Analyze Various Student Performance Prediction Methods And Plot The Rank Based On Performance

Farsana Jasmin
Dept. of computer Applications
Amal Jyothi college of Engineering
Kanjirappally, Kottayam
farsanajasu123@gmail.com

Ms. Navyamol K. T.
Asst. Professor
Department of Computer Applications
Amal Jyothi college of Engineering
Kanjirappally, Kottayam
navyamolkt@amaljyothi.ac.in

***Abstract* –The amount of data generated these days is enormous. Analyzing and generating useful insights from large amounts of data is difficult. Student data analysis plays a crucial role in providing an overall view of what students know and should know, as well as what may be done to satisfy their academic needs. With this information, the school can make decisions to improve students' academic performance. The significance of this work is to investigate the dataset, which includes student-related characteristics, using various methods such as Logistic Regression, Naive Bayes, and Extreme Gradient Boosting. This article also includes a comparison of these methods with multiple linear regression algorithms.**

Keywords: Prediction, Machine Learning, Performance, Algorithm,

I. INTRODUCTION

Machine learning is an area of computer science that allows computers to learn without having to be programmed directly. Machine learning is one of the most fascinating technology that has ever been discovered. It gives the computer the ability to learn, making it more human-like, as the name implies.

Multiple statistical approaches have been employed to examine and predict students performance from various perspectives over the years. One of the most significant difficulties facing higher education is the lack of decisions. Today's task is to forecast students' progress through the educational system. Predicting successful students' results early in the course depends on a variety of factors. data mining techniques could be applied. In the educational industry, data mining techniques are commonly utilized to uncover new hidden patterns in student data. The underlying patterns that are uncovered can be used to comprehend the educational issues that develop.

Based on the same dataset available in the public domain, this study seeks to provide a comparative analysis on several features of Nave Bayes, XGBoost, and Logistic regression in predicting student performance. Machine learning classification algorithms are used to apply the classification technique to the dataset. These models are used to improve the classification technique's accuracy. This model is capable of both classification and prediction. The Python Programming Language is used to create these models. The Nave Bayes Classifier is based on the Bayes Theorem and the concept of probability. It frequently plays an important part in the decision-making process. The K-nearest neighbor The feature similarity principle is used by the classifier. It can be used to solve classification and regression difficulties. In this post, you learned about the XGBoost algorithm for applied machine learning. That XGBoost is a library for quickly constructing high-performance gradient boosting tree models. On a range of difficult machine learning problems, XGBoost surpasses the competition. The method of modelling the probability of a discrete result given an input variable is known as logistic regression. The most frequent logistic regression models have a binary outcome, which might be true or false, yes or no, and so forth.

II. LITERATURE SURVEY

Febrianti Widya ha stuti,VianyUtamiTjhin et al[1] in order to Prediction of Student Performance using Machine Learning The study and prediction of student performance is extremely important.

The analysis and prediction of student performance is extremely beneficial to both the student and the institution. It can assist the student in better understanding his current performance and planning his time accordingly in order to improve his performance in the next tests. Institutions can adjust their coursework or include any other methods to improve student performance from the standpoint of the institution. We need effective data analysis and prediction to attain these valuable findings. The analysis and prediction of student performance is extremely beneficial to both the student and the institution. It can assist the student in better understanding his current performance and planning his time accordingly in order to improve his performance in the next tests.

III. MOTIVATION

Predicting student performance has become a critical issue in most educational bodies and universities. This is vital to help at-risk students and assure their retention, as well as to provide great learning materials and experiences, and to improve the university's rating and reputation. Student performance is a major problem in educational institutions, as a range of factors can influence student progress. The following three components are necessary for prediction: student performance-related parameters, student performance-related parameters, and student performance-related parameters. The third option is to use a data mining tool. Every year, a large amount of student data is entered into

a database, but it is not organized properly. There is a need for data mining to handle and overcome these issues. Then there is enough data for better planning, appraisal, and decision-making. Data mining is the process of extracting hidden information from data. This information will be useful to institutes because it will be stored in a student enrollment database. After that, a better &There is no additional requirement because mined knowledge is already stored in a database. The motivation for this paper is to use a classification model to forecast student performance based on a variety of characteristics such as failures, absences, and ranking based on their grades. This is helpful in determining which pupils are likely to be interested.to be aware of their performance We will discover some interesting patterns as a result of this research. Institutes may find it beneficial.

IV. METHODOLOGY

A. Data Source

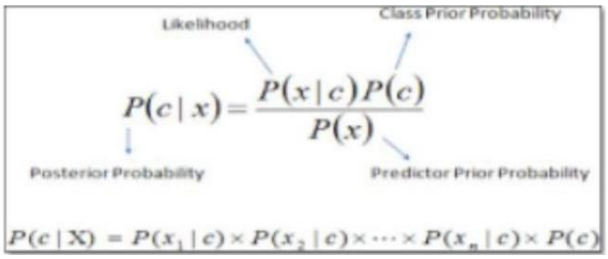
The dataset used here for predicting student performance is taken from UCI Machine learning repository. UCI is a collection of databases that are used for implementing machine learning algorithms. The dataset used here is student dataset. The dataset consists of 395 instances of data with the appropriate 22 student parameters

B.Naive Bayes Classifier

It is a classification approach based on the assumption of indicator autonomy. In simple terms, a Naive Bayes classifier accepts that the proximity of one element in a class is unrelated to the proximity of another element.

hese Bayesian probabilities are the key component used to determine the most likely next event for the given instance given all the training data. Conditional probabilities are determined from the training data. The Naive Bayes model is based on the conditional independence model of each predictor give the target class.

Results: Naive Bayes classification of Data Mining has been proven that can be effective for the prediction student performance. The accuracy or prediction rate of Naive Bayes is accuracy of 0.8125



C. XGBoost

The XGBoost package makes it possible to calculate the gradient boosting decision tree in real time.

a) Gradient boosting, multiple added substance relapse trees, stochastic inclination boosting, and angle boosting machines are some of the terms used to describe this calculation.

b) Boosting is an outfit technique in which new models are added to correct existing models' mistakes. Models are added one by one until no more improvements can be made. The AdaBoost calculation is a popular methodology for loading information that is difficult to predict. Angle boosting is an approach in which fresh models are created that anticipate the residuals or errors of previous models and then added together to get the final expectation. Because it uses an angle plunge calculation to limit the movement, it's termed inclination boosting.

c) Results: XGBoost of Data Mining has been proven that can be effective for the prediction student performance . Using the XG Boost algorithm we have trained the data. Using the predictions on the test data we have attained an accuracy of 0.844.

D. Logistic Regression

a) The logistic model is a widely used factual model in statistics. Its basic structure employs a strategic capability to present a twofold required variable, while more complicated augmentations are available.

Calculated regression is a method of evaluating the parameters of a strategy model in regression analysis (a type of binomial relapse).

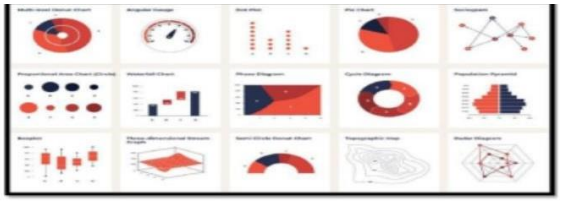
A needed variable with two conceivable qualities, such as pass/fizzle, win/lose, alive/dead, or solid/liquid, is spoken to by a marker variable, where the two qualities are called "0" and "1" in a double strategic model.

b)Results: Logistic Regression of Data Mining has been proven that can be effective for the prediction student performance. We attained an accuracy of 0.802 after training and predicting values using test data.

E.Multiple Linear Regression

The link between two or more independent variables and one dependent variable is estimated using multiple linear regression. Multiple linear can be used. By fitting a line to the observed data, regression models are used to describe relationships between variables. You can use regression to predict how a dependent variable will change as the independent variable changes.

Results: Multiple Linear Regression of Data Mining has been proven that can be effective for the prediction student performance. We attained an accuracy of 0.95.664after training and predicting values using test data.



V. BUILD MODEL

The model building, I the main step in student performance analysis while building the model user use the algorithms.The steps involved are:

Import the packages that are necessary.

```
import pandas as pd
import numpy as np
import pandas_profiling
import sklearn
from sklearn import linear_model
import matplotlib.pyplot as pyplot
import pickle
from matplotlib import style
```

Add data into a Data frame, then get the shape of data.

```
data = pd.read_csv("student-mat.csv", sep=";")

data = data[["G1", "G2", "G3", "studytime", "failures", "absences"]]
#print(data.head())

predict = "G3"
```

Then split the dataset into training and testing datasets.

```
x = np.array(data.drop([predict], 1))
y = np.array(data[predict])
x_train, x_test, y_train, y_test = sklearn.model_selection.train_test_split(x, y, test_size=0.1)


best = 0
for _ in range(30):
    x_train, x_test, y_train, y_test = sklearn.model_selection.train_test_split(x, y, test_size=0.1)
```

Then, fit and transform train and test set.

```
#Declare the models
linear.fit(x_train, y_train)
```

Then identify on test set and calculate the accuracy of the model

```
accuracy = linear.score(x_test, y_test)
# print(accuracy)
```

The accuracy is 95.6%

6.Plotting the data

```
#Plotting the data
p1 = "G2"
p2 = "G3"
style.use("ggplot")
pyplot.scatter(data[p1], data["G3"])
pyplot.xlabel(p2)
pyplot.ylabel("Final Grade")
pyplot.show()
print("mean:")
print(data['G3'].mean())
print("best")
print(best)

#profile=data.profile_report(title="hello")
```

7. sorting and ranking the student data

```
#profile.to_file(output_file="hello.html")
print("data with greater G3")
print(data.sort_values("G3", ascending=False))

print(data["failures"].rank(ascending=False))
data['rank']=data['G3'].rank(ascending=0)
data=data.set_index('rank')
data1=data.sort_index()
print(data1)
```

8. Plotting the rank

```
rank = data1[0]
G3 = data1.iloc[:, 2].values
rank=data1.index.tolist()
pyplot.plot(rank, G3)
pyplot.title('Upprediction')
pyplot.xlabel('rank')
pyplot.ylabel('G3')
pyplot.show()
```

VI. RESULT AND DISCUSSION

The outcome depicts the students' final test achievement based on factors such as study time, failure, absence, grade 1, grade 2, and grade 2. The model correctly predicted the grade 95% of the time. Data mining's Naive Bayes classification has been shown to be effective in predicting student achievement. Naive Bayes has an accuracy of 0.8125 in terms of prediction rate. Data mining's XGBoost has been shown to be effective in predicting student achievement. The data was trained using the XG Boost method. We achieved an accuracy of 0.844 using the predictions on the test data.

.

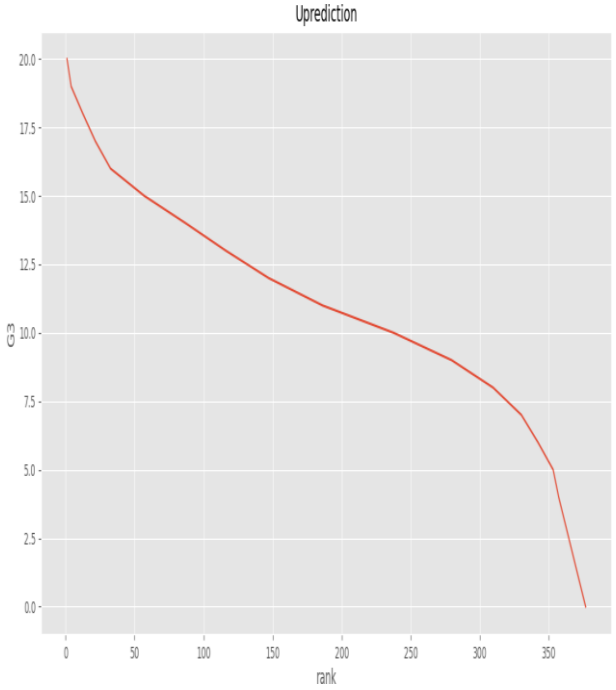
1. Predicted Accuracy

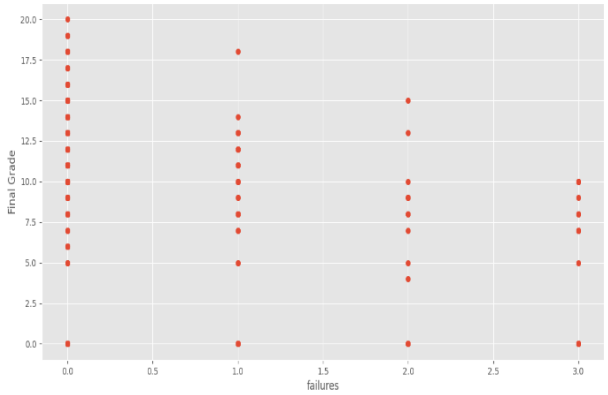
```
Coefficient [ 0.16308114  0.97229111 -0.19248231 -0.28608865  0.04564646]
intercept -1.5727770763520272
10.816810242063571 [ 9  9  2  0 56] 8
6.705607663746122 [6 8 1 1 0] 8
11.079110511998227 [11 11  2  0 12] 11
18.15279643557733 [16 18  2  0  0] 19
11.12193888615516 [ 9 12  3  0  3] 11
12.201768196373498 [10 13  4  0  6] 13
20.100196731431787 [18 19  1  0 10] 19
13.855785036871426 [13 14  3  0  6] 14
14.46572255091108 [15 14  2  0  8] 14
13.002650589286834 [15 13  3  2 14] 13
10.622645916252754 [11 11  2  0  2] 10
13.137867991191044 [14 13  3  0  8] 14
6.821396762830526 [ 7  7  1  0 14] 5
15.164134898682219 [15 15  2  0  2] 16
10.205190716325768 [ 9 11  2  0  0] 12
13.968552919347628 [11 12  2  0 54] 11
4.156079421295693 [6 5 2 0 6] 6
11.960108698067552 [11 12  2  0 10] 13
14.909760839144178 [14 15  2  0  0] 15
7.907863136804924 [10  8  2  0 10] 8
8.100345450917594 [10  8  1  0 10] 9
13.738350356057028 [12 14  3  0  7] 14
11.696126416434475 [11 12  1  0  0] 10
8.280113203798171 [8 9 2 0 4] 10
```

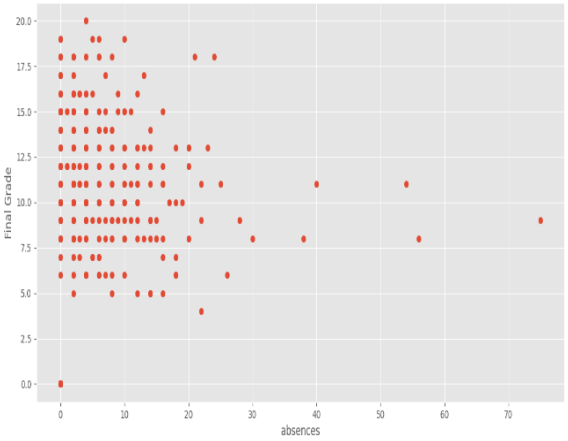
2. Sorting and ranking

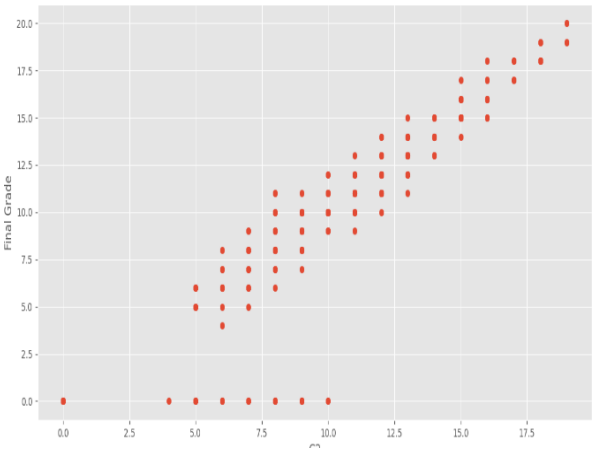
```
10.171970564283598 [10 10  2  0 17] 10
7.705772600597491 [11  8  2  0  2] 8
8.534487263336212 [9 9 2 0 6] 10
mean:
10.415189873417722
best
0.9459084637334887
data with greater G3
       G1  G2  G3  studytime  failures  absences
rank
1.0    19  19  20          4         0         4
4.0    19  18  19          3         0         0
4.0    18  19  19          1         0        10
4.0    18  19  19          1         0         6
4.0    16  18  19          2         0         0
...    ..  ..  ..        ...       ...       ...
376.5  11   0   0          3         0         0
376.5   9   0   0          2         0         0
376.5   8   0   0          1         0         0
376.5   6   5   0          1         3         0
376.5  10   9   0          4         0         0

[395 rows x 6 columns]
```

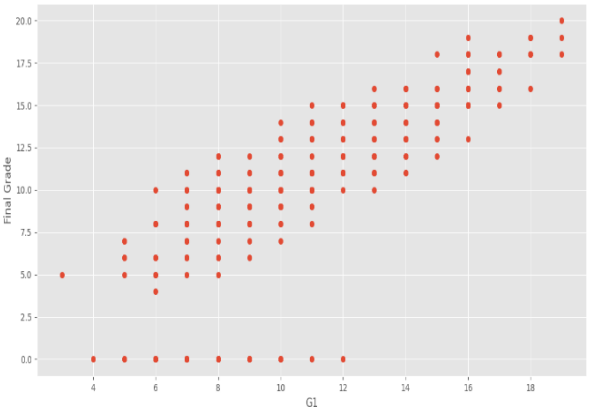
3. plotting the ranks and marks based on performance











VII. CONCLUSION

In this research conducted using the naïve bayes with accuracy 0.8125 and with logistic regression get accuracy 0.802 and using the XG Boost get the result 0.844 .Then when we compare the accuracy of these models is not greater than The accuracy of the model multi linear regression that is 95.6%.so this model is more efficient than other models.

REFERENCES

1. Febrianti Widya ha stuti,VianyUtamiTjhin, " Prediction of Student Performance using Machine Learning":
Volume-8 Issue-6, August, 2019

2. R. D. Ibrahim Z, "Predicting Students' Acdemic Performance:Comparing Artificial Neural Network, Decision Tree and Linear Regression, 2007," in 21st Annual SAS Malaysia Forum,5th September 2007, Shangri-La Hotel, Kuala Lumpur, 2007.

3. W. M. C. Heather J, Nicola Price, "Predictors of academic success in a Bachelor of Nursing course," Elsevier Sciences, Nurse Education Today, vol. 23, pp. 246-254, 2003
.

4. M. A. Wang T, "Using Neural Networks to Predict Student'sPerformance," in Proceedings of the International Conference on Computers in Education (ICCE'02), 2002.

5. I. Z. Rusli N M, Janor RM, "Predicting Students' Academic Achievement: Comparison between Logistic Regression, Artificial Neural Network, and Neuro-Fuzzy," in International Conference on Computers, 2008.