Predictive Analytics for Academic Performance at BrightPath Academy

Prepared by: Trent Evans, Demica Smit, Bianca Long, Jade Riley

Course: MLG382 Project 1 | 2025

Platform: Deployed with Dash on Render.com

---

1. Problem Statement

BrightPath Academy is dedicated to delivering a personalized learning experience and aims to provide personalised education and early academic intervention to ensure every learner reaches their full potential. Despite this mission, the school currently faces challenges such as delayed identification of students who are at risk, limited tools for tailored intervention, unclear influence of extracurricular involvement, and an overwhelming amount of data with little actionable value. The purpose of this project is to address these issues by analysing student performance data to identify patterns, predict academic grade classification, and generate meaningful insights. The ultimate goal is to equip BrightPath Academy with a predictive model that simplifies research and supports data-driven decision-making, moving the school closer to its vision of empowering every learner.

---

2. Hypothesis Generation

Student's academic performance can be effectively predicted using a combination of demographic details, study patterns, parental involvement, and extracurricular involvement. It is expected that students who study more hours weekly, receive tutoring, and have strong parental support are more likely to achieve higher grades. In contrast, frequent absences are anticipated to negatively impact performance. Additionally, involvement in extracurricular activities is presumed to enhance academic outcomes. By uncovering these patterns, the model aims to help BrightPath Academy identify students who are at risk early and implement targeted academic interventions.

---

3. Getting the System Ready and Loading the Data

The python programming environment was set up using a Jupyter Notebook and libraries such as Pandas, NumPy, Matplotlib, and Seaborn, Scikit-learn were imported for data handling and visualization

The data set provided, Student_performance_data, was then loaded into the environment.

After successfully loading the dataset, the processes of data understanding, analysis, and predictive modelling can commence.

---

4. Understanding the Data

The dataset includes **demographic**, **academic**, **activity**, and **parental involvement** features for over 2,300 high school students.

Key columns:

- **Target**: GradeClass (A–F: encoded 0–4)
- **Features**: GPA, StudyTimeWeekly, Absences, ParentalSupport, Extracurricular, Sports, Music, Volunteering, etc.

---

5. Exploratory Data Analysis

i. Univariate Analysis

Visualizations:

- GPA distribution peaks between 3.0 and 3.5 (Grade B)
- Study time is right-skewed (majority <10 hrs/week)
- Most students have some level of parental support

ii. Bivariate Analysis

- GPA vs Parental Support: strong positive trend
- GPA vs Absences: clear negative trend
- GradeClass vs Extracurriculars: involvement increases likelihood of A/B grades

---

6. Missing Value and Outlier Treatment

- Missing values were minimal: filled using median for continuous variables and mode for categorical.
- Outliers in Absences and StudyTimeWeekly treated using IQR-based capping.

---

7. Evaluation Metrics for Classification Problem

- **Accuracy Score:** Measures the rate at which the model's predictions are correct overall. This helps to give a quick overall performance measure.
- **Confusion Matrix:** Shows the number of students who were correctly or wrongly predicted for each grade class. This matrix give detailed information on where the models fail or succeed, which is needed in identifying misclassified students who may need intervention.

---

8. Feature Engineering

Selected Features:

StudyTimeWeekly: Represents the amount of time in hours per week that a student spends studying.

GPA: An important measure of academic performance.

Data Scaling:

I used Standard Scaler to scale the numeric features before training the models. Because some models like Logistic Regression are affected by differences in feature scale. Standardizing the data (making sure it's on the same scale) ensures more stable and consistent performance. Because GPA is 0-4 and StudyTimeWeekly is 0-20.

---

9. Model Building: Part 1 – Baseline Classification Algorithms

1. Logistic Regression

This model calculates the probability of each class using a logistic function and chooses the class with the highest probability. Useful for identifying linear relationships between student variables and grade outcomes.

2. Random Forest

This model builds multiple decision trees on random subsets of data and combines their predictions through majority voting. Captures non-linear impacts such as varying impacts of study time or GPA for various students. This model handles complex patterns and is less prone to overfitting.

3. XGBoost

This model is extremely accurate. It trains decision trees sequentially, with each new tree trying to correct errors of the preceding ones using gradient descent. Effectively identifies patterns in poorly performing students over time, enabling early intervention.

---

10. Model Building: Part 2 – Deep Learning Classification

4. MLPClassifier (Multi-layer Perceptron)

Neural networks can learn from complex patterns and feature interactions. This model uses stacks of interconnected neurons to learn weights using a process called backpropagation. It captures deeper patterns in the performance of students.

---

11. Model Deployment – Dash App on Render

The final solution was deployed using **Dash** and hosted on **Render**.

Features of the App:

- Home Page
- Information page – Shows the dataset and statistical information
- Visualize (used to show comparisons between the data we used):
    - Scatter Plot
    - Histogram
    - Box plots

    - correlation matrix
    - Performance monitoring between algorithms
- Student-grade prediction using our best performance model (RandomForest)
- Feedback form

Example Visual:

Scatter Plot: shows how the outcome for the GPA and Study time
Correlation matrix: shows how strongly and in what direction different numerical features are linearly related to each other.

Conclusion & Impact

This project provides **BrightPath Academy** with a data-driven system to:

- **Proactively identify** at-risk students
- **Tailor interventions** based on metrics
- Enable **early warning dashboards** for educators

With ML-integrated insights and visual analytics, BrightPath can now **intervene earlier, act smarter, and support holistically**