

Article Citation Format

Akinrotimi, A.O. & Aremu, D.R. (2018): Student Performance Prediction Using Randomtree and C4.5 Algorithm
Journal of Digital Innovations & Contemp Res. In Sc., Eng & Tech. Vol. 6, No. 3. Pp 23-34

Article Progress Time Stamps

Article Type: Research Article
Manuscript Received: 11th June, 2018
Review Type: Blind
Final Acceptance:: 25th August, 2018
DOI Prefix: 10.22624

Student Performance Prediction Using Randomtree and C4.5 Algorithm

Akinrotimi, Akinyemi Omololu & Aremu, Dayo Reuben

Department of Computer Science

Faculty of Communication and Information Sciences

University of Ilorin, Ilorin Nigeria

P.M.B 1515, Ilorin, Kwara State, Nigeria.

E-mail: timiakin2011@yahoo.com, draremu2006@gmail.com

ABSTRACT

Machine learning has been applied to various domains such as pattern recognition, image recognition, fraud detection, medical diagnosis, banking, bioinformatics, commodity trading, computer games and various control applications. Recently, there has been a paradigm shift towards the educational sector in evaluating higher education tasks. The focus of this work is on identifying a suitable algorithm for predicting tertiary students' academic performance, based on their family background factors and previous academic achievements. A university case study with 204 enrolment records of students admitted into computer science program in Kwara State University, Nigeria between 2013 and 2015 was used. The students' first year academic performance was measured by Cumulative Grade Point Average (CGPA), at the end of the first session and the previous academic achievement was measured by SSCE grade score and UME score. The MATLAB programming language was used to carry out the experimental research with emphasis on C4.5 and Randomtree algorithm. These algorithms were compared, using the holdout method, considering, accuracy level, confusion matrices and CPU time to determine the optimal model. Also, a framework of a predictive system based on the rules generated from the optimal model was designed.

Keywords: Decision Tree, Machine Learning, Encryption, Prediction, Data mining.

1. INTRODUCTION

There is an ever growing need to improve the performance of students by various institutions, but students' academic performance hinges on diverse factors like socio-economic, personal, psychological, family background and other environmental variables. There has been progression over the years towards the use of various means to improve and then predict students' performance [1]. Machine learning is a method used to devise complex models and algorithms that are useful for prediction. As such, this method is often used in predictive analysis. In using this technique, the analytical models built, allow researchers to produce reliable, results and gain deeper insight into historical relationships between different data components. This technique has been successfully applied in areas such as banking, bioinformatics, computer games, medical diagnosis and various control applications [4].

In higher learning institutions, diverse problems which keep the institutions from achieving their quality objectives are encountered. Some of the problems stem from knowledge gap. Knowledge gap is the lack of significant knowledge in educational main processes such as counselling, planning, registration, evaluation and marketing. For example, many learning institutions do not have access to the necessary information to counsel students. Therefore they are not able to give suitable recommendation to the students. The hidden patterns, associations and anomalies that are discovered by machine learning techniques can help bridge this knowledge gap in learning institutions. This knowledge would enable the higher learning institutions in making better decisions, having more advanced planning in directing students, predicting individual behaviours with higher accuracy [5].

The differential students' performance in tertiary institutions is a source of great concern and research interest to the higher education managements, government, parents and other stakeholders because of the importance of education to national development. As such, academic institutions are increasingly required to monitor both their performance and that of their students [6]. This gives rise to a need to extract useful information from the available students' large datasets to inform academic policies on how best to improve student retention rates, allocate teaching and support resources, or create intervention strategies to mitigate factors that affect student performance and adversely maximizing the potential of students. This paper is focused on suggesting a viable algorithm for predicting students' academic performance.

2. LITERATURE REVIEW

In [1] the capabilities of data mining techniques in context of higher education are justified. Here, a decision tree is used to evaluate students' performance at the end of semester. Variables considered are previous semester marks, class test grade, seminar performance, assignment, general proficiency, attendance, lab work, and end of semester remarks. The classification task used, is able to predict the student division on the basis of the previous database. This helps to reduce failure ratio because early identification will enable appropriate action. In another study [2] the authors focus on using the Bayesian classification algorithm to predict students' academic performance in BCA department of Indian University. Variables considered are Sex, Category, medium of teaching, student food habit, other habits, living condition, accommodation, family size, family status, family annual income, grade in senior secondary school, students' college type, father's qualification, mother's qualification, father's occupation, mother's occupation and grade obtained in BCA.

The Naïve Bayes classification algorithm was used as a technique, to design the student performance prediction model. It is found that grade in senior secondary school, living condition, medium of teaching, mother's qualification, students' habit, family income and family status were high potential variables for student performance. The investigation shows that other factors outside students' effort have significant influence over students' performance. In [3] generating predictive models are used for student retention management using decision tree algorithms (ID3, C4.5 and ADT) in WEKA. Study shows that intervention programs can have significant effects on retention, especially for the first year. Machine learning algorithms were applied to analyze and extract information from existing student data to establish predictive models. The predictive models are then used to identify among new incoming first year students, those who are most likely to benefit from the support of the student retention program. The empirical results show that short but accurate prediction list for the student retention purpose can be produced by applying the predictive models to the records of incoming new students. The study identifies students which needed special attention to reduce drop-out rate. [8] used K-Means clustering algorithm, to predict the pass percentage and fail percentage of the overall students that appeared for a particular examination in an institution.

Their experimental analysis revealed that comparison between Naviebayes algorithm and decision stump tree technique shows that the Naviebayes techniques produce accurate result in making the desired prediction, than the other. In [9], the K-means algorithm was used to get the cluster of students from different aspects, get the specific differences between students of different clusters, and based on this, obtained association between the sex, achievement, consumption and other behaviours of students, so as to improve the management of a vast number of students. Also in [10] the k-means clustering algorithm was applied in analyzing the students result data and predicting the students' performance. The results showed that the proper data mining application on student's performance can be efficiently used for hidden knowledge / information retrieval from a vast data, which can in turn, be used for the process of decision making by the management of an educational institution.

In [11], the researchers proposed a technique based on decision tree of data mining techniques and k-means partitioning of clustering methods for enhancing the quality of educational system by analyzing and improving student's performance. [12], employed the Decision tree and K-means data mining algorithms to model an approach to predict the performance of students in advance, so as to devise mechanisms of alleviating student dropout rates and improve on performance. [13], proposed a method of improving and defining clusters automatically by assigning required clusters to un-clustered points, and by enhancing the cluster technique, using the normalization procedure, thereby reducing clustering time and improving predictions and [14] used various data mining approaches like Clustering, classification and regression to predict, in advance, students' performance in examinations, so that necessary measures can be taken to help improve their performance. In their work, a hybrid approach of Enhanced K-strange points clustering algorithm and NaïveBayes classification algorithm is presented, implemented and compared with the K-means clustering algorithm and Decision tree. The multiple linear regression was then used to predict student performance.

3. PROPOSED SYSTEM

The techniques and procedures will take a stepwise approach which is as follows:

3.1 Data Set Acquisition

The system will be provided with a training dataset consisting of information about students admitted to the first year. The dataset used was collected from the Students' of the Department of Computer Science of Kwara State University, Malete, Nigeria. The dataset consists of sex of the students, Age of the students, student entry grades in secondary school (which is the Ordinary level results), entrance examination scores and the grade obtained at graduation (B.Sc) for all Computer Science graduates.

3.2 Data set Pre-processing and Normalization.

Getting rid of errors and outliers that may be present in the data are parts of the pre-processing task that was done to make the data suitable for modelling. This research work ensured that every inconsistent set of the data was filtered out, thereby enhancing a smooth operation on the dataset for better result optimization.

3.3 Feature Selection.

To select features with high discriminate power, the information gain filter selection technique was used, to optimize the features with high predictive information, thereby selecting the best sub-set of the dataset.

3.4 Feature Classification.

In this research work, students' final GPA was predicted, based on their grades on mandatory courses. This gave us an insight on how many specific courses affect the students' graduation grades. We chose to use classification because the objective of classification techniques in educational data mining is to identify what important factors that contributes to categorizing students' final grades. Decision trees are the most popular classification technique in data mining. They represent the group of classification rules in a tree form, and they have several advantages over other techniques as stated in (Vandamme, 2007). In this research work, the Randomtree and C4.5 algorithm is used to classify the selected features obtained from the feature selection phase.

3.5 Model Comparison.

To determine the more effective model between the RandomTree and C4.5 algorithms, a statistical comparative approach is used, to analyse the best as constrained to the given data set, with measures like the classification accuracy, mean square error and computational timing.

3.6 Recommender System

An application interface is also developed with the MATLAB programming language in order to create a user friendly interface that will cater to unfamiliar dataset that will be analysed on individual basis, by the system.

3.7 System Flow Chart

The system algorithm flowchart is shown below.

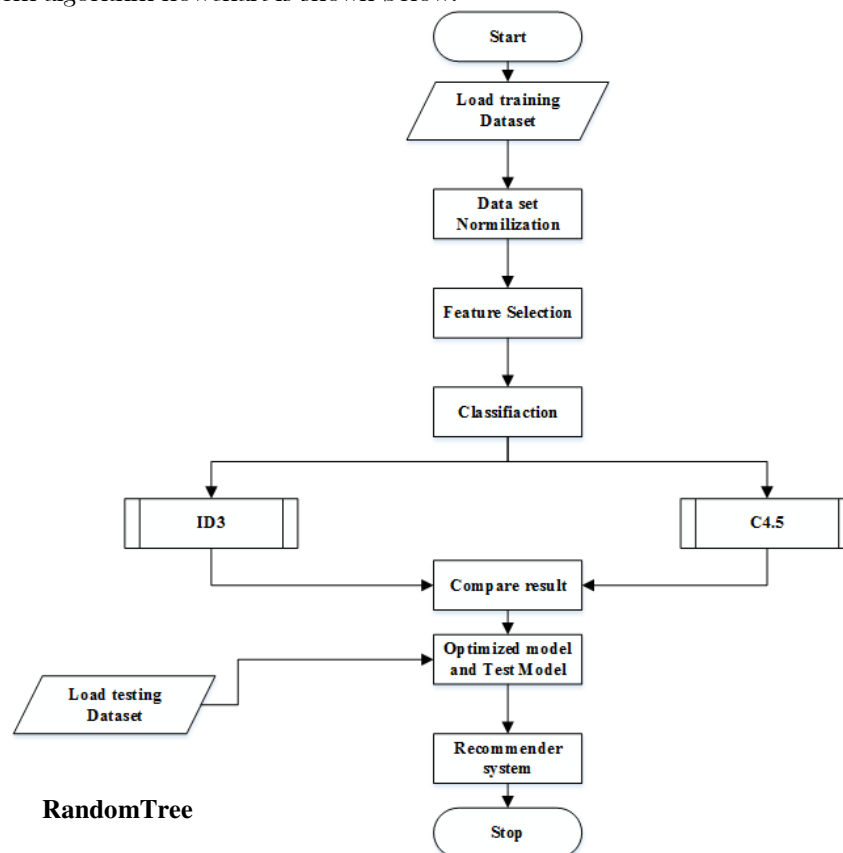


Figure 1.0: System Flow Chart

4. IMPLEMENTATION AND RESULTS

The implemented model and developed application uses the two way test filter feature selection technique, which selects attributes with high predictive power, and also uses the Randomtree and C4.5 decision for classification of the selected attributes into their appropriate class group.

4.1 Feature selection

The two way test is used to determine the significance between the predictors and response variable so as to determine the level of descriptive strength between the dependent variables and the independent variables.

The selected Attributes are as follows:-

1. Jamb score
2. Sponsor.
3. Fathers Education
4. Fathers Occupation
5. Mothers Occupation
6. English
7. Biology
8. Physics
9. Economics
10. Family Type
11. Chemistry
12. Maths
13. Marital status of parents

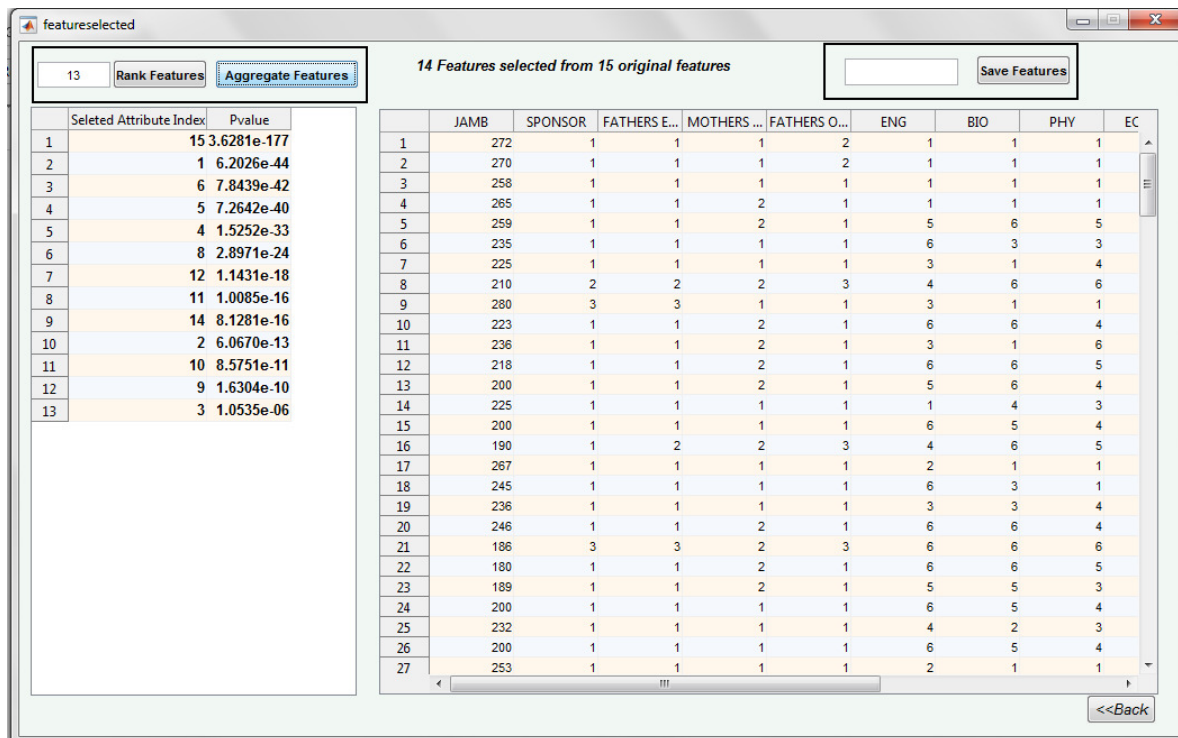


Figure 4.1: Collation of the selected features

4.2 Training of the Support Vector Machine

The dataset was divided into training and testing at a percentage of 75%:25% in the ratio of Train and testing. The train dataset was used to train the support vector machine so as to create an experimental knowledge.

4.2.1 C4.5 MODELLING AND TESTING

The selected data was passed into the C4.5 algorithm which was in turn trained to extract knowledge discovery from the dataset, after the training had been done, the held out data for testing was used to determine how well our model has memorized experimental knowledge. The result obtained is presented below.

4.2.1.1 Confusion Matrix

The confusion matrix shows the percentage classification of class label. Class 1 represent first class, Class 2 second class upper, Class 3 second class lower, and Class 4 third class. The figures 4.2 and 4.3 below show the percentage accuracy of the classification and misclassification rate of each of these classes.

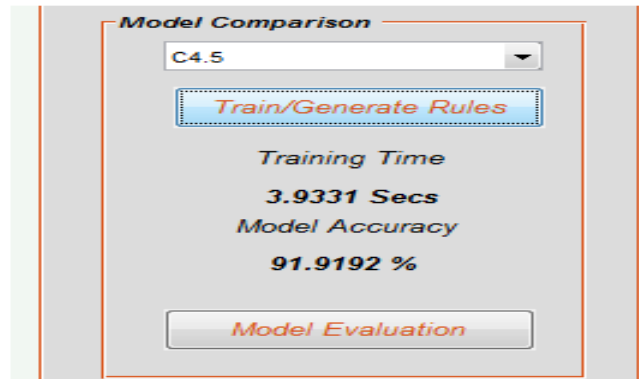


Figure 4.2: C4.5 Classification Result

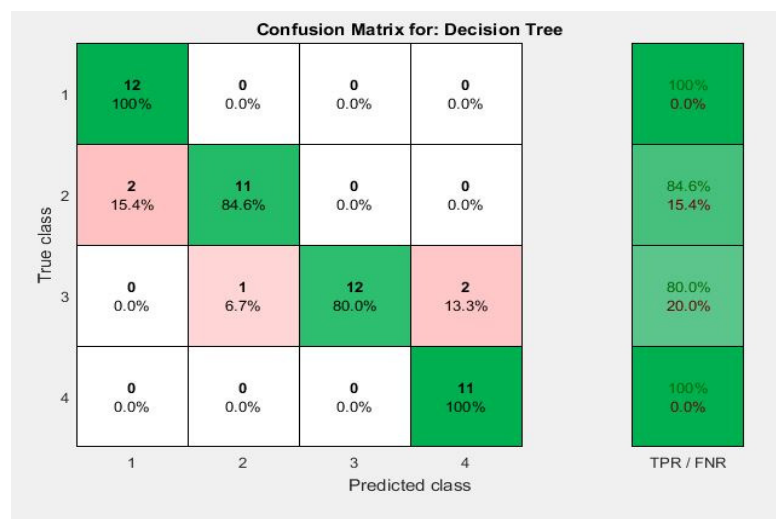


Figure 4.3: C4.5 Classification and misclassification accuracy

4.2.1.2 Tree Prunning

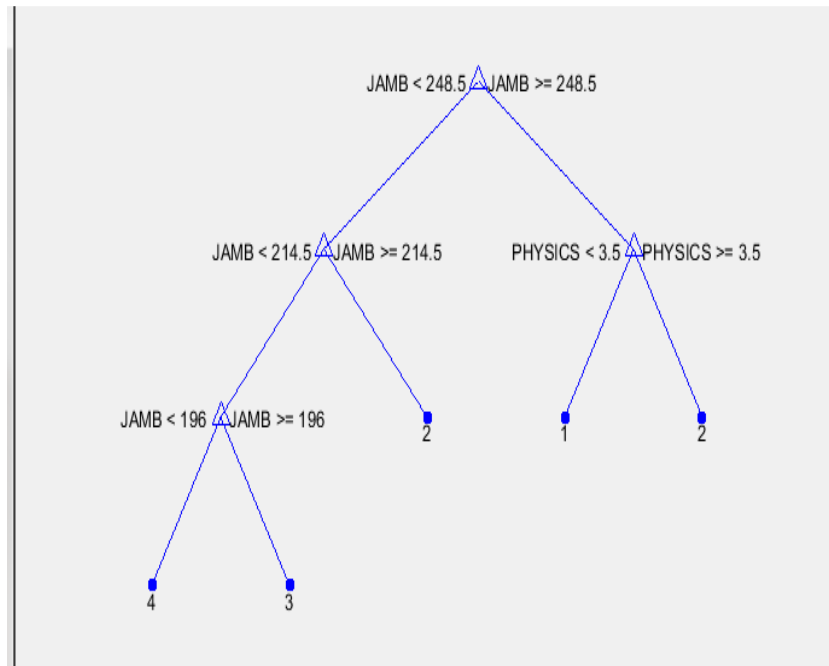


Figure 4.4: C4.5 Tree pruning

4.2.2 Randomtree Modelling And Testing

The selected data was then passed into the Randomtree algorithm which was in turn trained to extract knowledge discovery from the dataset, after the training was done the held out data for testing was also used to determine the how well our model has memorized experimental knowledge. The result obtained is presented below.

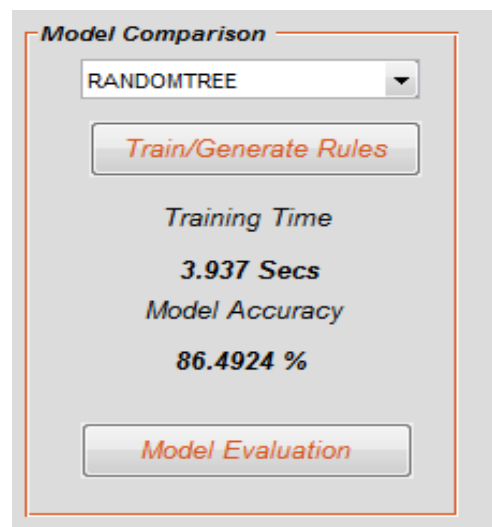


Figure 4.5: Randomtree classification

4.2.2.1 Confusion Matrix

As, in the confusion matrix in Figure 4.3, the confusion matrix here also shows the percentage classification of class label. Class 1 represent first class, Class 2 second class upper, Class 3 second class lower and Class 4 third class. The figure below shows the percentage accuracy of the classification and misclassification rate of each of these classes.

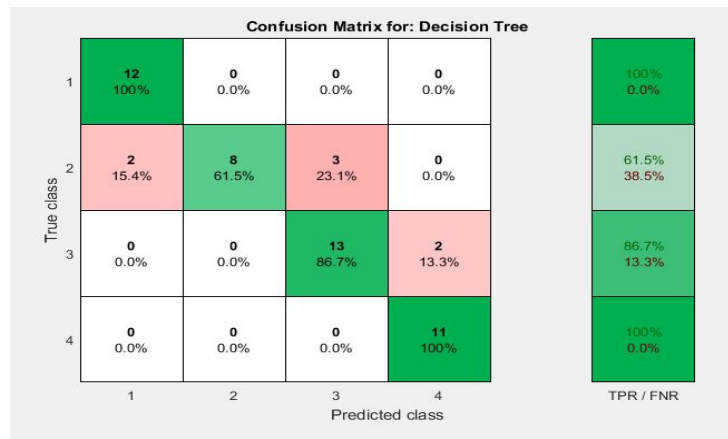


Figure 4.6: Randomtree Classification and misclassification accuracy

4.2.2.2 Tree Pruning

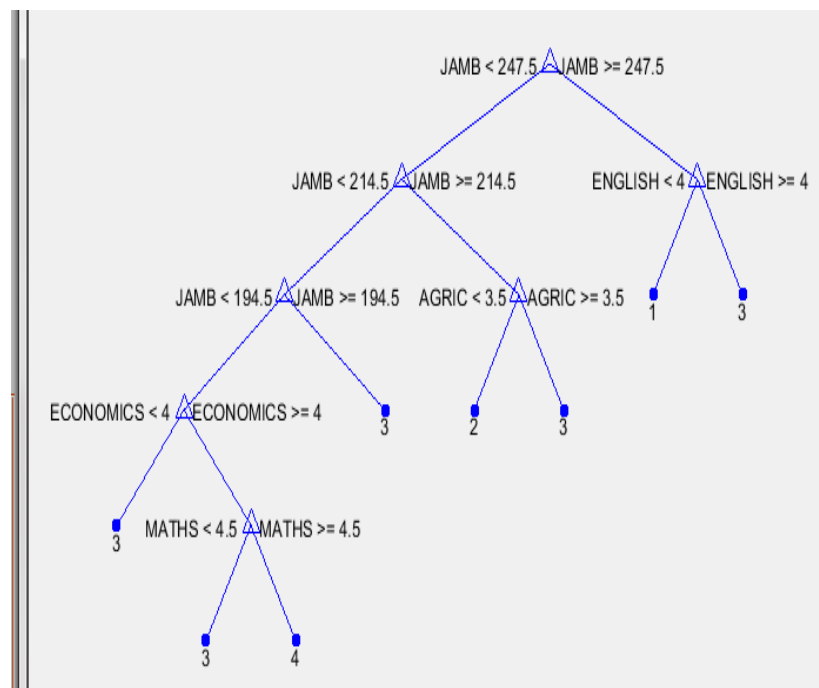


Figure 4.7: RadomTree Pruning

4.3 COMPARATIVE ANALYSIS

A. Classification Accuracy

This phase presents a comparative analysis of the Randomtree and C4.5 decision tree.

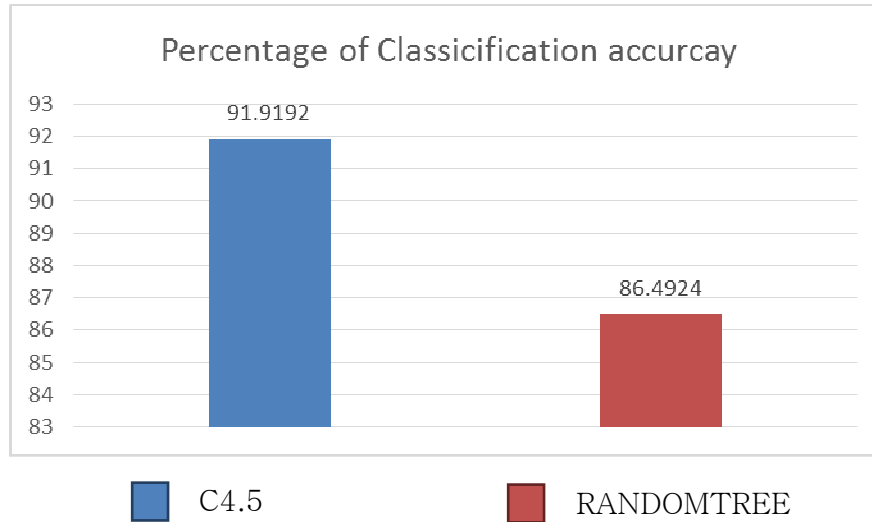


Figure 4.8: Classification Accuracy

B. Misclassification Accuracy

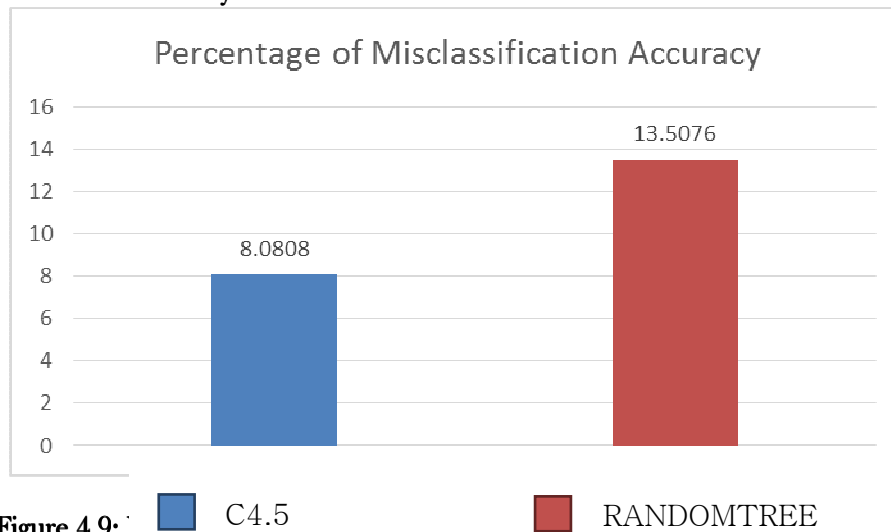
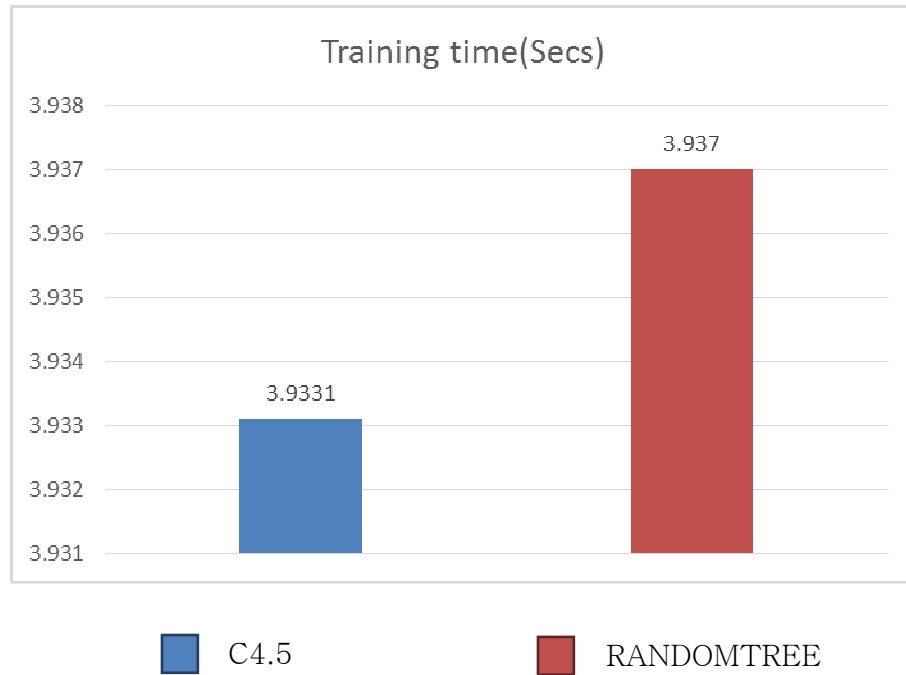


Figure 4.9:

C. Training Time

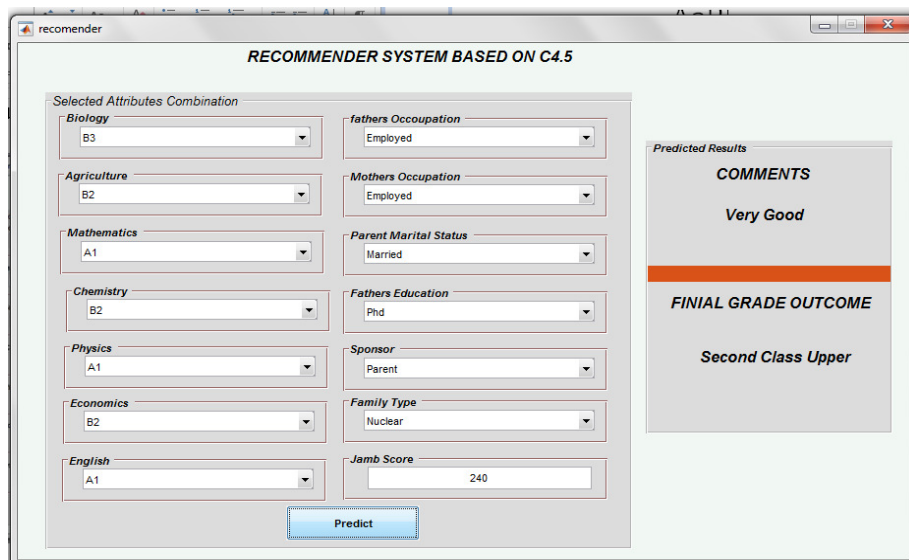
From the result obtained it is evident that the C4.5 decision was quite faster than the Randomtree as its training time was lower than that of the Randomtree decision tree.



From the evaluation of our results, it was observed that the C4.5 model was able to scale well in terms of their accuracy. The C4.5 also gave a better optimized time during the training phase. The recommender system was then developed based on the C4.5 algorithm. The interface is shown in figure 4.20 below:

4.3.3 Recommender System

The figure below presents the recommender system which scales on the C4.5 algorithm and the selected data obtained during feature selection phase.



RECOMMENDER SYSTEM BASED ON C4.5

Selected Attributes Combination

Biology B3	Fathers Occupation Employed
Agriculture B2	Mothers Occupation Employed
Mathematics A1	Parent Marital Status Married
Chemistry B2	Fathers Education Phd
Physics A1	Sponsor Parent
Economics B2	Family Type Nuclear
English A1	Jamb Score 240

Predicted Results

COMMENTS
Very Good

FINIAL GRADE OUTCOME
Second Class Upper

Predict

Figure 4.20: Recommender System Predicted Results

5. CONCLUSION

This study concludes that, the potential use of the classification model which is a support vector machine to predict probation status of students, does so, in a holistic and accurate manner. Also, higher education institutions would stand to benefit to the maximum, if this study is used to work on data sets consisting of actual number of students who have been dropped out of university. Despite the favourable outcome of the study, it is suggested that, working on larger data sets before using the SVM classifier for prediction is vital and could produce a more granular result.

REFERENCES

- 1) Amiena Bayat, (2014). "The Impact of Socio-economic Factors on the Performance of Selected High School Learners, Department of Economics, Faculty of Economic and Management Sciences.
- 2) Bhardwaj B. K. and Pal S. (2011) Data Mining: A prediction for performance improvement using classification, (IJCSIS) International Journal of Computer Science and Information Security, Vol. 9.
- 3) Yadav S. K., Bhardwaj B. K. and Pal S. (2012) Mining Education Data to Predict Student's Retention: A comparative Study, International Journal of Computer Science and Information Security (IJCSIS), Vol. 10, No. 2, 113-117.
- 4) Kuyoro 'Shade O., Nicolae Goga, (2013), "An optimal algorithm for predicting students' academic performance. International Journal of Computers & Technology www.cirworld.com Volume 4 No. 1.
- 5) Naeimeh Delavari, (2008), "Data Mining Application in Higher Learning Institution Informatics in Education - International Journal.
- 6) Vialardi C., Bravo J., Shafiti L. and Ortigosa A. (2009). Recommendation in Higher Education Using Data Mining Techniques, Educational Data Mining, pp190-199, 2009.
- 7) Vandamme, J.P., N. Meskens and J.F. Superby, (2007). Predicting academic performance by data mining methods. Educ. Econ., 15: 405-417.
- 8) Durairaj M., Vijitha C. (2014), Educational Data mining for Prediction of Student Performance Using Clustering Algorithms (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (4) , 2014, 5987-5991.
- 9) Zhongxiang F., Yan S. and Hong L. (2017), Clustering of College Students Based on Improved K-means Algorithm, Journal of Computers Vol. 28, No. 6, 2017, pp. 195-203
doi:10.3966/199115992017122806017
- 10) Nagesh S., Satyamurty S. (2018), International Journal of Computer Sciences and Engineering Application of clustering algorithm for analysis of Student Academic Performance (JSCE), Volume-6, Issue-1 E-ISSN: 2347-2693.
- 11) Sreedevi K., Chandra S. (2014), Analyzing the Student's Academic Performance by using Clustering Methods in Data Mining. International Journal of Scientific & Engineering Research, Volume 5, Issue 6, June-2014, ISSN 2229-5518.
- 12) Thaddeus O., Wilson C., George O. (2015), A Model for Predicting Students' Academic Performance using a Hybrid of K-means and Decision tree Algorithms, International Journal of Computer Applications Technology and Research, Volume 4- Issue 9, 693 - 697, 2015, ISSN: 2319-8656.
- 13) Vikas V., Shaweta B. and Harjit S. (2016), A Hybrid K-Mean Clustering Algorithm for Prediction Analysis, Indian Journal of Science and Technology, Vol. 9(28), DOI: 10.17485/IJST/2016/v9i28/98392.
- 14) Snehal B., Kedar S., Purva N., Rubana S., Odelia D., Saylee D. (2017), Predicting Student Performance Based On Clustering And Classification. IOSR Journal of Computer Engineering (IOSR-JCE), e-ISSN: 2278-0661, p-ISSN: 2278-8727, Volume 19, Issue 3, Ver. V (May-June 2017), PP 49-52.