2020 International Conference for Emerging Technology (INCET)
Belgaum, India. Jun 5-7, 2020

Prediction of Student Performance Using Linear Regression

Boddeti Sravani
Bachelor Student, Computer Science and Engineering
Institute of Aeronautical Engineering,
Hyderabad, Telangana, India
sravaniboddeti1502@gmail.com

Myneni Madhu Bala
Professor, Computer Science and Engineering
Institute of Aeronautical Engineering,
Hyderabad, Telangana, India
baladandamudi@gmail.com

***Abstract*— This paper is about how the application of machine Learning have huge impact in teaching and learning for further improvement in learning environment in higher education. Due to the interest of students in online and digital courses increased rapidly websites such as Course Era, Udemy etc became very influential. We implement the new applications of machine learning in teaching and learning considering the students background, students past academic score and considering other attributes. As the sizes of classes are large, it would be difficult to assist each individual student in each open learning course, this can increase the bar of the dropout rate at the end of the course. In this paper we are implementing linear regression which is a machine learning algorithm to predict the student's performance in academics**

Keywords*— *Classification Prediction, Machine Learning, Data cleaning, Data Processing, Linear Regression.

I. INTRODUCTION

As the usage of computers and internet is everywhere, the availability of data that can be analysed rapidly increased. Data can be anything related to population, academic data of students, and interests of people. We can see that new data emerging from time to time. Analysing the data is the difficult task for humans. So here comes the computer which can analyse the data more efficiently than humans because it stores the data digitally in a well-formatted way.

This is where the machine learning emerged. Machine learning is the branch of Artificial Intelligence that provides ability to automatically learn from past experiences. Here the machines do get programmed explicitly. As the name suggests, it gives the ability to the computer that makes humans and machines look alike in the aspect of learning. On the basis of the nature of the learning signal, machine learning is classified into supervised learning and unsupervised learning. This study focuses on supervised learning, more specifically on predictive analysis. Whenever the predictions of future outcomes are done, predictive analysis plays an important role. The range of applications of predictive analysis is very vast. Predicting student's academic performance is very important because it can indicate the teachers about the students who are possibly to drop out from the course and prediction can provide additional assistance to the students who need to improve their academic performance.

This study is on implementation of machine learning in education. The outcome of this study is to predict student's academic performance. The data of students is used to develop a model that can predict student's performance in academics considering some background data of the student. The input data for this study should be student's dataset. This dataset should be taken as a tabular format which contains information related to students (that is age, gender, academics record, medical information). Various algorithms can be used to create the model which gives the output for this thesis. There are many algorithms that are used in predictive models. In this study we focus on how linear regression is implemented to the student's academic performance considering student's dataset

II. LITERATURE REVIEW

To specify the thesis as well-structured idea, I have alluded to many research papers that are similar to the thesis. Conclusion details of few of the papers are as follows. This research study describes how the linear regression approach is used in predicting student's academic performance.

[1] In this research paper author implemented the thesis using SVM approach in java, decision tree, C4.5, Naive Bayes, LibSVM, Logistic Regression and Hybrid approach LMT and compared the accuracy of performance prediction among the hybrid approaches. The above methods are implemented by considering suitable attributes.

[2] In this research paper it is observed that the author used some of the most popular algorithms and regression algorithms. The experiment was carried using administrate data from the University Polo considering 700 courses. The paper concludes that best results are obtained by decision Trees and SVM.The main contribute of this paper is to compare the accuracy levels of different algorithms.

[3] The research is focused on predicting student's performance using personalized analytics. This paper presents two different approaches to work on the thesis. The first approach used by the author is Regression Algorithm, which is one of the data mining function. Error rate of the regression algorithm is also calculated by using the approach called root mean square.

[4] In this paper the author worked on how to improve the prediction algorithms which are used to analyze and predict the student's performance. The work of this paper is carried using decision trees algorithm

[5] This paper proposed the student Academic performance prediction using Support Vector Machine. Here the author compared SVM with other ML techniques such as linear regression, Decision Trees, KNN and concluded that SVM outperforms other ML algorithms.

978-1-7281-6221-8/20/$31.00 ©2020 IEEE

III. METHODOLOGY

The foremost step in the implementation is to collect the data set required for the research work. The methodology is applied to the dataset containing the information of the students. To reduce our analyzation, we can identify the unique attributes from the data set and remove as those cannot be used for analyzation. After collecting the data, the data is transformed into the desired form. This process is called as pre-processing of data. It is the most important step in order to get the particular desired data from the raw data. More the rate of the accuracy of Pre-processing of the raw data, more the rate accuracy of suitable data.

The next step after pre-processing the data is to find the incomplete, irrelevant data in the dataset and remove it in order to obtain the accurate results of the work. Removing of the unwanted data phase is called as Data Cleaning. Next, we can choose any one of the algorithms such as linear regression, Support vector machine, Naive Bayes Standard Classification, decision tree algorithms for better classification.

Here, in this paper linear regression algorithm is chosen for the implementation. Further, we have to choose training set from the dataset and identify the Result attributes that decides the output and start classification.

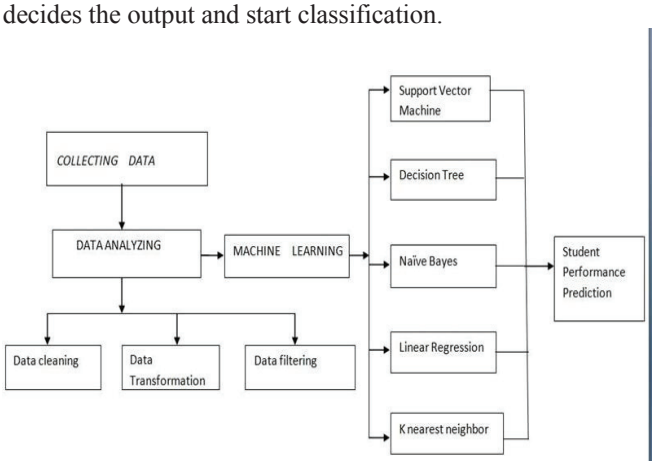

Fig. 1. Implementation of the model

IV. DATA DESCRIPTION

The data used in this paper is the sample data. The sample dataset is comprised of 100 students. In this study we deal with 100 instances and 10 attributes. All the dependent variables and independent variables are given in figure2.

S.NO	VARIABLE	DESCRIPTION	POSSIBLE VALUES	DATA TYPE
1	GENDER	GENDER OF THE STUDENT	F,M	CHAR
	AGE	AGE OF THE STUDENT	15,16,17,18	INTEGER
2	PARENT_EDUCATION	EDUCATION OF THE PARENTS	ASSOCIATE'S DEGREE,BACHELOR'S DEGREE,HIGHSCHOOL,MASTER'S DEGREE	CHAR
3	TEST_PREP	WHETHER THE STUDENT'S PREPARATION IS COMPLETED OR NOT	COMPLETED,NONE	CHAR
4	FAM_SIZE	SIZE OF THE FAMILY	2,3,4,5,6	INTEGER
5	FATHER_JOB	OCCUPATION OF THE FATHER	AT_HOME,HEALTH,OTHER,SERVICES, TEACHER	CHAR
6	MOTHER_JOB	OCCUPATION OF THE MOTHER	AT_HOME,HEALTH,OTHER,SERVICES, TEACHER	CHAR
7	ABSENT_DAYS	NO.OF DAYS THE STUDENT IS ABSENT	4,5,6,7,10,12,13,14,15,16,17,18,22,23,24,25,26,27 ,28,30,35,36,4,45,46,54,56	INTEGER
8	PARENT_STATUS	STATUS OF THE PARENTS RELATIONSHIP	TOGETHER,APART	CHAR
9	TRAVEL_TIME	TIME TAKEN BY THE STUDENT TO TRAVEL	1,2,3	INTEGER
10	ACADEMICSCORE	ACADEMICSCORE OF THE STUDENT	30,40,50,60,70,80,90	INTEGER

Fig. 2. Student related dataset

Here the attributes used in the dataset are gender, age, parent education, family size, test preparation, father job, mother job, absent days, parent status, travel time, academic scores.

V. ALGORITHM USED

There are many algorithms that are used to implement the thesis. In this thesis we use linear regression. Though they are all used to predict the dependent variable based on independent variables, they differ in implementation of the algorithm.

A. Linear Regression

Linear regression is one of the machine learning algorithms. It is based on supervised learning is one of the algorithms that is widely known and it is easily understood even by the person who is not so familiar with machine learning algorithms. As the name suggests linear regression performs regression. It defines the relationship between the two variables by fitting regression line to the data.one of the two variables is dependent variable which is dependent on another variable called as independent variable. One should make sure that there exists a relationship between the dependent and independent variables before modelling. Strength of the relationship between the variables can be known by using the scatterplot.

Linear regression line is represented in the form of: Y=a*X+b

- Y-Dependent Variable
- a-slope
- X-Independent Variable
- b-Intercept

With the best fit regression line to the data the error rate between the predicted and true values can be minimized. Linear Regression is classified into two types. One of it is Simple Linear Regression, in which only one independent variable is used and the second type of regression is Multiple Linear Regression. In this type of regression multiple independent variables are used which we are presently using for the thesis.

B. Implementation

AGE	
LENGTH	100
CLASS	CHARACTER
MODE	CHARACTER
MEAN	15.55
3RD QU	16.00
MAX	18.00

GENDER	
Length	100
Class	CHARACTER
Mode	CHARACTER

FAM_SIZE	
Min.	2.00
1st Qu.	3.00
Median	4.00
Mean	4.07
3rdQu.	5.00
Max.	6.00

TRAVEL_TIME	
Min.	1.00
1st QU.	1.00
Median	2.00
Mean	1.71
3rdQu.	2.00
Max.	3.00

Fig. 3. Description of variables

The first step is to read our data set into R and explore its summary and structure. The summary () functions provides us with information on each variable such as type of data: character, numerical, and if numerical, we find basic descriptive statistics such as measure of central tendency and spread. It also provides as with information on missing values (NA values).

C. *Data Visualization*

The primary purpose of Visualization is to find visual patterns. We are going to plot academic score versus gender, age, parent status using GGPLOT package. GGPLOT requires three key components:

- Define data in form of data frame
- Describe aesthetics for the visualization or how to map the attributes.
- Define the geometry or type of graphics to be used

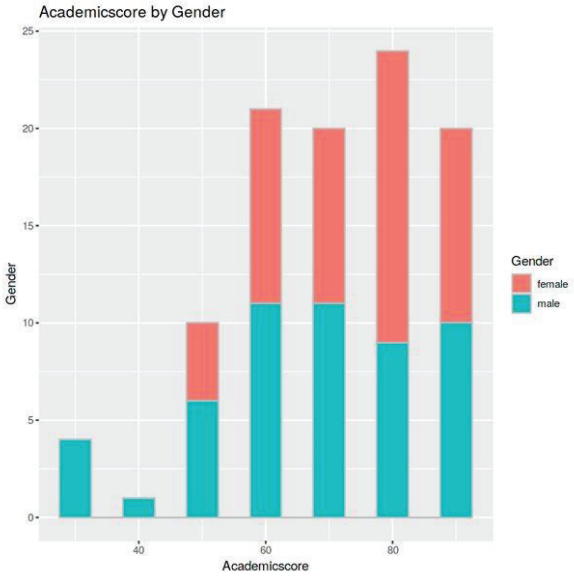

Fig. 4. Data visualization between Academic scores and gender of the students.

We can observe the percentage of the academic scores regarding the gender of the student. We can visualize the academic scores with every attribute we considered using the GGPLOT.

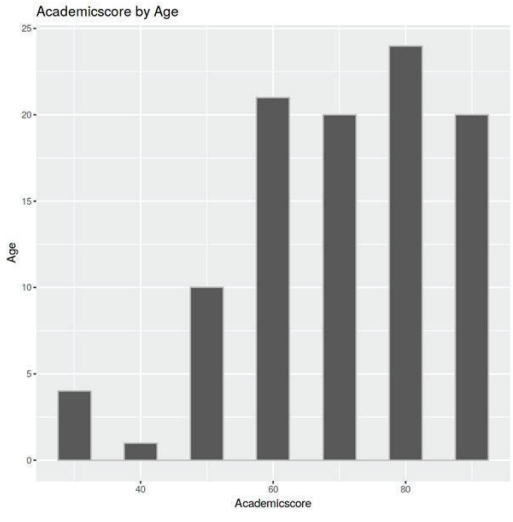

Fig. 5. Data visualization between the academic scores and age of the students

D. *Boxplot Representation*

Box plot is a graph that provides us with the measures of central tendency, spread and visual of out-liners:

- Median: The middle value of the data set.
- First Quartile: The middle number between the smallest number and the median of the dataset.
- Third Quartile: the middle value between the median and the highest value of the dataset.
- Interquartile range: 25th to the 75th percentile.
- outliers, maximum, minimum.

Summary for boxplot visualizations:

- students who completed the prep class had better academic scores.
- Female students scored more than the male students.

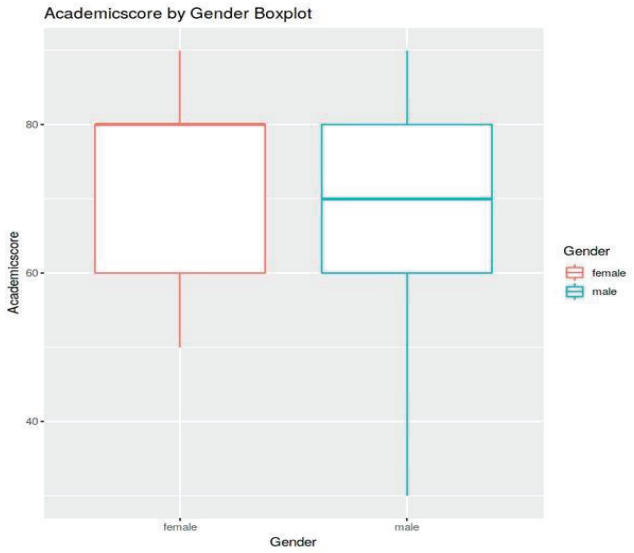

Fig. 6. Boxplot representation of the academics and the gender.

Fig. 7. Boxplot representation of academics and age of the student.

VI. RESULT

In this section we are going to build a linear regression model, predicting academic scores. Academic scores-dependent variable(Y), Gender, Age, Mother job, Parent Education, Test Prep, Father job, Parent status, travel time, absent days independent variables (X). First, we will divide our dataset into two parts as a training and testing data sets. Then, we will apply lm() function on" training" data and predict() function on" testing" data, and create a visualization of our regression model with regression line and 95% confidence intervals.

TABLE I. VALUES OF THE VARIABLES

Variable	Estimate	Error	t value	Pr(>|t|)
Intercept	63.8522	51.4755	1.240	0.221
Gender male	3.6258	4.3898	-0.826	0.413
Age	0.1288	3.1379	0.041	0.967
Parent_Educationbachelor's	-1.8353	6.0710	-0.302	0.764
Parent_Educationhigh school	2.7853	6.6926	0.416	0.679
Parent_Educationmaster's	-2.4314	6.7567	-0.360	0.721
Test_Prepnone	-1.4164	4.6634	-0.304	0.763
Fam_size	1.2378	2.5392	0.487	0.628
Father_jobHealth	-2.6580	10.5445	-0.252	0.802
Father_jobOther	7.9223	9.6146	0.824	0.414
Father_jobServices	3.4360	10.5184	0.327	0.745
Father_jobTeacher	-4.8843	9.7565	-0.501	0.619
Mother_jobHealth	-5.4085	7.8407	-0.690	0.494
Mother_jobOther	2.7027	6.5788	0.411	0.683
Mother_jobServices	3.9748	6.2236	0.639	0.526
Mother_jobTeacher	5.5508	6.1723	0.899	0.373
Parent_statusTogether	3.1105	4.4013	0.707	0.483
Travel time	-1.8446	3.2965	-0.560	0.578

The table describes estimated value, error value, tvalue. Here tvalue refers to the value of relative difference of the variation in the data.

The P value defines the statistically predictive capability of the independent variable. Probability of the predictive capability and the impact of the variable on the output are inversely related.

TABLE II. VALUES OF FITTING OF THE DATA

FIT	LWR	UPR
Min. :60.00	Min. :22.65	Min. : 97.25
1st Qu.:67.50	1st Qu.:31.45	1st Qu.:105.07
Median :72.81	Median :36.37	Median :109.29
Mean :72.04	Mean :35.18	Mean :108.91
3rd Qu.:75.81	3rd Qu.:38.98	3rd Qu.:113.15
Max. :83.73	Max. :48.34	Max. :120.85

The prediction interval is rather similar to the confidence interval in calculation. The prediction interval equation is defined as:

$$Y_{h\square} \pm t_{\alpha/2\underline{n}-2}\sqrt{MSE(1+\frac{1}{n}+\frac{(x_k-x)^2}{\sum(x_i-x)^2})}$$

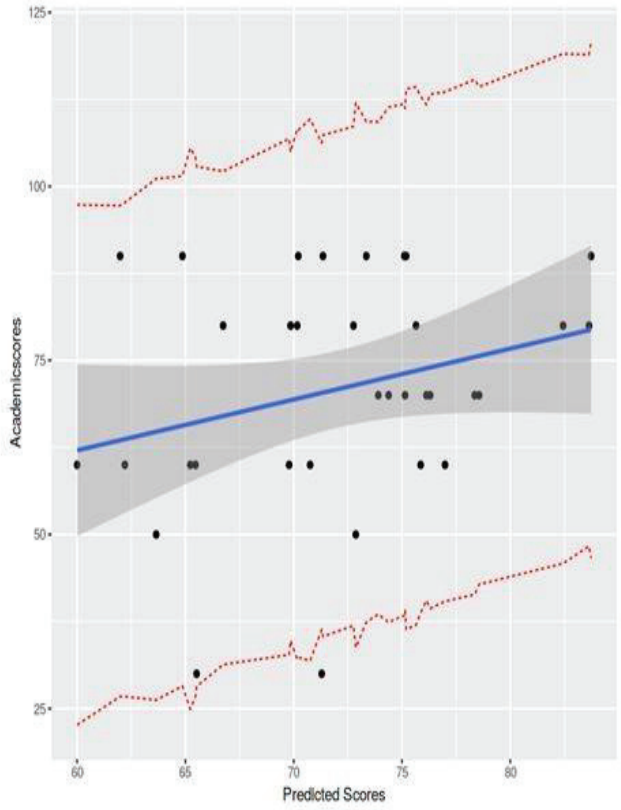

Fig. 8. Visualization of predicted scores

VII. CONCLUSION

The effectiveness of using machine learning in education field depends on the algorithm and the usage of the data. Choosing the algorithm to implement predicting the students' performance is important. The accuracy of the result depends on the Machine learning algorithm. The algorithm used to prove the thesis in this paper is Linear Regression. Present studies show that academic performances of the students are also dependent on student's background and other attributes. Many research works confirm that apart from the past academic performances student's background and other attributes indeed got a significant influence over students' performance. Machine learning has an emerging role in recent times in every sector, and it can also be used effectively in academia. In the future, many applications with improved ability and efficiency may become an integrated part of every academic institutions.

ACKNOWLEDGMENT

This research contribution is a part of undergraduate research and content development at Institute of Aeronautical Engineering.

REFERENCES

[1] Amandepp Kaur, Nitin Umesh, Barjinder Singh" Machine Learning approach to predict Student Academic Performance, International Journal for Research in Applied Science Engineering Technology (IJRASET), Volume.6 Issue IV, April 2018.

[2] Pedro Strecht, Luis Cruz, Carlos Soares, João Mendes-Moreira and Rui Abreu "A comparative study of classification and regression algorithms for Modelling student's Academic performance", Proceedings of the 8th International Conference on Educational Data Mining,2015.

[3] G. Sujatha, S. Sindhu and P. Savaridassan "Predicting student's performance using personalized analytics", Volume.119 No. 12, 2018.

[4] Ankitha A Nichat, Dr. Anjali B Raut "predicting and Analysis of student Performance Using Decision Tree Technique", International Journal of Innovative Research in Computer and Communication Engineering.Vol.5, Issue 4, April 2017.

[5] S.A. Oloruntoba, J.L. Akinode "Student Academic Performance Prediction Using Support Vector Machine", December, 2017

[6] Dhanashree Mane, Pranali Namdas, Pooja Gargade, Dnyaneshwari Jagtap, S.S. Rathi" Predicting student Performance Using Machine Learning Approach". VJER Vishwakarma Journal of Engineering Research, Volume 2 Issue 4, December 2018.

[7] Havan Agarwal, Harshil Mavani" Student Performance Prediction Using Machine learning", International Journal of Engineering Research and Technology (Ijert), Vol. 4 Issue 03, March-2015.

[8] Raheela Asif, Agathe Merceron, and Mahmood. K Pathan, "Predicting student academic performance at degree level: A case study, International Journal of Intelligent Systems and Applications" Vol.7, No.1, 2014.

[9] Murat Pojon "Using Machine Learning to Predict Student Performance" June 2017.

[10] Erkan Er" Identifying At-Risk Students Using Machine Learning Techniques: A Case Study with IS 100", International Journal of Machine Learning and Computing, Vol.2, No.4, August 2012