

MLG382: Guided Project Report

Ann-Mari De Beer

Jason Romario Maracha Bond

Tiaan Kemp

Jeané Christine Thomas - 577864

**Tuesday, 22 April 2025**

Dino Giovannoni • Hellen Machingauta • Francois Venter

Table of Contents

[Report title 3](#_Toc195986503)

[Problem statement 3](#_Toc195986504)

[Hypothesis Generation 3](#_Toc195986505)

[Understanding the Data 3](#_Toc195986506)

[Exploratory Data Analysis 5](#_Toc195986507)

[1. Univariate Analysis 5](#_Toc195986508)

[2. Bivariate Analysis 11](#_Toc195986509)

# Report title

*GitHub link:  
Render link:*

## Problem statement

Enhancing the academic experience of BrightPath Academy learners through personalised education and early academic interventions.

## Hypothesis Generation

To predict a Leaner's GradeClass, a multi-class classification model is used. The probability of a Leaner belonging to a particular class is calculated using the softmax function:

The probability for each class is given by the softmax function:

Where:

: actual GradeClass

: Possible class

: Input feature vector

: Vector of weights for class

: Predicted probability that a specific Leaner belongs to class

## Understanding the Data

The dataset contains the details of BrightPath Academy learners. These details include demographic attributes such as age, gender, and ethnicity. The next data section contains academic information, and this includes the learners’ academic experience, parental education level, the amount of study time a week, number of absences, and whether they receive tutoring and parental support. The dataset also includes binary indicators (yes/no) for leaners are involved in extracurricular activities, sports, music and volunteering.

The final columns are the learners’ GPA and grade class, which represent their respective academic performance and classification level. These variables serve as the key indicators for evaluating the Leaner outcomes.

The dataset aims to support early academic intervention strategies by identifying potential factors that my influence the leaner’s success.

Each learner is uniquely identified by a LeanerID, which allows for tracking without revealing any personal details about the leaner.

Target variable: ***GradeClass***

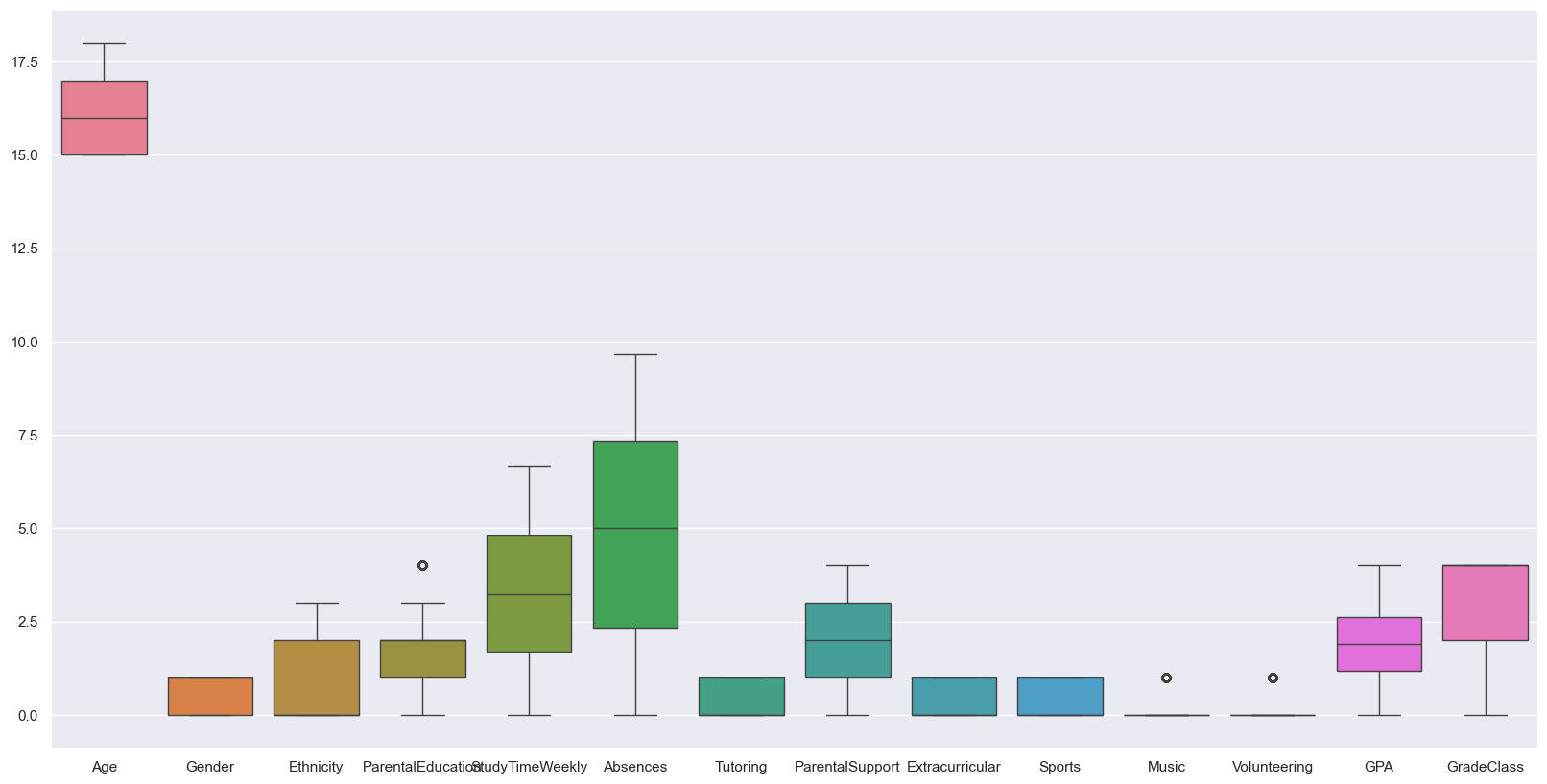


Figure 1

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Column** | **Non-Null Count** | **Data type** |
| 0 | LeanerID | 2392 | int64 |
| 1 | Age | 2392 | int64 |
| 2 | Gender | 2392 | int64 |
| 3 | Ethnicity | 2392 | int64 |
| 4 | ParentalEducation | 2392 | int64 |
| 5 | StudyTimeWeekly | 2392 | float64 |
| 6 | Absences | 2392 | int64 |
| 7 | Tutoring | 2392 | int64 |
| 8 | ParentalSupport | 2392 | int64 |
| 9 | Extracurricular | 2392 | int64 |
| 10 | Sports | 2392 | int64 |
| 11 | Music | 2392 | int64 |
| 12 | Volunteering | 2392 | int64 |
| 13 | GPA | 2392 | float64 |
| 14 | GradeClass | 2392 | float64 |

## Exploratory Data Analysis

### 1. Univariate Analysis

#### 1.1 Gender Distribution

A graph of a person's birth rate

AI-generated content may be incorrect.

Figure 2: Gender Distribution

**Observation:** Figure 2 shows there are more female learners than male learners.

#### 1.2 Ethnicity Distribution

A graph showing different colored squares

AI-generated content may be incorrect.

Figure 3: Ethnicity Distribution

**Observation:** Figure 3 displays that a majority of learners are Caucasian.

#### 1.3 Parental Education Distribution

A graph of a graph showing different colored squares

AI-generated content may be incorrect.

Figure 4: Parental Education Distribution

**Observation:** The majority of learners’ parents have an educational background as seen in Figure 4.

#### 1.4 Tutoring Distribution

A graph with a bar and a number of squares

AI-generated content may be incorrect.

Figure 5: Tutoring Distribution

**Observation:** Figure 5 shows that many learners do not receive tutoring.

#### 1.5 Parental Support Distribution

A graph of support distribution

AI-generated content may be incorrect.

Figure 6: Parental Support Distribution

**Observation:** Figure 6 shows that the data is varied.

#### 1.6 Extracurricular Participation

A graph showing a number of different colored squares

AI-generated content may be incorrect.

Figure 7: Extracurricular Participation

**Observation:** Majority of learners do not participate in extracurricular activities (Figure 7).

#### 1.7 Sports Participation

A graph showing a number of persons

AI-generated content may be incorrect.

Figure 8: Sports Participation

**Observation:** As seen in figure 8, the majority of learners do not participate in any sports.

#### 1.8 Music Participation

A graph with a green and orange rectangle

AI-generated content may be incorrect.

Figure 9: Music Participation

**Observation:** Majority of learners do not take music as an extracurricular activity (Figure 9).

#### 1.9 Volunteering Participation

A graph with a green and orange rectangle

AI-generated content may be incorrect.

Figure 10: Volunteering Participation

**Observation:** Figure 10 shows that a majority of learners do not volunteer, and as a result, volunteering is the least chosen extracurricular activity.

#### 1.10 Numerical Variables Distribution: Age

A graph showing the distribution of age

AI-generated content may be incorrect.

Figure 11: Numerical Variables Distribution of Age

**Observation:** Age is uniform, slightly more 15-year-olds (Figure 11).

#### 1.11 Numerical Variables Distribution: Study Time Weekly

A graph showing a number of data

AI-generated content may be incorrect.

Figure 12: Numerical Variables Distribution of Study Time Weekly

**Observation:** Study time has uniform distribution, slightly right skewed (Figure 12).

#### 1.12 Numerical Variables Distribution: Absences

A graph showing the distribution of absences

AI-generated content may be incorrect.

Figure 13: Numerical Variables Distribution of Absences

**Observation:** Absences are uniform, slightly left skewed (Figure 13).

#### 1.13 Numerical Variables Distribution: GPA

**A graph showing a distribution of gpa

AI-generated content may be incorrect.**

Figure 14: Numerical Variables Distribution of GPA

**Observation:** GPA has a normal distribution (Figure 14).

### 2. Bivariate Analysis

#### 2.1 Categorical vs Gradeclass: Gender

A graph showing genders and genders

AI-generated content may be incorrect.

Figure 15: Gender vs Gradeclass

**Observation:** Figure 15 shows that female Leaners (Gender = 1) show a slightly higher median GPA and narrower spread, suggesting better overall academic consistency compared to male Leaners.

#### 2.2 Categorical vs Gradeclass: Ethnicity

A graph of different colored bars

AI-generated content may be incorrect.

Figure 16: Ethnicity vs GradeClass

**Observation:** There is visible variation across ethnic groups, as seen in Figure 16. Some groups show tighter GPA clusters around higher values, indicating potential differences in performance that could be explored further with support and engagement factors.

#### 2.3 Categorical vs Gradeclass: Parental Education

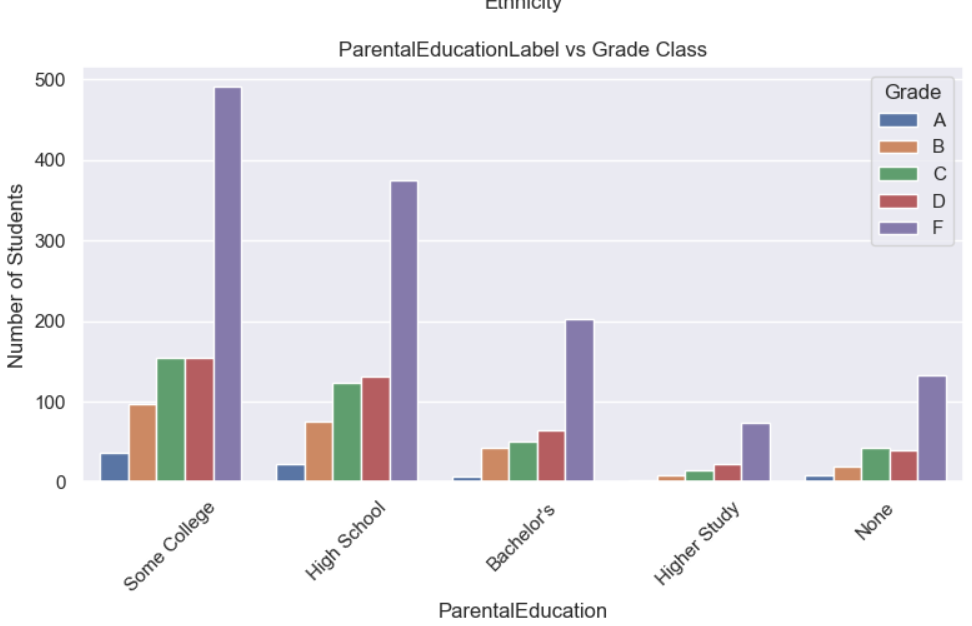


Figure 17: Parental Education vs GradeClass

**Observation:** Leaners whose parents have higher education levels tend to have higher GPAs. The GPA distribution shifts upward with increasing parental education, suggesting parental academic background plays a role in Leaner success (Figure 17).

#### 2.4 Categorical vs Gradeclass: Parental Support

A graph of different colored bars

AI-generated content may be incorrect.

Figure 18: Parental Support vs GradeClass

**Observation:** GPA tends to increase with higher levels of parental support. The violin plot in Figure 18 shows wider GPA distributions for Leaners with little or no support, and tighter, higher distributions with high support.

#### 2.5 Categorical vs Gradeclass: Tutoring

A graph of a bar graph

AI-generated content may be incorrect.

Figure 19: Tutoring vs GradeClass

**Observation:** Leaners receiving tutoring (Tutoring = 1) have a higher median GPA. This supports the effectiveness of academic intervention strategies, as displayed in Figure 19.

#### 2.6 Categorical vs Gradeclass: Extracurricular

A graph of different colored bars

AI-generated content may be incorrect.

Figure 20: Extracurricular vs GradeClass

**Observation:** Leaners involved in extracurricular activities show a slightly higher GPA median, though the overall impact appears moderate. Participation may support balanced development (Figure 20).

#### 2.7 Categorical vs Gradeclass: Sports

A graph of different colored bars

AI-generated content may be incorrect.

Figure 21: Sports vs GradeClass

**Observation:** Sports participation shows mixed results, evident in Figure 21. The GPA distribution is wide for both groups, indicating that athletic involvement doesn’t strongly correlate with academic outcomes alone.

#### 2.8 Categorical vs Gradeclass: Music

A graph of different colored squares

AI-generated content may be incorrect.

Figure 22: Music vs GradeClass

**Observation:** Leaners participating in music show higher GPA medians compared to non-participants. This may reflect benefits of discipline, structure, or creativity linked with music (Figure 22).

#### 2.9 Categorical vs Gradeclass: Volunteering

A graph of a graph showing a number of different colored bars

AI-generated content may be incorrect.

Figure 23: Volunteering vs GradeClass

**Observation:** Leaners who volunteer show a slightly higher GPA median, though the effect is subtle (Figure 23). Volunteering may indirectly support academic performance through skill development.

#### 2.10 Numerical vs GradeClass

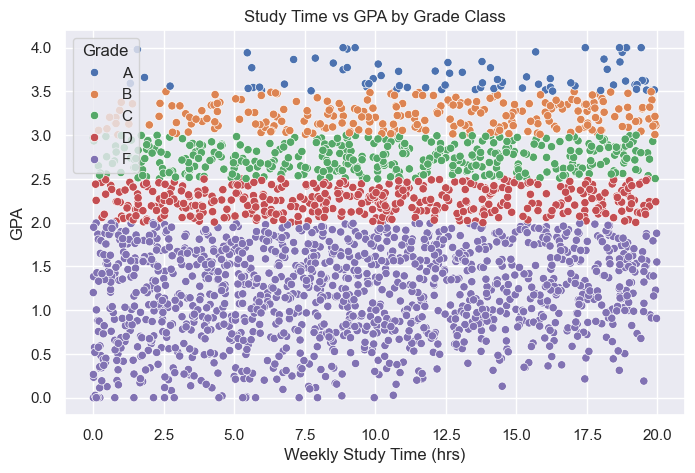


Figure 24: Numerical vs GradeClass

**Observation:** Learners who dedicate more hours to weekly study generally achieve higher GPAs. There is a clear positive trend though with some spread. So, while more study time helps, quality and other factors also play a role (Figure 24).

#### 2.11 Absences vs GPA by GradeClass

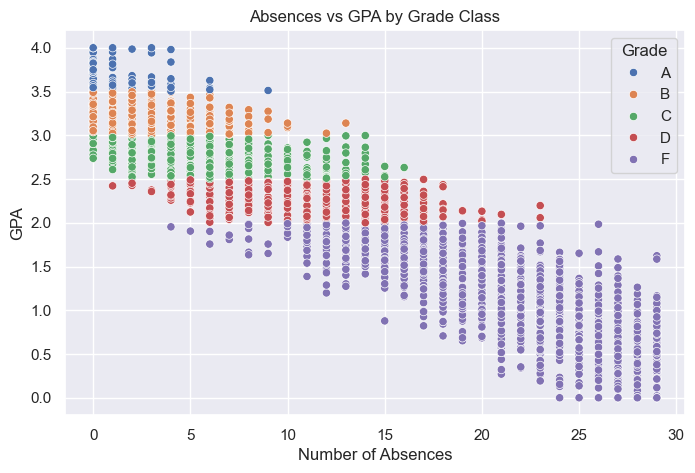


Figure 25: Absences vs GPA by GradeClass

**Observation:** A noticeable negative relationship exists between absences and GPA. Learners with frequent absences tend to have lower GPA scores, highlighting that consistent attendance is important for academic success (Figure 25).

#### 2.12 Age vs GPA by GradeClass

A graph of age vs gpa by grade

AI-generated content may be incorrect.

Figure 26: Age vs GPA by GradeClass

**Observation:** The majority of students are between 15 and 18 years old. GPA distributions across this age range appear relatively uniform, suggesting age does not have a strong linear impact on academic performance (Figure 26).

#### 2.13 GPA vs Weekly Study Time

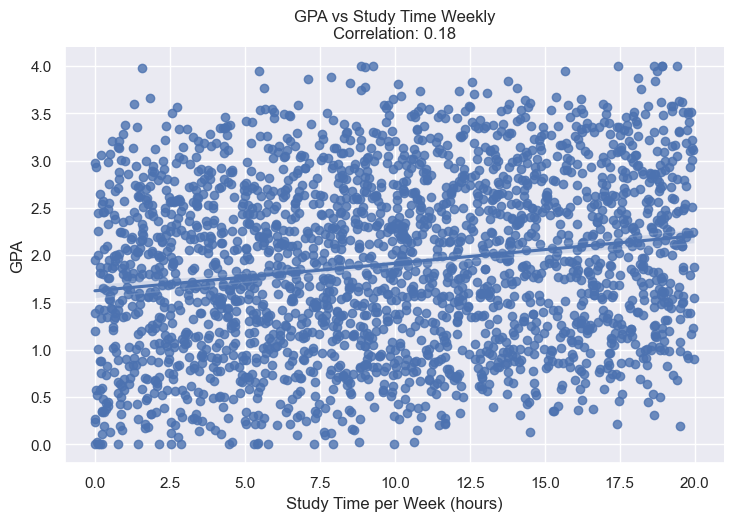


Figure 27: GPA vs Weekly Study Time

**Observation:** There is a positive correlation between study time and GPA. As students spend more hours studying each week, their GPA tends to increase. However, the correlation is moderate 0.18, suggesting that while study time matters, other factors like study effectiveness or support also play a role (Figure 27).

#### 2.14 GPA vs Absences

A graph showing a graph of a graph

AI-generated content may be incorrect.

Figure 28: GPA vs Absences

**Observation:** This relationship shows a negative correlation students with more absences generally have lower GPAs. The trend line clearly slopes downward, indicating that consistent attendance contributes to academic success (Figure 28).

#### 2.15 GPA vs Age

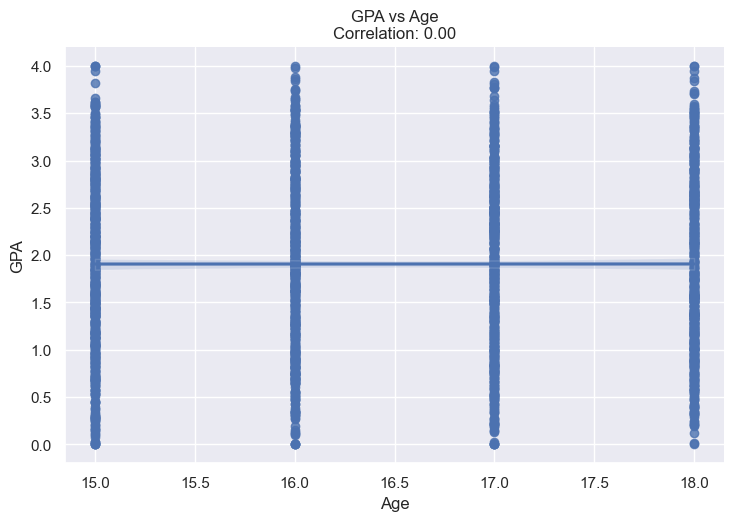


Figure 29: GPA vs Age

**Observation:** The correlation between GPA and age is weak or nearly flat. This suggests that age within the 15–18 range in this dataset has minimal effect on academic performance. GPA appears to be influenced more by behavioural and support factors than age alone. (Figure 29).

## Missing value and outlier treatment

|  |  |  |
| --- | --- | --- |
| **Column** | **Missing values** | **Outliers** |
| StudentID | 0 | 0 |
| Age | 0 | 0 |
| Gender | 0 | 0 |
| Ethnicity | 0 | 0 |
| ParentalEducation | 0 | 0 |
| StudyTimeWeekly | 0 | 0 |
| Absences | 0 | 0 |
| Tutoring | 0 | 0 |
| ParentalSupport | 0 | 0 |
| Extracurricular | 0 | 0 |
| Sports | 0 | 0 |
| Music | 0 | 0 |
| Volunteering | 0 | 0 |
| GPA | 0 | 0 |
| GradeClass | 0 | 0 |
| GenderLabel | 0 | 0 |
| EthnicityLabel | 0 | 0 |
| ParentalEducationLabel | 0 | 0 |
| ParentalSupportLabel | 0 | 0 |
| TutoringLabel | 0 | 0 |
| ExtracurricularLabel | 0 | 0 |
| SportsLabel | 0 | 0 |
| MusicLabel | 0 | 0 |
| VolunteeringLabel | 0 | 0 |
| GradeClassLabel | 0 | 0 |

## Evaluation Metrics for classification problem