Report Machine Learning Project

Predicting High school Student Performance

GitHub Repository: `https://github.com/lilian95520/ML-Predicting-high-school-student-performance`

Laura ABOUKRAT, Laura DELEUZE, Lilian ALLIO

DIA 1 - Group 4 - December 2025

Contents

1 Introduction **2**
- 1.1 Context . . . 2
- 1.2 Problem Statement . . . 2
- 1.3 Project Objectives . . . 2

2 Data presentation **3**
- 2.1 Data source . . . 3
- 2.2 Dataset structure . . . 3
- 2.3 Target variable . . . 3
- 2.4 Explanatory variables . . . 3
- 2.5 Exploratory Analysis . . . 4
  - 2.5.1 Distribution of the Target variable . . . 4
  - 2.5.2 Distribution of key variables . . . 5
  - 2.5.3 Relationship between key variables and final grades . . . 5
  - 2.5.4 Correlations between Explanatory variables and the Target . . . 8

3 Data preprocessing **9**
- 3.1 Data cleaning . . . 9
- 3.2 Encoding . . . 9
- 3.3 Creation of new features . . . 11
- 3.4 Feature selection . . . 11
  - 3.4.1 Feature Importance . . . 11
  - 3.4.2 Feature Selection . . . 12
- 3.5 Feature scaling . . . 13

4 Modelling and evaluation **13**
- 4.1 Implementation of a baseline model . . . 13
  - 4.1.1 Problem formulation and model selection . . . 13
  - 4.1.2 Model performance comparison . . . 13
  - 4.1.3 Overfitting/underfitting risk analysis . . . 14
- 4.2 Model's hyperparameter's tuning . . . 14
  - 4.2.1 Gradient Boosting Hyperparameter Tuning . . . 14
  - 4.2.2 SVM Hyperparameter Tuning . . . 15
  - 4.2.3 Random Forest Hyperparameter Tuning . . . 15
  - 4.2.4 Logistic Regression Hyperparameter Tuning . . . 15

5 Ensemble Model **16**

6 To go further: test yourself ! **16**

7 Difficulties and limitations **17**
- 7.1 Dataset selection and quality . . . 17
- 7.2 Data availability and target definition . . . 17
- 7.3 Model and methodological limitations . . . 17
- 7.4 Future improvements . . . 17

8 Conclusion **18**

1 Introduction

1.1 Context

High school education faces significant challenges worldwide, particularly in terms of student dropout and academic underperformance. In Europe, around 9.3 % of young people (aged 18-24) left education early in 2024, without any diploma meaning they had not completed high school[1]. In France, the rate of students leaving education before completing secondary education was about 7.7 % in 2024 [2]. In the United States, according to the National Center for Education Statistics, the high school dropout rate was around 5.3 % in 2022 for 16-24 year olds who were not enrolled in school [3].

These numbers highlight the existence of strong disparities in academic success among students.

Academic performance is influenced by a wide range of factors, including demographic characteristics, academic background, social environment and lifestyle habits. However, identifying the precise origin of these performance gaps remains complex, making it difficult for educational institutions to intervene effectively and at an early stage.

1.2 Problem Statement

The main challenge addressed in this project is the early identification of students at risk of academic underperformance. Educational institutions often lack effective tools to anticipate a decline in student performance before it leads to failure or dropout.

This project seeks to answer the following key questions:

- Which factors have the strongest influence on high school students' academic success?
- Is it possible to accurately predict a student's final academic level before academic failure occurs?

Addressing these questions is crucial in order to move from a reactive to a proactive approach in education, allowing educators to detect, prevent and intervene earlier through targeted actions.

1.3 Project Objectives

The main objective of this project is to develop a machine learning model capable of predicting high school students' final grades, categorized into three classes (A, B or C).

More specifically, the project aims to:

- Identify the key determinants of academic success
- Provide personalized predictions of students' final performance
- Propose an interactive tool capable of simulating the impact of behavioral changes on academic outcomes, for example by evaluating how reduced absences or increased study time may affect final grades.

Such a predictive approach can serve as a decision-support tool for educators and academic advisors, allowing early identification of students at risk and providing targeted academic or psychological support.

Through these objectives, the project seeks to demonstrate how machine learning can support early intervention strategies and contribute to more personalized and data-driven educational support.

2 Data presentation

2.1 Data source

The dataset used for this project was found on Kaggle and is called Students Performance Dataset[4]. It contains academic, demographic, social and lifestyle information collected to analyze the factors influencing students' academic performance.

2.2 Dataset structure

The dataset contains 2392 student records and 14 variables. Each row corresponds to an individual student, while each column represents a specific attribute related to the student's profile or academic background. The variables include both numerical features (such as age, number of absences and study time) and categorical features (such as gender, family support and parental education).

2.3 Target variable

The target variable of this study is **GradeClass**, which represents the final grade obtained by students. Initially, this variable consisted of five categories (A, B, C, D, and E) created from the variable **GPA**. However, this configuration did not suit us because, with only 2,392 rows and 14 columns, having five different categories for the target variable would have divided the data too much, leading to excessive fragmentation. This overly fine division of classes would have reduced the number of examples per category, making model learning less reliable and increasing the risk of overfitting. In order to obtain more balanced classes and more robust predictions, we therefore chose to group the initial categories into three final classes (A, B, and C).

These marks were calculated from the GPA as follows:

- A: GPA >= 3
- B: 2 <= GPA < 3
- C: 0 <= GPA < 2

This categorization allows the problem to be treated as a supervised multi-class classification task.

2.4 Explanatory variables

The explanatory variables used in this study describe several dimensions of students' profiles and are grouped into 4 main categories: demographic, academic, social and lifestyle factors.

Demographic variables include basic information such as gender and family background. These variables provide contextual information that may indirectly influence academic outcomes.

Academic variables are directly related to students' school performance and behavior, such as the number of absences and study time. These features are expected to have a strong impact on final academic results.

Social variables capture the students' environment, including parental support, family relationships and access to educational resources. A supportive social context can positively affect motivation and engagement at school.

Lifestyle variables reflect students' daily habits, such as free time activities and sports. These factors may influence concentration, fatigue and overall academic performance.

Together, these explanatory variables allow the model to capture both academic and non-academic influences on student performance, providing a comprehensive representation of the factors that contribute to final grades.

2.5 Exploratory Analysis

2.5.1 Distribution of the Target variable

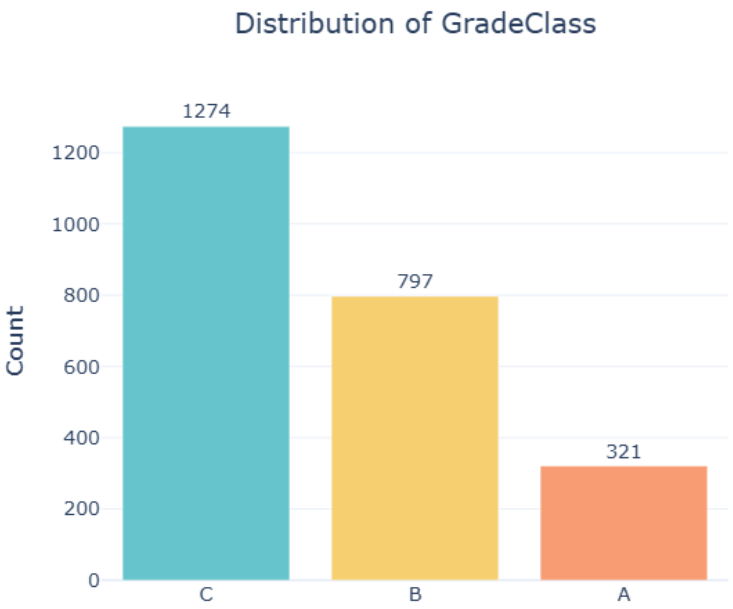


Figure 1: Distribution of the target GradeClass

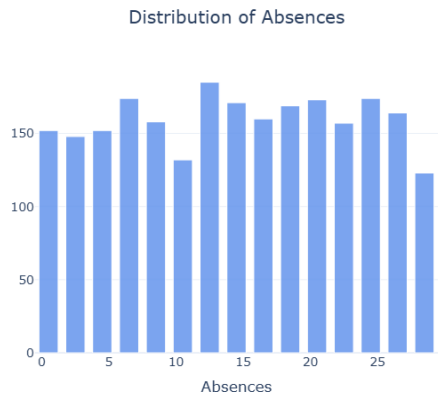
The distribution of GradeClass shows a **class imbalance.**

Class C is the most represented with 1274 observations, followed by class B (797 students), while class A is the least frequent with only 321 observations.

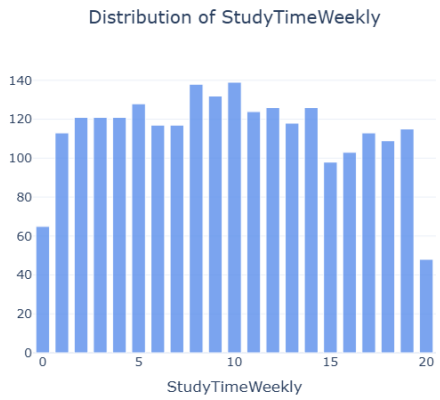
This imbalance indicates that low-performing students are more frequent in the dataset, whereas high-performing students are underrepresented.

Such an imbalance may bias the models toward the majority class (C) and negatively impact the prediction performance for class A. Therefore, we will test certain models with the parameter `class_weight="balanced"` which automatically assigns higher weights to minority classes and lower weights to majority classes during model training. If this parameter significantly improves the macro f1-score, we will keep it.

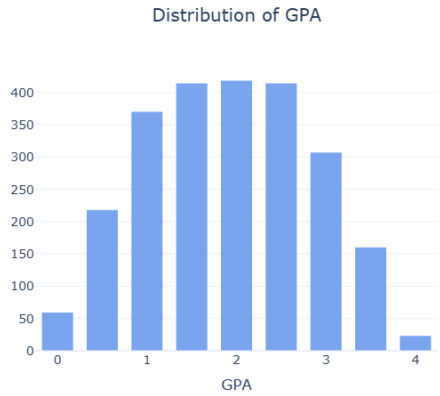
2.5.2 Distribution of key variables



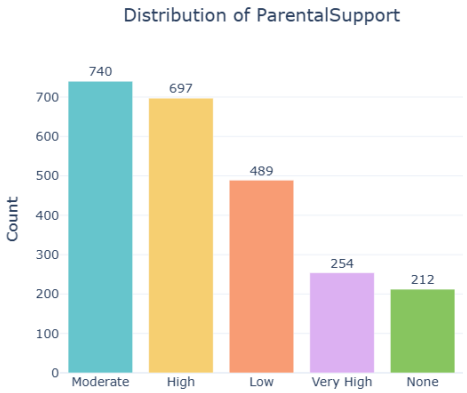
(a) Distribution of Absences



(b) Distribution of Study Time Weekly



(c) Distribution of GPA



(d) Distribution of ParentalSupport

Figure 2: Distribution of key variables

The distributions of the main variables provide useful insights into students' academic behavior.

The number of absences is relatively spread across the observed range, with no strong concentration around a single value. This suggests heterogeneous attendance patterns among students.

The weekly study time is mainly concentrated between moderate values, while very low and very high study times are less frequent. This indicates that most students dedicate a reasonable amount of time to studying.

The GPA distribution is centered around intermediate values and follows a normal law, with fewer students at the extremes.

Parental support is mostly reported as moderate or high, whereas very high or no support is less common. So most students benefit from some level of family involvement, which could play a role in their academic success.

2.5.3 Relationship between key variables and final grades

Average absences by grade

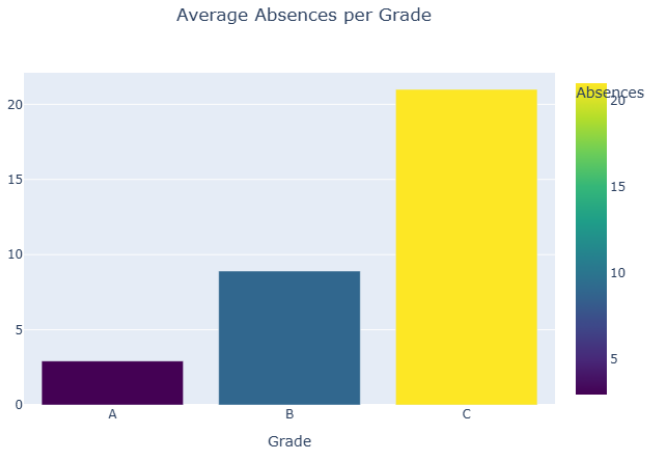


Figure 3: Average absences per grade

The graph confirms a clear relationship between academic level and the average number of absences.

Low-level students are the most absent, with more than 21 absences on average, highlighting a strong link between absenteeism and poor performance.

Medium-level students show fewer absences (around 9), placing them between struggling and high-performing students.

High-level students are the least absent, with an average of about 3 absences, supporting the idea that regular attendance contributes to better results.

So the graph shows that the higher the academic level, the fewer absences students have. Absenteeism appears to be one of the strongest indicators of academic performance.

Average Study Time Weekly per grade

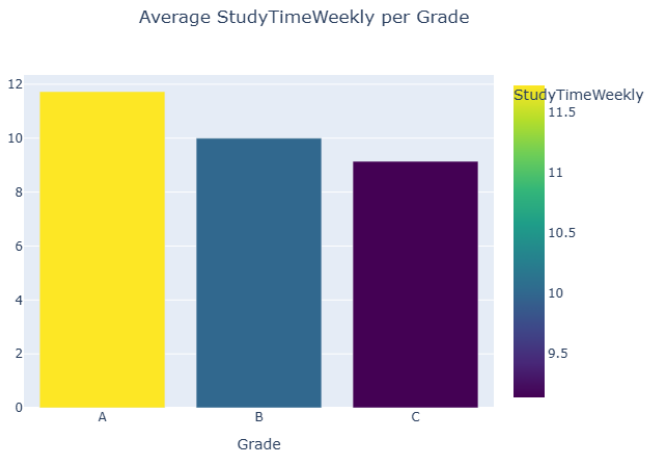


Figure 4: Average StudyTimeWeekly per grade

The graph confirms a relationship between academic level and the study time weekly.

Students with a High grade spend on average slightly more time studying each week than those with a Medium or Low grade.

The Low category studies the least, while High has the highest weekly study time, suggesting a positive relationship between study time and performance level.

GPA based on ParentalSupport

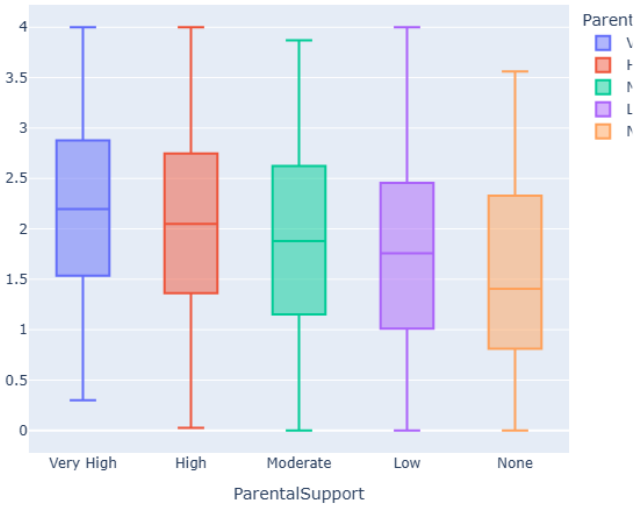


Figure 5: Boxplot: GPA based on ParentalSupport

There is a general trend: **the higher the parental support, the higher the GPA tends to be.**

Students with Very High support have the highest GPA medians and a higher overall distribution, while the Low and None categories show lower medians.

This suggests that parental support plays a positive role in academic success.

GPA based on Tutoring

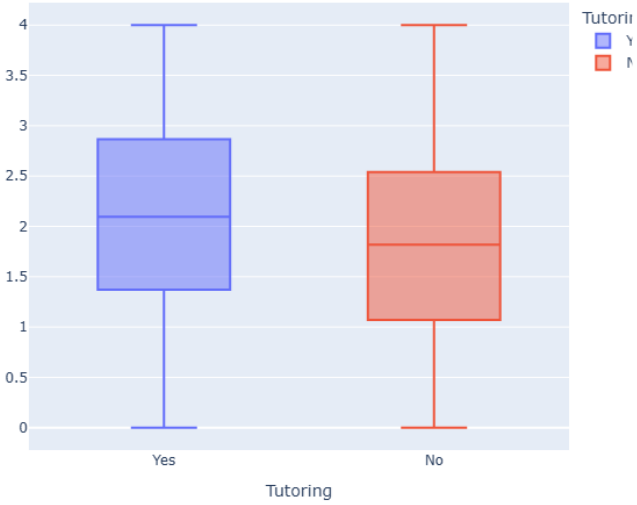


Figure 6: Boxplot: GPA based on Tutoring

Students receiving tutoring have a **slightly higher GPA on average**, with their median score above that of students without tutoring. However, the difference remains moderate, and the variability of the two groups is fairly similar.

Tutoring therefore appears to have a positive but limited effect on academic performance.

2.5.4 Correlations between Explanatory variables and the Target

Correlation Matrix

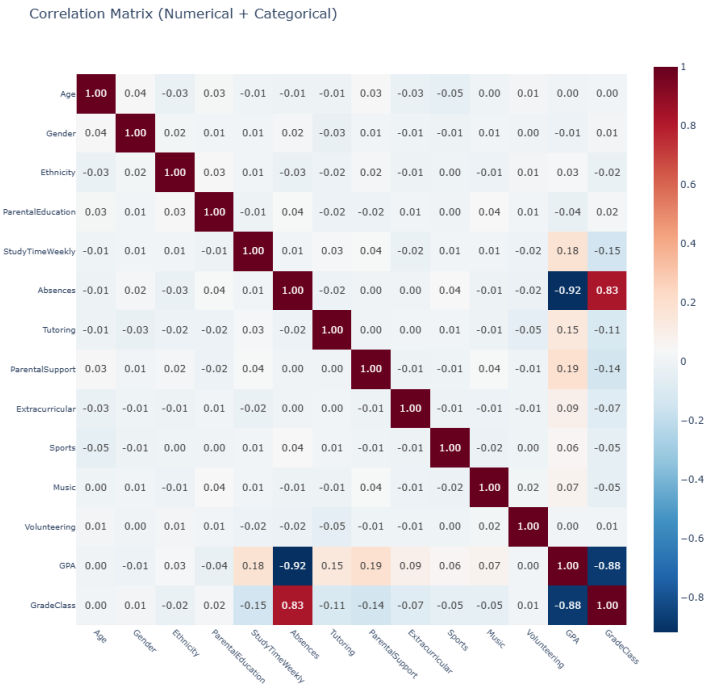


Figure 7: Correlation Matrix

Main correlations with GPA

- **Absences → GPA** (−0.92)
  Extremely strong negative correlation: absenteeism is by far the most decisive factor.
- **StudyTimeWeekly → GPA** (+0.18)
  Moderate positive effect: studying more helps, but remains secondary.
- **Tutoring → GPA** (+0.15)
  Positive impact, consistent with the box plot results.
- **ParentalSupport → GPA** (+0.19)
  Weak but positive influence.
- **Volunteering → GPA** (+0.09)
  Minor positive effect, possibly reflecting more organized students.

Relationships between explanatory variables

- **StudyTimeWeekly and Absences** (−0.15)
  Students who study more tend to be slightly less absent.
- **Extracurricular activities (sports, music, etc.)**
  Correlations close to zero, indicating no significant impact on GPA.

Demographic variables

Age, gender, ethnicity and parental education show very weak or no correlations with GPA.

Conclusion

The most influential factors are:

- Absences
- StudyTimeWeekly
- Tutoring
- ParentalSupport
- Volunteering

Note: Correlations with the target variable *GradeClass* cannot be interpreted reliably due to its reversed encoding (0 = A, 2 = C). This leads to misleading correlation signs, such as a positive correlation between absences and grade level, which is not meaningful.

3 Data preprocessing

3.1 Data cleaning

Checking for missing values: We had no missing value in our dataset.

Checking for duplicas: We had no duplicas in our dataset.

3.2 Encoding

In our dataset, all categorical variables were already encoded. When creating the target GradeClass, we also encoded this variable.

Here are the encoding-label correspondences:

- **Gender**
  - 0 → Male
  - 1 → Female
- **Ethnicity**
  - 0 → Caucasian
  - 1 → African American
  - 2 → Asian
  - 3 → Other
- **ParentalEducation**
  - 0 → None
  - 1 → High School
  - 2 → Some College
  - 3 → Bachelor's degree
  - 4 → Higher education
- **Tutoring**
  - 0 → No
  - 1 → Yes
- **ParentalSupport**
  - 0 → None
  - 1 → Low
  - 2 → Moderate
  - 3 → High
  - 4 → Very High
- **Extracurricular activities, Sports, Music, Volunteering**
  - 0 → No
  - 1 → Yes
- **GradeClass** (target variable)
  - 0 → A
  - 1 → B
  - 2 → C

It is important to note that the encoding of the target variable is reversed with respect to academic performance, since 0 correspond to better grades. This choice affects the interpretation of correlation analyses.

3.3 Creation of new features

In order to capture more complex patterns in students' behavior, additional features were engineered from existing variables.

StudySupportInteraction : A new feature was created as follows:

$$\text{StudySupportInteraction} = \text{StudyTimeWeekly} \times \text{ParentalSupport}$$

This feature captures the combined effect of personal study effort and the level of support received at home. While study time and parental support individually influence academic performance, their interaction highlights situations where these two factors reinforce each other.

- **High values** correspond to students who study regularly and benefit from strong parental support, indicating a highly favorable learning environment.
- **Low values** indicate limited study time and/or weak support, suggesting a higher risk of underperformance.

This interaction variable allows us to assess whether parental support amplifies the positive effect of studying.

EngagementScore : An overall engagement indicator was constructed as:

$$\text{EngagementScore} = \text{StudyTimeWeekly} + \text{Extracurricular} + \text{Sports} + \text{Music} + \text{Volunteering}$$

This feature measures the global level of student engagement by combining academic investment with participation in extracurricular activities. It reflects motivation, time management skills and involvement both inside and outside the classroom.

- **High values** correspond to highly engaged students who study regularly and participate in several activities, suggesting a balanced and motivated profile.
- **Low values** indicate low study time and limited involvement, which may reflect disengagement or lack of structure.

This score helps identify students who are strongly invested in their academic and personal development.

Rejected engineered features Several additional engineered features were also tested. However, some of them were strongly derived from existing variables. These features did not introduce new information and mainly reinforced patterns already present in the dataset.

As a result, they increased redundancy and risked adding noise to the models without improving predictive performance. For this reason, only the most informative and interpretable engineered features were retained.

3.4 Feature selection

3.4.1 Feature Importance

We first tested the importance of each feature on the predictive power of a Random Forest.

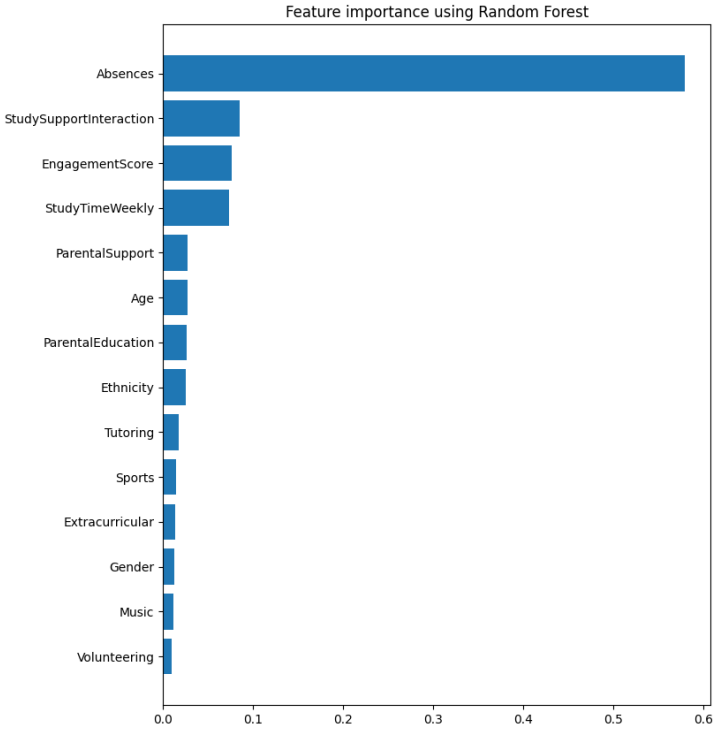


Figure 8: Features Importance

Dominant variable

Absences (39.4%) is by far the most influential feature for predicting GradeClass. This confirms previous observations: attendance is the strongest indicator of academic success. Higher absenteeism is strongly associated with lower academic performance.

Secondary variables

Several variables have a moderate but meaningful impact:

- **StudySupportInteraction (6.4%)**: students who study more and receive strong parental support tend to perform better.
- **StudyTimeWeekly (5.9%)**: individual effort remains an important factor.
- **EngagementScore (5.7%)**: overall engagement shows a positive but limited influence.

Low-impact variables

Other features have limited importance (between 1% and 3%), including *ParentalSupport*, *ParentalEducation*, *Age*, *Tutoring* and extracurricular activities.

So direct academic behavior (attendance and study effort) clearly outweighs socio-cultural and extracurricular factors in predicting student performance.

3.4.2 Feature Selection

The aim of the project is to predict a student's final grade by taking into account all of that student's characteristics and assessing which ones are most important. We have therefore chosen to keep all explanatory variables except **GradeClass**, obviously, and **GPA**, as the target is derived from GPA, which would constitute a data leak.

3.5 Feature scaling

We standardised the numerical variables using `StandardScaler` where necessary, particularly for KNN, SVM and Logistic Regression, as these models are sensitive to distances.

Without standardisation, variables with larger numerical ranges (such as absences or study time) would dominate the distance or optimization calculations, leading to biased predictions.

4 Modelling and evaluation

4.1 Implementation of a baseline model

4.1.1 Problem formulation and model selection

Predicting *GradeClass* is a **multi-class classification problem**. The target variable represents 3 academic performance categories (A, B and C, encoded as 0, 1, 2). Since the goal is to assign each student to one of these discrete classes, classification models are required.

Several models were considered:

- **Logistic Regression**: simple and interpretable, used as a baseline model.
- **Random Forest**: robust to overfitting, handles non-linearity and provides feature importance.
- **Gradient Boosting/XGBoost**: high-performance models for complex relationships.
- **SVM**: effective for well-separated classes and high-dimensional data.
- **K-Nearest Neighbors**: distance-based method used mainly for comparison.

4.1.2 Model performance comparison

Most models achieve strong performance, with test accuracies close to **0.89**, while KNN clearly underperforms.

Logistic Regression achieves the best overall accuracy (**0.894**). It shows balanced precision and recall across all classes, particularly for class C (F1-score = **0.94**), and handles class A better than most models. It is a strong and reliable baseline.

Random Forest and **SVM** obtain identical accuracies (**0.889**). Random Forest provides robust predictions, but shows weaker recall for class A. SVM performs slightly better on class C, but remains sensitive to class imbalance.

Gradient Boosting and **XGBoost** achieve slightly lower accuracies (around **0.87-0.88**). While they capture non-linear relationships, their performance does not surpass simpler models on this dataset.

In contrast, **KNN** performs poorly (accuracy **0.73**), especially for class A, confirming that distance-based methods are not well suited to this problem.

To conclude, these results indicate that **Logistic Regression offers the best trade-off between accuracy, stability and interpretability**, while Random Forest and SVM remain strong alternatives.

Model ranking

- **Logistic Regression** — Accuracy = 0.89
- **Random Forest** — Accuracy = 0.88

- **SVM** — Accuracy = 0.87
- **Gradient Boosting/XGBoost** — Accuracy = 0.85
- **KNN** — Accuracy = 0.83

Choice of class weighting

We tested the option `class_weight="balanced"` for models that support it. However, we decided not to keep it in the final models.

Although class balancing can be useful when classes are highly imbalanced, in our case the class distribution was not extreme. Using balanced weights tended to over-penalize misclassifications of the minority class, which led to a decrease in overall accuracy. It did not lead to a significant improvement in the macro F1-score.

Since our main objective was to maximize predictive performance measured by accuracy, we obtained better and more stable results without class weighting.

4.1.3 Overfitting/underfitting risk analysis

To assess model generalization, we compared training and testing accuracies. A large difference indicates overfitting, while low performance suggests underfitting.

- **Logistic Regression** shows almost identical train and test accuracies (difference = 0.017), indicating excellent generalization.
- **Gradient Boosting** presents a moderate gap (0.080), but remains acceptable with no strong overfitting.
- **SVM** shows a small difference (0.054), corresponding to slight but acceptable overfitting. This is normal for SVM models, as they slightly adapt to the training data in order to learn a good decision boundary.
- **Random Forest** and **XGBoost** achieve perfect training accuracy but lower test accuracy, revealing significant overfitting.
- **KNN** has low training accuracy and a large gap, indicating underfitting and poor generalization.

Based on these results, we retain **Logistic Regression, SVM, Gradient Boosting and Random Forest** for further hyperparameter tuning.

4.2 Model's hyperparameter's tuning

4.2.1 Gradient Boosting Hyperparameter Tuning

Using `RandomizedSearchCV`, the best hyperparameters obtained are:

- `n_estimators = 200`
- `max_depth = 2`
- `learning_rate = 0.1`
- `subsample = 1.0`

The best cross-validation score achieved is **0.8902**. The low tree depth limits model complexity and reduces overfitting, while the chosen learning rate ensures stable training.

Why RandomizedSearchCV? It is faster than GridSearch, explores the hyperparameter space efficiently, and is well suited for computationally expensive models such as Gradient Boosting.

4.2.2 SVM Hyperparameter Tuning

Using `GridSearchCV`, the best hyperparameters obtained are:

- `C = 1`
- `kernel = rbf`
- `gamma = scale`

The best cross-validation score achieved is **0.8787**. This configuration provides a good trade-off between model complexity and generalization. The RBF kernel captures non-linear relationships, while `C=1` prevents overfitting by limiting sensitivity to noise.

Why GridSearchCV? It performs an exhaustive search over the parameter grid and is efficient for SVM models, which are relatively fast to train.

4.2.3 Random Forest Hyperparameter Tuning

After hyperparameter tuning, the Random Forest model shows a much better balance between training and test performance.

The best parameters are:

- max_depth = 10
- min_samples_split = 10
- min_samples_leaf = 5
- n_estimators = 300

These parameters reduce model complexity and limit overfitting:

- Limiting the tree depth prevents overly complex trees
- Increasing the minimum number of samples per split and per leaf forces the model to learn more general patterns

The gap between training accuracy (0.94) and test accuracy (0.89) is now small (0.053), whereas the untuned model showed significant overfitting. This indicates that the model generalizes much better after tuning.

4.2.4 Logistic Regression Hyperparameter Tuning

After hyperparameter tuning using `GridSearchCV`, Logistic Regression achieved the best cross-validation score among all tested models (**CV accuracy = 0.9043**).

Best parameters:

- `C = 10`: weaker regularization, allowing a more flexible decision boundary
- `penalty = l2`: stabilizes the model by preventing large coefficients
- `solver = lbfgs`: efficient and well suited for multiclass classification with L2 regularization

Logistic Regression remains simple but highly effective. It shows excellent generalization, with a good bias-variance trade-off, suggesting that a linear model fits the structure of the dataset very well.

5 Ensemble Model

To further improve performance, we combined the best individual models into an ensemble using a **soft voting**. The ensemble includes:

- Logistic Regression
- Random Forest
- SVM

Soft voting aggregates the predicted class probabilities of each model and selects the class with the highest weighted average proba. Higher weight was assigned to Logistic Regression, as it achieved the best individual performance.

Ensemble performance:

- Test accuracy: **0.898**
- Train accuracy: 0.937
- Generalization gap: 0.039

The confusion matrix shows balanced and strong performance across all 3 classes, with particularly high recall for class C, which represents students with the lowest academic performance. This indicates that the ensemble model is especially effective at identifying students at risk of underperformance.

Compared to individual models, the ensemble slightly improves overall accuracy:

- Logistic Regression: 0.891
- SVM: 0.889
- Random Forest: 0.885

Therefore, the ensemble model achieves the **best predictive performance** while maintaining good generalization and a limited overfitting risk. This confirms that combining complementary models helps reduce individual weaknesses and improve robustness.

6 To go further: test yourself !

To meet the needs of educational institutions in predicting students at risk of academic failure, we developed a simulation tool. By entering key student information (such as study time, absences, parental support), students or educational staff can predict the student's expected grade and adjust their strategy if the student is at risk of failure.

In addition to predicting the grade, the tool also highlights the factors that had the greatest impact on the prediction. This allows students to better understand their situation and take action, for example by reducing absences if they are identified as a major risk factor.

7 Difficulties and limitations

7.1 Dataset selection and quality

One of the main difficulties at the beginning of the project was the choice of the dataset. We initially selected a dataset related to students' performance in college, but it had been **artificially generated**. As a result, all variables were almost uniformly distributed, which strongly limited the learning capacity of the models.

No meaningful patterns could be extracted, and the accuracy never exceeded **0.34**, regardless of the model used. This highlighted the importance of using a dataset that reflects realistic relationships between variables.

Consequently, we decided to **change the dataset** and adapt our entire pipeline (preprocessing, feature engineering and modeling) to a new data source.

7.2 Data availability and target definition

Finding a suitable dataset with **enough observations and a sufficient number of relevant variables** was also challenging. Many datasets were either too small or lacked important academic, social or behavioral features.

For this reason, we chose the current dataset and made the decision to **reduce the dimensionality of the target variable**. Originally, the grades were more finely detailed, but due to the limited number of samples per class, this led to poor performance and unstable models. Grouping the target into three classes (A, B, C) allowed us to obtain more reliable and robust results.

7.3 Model and methodological limitations

Despite good overall performance, several limitations remain:

- The student performance is observed at a single point in time. Temporal effects and long-term trends cannot be captured
- Some variables rely on **self-reported data** (study time, extracurricular activities), which may introduce bias or measurement errors
- Correlation-based analysis does not imply causality: for example, high absenteeism is strongly associated with low performance, but it may also be a consequence rather than a cause
- The models were optimized mainly for accuracy; other objectives such as fairness or interpretability could be further explored

7.4 Future improvements

With more data, several improvements could be considered:

- Using a larger dataset to keep a more detailed grading scale
- Incorporating longitudinal data to model performance evolution over time
- Exploring more advanced ensemble or neural models if additional data becomes available

8 Conclusion

This project focused on predicting high school students' academic performance by classifying final grades into three categories (A, B, C). Using academic, demographic and behavioral data, we aimed to identify key factors influencing student success and to build effective predictive models.

The analysis showed that academic behavior, especially absenteeism, is the most important factor affecting performance. Study time, parental support and tutoring also contribute positively, while demographic and extracurricular variables have a more limited impact.

Several classification models were evaluated. Logistic Regression provided the best trade-off between accuracy, stability and interpretability, while Random Forest and SVM also achieved strong results after tuning. The ensemble model slightly improved performance and proved particularly effective at identifying students at risk of underperformance.

To conclude, this project highlights the potential of machine learning as a decision-support tool for education. Despite some data limitations, the results are promising and could be further improved with larger and longitudinal datasets.

References

[1] Eurostat, *Early leavers from education and training - Europe*, European Commission, 2024. Available at: https://ec.europa.eu/eurostat

[2] Eurostat, *Early leavers from education and training - France*, European Commission, 2024. Available at: https://ec.europa.eu/eurostat

[3] National Center for Education Statistics, *Status dropout rates*, U.S. Department of Education, 2022. Available at: `https://nces.ed.gov`

[4] Kaggle, *Students Performance Dataset*, Available at: `https://www.kaggle.com/datasets/rabieelkharoua/students-performance-dataset`