Factors Affecting Students Academic Performance

Mohanraj[a], Parul Vats[b], Swathi Rapeti[c], and Megha Sharma[d]

[a] Assistant Professor (CSE), SRMIST, DELHI-NCR Campus, Modinagar
[b,c,d] M.Tech Student (CSE), SRMIST,DELHI-NCR Campus, Modinagar

Article History: Received: 10 January 2021; Revised: 12 February 2021; Accepted: 27 March 2021; Published online: 16 April 2021

Abstract: Many studies are carried out to find out the factors affecting students' performance Students' academic performance is impacted by many factors. Many factors may have an effect on student performance. To draw inference about whether factors like gender of the student, the race/ethnicity of the student, the level of education of their parents, the type of lunch they ate and whether the test preparation course has any impact on the scores obtained by the student in the tests.

Keywords: EDA, Random forest, Decision Tree, Logistic Regression.

1. Introduction

Student Performance Analysis Trending Field. Performance in schools is something that everyone expects to be of good quality. The social and economic development of the country is directly linked to the academic performance of students. Literacy rates and education are improving and most companies are producing educated, competitive and skilled people. Students' performance was affected due to lack of money, allocation of time for studies, parental level, maternal education, paternal education, address, additional educational assistance and gender. Exploratory data analysis is implemented to improve student performance in education systems. Pre-assessment and analysis of risk and student identification can help students and teachers. And factors involved in student performance. The academic performance of many students has been studied in the past on various topics such as class schedule, class size, English textbook, homework. The goal of compulsory education is to ensure equality at an early stage of one's education, and the principle must have its urgency and uniformity, minimizing the effects of family background on a child's school.

2. Literature Survey

Mehil B Shah, Yogesh Gupta, Maheeka Kaisth.[1] The problem statement of the paper is to develop models that can assess students' performance and classes, while at the same time taking into account other similar personality factors such as interests, traits and opinions (IAO variables) that shape their lifestyle. They affect. methodology They implemented four classification machine learning models and three algorithms that enhance machine learning. The result is that the graduate boosting algorithm was found to increase the accuracy to 93.8%, which is the best of all results. The boosting algorithm exhibits better performance because it uses hyper-parameters that increase performance by fine-tune. The logistic regression they find reflects values between 0 and 1, but does not exceed those limits and leads to wastage.

Ishwank Singh, A Sai Sabitha , Abhay Bansal[2] There are big challenges to the admission process and the curriculum. Two emergency processes that collect and analyze data. The methodology used in this paper is, first, the Crux data collection of any data mining project. In this project, data is considered for a section of the computer. The collected data is pre- processed and the missing values are deleted. The data were generalized and given appropriate weight for the relevant properties of the data The clusters required in the K-Mean clustering algorithm are intended to use the maximum silhouette measurement between all values of 'k'. 5K-ie algorithms are applied to pre-processed data sets to obtain clusters. The result they found was that data mining algorithms set good standards for understanding whether there was continuous improvement in student performance.

V. Shanmugarajeshwari, R. Lawrance[3] The paper is targeted. In it, they evaluate student performance using a variety of attributes to see if they pass or demonstrate attributes such as student names, roll numbers, previous semester marks, attendance, assignments, seminar presentations. The classification process used is based on the C5.05 algorithm and removes the missing data during the preprocessing stage. Properties are converted to the taxonomy format using the taxonomy step. Errors were removed during preprocessing. Feature selection methods incorporate preprocessing data sets and associated features. The profit ratio attribute was selected over the data ratio. This function can be implemented in cloud computing to achieve greater security for diverse datasets.

Muhammad Faisal Masood, Aimal Khan , Farhan Hussain, Arslan Shaukat, Babar Zeb.[4] In order to assess the student problem and its results, with experiments, we can easily identify vulnerable students and take

appropriate precautions that can improve performance. And which model is the best. They took eleven models and tried to figure out which was the best. So that the students have a problem in assessing their result. With exams, vulnerable students can be easily identified and appropriate precautions can be taken to improve performance. And which model is the best. They took eleven models and tried to figure out which was the best. The method they used defined the data extraction model in which the data extraction elements were classified into different categories to perform accuracy calculations for different models. Decision tree and random forest patterns are used. Conclusions and future scope are best based on the "Decision Tree" and "Random Forest" accuracy scores. In it, we want to recalculate the outcome and assess their future outcomes, based on some important steps that can enhance students 'progress and these steps. The issue is related to the purpose of security, permitted only for the undergraduate and higher management of student development due to its privacy setting. The proposed system maintains an assessment system that can assess student performance. These faculty were able to see the performance of all semesters while exploring the performance of students who provide a simple rule when evaluating student results. This system helps faculty to assess students' failure in the course. Its future scope is an estimate made using the Decision Tree produced by WEKA, which is not a redesign of the Dynamic Prediction Model, the Prediction Model can be applied by train

Ching-Chieh Kiu[4] In this paper, they analyze the main impact of student research in assessing student background, student social activities and student academic performance. The following features were applied prior to the evaluation of the data exchange and normalization assessment model. Webca data mining tools are used for analysis on datasets. Four supervised data mining methods, Nav Bayesian, Multilayer Perceptron, Decision Tree J48 and Random Forests are used to assess. In this paper, it was found that student background and social activities are important in assessing the early stages of student performance and can also be used to identify students at risk. In future work, unlicensed education data mining methods will be applied to find the correlation and effects of traits in groups.

Proposed Model

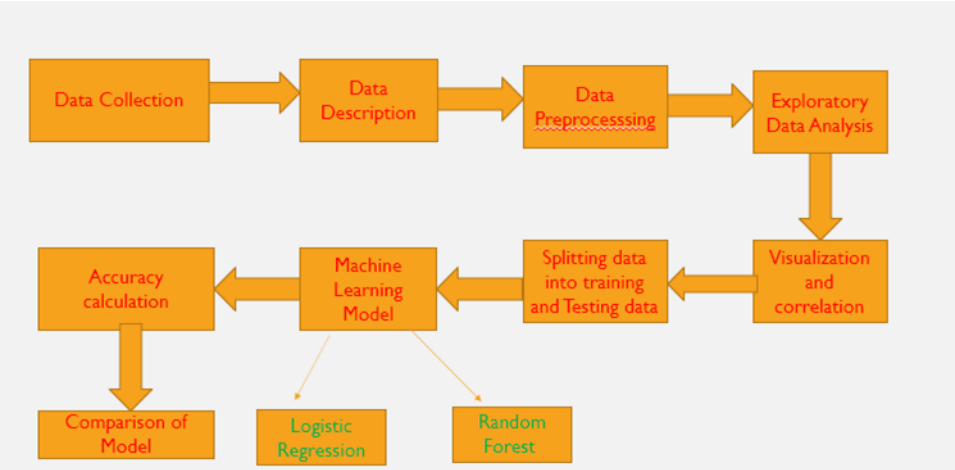


Fig.1 Proposed Model

DATASET DESCRIPTION

DataCollection:We got our dataset of Secondary schools. The dataset was collected using the school reported data in which student's social background data is present.

Data Description : We have used 8 attributes in this dataset. These attributes are gender ,race/ethnicity ,parental level of education, lunch, test preparation course, math score reading score , writing score .

Data Preprocessing: Since the data was clean, with null values or attributes ,we skipped the data cleaning phase ,and started preprocessing the data. In the Data Preprocessing phase, we added the Total score and pass/Fail column and assigned values to it depending on the total scores, and we transformed it into numerical values using Label Encoder. We concluded on Pass/Fail as our dependent variable and all other variables as the predictor variable.

3. Methodology :

A. Exploratory Data Analysis

We started our analysis by plotting various graphs on our dataset.

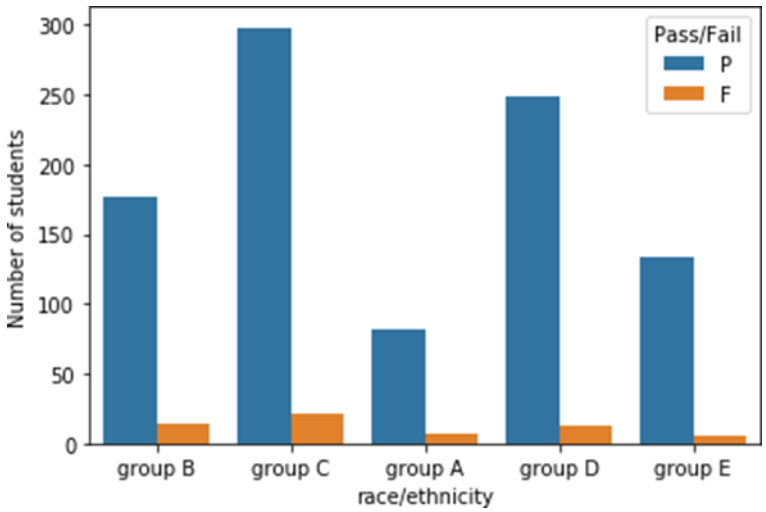
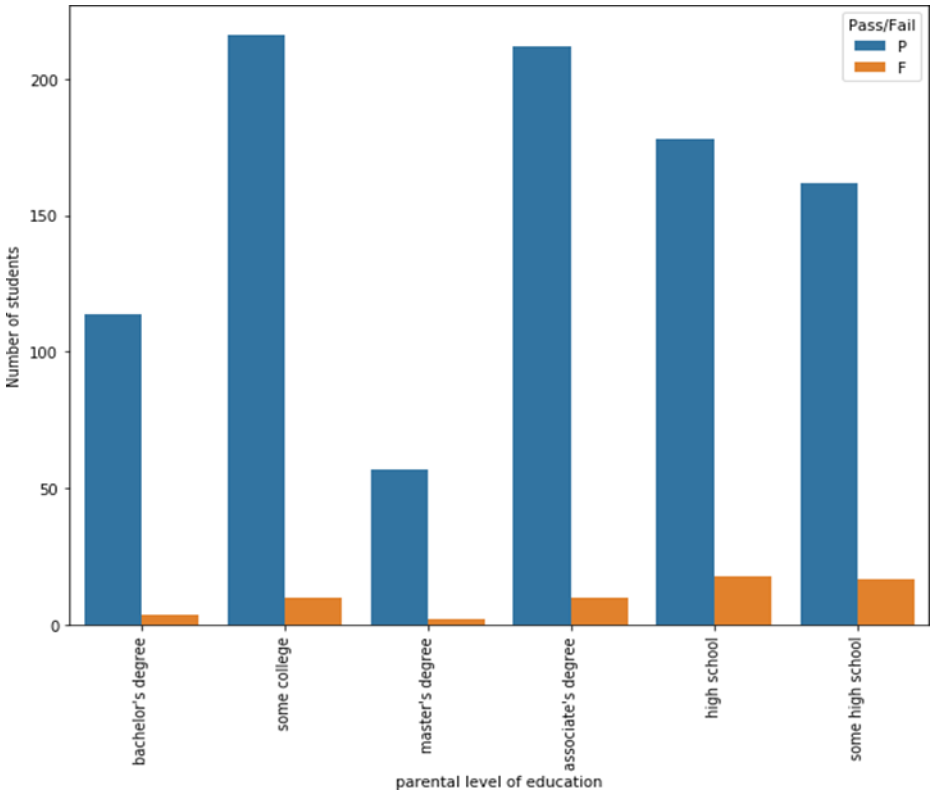
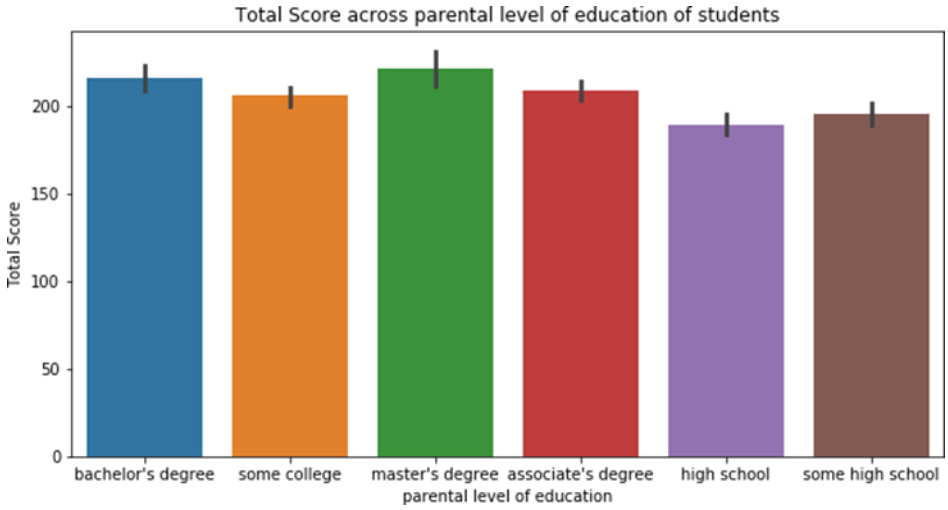


Fig.2 Histogram for race/ethenicity and number of student

Thus from the above analysis we can observe that the race/ethnicity groupC has performed better than all other groups and the group 'group A' has performed poorer than any other groups .It can also be observed that the performance of students in race/ethinicity group gets better as we move 'group A' to 'group E'.

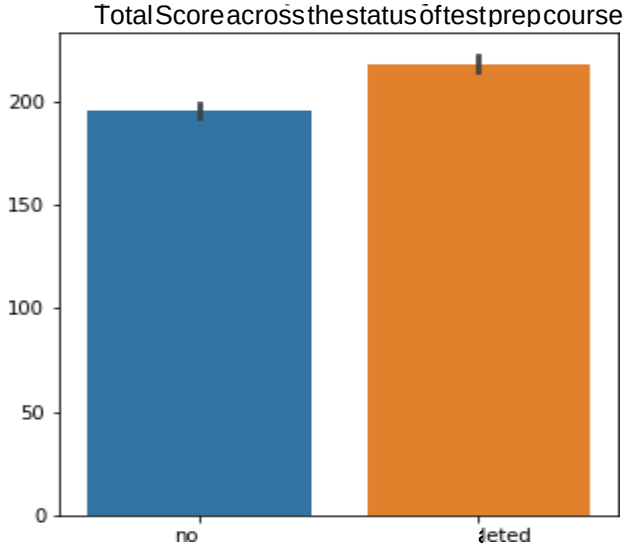


Thus among the 1000 students, 226 students have parents with 'some college' background, 222 with 'associate's degree',196 have 'high school' background, 179 have parents with 'some high school' background, 118 with 'bachelor's degree',59 with 'master's degree' background. Now we will try to analyze how the performance of the students vary depending on their parents educational background.

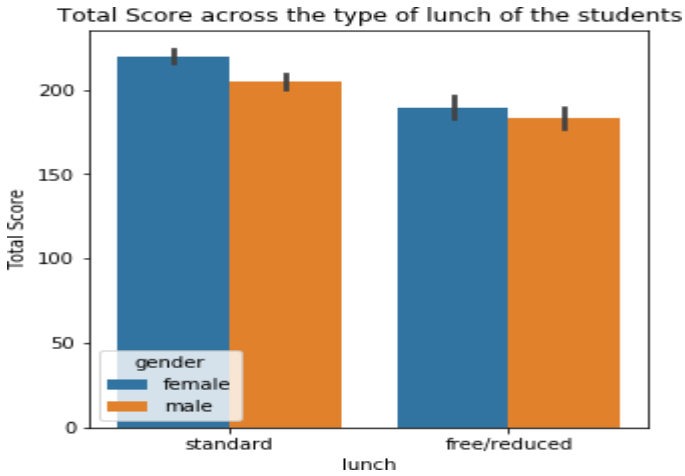


As can be observed from the above plot that there is some influence the parent's background have on the student's performance. As can be seen, that students having parents with master's degree performed better than other and students with parents having some high school level of education performed poorer than the other groups.

As can be noted from the below graph that the test preparation course has an impact on the performance of the students, 97.21% of the students who completed the 'Test Preparation Course'passed whereas 92.06% of the students who didn't complete 'Test Preparation Course' passed.

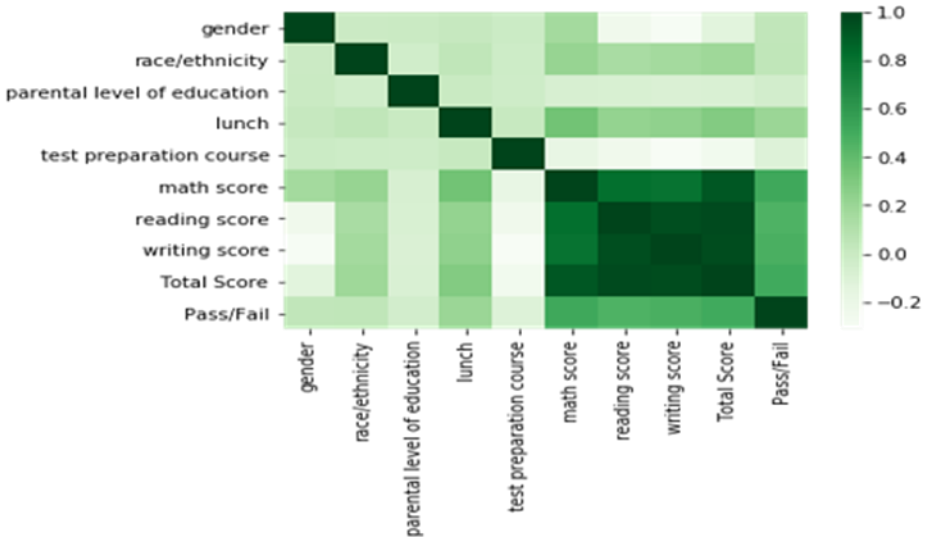


Status of Test Prep Course



So as we can observe from the above plot, the type of lunch has an impact on the scores of the students. The students with 'standard' lunch performed better than the student with 'free/reduced' lunch.

From the below corelation matrix we can make derived



• The grades of subjects have a very high correlation with the differentsubject.
• There is a correlation between lunch and math grades.
• There is a corelation between math score and race/ethenicity of student.
• There is no correlation between teset prepration of course and grades.
• There is correlation between between reading and mathscore
• We can observe from the correlation matrix there is no correlation between gender , parental level of education with subject score

Machine learning algorithm

We implemented two machine learning model on our dataset.

Logistic Regression

This is the most commonly used algorithm for classification. Logistic regression is a supervised learning algorithm that is applied to estimate taxonomic variables or discrete values. This can be applied for classification and the output of the logistic regression algorithm can be yes or no, 0 or 1 etc.

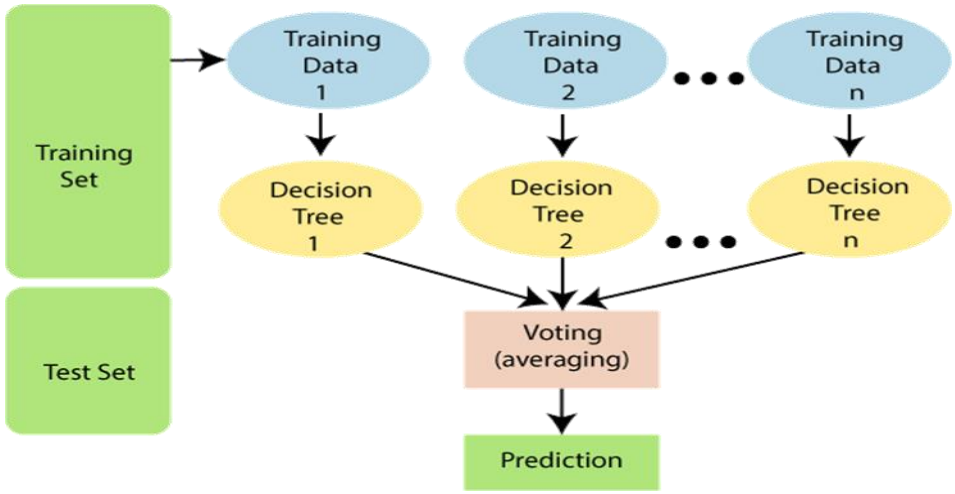
Logistic regression is similar to linear regression, except that they are applied, such as linear regression to solve the regression problem and to estimate the fixed values, while logistic regression is used to obtain the classification problem. Discerned values are valued and appreciated. It works on the sigmoid function.

Random Forest Classifier

Random Forest is a machine learning algorithm demanded by a member of the supervised learning methods. It can be used for both classification and regression problems in machine learning. This ensemble is built on the theme of learning, which is the act of merging different classifiers to solve an overall problem and increase the performance of a model.

As the name implies, "Random Forest is a taxonomy that imposes multiple types of decision trees on different subsets of a given dataset and takes an average to maximize the estimated accuracy of that dataset." Instead of relying on the judgment tree, the random forest takes an estimate from each tree and builds the majority of the predictor votes and it estimates the final product.

Large numbers of trees in the forest lead to high accuracy and avoid the problem of over-fitting.



4. Results

We have visualize from the above graph that the background of student also affect the performance of student. Lunch and level of parental education also affects the performance of student .To get the higher performance if student we should provide attention on the background and social factors which also influenced the performance.

Apart from academic studies, performance is influenced by social background factor

From the table 1 we can identify that the accuracy of logistic regression is more as compare to random forest .**Models**	**Accuracy**
Logistic Regression	92.8%
Random Forest	90.8%

Table 1 Accuracies of Model

Logistic regression is performing better among the models.

5. Conclusion:

Our plan for the upcoming is to get more data and train the model on that data to get more accuracy .By giving more data we can enhance the accuracy as well this research will be helpful for the faculties as predicting the student those are at risk at the early stage ,As predicting the students status will aids them to get the mandatory action and help for the faculties to determine the student those need more attention and help.

References

1. Mehil B Shah and Yogesh Gupta " Student Performance Assessment and Prediction System using Machine Learning " IEEE (2019).
2. IshwankSingh,ASai Sabitha ,AbhayBansal "Student performance analysis using clustering algorithms " IEEE (2017).
3. V.Shanmugarajeshwari,R.Lawrance"AnalysisofStudents'PerformanceEvaluationusing Classification Techniques " IEEE(2018).
4. Muhammad FaisalMasood,AimalKhan,FarhanHussain,ArslanShaukat,BabarZeb"Towards the Selection of Best Machine Learning Model for Student Performance" IEEE (2017).
5. ChewLiSa,DayangHananibt.AbangIbrahim,EmmyDahlianaHossain,Mohammadbin Hossin" Student Performance Analysis System(SPAS)" IEEE(2018).
6. Ching-ChiehKiu "DataMining Analysis onStudent's Academic Performance through Exploration of Student's Backgroundand Social Activities"IEEE(2017).
7. ismailduru"An Overview OfStudies About Students Performance Analysisand Learning AnalyticsinMOOC's(IEEEInternationalConferenceonBigData(BigData))(2016).
8. IrfanMushtaq&Shabana Nawaz Khan "FactorsAffecting Students' Academic Performance" (Global Journal ofManagement and Business Research)(2017).
9. https://www.ijert.org/.
10. https://en.wikipedia.org/. .