

BrightPath Grade Predictor

Predictive Maintenance



Institution: Belgium Campus

Course: MLG382

Group S:

* Jeremia Fourie
* Juan Oosthuizen
* Busisiwe Radebe
* Phumlani Ntuli

Submission Date: 22 April 2025, 12:00 AM

# Problem Statement

BrightPath Academy faces challenges in identifying at-risk students early, understanding how extracurricular activities influence grades, and developing targeted support strategies. This project addresses these issues by building a predictive model for GradeClass and analyzing key factors affecting student outcomes.

# Hypotheses

We hypothesize that:

* Students with higher StudyTimeWeekly are more likely to achieve better grades.
* Higher Absences correlate with lower grades.
* Participation in Extracurricular activities positively impacts grades.
* ParentalSupport levels significantly influence student performance.

These hypotheses will be explored and tested in the notebooks/eda.ipynb notebook through data visualizations and statistical analysis.

# Preparing Data

In the Preparing Data phase, we first loaded the raw student dataset (2,392 records, 15 columns), validated that StudentID was unique for every row—confirming it served only as an identifier—and then dropped it to avoid introducing non‑predictive noise. Next, we ran data.info() to verify there were no missing values and that all remaining features were numeric, and used data.describe() to see that age clustered tightly (15–18 yrs, median 16), StudyTimeWeekly varied widely (IQR ≈ 9.7 hrs, max ≈ 20 hrs), and absences were right‑skewed (median 15 days, max 29). We also checked for duplicate rows (none found), applied IQR‑based outlier detection to flag extreme study‑time and absence cases for review, and computed skewness and kurtosis to highlight any features needing transformation or scaling.

# Exploratory Data Analysis (EDA)

To understand our cleaned student dataset, we first split variables into three types—numerical (StudyTimeWeekly, Absences, GPA), categorical (Gender, Ethnicity, Tutoring, Extracurricular, Sports, Music, Volunteering) and ordinal (Age, ParentalEducation, ParentalSupport, GradeClass)—so that each group could be visualized with the most informative chart: histograms/KDEs for continuous measures, pie charts for binary flags, and countplots for ordered categories. This targeted approach reveals both the shape of each feature’s distribution and its relationship to student performance.

## Univariate Analysis

* StudyTimeWeekly: Most students study between 0–10 hrs/week; a long right tail out to ~20 hrs identifies a small “super‑studier” subgroup whose habits may merit separate analysis or outlier treatment.
* Absences: Centered around 10–20 days with a tail up to 29, indicating a handful of chronic absentees; these extreme cases should be flagged for potential capping or deeper review.
* GPA: Peaks at ~1.5–2.0, but also shows a secondary spike at the maximum (4.0), suggesting ceiling effects or grade inflation among high achievers.
* Gender & Ethnicity: Gender is nearly balanced (51 % vs. 49 %), while Ethnicity is dominated by category 0 (50.5 %), with three smaller groups making up the rest—allowing us to keep all levels without severe sparsity.
* Participation Flags: Tutoring (~30 %), Extracurricular (~38 %), and Sports (~30 %) show healthy variation; Music (~20 %) and Volunteering (~16 %) are less common but still sufficiently represented to include as features.
* Age, ParentalEducation & ParentalSupport: Ages 15–18 are almost evenly split (slight peak at 15). ParentalEducation clusters at level 2 (≈ 940 records) and ParentalSupport at level 2–3 (≈ 700 each), indicating most students come from moderately educated, moderately supportive homes.
* GradeClass: Strongly skewed toward the top class (4) with ~1,210 students versus only ~110 in class 0, flagging a class‐imbalance issue for modeling.

## Bivariate Analysis

* Numeric vs. GradeClass (Boxplots): Both median StudyTimeWeekly and GPA rise steadily from lower to higher GradeClass, while median Absences fall—confirming our hypotheses that more study time and better attendance relate to higher grades.
* Categorical vs. GradeClass (Grouped Bars): The proportion of students in tutoring and extracurriculars grows with GradeClass, suggesting these supports are associated with better outcomes. Music and volunteering show smaller but consistent gains.
* Correlation Matrix: A very strong negative correlation between Absences and GPA (≈ –0.92) and a modest positive link between StudyTimeWeekly and GPA (≈ +0.18) highlight attendance as the single most powerful univariate GPA predictor, with study time also important but to a lesser degree. ParentalSupport shows a weaker yet meaningful positive correlation with GPA (≈ +0.19).
* Pairwise Scatter & KDE by GradeClass: Lower‑class students cluster in the high‐absence, low‑GPA region, while upper‑class students concentrate at low absences and high GPA. Overlap in StudyTime vs. GPA suggests some high‑achievers study less, warranting deeper segmentation.
* Group-Level Summary: As GradeClass increases from 0 to 4, average study time and GPA both decline while absences climb dramatically—underscoring how worsening attendance and shrinking study habits drive lower performance.

A screenshot of a computer screen

AI-generated content may be incorrect.

# Preprocessing Data

In this phase we transform our cleaned student dataset into its final, model‐ready form. First, we create the target variable GradeClass by applying our gpa\_to\_grade\_class mapping to the continuous GPA scores—binning each student into one of five ordered categories that reflect performance bands. Once the target is set, we drop the original GPA column to prevent leakage. The result is a fully supervised dataset with 2,392 students, 14 feature columns (all numeric or properly encoded) plus our new ordinal target. We save this “engineered” data as engineered\_data.csv for downstream modeling.

# Evaluation Metrics

# Model Building: Part 1 (Baseline Models)

# Model Building: Part 2 (Deep Learning)

# Model Deployment