STUDENT PERFORMANCE: GRADE CLASS PREDICTION

Group 5 :
1. Malvin Ferdinand Tanzil - 2702208700
2. Marcelline Cathrine Wilison - 2702210604
3. Miecel Alicia Angel J - 2703327601
4. William Darma Wijaya - 2702218645

1 Problem Description

Classifying students into grade categories (A, B, C, D, F) is challenging due to the complex interplay of various factors influencing academic performance. This project uses machine learning to **streamline the classification process**, enabling targeted support for students.

2 Data Description

The **Student Performance Dataset** contains 2,392 student records including :

Student Information
Age, gender, ethnicity, study time, absences, tutoring.

Parental Involvement
Parental education and support levels.

School Activities
Participation in extracurriculars, sports, music, and volunteering.

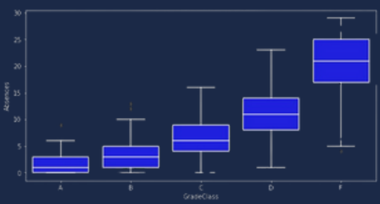
Academic Performance
GPA and GradeClass.

3 Feature Engineering

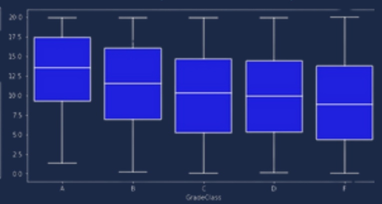
The target variable (GradeClass) was engineered to categorize students based on their GPA, making it easier to classify and interpret their academic performance.

Here are the distributions of selected crucial data columns that likely influence GradeClass:

Absences



Study Time Weekly



Students with **frequent absences** tend to have **lower GradeClass** categories, while those with **higher weekly study time** are more likely to achieve **higher GradeClass** categories.

4 Model & Justification

CatBoost:
It efficiently handles categorical features, making it ideal for datasets with many categorical variables, leading to better performance in such cases.

XGBoost:
Known for its high accuracy and versatility in handling structured data. Key advantages include:
1. Excellent performance and accuracy across most use cases.
2. Provides feature importance scores for interpretability.
3. Built-in regularization to prevent overfitting, ensuring robust models.

Evaluation Method 5

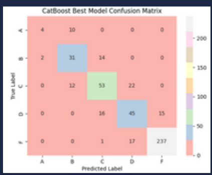
Classification Report:
Provides precision, recall, and F1-score for each class, along with their averages, making it a reliable method for evaluating model performance accurately.

Confusion Matrix:
Displays the prediction results for each class, showing the number of correct and incorrect predictions, helping to assess model performance comprehensively.

MODEL RESULTS

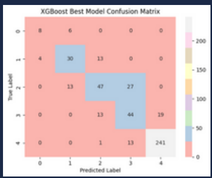
CatBoost

	precision	recall	f1-score	support
A	0.67	0.29	0.40	14
B	0.58	0.66	0.62	47
C	0.63	0.61	0.62	87
D	0.54	0.59	0.56	76
F	0.94	0.93	0.93	255
accuracy			0.77	479
macro avg	0.67	0.62	0.63	479
weighted avg	0.78	0.77	0.77	479



XGBoost

	precision	recall	f1-score	support
0	0.67	0.57	0.62	14
1	0.61	0.64	0.62	47
2	0.64	0.54	0.58	87
3	0.52	0.58	0.55	76
4	0.93	0.95	0.94	255
accuracy			0.77	479
macro avg	0.67	0.65	0.66	479
weighted avg	0.77	0.77	0.77	479



Overall Model Conclusion

Overall, both models performed similarly, **but XGBoost slightly outperformed CatBoost after fine-tuning**, with higher precision, recall, and F1-scores for predicting student GradeClass based on the available variables. Based on the evaluation results, **both models are reasonably effective for prediction, though not exceptionally strong.**