Big Data Management Systems & Tools

Group Project

An Analysis for Students' Performance in School

Institution: University of Waterloo/University of Toronto

Program: Data Science Certificate Program

Course: Big Data management Systems and Tools

Group Number: 3

Group members:
Diana Kao
David Blachut
Estelle Chettiar
Frank Miceli
G M Ashikur Rahman

Date: August 2024

1. Introduction

1.1 Background

Understanding the factors that impact academic performance, such as socioeconomic status, parental involvement, ethnicity, gender, study habits, attendance, and extracurricular activities, is a useful way to enhance ways to improve educational outcomes. The Students Performance Dataset contains data on 2,392 high school students, covering demographics, study habits, parental involvement, extracurricular activities, and academic performance. It includes a dependent features, GPA and GradeClass. GradeClass was classified into five categories based on GPA. This dataset contains useful details for predictive modeling, incorporating factors such as study time, absences, parental support, and extracurricular involvement.

This report explores these factors by analyzing student performance data and applying machine learning (ML) techniques to predict academic success. By using PySpark in Databricks and Google Colab Notebook, we performed data preparation, descriptive statistical analysis, correlation and multicollinearity assessments, feature engineering, and predictive model building. Our objectives included forecasting GPA and GradeClass, and evaluating the accuracy of different ML models. We built and tested predictive models using linear regression, random forest, Naïve Bayes and KNN and carried out unsupervised learning and Principal Component Analysis. Through this analysis we compared the effectiveness of these models in predicting academic performance and identified the most accurate approach.

2. Data Review

2.1 Data Description

The Students Performance Dataset (link) contains data from 2,392 high school students covering demographics, study habits, parental involvement, gender, ethnicity, among other factors. This dataset has two dependent features, GPA and GradeClass.

2.2 Initial Data Exploration and Data Quality Assessment

To explore correlations between features, a Pearson correlation and heat map were created. This map showed that features “StudentID”, “Age”, “Gender”, “Ethnicity”, “ParentalEducation”, “StudyTimeWeekly”, “Abscences”, “Tutoring”, “ParentalSupport”, “Extracurricular”, “Sports”, “Music”, “Volunteering”, “GPA”, and “GradeClass” were correlated with each other. All of the above features were kept in the analysis.

The original dataset was already cleaned, thus no significant training was required. We noted no missing or null values and as a result did not need to manipulate the dataset to normalize for missing values.

3. Methodology and Process

3.1 Data Preparation

Dataset was imported and transformed into a data frame. Following that, an “index” feature was added to the data frame so that the first row, which contained features’ names, could be removed. After that, “index” feature was removed, as it wouldn’t be included in the analysis. All variables were converted from type string to type double. Features “StudyTimeWeekly” and “GPA” were rounded up to 2 decimal places. “StudyTimeWeekly” and “GPA” were dropped after they were rounded up and renamed to “StudyTimeWeekly2” and “GPA2”.

3.2 Feature Engineering

For the Linear Regression model, correlation matrix was used to investigate which independent features were correlated with other features. GPA and Absences had the highest

correlation.Features “StudentID”, “GPA2”, and “GradeClass” were labeled as non-feature columns; the rest of features were designated as feature columns.

3.3 Model Training and Testing

Splitting data into training and testing datasets

The students’ performance dataset was randomly split into training and tests datasets with 70% and 30% in training and test datasets, respectively.

3.4 Predictive Modeling

3.4.1 Classification ML Algorithms:

Four machine learning algorithms were chosen and applied for predictive model building to predict students’ academic performance in school.

- **Linear regression:** Linear regression was applied to predict students’ grade point averages (GPAs), with mean squared error (MSE), mean absolute error (MAE), root mean squared error (RMSE), and adjusted R squared calculated to compare the accuracy of predicted GPAs. To simulate MLOps, the model created was loaded into a simple Flask App providing REST API endpoint to calculate a prediction of GPA, This interface would inform students and educators of a possible impact of absent days to the student’s grade.
- **Random Forest**: Random forest classifier was used to predict students’ grade classes. Confusion matrix was calculated to compare the predicted and the actual grade classes.
- **Naïve Bayes**: Naïve Bayes classifier was also applied to predict the grade classes of students, with accuracy on the test calculated.
- **Gradient-boosted tree regression:** Gradient-boosted tree regression was used to predict students’ GPAs, with root mean squared error (RMSE) and mean absolute error (MAE) calculated to evaluate the predictive model derived from gradient-boosted tree regression. Hyperparameter tuning and model section were conducted for gradient-boosted tree (GBT) regression to weak model performance for optimal results.

All analyses were conducted by using PySpark in Databricks Communication Edition and Google Colab Notebook.

3.5 Unsupervised Algorithms

We also applied different unsupervised algorithms for our analysis in this project.

3.5.1 Principal Component Analysis (PCA):

To get simplified visualization and reduce complexity of the dataset, we applied PCA to convert the dataset into two principal components. PCA is a dimensionality reduction technique that transforms the original variables into a new set of uncorrelated variables (known as principal components), capturing the maximum variance in the data with the fewest components.

3.5.2. K-Means Clustering

We used K-Means clustering with various feature columns and 2D data obtained from PCA. K-Means is an iterative algorithm that partitions the dataset into k distinct, non-overlapping clusters. Each data point is assigned to the cluster with the nearest centroid, and the centroids are updated iteratively until desired convergence.

The Elbow Method was used to determine the optimal number of clusters (k) for K-Means clustering algorithm where we ran K-Means for different values of k and plotting the Within-Cluster Sum of Squares (WCSS) against the number of clusters. The point where the reduction in WCSS begins to slow down significantly (the "elbow" point) suggests the optimal number of clusters.

4. Results

4.1 Descriptive Statistics

Figure 1 shows distributions of independent and dependent features. Independent features are defined as age, gender, ethnicity, parental education, study time weekly, absences, tutoring, parental support, extracurricular, sports, music, and volunteering. GPA and grade class were the dependent features.

The students were aged between 15 and 18 years old. There were more male students than female students. The weekly number of study hours ranged between 0 and 20 hours. The

students had as many as close to 30 absences. Notably, a majority of students had a GPA between 1 and 3, and a grade class of 4.

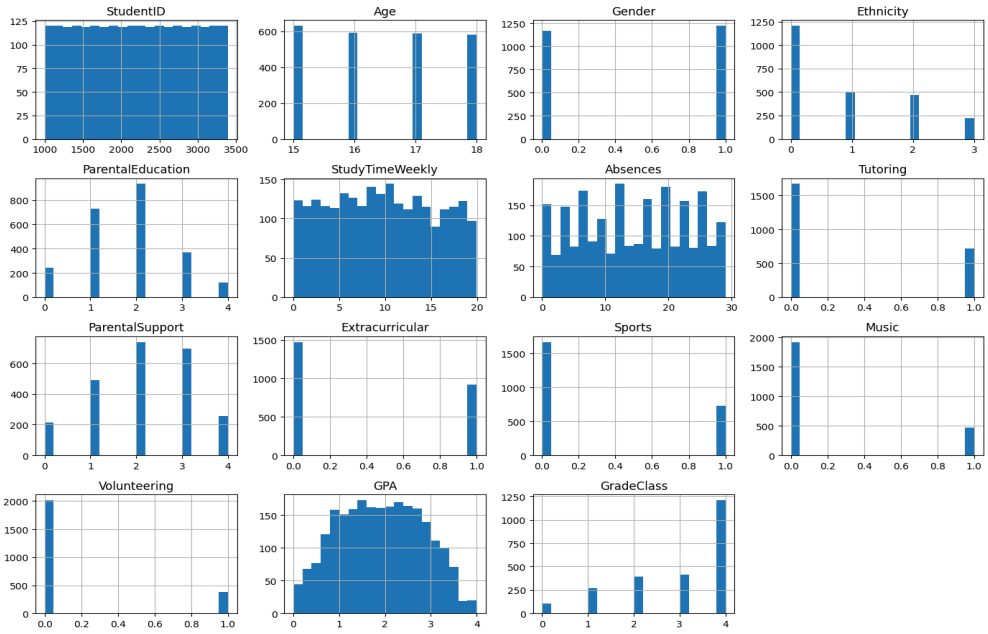

Figure 1. Distributions of features included in predictive modeling.

As the number of absences increased, GPA decreased.

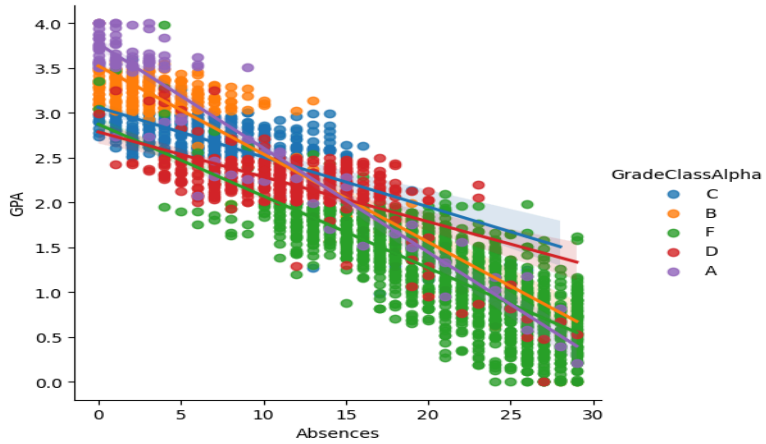

Figure 2. GPA by number of absences. Grouped by Grade Class.

As weekly study hours increased, GPA increased moderately.

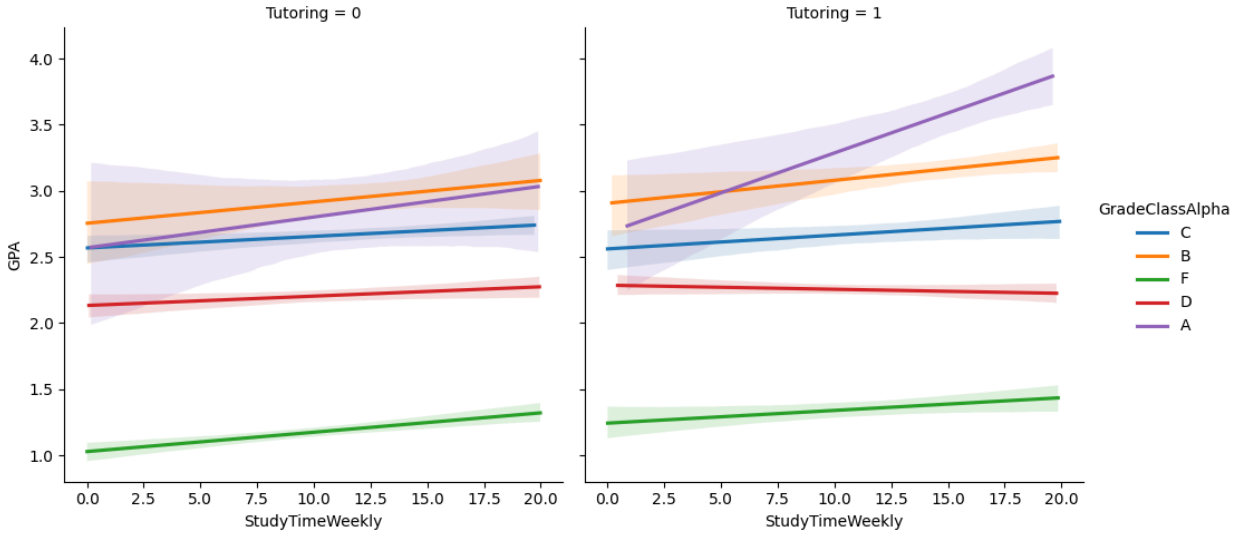

Figure 3. GPA by number of weekly study hours grouped by tutoring status

A correlation matrix was created to quickly illustrate relationships, In this data set there was a strong correlation between Absent days and GPA.

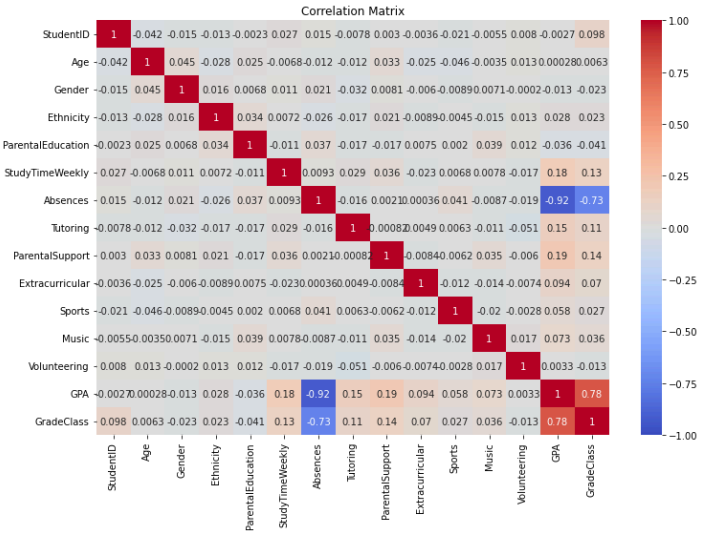

Figure 4. A correlation matrix of features, with -1 and 1 indicating a strong correlation.

4.2 Model Performance

- **Linear Regression**:

  The R squared value shows that 95.1% of data can be explained by the linear regression model.

Metrics	Value (rounded up to 3 decimal places)
MSE	0.0425
MAE	0.165
RMSE	0.206
R squared	0.951
Explained variance	0.828

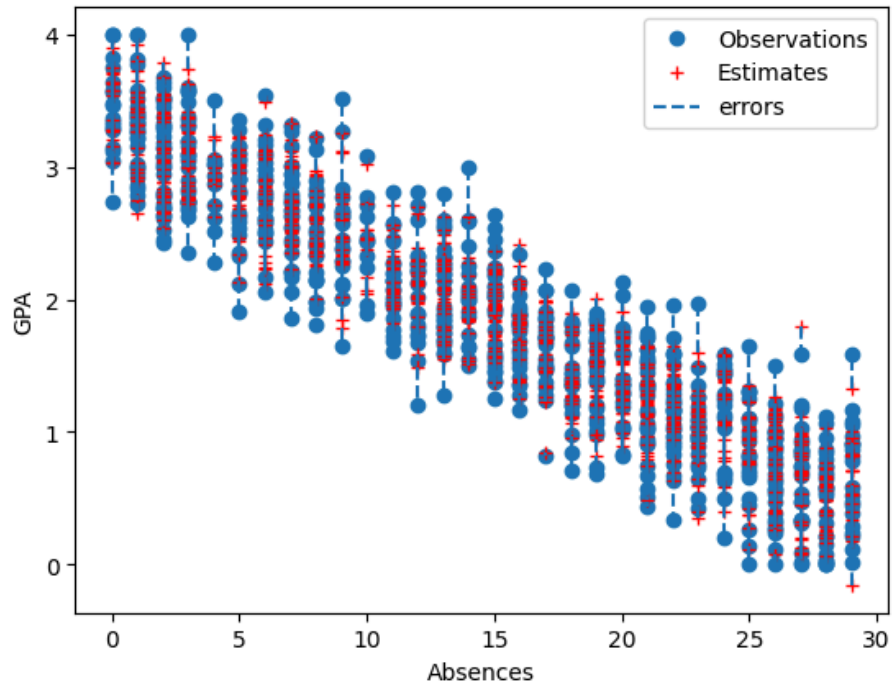

Figure 5.. Estimates vs actuals for the linear regression

- **Random Forest**:

Table 1 shows the accuracy and test error for students' grade classes predicted by random forest classifier method. The accuracy was 0.975, and test error was 0.025.

Metrics	Value (rounded up to 3 decimal places)
Accuracy	0.975
Test error	0.025

Table 1. Accuracy and test error for predicted students' grade classes by random forest classifier method.

Table 2 shows confusion matrix for the grade classes predicted by random forest classifier method. A total of 663 of 690 grade classes were correctly predicted by random forest model, while 25 were incorrectly predicted.

380	0	0	0	0
0	128	0	0	0
0	0	109	0	0
0	0	3	75	0
0	0	1	13	20

Table 2. Confusion matrix for comparing the grade classes predicted by random forest classifier model to the actual grade classes.

- **Naïve Bayes**:

  Table 3 shows the accuracy of grade classes predicted by Naïve Bayes classifier method. The accuracy was 0.794, indicating 79.4% of grade classes were correctly predicted.

Metrics	Value (rounded up to 3 decimal places)
Accuracy	0.794

Table 3. Accuracy of grade classes predicted by Naïve Bayes classifier method.

Figure 5 demonstrates a second Naive Bayes model built to predict GPA values. The confusion matrix shows a total of 548 out of 722 GPAs were correctly predicted.

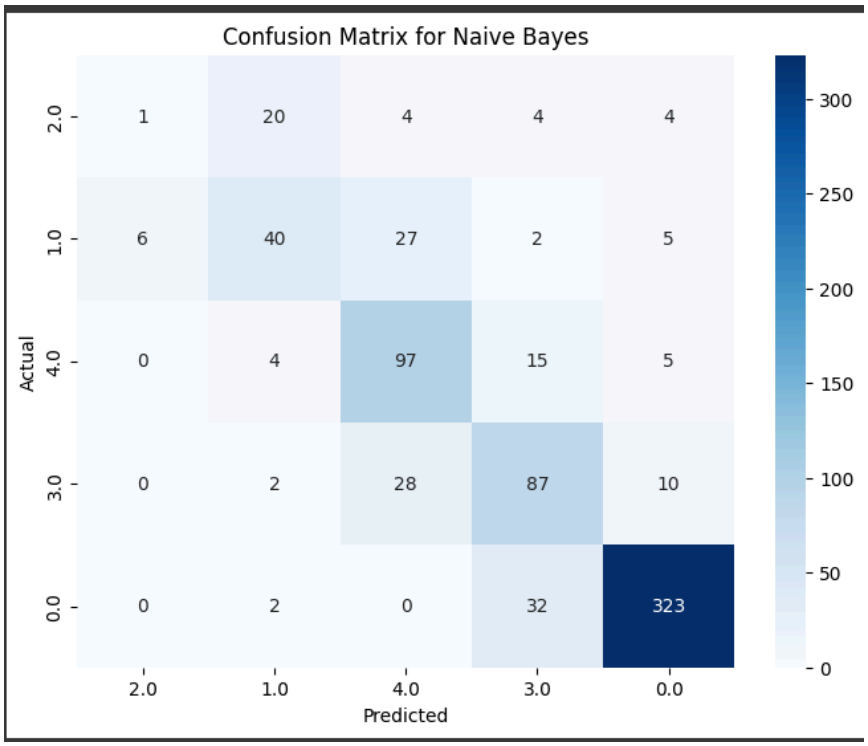

Figure 6. A Naive Bayes confusion matrix listing actual values against predicted GPA values.

- **Gradient-boosted Tree Regression:**

Table 4 shows RMSE and MAE of GPAs predicted by GBT regression. A low RMSE

(0.05) indicates that the GBT regression model's accuracy was high.

Metrics	Value
RMSE	0.05
MAE	0.03

Table 4. Root mean squared error and mean absolute error for comparing the accuracy of students' GPAs predicted by GBT regression.

Table 5 shows the RMSE derived by cross validation in hyperparameter tuning and model selection was 0.05. The relatively low RMSE indicates GBT hyperparameter tuning and model selection provided a high accuracy of GPA prediction.

Metrics	Value
Cross validator RMSE	0.05

Table 5. RMSE for cross validation of GBT hyperparameter tuning and model selection.

- **K-Means Clustering:**

  After performing the elbow test it was determined that k=4 is the optimal number of clusters. The elbow test figure is shown below.

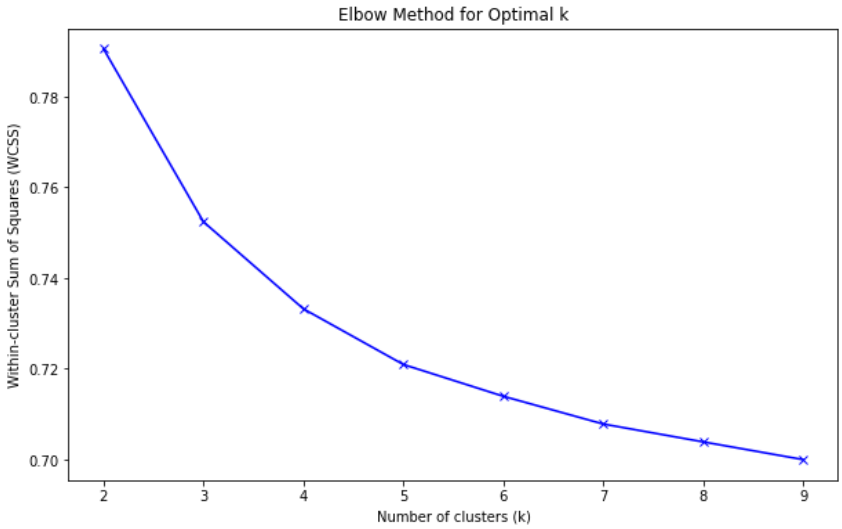

Figure 7. WCSS versus K-Means K-Means clustering was applied to the reduced dimensional data using PCA algorithm. Similar result was observed (Figure 2).

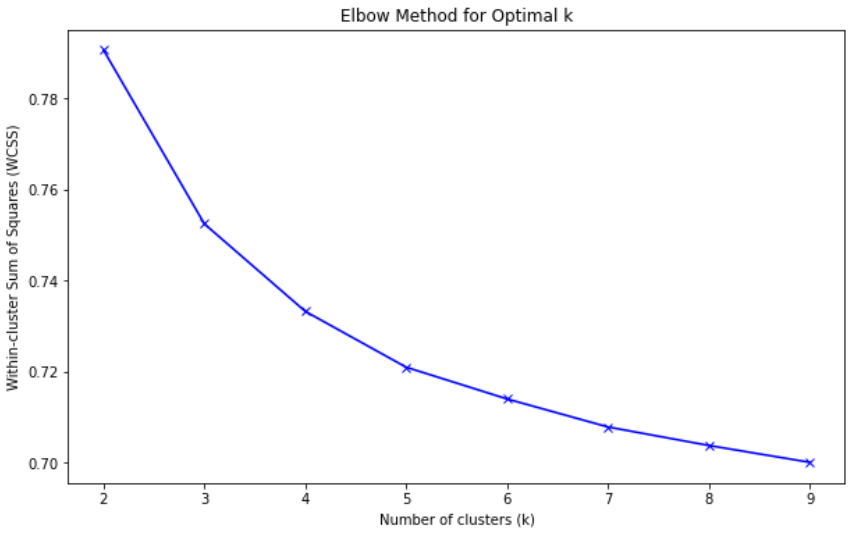

Figure 8. WCSS versus K-Means (Reduced dimensional data) The resulting clusters were visualized in a 2D scatter plot where different colors are representing four different clusters.

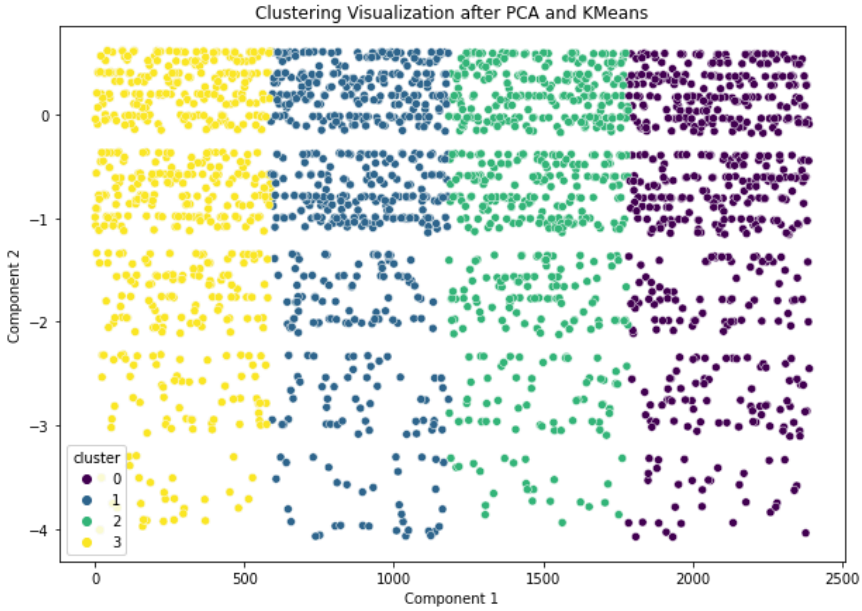

Figure 9. The visualization of clusters in the figure above created groupings of students where each cluster represents a distinct subgroup within the dataset, based on different combinations of study habits, absences, and parental support etc..

5. Conclusions

5.1 Summary of Findings

In the predictions for students' GPAs, Linear Regression's and GBT regression's accuracies were high. In predicting students' grade classes, random forest classifier had a higher accuracy than Naïve Bayes classifier. The results achieved using unsupervised algorithm can be helpful to get a better understanding of student's behaviors and to inform interventions aimed at improving academic outcomes. Future work could involve exploring other clustering algorithms or adding more features to enhance the analysis.

5.2 MLOps Simulation

The REST API was created to provide a simple read interface to the model. The linear regression model was embedded into a Flask application in a jupyter notebook via spark. This interface could be leveraged by other systems to inform students and educators of the possible impact of Absenteeism on the student's GPA.

The application will load the saved model created earlier. In a production setting containers would be used to deploy updated models and their respective applications.

```
Request: POST /predict

Request Body: {"Age": 15, "StudyTimeWeekly": 10.1, "Absences": 6, "ParentalEducationLevel": 2,
"Support": 3, "Activities": 1}

Response Headers: {'Server': 'Werkzeug/3.0.3 Python/3.10.12', 'Date': 'Fri, 09 Aug 2024
23:38:50 GMT', 'Content-Type': 'text/html; charset=utf-8', 'Content-Length': '34',
'Connection': 'close'}

Response body: '{"prediction": 2.8557700037779234}'
```

```
Request: POST /predict

Request Body: {"Age": 15, "StudyTimeWeekly": 10.1, "Absences": 15, "ParentalEducationLevel":
2, "Support": 3, "Activities": 1}

Response Headers: {'Server': 'Werkzeug/3.0.3 Python/3.10.12', 'Date': 'Fri, 09 Aug 2024
23:38:57 GMT', 'Content-Type': 'text/html; charset=utf-8', 'Content-Length': '34',
'Connection': 'close'}
```

```
Response Body: '{"prediction": 1.9544748902267568}'
```

6. References

1. Kaggle Students Performance Dataset
   https://www.kaggle.com/datasets/rabieelkharoua/students-performance-dataset

2. Ashikur Rahman's Databricks Link:
   https://community.cloud.databricks.com/?o=1508783109418 8#notebook/4094871093664272/command/4094871093664287

3. David Blachut's Databricks link:
   https://databricks-prod-cloudfront.cloud.databricks.com/public/4027ec902e239c93eaaa8714f173bcfc/4065297464460781/1537349993340573/4664763525987471/latest.html

4. Diana Kao's Databricks link:
   https://databricks-prod-cloudfront.cloud.databricks.com/public/4027ec902e239c93eaaa8714f173bcfc/4181377907373995/972147995344617/4533145134735469/latest.html

5. Estelle Chettiar's Google Colab:
   https://colab.research.google.com/drive/10KaQB2H4ZPi6zvwgRsanZLhJy74PHqaM?usp=sharing

6. Frank Miceli's Google Colab:
   https://colab.research.google.com/drive/1JU2yiYtqiDaZ5tWh7mR7zBDafNVLDlAm

7. Frank Miceli's sample App for linear regression:
   https://colab.research.google.com/drive/1jeFY2_cR8phaUcLqSnF36OczlaaMZ1Hs