Prediction of students’ performance in education system based on artificial intelligence

Pooja Pathak¹, Nazia Farheen² and Avinash Dubey²

¹ Department of Mathematics, GLA University, Mathura, UP, India
² Department of Computer Engineering and Applications, GLA University, Mathura, UP, India

Abstract. The main objective of any educational system is to provide the best knowledge to students. To achieve this goal this is important to identify the weak students who need more support and take correct decisions to improve their performance. In this research for predicting the students’ performance, four techniques of machine learning are used. For this technique, we take the data from computer science students of GLA University, Mathura, U.P. INDIA. These machine learning techniques include various processes such as Artificial Neural Networks (ANN), Naive Bayes (NB), Logistic Regression, and Decision Tree. In this model, we put more efforts to know the time attended by the students on the internet for learning and social media. Also, various measurements have been done such as precision, F measure, recall, and classification errors. We used the dataset for building a model depending upon the survey that given to the all-computer science students and grade copy of students. The decision tree identified four main attributes that influence the performance of students a lot. This helps us to achieve an accurate prediction of around 98%.

1. Introduction

Education, the basic need of everyone to explore this huge universe with their small body capacity. Education plays an important function in motivating ourselves and also in contributing important components in our daily life. Education, the most essential part of our life which exists everywhere and in every moment of our daily life. For this, we made huge buildings to teach peacefully to make this world a better place. Since starting, we just want to improve our education technique in which prediction of student's performance plays an important role. Like many countries, India also provides educational loans to that student who is unable to pay the fees on time. If any student fails to attempt his graduation on time then it gives extra expenses to the government, so, yes it also affects our economic system.

As said Prevention is better than cure, the government confirms to graduate all the students on time to avoid all extra expenses. The technique of ML is used to predict the performance of students and also help to find the student who is at risk as soon as possible. The most important point is to select attributes or features used as input. This input can be grades of the test, end term exam, sports activity, cultural activity, educational background.

In this research, authors focus on the result of using the internet by students for learning and skill development on students’ performance.

2. Literature Review

Dorina proposed a model in which binary class (0, 1) was used as predicting the chances of success and failure of the student. This model was constructed under the CRISP-DM research approach. With the help
of classification algorithms such as one, MLP, J48, and IBK in the dataset, the highest accuracy was achieved at 73.59%. The drawback of this model to work with high dimension data and balance the difficulty of classes. Edin Osmanbegovic et al. [2] by lessening the data dimensionality problem. In this study NB, MLP and J48 were evaluated as machine learning classifiers. This method gained the highest result accuracy of 76.65%. The drawback of this model is unable to handle the class imbalance problem. Carlos et al. [3] use ML to reduce the problem of data dimensionality and class imbalance. In this data set, ten classifiers were applied and achieved the highest accuracy of 92.7%. The drawback of this model is that it is unable to handle the varying student's standards at every educational stage and tested failure. Ajay et al. [4]. He introduced a new factor known as CAT. It is a social factor that describes how studies are affected by the Indian social status which is divided into four groups. It shows the highest accuracy of 82%. In Alaa Khalaf et al. [5]. model, three types of Decision tree classifiers Reptree, J48, and holding tree are used. This model achieves the highest accuracy of 91.47% by Reptree. Data high dimensionality and class balancing problem is the drawback of this model also. Dech Thammasiri et al. [6] gives a research model that classifies the poor performance of new students. The Three Balancing Method with four classification techniques is used to resolve the problem of class imbalance. This model achieved the highest accuracy of 90.24%. Camilo proposed a model by overcoming the problem of class imbalance to predict the students’ performance. The algorithms of Naive Bayes and decision trees are used in this model. This model got the highest accuracy of 85%. Bilal proposed a prediction model which main aim is to find failure student. With the help of it we can identify early at-risk students. This model uses the most important attributes and got the highest accuracy of 79.23%.

3. Proposed Model

The first step to collect data from the source of data. In our research, the data source is collected from the grade book of students and by the given survey to the students. The next step is to mine all the data and collect all the raw data. The third step is to use machine learning algorithms that build the classifier, then in the fourth step trained classifier classified all the data and predicted the student performance as shown in figure 1. All the important points are shown below-

1. To collect the data.
2. Data preprocessing and labeling
3. Machine learning algorithms
4. Trained Classifier
5. Predicted label
4. Methodology

In this part, we describe all the techniques of machine learning used in this model.

4.1. Data pre-processing

It is an important part of Data Mining. The conversion of unprepared data into an easy form that can be easily understand by the mining algorithms is the main aim of data preprocessing.

There are several tasks performed in data preprocessing defined below-

4.1.1. Data integration: - Collection of data from various sources is the main aim of data integration.

4.1.2. Data cleaning: - This task helps to achieve the missing data and handle housing data.

4.1.3. Discretization: - This mechanism is used to convert all the nominal data into the desired numerical data.

4.2. Artificial neural network

In ANN the set of output and input units are connected by heavy connections as shown in figure 2. Backpropagation is the most famous algorithm used in ANN.

It has various advantages given below-

- High resistance to the dataset’s noise.

- On patterns of classification.
Easy to use with a less knowledge of the connection between the features and class labels in the datasets.

All the main layers such as an input layer, hidden layer and the last output layer are shown below in figure 2.

![Architecture of ANN](image)

**Figure 2. Architecture of ANN**

### 4.3 Decision Tree

It is a tree-like structure which is identical to a flowchart. In this method, the branches demonstrate the result of the test where the node represents a test on attributes of data. The target feature represents each node of the leaf and root node symbolized by the upper first node of the tree. In a Decision tree, there is no need to know the domain of difficulty that makes it a very popular classification technique. It is easy to convert in classifications rules and understand. During the construction of the decision tree, the algorithm uses a measurement of attribute selection that is used in picking the form that helps to divide the illustration of the dataset into different target classes at the beat level.

### 4.4 Logistic regression

It is the suitable expulsion analysis to conduct when the dependent variable is binary also known as **Dichotomous**. Logistic regression is also a predictive analysis like all expulsion analyses. The main function of logistic regression is to illustrate the relationship between one binary variable that is dependent and one or more ordinal, nominal, ratio-level or, interval that are independent variables and to depict the data. Logistic regressions are not much easy to interpret sometimes; the tool name **Intellects Statistics** allows us to do the analysis, then the output is interpreted in plain English.
4.5. Naive Bayes

Naive Bayes is a technique of classification based on Theorem of BAYES in which there is an acceptance of originality among all predictors. In easy words, a NB classifier imagines the existence of a specific property in a class that is not related to the existence of any other feature. Each instance y in the dataset restrain the value of the attribute. NB model circumstantially useful for huge data sets and very convenient to build. Along with naivety, Naive Bayes is also famous to submerge even highly complex classification methods.

5. The Experiment

All the basic processes that we are going to do are shown in the above figure 3. The dataset we used in our research is collected from our GLA university of the academic year 2018-19. We used two data sources, a survey collected from the computer science students, and a grade book record of that students. This Dataset contains a record of 150 students and 20 attributes. Each student is labelled as a good and bad student as per their grade marks. If the student got 60 or less than 60 out of 100 then that student is considered a Weak Student. The student got 60 grades out or more than that out of 100 that student considered as a Good student. There are 86 students marked as good and remaining marks as 64 students mark as weak students. Our main motive is to identify weak students so we target it as a positive and good student as negative. Now, CS-grade 1 attributes represent the average of the first and second midterm exams which are held before the end term examination. The academic year contains two end term examinations. So, to get early notification of the students who need early support, we use the CPI of the first semester and the average of the first two midterms of the second end term exam taken as an attribute because it is the basic prediction of students’ performance in the final exam.
With the help of these poor students get an early chance to improve their study habits, materials, and time table, etc. as well as teachers also get the chance and time to give more effort and efficient ways to improve their skills.

This is all about academic performance. Now we will work on a job and for the job, the most important skill is communication skills in a proper way and exactly correct English knowledge. So now we take another attribute named English-grade 1 that represents marks in English in the first semester.

Now the next one is doing activities in colleges. There are many clubs in GLA University Mathura like a dance club, singing club, culture club, and many more that organize various functions on different occasions. In these functions students that participate in it and the students who are part of this club manage all things together that build a group working skill in them and many more. So we take culture-grade 1 as an attribute.

According to the dataset 73 students who are working on these clubs therefore 48.66% of 150 students are taking part in these activities and 43.1% of them are weak students.

There are a various department in GLA UNIVERSITY MATHURA-

- Computer Science and Engineering
- Electronics and Communication
- Electrical Engineering
- Mechanical Engineering
- Civil Engineering
- English
- Physics
- Chemistry
- Biotechnology
- Mathematics
- Agriculture Science
- Physical Education

We took data of only computer science students in which 113 students, out of 150 are good, and the remaining 37 students are weak, meaning overall 24.67% of students are weak. The policy of the college to place the students in various sections of different departments based on their CPI/GPA they have scored in their high school and intermediate, so we take department as an indicator attribute.

The lifestyle of all students also affects the performance of them. Fernando et al. [7] found in his study that there is an earnest connection between the High academic performance and formal sports activities.

5.1. Data Pre-processing and machine learning software

In the dataset value of each attribute has been calculated by deducting the mean of attribute from it and divided by the difference between the highest and lowest attribute value.
5.2. Method of Validation

In this research, we used the method of three folds cross-validation. In this process, the dataset has been divided into three sets of same size. Testing and learning are carried out three times. In this method at each carry out process, the algorithms of ML take only one set as a test set and another two sets as a training set.

5.3. Accuracy and measurement of performance

To figure out the final accuracy and performance of the model, the measurement of performance and the accuracy is agglomerated over all the folds. The ROC index is used to measure the accuracy of this model by comparing it with other models. With the help of the ROC index, many other useful measurements are done such as F measure, classification errors, and accuracy, the most important measurement. Larger the ROC index, better the classification model. The measurement has criteria. If the Roc index of a model is higher than 0.7 then this model is presumed a strong model whereas if a model has a ROC index less than 0.6 then it is presumed as a weak model. [8]

\[ F \text{ Measure} = \frac{2 \times (\text{Precision} \times \text{Recall})}{(\text{Precision} + \text{Recall})} \]

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

\[ \text{Recall} = \frac{TP}{(TP + FN)} \]

TP stand for true positive which shows the number of data in any set of test as positive target that also had a positive target. TN stands for true negative which predict negative target data of any test set as negative. FP stands for false positive which predict negative data target of any set as positive. FN stands for false negative which predict positive target of any set of test as negative. The conclusion is TP and TN are showing correct result while FP ad FN are like errors. The final value of TP, TN, FP, FN of all four models are given in table 1 as shown in result.

6. Result

Authors have created and tested four classification methods using four techniques of machine learning that are ANN, Logistic regression, Naive Bayes, and Decision Tree. After working in all four techniques with our collected data as shown in table 1 we get the highest accuracy of 94.71\% of ANN model techniques. Artificial Neural Network model has the highest ROC index of 0.899 and the least classification error of 5.29\% while logistic regression gets the lowest ROC index of 0.650 with the highest classification error of 24.51\%. The decision tree model gets a ROC index of 0.735 with the classification error of 14.63\% and Naive Bayes gets the ROC index of 0.765 with the classification error of 22.87\%. Hence the most accurate method is the ANN technique. Even though the decision tree model has not the highest accuracy but it helps us find that not all attributes affect the classification of student’s performance. There are only four main attributes that affect the performance of students are Computer Grade Course-1, interest in Computer, the time they do for practice, and environment of study in their home as well as in the hostel. Therefore, to improve the performance of weak students all the faculty members should focus on these attributes and also
discuss these problems with the parents of those students who are living in their homes not in the hostel because not every home can able to make the environment of study.

Table 1. The measurement of accuracy and performance of all four models

| Model               | TP  | FP  | TN  | FN  | Precision | Recall | F-Measure | Accuracy | Classification Error |
|---------------------|-----|-----|-----|-----|-----------|--------|-----------|----------|----------------------|
| ANN                 | 69  | 12  | 60  | 11  | 0.85      | 0.86   | 0.85      | 94.71    | 5.29                 |
| DT                  | 69  | 13  | 59  | 09  | 0.84      | 0.88   | 0.86      | 85.37    | 14.63                |
| Logistic regression | 62  | 18  | 49  | 21  | 0.77      | 0.74   | 0.75      | 75.49    | 24.51                |
| Naïve Bayes         | 65  | 15  | 48  | 22  | 0.81      | 0.75   | 0.78      | 81.23    | 22.87                |

To enhance the performance of student’s faculty members should give more attention to weak students, make their subject more interesting, and always ask them for their problems if they get any difficulty and try to help them.

7. Conclusion

Nowadays every educational institute demands an accurate model for the prediction of student’s performance. Data management and selecting a valuable attribute is the most important point. This research article presents a model for prediction of student’s performance based on Machine learning and artificial intelligence. All four techniques of ML i.e. ANN, Logistic regression, DT and Naive Bayes have been applied to build this model. ROC index is used to check the accuracy of the model. The dataset was collected from the students of GLA University Mathura, U.P., India during the 2018 and 2019 academic year, by using the survey of grade books of students. This research model includes data collecting, data preprocessing, creating students’ datasets, and evaluating the best model by checking the accuracy level. This model shows the highest accuracy of 98%.

It is important to recognize and fix your mistakes in order to grow fundamentally. With this high accuracy it can predict performance of students more accurate that can identify student at high risk in minutes in entire university. There are many sightless patterns in data of students that human might not be able to recognise but that computer can easily. Then we have enough time for faculty member to focus on weak student and to assist them.

8. References

[1] Imran, M., Latif, S., Mehmood, D., & Shah, M. S. 2019 International Journal of Emerging Technologies in Learning (iJET), 14(14), 92-104.
[2] Osmanbegovic, E., & Suljic, M. 2012 *Economic Review: Journal of Economics and Business, 10*(1), 3-12.

[3] Márquez-Vera, C., Cano, A., Romero, C., & Ventura, S. 2013 *Applied intelligence, 38*(3), 315-330.

[4] Pal, A. K., & Pal, S. 2013 *International Journal of Computer and Information Technology, 2*(06).

[5] Hamoud, A. 2016 *American International Journal of Research in Science, Technology, Engineering & Mathematics, 16*(1), 26-32.

[6] Thammasiri, D., Delen, D., Meesad, P., & Kasap, N. 2014 *Expert Systems with Applications, 41*(2), 321-330.

[7] Muñoz-Bullón, F., Sanchez-Bueno, M. J., & Vos-Saz, A. 2017 *Sport Management Review, 20*(4), 365-378.

[8] Kelleher, J. D., Mac Namee, B., & D'arcy, A. 2015 *MIT press.*