Student Academic Performance Prediction Leveraging Linear Regression and Multi-Class Features as Descriptive Features

Victor Solomon
PhD student, Department of Computer Science,
Georgia State University
vsoldev@gmail.com

Walid Abdullahi
BSc student, Department of Computer Science,
Georgia State University
wabdullahi0529@gmail.com

***Abstract*— This work is an attempt to use machine learning algorithms to build predictive models that can be utilized to predict the academic performance of a student in a course. We collected data from Kaggle which contains 32 features, 3 of which are target class features. This study can help academic institutions take proactive actions that would help students to perform better in their academics. In this paper, we use Linear Regression to predict the performance of the student by assigning the predictions made by the model for previous grades in the course- this is a novel approach we have considered in this domain of research- academic performance prediction.**

Keywords— Linear Regression, Prediction, Machine Learning, Data Processing.

I. BACKGROUND STUDY AND INTRODUCTION

The advancement of technology and presence of the web has led to a great availability of data in our contemporary world. As a result of this, tons of data is generated on a regular basis in different areas where data is generated. It is said that we are in the information age while in fact we are in the data age because although we have unlimited access to data, they are of no purpose if we can not analyze the humongous data sts to help us make decisions. In this study, we go ahead to take student data from an academic environment which have been gathered by their academic involvements and personal records to make predictions of the score a student will have in a course. Machine learning algorithms can take data as input to induce machine learning models that are capable of making predictions. Machine learning algorithms can be supervised , unsupervised and semi-supervised. In this study, we focus on supervised learning and particularly we use the linear regression algorithm. Predicting the performance of a student will go a long way to help the school authority and teachers to be able to know which students will need help in their course before they get deep into the course module.

Having a model such as the one proposed in this study can help build a robust student performance that would help academic institutions monitor student's performance and take precautionary measures proactively when necessary..

II. RELATED WORK

The study carried out in this paper is an interesting one as it leverages some analytical techniques to improve on the vanilla linear algorithm but prior work has been conducted by other researchers in the area of predicting student academic performance. We go ahead to show some of these works and what their authors aimed to achieve in their individual studies.

Raza et al. [1] proposed a methodology to predict student performance using decision trees. Their study showed that the Random forest algorithm outperforms decision trees for prediction as it gives 100% accuracy and they also used the WEKA in the data mining process. Myneni et al [2] induces a predictive model from student data to predict the performance of students using linear regression

algorithm. The linear regression algorithm is an exploratory machine learning algorithm used in predictive analytic that predicts the value of unknown data also called a dependent variable by using another related and known data value known as independent variable. It induces models that are linear functions to get the line of best fit for the data. Muhammad et al. [3] attempted to build models that would predict student scores by using supervised learning techniques. Three supervised learning algorithms were experimented upon and the J48 algorithm showed a higher level of accuracy. Bilal et al. [4] proposed a student performance prediction model which pointed out the possible reasons students fail a class. Other models that were evaluated included logistic regression, CART algorithm, and naïve Bayes. In their comprehensive study, Xinnning et al [5] proposed a comprehensive and high-performance prediction model to probe Student Academic Performance (SAP) which they referred to as (ProbSAP). They also applied efficient methods for improving on the issue of imbalance of data in student academic performance prediction .

III. METHODOLOGY AND DATA

The initial step in the implementation of this project involves gathering the necessary dataset for the research. The suitable dataset for this work must be one containing student information. To streamline our analysis, we try to pinpoint unique attributes within the student dataset and eliminate those unsuitable for analysis- this is done through a feature selection process. Once the data is collected, it undergoes a transformation into the desired format, a crucial process known as data pre-processing. This step is paramount in extracting specific desired data from the raw dataset; the higher the accuracy of pre-processing the raw data, the greater the accuracy of the resulting suitable data.

Following pre-processing, the subsequent step is to identify and eliminate incomplete or irrelevant data within the dataset to ensure the accuracy of the research outcomes. This phase, known as data cleaning, involves removing unwanted data. Subsequently, a choice of algorithms, such as linear regression, Naive Bayes classification, or decision tree algorithms, can be made for improved classification. In this paper, the linear regression algorithm is selected for implementation.

Moreover, it is necessary to select a training set from the dataset and identify the result attributes that determine the output to initiate the classification process.

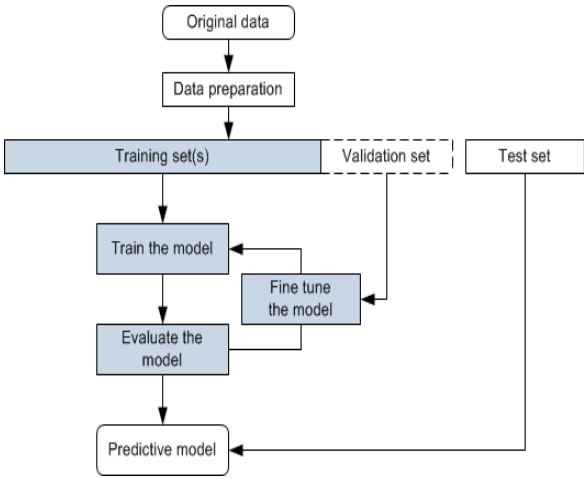

Fig 1. Implementation of the model

The information utilized in this manuscript corresponds to the sample dataset, which consists of 612 instances, each characterized by 9 attributes. Figure 2 provides a comprehensive overview of both the dependent and independent variables.

In this dataset, the included attributes encompass Mother's education, Failures, Study time, Absences, Grade 1, and Grade 2 and a final grade- G3. Several variations of the Linear regression algorithm are employed in executing this project.

A. Implementation

Linear regression stands as one of the machine learning algorithms, operating on the principles of supervised learning. Widely recognized, it is known for its accessibility, making it comprehensible even for those not well-versed in machine learning. True to its name, linear regression is designed for regression tasks, delineating the relationship between two

variables by fitting a regression line to the data. The dependent variable relies on another variable, termed the independent variable.

Before delving into modeling, it is imperative to ensure a discernible relationship between the dependent and independent variables. The strength of this relationship is often assessed through a scatter plot. The linear regression line is mathematically expressed as:

$$Y = m * X + b,$$

where Y is the dependent variable, X is the independent variable, m is the slope, and b is the intercept.

By optimizing the fit of the regression line to the data, the error rate between predicted and true values can be minimized. Linear regression is dichotomized into two categories: Simple Linear Regression, utilizing a single independent variable, and Multiple Linear Regression, the latter of which involves multiple independent variables—a characteristic pertinent to our present thesis.

The initial stage involves importing our dataset into Python and examining its summary and structure. The 'describe()' function furnishes details about each variable, including the data type (character or numerical). For the numerical variables, it offers basic descriptive statistics, encompassing measures of central tendency and spread. Additionally, the function informs us about the presence of missing values.

The novel approach we try to implement in this project uses other target features as input descriptive features to our linear regression algorithm and then we compare the results to see if the accuracy and error was better.

S/NO	VARIABLE	DESCRIPTION	RANGE	DATA TYPE
1	Failures	Number of previous failures	0 - 3	Integer
2	Medu	Mother's education level	0 - 4	Integer
3	Studytime	Amount of time used to study	1 - 4	Integer
4	Absences	number of times student was absent from class	0 - 30	Integer
5	G1	Grade in the first exam	0 - 20	Integer
6	G2	Grade in the second exam	0 - 20	Integer
7	G3	Grade in the final exam (target value)	0 - 20	Integer

Fig 2. Implementation of the model

B. Data Visualization

The main objective of Visualization is to identify visual patterns. We intend to create plots of academic scores versus gender, age, and parent status histogram and show distributions of data using boxplots.

We show a few visualizations for the sake of this text. They include boxplots and histograms. Data visualization tools provide better information about a dataset. It shows whether the data is skewed, if there are outliers etc.

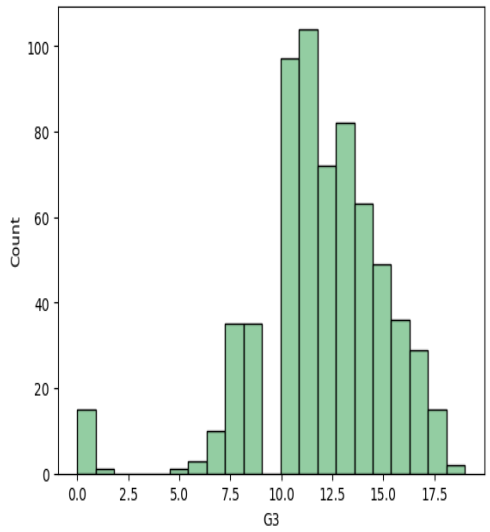

Fig. 3. Histogram of the distribution of the target class feature.

From the histogram in Fig 3, it can be seen that the distribution for the target class is skewed to the right which is a good thing by considering domain knowledge- since a student's score should be about average or more in a typical scenario.

Again, we show a box plot visualization of the fathers education in Fig 4 below.

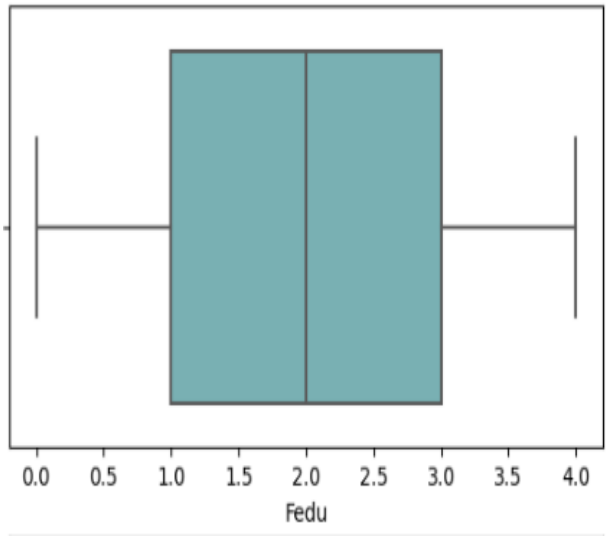

Fig. 4. Box Plot visualization of the father education feature

A box plot is a graphical representation that conveys information about central tendency, spread, and the visualization of outliers. It includes the following components:

- Median: The central value of the dataset.
- First Quartile: The middle number between the smallest value and the median.
- Third Quartile: The middle value between the median and the highest value.
- Interquartile Range: The range between the 25th and 75th percentiles.
- Outliers, maximum, minimum.

Key observations from boxplot visualization is that students whose fathers have higher education seemed to have better performance in class.

IV. RESULTS

Within this section, we will construct a linear regression model aimed at predicting academic scores. The dependent variable (Y) is final grade- G3, while the independent variables (X) include Mother's education, Failures, Study time, Absences, Grade 1, and Grade 2. Initially, we will partition our dataset into training and testing sets using the K-Fold Cross validation technique. Subsequently, we train the model using the training data for each partition, and the predict for the test data. Again we show that the Bagging variation of the model performs better than the Base Linear Regression Model.

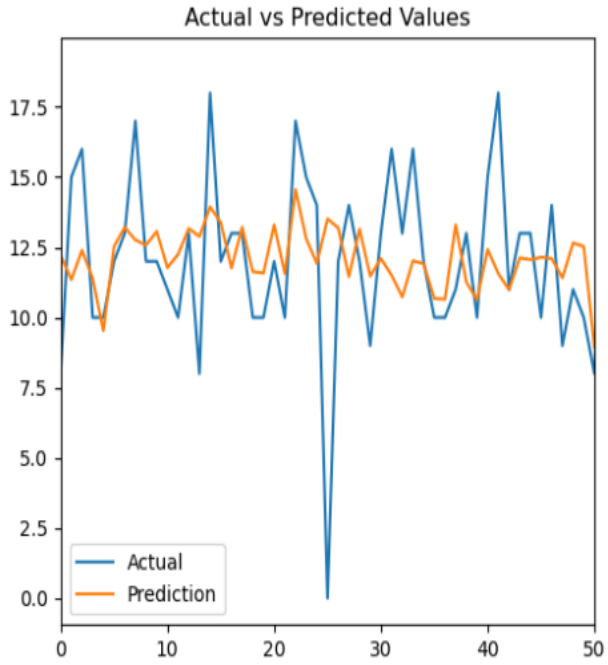

Fig 5. Actual vs Predicted value using the current approach

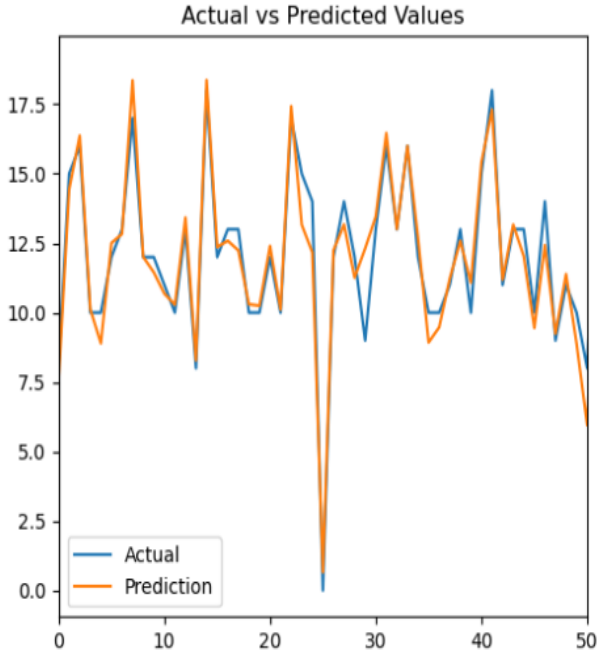

Fig 6. Actual vs Predicted value using our novel approach.

From Fig 5 and Fig 6 above, it can be seen that the new approach that has been implemented performs way better than the previous approach. This goes to show that when we have multi-classed target predictions, we can always attempt to introduce some or all of the target features as descriptive features to a machine learning algorithm for better performance.

We have also shown a table in Fig 7 showing the performance of the different models used in this project. From the table, it can be said that our approach performs better with a lower RMSE and thus this project was effective.

Model	**RMSE**	**R2-Score**
Linear Regression (Initial)	3.09	0.19
Base-Linear Regression(novel)	1.59	0.78

V. CONCLUSION

The efficacy of applying machine learning in the field of education relies on both the choice of algorithm and how the data is utilized and sampled. Selecting an algorithm for predicting students' performance is a crucial decision, as the accuracy of the outcome hinges on the machine learning algorithm employed. The algorithm utilized to substantiate the idea in this paper is Linear Regression leveraging the concept of taking a couple of target features as descriptive features to enhance the performance of the model.

Machine learning has progressively gained prominence across various sectors, including academia. It is anticipated that in the future, applications leveraging enhanced capabilities and efficiency will become integral components of academic institutions.

REFERENCES

[1] Raza et al " Student Academic Performance Prediction by using Decision Tree Algorithm", 4th International Conference on Computer and Information Sciences (ICCOINS), 2018.

[2] Boddeti et al " Prediction of Student Performance Using Linear Regression", International Conference for Emerging Technology (INCET), 2020.

[3] Muhammad et al "Student Academic Performance Prediction using Supervised Learning Techniques", International Journal of Emerging Technologies in Learning (iJET), 2019.

[4] BIlal et al "Student Performance Prediction and Risk Analysis by Using Data Mining Approach", Journal of Intelligent Computing Volume 8 Number 2, 2017.

[5] Xinning et al "ProbSAP: A comprehensive and high-performance system for student academic performance prediction", Elsevier Publishing, 2023.