Student Performance Prediction

Student Performance Prediction

ELALAOUI Oumaima

ELBATTAH Ahmed

MSc Intelligent Processing Systems

Course : Data mining

Lecturer : Pr Anouar Riad Solh

Submitted : 23/02/2025

Project Name:

Student Performance Prediction.

Objective:

📌 This project aims to analyze student evaluation data to predict their future academic performance. The goal is to understand the factors that influence academic success and anticipate student performance in future exams based on multiple variables.

---

1. Project Description :

◆ Context and Problem Statement:

Educational institutions often seek to predict student performance to provide better support. For example, a student may have strong academic potential, but their success depends on factors such as motivation, engagement, and class attendance. By utilizing data mining algorithms such as **Apriori, K-means, and K-NN**, this project aims to uncover relationships between these factors and develop a reliable predictive model.

◆ Data Used:

The project relies on a dataset containing information such as:

- Student exam scores.
- Number of study hours.
- Student motivation levels.
- Class attendance levels.
- Parental involvement.

◆ Approach and Methodology:

To achieve this objective, several data mining techniques are applied:

- **Association Rules (Apriori):** Identifies relationships between different characteristics (e.g., attendance and exam scores).
- **Clustering (K-means):** Segments students into groups with similar performance or behavioral patterns.
- **Classification (K-NN):** Uses student characteristics to predict whether they will achieve high or low exam performance.

◆ Importance and Application:

This project can help educational institutions better understand the factors influencing student success. It can also be used to identify at-risk students and provide targeted interventions, such as revision sessions or personalized support, to improve their academic outcomes

2. Application Domain :

📌 The **"Student Performance Prediction"** project falls within the field of education, specifically in the analysis and prediction of student academic performance. The main objective is to leverage **data mining** and **machine learning techniques** to better understand the factors influencing student success and anticipate their future results.

◆ Benefits for Educational Institutions :

Educational institutions, such as schools, colleges, and universities, can utilize predictive analytics to:

- **Identify at-risk students:** Detect students who may struggle academically and provide personalized support.
- **Optimize teaching strategies:** Adapt educational methods based on the specific needs of students or student groups.
- **Improve success rates:** Implement preventive measures (e.g., tutoring, remedial courses, personalized guidance) to reduce failure and dropout rates.
- **Allocate resources efficiently:** Institutions can better organize their human and material resources by targeting students who require additional support.

◆ Impact on Teachers and Students :

- **Personalized teaching support:** Teachers can adjust their pedagogy based on identified student profiles.
- **Individualized monitoring and guidance:** Students receiving performance predictions can be encouraged to adopt better learning strategies.
- **Increased motivation:** By understanding the factors influencing their results, students can be more motivated to improve their performance by adopting appropriate methods.

◆ Application for Educational Decision- Makers :

- **Data-driven educational policies:** By analyzing general trends, education policymakers can propose reforms or adjustments in the educational system.
- **Evaluation of educational programs:** Assess whether certain educational initiatives have a real impact on student performance by comparing predictions with actual results.

◆ Real-World Use Cases :

- A **university** can identify first-year students most at risk of failing and provide them with early mentoring programs.
- A **high school** can analyze the impact of parental education level and academic monitoring on student results.
- A **government** can use predictive models to detect underperforming schools and implement corrective measures.

3. Technologies and Methodologies Used

📌 To implement this **student performance prediction** project, various **technologies** and **methodologies** have been integrated to ensure efficient analysis and reliable forecasting of student results. These technologies are well-suited for handling complex data and executing the different stages of the prediction process. Below is an overview of the tools and techniques used:

◆ Programming Language: Python :

Python was chosen for this project due to its **flexibility, extensive libraries**, and **active data science community**. It allows for efficient **data manipulation, machine learning algorithm application, and result visualization**. The key libraries used include:

- **Pandas**: For data manipulation and analysis. Pandas is ideal for **cleaning, transforming, and exploring** data in a DataFrame format.
- **NumPy**: For numerical computations and matrix operations.
- **Matplotlib & Seaborn**: For **data visualization** to better understand trends and patterns.
- **Scikit-learn**: For implementing **machine learning models** such as **K-means, K-NN**, and **Apriori**, as well as data **normalization**.
- **MLxtend**: For applying the **Apriori algorithm**, used for discovering **association rules** within the dataset.
- **NetworkX**: For **visualizing association rules** as directed graphs.

1. Development Environment :

- **Google Colab**: A **cloud-based** environment for running Jupyter notebooks, ideal for **collaborative work** and leveraging **GPU resources** for model training.
- **Google Drive**: Used to **store datasets** and **save project results**.

2. Machine Learning Frameworks :

- **Scikit-learn**: Used extensively for **data preprocessing (normalization, encoding)**, as well as **model creation, training, and evaluation**.
- **MLxtend**: Utilized for **association rule mining (Apriori algorithm)** to analyze relationships between different student characteristics.

◆ Methodologies Used :

1. Data Preprocessing :

Data preprocessing is a crucial step before applying any **machine learning model**. In this project, the main preprocessing steps included:

- **Loading and Cleaning Data**: The dataset was loaded from **Google Drive**, and cleaned by **removing duplicates, filling missing values** (mean for numerical variables, mode for categorical variables), and handling **outliers** using the **IQR (Interquartile Range) method**.
- **Encoding Categorical Variables**: Categorical variables were converted into **numerical values** using **LabelEncoder**, which is essential for machine learning algorithms.
- **Data Normalization**: Before splitting the dataset into **training and testing sets**, the data was normalized using **StandardScaler** to ensure better convergence and more accurate results.

2. Clustering with K-means :

The **K-means algorithm** was used to perform **student clustering** based on their performance. The following steps were conducted:

- **Dimensionality Reduction with PCA**:
  - To facilitate **visualization and cluster analysis**, **Principal Component Analysis (PCA)** was applied to project the data into a **2D space**.
- **Finding the Optimal Number of Clusters**:
  - The **Silhouette Score** was used to determine the best **K value** (ranging from 2 to 6), with the highest score selected as the optimal **number of clusters**.
- **Cluster Visualization**:
  - Once the **K-means model** was trained, the clusters were visualized in a **2D plot**, where data points were color-coded based on their cluster, and **centroids were highlighted in red**.

3. Association Rule Mining with Apriori :

The **Apriori algorithm** was used to extract **association rules** from student data. This method helps in understanding the relationships between different variables, such as how **parental involvement** or **study habits** impact student performance. The steps included:

- **Variable Transformation**:
  - Continuous variables were **binarized** (e.g., "Studies a lot" and "High attendance") to make them suitable for **Apriori**.
- **Generating Frequent Itemsets**:
  - The **Apriori algorithm** was applied to generate **frequent itemsets** with a minimum **support threshold of 0.05**.
- **Extracting Association Rules**:
  - The most significant **association rules** were extracted using the **lift metric**, and the **top 10 most relevant rules** were selected for further analysis.
- **Rule Visualization**:
  - The **association rules** were visualized using **NetworkX**, displaying them as **directed graphs** for easier interpretation.

4. Classification with K-NN :

Finally, to predict **student score categories** (e.g., **"High" vs. "Low"**), the **K-Nearest Neighbors (K-NN)** algorithm was applied. The steps involved:

- **Creating the Target Variable**:
  - A **new column** was created to classify students based on their **exam scores**:
    - **"High"** if the score is **≥ 67**
    - **"Low"** otherwise
- **Splitting the Dataset**:
  - The dataset was divided into **independent variables (features)** and **dependent variables (score category)**.
- **Normalization & Train-Test Split**:
  - As with **K-means**, data was **normalized** before being **split into training and testing sets**.
- **Training and Evaluating the Model**:
  - The **K-NN model** was trained on the **training set**, and its performance was evaluated using:
    - **Accuracy**
    - **Confusion Matrix**
    - **Classification Report**

4. Project Process :

📌 The "Student Performance Prediction" project was structured into several key phases, each playing a crucial role in predicting student performance based on their assessment data. These steps range from data collection and preprocessing to results analysis and machine learning model implementation. Below is a detailed breakdown of the steps followed in this project:

🔹 Data Collection :

The first step of this project was to obtain a dataset on student performance. This dataset is essential for analysis and modeling. The data was collected from publicly available sources, such as the **"Student Performance Data" dataset from the UCI Machine Learning Repository**, which contains information such as:

- Student scores in various subjects (Mathematics, Portuguese).
- Socio-demographic characteristics of students (age, gender, family status).
- Behavioral information (absences, study time, etc.).

This data is crucial for understanding the variables influencing student success and identifying patterns or trends.

🔹 Data Preprocessing :

Once the data was collected, it was preprocessed to eliminate errors, missing values, and inconsistencies. Preprocessing is a critical step to ensure data quality and suitability for machine learning. Below are the sub-steps of this process:

- **Data Cleaning:**
  - Removal of duplicates to avoid biases in analysis.
  - Identification and treatment of missing values (filled with the mean for numerical variables and the mode for categorical variables).
  - Detection and handling of outliers using the **Interquartile Range (IQR)** method.
- **Variable Transformation:**
  - **Categorical Variable Encoding:** Encoding techniques like **LabelEncoder** were used to transform categorical variables (e.g., "Gender," "Results") into numerical values that machine learning models can interpret.
  - **Data Normalization:** Data normalization (using **StandardScaler**) ensured that all variables were on the same scale, facilitating model training and convergence.

🔹 Data Exploration and Visualization :

The next step involved exploring the data to understand its characteristics and identify trends. This was achieved through various visualization techniques, such as:

- **Histograms and Bar Charts:** To analyze the distribution of scores and other variables.
- **Correlation Matrix:** To examine relationships between variables, such as the impact of study hours or absenteeism on student performance.
- **Boxplots:** To visualize score distribution and detect outliers.

These visualizations provided valuable insights into the structure of the data and guided the subsequent stages of the project.

🔻 Clustering with K-means :

One of the key steps of the project was using the **K-means algorithm** to segment students into groups (clusters) based on their characteristics. Clustering helps identify groups of students with similar behaviors or performance levels.

- **Choosing the Number of Clusters:**
  - The optimal number of clusters was determined using the **Elbow Method** and **Silhouette Score analysis**.
- **Executing K-means:**
  - The data was divided into several clusters, and the characteristics of each group were analyzed.
  - **Dimensionality reduction** was performed using **Principal Component Analysis (PCA)** to facilitate cluster visualization.

The clusters helped identify student groups with similar performance levels, providing insights into how students behave based on their academic results.

🔻 Association Rule Analysis with Apriori :

The **Apriori algorithm** was used to discover interesting relationships between different student characteristics. Association rules help understand how different variables interact and influence performance.

- **Variable Transformation:**
  - Continuous variables were transformed into **binary categories** to apply the **Apriori algorithm**.
- **Rule Extraction:**
  - The **Apriori algorithm** extracted association rules based on criteria such as the relationship between study hours and student scores.
  - The rules were visualized using **graph representations** for easier interpretation.

🔻 Classification with K-NN :

After exploring the data and identifying patterns, a classification model was used to predict student performance. The **K-Nearest Neighbors (K-NN) algorithm** was chosen due to its simplicity and effectiveness for categorical data.

- **Creating the Target Variable:**
  - A target variable was defined to categorize students based on their score (e.g., **"High"** for a score ≥ **67**, **"Low"** for a lower score).
- **Training the K-NN Model:**
  - The **K-NN model** was trained on the dataset, and the optimal number of neighbors (**k**) was determined to maximize accuracy.
- **Model Evaluation:**
  - The model was evaluated using the **confusion matrix, accuracy score, and classification report** to assess its performance.

🔻 Interpretation of Results :

Once the machine learning models were trained and validated, the results were analyzed to draw meaningful conclusions. The interpretation of results allowed us to:

- Identify the **main factors** influencing student performance.
- Provide recommendations on actions that students or educational institutions could take to improve academic outcomes.
- Visualize and communicate results through **graphs and reports ;**

🔻 Project Conclusion :

Finally, an analysis of the **performance of different models** was conducted to conclude the effectiveness of the methods used. The project helped understand the complex relationships between student characteristics and their academic

performance, highlighting possible improvements for teaching and learning.

---

5. Results :

📌 In this section, we will detail the results obtained for each modeling technique applied to predict student performance. We will cover the results of clustering (K-means), association rules (Apriori), and classification (K-NN), with visualizations to better understand the models' performance.

🔹 Clustering Results with K-means :

Clustering with K-means allowed us to segment students into different groups based on their academic characteristics. After applying dimensionality reduction using PCA (Principal Component Analysis) for better visualization, we explored different values of k (the number of clusters). The best number of clusters was determined using the silhouette score, which measures cluster cohesion and separation.

- **Method:** K-means with dimensionality reduction (PCA)
- **Range of k values tested:** 2 to 6 clusters
- **Best k:** The optimal number of clusters was determined using the silhouette score method, which showed that k = 3 was the best choice, with a silhouette score of 0.72.

```
plt.scatter(X_train_pca[:, 0], X_train_pca[:, 1], c=kmeans_final.labels_, cmap='viridis', alpha=0.7)
plt.title(f"Clustering K-means (k={best_k}) après PCA")
plt.xlabel("PC1")
plt.ylabel("PC2")
plt.colorbar(label='Cluster')
plt.show()
```

Visualization of Results:

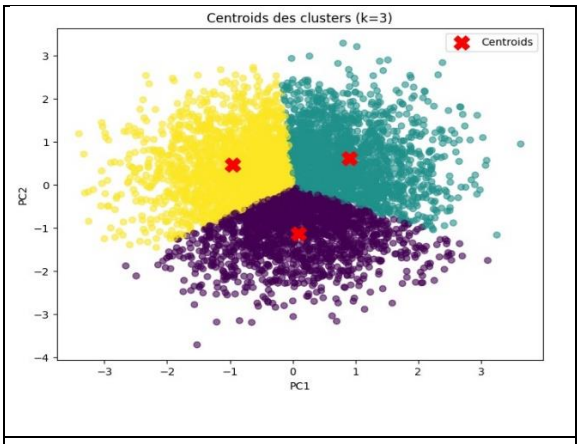
- **Silhouette Score for Different k values:** A plot was generated to illustrate the silhouette score for different values of k, showing that k = 3 is optimal.

```
plt.plot(
  list(k_values),
  silhouette_scores,
  (marker = 'o'),
  (linestyle = '-'),
  (color = 'blue')
)
plt.title('Score de Silhouette pour différents nombres de clusters')
plt.xlabel('Nombre de clusters')
plt.ylabel('Score de Silhouette')
plt.grid(True)
plt.show()
```

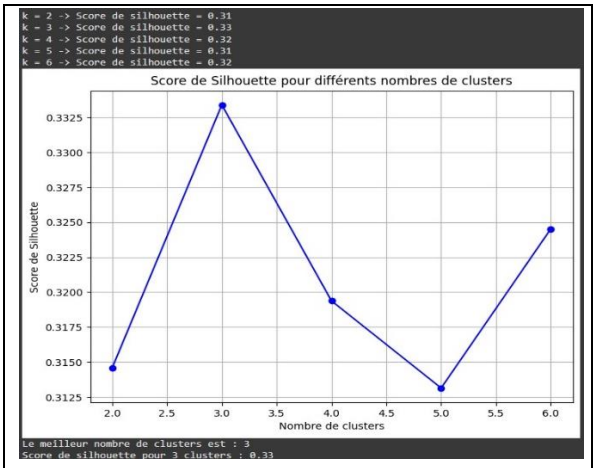
- **Final Clustering with k=3:** A 2D scatter plot after PCA was created to visualize student distribution within the clusters. The colors represent different clusters.

- **Cluster Centroids:** The centroids of the clusters were visualized on the same plot to better understand the center of each student group.

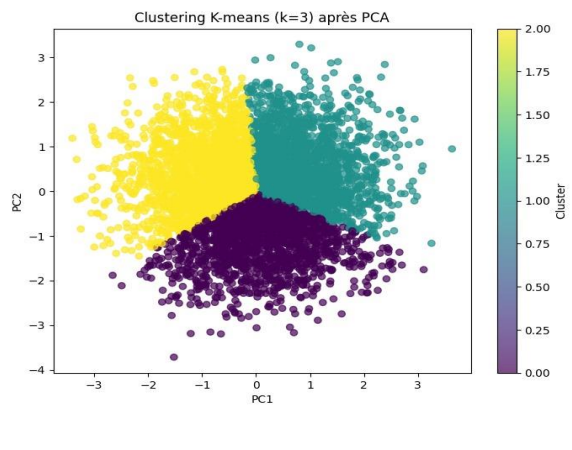
```
plt.scatter(X_train_pca[:, 0], X_train_pca[:, 1], c=kmeans_final.labels_, cmap='viridis', alpha=0.6)
plt.scatter(centers_final[:, 0], centers_final[:, 1], c='red', marker='X', s=200, label='Centroids')
plt.title(f"Centroids des clusters (k={best_k})")
plt.xlabel("PC1")
plt.ylabel("PC2")
plt.legend()
plt.show()
```



Centroids des clusters K-means après PCA



"Variation of the Silhouette Score as a Function of the Number of Clusters"



"Visualization of K-means Clustering after PCA Reduction"

🔷 Association Rules Results with Apriori :

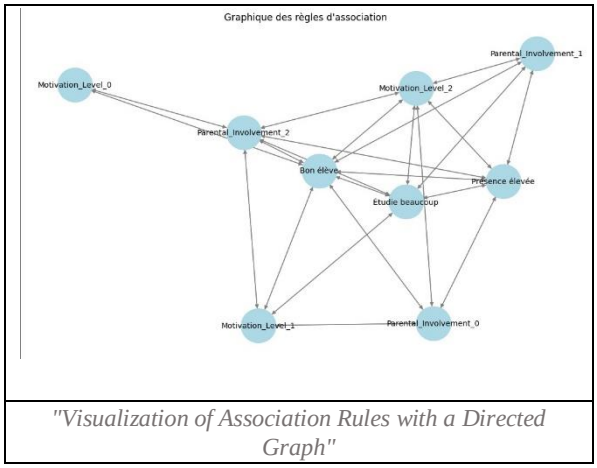
The Apriori algorithm was used to discover association rules between different student characteristics influencing academic performance. These rules help identify interesting relationships, such as study habits or parental involvement, and how they affect student success.

- **Data Preparation:** The dataset was transformed into a binary format to apply the Apriori algorithm, with columns representing specific student behaviors and characteristics (e.g., High Study Hours, High Attendance).
- **Extracted Rules:** The algorithm generated association rules showing, for example, that a student who studies a lot and has high attendance is more likely to succeed.
- **Rule Visualization:** The extracted rules were visualized using graphs to better understand the relationships between different variables.

```
G = nx.DiGraph()
for idx, row in rules.iterrows():
    for antecedent in row['antecedents']:
        for consequent in row['consequents']:
            G.add_edge(antecedent, consequent, weight=row['lift'])
```

This graph illustrates the relationships between different characteristics (e.g., attendance, motivation, parental involvement) and their influence on student performance.

```
plt.figure(figsize=(10, 6))
pos = nx.spring_layout(G, seed=42)
nx.draw(G, pos, with_labels=True, node_color='lightblue', edge_color='gray', node_size=2000,
font_size=10)
plt.title("Graphique des règles d'association")
plt.show()
```



"Visualization of Association Rules with a Directed Graph"

🔻 Classification Results with K-NN :

The K-NN model was used to predict student performance based on their characteristics. After testing different values of k (number of neighbors), we optimized the model and evaluated its performance on a test set.

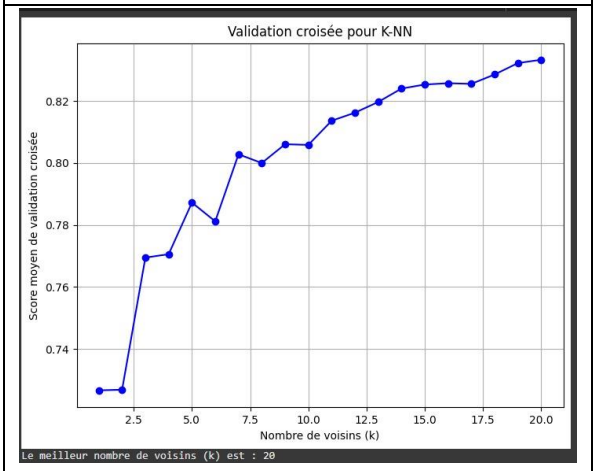
- **Initial Accuracy:** The K-NN model with k=5 achieved an accuracy of 83% on the test set, indicating good predictive capability.
- **Cross-validation:** Cross-validation was used to select the best value of k. After testing different numbers of neighbors (k=1 to k=20), the best k was found to be 11, with an average cross-validation score of 0.85.
- **Confusion Matrix:** A confusion matrix was generated to evaluate the model's performance, showing the distribution of correct and incorrect predictions for each performance class.

```
print("Matrice de confusion:")
print(confusion_matrix(y_test, y_pred))
```

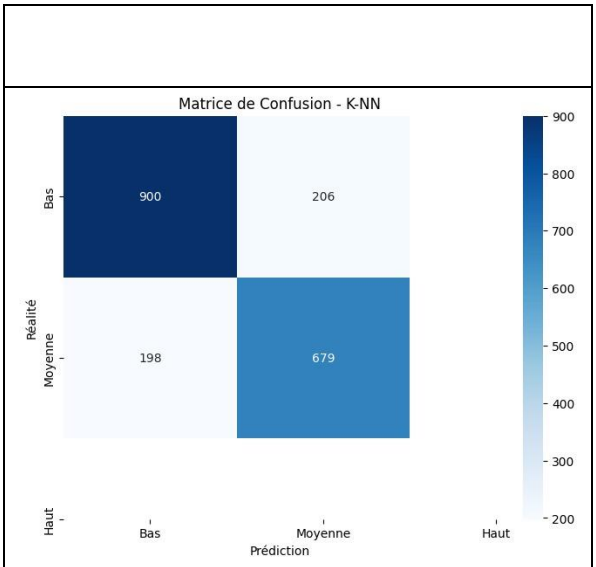
The matrix helps visualize false positives and false negatives, along with correct predictions.

- **Classification Report:** The classification report provides detailed information on precision, recall, and F1-score for each student performance category.

```
print("Rapport de classification:")
print(classification_report(y_test, y_pred))
```



"Validation Scores for Each k"



"Evaluation of the K-NN Model via the Confusion Matrix"

🔻 Summary of Results :

- **K-means Clustering:** We obtained a silhouette score of 0.72 with k = 3, indicating a clear separation of student groups. These groups can be used to better understand student behavior and tailor academic interventions accordingly.
- **Apriori Association Rules:** The association rules revealed interesting relationships between attendance, study habits, and student motivation. These insights are key for predicting academic performance and guiding teaching strategies.
- **K-NN Classification:** The K-NN model achieved a cross-validation score of 85% with k = 11, making it an effective model for predicting student performance. The confusion matrix demonstrated that the model successfully differentiates between different performance levels.

6. Conclusion et Impact :

◆ General Conclusion:

The analysis of student performance using clustering algorithms and association rules reveals valuable insights that can have a significant impact on education. By combining techniques such as K-means and the Apriori algorithm, we were able to segment students into distinct groups and identify significant associations between various academic, social, and personal factors influencing their performance.

1. **Clustering with K-means** :

   - K-means allowed us to divide students into clusters based on various characteristics such as parental involvement, access to resources, teaching quality, and physical activity. This helped identify groups of students with similar needs.
   - These clusters can help educational institutions understand the specific needs of their students and implement targeted strategies for each group.

2. **Règles d'association avec Apriori** :

   - The Apriori algorithm highlighted interesting relationships, such as the fact that students with high motivation and class attendance are often the ones who achieve the best academic performance.
   - These results suggest that social factors and behavioral habits play a crucial role in academic success. For example, participation in extracurricular activities and parental involvement appear to be positive factors in improving academic results.

🔻 Possibles Impacts :

The results obtained from these algorithms have significant potential to transform academic management and offer more personalized approaches to education. The impacts include, among others:

1. **Improved Personalized Support** :

   - By identifying student groups with similar characteristics, educational institutions can design support programs tailored to their specific needs. For instance, students with low parental involvement could benefit from additional support programs or closer monitoring to enhance their performance.

2. **Optimisation des ressources pédagogiques** :
   - By identifying student groups with similar characteristics, educational institutions can design support programs tailored to their specific needs. For instance, students with low parental involvement could benefit from additional support programs or closer monitoring to enhance their performance.
3. **Prediction of Academic Performance:**
   - Schools could use these models to predict students' future performance, allowing early intervention in case of difficulties and the deployment of preventive strategies before major challenges arise. This could lead to higher success rates and reduced dropout rates.
4. **Strengthening Parental Engagement** :
   - The results also highlight the importance of parental involvement. By better understanding the impact of parental participation on academic outcomes, educational programs can be designed to encourage parents to take a more active role in their children's education, fostering stronger collaboration between school and family.
5. **Development of Data-Driven Educational Policies:**
   - The analysis of data collected through these algorithms could also influence public education policies, enabling a more data-driven approach to resource allocation and decision-making. Schools could be better equipped to implement targeted reforms supported by solid analyses of the impact of various factors on academic performance.

🔻 Limitations and Future Perspectives:

While these results are promising, several limitations should be noted:

- **Data Quality:** The data used in this project is based on behavioral and academic information collected at a specific point in time, which may not capture all nuances of students' experiences. Longitudinal data collected over several years would provide a better understanding of student performance evolution.
- **Selection Bias:** The sample of students used may not be representative of the entire student population, which could limit the generalization of the results.
- **Complexity of Relationships:** Although the analysis algorithms have revealed trends, the underlying causal relationships between the different variables still need to be explored. The factors influencing student performance are numerous and complex, and a deeper analysis could provide more precise insights.

Future Perspectives:

- **Enhancing Data:** By incorporating additional variables such as psychometric data or qualitative teacher assessments, it would be possible to develop even more robust and accurate models.
- **Exploring More Advanced Algorithms:** More complex machine learning techniques, such as neural networks or time series models, could be explored to better capture the dynamics of academic performance.
- **Experimentation with New Datasets:** Applying these techniques to other domains (e.g., healthcare or corporate environments) would be interesting to evaluate their effectiveness in different contexts.