Available *online* at : http://jitce.fti.unand.ac.id/

JITCE (Journal of Information Technology and Computer Engineering)

| ISSN (Online) 2599-1663 |

Research Article

Analysis and Prediction of Student Academic Performance Using Machine Learning

Ajibola O. Oyedeji, Abdulrazaq M. Salami, Olaolu Folorunsho, Olatilewa R. Abolade

Department of Computer Engineering, Olabisi Onabanjo University, Ago-Iwoye, Nigeria

ARTICLE INFORMATION

Received: January 25th, 2020
Revised: February 24th, 2020
Available online: march 30th , 2020

KEYWORDS

Student's performance, machine learning, linear regression, neural network, MAE.

CORRESPONDENCE

Phone: +234(905)6728637

E-mail: oyedeji.ajibola@oouagoiwoye.edu.ng

ABSTRACT

Analyzing the academic performance of students is of utmost importance for academic institutions and educationists, so as to know the ways of improving individual student's performance. The project analyzed the past results of students including their individual attributes including age, demographic distribution, family background and attitude to study and tests this data using machine learning tools. Three models which are; Linear regression for supervised learning, linear regression with deep learning and neural network were tested using the test and train data with the Linear regression for supervised learning having the best mean average error (MAE).

INTRODUCTION

Student performance is of deep concern in the educational institutes where various factors may affect the performance of students. Predicting student academic performance has long been an important research topic in many academic disciplines [1] [2]. Based on the results of a predictive model, the instructor can take proactive measures to improve student learning, especially for those low-performing students [3] [4].

Data science and Machine learning over the years have proven very efficient and decisive in many sectors including education. This project is all about analysis of the performance of some students for a particular semester considering different factors ranging from psychological, personal, and environmental, etc. and prediction of their performance for another semester using Python programming language for data science and Machine learning [4]. Data science is a multidisciplinary field dealing with structured and unstructured data, using scientific methods, processes, algorithms, and systems to extract knowledge and insights from the data. Data science comprises the field of data mining and big data [1] [5].

Machine learning is an aspect of Artificial Intelligence (AI) in which a computing system is able to learn from data and make decisions [6]. Machine learning has become one of the most important fields within development organizations that are looking for innovative ways to grasp data assets to help the business attain a new level of understanding. Applications of machine learning include but are not limited to fraud detection, prediction of equipment failures, pattern and image recognition, etc. [7] [8].

The aim of this paper is to test the performance of different predictive models on the students' performances and to accurately predict students' course scores.

METHOD

Model Selection

The three (3) regression algorithms or models to be used for the purpose of this project are listed below;

Model 1 - Linear regression for supervised learning implemented with scikit-learn.

Model 2 - Linear regression for deep learning implemented with TensorFlow and Keras.

Model 3 - Neural network for deep learning implemented with TensorFlow and Keras.

https://doi.org/10.25077/jitce.4.01.10-15.2020

Model Training and Building

Before selecting an algorithm, applying a practical approach that can be used to solve most machine learning problems results in the machine learning process as shown in Figure 1 below.

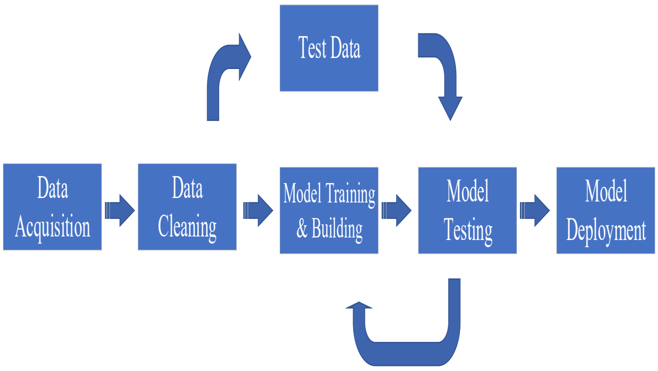


Figure 1. Machine Learning Process

Data Analysis: In this step, descriptive statistics with the aid of visualization and plots is performed.

Data Cleaning: include pre-processing, profiling, cleaning.

Data Transformation: involves transforming data from a raw state to a state suitable for modeling is where feature engineering fits in.

Model Selection: After categorizing the problem and understand the data, the next milestone is identifying the algorithms that are applicable and practical to implement in a reasonable time.

Implement the Model: Set up a machine learning model that analyzes the performance of each algorithm on the dataset using a set of carefully selected evaluation criteria.

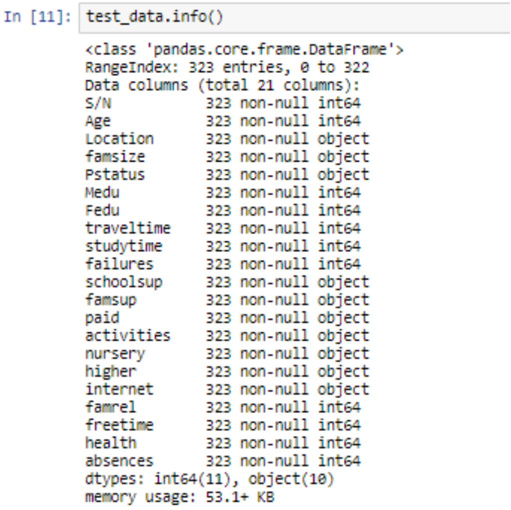
Data Collection and Descriptive Analysis

The data for this project was collected from Kaggle with 648 data set and twenty two 22 attributes. The data was divided into 2 parts with 325 data sets for training titled Train.CSV and 323 serving as test data named Test.CSV. This data captures the performance of randomly selected students. The data attributes are shown as presented in Table 1 below. The student performance dataset contains twenty two (22) factors ranging from psychological, personal and environment. The factors include the level of student attendance, distance from home to school, reading hours, educational support, health status, Father's and mother's education level and more.

Table 1. The Student Data Set Attributes

Attribute	Description
Gender	Student's gender (binary: Female 'F' - 0 or Male 'M' - 1)
Age	Student's age (numeric: from 10 to 17)
Location	Student's home address type (binary: Urban 'U' - 1 or Rural 'R' - 0)
Famsize	Family size (binary: 'LE3' - less or equal to 3 or 'GT3' - greater than 3)
Pstatus	Parent's cohabitation status (binary: Living Together 'T' - 1 or Apart 'A' - 0)
Medu	Mother's education (numeric: 0 - none, 1 – Lower Primary, 2 - Upper Primary to JSS3, 3 - SSCE level or 4 Higher education)
Fedu	Father's education (numeric: 0 - none, 1 – Lower Primary 2 Upper Primary to JSS3, 3 SSCE level or 4 Higher education)
traveltime	Home to school travel time (numeric: 1 - <15 min., 2 - 15 to 30 min., 3 - 30 min. to 1 hour, or 4 - >1 hour)
studytime	Weekly study time (numeric: 1 - <2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - >10 hours)
Failures	Number of past class failures (numeric: n if 1<=n<3, else 4)
schoolsup	Extra educational support (binary: 'yes' - 1 or 'no' - 0)
Famsup	Educational support from family (binary: 'yes' - 1 or 'no' - 0)
paid	Extra paid classes within the course subject (binary: 'yes' - 1 or 'no' - 0)
activities	Extra-curricular activities (binary: 'yes' - 1 or 'no' - 0)
nursery	Attended nursery school (binary: 'yes' - 1 or 'no' - 0)
higher	Desire for higher education (binary: 'yes' - 1 or 'no' - 0)
internet	Internet access at home (binary: 'yes' - 1 or 'no' - 0)
famrel	The quality of family relationships (numeric: from 1 - very bad to 5 - excellent)
freetime	Free time after school (numeric: from 1 - very low to 5 - very high)
Health	health status of student (numeric: from 1 - very bad to 5 - very good)
absences	Number of school absences (numeric: from 0 to 93)
Score	(numeric, 0-60)

The summary of the data used are presented in figures 2 and 3 below. Figures 2 and 3 show the result of exploratory analysis done indicating each information of each column in the train and test dataset e.g. the number of columns, type of value in each column (int or object).

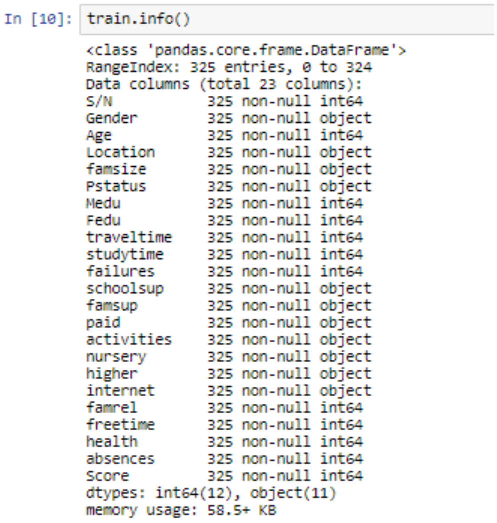


```
In [11]: test_data.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 323 entries, 0 to 322
Data columns (total 21 columns):
S/N            323 non-null int64
Age            323 non-null int64
Location       323 non-null object
famsize        323 non-null object
Pstatus        323 non-null object
Medu           323 non-null int64
Fedu           323 non-null int64
traveltime     323 non-null int64
studytime      323 non-null int64
failures       323 non-null int64
schoolsup      323 non-null object
famsup         323 non-null object
paid           323 non-null object
activities     323 non-null object
nursery        323 non-null object
higher         323 non-null object
internet       323 non-null object
famrel         323 non-null int64
freetime       323 non-null int64
health         323 non-null int64
absences       323 non-null int64
dtypes: int64(11), object(10)
memory usage: 53.1+ KB
```

Figure 2. Test Data Attributes Information

https://doi.org/10.25077/jitce.4.01.5-9.2020



```
In [10]: train.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 325 entries, 0 to 324
Data columns (total 23 columns):
S/N           325 non-null int64
Gender        325 non-null object
Age           325 non-null int64
Location      325 non-null object
famsize       325 non-null object
Pstatus       325 non-null object
Medu          325 non-null int64
Fedu          325 non-null int64
traveltime    325 non-null int64
studytime     325 non-null int64
failures      325 non-null int64
schoolsup     325 non-null object
famsup        325 non-null object
paid          325 non-null object
activities    325 non-null object
nursery       325 non-null object
higher        325 non-null object
internet      325 non-null object
famrel        325 non-null int64
freetime      325 non-null int64
health        325 non-null int64
absences      325 non-null int64
Score         325 non-null int64
dtypes: int64(12), object(11)
memory usage: 58.5+ KB
```

Figure 3. Training Data Attributes Information

Figures 4 and 5 shows the results of the descriptive analysis performed on the test data set. Figure 4 presents the gender distribution plot of the total count while Figure 5 shows the age distribution of the students between the ages of 10 to 17 with the highest frequency being 11 years old.

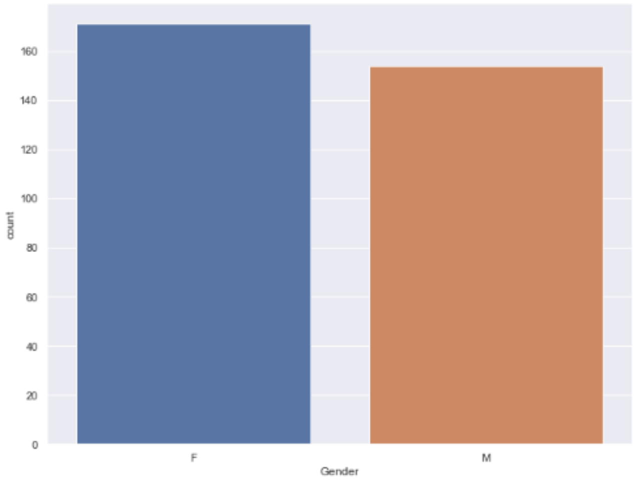


Figure 4. Frequency Plot for Gender of Students

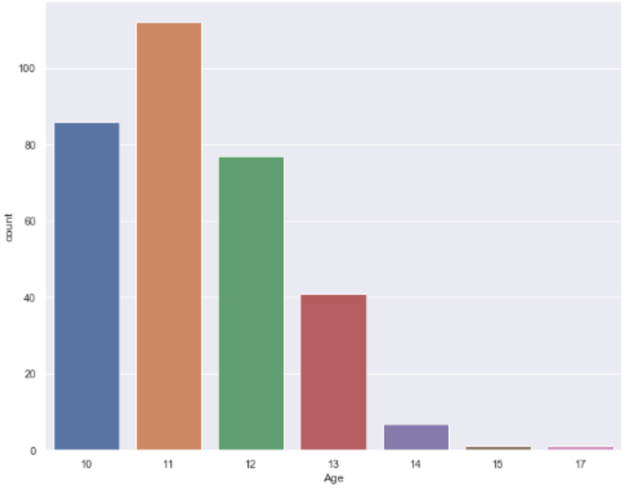


Figure 5. Frequency Plot for Age of Students

The performance was based on 60 pass mark for the particular semester. The grades are as follows **A** = {50-60}, **B** = {40-49}, **C** = {30-39}, **D** = {20-29}, **E** = {10-19}, **F** = {0-9}. Where **A-E** is **PASS** and **F** is **FAIL.** The breakdown of the student performances is shown in Figure 6 with no record of students scoring E and F.

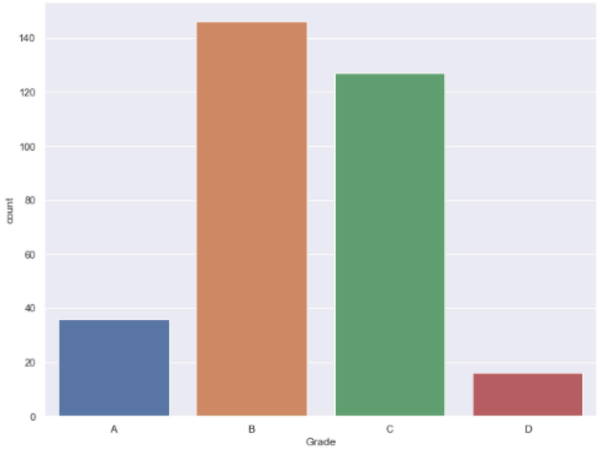
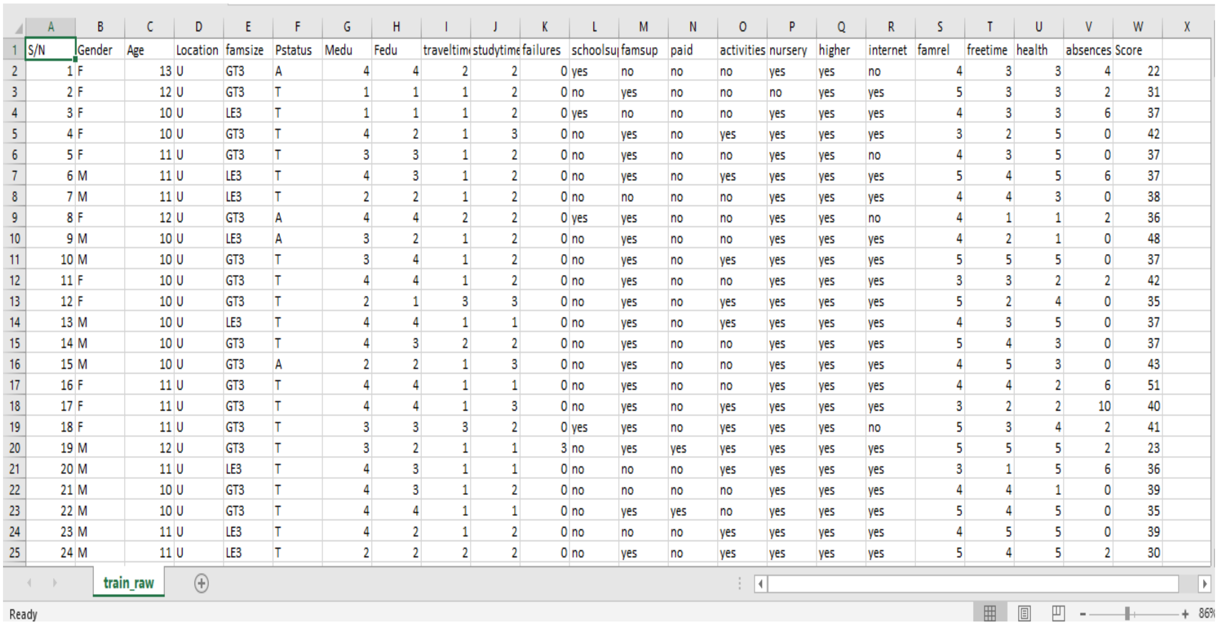


Figure 6. Frequency Plot for Grades of Students

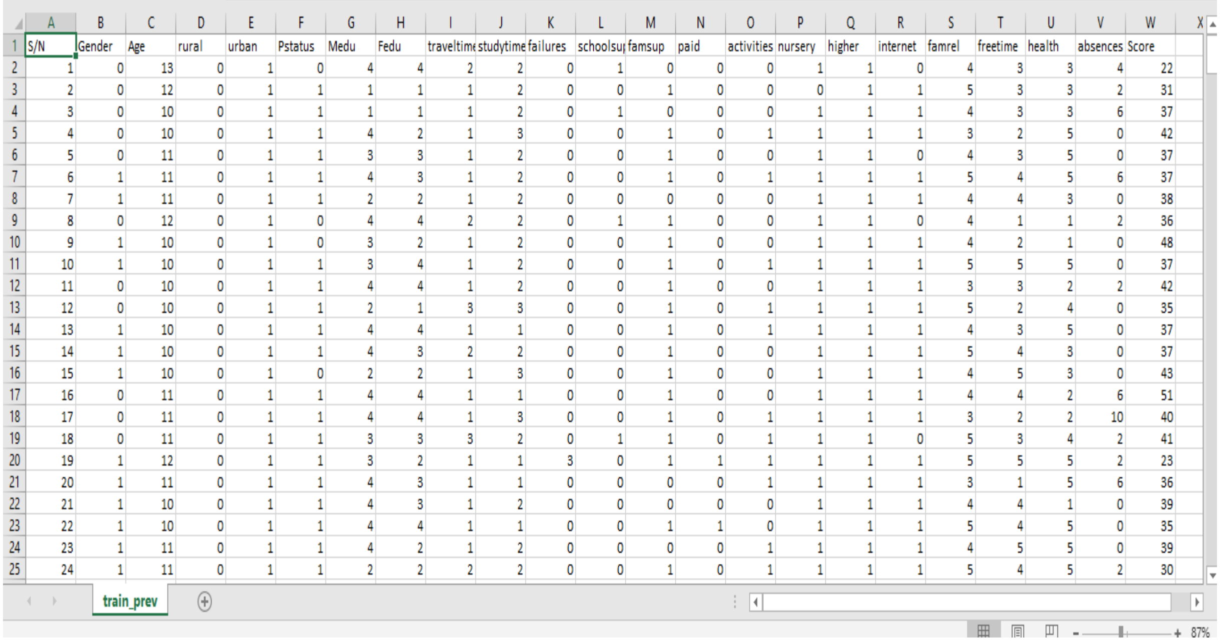
After the data analysis has been completed, the next step is the Data Transformation. Figures 7 and 8 show samples of the original train dataset and transformed train dataset respectively. Figure 7 shows the train dataset for this project indicating all parameters from the S/N to the SCORE. Models will be built with this dataset, then used to make predictions on the test dataset while Figure 8 is the transformed train data converting all object values to integers.

https://doi.org/10.25077/jitce.4.01.5-9.2020



S/N	Gender	Age	Location	famsize	Pstatus	Medu	Fedu	traveltim	studytime	failures	schoolsu	famsup	paid	activities	nursery	higher	internet	famrel	freetime	health	absences	Score
1	F	13	U	GT3	A	4	4	2	2	0	yes	no	no	no	yes	yes	no	4	3	3	4	22
2	F	12	U	GT3	T	1	1	1	2	0	no	yes	no	no	no	yes	yes	5	3	3	2	31
3	F	10	U	LE3	T	1	1	1	2	0	yes	no	no	no	yes	yes	yes	4	3	3	6	37
4	F	10	U	GT3	T	4	2	1	3	0	no	yes	no	yes	yes	yes	yes	3	2	5	0	42
5	F	11	U	GT3	T	3	3	1	2	0	no	yes	no	no	yes	yes	no	4	3	5	0	37
6	M	11	U	LE3	T	4	3	1	2	0	no	yes	no	yes	yes	yes	yes	5	4	5	6	37
7	M	11	U	LE3	T	2	2	1	2	0	no	no	no	no	yes	yes	yes	4	4	3	0	38
8	F	12	U	GT3	A	4	4	2	2	0	yes	yes	no	no	yes	yes	no	4	1	1	2	36
9	M	10	U	LE3	A	3	2	1	2	0	no	yes	no	no	yes	yes	yes	4	2	1	0	48
10	M	10	U	GT3	T	3	4	1	2	0	no	yes	no	yes	yes	yes	yes	5	5	5	0	37
11	F	10	U	GT3	T	4	4	1	2	0	no	yes	no	no	yes	yes	yes	3	3	2	2	42
12	F	10	U	GT3	T	2	1	3	3	0	no	yes	no	yes	yes	yes	yes	5	2	4	0	35
13	M	10	U	LE3	T	4	4	1	1	0	no	yes	no	yes	yes	yes	yes	4	3	5	0	37
14	M	10	U	GT3	T	4	3	2	2	0	no	yes	no	no	yes	yes	yes	5	4	3	0	37
15	M	10	U	GT3	A	2	2	1	3	0	no	yes	no	no	yes	yes	yes	4	5	3	0	43
16	F	11	U	GT3	T	4	4	1	1	0	no	yes	no	no	yes	yes	yes	4	4	2	6	51
17	F	11	U	GT3	T	4	4	1	3	0	no	yes	no	yes	yes	yes	yes	3	2	2	10	40
18	F	11	U	GT3	T	3	3	3	2	0	yes	yes	no	yes	yes	yes	no	5	3	4	2	41
19	M	12	U	GT3	T	3	2	1	1	3	no	yes	yes	yes	yes	yes	yes	5	5	5	2	23
20	M	11	U	LE3	T	4	3	1	1	0	no	no	no	yes	yes	yes	yes	3	1	5	6	36
21	M	10	U	GT3	T	4	3	1	2	0	no	no	no	no	yes	yes	yes	4	4	1	0	39
22	M	10	U	GT3	T	4	4	1	1	0	no	yes	yes	no	yes	yes	yes	5	4	5	0	35
23	M	11	U	LE3	T	4	2	1	2	0	no	no	no	yes	yes	yes	yes	4	5	5	0	39
24	M	11	U	LE3	T	2	2	2	2	0	no	yes	no	yes	yes	yes	yes	5	4	5	2	30

Figure 7. Sample Original Train Dataset



S/N	Gender	Age	rural	urban	Pstatus	Medu	Fedu	traveltim	studytime	failures	schoolsu	famsup	paid	activities	nursery	higher	internet	famrel	freetime	health	absences	Score
1	0	13	0	1	0	4	4	2	2	0	1	0	0	0	1	1	0	4	3	3	4	22
2	0	12	0	1	1	1	1	1	2	0	0	1	0	0	0	1	1	5	3	3	2	31
3	0	10	0	1	1	1	1	1	2	0	1	0	0	0	1	1	1	4	3	3	6	37
4	0	10	0	1	1	4	2	1	3	0	0	1	0	1	1	1	1	3	2	5	0	42
5	0	11	0	1	1	3	3	1	2	0	0	1	0	0	1	1	0	4	3	5	0	37
6	1	11	0	1	1	4	3	1	2	0	0	1	0	1	1	1	1	5	4	5	6	37
7	1	11	0	1	1	2	2	1	2	0	0	0	0	0	1	1	1	4	4	3	0	38
8	0	12	0	1	0	4	4	2	2	0	1	1	0	0	1	1	0	4	1	1	2	36
9	1	10	0	1	0	3	2	1	2	0	0	1	0	0	1	1	1	4	2	1	0	48
10	1	10	0	1	1	3	4	1	2	0	0	1	0	1	1	1	1	5	5	5	0	37
11	0	10	0	1	1	4	4	1	2	0	0	1	0	0	1	1	1	3	3	2	2	42
12	0	10	0	1	1	2	1	3	3	0	0	1	0	1	1	1	1	5	2	4	0	35
13	1	10	0	1	1	4	4	1	1	0	0	1	0	1	1	1	1	4	3	5	0	37
14	1	10	0	1	1	4	3	2	2	0	0	1	0	0	1	1	1	5	4	3	0	37
15	1	10	0	1	0	2	2	1	3	0	0	1	0	0	1	1	1	4	5	3	0	43
16	0	11	0	1	1	4	4	1	1	0	0	1	0	0	1	1	1	4	4	2	6	51
17	0	11	0	1	1	4	4	1	3	0	0	1	0	1	1	1	1	3	2	2	10	40
18	0	11	0	1	1	3	3	3	2	0	1	1	0	1	1	1	0	5	3	4	2	41
19	1	12	0	1	1	3	2	1	1	3	0	1	1	1	1	1	1	5	5	5	2	23
20	1	11	0	1	1	4	3	1	1	0	0	0	0	1	1	1	1	3	1	5	6	36
21	1	10	0	1	1	4	3	1	2	0	0	0	0	0	1	1	1	4	4	1	0	39
22	1	10	0	1	1	4	4	1	1	0	0	1	1	0	1	1	1	5	4	5	0	35
23	1	11	0	1	1	4	2	1	2	0	0	0	0	1	1	1	1	4	5	5	0	39
24	1	11	0	1	1	2	2	2	2	0	0	1	0	1	1	1	1	5	4	5	2	30

Figure 8. Sample Transformed Train Dataset

Regression

Regression analysis is one of the most important fields in statistics and machine learning. Regression problems usually have one continuous and unbounded dependent variable. Linear regression is one of the most important and widely used regression techniques because of the ease of interpreting results. A linear relationship exists between some dependent variable y on a set of independent variables $x = (x_1,\ldots,x_r)$, where r is the number of predictors as shown in equation 1:

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_r x_r + \varepsilon \qquad (1)$$

This equation is the regression equation. β_0, β_1,…, β_r are the regression coefficients, y is the predicted output, and ε is the random error.

$$\text{Student\_score} = \theta_0 + \theta_1 \text{ Study time} + \ldots + \text{class attendance } x_2 + \cdots \theta n\, xn \qquad (2)$$

- y = Predicted score
- θ = Data
- $x1$,x2 = dependent variables

The metric utilized for the measurement and determination of the performance of the models used is the Mean Absolute Error (MAE), which describes the average difference between two (2) variables over the test sample where all individual differences have equal weight as described mathematically in equation 3.

$$\text{MAE} = \frac{1}{n}\sum_{j=1}^{n} |y_i - x_i| \qquad (3)$$

RESULTS AND DISCUSSION

Model 1 - Linear regression for supervised learning with scikit-learn

All 21 attributes with the exception of the "Score" are used as inputs (independent variables) to the system while the output (dependent variable) was the "Score" to get the Beta coefficients of the training dataset. Thereafter, the beta coefficients was used alond with all the 21 independent variables to predict students scores using the test dataset.

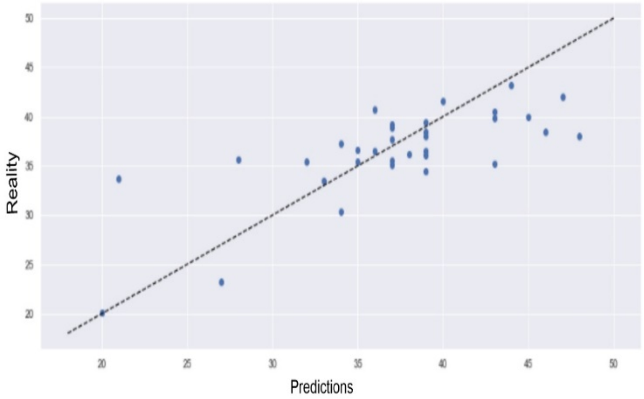


Figure 9. Regression performance output and linear regression graph for Model 1

Figure 9 displays the linear regression graph for the first model plotting the actual score against the predicted score with metrics: MAE = 3.26. The model performed averagely on the prediction with a mean absolute error of 3.26.

Model 2 - Linear Regression Model for Deep learning with TensorFlow and Keras

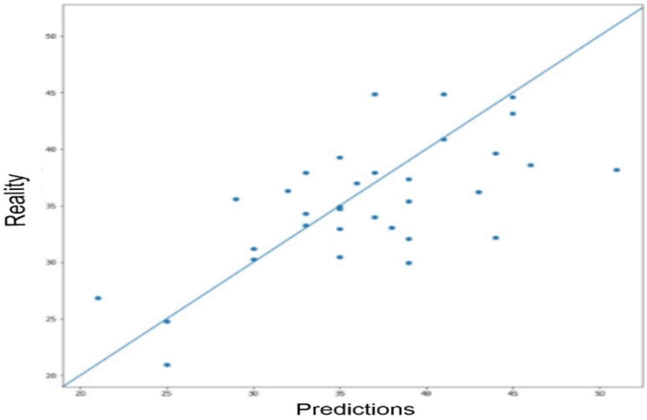


Figure 10. Regression performance output and linear regression graph for Model 2

Figure 10 displays the linear regression graph for the second model plotting the actual score against the predicted score with metric: MAE = 4.61. The model did not perform better than the first model because of the increase in MAE.

Model 3 - Neural Network Model with Deep Learning

To rescale our data we will use the function MinMaxScaler of Scikit-learn. With tf.contrib.learn it is very easy to implement a Deep Neural Network. 5 hidden layers with 200, 100, 50, 25 and 12 units respectively and the function of activation Relu with learning rate of 0.01 was implemented. The optimizer used is an AdaDelta optimizer.

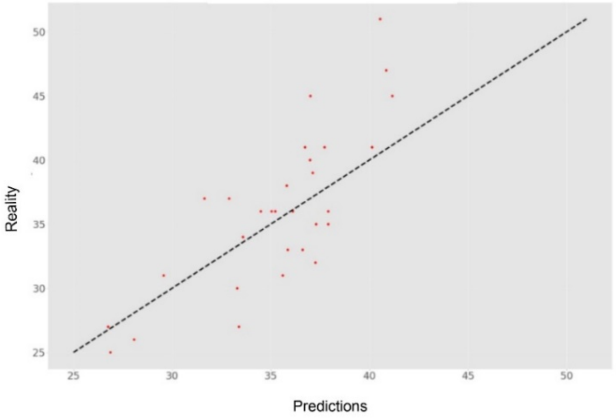


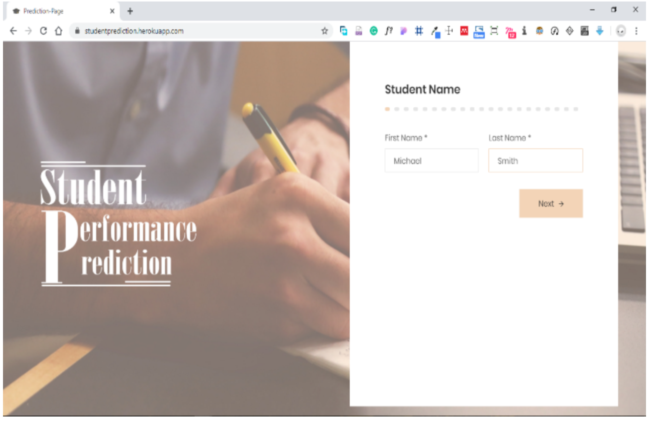
Figure 11. Regression performance output and linear regression graph Model 3

Figure 11 displays the linear regression graph for the third model plotting the actual score against the predicted score with metric: MAE = 6.23 and the linear regression graph for the first model plotting the actual score against the predicted score.

The model did not perform better than the first and second models because it has a higher MAE than the other models. *After* going through the three models and how they performed on the train data. The Linear regression model for supervised learning with Sci-kit Learn was the best of the three having the best MAE.

Platform Model Deployment

The Student Performance Prediction platform was designed using Heroku. The attributes are selected using the easy to use graphical user interface and prediction carried out. The graphical user interfaces are shown in the Figures 12 -14 below.



Student Name

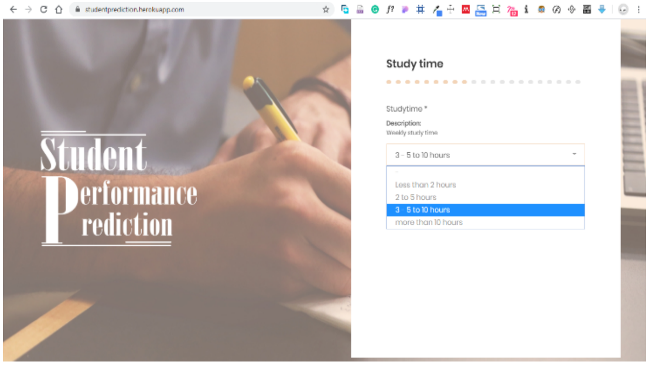
First Name *

Michael

Last Name *

Smith

Figure 12. Input Page for Model Testing



Study time

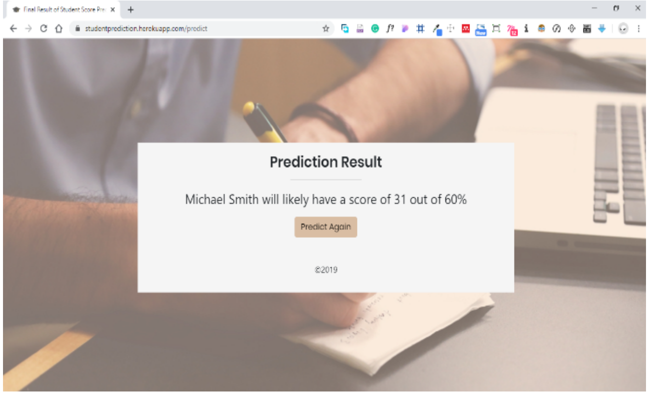
Studytime *

Description:
Weekly study time

3 - 5 to 10 hours

Less than 2 hours
2 to 5 hours
3 - 5 to 10 hours
more than 10 hours

Figure 13. Attribute Input Page for Model Testing


Final Result of Student Score Pre
studentprediction.herokuapp.com/predict
Predict Again

Prediction Result

Michael Smith will likely have a score of 31 out of 60%

©2019

Figure 14. Result Page for Model Testing

Figures 12 - 14 displays the pages designed to make a prediction for a student by inputting some of the factors and the result after inputting all the factors and making the prediction.

CONCLUSION

To achieve student performance prediction with a well-labeled dataset which makes it a regression problem and a supervised learning data, some of the best regression algorithms were used to solve the problem. Linear regression for supervised learning with Sci-kit learn gave us the most preferable model with a Mean Absolute Error of 3.26.

All these were achieved with our dataset of 325 entries; however, the result could be better if the dataset was much bigger, let say having thousands more entries or more. Machine learning algorithms try to find patterns in data that make them efficient in the use of huge datasets. This does not mean machine learning can't be effective on small datasets.

For further researches on student performance, I recommend the use of very large data, at least a 5,000 entries dataset. Huge data gives room for more Train data helping the algorithms to find patterns and make accurate predictions.

REFERENCES

[1] S. Huang and N. Fang, "Predicting student academic performance in an engineering dynamics course: A comparison of four types of predictive mathematical models," *Computers & Education,* vol. 61, pp. 133-145, 2013.

[2] L. Cohen, L. Manion and K. Morrison, Research Methods in Education, 6th ed., Routledge: Oxon, UK, 2007.

[3] W. B. Ware and J. P. Galassi, "Using Correlational and Prediction Data to Enhance Student Achievement in K-12 Schools: A Practical Application for School Counselors," *Professional School Counselling,* vol. 9, no. 5, pp. 344-356, 2006.

[4] S. K. Yadav and S. Pal, "Data Mining: A Prediction for Performance Improvement of Engineering Students using Classification," *World of Computer Science and Information Technology Journal,* vol. 2, no. 2, pp. 51-56, 2012.

[5] P. Cortez and A. Silva, "Using Data Mining to Predict Secondary School Student Performance".

[6] M. Pandey and V. K. Sharma, "A Decision Tree Algorithm Pertaining to the Student Performance Analysis and Prediction," *International Journal of Computer Applications,* vol. 61, no. 13, pp. 1-5, 2013.

[7] D. Kabakchieva, "Predicting Student Performance by Using Data Mining Methods for Classification," *Cybernetics and Information Technologies,* vol. 13, no. 1, pp. 61-72, 2013.

[8] B. A. Kalejaye, O. Folorunsho and O. L. Usman, "Predicting Students' Grade Scores using Training Functions of Artificial Neural Network," *Journal of Natural Science, Engineering and Technology,* vol. 14, no. 1, pp. 25-42, 2015.

https://doi.org/10.25077/jitce.4.01.5-9.2020