COS6008 Introduction to Data Science

Assignment 2, 2023, Semester 1

Student Performance Analysis and Prediction

Student Details:

- Name: Thi Thanh Thuy Tran
- Student ID: 103514782
- Email: 103514782@student.swin.edu.au

- Submission Date: Sunday 28 May 2023
- TuteLab Class: Thursday 17:30

Abstract:

This report focuses on the analysis and prediction of student performance in secondary education using a dataset from two Portuguese schools [1]. The dataset includes information about student grades as well as demographic, social, and school-related features. The dataset specifically demonstrates student performance in the Mathematics (mat) subject. It was modeled using five-level classification approaches including K closest neighbor (KNN) and decision tree algorithms. Overall, the target attribute, the final grade that determines the success or failure of the school year G3 has a high correlation with the attributes G2 and G1. This implies that predicting G3 without considering G2 and G1 is challenging, but the prediction becomes significantly more useful. This paper will look into the benefits of using such a dataset for student performance analysis and prediction by implementing various methodologies and tools.

Introduction:

The analysis and prediction of student performance in secondary education play a crucial role in understanding educational outcomes and identifying factors that contribute to academic achievement. This report will analyze student performance analysis and prediction based on a dataset from two Portuguese schools by utilizing analytics and machine learning algorithms to obtain insights into student performance and forecast future outcomes. The dataset is made up of information gathered through school reports and questionnaires, with an emphasis on student achievement in the field of Mathematics. We will go through problem formulation, data acquisition and preparation, data exploration and modeling process.

Task 1 – Problem Formulation, Data Acquisition and Preparation

This task focuses on the selection of "Student Performance Data Set" from the UCI Machine Learning Repository. The dataset has been chosen as it satisfies the given criteria and contains relevant columns for our analysis and prediction of student performance. It consists of a total of 395 rows and 10 selected columns and the details in selected columns are as follows:

- G3 - final grade (numeric: from 0 to 20, output target)
- sex - student's sex (binary: 'F' - female or 'M' - male)
- age - student's age (numeric: from 15 to 22)
- studytime - weekly study time (numeric: 1 - <2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - >10 hours)
- health - current health status (numeric: from 1 - very bad to 5 - very good)
- failures - number of past class failures (numeric: n if 1<=n<3, else 4)
- Medu - mother's education (numeric: 0 - none, 1 - primary education (4th grade), 2 - 5th to 9th grade, 3 - secondary education 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)
- G1 - first period grade (numeric: from 0 to 20)
- G2 - second period grade (numeric: from 0 to 20)

These columns include information on the students' grades, demographics, and other important aspects that can affect their performance.

1.1 Problem Formulation:

The goal of this task is to analyze and predict student performance in secondary education based on the selected columns from the Student Performance Data Set. It is expected to gain insights into the factors that influence student performance and build models to predict the final grade ('G3') of the students by examining this dataset. This analysis can assist educators and policymakers better identify the major factors influencing academic achievement and give specific support to students who are struggling academically.

1.2 Data Acquisition:

The Student Performance Data Set was acquired from UCI Machine Learning Repository (https://archive.ics.uci.edu/ml/datasets.php) after exploring all the available datasets. The "Student Performance Data Set" is considered as a suitable choice based on its compatibility with the given criteria.

1.3 Data Preparation:

In the data preparation phase, several operations were performed on the dataset to ensure its suitability for analysis and modeling.

Data Cleaning:

After data is acquired with 10 selected columns, missing values are checked in the DataFrame using the isnull().sum() method. Fortunately, no missing values were found. Additionally, we identified and dropped duplicate rows using the duplicated() method and drop_duplicates() function.

1.4 Feature Engineering:

In this step, we performed feature engineering to enhance the dataset for analysis and modeling. This involved converting categorical columns into numerical format using label encoding.

By applying label encoding to the 'sex' and 'school' columns, we transformed them from categorical values (e.g., 'F': 0, 'M': 1 for 'sex' column, 'GP': 0, 'MS': 1 for 'school' column) into numerical representations. This conversion enables me to utilize these columns in numerical form for further analysis and modeling.

Moreover, to enrich the dataset and potentially enhance model performance, I created a new column named 'G3_classification' based on the 'G3' column in order to apply five-level classification approaches modeling. The 'G3_classification' column assigns a classification label to each student based on their final grade ('G3'). We used a lambda function to apply the classification logic, where we can analyze the students' grades by 5-level classification:

- grades between 0-9 = 5
- grades between 10-11 = 4
- grades between 12-13 = 3
- grades between 14-15 = 2
- grades between 16-20 = 1

Our final data will look like this:

	G1	G2	school	sex	age	health	Medu	failures	studytime	G3	G3_classification
0	5	6	0	0	18	3	4	0	2	6	5
1	5	5	0	0	17	3	1	0	2	6	5
2	7	8	0	0	15	3	1	3	2	10	4
3	15	14	0	0	15	5	4	0	3	15	2
4	6	10	0	0	16	5	3	0	2	10	4

Figure 1. First 5 rows of student performance dataset

Task 2 – Data Exploration

To explore the data loaded and prepared in Task 1, various analyses and visualizations will be performed to gain insights into the dataset.

2.1 Exploring each column
In this step, I will analyze each column in the dataset individually to understand its distribution and gain insights by implementing descriptive statistics and graphical visualizations.

Descriptive statistics:
The descriptive statistic table provides key statistical measures such as count, mean, standard deviation, minimum, maximum, and quartiles for each column.

	G1	G2	school	sex	age	health	Medu	failures	studytime	G3	G3_classification
count	393.000000	393.000000	393.000000	393.000000	393.000000	393.000000	393.000000	393.000000	393.000000	393.000000	393.000000
mean	10.895674	10.704835	0.117048	0.473282	16.694656	3.549618	2.743003	0.335878	2.035623	10.399491	3.569975
std	3.320746	3.764686	0.321888	0.499922	1.277113	1.391742	1.093864	0.745161	0.841375	4.584323	1.346126
min	3.000000	0.000000	0.000000	0.000000	15.000000	1.000000	0.000000	0.000000	1.000000	0.000000	1.000000
25%	8.000000	9.000000	0.000000	0.000000	16.000000	3.000000	2.000000	0.000000	1.000000	8.000000	2.000000
50%	11.000000	11.000000	0.000000	0.000000	17.000000	4.000000	3.000000	0.000000	2.000000	11.000000	4.000000
75%	13.000000	13.000000	0.000000	1.000000	18.000000	5.000000	4.000000	0.000000	2.000000	14.000000	5.000000
max	19.000000	19.000000	1.000000	1.000000	22.000000	5.000000	4.000000	3.000000	4.000000	20.000000	5.000000

Figure 2: Descriptive statistics of student attributes

From figure 2, it can be seen that the average grades for G1, G2, and G3 are roughly 10.9, 10.7, and 10.4, with minor variation among students. The dataset contains records from two separate schools, with one institution accounting for around 11.7% of the data. Male and female

students are almost equally represented. Students' ages range from 15 to 22, with an average of 16.7 years. Other attributes such as health, mother's education level (Medu), failures, and study time also provide insights into the students' well-being, educational background, and habits. These initial observations highlight the diversity and key characteristics of the student population in the dataset.

Exploring Data(visualization)

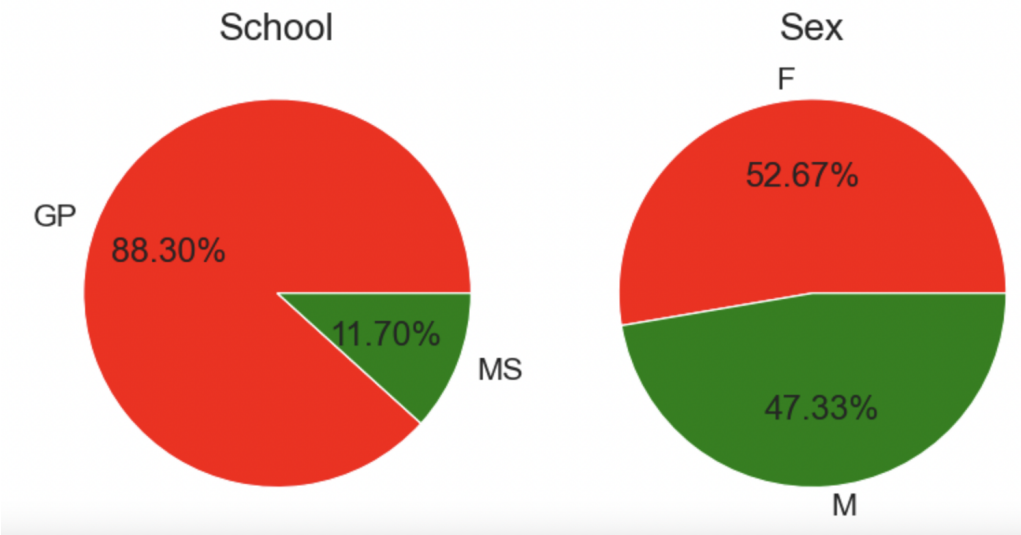

Figure 3: Distribution of school and sex

- Figure 3 provides insights into the distribution of students across different schools and genders. It reveals that the majority of students, approximately 88.30% of the dataset, come from the GP school. Due to the inequality representation in the dataset contributed from 2 schools, there could be a potential bias towards this particular school in the data collection process.
- On the other hand, male and female students show a roughly equal distribution with females accounting for around 47.33% of the overall dataset and boys accounting for approximately 52.67%. This indicates that the dataset has a balanced gender representation, implying that gender-based analysis and comparisons can be performed effectively without severe gender imbalances.

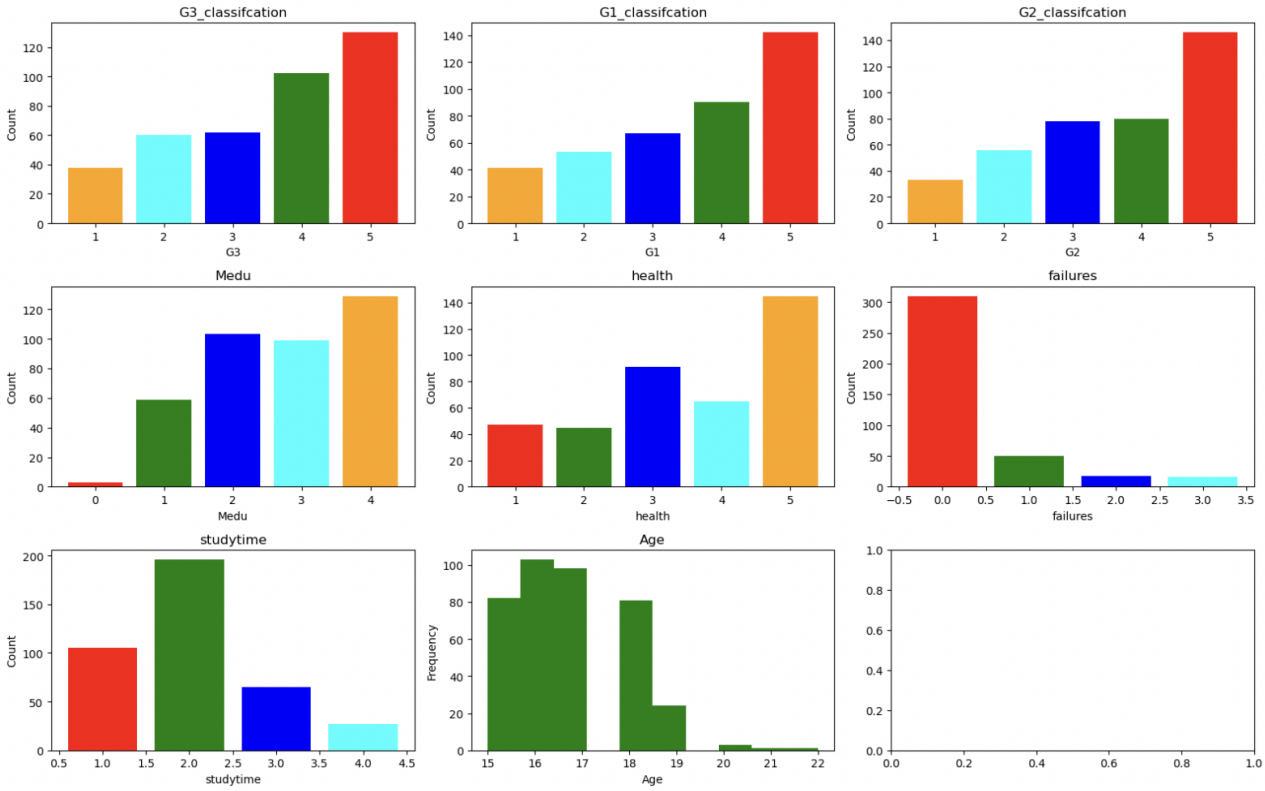

Figure 4: Distribution of student attributes

- The bar chart and histogram provide visual representations of the distribution of various student attributes. In the bar chart, students received grades between 0-9, indicating a failure account for the largest portion for G3, G1, and G2 in terms of 5 level classification. This highlights the need for academic support and intervention to improve student performance.
- In terms of mother's education level, the bar chart reveals a considerable number of students who have mothers with a higher education degree, followed by those with 5th to 9th grade education and secondary education. This shows that the students' mothers have a typically positive educational background, which may influence their academic achievement.
- A substantial number of students report having excellent health. This shows that the majority of students in the dataset believe they are in good physical health, which can improve their overall academic performance and attendance.
- In terms of failures, the majority of students do not have any recorded failures. This suggests that most students have been able to maintain a satisfactory academic track record without encountering significant setbacks or repeating grades.
- It can be seen that the majority of students study for 2-5 hours each week, followed by those who study for less than 2 hours. It helps to gain insights into study habits and time management among students, with the majority devoting a reasonable amount of time to studying.
- Lastly, the histogram depicting the age distribution of students highlights that the ages range from 15 to 22 years. There is a noticeable gap between the age groups of 17-18

and 19-20, which could be attributed to students graduating from high school and entering college or other higher education institutions.

Overall, I had explored some useful insights into the distribution and features of student variables in the dataset which help to understand the overall profile of the student population and identify areas that may require additional attention or further analysis.

2.2 Exploring the relationships between all pairs of columns

To explore the relationships between different pairs of columns, I have selected 10 pairs of columns to analyze. The pairs I have chosen include the final grades (G3) with other attributes, as well as the age and failures columns.

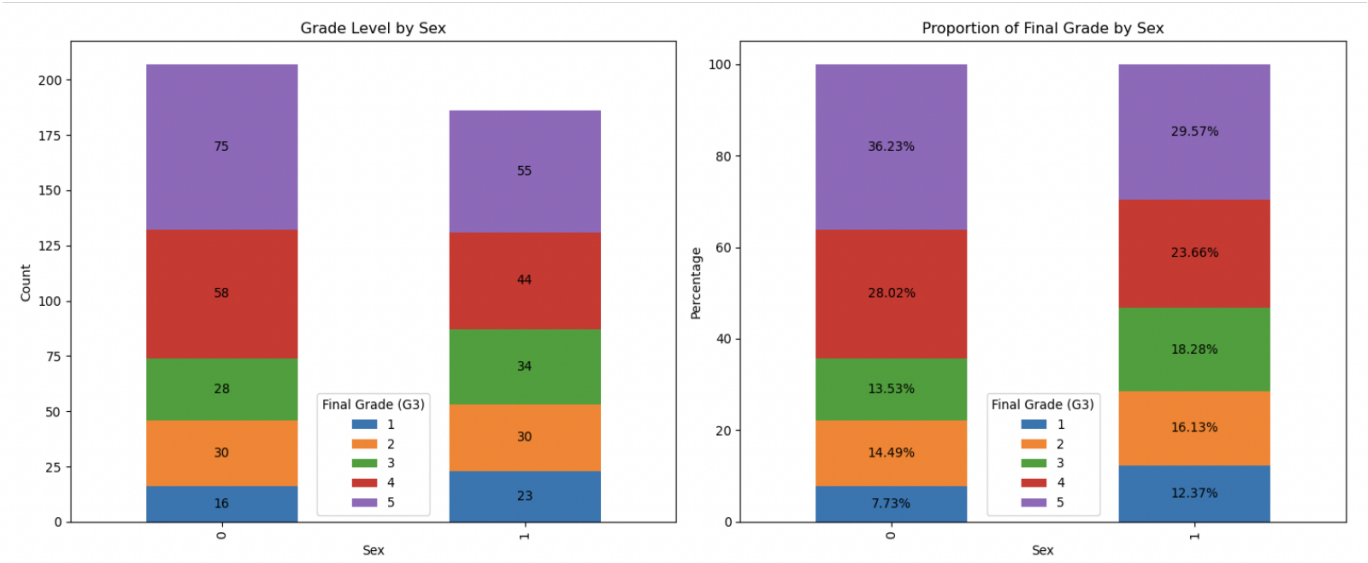

Figure 5: Proportion of Final Grade by Sex

When the relationship between final grades (G3) and gender is examined, it becomes clear that male students' academic performance is better compared to female's. According to figure 5, In comparison to male students (29.57%), a higher proportion of female students (36.23%) obtain grades ranging from 0 to 9, indicating a failure in the course. There are 58 female students and 44 male students with passing grades (10-12), demonstrating a larger pass rate for females. Interestingly, the proportion of female students attaining grades 14-16 is the same as that of male students, while the proportion of male students (12.37%) achieving the highest grades (16-20) is nearly double that of females (7.73%). These findings suggest a disparity in academic achievement, with male students performing marginally better overall.

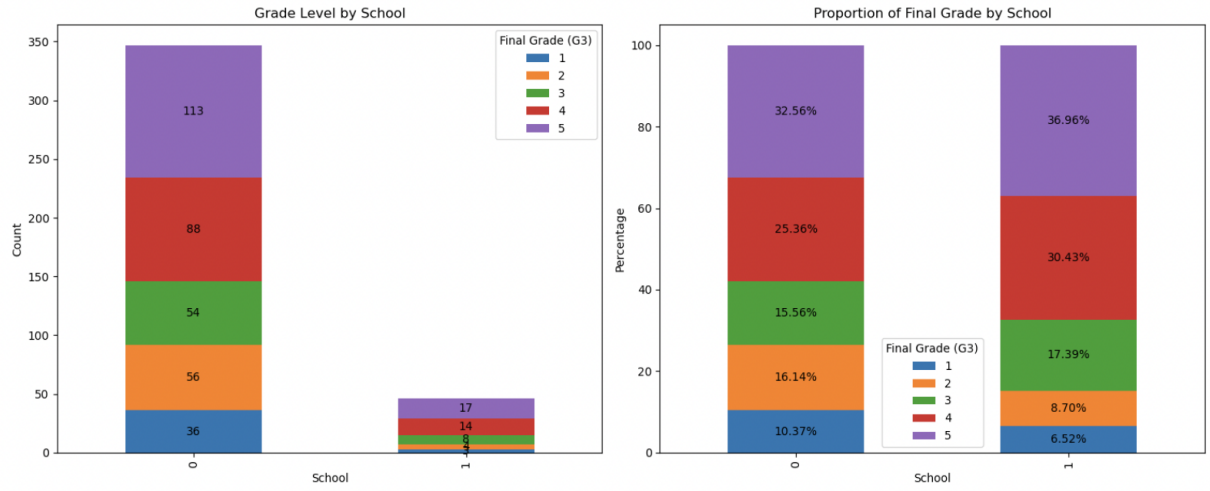

Figure 6: Proportion of Final Grade by School

Figure 6 highlights that the majority of students come from the GP school. This suggests a potential bias towards this specific school in the data collection process. Despite this, there does not appear to be a major difference in grades between the two institutions. As a result, the school of choice does not appear to have a significant impact on student performance as indicated in their final grades.

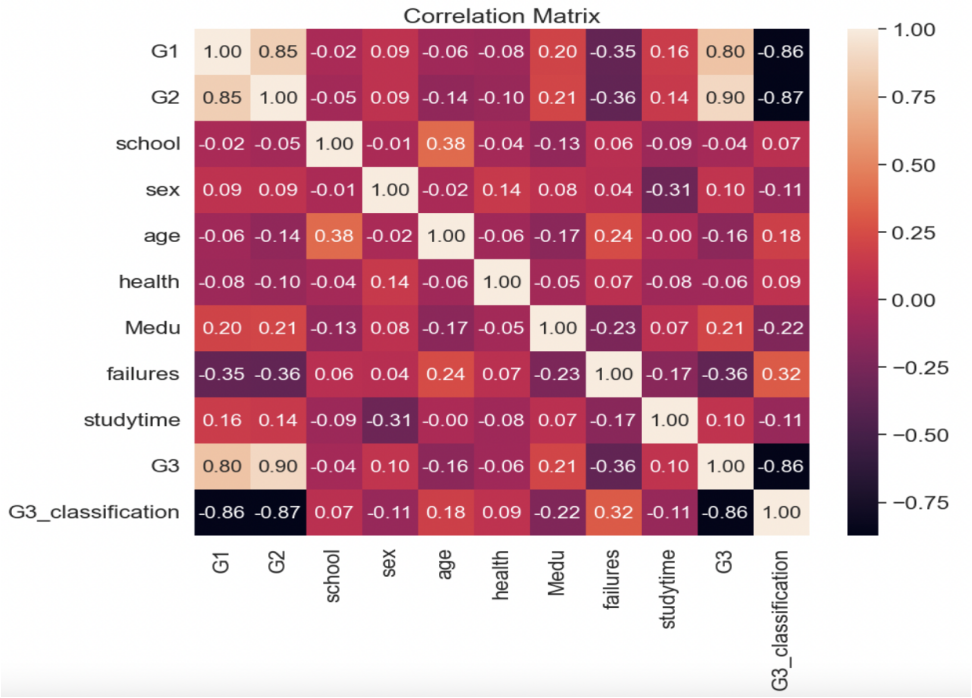

Figure 7: Attributes correlated to final grade

To further explore the relationship between attributes and the success or failure of the school year, a correlation matrix was utilized. Together with scatter plots which helps to visualize relationships between pairs easier.

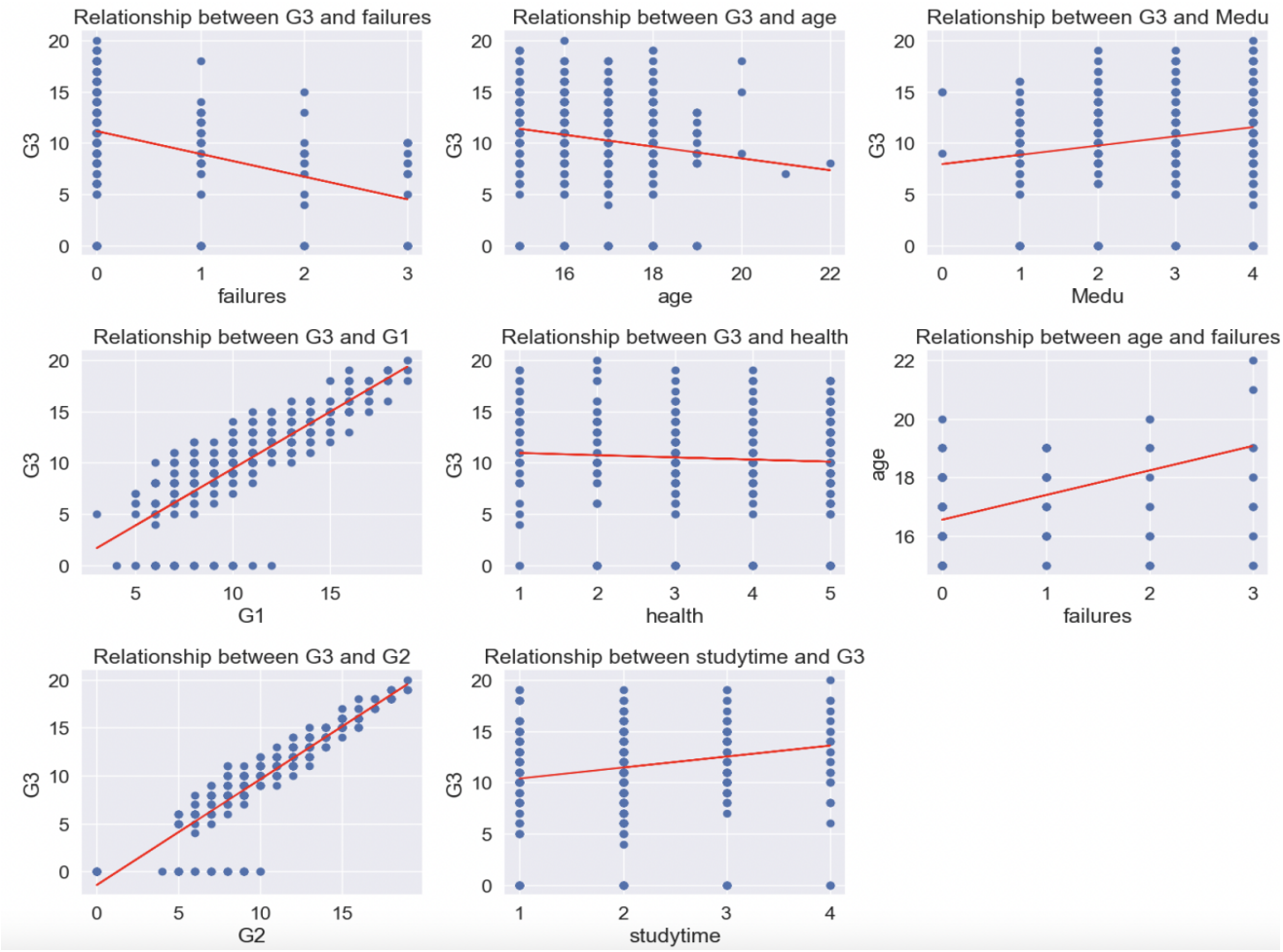

Figure 8: Scatter plots between pairs

G1, G2, and G3: The attributes with the highest absolute correlations (both positive and negative) with the final grade (G3) were identified. Among these attributes, G1 and G2 exhibited a correlation ranging from 0.8 to 0.9. Additionally, scatter plots were generated to visually depict the relationship between G1, G2, and G3. The scatter plots showed a near-perfect distribution, emphasizing the strong correlation between these variables, despite the presence of a few outliers. This analysis confirms the significant influence of previous grades (G1 and G2) on the final grade (G3) and supports the predictability of student performance based on their past academic achievements.

G3 and Age: The scatter plot shows a slightly negative correlation between age and G3 (final grade). As a result, as students become older, their final grades tend to fall. This discovery implies that age may have an impact on academic achievement, and it may be worthwhile to investigate the elements that contribute to this tendency.

G3 and Medu, studytime and G3: The correlation study finds a positive association with a coefficient of 0.21 between G3 (final grade) and Medu (mother's education), showing a mild positive relationship. Likewise, there is a 0.1 weak positive connection between study duration

and G3. However, when the scatter plots are examined, both associations appear to be slightly positive. This shows that, while the mother's education and study time may have an impact on final grades, there are certainly other factors at work as well.

G3 and health: The correlation study and scatter plot of G3 (final grade) and health show no discernible association. This means that a student's level of health does not appear to have a substantial impact on their final grade. Other factors may be more important in influencing academic performance, and greater research into these factors may be beneficial.

Age and Failures: The positive correlation and linear association between age and failures is another intriguing finding from the correlation analysis. As students get older, they tend to have more failures in the past. This insight emphasizes the importance of age in evaluating academic achievement and addressing potential problems.

Overall, the study found that male students outperformed female students in academic performance. The school of choice had no effect on academic performance. Previous grades (G1 and G2) were strongly associated with the final grade (G3), although age was inversely related. Mother's education and study time had a minor favorable influence. These studies shed light on the elements that influence student achievement.

2.3 Posing one meaningful question and exploring the data

In order to better understand insights on the performance of students, I had come up with the question if there is a relationship between the amount of study hours and final grades for students who had failed at least one course compared to those who had not failed any courses. To get the answer, I have calculated the average study hours and final grades for both groups and a t-test to determine whether there is a significant difference.

The result shows the average study hours for failed students (failures > 0) were 1.78 hours, with a final grade of 7.27. On the other hand, students with failures equal 0, had an average study time of 2.10 hours and a final grade of 11.24. The t-statistic of -3.11 suggests that there is a significant difference between the two groups. Furthermore, the p-value of 0.002 indicates that this difference is unlikely to be coincidental.

Based on these findings, it can be concluded that students who have failed at least one course have lower average study hours and poorer average final marks than those who have not failed any courses. This implies that the number of study hours may influence academic achievement, particularly for individuals who have previously failed. It emphasizes the need of sufficient time for study as well as the need for additional support and interventions for students who have had academic failures.

Task 3 – Data Modeling

3.1 Data Splitting:

In this task, data modeling is performed using two classification models: Decision Tree Classifier and K-Nearest Neighbors (KNN). The data is split into a training set and a test set and three different suites of training and test sets with the purpose to evaluate the performance of each model on the different suites and compare their results:

- Suite 1 (50% for training, 50% for testing)
- Suite 2 (60% for training, 40% for testing)
- Suite 3 (80% for training, 20% for testing)

To ensure the reproducibility, the train_test_split function was used from the scikit-learn package with the random_state parameter to 42 in each split to obtain consistent results across different runs.

3.2 Model Training and Evaluation:

Model Training:

Decision Tree Classifier:

- Parameter selection: To select the appropriate model parameters, GridSearchCV method from scikit-learn is implemented to find the best parameters in order to avoid overfitting or underfitting. GridSearchCV performs an exhaustive search over a specified parameter grid and uses cross-validation to determine the best parameters for the model. Parameters for Decision Tree Classifier include max_depth (I set range from 1 to 6), min_samples_split, and criterion.
- Model training: After selecting the best parameters, the Decision Tree Classifier is trained using the identified method with the chosen parameters on each suite.

K-Nearest Neighbors (KNN):

- Parameter selection: For parameter selection, I defined a parameter grid containing different values for key parameters like n_neighbors (set range from 1 to 100 to find the best fit n_neighbors) and weights. I then used the GridSearchCV method to find the best parameters, especially n_neighbor for the KNN model.
- Model training: After selecting the best parameters, I trained the KNN model using the identified method with the chosen parameters on each suite.

Evaluation:

The performance of the results for each model and suite will be evaluated respectively, in terms of confusion matrix, classification accuracy, precision, recall, and F1 score.

The performances of the models on the training and test sets, along with the evaluation metrics are shown below:

Suite 1 (50% for training, 50% for testing)

DecisionTreeClassifier

```
Confusion Matrix:
[[13  5  0  0  0]
 [ 1 24  6  1  0]
 [ 0  2 15 17  0]
 [ 0  0  1 31 19]
 [ 0  0  0  1 61]]
Classification Accuracy: 0.7309644670050761
Precision: 0.7287578550481776
Recall: 0.7309644670050761
F1 Score: 0.7197704285386305
```

KNN

```
Confusion Matrix:
[[10  8  0  0  0]
 [ 3 19  8  2  0]
 [ 0  4 17 13  0]
 [ 0  0  3 37 11]
 [ 0  0  0  4 58]]
Accuracy: 0.7157360406091371
Precision: 0.6981141696078582
Recall: 0.6620559245203458
F1 Score: 0.6747615917048426
```

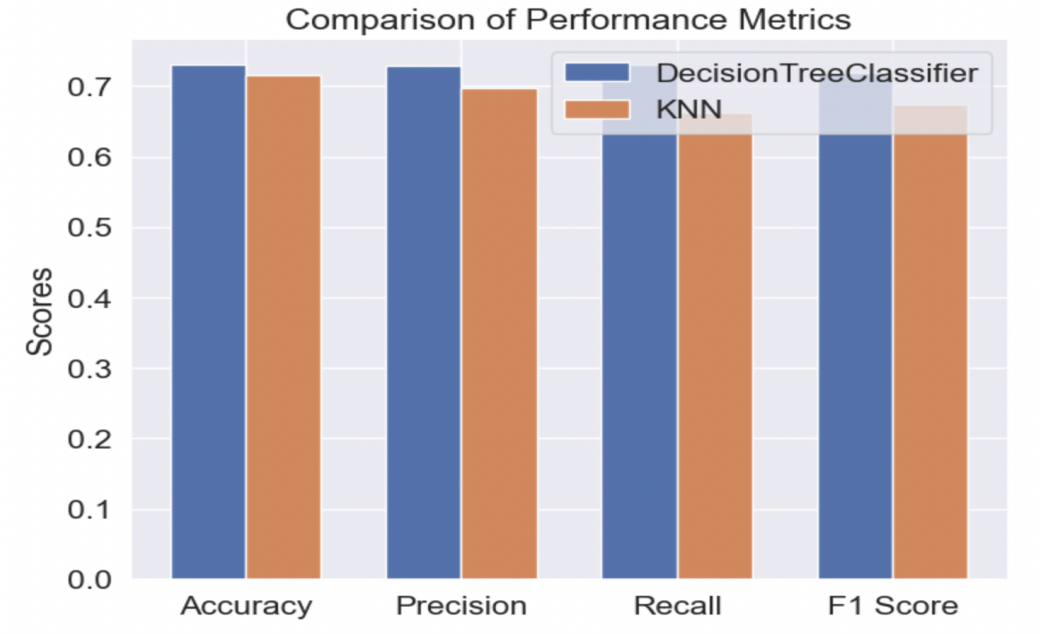

Figure 9: Comparison of Performance Metrics for Suit 1

In Suit 1, both the Decision Tree Classifier and KNN achieved relatively similar performance. The Decision Tree Classifier had a slightly higher accuracy (0.7309644670050761) and F1 score (0.7197704285386305) compared to KNN. However, KNN had a higher precision (0.6981141696078582) and recall (0.6620559245203458) compared to the Decision Tree Classifier.

Suite 2 (60% for training, 40% for testing)

DecisionTreeClassifier

```
Confusion Matrix:
[[12  3  0  0  0]
 [ 0 22  5  0  0]
 [ 0  2 20  4  0]
 [ 0  0  6 26 10]
 [ 0  0  0  2 46]]
Classification Accuracy: 0.7974683544303798
Precision: 0.8058719302339148
Recall: 0.7974683544303798
F1 Score: 0.7946455654650192
```

KNN

```
Confusion Matrix:
[[ 9  6  0  0  0]
 [ 0 18  8  1  0]
 [ 0  3 14  9  0]
 [ 0  0  7 27  8]
 [ 0  0  0  7 41]]
Accuracy: 0.689873417721519
Precision: 0.7199592689740474
Recall: 0.6604304029304029
F1 Score: 0.679805075448806
```

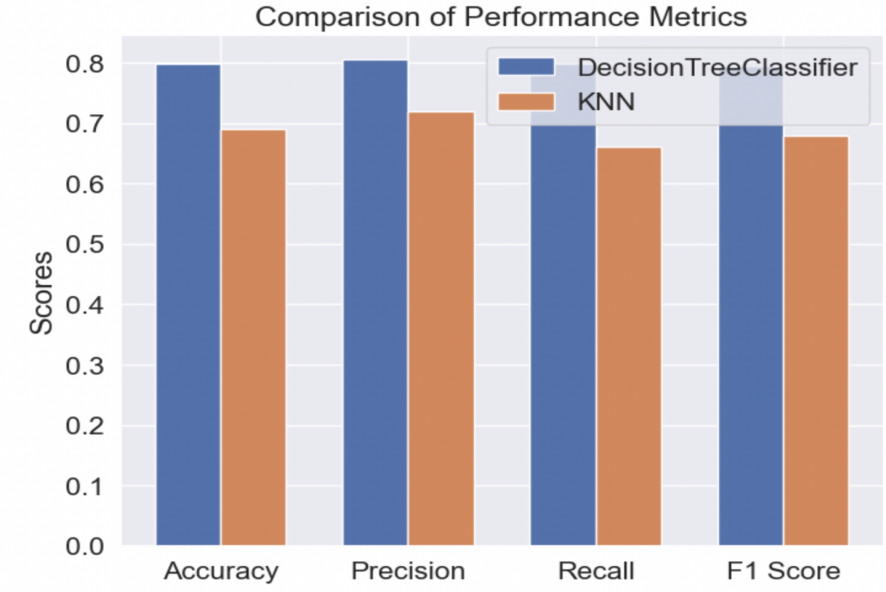

Figure 10: Comparison of Performance Metrics for Suit 2

In Suit 2, the Decision Tree Classifier outperformed KNN in all performance metrics. The Decision Tree Classifier achieved a higher accuracy (0.7974683544303798), precision (0.8058719302339148), recall (0.7974683544303798), and F1 score (0.7946455654650192) compared to KNN.

Suite 3 (80% for training, 20% for testing)

DecisionTreeClassifier

```
Confusion Matrix:
[[ 7  2  0  0  0]
 [ 0 13  3  0  0]
 [ 0  1  9  3  0]
 [ 0  0  1 11  5]
 [ 0  0  0  2 22]]
Classification Accuracy: 0.7848101265822784
Precision: 0.7878867791842475
Recall: 0.7848101265822784
F1 Score: 0.7837242491933483
```

KNN

```
Confusion Matrix:
[[ 7  2  0  0  0]
 [ 0 12  3  1  0]
 [ 0  0 10  3  0]
 [ 0  0  1 12  4]
 [ 0  0  0  4 20]]
Accuracy: 0.7721518987341772
Precision: 0.800952380952381
Recall: 0.7672448466566114
F1 Score: 0.7795445445445446
```

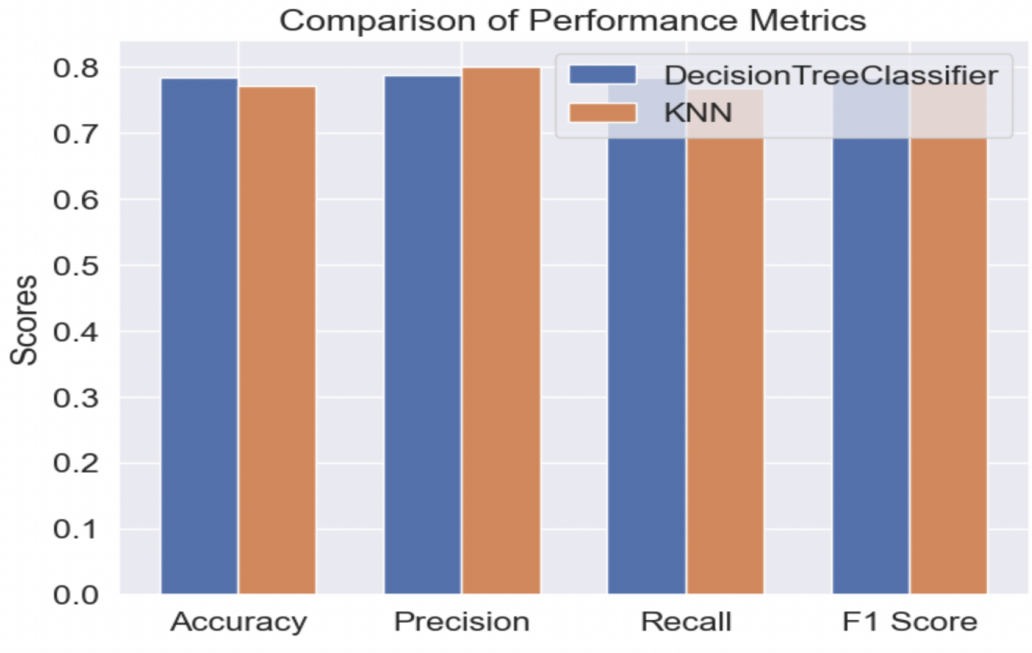

Figure 11: Comparison of Performance Metrics for Suit 3

In Suit 3, the Decision Tree Classifier and KNN performed relatively close to each other. The Decision Tree Classifier achieved a slightly higher accuracy (0.7848101265822784), precision (0.7878867791842475), and F1 score (0.7837242491933483) compared to KNN. However, KNN had a higher recall (0.7672448466566114) compared to the Decision Tree Classifier.

Comparing the results across the three suites:

- Decision Tree Classifier:
    - Suit 2 achieved the highest classification accuracy (0.7974683544303798), precision (0.8058719302339148), recall (0.7974683544303798), and F1 score (0.7946455654650192), indicating the best overall performance among the three suites.

    - Suit 1 and Suit 3 had slightly lower performance metrics but were still relatively close in terms of accuracy, precision, recall, and F1 score.
- KNN:
    - Suit 3 achieved the highest classification accuracy (0.7721518987341772), precision (0.800952380952381), and F1 score (0.7795445445445446) among the three suites.
    - Suit 1 and Suit 2 had slightly lower performance metrics but were relatively close in terms of accuracy, precision, recall, and F1 score.

Overall, if we consider both models, Suit 2 performed the best with the Decision Tree Classifier, while Suit 3 performed the best with the KNN model. Therefore, if the dataset used for evaluation is similar to the one used for comparison, using the Decision Tree Classifier with Suit 2 would be recommended for achieving the highest performance.

Conclusion

In conclusion, this study successfully employed a dataset from two Portuguese schools in order to perform analysis and prediction of student performance. This report has gone through problem formulation, data acquisition and preparation, data exploration and looking into the algorithms for modeling process. By choosing Decision Tree and KNN algorithms, when these models were compared, it was discovered that Decision Tree classification performed better than KNN classification. Based on the results, Decision Tree classification is identified as the most suitable approach for addressing the problem, providing an accuracy of 77.2% and yielding the most accurate results. These conclusions emphasize the importance of utilizing appropriate prediction models to improve student performance assessment and inform educational strategies.

References:

[1] Cortez, Paulo and Silva, António. 2008. Student Performance Dataset. UCI Machine Learning Repository. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/student+performance