e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:03/Issue:12/December-2021 Impact Factor- 6.752 www.irjmets.com

PERFORMANCE ANALYSIS AND PREDICTION OF STUDENT RESULT USING MACHINE LEARNING

Prakash*1, Mr. Sachin Garg*2

*1Information Technology, Maharaja Agrasen Institute Of Technology, Rohini, Delhi, India.

*2Assistant Professor, Department Of Information Technology, Maharaja Agrasen Institute Of Technology, Rohini, Delhi, India.

ABSTRACT

There is a lot of research work going on to enhance the Learning Management system. Nowadays, educational institutes have many tasks to be completed in a given timeline. In today's scenario, educational institutes need to analyze student results manually, and sometimes errors may occur during analysis. This process takes a lot of time and effort from faculties who need to analyze the students' results individually. Hence, to simplify this task, a system is introduced that uses machine learning to analyze the student performance and predict future results based on the student's previous performance while considering other factors about the student. This paper proposes a machine learning model that can predict the student's future grade by analyzing previously obtained marks and other socio-economic factors like student attendance, parent's education, travel time, study time, frequency of going out. For this purpose, this model uses various machine learning algorithms. With the help of this model, faculties can analyze student performance and guide their students so that he/she can reach expected marks. The web version of this project can also be implemented on existing college websites to help both students and faculty.

Keywords: Machine Learning, Pandas, Numpy, Matplotlib, Multivariate Linear Regression, Random Forest, RMSE.

I. INTRODUCTION

Machine learning is a subset of Artificial Intelligence (AI) that enables applications to make predictions without having to be explicitly coded. Through this, application programs can achieve a great level of accuracy in the final result. Machine learning uses historical data or information as input to predict new results. Prediction system is one of the primary use cases of machine learning. Other popular uses are email spam filtering, speech recognition, recommendation system. Machine learning also helps in the education field, where prediction become an important task for both student and faculties. In learning management system (LMS), algorithms are designed in such a way that allows the model to take input as data, train according to data, and produce a required range result. In the field of education, student final grade prediction is done by only considering the study hours. That model is good, but study time is not the only factor that decides the final grade of the student, other real-life factors majorly affect the student's final grade, which is sometimes not considered. In this paper, students' previous performance and other factors are considered to predict the final grade of the student. Performance evaluation is an essential tool for educators for helping students in attaining their desired goals. Through performance analysis, teachers can divert their attention to the necessary areas that need it most, offer advice and guidance, and recognize and reward their accomplishments. For this task, machine learning comes into the picture, through which various algorithms accuracy can be measured, and the best one we will take into consideration for the prediction task.

II. LITERATURE REVIEW

Student grade prediction is one of the essential research topics in education. Several other authors have worked on this topic and found different insights:

B. k. Bhardwaj and S. Pal [1] did a study on predicting the student's performance by choosing over 300 students from six-degree colleges conducting BCA (Bachelor of Computer Application) course in Dr. R. M. L. Awadh University, Faizabad, in India. Using the Bayesian classification technique on seventeen attributes, they showed that students' academic performance corresponds to both the academic and non-academic attributes like family annual income and student's family overall status, etc.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:03/Issue:12/December-2021 Impact Factor- 6.752 www.irjmets.com

Authors in [2] proposed a model on Prediction of Students Performance using Machine learning in which they used previously obtained marks by the students of class 10th, 12th, and their semester marks. The scope of this paper was to predict the result and find out how many students got marks below 50% in 10th and 12th, students who failed in the internal exam, and the students had less attendance percentage.

S. Huang and N. Fang[3] examined various mathematical and machine learning techniques and applied four different mathematical modeling techniques: multivariate linear regression, multilayer perceptron neural networks, radial basis function neural networks, and support vector machines to predict student performance. The dataset contains 1,938 data records that were collected from 323 undergraduates in four semesters. This study concludes that there is no such difference between these methods. Their model was able to get more than 80% of accuracy.

J. Gamulin, O. Gamulin, and D. Kermek[4] collected the student data which is generated through Learning Management System(LMS), web-based formative and summative assessments during the traditional teaching in the classroom. The dataset contains a huge amount of data on students' behavior and grade/percentage at the point of time when the course is still in progress. Considering this dataset and by applying various classification algorithms and genetic algorithms, the author proposed a model for predicting the performance of students in the final examination.

III. PROBLEM STATEMENT

The problem statement can be defined as "Given a Portuguese school student dataset[5], analyze the performance of the student and predict the final grade of the student by considering previously obtained grade along with other socio-economic factors such as parent's education, travel time, study time, attendance, family relationship, alcohol consumption, etc. by applying different machine learning algorithms on the dataset."

IV. METHODOLOGY

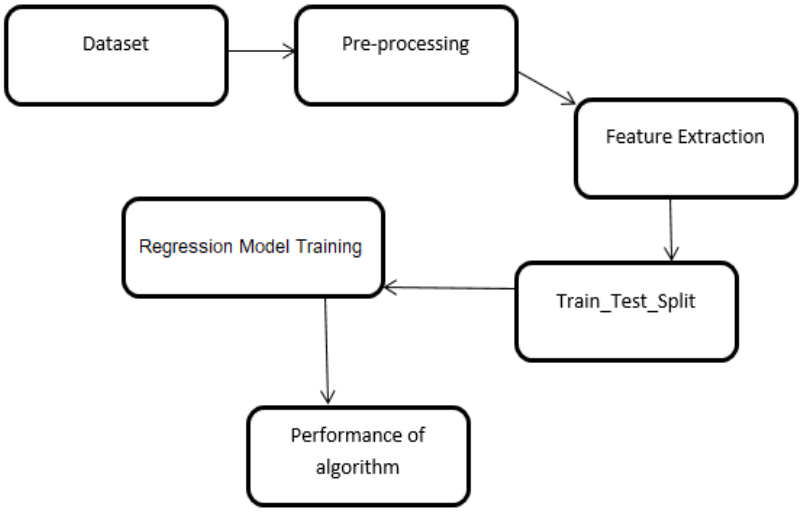

Figure 1: Steps followed for obtaining results.

A. Dataset

The dataset is taken from the UCI Machine Learning Repository[5] of two different schools in Portugal comprising secondary education. The dataset contains information about student performance with several parameters: previously obtained grades, study time, past failures, parent's education, presence in class, etc., and it is the result of data collected using school reports and questionnaires. The datasets are of two subjects: Mathematics and Portuguese language, which were designed according to binary three-level classification and regression.

B. Data Preprocessing

First, we'll check for potential null or nan values in our dataset. Since our dataset is already clean, no null values or missing values present in the dataset. so we leave out the data cleaning phase and directly proceed with the data Preprocessing phase.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:03/Issue:12/December-2021 Impact Factor- 6.752 www.irjmets.com

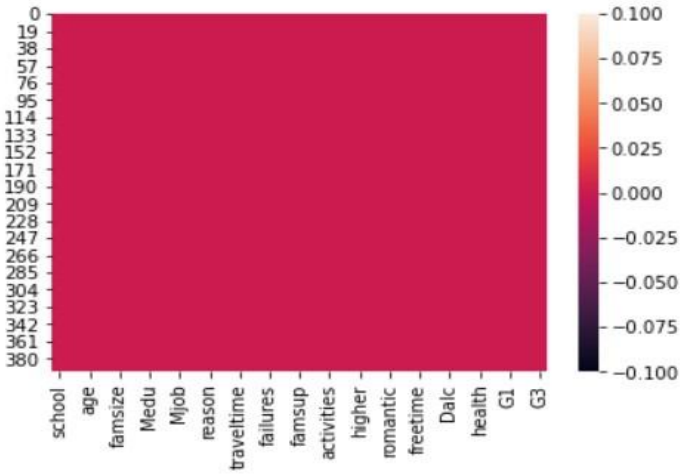

Figure 2: Heatmap showing no null values in dataset

Our machine learning model needs data to be present in numerical form. So we need to transform categorical values into numerical values, for this purpose we have used ordinal Encoding.

```
d = {'yes': 1, 'no': 0}
dataset['schoolsup'] = dataset['schoolsup'].map(d)
dataset['famsup'] = dataset['famsup'].map(d)
dataset['paid'] = dataset['paid'].map(d)
dataset['activities'] = dataset['activities'].map(d)
dataset['nursery'] = dataset['nursery'].map(d)
dataset['higher'] = dataset['higher'].map(d)
dataset['internet'] = dataset['internet'].map(d)
dataset['romantic'] = dataset['romantic'].map(d)
```

```
# map the school data
d = {'GP': 1, 'MS': 0}
dataset['school'] = dataset['school'].map(d)

# map the sex data
d = {'F': 1, 'M': 0}
dataset['sex'] = dataset['sex'].map(d)

# map the address data
d = {'U': 1, 'R': 0}
dataset['address'] = dataset['address'].map(d)

# map the famili size data
d = {'LE3': 1, 'GT3': 0}
dataset['famsize'] = dataset['famsize'].map(d)

# map the parent's status
d = {'T': 1, 'A': 0}
dataset['Pstatus'] = dataset['Pstatus'].map(d)

# map the parent's job
d = {'teacher': 0, 'health': 1, 'services': 2,'at_home': 3,'other': 4}
dataset['Mjob'] = dataset['Mjob'].map(d)
dataset['Fjob'] = dataset['Fjob'].map(d)
```

Figure 3: Ordinal encoding

C. Exploratory Data Analysis

Before applying the machine learning model, some performance analysis needs to be done to analyze which factor affects the student performance most and how dependent and independent features are related to each other. In EDA, various graphs are plotted based on the dataset. The barplot below indicates that the majority number of students have fewer absences and tend to perform better. At the same time, the students who are absent for most of the classes will get their grades low.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:03/Issue:12/December-2021 Impact Factor- 6.752 www.irjmets.com

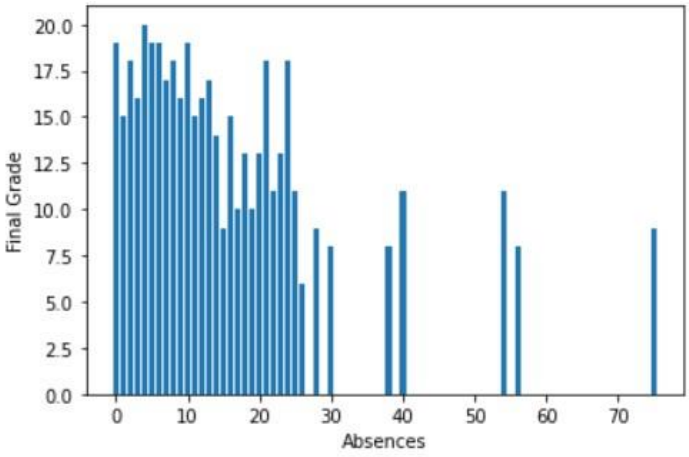

Figure 4: Bar graph showing student absence in class vs final grade

A box plot is a statistical plot to visualize descriptive statistics, including mean, median, quartile 1, quartile 2, minimum, maximum values. From Fig. 5, it can be inferred that students with study time belonging to group 3(5-10 hours) or group 4(>10 hours) will perform better in the final exams as they have the maximum median final score. Students who belong to group 1(1-2 hours) and group 2(2-5 hours) have the lower final grade. This led to the inference that more study time will always result in better final grades.

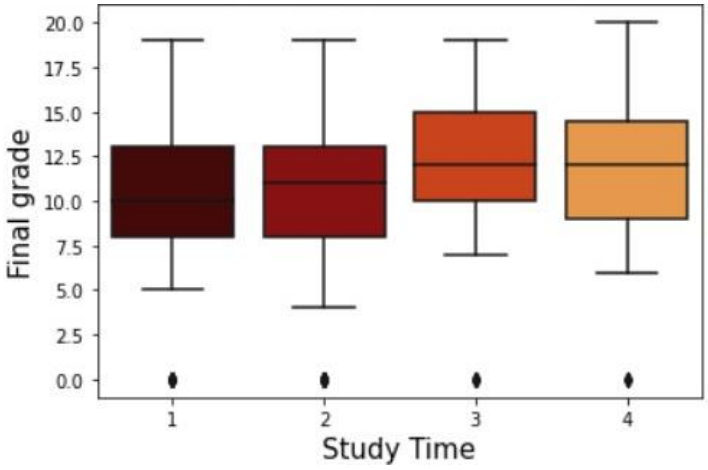

Figure 5: Box plot showing Study time vs Final grade.

From the below bar graph, we can see that the parent's education level plays a significant role in student grades. The below graph shows that the connection between the father's education and student's grades is weak, but for the mothers who are educated, their student's final grade increased. It may be because mothers teach their children and spend most of their time with them.

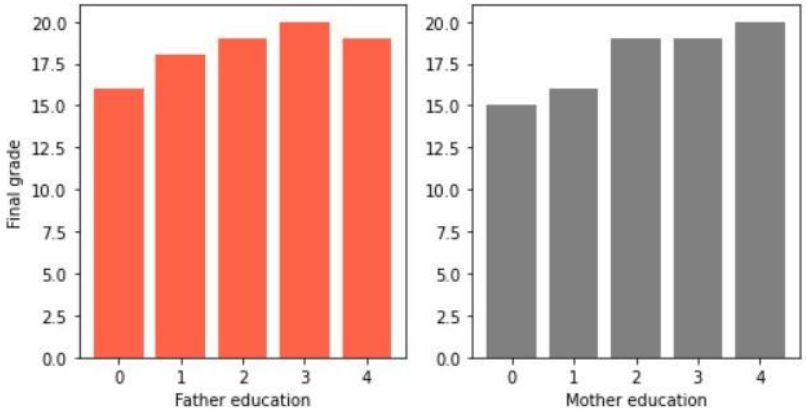

Figure 6: Bar graph showing Parent's education vs Final grade

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:03/Issue:12/December-2021 Impact Factor- 6.752 www.irjmets.com

The below heatmap shows that parents' education, number of past class failures, study time, alcohol consumption, travel time, the student who wants to pursue higher education, class attendance, and previously obtained grades are the eight most correlated factors with G3. Here, G1 and G2 (independent feature) are highly correlated with G3 (dependent feature), but G1 and G2 are highly correlated with each other, so we can remove one of them to avoid overfitting, in this case, we'll only consider G1 (previously obtained grade) to predict G3 (final grade).

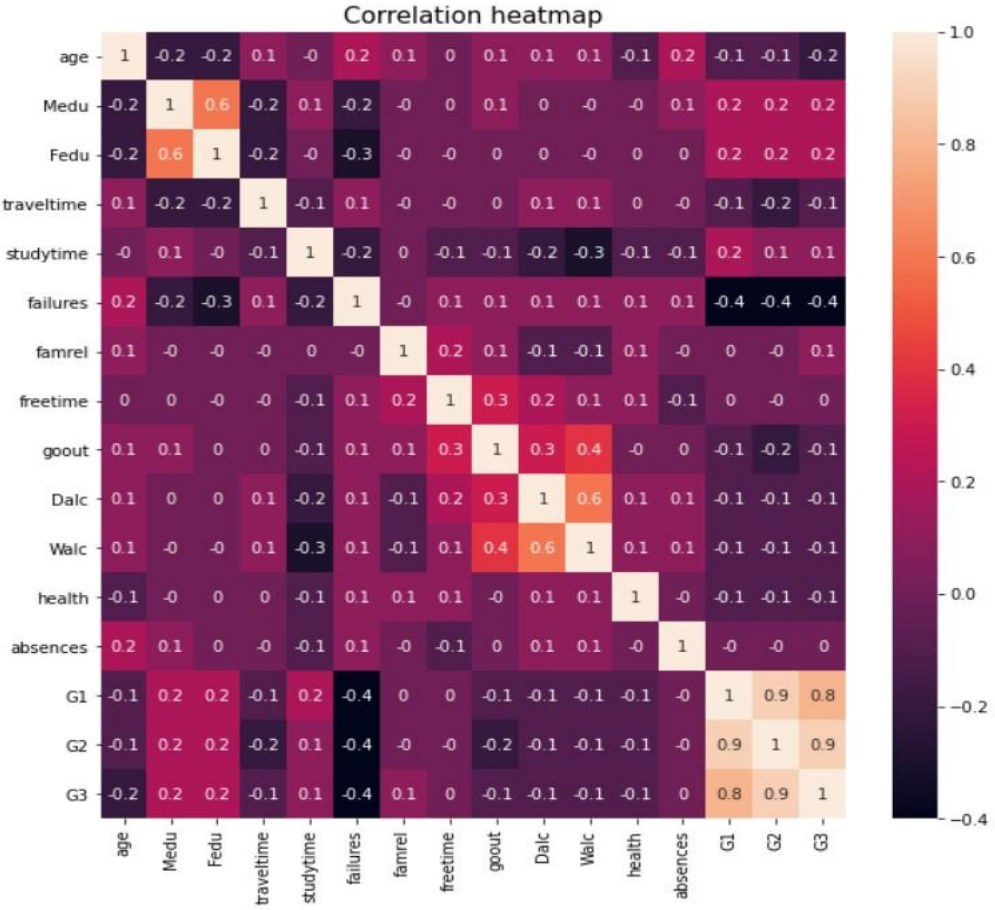

Figure 7: Correlation Heatmap

By seeing the correlation heatmap, we can observe that:

1. The final grade G3 is highly correlated with the parent's education, and since both Mother education and father education are highly correlated, we will only consider mother education.

2. Final Grade G3 negatively correlates with past failures, frequency of going out, and daily alcohol consumption.

3. Previously obtained grades G1 and G2 are highly correlated with final grade G3.

D. Various Machine Learning Algorithms

For predicting the student's final grade, we implemented four different regression machine learning Algorithms on our dataset, i.e., multivariate linear regression, random forest, gradient boosting, and bayesian ridge regression.

Testing and Training the data

For training and testing, we use the 8:2 ratio, i.e., 80% data for training and the remaining 20% data for testing. Using the scikit-learn library, we split the data into X_train, y_train, X_test, and y_test. Training data is used for training the model, and testing data is used for predicting and checking the accuracy against training data. At each run, the final result and accuracy score may vary. To avoid this, we set the random_state attribute to 0.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:03/Issue:12/December-2021 Impact Factor- 6.752 www.irjmets.com

Splitting the data into training and testing data

```
df = dataset[["Medu", "traveltime", "failures", "studytime", "higher", "Dalc", "absences", "G1", "G3"]]

X = df.iloc[:, 0:8].values

y = df.iloc[:, -1].values

from sklearn.model_selection import train_test_split

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state = 0)

print("X_train", X_train.shape)
print("X_test", X_test.shape)
print("y_train", y_train.shape)
print("y_test", y_test.shape)
```

```
X_train (316, 8)
X_test (79, 8)
y_train (316,)
y_test (79,)
```

Figure 8: train_test_split

1. Linear regression

Linear regression is a regression technique where a straight line is used to model the relationship between dependent variable and independent variable. More than one feature is used in multivariate linear regression to predict the target value. As to predict the value of the dependent variable, it requires the set of data, so it belongs to the supervised learning category.

```
from sklearn.linear_model import LinearRegression
model1 = LinearRegression()
model1.fit(X_train, y_train)
pred1 = model1.predict(X_test)
```

2. Random forest

Random Forest is a basic yet adaptable machine learning method that delivers excellent results in the vast majority of cases. Due to its simplicity, it is widely used in a variety of problem statements, and it is best suited for both regression and classification issues. Forest is a collection of Decision Trees that have been trained mostly using the "bagging" approach, in which the combination of various learning algorithms helps in improving accuracy.

```
from sklearn.ensemble import RandomForestRegressor
model2 = RandomForestRegressor(n_estimators = 100, random_state = 0, bootstrap = True,
model2.fit(X_train, y_train)
pred2 = model2.predict(X_test)
```

3. Gradient Boosting Regression

Gradient boosting is another machine learning algorithm in which many models are trained consecutively. Using the Gradient Descent approach, each new model gradually reduces the loss function (y = axe + b + e, where 'e' is the error component) of the entire system. The learning process fits new models in a sequence to provide a more precise estimate of the response variable. Gradient boosting regression is also used for both regression and classification problems.

```
from sklearn.ensemble import GradientBoostingRegressor
model3 = GradientBoostingRegressor(max_depth = 2, n_estimators = 50, learning_rate = 0.5)
model3.fit(X_train, y_train)
pred3 = model3.predict(X_test)
```

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:03/Issue:12/December-2021 Impact Factor- 6.752 www.irjmets.com

4. Bayesian Ridge Regression

Bayesian Regression is a regression algorithm that comes in handy when there isn't enough data in a dataset or when the data isn't evenly distributed. In contrast to traditional regression techniques, where the output is derived from a single value of each attribute, the output of a Bayesian Regression model is derived from a probability distribution. A normal distribution (where mean and variance are normalized) is used to generate the output.

```
from sklearn.linear_model import BayesianRidge
model4 = BayesianRidge()
model4.fit(X_train, y_train)
pred4 = model4.predict(X_test)
```

V. RESULTS

While considering the student grade dataset, four algorithms, linear regression, random forest, gradient boosting, and bayesian ridge, are tested and under different permutations and combinations. The below bar graph shows the accuracy of each algorithm under these parameters, and gradient boosting comes out to be the best algorithm to predict the student grade based on the given parameters with an accuracy of 79%. The random forest has the second-best result with an accuracy of 74%. Linear regression and bayesian ridge regression show the least accuracy, i.e., 69%.

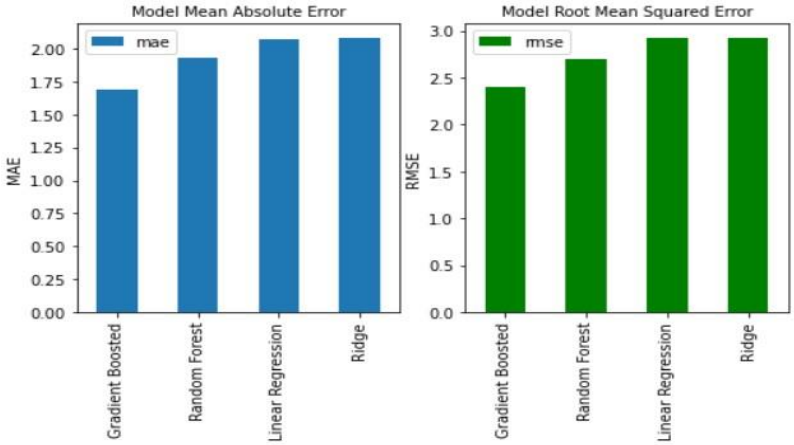

Figure 9: Bar graph showing MAE and RMSE score

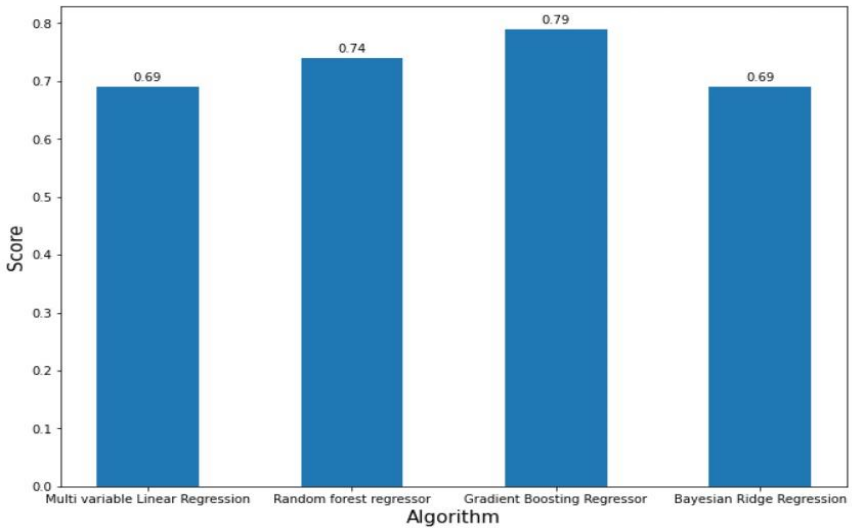

Figure 10: Bar graph showing accuracy score of different algorithms.

VI. CONCLUSION

After evaluating all the algorithms on different parameters, we have managed to propose a model that can predict the grade more accurately using gradient boosting regression algorithm. This model helps both the educator and academic institutions to analyze student performance with the help of various graphs through which they can easily decide about the students' performance and suggest a better method for improving their academics. In the future, an end-to-end website can be developed using which the end-user can come to check the predictions more easily. The performance of the model will vary as per the features changes. We have proposed a model that can give consistent and accurate results to the end-user, which satisfies their need by showing the correct output and helps to take better steps toward the study. In the future, more features and different regression algorithms can be used to improve the model accuracy.

VII. REFERENCES

[1] Baradwaj, Brijesh & Pal, Saurabh. (2011). Mining Educational Data to Analyze Students' Performance. International Journal of Advanced Computer Science and Applications.
2. 63-69. 10.14569/IJACSA.2011.020609.

[2] Dhilipan, J., Vijayalakshmi, N., Suriya, S., & Christopher, A. (2021). Prediction of Students Performance using Machine learning. IOP Conference Series: Materials Science and Engineering,
1055(1), 012122.doi:10.1088/1757-899x/1055/1/012122.

[3] S. Huang and N. Fang, "Work in progress: Early prediction of students' academic performance in an introductory engineering course through different mathematical modeling techniques," 2012 Frontiers in Education Conference Proceedings, 2012, pp. 1-2, doi: 10.1109/FIE.2012.6462242.

[4] J. Gamulin, O. Gamulin and D. Kermek, "Comparing classification models in the final exam performance prediction," 2014 37th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), 2014, pp. 663-668, doi: 10.1109/MIPRO.2014.6859650.

[5] https://archive.ics.uci.edu/ml/datasets/student+performance

[6] Ajay Ohri (2017, Feb 16). Popular regression algorithms [Online]. Available:
https://www.jigsawacademy.com/popular-regression-algorithms-ml/ accessed on 25.10.2021.

[7] A. M. Shahiri, W. Husain, and N. A. Rashid, “A Review on Predicting Student’s Performance Using Data Mining Techniques,” in Procedia Computer Science, 2015.

[8] P. Guleria, N. Thakur, and M. Sood, “Predicting student performance using decision tree classifiers and information gain,” Proc. 2014 3rd Int. Conf. Parallel, Distrib. Grid Comput. PDGC 2014, pp. 126–129, 2015.

[9] Z. Liu and X. Zhang. Prediction and analysis for students’ marks based on decision tree algorithm. In 2010 Third International Conference on Intelligent Networks and Intelligent Systems, page 338341, Nov 2010

[10] P. Kaur and W. Singh. Implementation of student sgpa prediction system (ssps) using optimal selection of classification algorithm. In 2016 International Conference on Inventive Computation Technologies (ICICT), volume 2, page 18, Aug 2016.