4/21/2025

BrightPath Machine learning report

Matt Takawira 600791

Charl Smuts 600270

Ruan Le Raux 600194

Vunene Khoza 600676

Table of Contents

[Github Link 2](#_Toc196239464)

[Problem statement 2](#_Toc196239465)

[Hypothesis Generation 2](#_Toc196239466)

[Demographics & Academic Performance: 2](#_Toc196239467)

[Study Habits & GPA: 2](#_Toc196239468)

[Parental Support & Academic Success: 2](#_Toc196239469)

[Extracurricular Activities & Holistic Development: 2](#_Toc196239470)

[Early Identification of At-Risk Students: 2](#_Toc196239471)

[Getting the system ready and reading the data 3](#_Toc196239472)

[Understanding the data 4](#_Toc196239473)

[Exploratory Data Analysis 5](#_Toc196239474)

[Missing value and outlier treatment 6](#_Toc196239475)

[Evaluation Metrics for classification problem 6](#_Toc196239476)

[Feature engineering 7](#_Toc196239477)

[Model Building: Part 1 7](#_Toc196239478)

[Model building: part 2 10](#_Toc196239479)

[Model deployment – 11](#_Toc196239480)

[Conclusion 11](#_Toc196239481)

## Github Link

<https://github.com/Ronin326/MLG-382-Project>

## Problem statement

BrightPath academy, while committed to empowering its students, faces issues including inability to identify at-risk students, the lack of strategies tailored for individual students, and difficulties relating extracurricular participation to academic achievement. To address these challenges data analytics will be used with the aim of identifying students who are at risk based on the data analysed.

## Hypothesis Generation

### Demographics & Academic Performance:

* The level of parental education positively impacts students' GPA, as they are likely to receive academic guidance from their parents.
* On average females tend to have higher GPAs than males.

### Study Habits & GPA:

* An increase in weekly study time will lead to higher GPAs.
* Higher rates of absence will result in lower GPAs due to missed coursework.
* Students with tutoring are likely to have an increased GPA from having assistance with areas that they struggle with.

### Parental Support & Academic Success:

* Increased parental support will have a positive effect on GPA.

### Extracurricular Activities & Holistic Development:

* Increased engagement in extracurricular activities contributes to a higher GPA as more involved students tend to put more effort into studies.

### Early Identification of At-Risk Students:

* Students with low parental education, minimal weekly study time, high absence rates, no tutoring, limited parental support, and low extracurricular engagement are more likely to be identified as at-risk.

## Getting the system ready and reading the data

* All necessary modules are imported into the notebook and initialized if required.

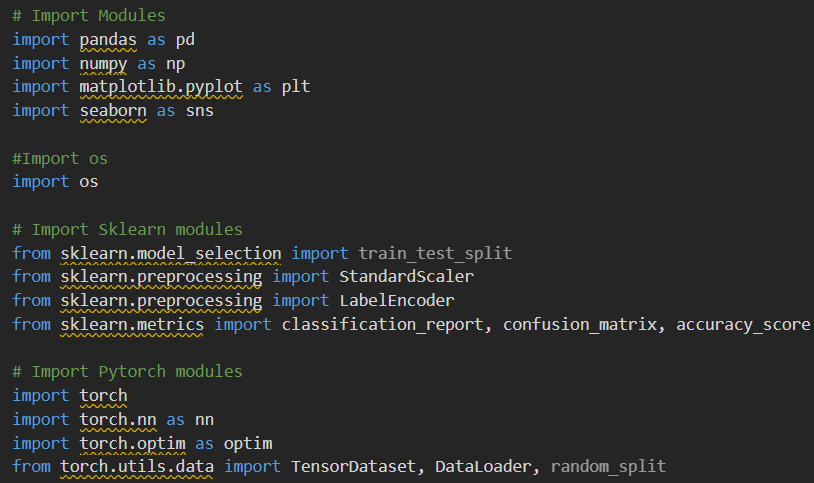


Figure 1 importing modules

* The dataset is loaded from the directory and stored on a data file for later use.

A computer screen shot of text

AI-generated content may be incorrect.

Figure 2 Loading dataset

## Understanding the data

* Here the basic information is viewed in order to gain more initial information.

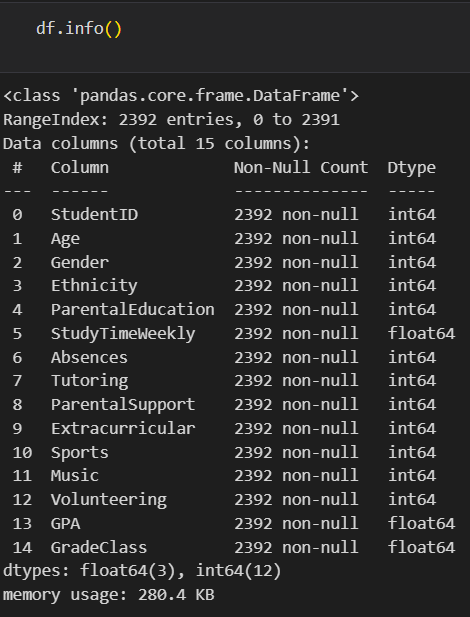


Figure 3 Viewing data information

## Exploratory Data Analysis

* Visualise the data to identify trends and correlations.

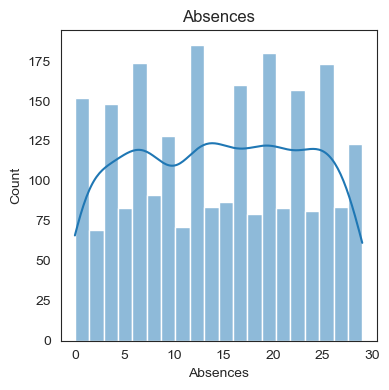


Figure 4 Graph of absences

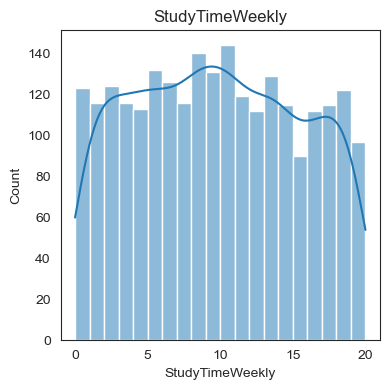


Figure 5 Graph of weekly study time

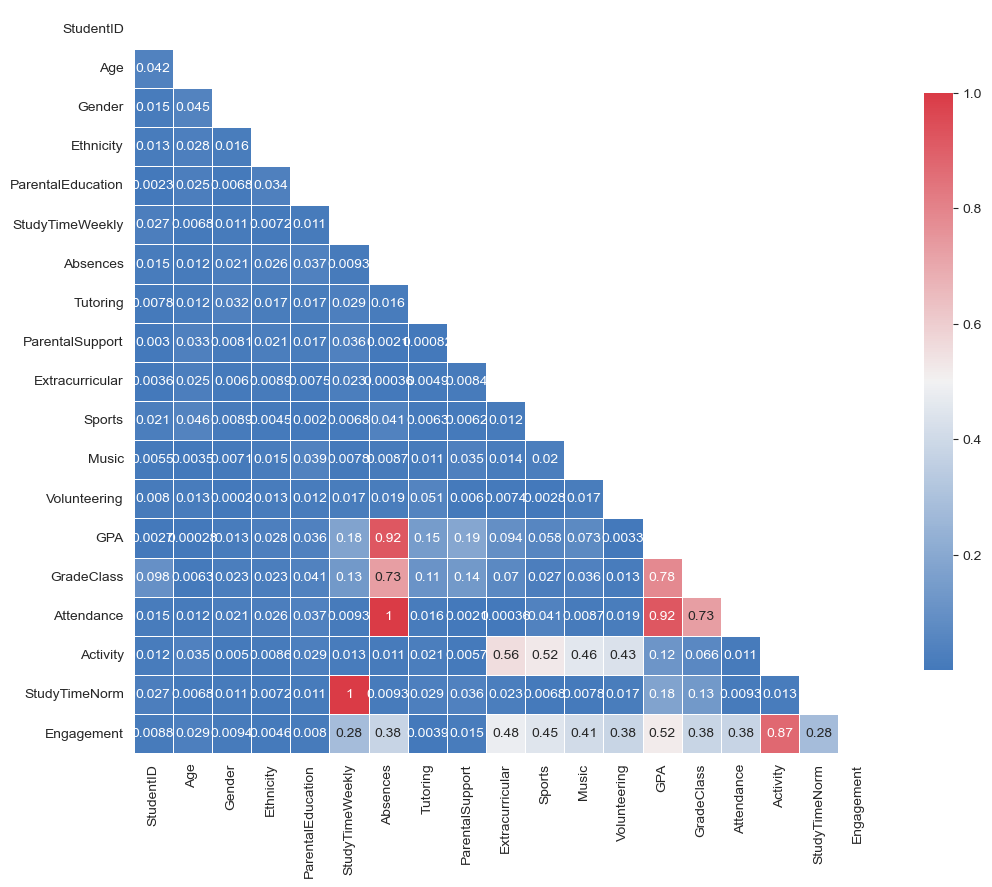


Figure 6 Correlation heatmap Graph of Input Variables

## Missing value and outlier treatment

* Treating the outliers to improve accuracy

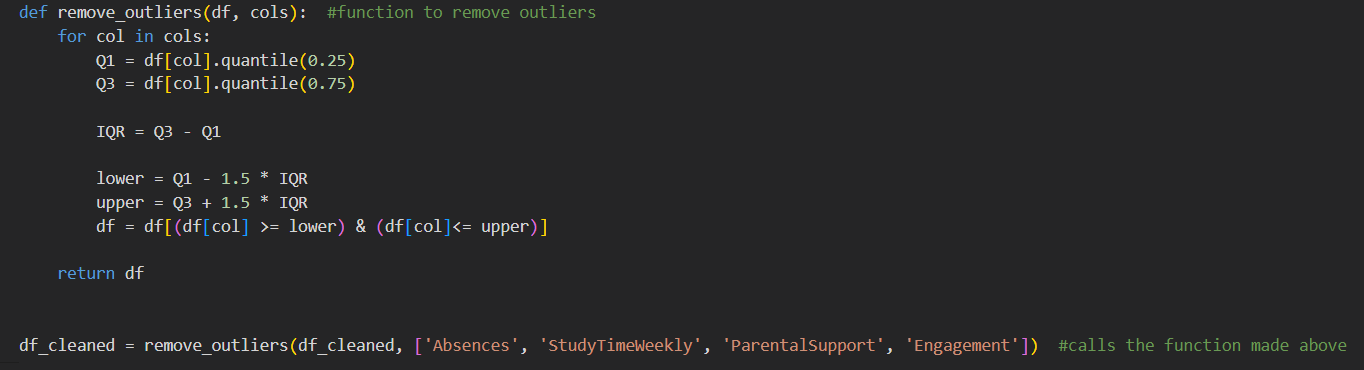


Figure 7 Outlier correction

## Evaluation Metrics for classification problem

* Confusion matrix used to check accuracy of deep learning model

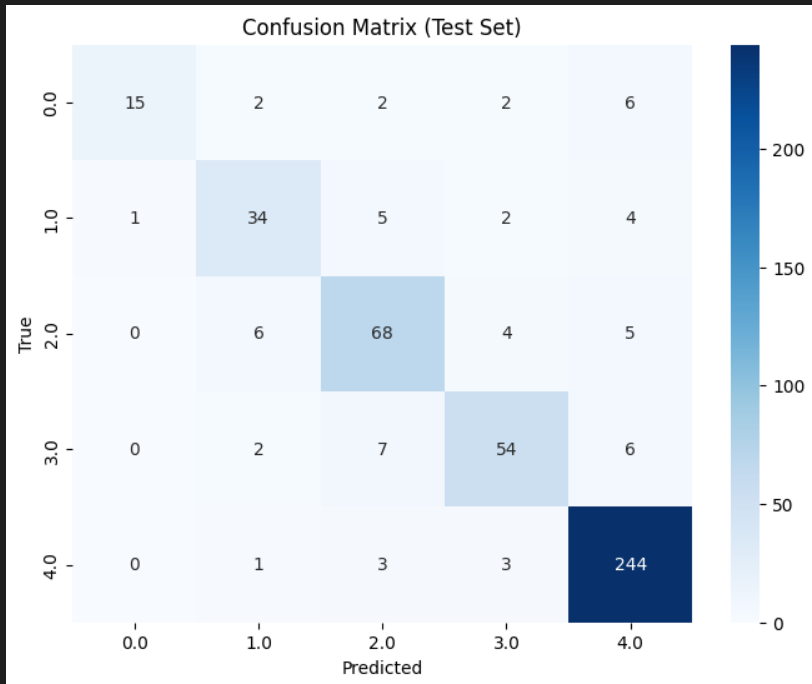


Figure 8 Confusing matrix

## Feature engineering

* Scaling of attendance, activities, study time and aggregating them into engagement for later use.

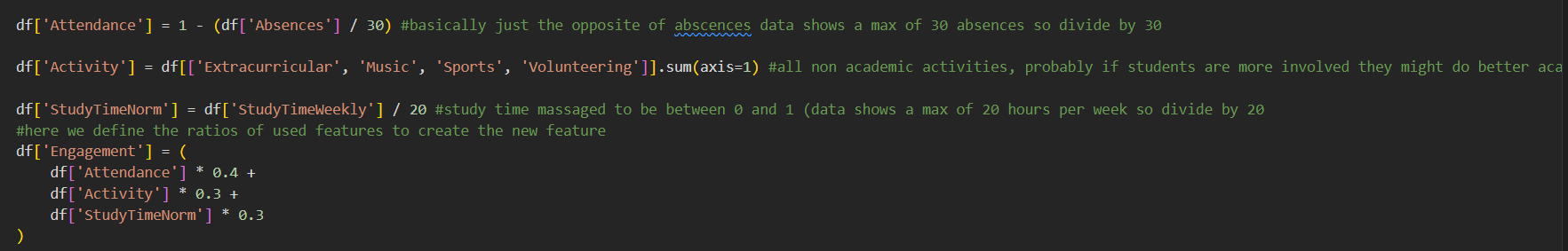


Figure 9 scaling of variables

## Model Building: Part 1

* The data is converted into Series data and split into training and test data. The split data is used to train an evaluate all the models.

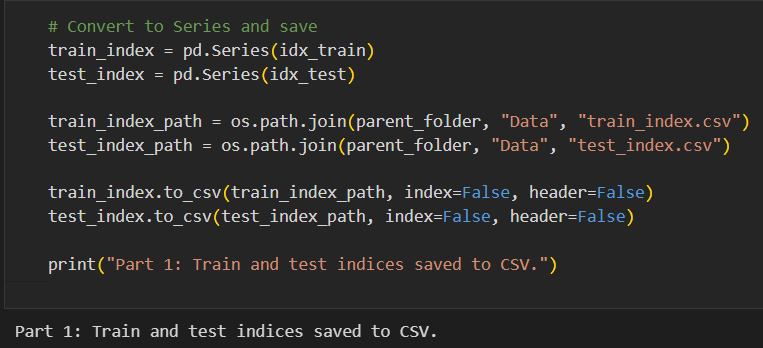


Figure 10 Saving trained data

* Logistic regression model is trained and evaluated showing 91% accuracy

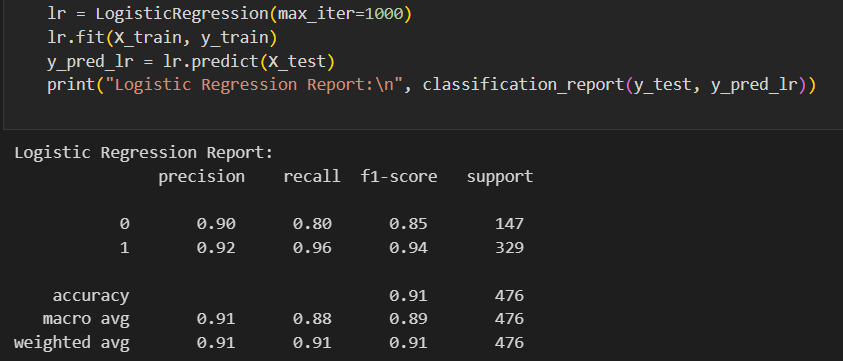


Figure 11 Using logistic regression

* Random forest is trained and evaluated showing 91% accuracy

A screen shot of a computer

AI-generated content may be incorrect.

Figure 12 Using random forest

A graph with blue squares

AI-generated content may be incorrect.

Figure 13 Graph of Random Forest features

## Model building: part 2

* The deep learning model undergoes the same preparation steps as the other two models, and after being trained it evaluated to only 87% accuracy. Tis is possible due to the simplicity of the dataset limiting its capabilities. Due to this the other two models were considers for the web application.

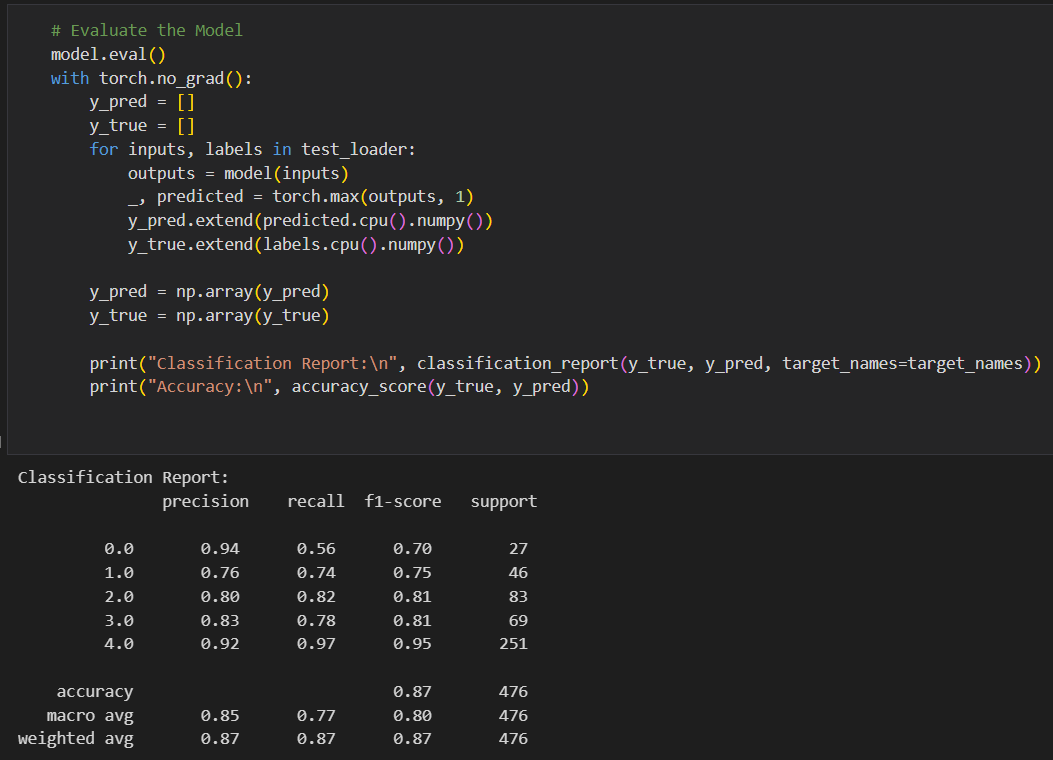


Figure 14 Evaluation using Deep learning Model

## Model deployment –

* The model has been deployed using Render and uses input boxes to receiver student information before processing and returning a prediction based on the trained model.

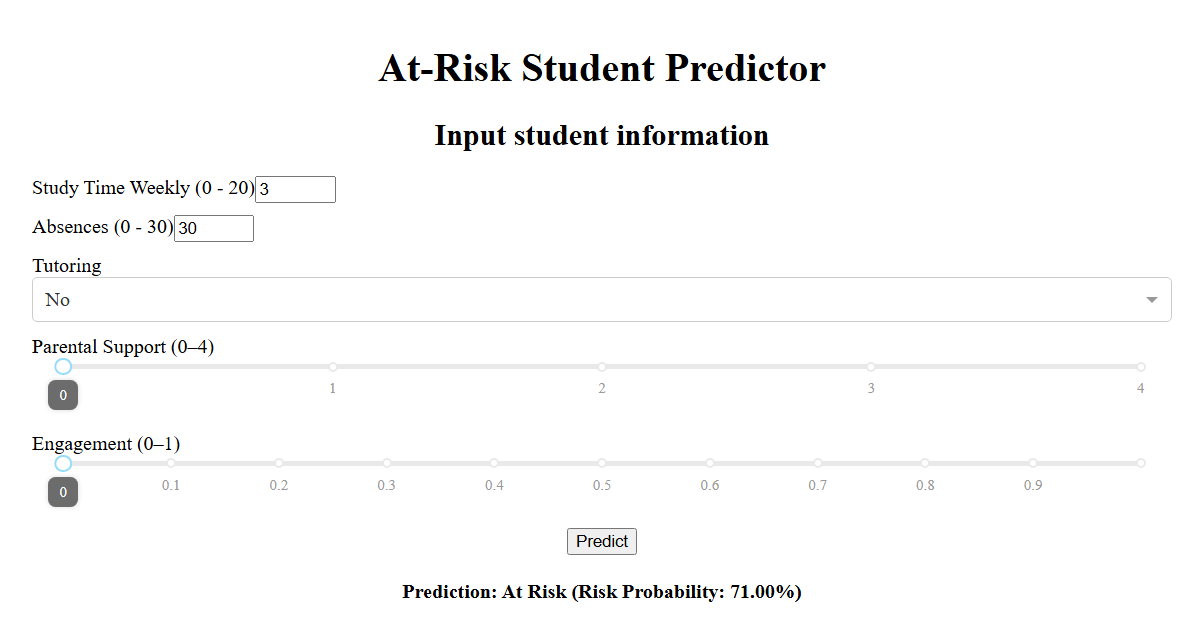


Figure 15 Web app interface

## Conclusion

BrightPath Academy's data analysis confirmed the hypotheses previously stated. This is supported by findings illustrated in Figure 4 (Graph of Absences), Figure 5 (Graph of Weekly Study Time), and Figure 6 (Correlation Heatmap of Input Variables). To accurately assess students at risk, various machine learning models, including Random Forest, Logistic Regression, and Deep Learning, were explored. Among these, Random Forest and Logistic Regression demonstrated the highest accuracy and Random Forest was selected as the primary model for the web application. With this deployed application, BrightPath Academy can input a student's basic information and receive an accurate prediction of whether the student is at risk.