PREDICTION OF STUDENTS PERFORMANCE

Submitted by

213244

SANTRA JOHNY

213204

AGY SUNNY

Under the guidance

DR MANOJ KUMAR T K

DIGITAL
UNIVERSITY
KERALA

Curating a responsible digital world

CONTENT

1)ABSTRACT

The future of a country depends on its millennials. Nurturing them and helping them grow by taking care of their weaknesses can help them grow. Student's performance prediction is essential to be conducted in all educational institutions in order to prevent the number of failures and to analyse the factors affecting their performance. With a strength of 40-50 or even higher number of students, it will be hard for teachers to analyse and provide the necessary care. Usage of a predictive model can ease the work of teachers by helping them save time and directly help with the cause of the result. This can also help the respective student to improve his or her weak point and hence helping them grow.

The above problem can be minimised by using a machine learning approach in which the data collected is analysed and predicted using a suitable model. Various models like Logistic regression, Gaussian Naïve Bayes, k-nearest neighbours, Decision Tree Algorithm and Random Forest are used. The model with highest accuracy is chosen for analysis of the next dataset when used in day-to-day life. Linear regression is used to predict the performance of the students with the existing dependant factors. Such an analysis can help the students to improve their learning strategies and institutions to identify the most promising students.

Our experiments show Logistic Regression and k-nearest neighbours to be the most accurate analysis of the dataset in prediction of success and failure.

2)INTRODUCTION

The change from manuscripts to paper pulps was a tiring slow journey. But then came the invention of watches and hence time became valuable as well. The drastic drift from conventional learning towards a student centric learning approach is emerging at a faster pace. Conventional learning approach focuses on delivering lecture and student passively absorbing it where the teacher uses different assessments to evaluate the performance of students [1]. This is purely based on the internal marks which classify students into front benchers and back benchers. But as quoted by Dr APJ Abdul Kalam: "The best brains of the nation may be on the last benches of the classroom". Marks should just be numbers. In fact, we need techniques to help students identify and develop their dominant skills.

Machine learning is the branch of Artificial Intelligence that provides ability to automatically learn from past experiences [2]. Machine learning can be classified into two, supervised learning and unsupervised learning.

In this project we are mainly focusing on supervised learning, more specifically on predictive analysis. Predicting academic performances of students help the instructors to understand their potentials and nurture them. This will also minimise the system of classifying students barely based on their marks. Marks should just be numbers rather than a parameter used to judge the future of students. The performance of a student depends on external factors like the environment he or she was brough up in, parental level of education, grasping power etc. Getting someone to look into all these factors might be difficult but instead common factors like first exam marks, parental level of education can be used to predict the future performances. It also helps to identify those who are likely to drop out from the course and help them modify the system. This prediction can also to identify students who need special guidance and care.

Machine learning models helps the teachers to identify the weak points of students by enabling an early warning system. In this project, we have applied different machine learning algorithms on the historic results of secondary school students to find out the prediction accuracy. Various models like Logistic regression, Gaussian Naïve Bayes, k-nearest neighbours, SVM and Random Forest are used.

3)PROBLEM STATEMENT

Traversing through various articles we understood the importance of not judging a student solely using his or her marks but rather understand other factors that would help them in expanding their potentials. As mentioned, it is not easy for a person to personally go through each of these decision factors and find a prediction and hence rising the demand for a proper algorithm. Taking all these factors into consideration we are trying to find an appropriate model that will give highest accuracy while predicting the model.

Hence, we use different machine learning algorithms like SVM, K-nearest neighbours, Naïve Bayes, Random Forest, Logistic and Linear Regression to compare the accuracy in prediction.

4)LITERATURE REVIEW

Here, we review about previously referred research papers and thesis papers for the organizing our project.

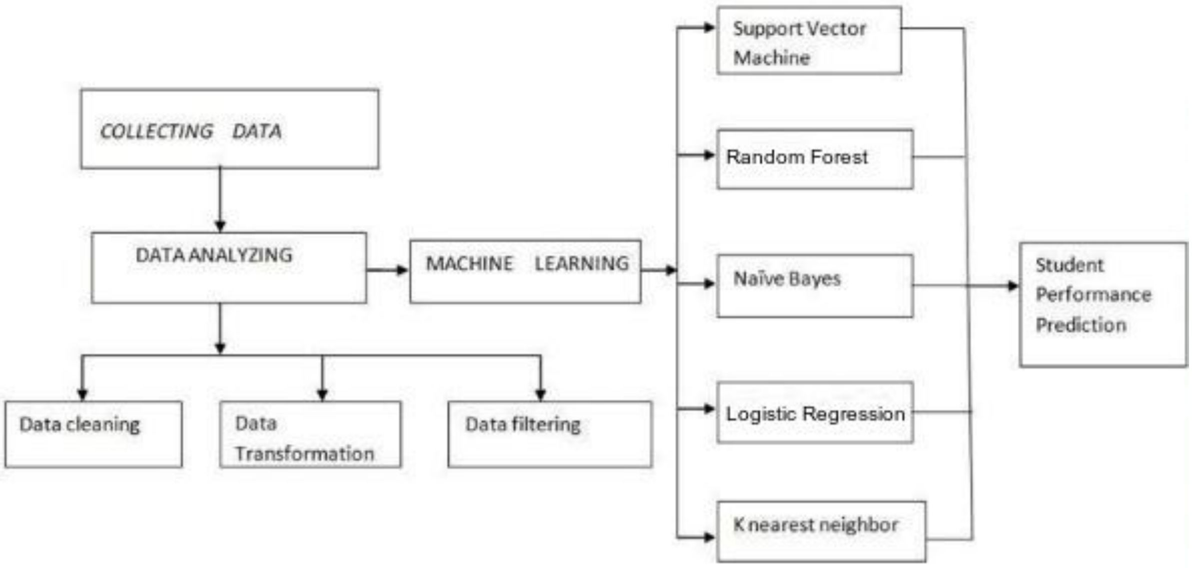
The first paper [1] uses the data of students, taking one of the undergraduate courses across the batches of 2016, 2017, 2018 and 2019.In this paper they implement Logistic Regression, Linear Discriminant Analysis, Support Vector Machine, Naïve Bayes, Decision Tree and K- Nearest Neighbour etc. Among these models LDA (0.81) and SVM (0.80) has the best accuracy.

The second paper [2] uses a sample data, which contains information regarding 100 students. They implemented linear regression to train the dataset. Linear regression is divided into two, simple linear regression and multiple linear regression, in this thesis they have implemented multiple linear regression and found that academic performance of students depend not only on their learning strategies but also on their family background and other attributes.

The third paper [3] uses the dataset which is collected through a questionnaire. A comparative approach is used in this thesis. They have applied Support Vector Machine and Naïve Bayes for comparing the dataset. They have selectively compared all the problem faced by students, teachers, administrators etc for the readers to understand the need of a machine learning algorithm which can predict student’s performance. As per this thesis, Naive Bayes algorithm (~ 0.92) has higher accuracy and less execution time compared to Support Vector Machine.

5)METHODOLOGY

Main purpose of methodology is to implement various models in greater depth to the dataset. Prior step is to collect an appropriate dataset which can be used for analysis. This step is called data collection. After data collection we transform the data into desired format by removing null values and outliers, dropping columns, filling null values with appropriate exploratory data analytics tools. This process is called data cleaning or data pre-processing. Basic EDA is done to remove null values and to find information about the data. Unnecessary data columns are removed as they do not help in concluding or prediction.



After which we start off by taking the average score of 'G1', 'G2', 'G3' and creating a column called "mean_score". This column is then classified into 4 parts as { 'fail':1, 'sufficient':2, 'satisfactory':3, 'good':4, 'excellent':}].

0.0 < Average score <= 9.5 is classified as fail.

9.5 < Average score <= 11.5 is classified as sufficient.

11.5 < Average score <= 13.5 is classified as satisfactory.

13.5 < Average score <= 11.5 is classified as good.

15.5 < Average score <= 20 is classified as excellent.

In Fig.5.1 we can see the plot of average score(score_eval) on x-axis and its count on y axis.

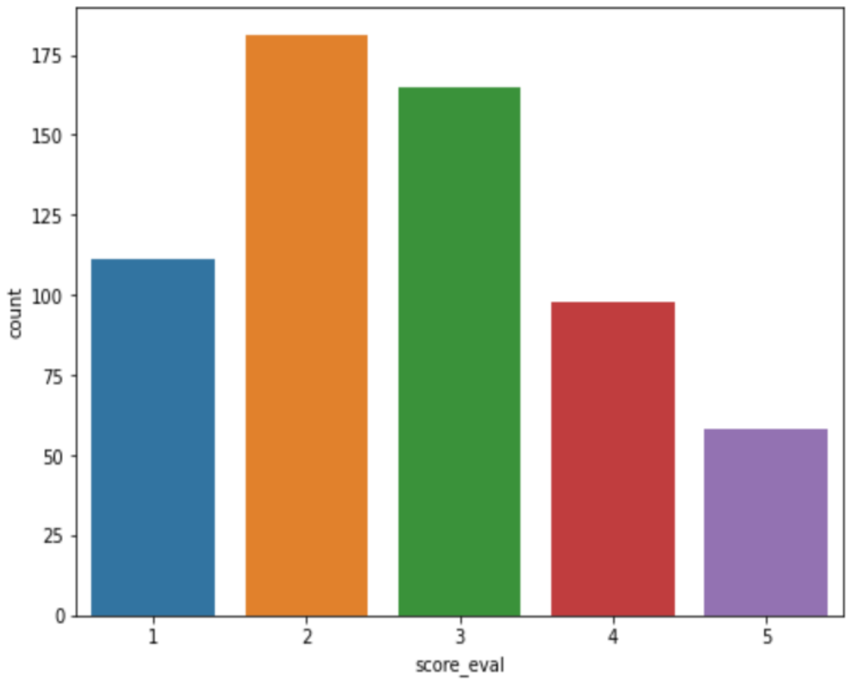


Figure.5.1

The main points that are drawn are:

1) The number of females who fail is greater than that of males but they do better overall

2) The performance of students in urban area is better than that in rural area

3) The ‘reputation’ can be considered as the main factor to choose a ‘good’ student.

4) This dataset disproves the common theory of students who are bookworms have better grades. We see that students who engage in various extracurricular activities have better grades.

In Fig.5.2 a boxplot of score_eval v/s mean_scores is plotted which shows that the there exist a few outliers below the minimum value of failures.

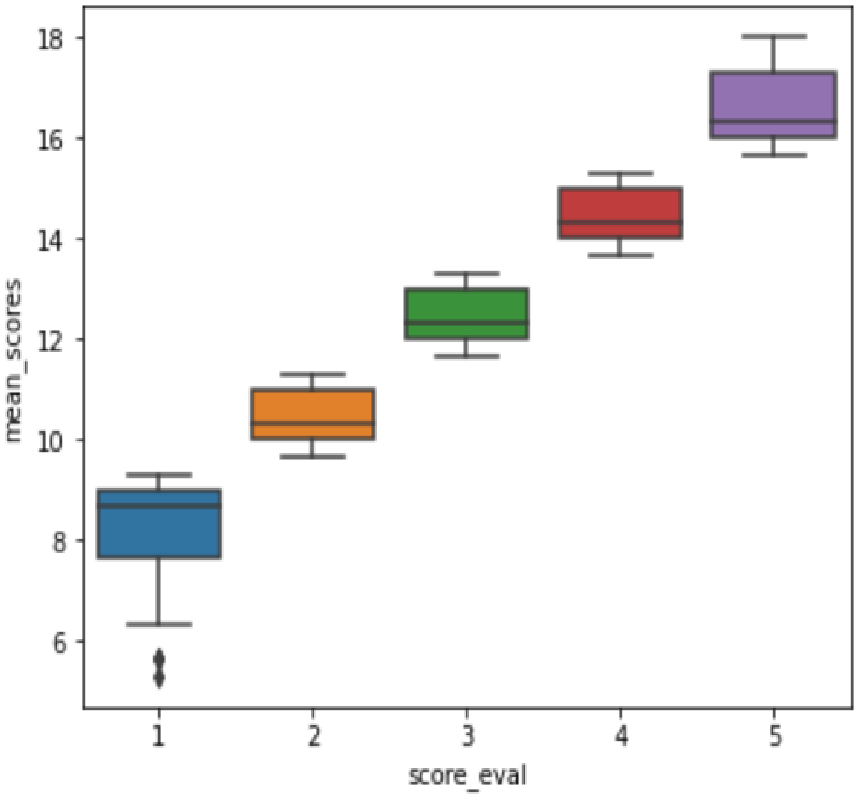


Fig.5.2

Outliers makes the data unstable and can cause large variations while predicting. Methods like Standard Deviation Method, Interquartile Range Method, Automatic Outlier Detection are used to remove outliers. Since there exist outliers, the next task is to remove the outliers. For this Q1 and Q3 is found. Using these values an upper limit and lower limit is found. Any value outside this range is considered as an outlier and is removed.

The presence of categorical data can make it hard to do the prediction model. On observing the data carefully, we can see that most of the columns are either of yes or no type or 0 and 1 or can be classified into different groups. Therefore, we convert these values into 0’s and 1’s or to numbers in the rang 0 to 5, thus converting the whole data into numerical data. The obtained dataset is then further used for splitting and training the model.

Modelling of the dataset is done using models such as Logistic Regression, K-nn, SVM, Naïve Bayes, Random Forest and AdaBoost.

6)ALGORITHMS

5.1) LINEAR REGRESSION

Regression analysis is a technique used in statistics for understanding the relationship between variables. It is a machine learning algorithm, which works on supervised learning. It helps to find out the relationship between variables and perform regression tasks. Linear regression is a model which is used to find the relationship between two variables by fitting a linear equation to the data. Here, one variable is considered to be an independent variable and the other is a dependent variable. Generally, x is considered as the independent variable and y as the dependant variable. Before fitting a linear model to observed data, we should check whether there is relationship between the variables of interest exist or not This can be identified by looking at the correlation between the variables. Regression technique helps to finds out a linear relationship between input and output.

The general format of a linear regression line is given by $y = \beta_0 + \beta_1 x + \varepsilon$.

Since the usage of x as independent and y as dependant may cause a confusion x is referred to as the predictor or regressor variable and y as the response variable. Since the above equation has only one regressor variable, it is called a simple linear regression model. The slope of the line is β_1, and β_0 is the intercept (the value of y when x = 0)

5.2) LOGISTIC REGRESSION:

Logistic regression is also a machine learning algorithm which comes under supervised learning. It is used for both regression and classification but mainly focuses on classification. Logistic regression apart from linear regression predicts whether something is true or false instead of predicting if something continuous like height. This algorithm predicts uses the given set of independent variables to predict the categorical dependent variable. Sine the model predicts the output of a categorical dependent variable, the outcome must be a categorical or discrete value. It ranges from Yes or No, 0 or 1, True or False, etc.

Both Logistic Regression and Linear Regression are same except for the method of implementation. As stated above Linear Regression is used for solving Regression problems, Logistic regression focuses mainly on the classification problems.

One of the biggest differences between Logistic and Linear regression is how the fitting of a regression line to the data is done, we fit an ‘S’ shaped logistic function to predict the output. It’s ability to classify datasets using continuous and discrete measurements makes it widely used machine learning method.

5.3) K- Nearest Neighbour:

K-Nearest Neighbour is one of the fundamental Machine Learning algorithms which works on Supervised Learning technique. Identification of various object that we see around us done with the help of their characteristics such as shape, size, colour etc. Similarly in K-NN algorithm the given input values pass through the trained pre-existing model and identification of the object takes place or the value is predicted. The classification is done by using at the k nearest neighbours. Among the k neighbours those which has higher similar characteristics are put into the same group and the input data is given that particular label. Usually, k is chosen to be odd. There is no particular algorithm to decide the value of k other than trial and error. Large values of k smoothen the data but when the data is small choice of k shouldn’t be large as it may out vote the data. Small values of k such as 1 and 2 can be noisy and maybe affected by outliers.

K-NN is a non-parametric algorithm, which means that there exist no prior underlying assumptions. Similar to logistic regression, K-NN is used for regression as well as classification problems, but mostly used in classification problems. It is also known as lazy learner algorithm.

Since it is a lazy learner algorithm, this algorithm doesn’t learn anything from the training set immediately but instead stores the input dataset and then later on perform actions on the dataset. At training phase, it just stores the dataset and when it gets new data it classifies it into categories.

5.4) SVM:

Support Vector Machine is one of the standard tools for machine learning algorithm which can be used for both classification as well as regression problems, but it is commonly used in classification problems.

SVM is generated using recent advances in statistical learning theory. The aim of SVM is to find a best line or decision boundary so that it segregate n dimensional space into classes so that it can classify new data point into correct category. this decision boundary is termed as hyperplane.

For creating hyperplane SVM uses extreme vectors which is termed as support vectors. SVM can be classified into two linear SVM and non-linear SVM. linear SVM is applied if our dataset is linearly separable. If a dataset can de classified into two classes using a single straight line, then it is known as linearly separable data. If we can’t classify our dataset into two classes using a single straight line, we apply nonlinear SVM.

5.5) NAÏVE BAYES CLASSIFIER

It is a supervised learning algorithm, which is used for solving classification problem, and works on the basis of Bayes theorem. It is comparatively simple and one of the most effective classification algorithms, which makes a quick prediction on dataset.

It predicts on the basis of probability of given object so this classifier is also known as probabilistic classifier. It is naïve because appearance of one feature is unrelated to appearance of others.

It works on Bayes theorem; it makes use of conditional probability. The Bayes theorem's formula is as follows:

$$P(A/B) = P(B/A)\ P(A) / P(B) \text{ ----- Probability of A given B}$$

Here we try to find the probability of a particular condition given a particular condition.

In Naïve Bayes, we have a dataset with pre-existing information the input data is then compared with this dataset and then classified.

5.6) RANDOM FOREST:

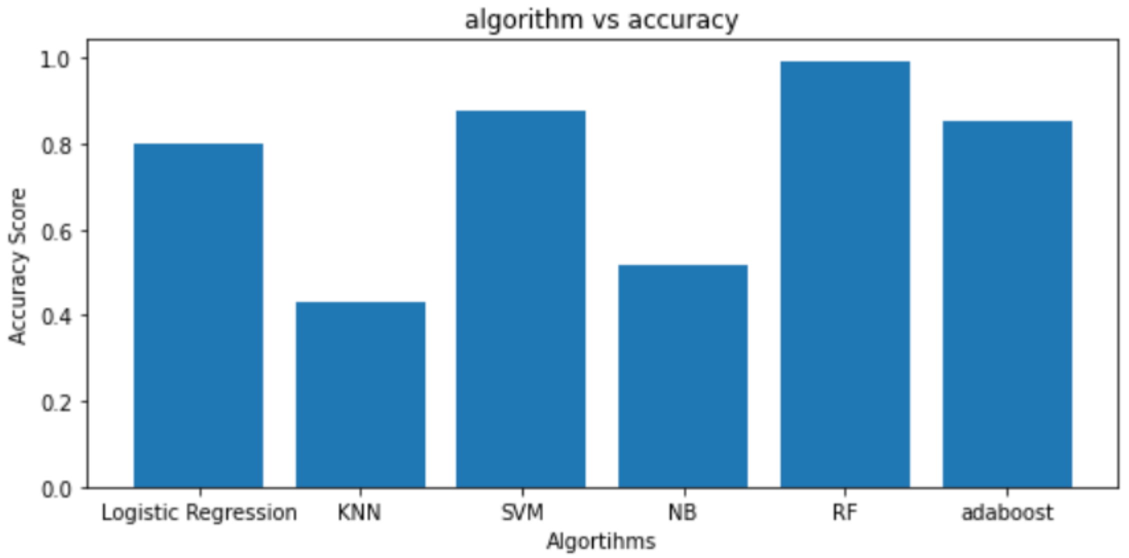
Random forest combines multiple decision trees together. It uses the bagging method on the training dataset. It can be viewed as a combination of many decision trees and each tree produces a class for prediction. The class that satisfies maximum condition will be chosen for the prediction model. The best part about this model is that each tree has its own individual areas and the errors of one tree doesn’t affect the other trees. But on working together they have higher precision and is more stable. The trees have low correlation with each other. When it comes to normal decision tree if we have to split a node, we will consider all possible features and choose the factor that causes the most separation between the left and right node. Whereas in random forest can only choose from the subset of the available features. Random forest uses bagging and feature randomness to built the individual trees so that an uncorrelated forest of trees is obtained.

7. RESULT AND EVALUATION

The accuracy obtained from different models is summarised in the table given below:

SL.NO	ALGORITHM USED	ACCURACY
1.	Support Vector Machine	87.66
2.	Logistic Regression	79.87
3.	Naïve Bayes	51.94
4.	K-nn	42.85

The plot of the above table is given by:



8) CONCLUSION AND FUTURE SCOPE

The future of a country depends on its youth and hence their upbringing is an important responsibility of every nation. This nurturing starts at the elementary level by helping students develop their potentials. Doing this is no easy task as your test scores always remain as numbers rather than helping one grow. Hence, we have come up with the accurate a model to predict the future scores and identify those students who need help and attention. We believe this is an important aspect that can help in developing the future of a country. Various models can easily be accessed and trained to find the perfect predicted value.

REFERENCES

[1] Improving Learning Experience of Students by Early Prediction of Student Performance using Machine Learning, 2020 IEEE International Conference for Innovation in Technology (INOCON).

[2] Prediction of Student Performance Using Linear Regression, 2020 International Conference for Emerging Technology (INCET) Belgaum, India. Jun 5-7, 2020.

[3] Naïve Bayes Classification Model for the Student Performance Prediction, 2019 2nd International Conference on Intelligent Computing, Instrumentation and Control Technologies (ICICICT).