A Conceptual Framework to Predict Academic Performance of Students using Classification Algorithm

Sujith Jayaprakash[1], Jaiganesh V[2]

[1]*Research Scholar, Dr. N.G.P Arts and Science College*
[2]*Professor, Department of PG & Research, Faculty of Computer Science, Dr. N.G.P Arts and Science College.*

Abstract - *Technological advancements have improved customer service and enhanced customer satisfaction to a large extent in the Industry and Service sectors. Education institutions across the globe are leveraging on the technology to produce high-quality graduates and improve the customer satisfaction level. Several researchers are striving hard to improvise the system through their innovative solutions. Education Data Mining (EDM) is an evolving field in Data mining due to its increasing demand in the higher education sector. Analyzing the students learning behavior and predicting their progression at the early stage will help the higher education institutions to produce quality graduates and to curb the student attrition. In this paper, we propose a conceptual framework that can act as a guide to develop a recommender system to predict the academic performance of students at the early stage by using classification algorithms. Various factors like Socioeconomic, Psychological, Cognitive, and Lifestyle are considered in analyzing the performance of students and predictions will be made based on their Semester GPA. Classification algorithms like Naïve Bayes, Random Forest and Bagging are used in finding the better prediction model.*

Key Words: Education Data Mining, Ensemble learning, Prediction framework, Boosting, Classification algorithm, Multivariate prediction analysis, Bagging, Enhanced Random Forest

1. INTRODUCTION

Higher Education Institutions across the world are churning out graduates and large populations of the graduate students are finding a job which is irrelevant to their course of study or found to be jobless. The reason behind this is that the majority of the institutions have failed to evaluate the quality of graduates produced; hence, they produce run-of-the-mill graduates who are unemployed or dissatisfied. Although every institution follows the traditional assessment models and grade students based on their performance there is no interim mechanism to find or evaluate their expectation, academic performance or level of understanding. Implementing such mechanisms will help the institution to make an early intervention to resolve the problems faced by students and improvise their performance. India's most progressive higher education sector was the engineering education but in recent times that is dwindling due to poor academic delivery in most of the engineering colleges and also due to the churn out of low-quality graduates.

- Why majority of the higher education institutions are not proactive?
- Why mechanisms are not put in place to evaluate the student's performance and make them job ready professionals?
- Why these low quality graduates are not warned at the early stage and helped them to improve their grades?

Every institution has to address these queries to produce graduates who can make great impact in the society through the education provided. Plethora of research work in education mining has given solutions which can address these issues to a large extent but the drawback is that these research works are not implemented as a full-fledged system to follow. Identifying the performance of students at the early stage of their studies helps institution to take decision on time. In recent times, several researchers have proposed solution by analysing the student's demography, socioeconomic factor, and their education level. Using various surveys and historic data it has been proved that the performance of a student can be predicted at the early stage and the various factors affecting the performance can also be identified. Although, majority of the research is carried out in the e-learning sector, few works have been performed on the traditional classroom teaching. Reason behind this is the lack of digitizing student information in higher education institutions. In e-learning system, every piece of information is recorded. Student's historic data, current performance, accessibility to the course module, active involvement in questionnaire sessions, interaction with peers are recorded and analysed using various algorithms which in-turn provides prolific results. Core objective of these research works are to identify a student's knowledge level and add them to a similar knowledge level group [1]. Significant contributions are made in the field of fraud detection, predicting customer behaviour, financial market, loan assessment, bankruptcy prediction, real-estate assessment and intrusion detection using Analysis and Prediction [2].

In this paper, student's academic performance is predicted using first semester GPA and various other factors like Socioeconomic, Psychological, Cognitive, and Lifestyle. Student historic data is collected from the University database and rest of the information is collected through survey. This paper focuses on Multiclass classification problem where in the predicting variable is classified into three classes. In this research work, classification algorithms are used to analyse the data to make early prediction about the academic performance of a student and various factors influencing his performance. A model framework is proposed using various classification algorithms like Naïve Bayes, Decission Tree and Ensemble learning algorithms. Ensemble learning algorithm provides good accuracy comparing the rest of the algorithms. This proposed framework is designed to build up a system which can help the institutions to capture student data and analyze their performance.

This research work aims to build a robust framework that can be developed as a recommendation system for predicting the performance of students in the higher education system. Furthermore, it aids the stakeholders to devise various strategies which can holistically develop a student’s performance.

In Section 2, we have discussed the related research works in this field, and Section 3 presents the detailed description of the data used for prediction and it’s pre-processing. Section 4 describes the proposed framework and various technologies to be used in developing it as a complete system. Section 5, 6, 7 and 8 will discuss about the implementation of various classification algorithms and its results. Section 9 will compare the results of all the algorithms and best algorithm for the model will be predicted. Finally, Section 9 discusses about the conclusion and future of the research.

2. RELATED WORK

Many studies on Education Data Mining have focused on predicting the performance of students using classification and regression algorithms. Several frameworks are proposed in line with this research work but implementations of such works are still at the nascent stage. Romero and Ventura [3], Amira and Wahida [4] and others have reviewed several research works in the last decade and justified the capabilities of education mining. Maria Goga et al. [3] has devised a framework that highlights poor performing students based on their academic performance in the first year. It's an early prediction mechanism which helps institutions to concentrate on the weak areas of students and steer them to score high in the sophomore stage. O. Adejo and T. Connolly [5] recommended a system that combines learners input and engagement while predicting their performance. Thus, this integrated framework extracts data from LMS as well as the Survey questionnaire filled by the student. Proposed framework makes use of the data collected from six (6) different domains like

- psychological,
- cognitive,
- economical,
- personality,
- demographic
- and institutional.

Data collected are analyzed using association rule mining and predictions are made by applying If-Then rule.

Fadhilah et al., [6] suggested a system to assess the performance of students based on their year one results. Few classification algorithms like the Decision Tree, the Naive Bayes and the Rule-Based algorithms were used to develop a prediction model. The final result shows that the Rule-Based algorithm outperformed the rest of the algorithms.

Z. Ibrahim and D. Rusli [7] used SAS Enterprise Miner to develop a predictive model which used student's demographic profile and the first semester's academic performance. Proposed models using Artificial Neural Networks, Decission tree algorithm and linear regression provided 80% accuracy. Upon building the model and evaluating it, ANN is predicted as the best model to predict the final CGPA. Carlos Villagrá-Arnedo et al., [8] attempted to study the learning progress of students using a Learning Management System. A learning platform has been developed to collect student data like usage of the platform, learning and training activities. Proposed model is build based on a classifier that uses Support Vector Machine. The study reveals that the student's behavioural data coupled with learning data attributes to a better prediction. In the learning management system, the student can upload the exercises to be auto evaluated and the feedback will be given on time. During the assessment process a list of concrete events that occur during the interaction between students and the system was considered. Such events are stored in the event log to analyze. Learning and Behavioral data are collected in order to analyze and predict the data. Sattar Ameri et al., [9] proposed a framework named "survival analysis framework", it’s an early prediction mechanism. The proposed model predicts the performance of a student based on the demographic details, socioeconomic status, high school information, enrollment details and the semester credits. Thus, the research shows that a student's pre and post-enrollment data attributes to a better prediction on his performance. Additionally, this framework helps institution to estimate the semester of dropout based on pre-enrollment attributes. Comparing the COX proportional hazards model and time-dependent COX (TD-COX) model, TD-Cox shows better performance accuracy. A Theoretical framework recommended by Raheela Asif et al., [10] can help the institution to make a clear human judgment as it consolidates distillation of data, clustering and the performance prediction. A Prediction model is developed using the student's four-year academic performance. Several models are built using ten different algorithms. Some of the attributes like socioeconomic status or demographic details have refrained in the model. Models developed using 1-nearest neighborhood, random forests with Gini index and the naive Bayes algorithms shown high accuracy. This research also revealed the strong indicators that influence the performance of students.

Recommendation model developed by Bo Guo et al., [11] named "Student Performance Prediction Network", was built on a deep learning algorithm. Algorithm helped to identify the complex representations of data and extracted the useful insights. The proposed network used six layers of the neural network to implement the deep learning algorithm. Results of models developed from the Multilevel Perceptron, the Naïve Bayes and the Support Vector Machine were compared with the results of SPPN. However, the research exhibits that the hybrid model has better accuracy results than conventional models.

3. METHODOLOGY

The objective of this research work is to develop an early intervention mechanism that could identify the students who are weak in their performance and classify different parameters that could influence the performance. A total of 155 student's records are randomly selected for this research work. We proposed a multiclass classification approach by classifying the performance of students based on their grades. We classified student's performance as First Class (FC) holders, Second Class (SC) holders and Third class (TC) based on their final grade point in the first semester.

4. DATA PREPROCESSING

Various factors like Socioeconomic, Psychological, Cognitive, and Lifestyle are considered in analyzing the performance of students along with their Semester 1 GPA. Using this dataset performance in Semester 2 will be predicted.

Table 1:- Socioeconomic Attributes

Gender	M/F
Family Size	Small, Medium and Large
Family Income	Between 200,000 and 500,000 Between 100,000 Rs. and 200,000 Rs. <100,000 Rs. Above 500,000 Rs. Between 200,000 Rs. and 500,000 Rs.
Parent Education Status	Both are educated Only father is educated Both are not educated Only mother is educated
Medium of Study	English, Tamil, Others
Average travel distance to school from Residence	Between 10KM to 25KM <10 KM Between 25KM to 50KM More than 50KM

Table 2:- Psychological Attributes

Group of Study	Science and Maths, Arts and Commerce, Computer Science, Biology
Rating of Reading Habit	Very Good, Good, Moderate, Poor, Very Poor
Rating of Concentration level during class	Very Good, Good, Moderate, Poor, Very Poor

Table 3:- Cognitive Attributes

Reason to choose this program	Own interest, Recommended
Usage of education resources like College Library, E-Library, etc.,	Very Often, Often, Sometime, Rarely, Never

Table 4:- Lifestyle Attributes

Usage of social media platform like Facebook, Twitter, Whatsapp and Instagram	Very Often, Often, Sometime, Rarely, Never
Rating of Social Skills	Very Good, Good, Moderate, Poor, Very Poor

Table 5:- Academic Performance Attributes

Grade obtained in Secondary School Level	FC, SC, TC
Grade obtained in Senior Secondary school level	FC, SC, TC
Grade obtained in Semester 1	FC, SC, TC

Aforementioned data are collected from student enrollment records, institutional surveys and from student academic record. From the data collected using the survey and from the databases, missing and erroneous data are removed as part of the Data Cleansing process. Semester 1 GPA are classified as First Class, Second Class and Third Class. The performance of a machine learning algorithm differs from one model to another. Algorithms are evaluated based on its prediction accuracy. If data in the model is imbalanced, then it will affect the performance of a model and provide a poor accuracy [12]. As real dataset is used in this research work the dataset found to be imbalanced due to lower number of First Class students comparing to Second Class and Third Class. In this research work, we used Synthetic Minority Oversampling technique to address the class imbalance problem.

a. **SMOTE**: Synthetic Minority Over-sampling Technique

The accuracy rate of a prediction model is highly dependent on the algorithm and the dataset. If the training dataset has imbalanced class then there is a high chance of inaccuracy in the prediction. Hence to overcome the problem of inaccuracy classifiers are evaluated. In this dataset only 15% of the students have scored first class comparing to the Second Class and Third Class. Hence, the minority class is oversampled using SMOTE Technique. Fig (a) and (b) below describe the dataset before and after using oversampling technique.

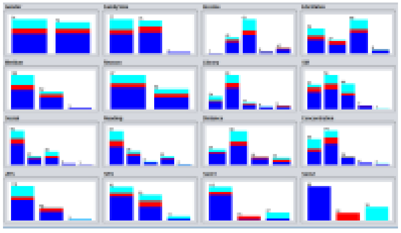


Fig (a) Before using SMOTE Technique

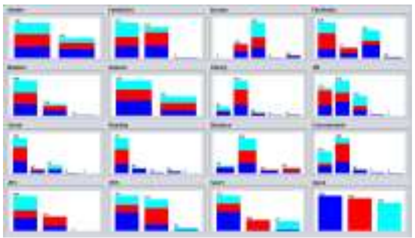


Fig (b) After using SMOTE Technique

A.T.M. Shakil Ahamed et al., has used SMOTE Boost technique to fix the skewed dataset [13]. Similarly, Chawla et al. identified the minority class in the dataset and improvised it using the SMOTE Technique and later applied the boosting algorithm to predict the performance [14].

5. CONCEPTUAL FRAMEWORK AND PROPOSED MODEL

This Conceptual framework helped us to build a model which can be implemented to find a viable solution. Based on the experiment conducted using different Machine learning algorithms, the proposed model will use an efficient algorithm that perfectly fits into the problem to produce a high accuracy.

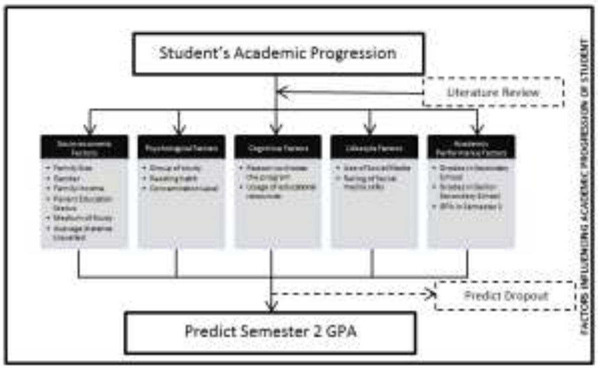

Student's Academic Progression
Literature Review
Socio-economic Factors
Family Size
Gender
Family Income
Parent Education Status
Medium of Study
Average Distance travelled
Psychological Factors
Group of study
Reading habit
Concentration level
Cognitive Factors
Reason to choose the program
Usage of educational resources
Lifestyle Factors
Use of Social Media
Rating of Social media skills
Academic Performance Factors
Grades in Secondary School
Grades in Senior Secondary School
GPA in Semester 1
Predict Dropout
Predict Semester 2 GPA
FACTORS INFLUENCING ACADEMIC PROGRESSION OF STUDENT

Fig (c): Conceptual Framework

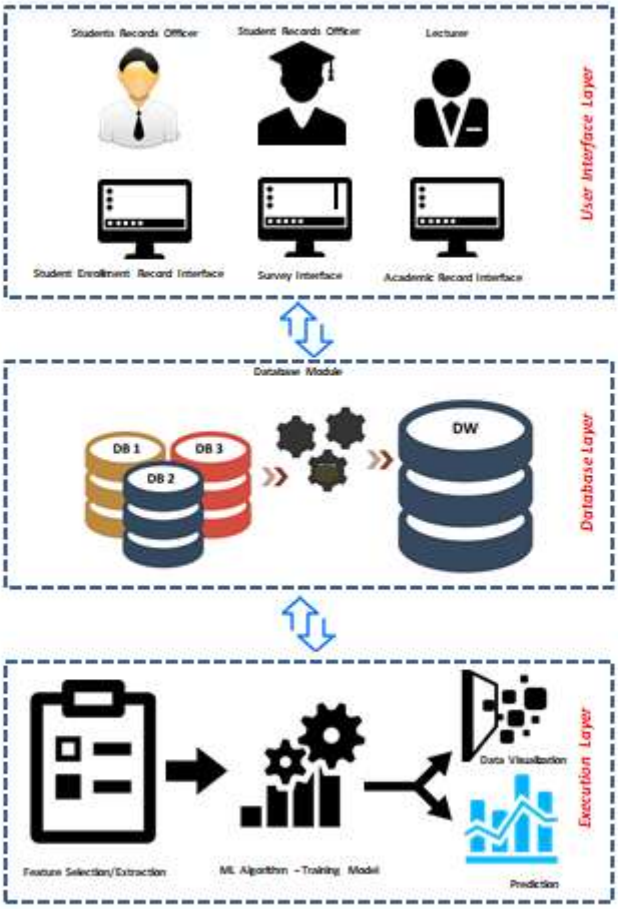


Fig. (d) –Model

6. IMPLEMENTATION OF NAÏVE BAYES ALGORITHM

Naïve Bayes Classifier is also known as simple Bayes or independence Bayes which is used to construct classifiers and to identify the membership probabilities of each class. Naïve Bayes algorithm works efficiently in a supervised environment and it is scalable. It is considered to be a simplistic and robust classification algorithm in predicting a variable. Naïve Bayes algorithm is largely used in predictions as it generally outperforms in a refined classification methods. Naïve Bayes is a conditional probability model. Considering the given dataset where x represents the independent features like Gender, Family Size, Income, Education status etc., the model predicts K which represents the Semester 2 GPA which is classified as First Class, Second Class or Third Class.

$$\text{posterior} = \frac{\text{prior} \times \text{likelihood}}{\text{evidence}} \qquad p(C_k \mid \mathbf{x}) = \frac{p(C_k)\, p(\mathbf{x} \mid C_k)}{p(\mathbf{x})}$$

In this proposed model, we used cross-validation to asses the model's performance. Both 5-fold and 10-fold cross validation are used to evaluate the model. In the former the model attains 83% accuracy while the latter get 82.5% accuracy.

7. IMPLEMENTATION OF RANDOM FOREST ALGORITHM

Decission tree fits in both classification and regression problems. It provide tree like structure with possible consequences that gives better understanding of the model. It works based on If-then condition and it's easy to interpret. Various types of decision tree classifiers are ID3, C4.5 and CART. Decision tree algorithms are highly useful to evaluate a complex dataset and the execution time is faster comparing to other machine learning algorithms. Random forest or random decision trees are ensemble learning method used for classification. Random Forest algorithm is used to handle complex datasets or when there is a deep tree structure. Multiple Decission trees are created and merge them to get better prediction accuracy. Hence, this algorithm becomes a preferred choice among conventional machine learning algorithms. Decission are constructed based on the information gain and Gigi index approach.

In this model, Random forest algorithm is used to classify and predict the output. Dataset is trained and tested based on the 5-fold and 10-fold cross validation. There is an high accuracy of 89.32% in the 10-fold cross validation comparing to 5-fold which has an accuracy rate of 88.2%

8. IMPLEMENTATION OF BAGGING ALGORITHM

Bagging is an ensemble technique in which various predictors are combined to make a better accuracy rate. Bagging is also known as a Bootstrap aggregation which uses multiple classifiers and the results are combined through model averaging technique. This is to reduce the over fitting of a model. Bootstrapping is a classical statistical technique which helps to learn a new subset of data by sampling the existing dataset. Different training sets are created from the existing dataset and it is tested. Bagging helps to reduce the complexity. In Bagging, all features are considered in splitting a node whereas Random Forest selects only subset of features.

Our model is tested using bagging algorithm and found that it provides 93.243% accuracy in 5-fold cross validation whereas 10-fold cross validation provides 94.208% accuracy.

9. COMPARISON OF RESULTS

The main focus of this research work is to identify the key parameters that influence the performance of students. Few supervised learning algorithms like the Naive Bayes, the Regression Tree and the Bagging classifiers are used to build a prediction model. We partitioned the dataset into different subsets. We partitioned the dataset into multiple subsets. Each subset is then tested with the training data to predict the accuracy. We also examined the model's performance using True positive (TP) rate, false positive (FP) rate, Precision and recall.

a. True Positive

When a model correctly predicts the outcome then it is called as True positive. Fig (c) shows the True Positive rate of all the models used. From below fig. it is evident that Bagging with 10-fold cross validation has highest prediction compared to the rest of the models.

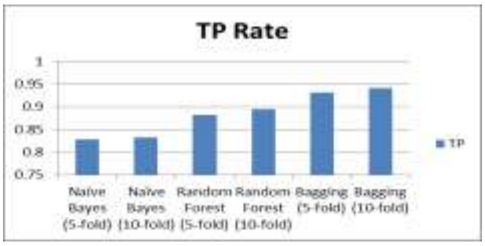
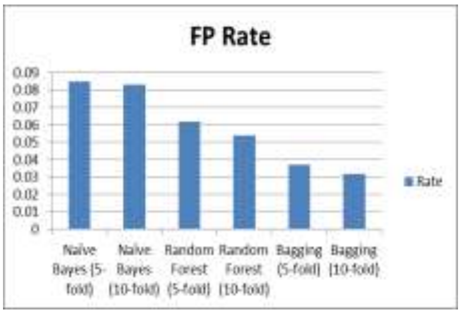


Fig (d) – True Positive Rate

b. False Positive

When a model is incorrectly predicts the outcome then it is called as False positive rate and the model which has lesser FP rate is considered to be more accurate. Comparatively, Naïve Bayes algorithm has shown high false positive rate comparing to the rest of the models.



Fig(e) – False Positive Rate

c. Precision

Precision is measured based on the number of True Positive divided by the number of True Positive and False positive rates. Fig. below shows the Precision rate of all the models. Bagging holds the highest precision value of .94 comparing to the rest of the algorithms.

Precision is defined as,

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

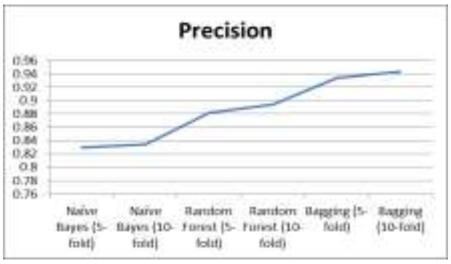


Fig. (f) – Precision Rate

d. Recall

Recall identifies the proportion of actual positives identified correctly. Recall is defined as below,

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

Fig. below shows the Recall rate of the models evaluated

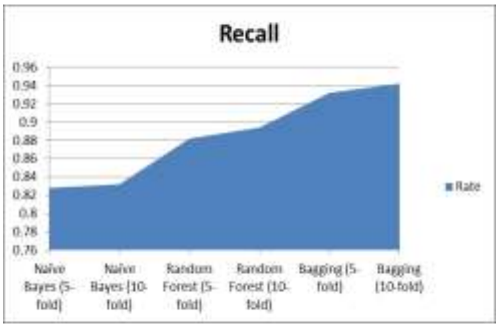


Fig. (g)-Recall Rate

e. F-Score

It is average of Precision and Recall. F-Score is defined as below,

$$F_1 = \left(\frac{\text{recall}^{-1} + \text{precision}^{-1}}{2}\right)^{-1} = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

Fig. below shows that F-Value of Bagging is higher than the other models.

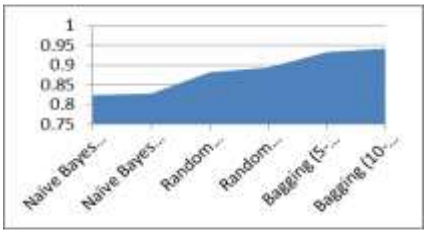


Fig. (h) – F-Score

10. COMPARISON OF RESULTS

Predicting the academic progression of students in higher education system is very crucial and eminent for the growth of any institution. This prediction not only helps the student to understand and better their performance but also helps the institution in assessing the quality of education provided. To a large extent these mechanisms can also reduce the student attrition rate. Factors such as Gender, Family income, parent's education, distance travelled, size of the family, reading habit, usage of social media skills and their academic performance can highly influence their performance. Higher Education Institutions should invest on analyzing these influential factors and aid the students who are not performing or in the verge of drop-out due to the influential factors. In this research work, we have tried analyzing these factors using three machine learning algorithms and predicted the output. As stated in No Free Lunch Theorem, performance of all algorithms is equivalent and it is purely based on the problem used. In some cases, Naïve Bayes can outperform the rest of the algorithms and vice versa. In this research work, we have used Naïve Bayes, Random Forest and Bagging to classify the problem and predict the result. From the experimental results, it's evident that Bagging outperforms the rest and it is the best suitable algorithm for the problem defined. Though the model's performance is satisfactory, the accuracy rate of the model can be still improved. Hence, our future research will be to identify the most relevant features in our dataset using feature ranking algorithms and increase the accuracy rate of our model. The framework proposed in this research work shows that the model used can identify the weak performers at the early stage. This intervention mechanism can help institutions to produce high-quality graduates and avoid attrition rates to a greater extent.

REFERENCES

1. Ayers, E., Nugent, R., Dean, N. (2009). A Comparison of Student Skill Knowledge Estimates. 2nd International Conference on Education Data Mining, Cordoba, Spain, pp. 1-10
2. Pooja, T., Mehta, A., Manisha. (2015) Performance Analysis and Prediction in Educational Data Mining: A Research Travelogue. International Journal of Computer Applications. vol. 110, no. 15
3. Cristobal Romero (2010), "Educational Data Mining: A Review of the State-of-the-Art", IEEE Transactions on. Systems, man and cybernetics- Part C: Applications and Reviews vol. 40 issue 6, pp 601-618.
4. Shahiri, Amirah & Husain, Wahidah & Abdul Rashid, Nur'Aini. (2015). A Review on Predicting Student's Performance Using Data Mining Techniques. Procedia Computer Science.
5. Adejo, Olugbenga & Connolly, Thomas. (2017). An Integrated System Framework for Predicting

Students' Academic Performance in Higher Educational Institutions. International Journal of Computer Science and Information Technology. 9. 149-157. 10.5121/ijcsit.2017.93013.

6. Fadhilah Ahmad, Nur Hafieza Ismail and Azwa Abdul Aziz. The Prediction of Students' Academic performance Using Classification Data Mining Techniques. Applied Mathematical Sciences, Vol. 9, 2015, no. 129, pp. 6415– 6426.
7. Ibrahim, Z. & Rusli, D. (2007). Predicting student's academic performance: Comparing artificial neural network, decision tree and linear regression. Paper presented in the 21st Annual SAS Malaysia Forum, 5th September 2007, Shangri-La Hotel, Kuala Lumpur.
8. Villagrá, Carlos & Durán, Francisco José & Rosique, Patricia & Llorens, Faraón & Molina-Carmona, Rafael. (2016). Predicting academic performance from Behavioural and learning data. International Journal of Design & Nature and Ecodynamics. 11. 239-249. 10.2495/DNE-V11-N3-239-249.
9. Ameri, Sattar & Jahanbani Fard, Mahtab & Chinnam, Ratna Babu & Reddy, Chandan. (2016). Survival Analysis based Framework for Early Prediction of Student Dropouts. 10.1145/2983323.2983351.
10. Asif, Raheela & Merceron, Agathe & K. Pathan, Mahmood. (2014). Predicting Student Academic Performance at Degree Level: A Case Study. International Journal of Intelligent Systems and Applications. 7. 49-61. 10.5815/ijisa.2015.01.05.
11. Guo, Bo & Zhang, Rui & Xu, Guang & Shi, Chuangming & Yang, Li. (2015). Predicting Students Performance in Educational Data Mining. 125-128.
12. Chawla, Nitesh & Bowyer, Kevin & O. Hall, Lawrence & Philip Kegelmeyer, W. (2002). SMOTE: Synthetic Minority Over-sampling Technique. J. Artif. Intell. Res. (JAIR). 16. 321-357. 10.1613/jair.953.
13. A. T. M. Shakil Ahamed, Navid Tanzeem Mahmood & Rashedur M Rahman (2017) An intelligent system to predict academic performance based on different factors during adolescence, Journal of Information and Telecommunication, 1:2, 155-175, DOI: 10.1080/24751839.2017.1323488
14. Chawla, N. V., Lazarevic, A., Hall, L. O., & Bowyer, K. W. (2003). SMOTEBoost: Improving prediction of the minority class in boosting. Knowledge Discovery in Databases: PKDD, 107–119. doi: 10.1007/978-3-540-39804-2_12

BIOGRAPHIES

Sujith Jayaprakash is a research scholar at Dr N. G. P College of Arts and Science. His area of research is in machine learning algorithms, the academic progression of students, web mining, Use of education apps etc. He has over a decade of experience in Education Administration and academia.

Dr. Jaiganesh V. is currently working as a Professor at Dr N. G. P College of Arts and Science. His area of specialization includes Data mining and Machine learning. He has 19 years of teaching experience and guided several research scholars in the field of Data mining.