Online Student Performance Prediction Using Machine Learning Approach

Neena Somar
PG Scholar of Amal Jyothi College of Engineering, Kanjirappally, Kerala
neenasomar@mca.ajce.in

T J Jobin
Asst. Professor of Amal Jyothi College of Engineering, Kanjirappally, Kerala
tjjobin@amaljyothi.ac.in

Abstract:

The paper is ready to predict scholars' overall performance on online medium the use of gadget studying. The present-day scenario of covid brings about turning lessons into the online platform.

Right here the paper is to predict scholar overall performance and analyze the numerous other factors affecting overall performance gadget mastering helps to expect student online class performance. based on the prediction either properly or bad for college kids can be determined out. This helps the government authority to take essential motions to improve their performance. right here we make use of various gadgets gaining knowledge of algorithms like regression models and correlation approaches.

***keywords*: device studying, regression version, and correlation technique**

I. INTRODUCTION

Machine learning is the approach of learning or predicting an output based totally on input. The enter is previously accumulated datasets. this is a part of artificial intelligence(AI). synthetic Intelligence is the drawback of human intelligence in gadgets that are programmed to think and act like humans.

Machine learning getting to know is important for lots of commercial enterprise organizations to view the customer fashion. machine gaining knowledge of are labelled into unique shapes based on predicting accuracy. Supervised learning, unsupervised learning, semi-supervised learning, and reinforcement learning are the four broad categories.

A. *Multivariate Regression*

Multivariate Regression is a technique for determining the degree to which many independent variables (predictors) and correlation between different variables influence one another (responses), which are linearly related. The method is broadly used to predict the behavior of the response variables associated with changes in the predictor variables, once a desired degree of relation has been established.

In this evaluation of students, the range of predicted values is smaller than that of the actual values,indicating improvements in the regression model

B. *Requirements:*

- Jupyter
- data/.csv datasets
- Appropriate Algorithm

C. *Correlation*

Correlation is the indicator of changes between two variables. The correlation coefficient is to find the difference between the estimated value and the original value.

Correlation is the mathematical summary of the relationship between variables and how to calculate it for different types of variables and relationships.

D. *Decision Tree*

The decision tree is termed to be Classification and Regression Trees. The paintings with the aid of mastering answers to a hierarchy of if/else questions main to a choice, those questions shape a tree-like structure and as a result the name.

a) *Features*

- The feature importance is clear and relations can be viewed simply. Regression trees are represented identically, simply they predict nonstop values like the charge of a house.
- Commonly, Decision Tree algorithms are pointed out as CART or Classification and Regression Trees.
- It's miles a tree-structured classifier, where internal nodes constitute the capabilities of a dataset,

DOI: 10.5281/zenodo.6228486
ISBN: 978-93-5607-317-3

branches constitute the choice regulations and every leaf node represents the final results.

E. Random Forest Regression Method

The random forest regression method is a type of supervised learning and works based on a decision tree algorithm. Mostly used for solving regression and classification problems. This focused on ensemble learning, combining many classified to solve complex problems

a) *Features*

- More accurate than decision tree
- Algorithm can handle missing data perfectly
- Solves issues of overfitting.

II. LITERATURE REVIEW

[1] The paper discusses online student performance prediction. Due to the pandemic situation of covid, all the offline classes are turned into online classes. The paper figures out the problems faced by students, predicting their performance, how students adapted to the change, and how to improve the current scenario of student performance and their health.

[2] Student performance analyzing during the covid situation. We know the fact that online class affects the health of student since they lack physical work and too many students affect them mentally and lacks concentration. We proposed trendy prediction-based students improving their performance. After general prediction, we term to all the areas students got engaged during an online class.

III. METHODOLOGY

In this Prediction, we make use of the JUPYTER NOTEBOOK tool and integrate it with Python programming. Jupyter is open-source software used for machine learning, data science, and statistical analysis. This contains a collection of code for visualization and predicting algorithms. and data analyzing tool. This helps in the user interface for graphical representation and an easier method for prediction.

Listed below are some formats that Jupiter supports, from where data can be imported:

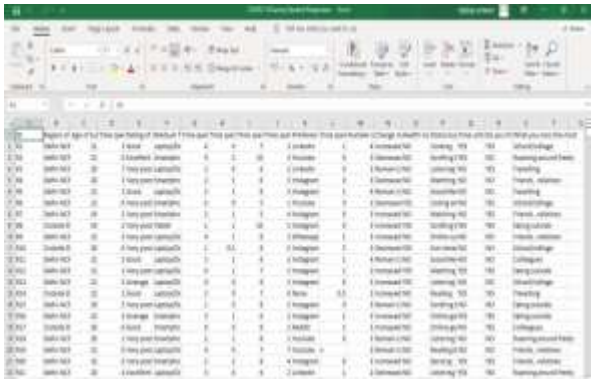
- CSV (Comma-Separated Values)

Jupyter supported a variety of programming languages, but we'll be using Python in this example. Jupyter facilitates data visualization, graphical representation, and algorithm implementation. Jupyter is a set of tools that can be used to:

- Association
- Data pre-processing
- Classification
- Visualization
- Regression
- Clustering

IV.DATASET

The implementation of prediction online student class performance has based on a set of datasets. The dataset is saved with an extension of .csv.The dataset contains different columns and corresponding variables. The dataset is from Kaggle.com.



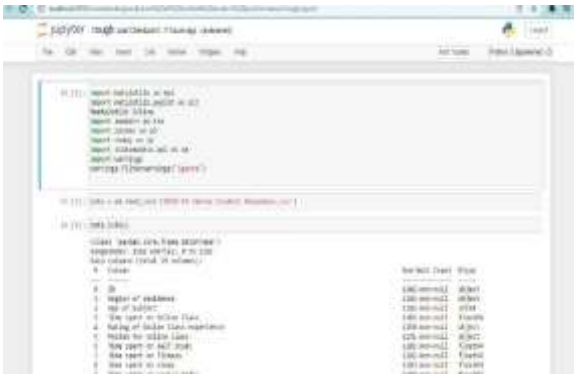
V. IMPLEMENTATION

First, open Jupyter Notebook and create a folder for the project. Take the dataset from Kaggle.com "COVID-19 observations Student Responses.csv" dataset. Load the dataset in the corresponding folder.

DOI: 10.5281/zenodo.6228486
ISBN: 978-93-5607-317-3 @2022 MCA, Amal Jyothi College of Engineering Kanjirappally, Kottayam

To load Jupyter ->on the command prompt of python folder type “Jupiter notebook”

Create a new document of python 3(ipykernal) and it has an extension of .ipyb.There we have the option to rename a new document.

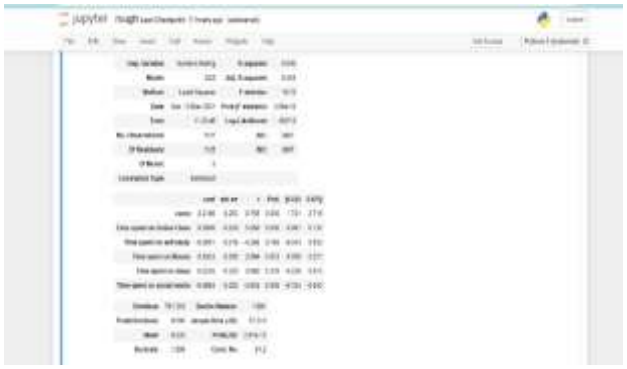


Implementing steps involved data cleansing, removing duplicates, removing null values, boxplot, correlation, creating regression model, and validating the prediction.

A. Multivariate Regression Model

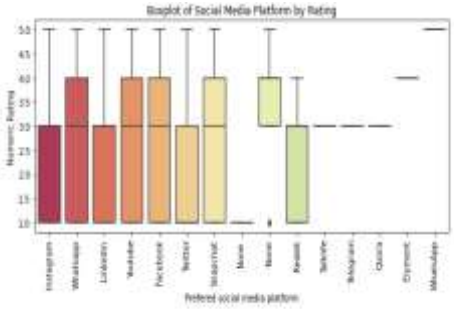
Multivariate Regression is a model used to measure the degree to which more than one independent variable and more than one dependent variable are linearly related.

Standard Errors guess that the covariance matrix of the errors is correctly named



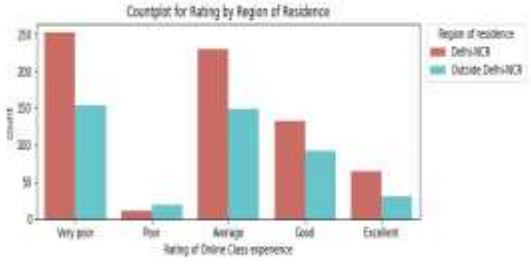
B. Boxplot for social media usage

- Instagram, WhatsApp, and YouTube are the top 3 selected social media platforms.
- WhatsApp and YouTube both have a median ranking of 3 ("Average") with a maximum rating of 4 ("Good"), while Instagram users rate their highest online class experience to be 3 ("Average").



C. Count plot for rating by region of reference

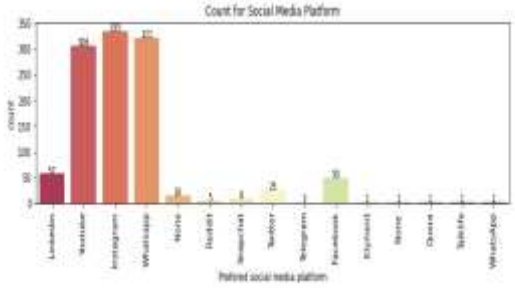
The count plot determines the graph of access of students to the online platform. The database is based on students in the region of Delhi and outside Delhi. The count plot determines the students of excellent way of counting than outside Delhi.



DOI: 10.5281/zenodo.6228486
ISBN: 978-93-5607-317-3 @2022 MCA, Amal Jyothi College of Engineering Kanjirappally, Kottayam

D. Count plot of students accessing social media during online class

Count plot graph determines WhatsApp, Instagram, YouTube are social media mainly students got engaged.



E. Heat Map on student spending time

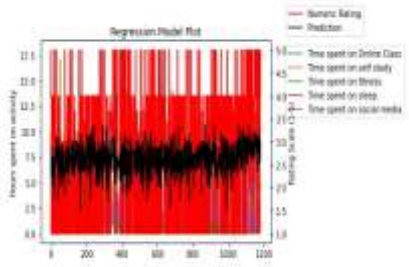
In this paper heat map determines the intensity of color variance based on students spending time on self-study, students spending time on sleep, online class, spent on fitness, spent on social media.

VI. RESULT

As we can see, the majority of predictions are put in the range between 1.5 to 3.5. It matches the count plot where the ratings "Very poor" and "Average" cover the largest proportion.

The range of predicted values is smaller than that of the actual values, indicating room for improvements in the regression model.

When applied to the regression model gives the result that the actual value varies from the predicted value, thus needing to get improved. The black mark determines the predicted value and red determine the numerical rating



VII. CONCLUSION

Due to the current scenario of covid-19, offline classes turned into online classes. This paper discusses predicting student performance on online platforms. Here we discuss the dataset oriented with the Delhi region. The dataset contains the data about the medium students used to have their online class, the time they spent in class, self-study, social media, engaging activity, health issues, meals they have taken. After undergoing machine learning concludes that students lack concentration, which affects their mental health, and more than that performance is weak corresponding to the offline class.

VIII. FUTURE WORK

When the dataset undergoes a regression model their carry variation in the predicted value and actual value. The future work is related to improving the accuracy of training and testing data.

IX. REFERENCES

[1] Nahid Khalil Elfaki-Department of Community and Mental Health

[2] Nursing, Nursing College, Najran University, KSA

[3] Ideal Abdulraheem-Department of Anatomy, Medicine College, Najran University, KSA.

[4] Lubna Salamat Lecturer, Department of Education The University of , Pakpattan Campus.

[5] Imran Latif Saifi Lecturer, Education The University of Campus

DOI: 10.5281/zenodo.6228486
ISBN: 978-93-5607-317-3 @2022 MCA, Amal Jyothi College of Engineering Kanjirappally, Kottayam