be used to analyze and predict the performance of a student in the first year of a degree program. Two data sets were used. The first used categorical data as the target data set variable. The classes created are those used within the university system namely: First class, 2nd class upper Division, 2nd class Lower Division, Third class, Pass and Withdrawal. The second used a regression (variance) model with continuous rather than categorical data. Here the raw CGPA scores after the first year exams are used. For the classification models DTREG implements the Entropy, Gini and Misclassification Cost tree splitting methods. The Entropy and Variance (regression) algorithms were used for the analysis of the data set. The summary of the results obtained is presented in Table 2.

Table 2. Summary of Results

	Entropy	Variance (Regression)
Type of Tree	Single Tree	Single Tree
Tree Size (Optimal)	7	5
Number of terminal nodes	4	3
Relative Importance Variable	UME: 15 SSCE: 100	UME: 50 SSCE: 100
Standard Error	0.0396	0.0765
Relative Error	0.7873	0.9546
Max Splitting Level	10	10
No of Predictor Variable	4	4
Target Variable	CGPA 100L	CGPA 100L

The classification tree obtained using the Entropy method is shown in Figure 2. The chart of Relative Importance of Variables generated using the Entropy method is shown in Figure 3. The classification tree obtained using the Variance (regression) method is shown in Figure 4. The chart of Relative Importance of Variables generated using the Variance (regression) method is shown in Figure 5.

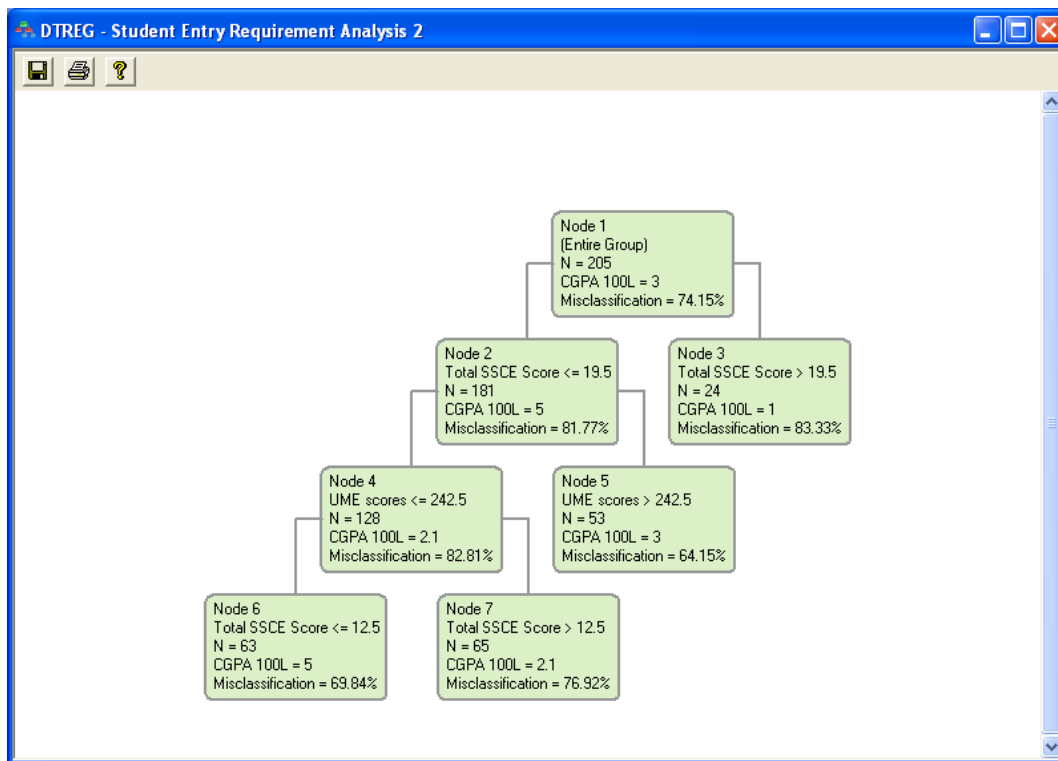


Figure 2. Classification Tree generated by Entropy method

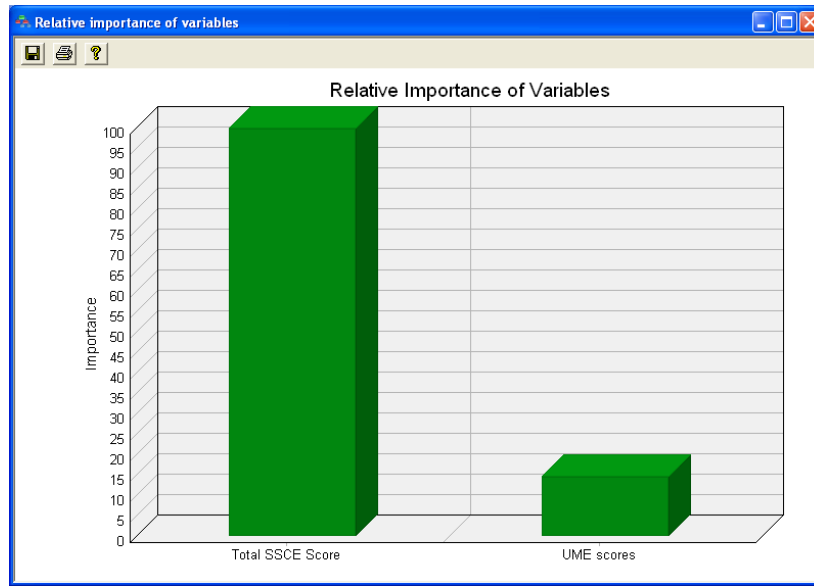


Figure 3. Chart of Relative Importance of Variables generated by Entropy method

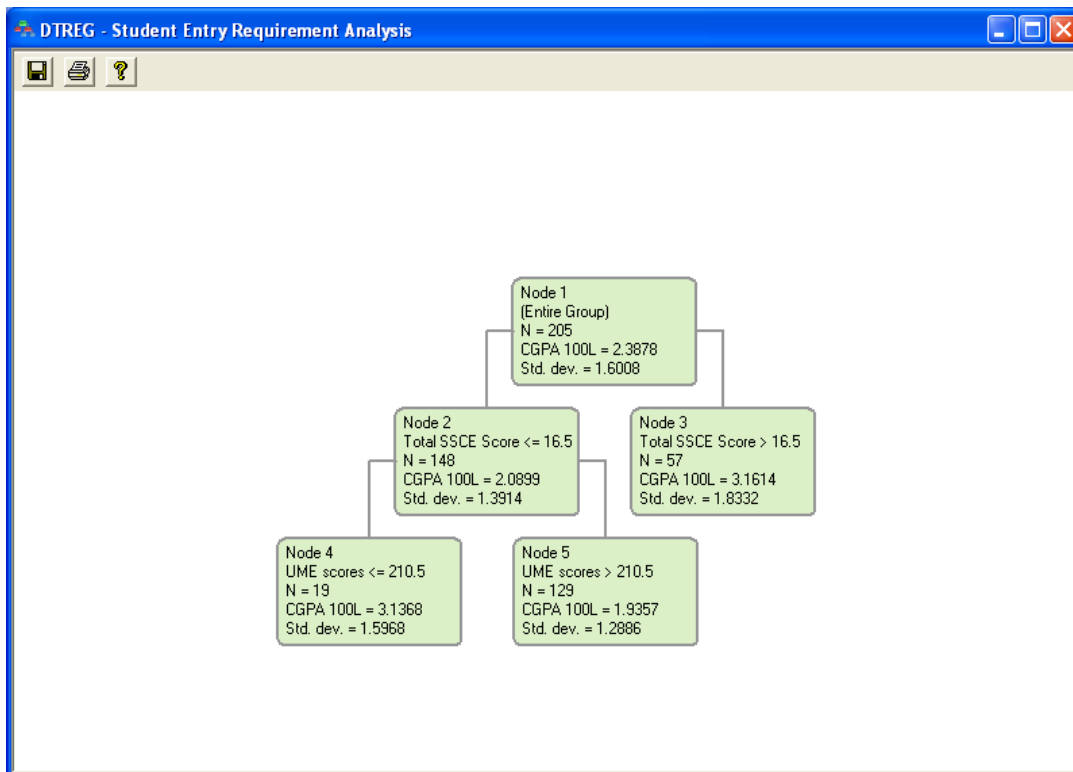


Figure 4. Classification Tree generated by Variance (regression) method

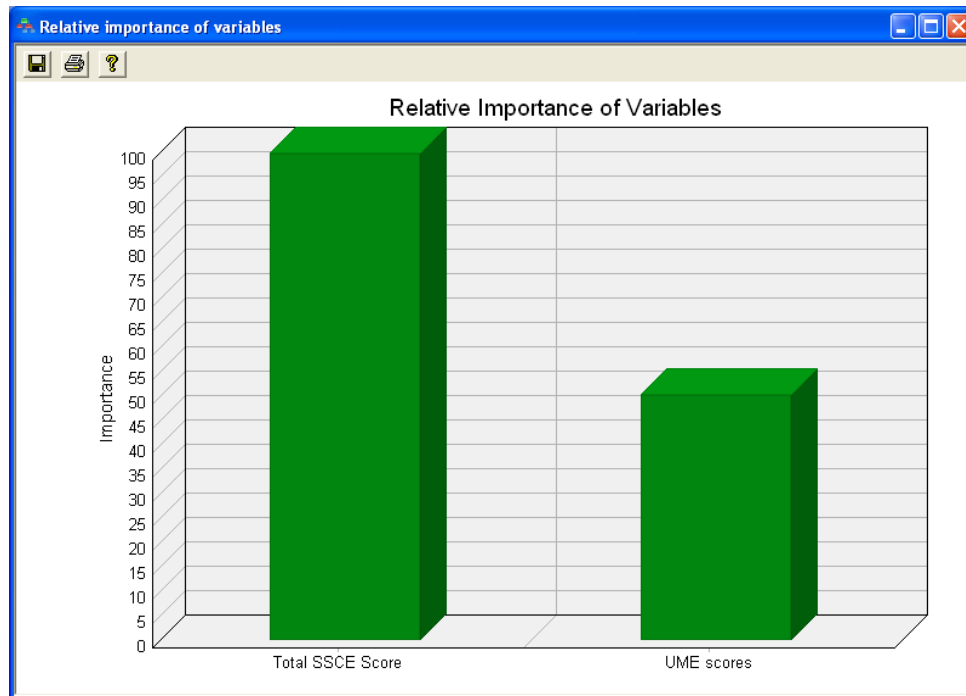


Figure 5. Chart of Relative Importance of Variables generated by Variance (regression) method

The Entropy and Gini models used categorical data in the target variable field, while the Variance model used continuous data in the target variable field. The Entropy model that used categorical data in the target variable field gave results similar to that obtained with the Variance (Regression) model that used continuous data in its target variable field. An examination of the Relative Importance of Variables analysis generated by the software shows that both models placed more emphasis on the SSCE score variable instead of the UME score variable in the analysis the performance of students taking the 100 Level degree examinations.

Conclusion

The result obtained from the analysis of students academic performance has showed that the decision tree data mining software was able to model the data set used. It can be concluded from the result of the analysis that the current attempt by stakeholders in the education sector to de-emphasize the UME scores and incorporate SSCE results of students as a variable in the selection process is a step in the right direction. It is expected that this preliminary work will serve as the basis of a more detailed and elaborate study involving more data collected from different academic units across the country.

References

- Berry, M. J. A., & Linoff, G. (1997). *Data Mining Techniques: For Marketing, Sales, and Customer Support*. Wiley Computer Publishing.
- Sherrod, P. H. (2006). *DTREG: Classification and Regression Trees and Support Vector Machines for Predictive Modeling and Forecasting*. Retrieved from <http://www.dtreg.com/DTREG.pdf>
- StatSoft, Inc., (2003). *Electronic Statistics Textbook*. Tulsa, OK: StatSoft. Retrieved from <http://www.statsoftinc.com/textbook/stathome.html>

-
- ¹ Mr. Adesesan Barnabas Adeyemo is a Lecturer at the Department of Computer Sciences, University of Ibadan, Ibadan, Oyo State, Nigeria. He can be reached at Number 17 Ikale Road, Ijapo Estate, Akure. Ondo State, Nigeria. Email: sesan_adeyemo@yahoo.com; Phone: +(234) 08052107367
- ² Ms. Gbemi Kuye is a graduate of the Computer Science Department of the University of Ibadan, Oyo State, Nigeria. She can be reached at: 21 Aina Street Ojodu, Begger, Lagos, Nigeria. Email: gbemi_k2001@yahoo.com; Phone: +(234) 8023386030