Predicting students’ performance using Machine Learning Techniques

D.VijayAnand, M.Deepika, C.S.M.Nandhini, N.Nishanthini

Abstract—The major task for universities is to thoroughly assess their performance, pinpoint their distinctiveness, and devise strategies for future growth and success. The educational system is aware of the potential for machine learning to significantly enhance the performance. The ability to predict student performance with a high degree of accuracy is helpful because it allows for the early identification of children who do poorly academically. In order to increase the performance of the students prediction is used to analyse how well the students are studying. In this paper we proposed the student performance prediction system using Machine Learning. The model is trained and tested with data-set using different algorithms such as Logistic Regression, Support Vector Machine and K-Nearest Neighbor were evaluated from among the various machine learning techniques to forecast and evaluate their performance in terms of accuracy and compared the accuracy of all algorithms used. Among three algorithms Support Vector Machine outperformed with 84% as accuracy.


I. INTRODUCTION

In the education system wants to make the students to improve the performance in studies. Machine Learning techniques can be used to predict students’ performance. As the techniques would help students improve their scores and it would allow teachers to identify individuals who might need additional assistance during performance. The performance of students in academics is very important, yet due to a lack of interest in the class and insufficient in the subject matter, many students may do well on the exam but fail or receive low grades. The students were unable to complete their courses, nevertheless. Machine Learning techniques can be used to predict students’ scores in various courses and that would help in their studies. The techniques would help students to improve students’ scores based on expected grades and it would allow teachers to identify individuals who might need additional assistance in the courses. Predictive analysis of the application has become eagerly looked after in higher education.

For all educational levels, for predictive modeling advanced analysis that would include in machine learning for deploying high-quality performance and useful data. Most individuals are aware that one of the major performance measures that can aid trainers in keeping track of their students’ academic progress is their grade. The variations are huge and have been countless of machine learning techniques in the education area over the last decade. To improve the accuracy of predicting student grades, process the imbalanced datasets to bring a significant barrier. This study compares multiple resampling methods to manage the unbalanced data problem. The distinction between multi-class and binary classification, as well as feature structure, are looked at. In the project a variety of machine learning classifiers, including K-Nearest-Neighbor, Support Vector Machine and Logistic Regression to be able to better access how well resampling methods perform in solving the imbalanced problem. The major goal is to decrease the rising failure rate by predicting students’ performance level and giving them the necessary training.

II. RELATED WORKS

A. Ghorbani (2021) [1] describes an experimental study of both supervised and unsupervised learning methods were used to conduct which gives higher prediction. Increases classification accuracy while minimizing the number of selection features. A. I. Adekitan (2020) [2] comprises the study of determine the relationship between extracurricular activities and student performance. Co-curricular activities are extracurricular activities that support and enhance the academic or core curriculum. E. Fernandes (2020) [3] describes the The main focus has been the detection of academic performance and dropouts, which are often related to the notions of success or failure in the courses. A. B. El Din Ahmed (2021) [4] shows a comprehensive analysis of machine learning techniques to predict the final student grades in the first semester courses by improving the performance of predictive accuracy. C. Máñuez-Vera (2019) [5] represents the scores of four courses served as key indicators to divide students into different groups. Through observing the dataset in each cluster, the authors reported that students tended to show the same kind of scores in all courses.
III. PROPOSED METHODOLOGY

Here, the downloaded student dataset from UPI Repository is used to make predictions whether pass or fail. The dataset contains student records with 31 attributes including school, sex, age, address, familysize, Pstatus, Medu, Fedu, Mjob, Fjob, reason, guardian, traveltime, studytime, failures, schoolsup, famsup, paid, activities, nursery, higher, internet, romantic, famrel, freetime, goout, Dalc, Walc, health, absences, and passed.

**Data Pre-processing**

Feature Selection – The process of reducing the column by using relevant data and get rid of noisy data. The process makes the data warpper and filtering.

**Classification**

Three different classification algorithms are used to analyse the dataset: K-Nearest Neighbor (KNN), Support Vector Machine (SVM) and Logistic Regression (LR).

**K-Nearest Neighbor** - The K-Nearest Neighbor is a popular learning technique that classifies an instance based on the instances that are similar to it. K-Nearest Neighbor is easy to implement and the calculation time is short. However, it is mainly used for classification predictive problems.

**Support Vector Machine** - Support Vector Machine, or SVM, is used to solve Classification and Regression problems. However, it is largely employed in Machine Learning Classification issues. When using the SVM algorithm, each data point is represented as a point in n-dimensional space (where n is the number of features you have), with each feature's value being the value of a certain coordinate. Then, we carry out classification by identifying the hyper-plane that effectively distinguishes the two classes. All the performance measures are calculated and tabulated for all these three algorithms for the dataset individually.

**Logistic Regression** - The likelihood of the target attribute is ascertained using the supervised approach technique of LR. Ordinal, interval, or ratio-level independent variables are either true or false, 0 or 1, etc. as a result of this regression. The distinction between a dependent variable and additional independent variables is described and explained using LR.

IV. RESULTS AND DISCUSSION

The simulation is done with the Ansys-HFSS using the appropriate design dimensions given in the Table 1. The return loss (S_11), voltage standing wave ratio (VSWR), gain, and radiation pattern are the criteria used to evaluate the planned antenna's performance.

**Parameters for Evaluation**

Three prediction models were used in this experiment to predict performance. Here, K-Nearest Neighbor, logistic regression, and SVM are employed as classification algorithms. The parameters which were used are Precision, Recall, F1-Score and Accuracy.

**Precision** - The proportion of accurately categorized positive samples (True Positive) to the total number of positively classified samples is known as precision (either correctly or incorrectly).

\[
\text{Precision} = \frac{TP}{TP + FP}
\]

TP: True Positive - Observation rightly detected as positive
FP: False Positive - Observation wrongly detected as positive

**Recall** - The recall is determined as the proportion of Positive samples that were correctly identified as Positive to all Positive samples. The recall measures how well the model can identify positive samples. The more positive samples that are identified, the larger the recall.

\[
\text{Recall} = \frac{TP}{TP + FP}
\]

FN: False Negative - Observation wrongly detected as negative
F1-Score - The weighted average of Precision and Recall is the F1 Score. Therefore, both false positives and false negatives are considered while calculating this score. Although F1 is generally more beneficial than accuracy, especially if you have an uneven class distribution, it is not intuitively as simple to understand as accuracy. When false positives and false negatives cost about the same, accuracy performs best. It is preferable to include both Precision and Recall if the costs of false positives and false negatives are significantly different.

\[
F1-Score = \frac{2 \times (Recall \times Precision)}{Recall + Precision}
\]  

Accuracy - The easiest performance metric to understand is accuracy, which is just the proportion of properly predicted observations to all observations. One would believe that if our model is accurate, it is the best. Yes, accuracy is an excellent indicator, but only when the values of the false positive and false negative rates are nearly equal in the datasets. As a result, you must consider other factors while assessing the effectiveness of your model.

\[
Accuracy = \frac{TP+TN}{TP+FP+FN+TN}
\]  

TN: True Negative - Observation rightly detected as negative

Performance of K-Nearest Neighbor

Fig. 2: Confusion matrix – K-Nearest Neighbor

Fig. 3: ROC Curve – K-Nearest Neighbor

Performance of Support Vector Machine

Fig. 4: Confusion matrix – Support Vector Machine
The Table 1 shows the accuracy, precision, recall and F1-Score of various classifier models. It indicates that the maximum accuracy was achieved by the Support Vector Machine classification model.

### Table 1 Comparing Three Algorithms

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Accuracy (%)</th>
<th>Precision (%)</th>
<th>Recall (%)</th>
<th>F1-Score (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM</td>
<td>84</td>
<td>91</td>
<td>78</td>
<td>82</td>
</tr>
<tr>
<td>KNN</td>
<td>80</td>
<td>80</td>
<td>99</td>
<td>88</td>
</tr>
<tr>
<td>LR</td>
<td>64</td>
<td>71</td>
<td>24</td>
<td>36</td>
</tr>
</tbody>
</table>
V. CONCLUSION

Machine learning techniques are concentrated on improving predictive models to achieve better accuracy in predicting successful attributes. This method has the amazing capacity to raise categorization and forecasting precision. SVM is considered as the best choice therefore it provides better accuracy than the other two models. In future, the developers can add any type and number of additional categories with better scalability on datasets. If the dataset, which is going to be collected in the future, has records of students’ from various countries all over the world, then its trained model will be more efficient than the proposed one. Moreover, the larger the dataset, more will be the benefit to be gained. As a result of that, the model may be recommended to more users irrespective of geographical locations and its conditions.

REFERENCES