ISSN: 0974-5823 Vol. 6 No. 3 December, 2021

International Journal of Mechanical Engineering

Undergraduate Student’s academic performance prediction using Support Vector Machine (SVM) and Random Forest algorithm

Rashmi V. Varade[1], Dr. Blessy Thankanchan[2]
[1,2] Jaipur National University, Jaipur

Abstract:
Educational data mining is the application of machine learning and data mining to predict student performance in the educational sphere, and it has long been a popular research topic. Early prediction of student performance may aid responsible organisations in providing solutions to kids who are performing poorly.

Many factors can influence a student's final test achievement (for example, past assignment grades, social life, parents' jobs, and absence frequency).

This research tries to forecast student academic performance in order to improve educational organisations' performance and help students achieve better academic achievements.

Support Vector Machines (SVM) and Random Forest were used as classification methods and methodologies in this study (RF). Both SVM and RF have been used to apply binary classification and regression techniques. In this paper, we use a dataset of 600 students and accuracy is calculated.

Keywords: Machine learning, support vector machine, supervised learning, random forest,

1 INTRODUCTION: -

If the massive amount of student data has been examined and translated into knowledge, it might be helpful and used in prediction.

This knowledge might help students thrive in their academic careers by improving the quality of teaching and learning. Educational Data Mining (EDM) is a subfield of data mining that focuses on educational datasets. [1]

EDM intends to create and utilise data mining, machine learning, and statistics approaches to evaluate data acquired from students over the course of their academic careers [2].

Applications of EDM prediction have been studied, and one of the applications is forecasting student performance. Predicting student performance can aid the responsible organisation in identifying students who may require further support.

We try to use machine learning algorithms to forecast students' academic achievement and help to improve the educational process and students' experience at the educational institution. We’ll use two classification algorithms to calculate a forecast for the students' final grade. These papers helped us determine the issues of student performance prediction and how they were measured in previous researches, and find other ways to use and apply the dataset to find different outcomes, based on our previous findings for the literature review of several papers that discuss the educational system.

The study's goal was to forecast students' academic success, identify concerns and problems that impact students' outcomes, and propose strategies to enhance the educational system so that students can achieve better academic achievements and the system's overall results.

Academic achievement can be influenced by a variety of circumstances.

We try to figure out what aspects in the educational system are impacting students' performance, as well as what other issues, such as the economy, society, and others, are negatively affecting pupils.

REVIEW OF RELATED LITERATURES: -

Because of the importance of education, governments and many organisations are paying close attention to it these days. Student performance is an important aspect of the learning process and one of the indicators of high-quality educational institutions.

In many situations, it is necessary to predict student performance in order to identify pupils who are more likely to have poor academic achievement.

Various aspects, such as social, economic, personal, and others, are used to evaluate performance.

The prediction of a student's academic performance aids in improving education quality and ensuring that the student receives an appropriate education environment.

It also aids instructors and educational planners in determining the most effective techniques for improving student academic performance. [3]

Many publications have been published that address student performance measurement utilising machine learning algorithms in a variety of scenarios, as detailed in the sections below.

Devasia et al. employed naive Bayes in [4], and the results revealed that it may assist students and lecturers in doing well, observing those students who need more attention, minimising the failure ratio, and suggesting appropriate action for the following semester.

The Naive Bayes algorithm outperformed the others.

Previous research did not consider accuracy, and models were designed to improve student performance based on the success/failure ratio.

Meanwhile, Vijayalakshmi et al. perform a field investigation in [5]. Using a deep neural network to develop a model of educational data mining System for predicting student success. They want to be able to anticipate the future.The model's performance is depending on the many inputs kept in the model.The dataset used to train and test the model .The information contained in the dataset offers 20 characteristics and information on 600 pupils.

with an accuracy of 84% Kabakchieva's research is similar to Vijayalakshmi's.[6]

MATERIAL AND METHODS: -

Description about Dataset: -Because the academic success rate of students is critical to the operation of any educational institution. As a result, data for this study was acquired from the University of Mumbai's undergraduate degree institutions.

Nearly 600 pupils' information has been gathered. For the years 2019-20 data on students is being gathered. Stream, age, religion, parents' education, parents' financial circumstances, marital status, and past academic performance are all characteristics included in this study.

The following characteristics are used in the study: Steam:

-There are three undergraduate streams: Arts, Commerce, and Science.

Table 1. Attributes used for analysis

Name of the attribute	**Description**
Age: -	Age of the student
Gender:	Gender of student e.g. Male or Female
Attendance: -	Attendance of the student
Educational Gap:	Is there any gap during education.
Internet Connection:	is there internet connection at home or not.
SSC %: -	percentage obtained at SSC
HSC%: -	percentage obtained at HSC
Travel time: -	Travel time from home to institution
Admission category: -	Scholarship or fees paid
Family Income: -	Income of the family
Family size: -	Size of the family
Fathers' qualification: -	Qualification of the father
Mothers Qualification: -	Qualification of the mother
Fathers' occupation: -	occupation of father.
Admission category: -	Scholarship or fees paid
Mothers' occupation: -	Occupation of mother
Friends: -	how many friends are there?
Marital status: -	marital status of the student.
Marks obtained by students :-	Sem1, sem2, sem3, sem 4: - semester exam marks
Remark: -	based on marks remark is given

According to students score they are classified into three categories

1. Excellent (E) – course grade of the student is between 75-100%
2. Good (G) - course grade of the student is between 51-74%
3. Poor (P) - course grade of the student is between <50% -35

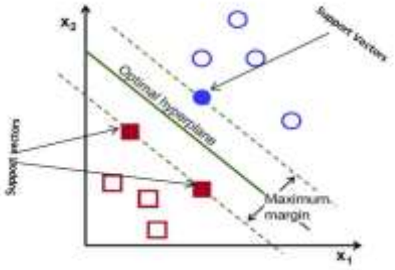
3.2 Machine Learning Model: -

Support Vector Machine: - Vapnick (1995) invented the Support Vector Machine (SVM) learning technique to handle prediction and pattern recognition problems, as well as analysis and mapping of both linear and non-linear functions.

In a high-dimensional space, it creates a hyperplane or group of hyperplanes (classes) that may subsequently be utilised for classification. [9] It divides objects into groups (above or below plane) depending on their characteristics, and it may transform nonlinear to linear before dividing using Kernel approaches. Its operations are based on statistical learning theory and the principle of structural risk minimization. [10]

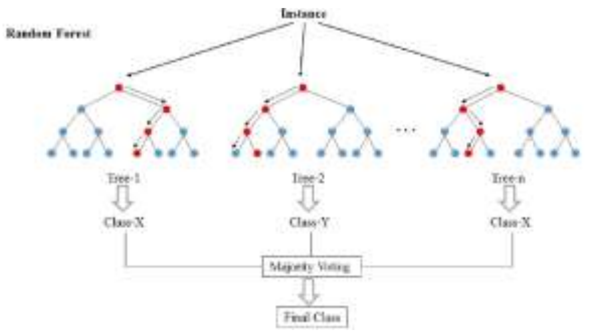
SVM is excellent for numerous predictive factors, especially when group prediction rather than individual prediction is desired. It was chosen for the study because of its broad range of applications and adaptability. Because of its ability to minimise mistakes, SVM can swiftly learn a bigger collection of patterns and is more accurate in generalisation. It also offers the capacity to dynamically adjust training patterns as new data becomes available.

SVM has been widely used in a variety of industries (banking, for example).



Random forest: -Random Forest is just a bagging strategy. Some of the row and feature samples are collected and delivered to one of the multiple basic learners in this method.

Learners are decision trees in a random forest foundation. This is essentially a bootstrap stage. After then, the data is compiled using majority voting.



Model Evaluation and Analysis: -The pre-processed data was divided into two sets: training (70 percent) and testing (20 percent) (30 percent). A training set was used to create prediction models, and a testing set was utilised to confirm them. On the training dataset, it chooses the optimal parameter that matches the estimator.

Traditional classification algorithms such as DT, NB, and KNN, as well as standard ensemble approaches such as Bagging, Boosting and Random Forest, were used to create the prediction models. The models' outputs were verified, and a comparison analysis was carried out in the following stage.

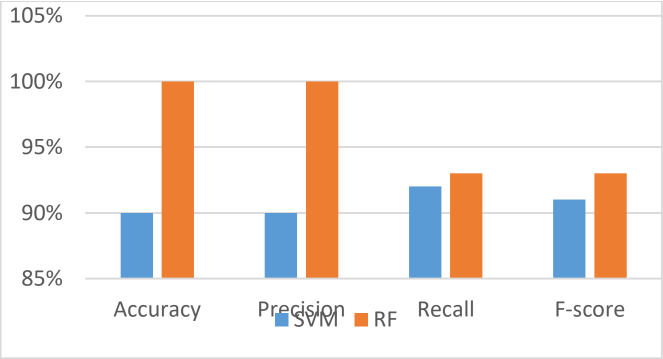
Evaluation Measure: - In the event of unbalanced classes, the model's accuracy cannot be used to evaluate the model.

As a result, the following assessment measures were used to assess the model:
The ratio of accurately categorised examples to the total number of occurrences is known as accuracy.

accuracy = T P + T N
T P + T N + FP + FN
Precision is defined as the ratio of correctly classified
positives cases and the positively predicted cases
precision = T P
T P + FP
Recall is defined as the ratio of correctly classified
positive cases and the total positive cases
recall = T P
T P + FN
F1-score is the harmonic mean of precision and recall
F1-score = 2 ∗ precision ∗ recall
precision + recall

Result: -We applied to algorithms to the dataset and we found that accuracy of SVM algorithm is 90% and precision is 90% recall is 92% and F-score is 91% While Random forest applied to dataset gives accuracy 93% precision 93% recall 92%and F-score also 92%.

Algorithm	Accuracy	Precision	Recall	F-score
SVM	90%	90%	92%	91%
RF	100%	100%	99%	99%



Accuracy of RF and SVM algorithm

Conclusion and future work:

In many areas, the ensemble learning methodology gives great accuracy. Its usage in education as a novel strategy is fast gaining traction. In conclusion, many factors can influence a student's academic success; nevertheless, in our research, we discovered that the preceding grade had the most influence on final grades. This study suggests that Random Forest algorithm gives highest accuracy for prediction of students performance.
The model will be evaluated on more datasets in the future, and it will can be upgraded to use a deep learning model.

Reference:

[1] Leena H. Alamri, Ranim S. Almuslim, Mona S. Alotibi, Dana K. Alkadi,Irfan Ullah Khan, and Nida Aslam. 2020. Predicting Student Academic Performance using Support Vector Machine and Random Forest. In 2020 3rd International Conference on Education Technology Management (ICETM 2020), December 17–19, 2020, London, United Kingdom. ACM, New York, NY, USA,8 pages. https://doi.org/10.1145/3446590.3446607

[2] E. Fernandesab, M. Holandaa, M. Victorinoa, V. Borgesa, R. Carvalhoa and G. V.Ervenac, "Educational data mining: Predictive analysis of academic performanceof public school students in the capital of Brazil," Journal of Business Research,pp. 335-343, January 2019.

[3] F. Berhanu and A. Abera, "Students' Performance Prediction based on theirAcademic Record," International Journal of Computer Applications, pp. pp27-35, December 2015.

[4] T. Devasia, V. TP and V. Hegde, "Prediction of Students Performance using Educational Data Mining," 2016.

[5] V. Vijayalakshmi and K. Venkatachalapathy, "Comparison of Predicting Student'sPerformance using Machine Learning Algorithms," vol. 11, no. 12, pp. 34-45, 2019.

[6] D. Kabakchieva, "Student Performance Prediction by Using Data Mining Classification Algorithms," International Journal of Computer Science and Management Research, vol. 1, p. 686–690, 4 November 2012.

[7] Vapnik, V. and Cortes, C., (1995). Support-vector networks. Machine learning, 20(3), 273297.

[8] Kotsiantis, S. B., & Pintelas, P. E. (2005, July). Predicting student's marks in hellenic open university. In Advanced learning technologies, 2005. ICALT 2005. fifth IEEE international conference on (pp. 664-668). IEEE.