PROJECT E2:
JANA KOTŠNOVA
KARINA PINAJEVA
ANN MARLEEN VARUL

Introduction to Data Science
8.12.2025

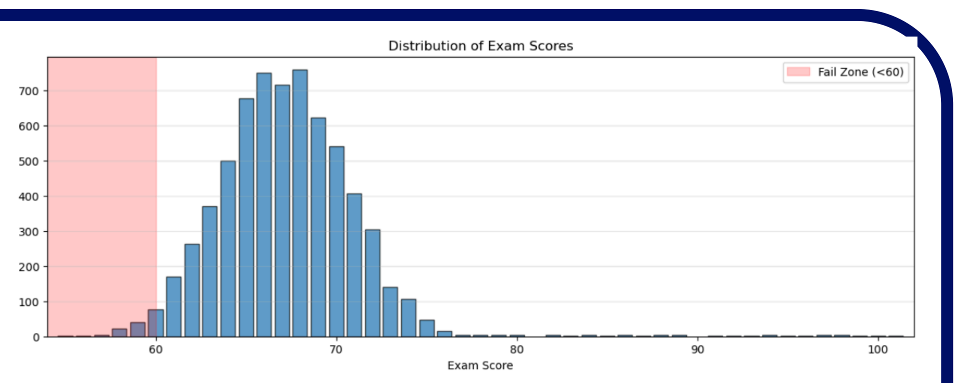
INSIGHTS INTO STUDENT PERFORMANCE:
Predicting Outcomes using Machine Learning and Identifying Risk Profiles by Clustering

Goals of this project

This projects aims to :

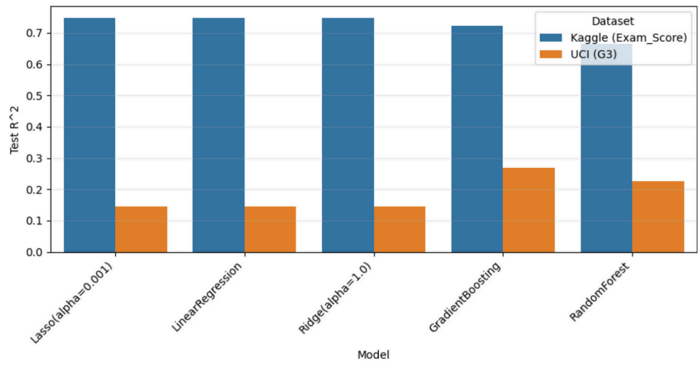
- use **regression algorithms** in Python to predict each student's final exam score based on their study habits, attendance, motivation, and background.
- apply **clustering** techniques to group students into categories such as high-performing, moderate-risk, and at-risk, based on their academic and behavioral patterns.
- train a simple **language model** that provides personalized feedback or warning messages for students who are predicted to be failing or trending toward low performance.

Data preprocessing



This project was built on an open **Kaggle dataset** (641.95 kB) with 6,590 students and the factors that may affect their academic performance, such as number of study hours, attendance percentage, level of involvement from parents, sleep hours, availability of study resources, and others. Data was cleaned: missing data was handled through **imputation** (mode for categorical variables), followed by **encoding** of categorical and binary features into numerical. **Statistical** parameters (max/min, average, median, mode, sd) were computed to detect potential outliers. The shape of the **distribution** and **cumulative distribution** function were visualized. The distribution of the Kaggle data is shown on the image above. For regression modeling, the dataset was **split into training and testing** sets to evaluate predictive performance. For clustering, meaningful **comparison groups** were created based on standard deviation thresholds, dividing students into At-Risk (below μ - σ), Average (μ ± σ), and High-Performing (above μ + σ) categories to then compare with results of clustering.

Regression model

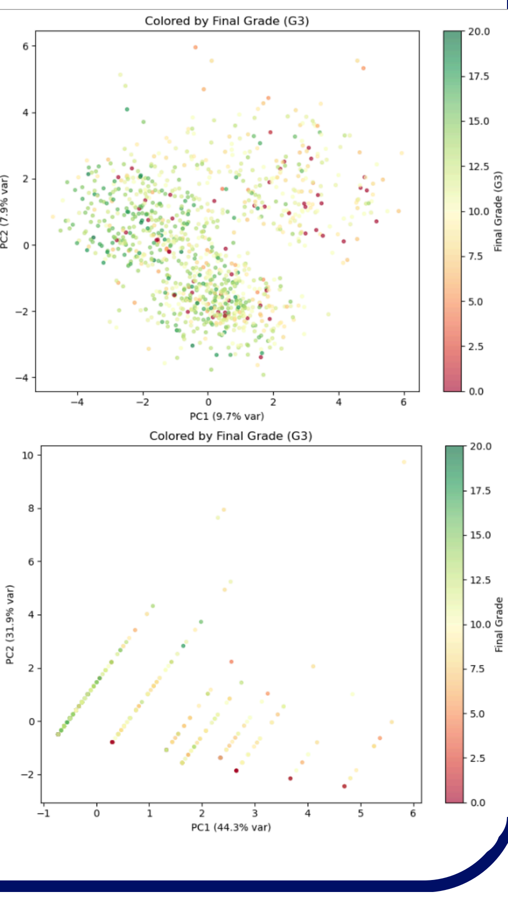


To address Goal 1 we framed the prediction of study performance as a **regression task**. For each dataset we used standardized numerical features and encoded categorical encoded categorical variables describing study habits (study time, previous scores, absences), attendance, lifestyle (sleep, free time, going out, alcohol use), and family or school background (parental education, support, school type, higher-education plans). Five regression algorithms were trained and compared: **Linear Regression**, **Ridge**, **Lasso**, **Random Forest**, and **Gradient Boosting**, with a simple mean baseline for reference. Models were evaluated on a held-out test set using **MAE**, **RMSE** and **R²**, and cross-validation was used to check that results were stable and not due to overfitting. The **bar chart** above summarizes how much of the variation in final exam scores **each model can explain** on the synthetic Kaggle dataset versus the real UCI school records, highlighting the **contrast** between artificially high accuracy on **synthetic data** and more modest but **realistic performance** on natural data.

Clustering

Three clustering techniques: **K-Means**, **Hierarchical**, and **Fuzzy C-Means** were applied to originally chosen Kaggle dataset. Upon discovering its poor quality (and synthetic nature), we decided to compare the results with clustering on the UCI Machine Learning Repository Student Performance dataset[2], collected via school reports and questionnaires. This dataset includes performance data for Mathematics and Portuguese courses, which we combined into a single dataset of 1 044 instances.

We applied all three clustering methods to the combined dataset, followed by **targeted K-Means** clustering using features identified as most important in Cortez and Silva's (2008) study[3]: **failures** (past class failures), **absences** (number of school absences), and **higher** (desire to pursue higher education). The real student data achieved **significantly better cluster separation** with a **Silhouette** Score of **0.653** compared to 0.352 for the artificial dataset. An interesting pattern emerged from the clustering visualization: **high-performing** students (green) form **distinct**, cohesive clusters with shared behavioral characteristics (low failures, good attendance, and educational aspirations), while **low-performing** students (red) are **scattered across clusters**, indicating that academic struggle stems from diverse factors such as excessive absences, past failures, lack of educational support, or combinations thereof. This finding suggests that while successful behaviors can be promoted uniformly, at-risk students require individualized assessment and support based on their **specific risk factors**. Notably, Hours_Studied, a primary determinant in the artificial dataset, showed neither high correlation nor importance in the real student data. Instead, **desire to pursue higher education** emerged as a critical predictor, demonstrating that educational aspirations, combined with **good attendance** and **low failure rates**, distinguish successful students from their struggling peers.



Finetuning GPT2

In this project, **GPT-2** was fine-tuned on two different student performance datasets, the UCI Student Performance dataset and a Kaggle student success dataset, to generate **personalized academic feedback** based on structured student profiles. For both datasets, rule-based methods were used to create thousands of **profile - feedback** pairs, allowing **GPT-2** to learn how specific student characteristics relate to appropriate guidance. The UCI dataset produced **1044** training examples focusing on factors such as study time, travel time, and past failures. The Kaggle dataset provided a much larger training set of **6607** examples with additional variables like motivation, attendance, and SD_Group. By training on these examples, GPT-2 learned how different combinations of student characteristics relate to appropriate feedback messages. The **fine-tuned models** can now generate consistent, context-aware feedback for **new profiles**, adapting their tone and recommendations based on performance levels (encouraging high-performing students, supporting at-risk students). This demonstrates how a **general-purpose language model** can be customized into a domain-specific educational **feedback system** through targeted fine-tuning.

Conclusion

Our analysis revealed that the **widely-used** Kaggle Student Performance dataset is **synthetic** and **unsuitable** for machine learning tasks due to **minimal score variation**. The UCI dataset[2], based on **real school records**, provided **meaningful results**: **Gradient Boosting achieved MAE of 2.5** points, with tree-based models explaining approximately **25% of grade variance** using only demographic and behavioral features. Clustering achieved a **Silhouette Score of 0.653** compared to 0.352 for the artificial data. The real data showed that **educational aspirations**, **past failures**, and **attendance** were the **strongest predictors**, not study hours as suggested by the synthetic dataset. Tree-based models outperformed linear approaches, indicating **non-linear relationships** between student characteristics and performance. Clustering revealed that **high-performing** students share **similar behavioral patterns**, while **at-risk** students fail for **various reasons**, suggesting that **intervention strategies** should differ: **standardized approaches** for promoting success behaviors, but **individualized support** for struggling students. While our models achieve **modest but realistic accuracy** on real data, synthetic datasets can lead to **incorrect conclusions** about which factors influence student performance.

Sources:

[1] https://www.kaggle.com/datasets/anassarfraz13/student-success-factors-and-insights
[2] https://archive.ics.uci.edu/dataset/320/student+performance
[3] Cortez, P., Silva, A. (2008). USING DATA MINING TO PREDICT SECONDARY SCHOOL STUDENT PERFORMANCE. *University of Minho*.