Student Intervention System Using ML

M. Ananya[1], K. Akhila[2], E. Vinay[3], Dr M. Rekha Sundari[4]

[1,2,3] B Tech Students, Department of Information Technology, Anil Neeru Konda Institute of Technology &Sciences, Visakhapatnam (Andhra Pradesh), INDIA
[4] Associate Professor, Department of Information Technology, Anil Neeru Konda Institute of Technology &Sciences, Visakhapatnam (Andhra Pradesh), INDIA

Abstract: In most universities, advisors guide their respective students in choosing their course, and warn people who may be in danger of being unemployed or placed on probation. The massive range of scholars makes it difficult for universities to identify those at risk, as it would get very time consuming and inaccurate. Hence, there's a desire for a system which will acknowledge these students at the tip of every semester. Machine learning can be used to identify students who are at risk and whether or not they are qualified for placements, allowing us to provide them with a more personalized and tailored approach throughout this work. The primary goal of this report is to forecast student performance on the test and to provide a prediction about whether or not the student will be placed.

Keywords: Gaussian Naïve Bayes, Intervention, K Nearest Neighbour's, Prediction, Support Vector Machine.

Introduction

It has become more common for educational institutions to gather huge amounts of data on a regular basis that includes information on students' grades, their personal information (such as their health and travel times), and records of their activities. We must assist students succeed at all levels of school, but we must not lower the education level. Teachers and administrators are looking for new ways to forecast success and failure early enough to stage suitable interventions and to determine the effectiveness of different treatments. It may be possible to use this data in the future to help students get better placements by intervening.

To do this, a decision must be taken where we need to answer some queries such as

1. Who are the students likely to getting not placed?
2. What are the courses that students may fail?
3. What really is the quality of the student's involvement to the grades he or she received?
4. Which courses must be prioritised in order to assist students get placed?
5. What are the easiest topics to do well on in a written placement test?

Based on the responses gathered during the analysis phase, a detailed choice may be made utilising various machine learning approaches and algorithms, allowing pupils to be steered in a precise path while maintaining educational standards. Many factors impact a student's academic achievement, including personal, economical, and other factors. Learning about these elements and how they affect student performance might help you cope with them. Estimating academic achievement has received a lot of attention. In educational institutions, the capacity to predict student success is critical. In educational institutions, increasing student success is an ongoing goal. Predicting pupils' academic performance ahead of time before placement drives is a significant accomplishment. At this stage, educational institutions should take certain steps to provide suitable assistance to low-performing students in order to help them strengthen their studies and advance.

EASE OF USE:

A) LITERATURE REVIEW

The primary goal of a literature review is to find new methods to interact with and extract new information from an existing data set. Children's academic performance has been studied extensively during the last several decades [1]. When it comes to student characteristics that impact academic performance and, more importantly, job placement, we'll take a look at a few study publications in this part. Each paper, article, or book chapter that is submitted for evaluation is given a priority rating out of 10. The world's educational institutions are very different. Diverse educational goals, structures, sizes, and methodologies offer students with a variety of options while also encouraging competition and new ideas. There is a growing demand for institutions to collect data that better reflects the complexity and dynamic nature of their environments [2]. J48, Simple cart, kstar, SMO, NaiveBayes, and OneR classification algorithms were used to classify data from Oktariani Nurul Pratiwi's high school. With a 79.61 percent accuracy rate, J48 and Simple Cart predicted the best of the rest. [3]. Using the course of students' midterm grades, lab test grade, and seminar performance as predictors of SAP, El Din Ahmed and his colleagues conducted a study in 2014. Fadhilah and Azwa (2015) utilised the Informix Database Management System of Unis SZA's Academic Department to collect data (DBMS). Besides gender and race, GPA and family wealth, university entrance style as well as grades in English and mathematics and grades in other areas, they used nine more factors. [4]. Self-Evaluation When it comes to predicting a student's future academic success, another study by Angeline (2013) looked at test scores together with assignment submission and grade along with correct response, self-confidence, interest in a certain course, and degree aspirations [5]. M. Abu Tair (2012) Halee analyzed Khan Younis Science and Technology College students' data. Gender and year of birth were only the beginning; they also selected a slew of additional characteristics, such as specialization, year of enrollment, year of graduation, and the city, state or country in which the person resided as well as an address and phone number. Nevertheless, after preprocessing, it was determined that factors including gender, area of study (SSC or SSC), city, and high school grade point average (GPA) were the most critical [6]. There is a correlation between academic achievement and statistics from Anchor Kutchhi Polytechnic in Chembur, Mumbai. A student's family history, parents' jobs, their board of study at SSC, their admission type and the medium and class at SSC were all criteria considered in determining which attributes were most important for admission [7].

At first, we looked at 24 different traits, but in the end, only those with the highest rankings were taken into account for classifying the data. CGPA, arrears, attendance, engineering cut-off, learning time, travel time, health, and so on are some of the criteria considered. Students' gender, parents' education, financial status, living location, medium of teaching, and family status all play a role in their placement and graduation. We found that in the majority of cases, these factors had a direct impact on students' placement and graduation. Personal, familial, academic, institutional, and social qualities are all subcategories of each other. For the first time in our history, we've introduced another element or criterion to ensure that the student will either obtain a job or not. The teachers and students will have an easier time keeping track of the number of students who are eligible for placement and have successfully completed their studies with the addition of these two features.

B) METHODOLOGY

In this segment, we'll go into great detail on the elements that might help predict a student's success.

Identifying kids who may require early assistance before they miss out on a spot is the purpose of this study. First, we need to determine if this is a classification or regression issue. An method for classifying new observations based on previously trained data is known as a Supervised Learning approach. To classify new data, a software uses the dataset or observations it has been provided to learn about the different types of data. For example, yes or no, 0 or 1, and so on. If we can accurately anticipate whether or not a youngster will be placed, we can provide teachers and others with the information they need to begin intervening early on. In other words, it's an issue of categorization because of the binary nature of the question. Furthermore, if we are attempting to forecast a student's continuous score and assessment, but we just know whether a student was placed or not, it may be deemed a regression challenge. As a result, it is a challenge of categorization. These algorithms include Decision Tree Classifier, Logistic regression, Naive Bayes and K-Nearest Neighbors, as well as Support Vector Machines. Nave Bayes, K-Nearest Neighbors, and Support Vector Machines are some of the other options under consideration for this issue.. An SVM model is better than a linear SVM in this case since there are 23 characteristics and hence the data cannot be linearly divided. For high-dimensional datasets like student data, Non-Linear SVM is a good choice. With SVM, the predictor becomes increasingly accurate as the data becomes more complicated. The nave bayes algorithm determines whether or not a data point falls into a certain category. The speed with which Gaussian Naive Bayes can categorize data was an important factor in the decision to use it in this situation. Because this project comprises categorical variables with various levels, information gain in decision trees is skewed in favour of those features with higher levels, which is why we didn't use Decision Tree Classifier.

Kaggle.com was used to gather the data. There are 325 student records included in this data collection. There are 19 characteristics and their domain values in each entry. The following is list of the qualities.

1. Every student should have a unique identification number.
2. The gender of the student (binary: "F" - female or "M" - male) (binary: "F" - female or "M" - male)
3. The mother's job: the mother's position (nominal: "teaching", "health", "services", "at home" or "other") (nominal: "teacher", "health", "services", "at home" or "other")
4. The function of a father: a father's work (nominal: "teacher", "health", "services", "at home" or "other") (nominal: "teacher", "health", "services", "at home" or "other")
5. why this school was picked out of the others (nominal: close to "home", school "reputation", "course" preference or "other")
6. Guardian: student's guardian (nominal: "mother", "father" or "other") (nominal: "mother", "father" or "other")

7. Time Travel: (numeric: 1 - less than 15 minutes, 2 - 15 to 30 minutes, 3 - up to an hour, and 4 - over an hour) (numeric: 1 - less than 15 minutes, 2 - 15 to 30 minutes, 3 - up to an hour, and 4 - over an hour)
8. One to two hours a week, two to five hours, five to 10 hours, or more than 10 a week is the range of weekly study time.
9. The number of previous class failures (numeric: n if 1=n3; else, 4)
10. Activities: extra-curricular activities (binary: yes or no) (binary: yes or no)
11. Having access to the Internet at home (binary: yes or no)
12. After school is out, you have some time to yourself (numeric: from 1 - very low to 5 - very high)
13. Going out with friends is a great way to meet new people (numeric: from 1 - very low to 5 - very high)
14. present state of health (numeric: from 1 - very bad to 5 - very good)
15. Absences: amount of school absences (numeric: from 0 to 93) (numeric: from 0 to 93)
16. Networking, DS, DBMS, OS, OOPS, JAVA (Alphabetic: 0 A+, A, B+, B, C, P, F)
17. Qualitative and Quantitative Assessments (QA & VA)
18. Is the student a successful graduate? (Binary: yes or no)
19. Was the student successfully placed? (Numerical: 1 or 0)

Seventy-seven percent of the dataset was used for training, but just 23 percent was used for testing. K Neighbour Classifier, Support Vector Machine, Gaussian Nearest Neighbor, and K-Nearest Neighbour [8].

C) ALGORITHMS

Support vector Machine (SVM)

Classification and regression are also possible applications of this form of supervised machine learning. SVM first learns the data and then divides it into two categories, termed labels, namely YES and NO, depending on whether or not the student requires help. If you are using SVM, a hyper plane is the line that divides the input variable space into yes and no regions on each side[9].

Gaussian Naïve Bayes

Naive Bayes is one of the quickest algorithms for learning new information. In contrast to other discriminative models, such as supplying regression, naive Bayes has the advantage of being faster, therefore we need less coaching information. For each class of input data, the Gaussian NB method determines the mean and standard deviation, as well as the probability for each class.

K Neighbour's Classifier

An unknown data point is classified using KNN by comparing it to its nearest neighbour, which has previously been classified. By specifying the k-value, the number of a sample data point's closest neighbours may be determined, and therefore the class of that data point is established. In light of the time invested here, numerical value-based groups and model-based training are the best options.

Workflow of the project:

1. At first, we make sure that the dataset is in designated format without any null values
2. After selecting the dataset, we have to pre-process the data
3. Next, the dataset should be divided into training and testing sets, with the training dataset being trained using different methods and the testing dataset being tested afterwards.
4. We must train the model in such a way that considering the set sizes as 100,200,250 we have to train the model with 100 set size at first and then 200 and then 250
5. Then final F1 Score should be compared between various algorithms and conclude the highest accuracy

Conclusion

We'll look at the most current studies in the subject of student success prediction in this section. Using the most accurate prediction methods and the most important aspects that might effect a student's performance, this meta-analysis has been carried out. Students' progress is gauged using the F1 score. It is possible to rate the F1 score from 1 to 0, with 1 being the best. F1 scores were acquired using data from 100 students, 200 students, and finally 250 records of the same model as the previous tests in order to find the best model. F1 scores are generated for each sample size considered. Figure 1 shows the F1 score of the support vector machine. In this, the F1 score is 0.87 for a sample of 100 records, 0.8439 for a sample of 200 recordings, and 0.851 for a sample of 250 records. This figure shows that the decision tree's performance decreases until a particular input limit is reached, and then it rises unexpectedly around the original score.

Table 1: Training and its prediction time of support vector machine

Training set size	100	200	251
Training Time	0.008	0.009	0.016
Prediction Time	0.006	0.014	0.023

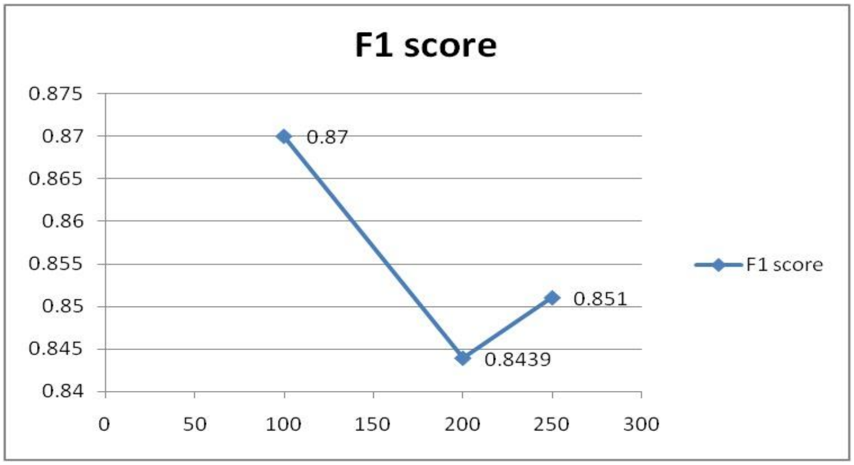

Figure 1: F1 Score for svm

F1 score increases from 0.5666 to 0.7590 for a sample of 100 records in the KNN approach and from 0.66866 to 0.66860 for a sample of 250 records.

Table 2: Training and its prediction time of K Neighbours classifier

Training set size	100	200	251
Training Time	0.005	0.006	0.006
Prediction Time	0.011	0.019	0.022

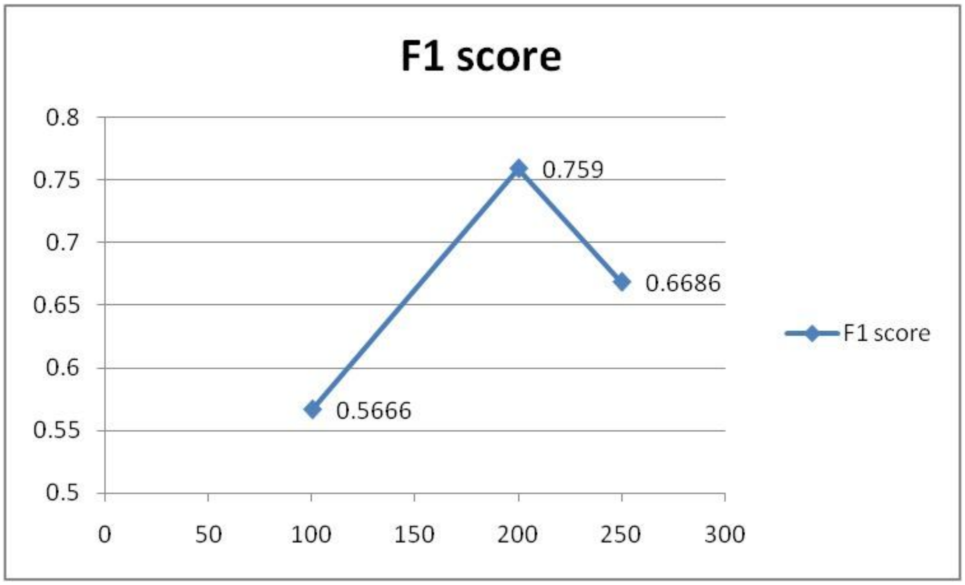

Figure 2: F1 Score for KNN

Naive Bays uses a sample of 100 recordings to calculate the F1 score, which rises to 0.854 for a sample of 200 records and to 0.87206 for a sample of 250 records. It is clear from this graph that Naive Bayes' performance improves with the size of the data collection.

Table 3: Training and its prediction time of Naive Bayes

Training set size	100	200	251
Training Time	0.006	0.006	0.007
Prediction Time	0.004	0.006	0.005

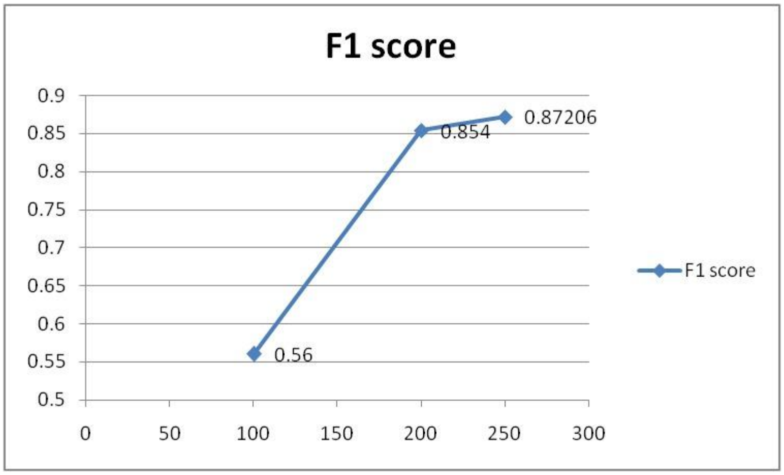

Figure 3: F1 Score for NB

F1 score increases from 0.5666 to 0.7590 for a sample of 100 records in the KNN approach and from 0.66866 to 0.66860 for a sample of 250 records.

Finally, the identical input data are used to compare all machine learning methods, and a comparison table is generated. For each method, the F1 score is shown in Table 4 along with its accuracy for both the training and testing datasets.

Table 4: Comparing SVM, KNN, Naïve Bayes

Algorithms	Support vector machine	Gaussian Naive Bayes	Neighbours Classifier
Training set size	250	250	250
F1 Score (Training Set)	0.8463	0.87027	0.77083
F1 Score (Test Set)	0.8689	0.50746	0.7244

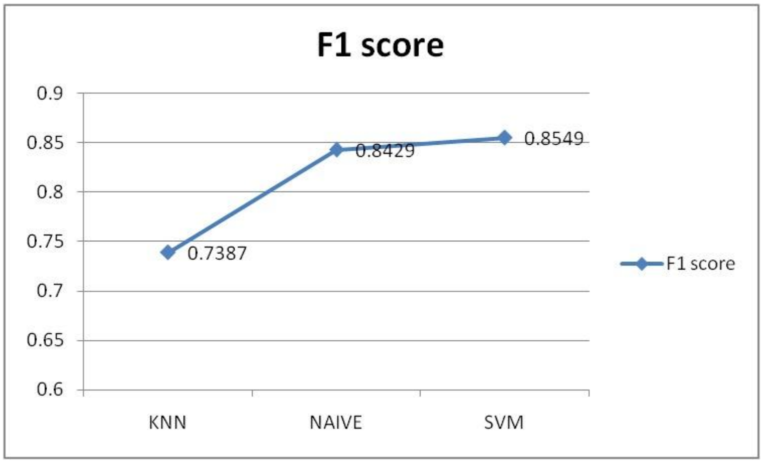

Figure 4: A comparison of student intervention system final F1 scores

You may see a visual depiction of every algorithm's F1 score in this chart. Using a support vector machine, the researchers found the greatest results. With more students, support vector machine is the best model since the prediction accuracy likewise rises with the number of students. Researchers in this article first proposed using machine learning to track down students' academic progress and improve the educational environment. These methods may be used by administrations to improve course outcomes and student performance. This study may be used by professors and managers to enhance performance. Other than that, this kind of learning helps administration to improve policies and plans as well as improve quality of the system itself. Students' learning activities may be predicted using Support Vector Machines, KNN, and Nave Bayes algorithm approaches. We hope that the data we've acquired will be useful to educators and students alike. To increase educational quality, this research takes the necessary actions at the right time and helps students perform better. There will be more students at the first-year academic level in the future that can be studied using this kind of study. Students' first-year academic performance was the focus of this research, which analysed the impact of family background characteristics and prior scholastic successes on the student's first-year academic performance.

REFERENCES:

[1] "Student Intervention System Using Machine Learning Techniques," a paper by Kshtij Gupta and Shubhangi Urkude was published in the International Journal of Engineering and Advanced Technology.

[2] Institutional research may serve as a bridge between academics and the general public.

[3] Oktariani Pratiwi Oktariani Nurul Teaching, Assessment, and Learning in Engineering (TALE) Conference

[4] The author of a recent literature review on the subject, Abeer Badr, suggests that data mining methods may be used to forecast student success in the classroom.

[5] The SIJ Transactions on Computer Science Engineering & its Applications (CSEA) 1 (1) (2013) p12–16 uses an a priori approach to construct association rules for student performance analysis.

[6] In February 2012, Mohammed M. Abu Tair and Alaa M. El-Halees performed a study on the application of data mining methods in education to enhance student performance.

[7] Volume 4 Issue 1 of the International Journal of Recent and Innovative Trends in Computing and Communication (ISSN 2321-8169) "Jyoti Banshee," as in Predicting Students' Academic Performance by Mining Educational Data

[8] These are the 11 most prevalent algorithms for machine learning. Presented by Soner Yildirim Soner, https://towardsdatascience.com/11 most popular machine learning methods in a nutshell.

[9] In the survey conducted by Ashis Pradhan Support Vector Machine.

[10] Anatol A. Heidari et al. provide the theory, literature study, and implementation of the Ant Lion Optimizer (ALO) in multi-layer perceptron neural networks (MPNs). The book is titled "Nature-Inspired Optimizers" and will be published by Springer in Cham, Switzerland, in 2020.