Beyond Grades - Impact Report

Unveiling the Factors Behind Students' Success

Enrico Vaccari - Applied Machine Learning (Regression) - Tomorrow University - Sep 2025

Intro

Education is one of the most powerful levers for building a sustainable and fair society. Yet, schools and universities often struggle to identify early which students might face challenges and which factors most strongly influence academic outcomes.

> Our Ridge Regression model explains **~95%** of GPA variance across students with an average error of **0.16 points**, enabling earlier identification of at-risk students and supporting fairer interventions for schools within one academic term.

This project explores how data science and machine learning can be used responsibly to predict student performance. By analyzing a dataset of student characteristics and academic results, the goal was to:

Build predictive models that estimate academic performance.

Identify the most relevant factors influencing student success.

Translate findings into insights that can support educators & institutions.

Table of Contents

1. Executive Summary

As just mentioned, ***Beyond Grades*** - *Unveiling the Factors Behind Students Success* - explores how machine learning can be applied responsibly to predict student performance and identify early signs of academic risk. Using a dataset of student records from Multan, Pakistan (2025), we developed regression models to estimate GPA based on academic, familial and socio-demographic factors.

1.1 Key Findings

- The final Ridge Regression model explained ~95% of the variance in GPA ($R^2 \approx 0.95$), with a prediction error of ~0.16 GPA points and RMSE ≈ 0.20 on the test set.
- Absences, weekly study time and parental education consistently emerged as the strongest predictors of academic performance.
- Demographic features such as gender and ethnicity contributed minimally once academic and familial variables were accounted for, reinforcing that interventions should focus on structural and behavioral factors rather than demographics.

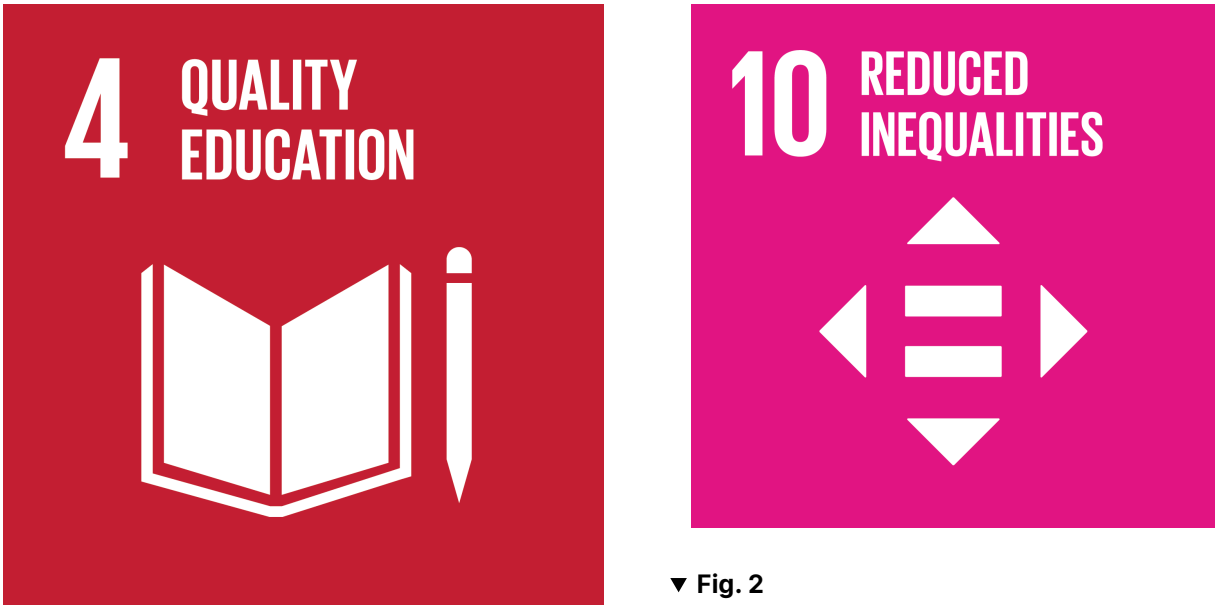
1.2 Social Impact

The social impact can be summarized into three main points:

> Early detection of at-risk students allows schools to reduce dropout risk and target tutoring or engagement interventions more efficiently.

> Educators and policymakers can prioritize support strategies around attendance, study habits and family engagement.

> From a sustainability perspective, improving academic outcomes contributes directly to **SDG 4 (Quality Education)** and **SDG 10 (Reduced Inequalities)**, ensuring that students from diverse backgrounds are given fair chances to succeed.



▼ **Fig. 2**

▼ **Fig. 1**

SDG 4 (Quality Education)

SDG 10 (Reduced Inequalities)

1.3 Deployment Readiness

The model is suitable for pilot deployment in an educational context, provided it is used as a decision-support tool rather than a stand-alone classifier. Human oversight is essential to prevent misuse.

1.4 Recommendations

Before diving into the project, here are a few recommendations on how to properly use the model:

🔎

- Launch a pilot integrating the Ridge model into school dashboards to flag students at risk.
- Expand the dataset to include additional variables (nutrition, sleep quality, mental health) to improve predictive power.
- Monitor model fairness and recalibrate quarterly to prevent drift or bias amplification.

2. Stakeholder Facing Summary

Imagine a teacher with a class of 30 students. Some show signs of disengagement, but without objective signals, it is hard to know who truly needs help.

The goal is to show that with simple student data - such as study time, attendance and parental education - we can predict academic outcomes with high reliability.

The strongest signal is clear: **students who attend regularly and study consistently are much more likely to succeed**, while demographic traits like gender matter far less. This is good news, because it means schools can act on concrete, modifiable factors instead of fixed labels.

For **teachers**, this means spotting struggling students earlier and guiding them toward tutoring or parental involvement. For **school leaders**, it means allocating resources where they will make the most difference. For **policy-makers**, it means evidence to support programs that keep students in class and engaged.

> This work contributes to **fairer, more inclusive education** - helping reduce inequalities by ensuring interventions are based on effort and support, not stereotypes.

2.1 Why predicting student performance matters

Student performance is a central indicator of educational outcomes, shaping future opportunities for individuals and the broader social fabric. Early identification of students at risk of underperforming allows schools to design **targeted interventions**, provide **personalized support** and ultimately promote **equity in education**.

2.2 Educational & Business Context

Education is central to building a sustainable future. In many contexts, inequalities in access to resources, parental support, or quality of teaching translate into performance gaps that widen over time. Moreover, student dropout and underperformance carry major social costs, reinforcing cycles of inequality. Schools often intervene too late, after academic struggles are already entrenched.

By uncovering which factors truly matter, data-driven insights can guide educators and policymakers to design interventions that are not only efficient but also socially just.

> This project aligns with Tomorrow University's mission of using **data science for social good**, emphasizing responsible use of technology in service of sustainability and educational fairness.

2.3 Stakeholder Map

Primary Stakeholders

1. **Students (primary beneficiaries)**
   - **Potential benefit**: Early identification of risk → targeted support programs.
   - **Potential concern**: Risk of being unfairly labeled "at-risk" without proper context.
2. **Teachers & Counselors**
   - **Benefit**: Actionable insights into attendance, study time and support factors → prioritize interventions.
   - **Concern**: Added workload or pressure if flagged students exceed available resources.
3. **School Leadership / Administrators**
   - **Benefit**: Aggregate view of at-risk patterns → guide resource allocation, program funding.
   - **Concern**: Misinterpretation of model outputs as definitive judgments rather than probabilistic guidance.

Secondary Stakeholders

1. **Parents & Families**
   - **Benefit**: Better communication of how support and engagement influence outcomes.
   - **Concern**: Stigmatization if family education/support variables are misused.
2. **Policy Makers / Education Boards**
   - **Benefit**: Data-driven basis for attendance policies, tutoring programs, equity initiatives.
   - **Concern**: Risk of reinforcing existing inequalities if models are applied without fairness checks.
3. **Researchers & EdTech Developers**
   - **Benefit**: Open dataset, methodology and pipeline → opportunities to improve or scale solutions.
   - **Concern**: Model drift or bias if applied in different cultural/educational contexts without validation.

Stakeholder	Type	Impact Level	Probability of Influence	Priority for Engagement	Notes
Students	Primary	High	High	Critical	Direct beneficiaries; need fair, actionable predictions
Teachers/Counselors	Primary	High	Medium	High	Use insights daily; require clear guidance and minimal extra workload
School Leadership	Primary	High	High	Critical	Allocate resources; risk of misinterpreting metrics
Parents/Families	Secondary	Medium	Medium	Medium	Can reinforce student support; risk of stigmatization
Policy Makers	Secondary	High	Low	Medium	Potential long-term influence on policy; fairness concerns
Researchers/EdTech	Secondary	Medium	Medium	Low	Value in methodology and replication; risk of misuse in other contexts

3. Problem Statement

🎯 The goal is to build a regression model that can **predict GPA scores** using a mix of academic, familial and socio-demographic variables.

3.1 Desired Outcome

Success means identifying the factors with the greatest predictive power and using these insights to inform **practical recommendations** for educators, parents, and policymakers. Ultimately, the outcome should help schools intervene earlier, reduce dropout risk and create fairer opportunities for students from diverse backgrounds.

3.2 Scope of Analysis

The scope:

- **includes** academic and family-related predictors such as parental education, study habits, attendance and parental support.
- **does not include** unobservable factors like motivation, mental health, or nutrition (absent in the dataset).

The analysis is limited to the **Student Performance Data (Multan, Pakistan, 2025)**, which constrains generalizability but offers a valuable case study for developing the approach.

3.3 Success Metrics

Progress will be measured by:

- Model performance on regression metrics (**$R^2 \geq 0.90$, RMSE ≤ 0.25 GPA points**) to ensure accuracy.
- Ability to rank key predictors (e.g., absences, study time, parental education).
- Ability to translate technical results into **clear, actionable insights** for stakeholders.
- Consideration of fairness, ensuring demographic variables like gender do not bias predictions.

3.4 Hypotheses & guiding questions

This study applies regression analysis to predict students' performance scores (measured as GPA) based on a set of academic, familial and socio-demographic variables. The central hypotheses are:

🔎

- **H1:** *Higher parental education level is positively correlated with higher student performance.*
- **H2:** *Stronger study habits and consistent attendance are strong predictors of better academic outcomes.*
- **H3:** *Once other factors are controlled for, gender has little to no significant predictive power.*

These hypotheses give rise to guiding questions:

1. Which personal, familial and academic features are the strongest predictors of student performance?
2. Do within-student factors (study time, mindset, motivation) outweigh structural or demographic factors (gender, ethnicity)?
3. How can predictive insights be translated into **actionable recommendations** for schools, parents, and policymakers?
4. What are the risks of misusing predictive models in education, and how can we ensure fairness and transparency?

Before diving into the dataset research, we thought of potential target and influencing factors to look for:

- **Target:** students' performance and success at school (in the form of GPA, StudentPerformance, Grade etc...)

🔎 - **Influencing factors:** parental education level, socio-economic status, quality of teaching and school environment, student motivation and mindset, study habits and time on task, attendance and engagement, nutrition and sleep quality, parental involvement and support, peer influence and social environment, mental health and stress levels.

4. Project Management & Workflow

A strong project is not only about models, but also about how work is structured and reproduced. To ensure rigor, this project was designed as a reproducible pipeline with clear version control environment management and transparent organization.

4.1 Tools, environment & version control

- **IDE & workflow:** Visual Studio Code, using integrated Jupyter notebooks for analysis and Git for version control.
- **Environment:** Conda-managed Python environment with pinned dependencies (`environment.yml`) to guarantee reproducibility.
- **Version control:** GitHub repository with structured commits, feature branches and issue tracking. This setup ensures both peer review and rollback capabilities.
- **Notebook order:** Six numbered notebooks (`01_data_loading` to `06_model_evaluation`) correspond to each stage of the pipeline, minimizing confusion.

4.2 Reproducibility and pipeline design

- **Determinism:** Random seeds fixed across data splits and cross-validation.
- **Leakage control:** Splits occur before preprocessing; encoders and scalers fitted only on training data.
- **Artifact saving:** Post-cleaning and preprocessing datasets saved to `/data/interim/` and `/data/processed/`
- **Metadata:** Config files (`config.yml`) track seeds, folds, and parameters; split metadata stored as `split_meta.json`.
- **Outputs:** Metrics, plots and reports exported to `/outputs/` for clarity and reproducibility.

```
graph TD
    A[Raw Data] --> B[Data Splitting]
    B --> C[Structural Cleaning]
    C --> D[Exploratory Data Analysis]
    D --> E[Preprocessing]
    E --> F[Model Training & Cross-Validation]
```

```
F --> G[Evaluation & Diagnostics]
G --> H[Export Models & Artifacts]
H --> I[Monitoring & Fairness Checks]
```

▼ **Fig. 3**

Pipeline adopted.

4.3 Repository Structure

The folder structure was designed to separate raw, interim, and processed data, and to keep notebooks modular.

```
project-root/
├── data/
│   ├── raw/          # original, untouched
│   ├── interim/      # cleaned data, split indices, encoders
│   ├── processed/    # final analysis-ready tables
│   ├── meta/         # split metadata
│   └── docs/         # feature summary
├── notebooks/
│   ├── 01_data_loading.ipynb
│   ├── 02_data_cleaning.ipynb
│   ├── 03_data_splitting.ipynb
│   ├── 04_eda.ipynb
│   ├── 05_preprocessing.ipynb
│   └── 06_model_evaluation.ipynb
├── src/
│   ├── data/         # data loading & cleaning scripts
│   ├── features/     # feature engineering scripts
│   ├── project/      # project-specific logic
│   └── utilities/    # helper functions
├── artifacts/        # final Ridge pipeline (.joblib)
├── outputs/
│   ├── figures/
│   │   ├── EDA/
│   │   └── modeling/
│   └── reports/
├── extra/            # extra log files
├── environment.yml   # reproducibility file
├── config.json       # paths, seeds, CV settings
├── README.md
├── requirements.txt
├── how_to_run.md
├── ethics.md
└── .gitignore
```

▼ **Fig. 4**

Repository Structure.

5. Data Discovery & Structural Cleaning

5.1 Dataset overview

The dataset used in this study is titled **“Student Performance Data”** (Kaggle, curated by Muhammad Azam). It provides a structured collection of student records from **Multan, Pakistan**, capturing personal, familial and academic characteristics alongside a continuous **performance score (GPA)**.

🔎

- **Dataset name:** Student_performance_data.csv
- **Source:** Kaggle
- **Author:** Muhammad Azam
- **License:** CC BY 4.0
- **Last updated:** June 2025
- **Geographic area:** Multan, Pakistan
- **Structure:** Structured tabular dataset (CSV, UTF-8 encoded)
- **Collection method:** Aggregated school records and local surveys conducted by an educational organization
- Why chosen: recent, clean, structured and highly relevant to educational and social sustainability research.

# StudentID Unique identifier for each student	# Age Age of the student in years	Gender Gender of the student (Male/Female/Other)	Ethnicity Ethnic background of the student	ParentalEducation Highest education level achieved by either parent	# StudyTimeWeekly Total study time per week in hours	# Absences Number of school days missed
1001 … 3392	15 … 18	2392 total values	2392 total values	2392 total values	0 … 20	0
1001	17	1	0	2	19.833722807854713	7
1002	18	0	0	1	15.40875605584674	0
1003	15	0	2	3	4.21056976881226	26
1004	17	1	0	3	10.028829473958215	14
1005	17	1	0	2	4.6724952729713385	17
1006	18	0	0	1	8.191218545250186	0
1007	15	0	1	1	15.601680474699295	10
1008	15	1	1	4	15.424496305808074	22
1009	17	0	0	0	4.562007558047703	1
1010	16	1	0	1	18.444466363097202	0
1011	17	0	0	1	11.851363655296536	11
1012	17	0	0	1	7.59848581924029	15

▼ **Fig. 5**

Chosen dataset (Kaggle):

Link here: https://www.kaggle.com/datasets/muhammadazam121/student-performance-data

5.2 Target Variable & Features

The dataset identifies 'GPA' as the key outcome of interest, typically expressed as a *continuous numeric score*. This makes it a suitable target variable for regression analysis, since the prediction task involves modeling continuous values rather than discrete categories. Regression methods can therefore be effectively employed to estimate how changes in independent variables, such as study time or family background, are associated with variations in overall student performance.

The dataset provides a wide range of features, including *demographic characteristics, parental education levels, study habits, attendance records and access to educational resources.* Among these, variables such as parental education, family income, attendance, and dedicated study time are likely to emerge as the *strongest predictors*. However, some features may add little value to the model, particularly if they exhibit very low variance, contain excessive missing values, or overlap significantly with other variables through multicollinearity. A correlation analysis and variance inflation factor (VIF) check will therefore be essential in determining which predictors should be retained and which may be dropped from the regression model.

Name	Role	Type	Description	Expected Predictive Power	Keep/Drop
StudentID	ID	/	/	/	Drop
Age	Feature	Numeric	Student's age in years (15–18).	Low–Moderate	Keep
Gender	Feature	Categorical	Student's gender.	Low	Keep (fairness check)
Ethnicity	Feature	Categorical	Student's ethnic background.	Low–Moderate (bias risk)	Likely Drop
ParentalEducation	Feature	Ordinal	Parent's education (0 = highest, 4 = none).	High	Keep
ParentalSupport	Feature	Categorical	Emotional/academic support from parents.	Moderate–High	Keep
Tutoring	Feature	Binary	Whether the student receives tutoring.	Moderate	Keep
StudyTimeWeekly	Feature	Numeric	Weekly study time (hours).	High	Keep
Absences	Feature	Numeric	Number of school days missed.	High	Keep
Sports	Feature	Binary	Participation in sports.	Low–Moderate	Keep (optional)

Name	Role	Type	Description	Expected Predictive Power	Keep/Drop
Music	Feature	Binary	Plays a musical instrument.	Low	Likely Drop
Extracurricular	Feature	Binary	Participation in extracurricular activities.	Moderate	Keep
Volunteering	Feature	Binary	Volunteering activities.	Low–Moderate	Keep (exploratory)
GradeClass	Auxiliary	Categorical	Binned GPA (Fail–Excellent).	–	Drop
GPA	Target	Numeric	Final GPA on 0–4 scale.	–	Target

5.3 Data Quality Assessment

Dimension	Assessment	Notes
Accuracy	High	Dataset pre-cleaned: duplicates removed, missing values handled, categorical labels standardized. Some self-reported survey variables may include minor inaccuracies.
Completeness	Good	Core predictors present (demographics, attendance, study habits, parental education). Missing potentially important variables like nutrition, sleep quality, and mental health.
Consistency	High	Column names, formats, and categorical labels are standardized (e.g., education levels, gender).
Timeliness	Very recent	Coverage: March-June 2025, with last update on June 14, 2025.
Relevance	Strong	Directly aligned with the research objective of predicting student performance based on socio-academic factors.
Representativeness	Moderate	Balanced by gender and academic variables, but geographically restricted to Multan. Cultural/systemic factors may limit generalization globally.

5.4 Ethical/Legal Considerations and Limitations/Assumptions

When analyzing student performance, it is crucial to balance technical rigor with ethical responsibility. This subsection highlights key ethical concerns and dataset limitations to ensure the findings are interpreted fairly.

Ethical Considerations

Limitations

🔎
- The dataset is publicly released under a CC BY 4.0 license, making it legally shareable and reusable with proper attribution. It has been anonymized and cleaned before publication, ensuring compliance with open data requirements.

🔎
- The dataset is specific to Multan, Pakistan, which restricts how far the findings can be generalized to other cultural, social, or educational contexts.

🔎
- Although anonymized, privacy and informed consent remain key considerations. Researchers must handle the data responsibly and ensure results are communicated in a way that avoids reinforcing stereotypes related to gender or socioeconomic background.

🔎
- Important elements such as nutrition, sleep quality and mental health are not included, limiting the ability to capture the full range of factors that affect student performance.

🔎
- Indicators such as GPA may reflect grading practices or teacher bias. For this reason, predictions should be interpreted as tools for supporting and empowering students, rather than for punitive purposes or exclusion.

🔎
- Both GPA and GradeClass are influenced by grading policies and teacher subjectivity, meaning they may not fully reflect true student ability and must be interpreted with caution.

5.5 Stakeholders, Beneficiaries & Impact

When developing predictive models for student performance, it is essential to consider **who will use the results, who will benefit, and what real impact the insights may have**.

- **Stakeholders**
  - *School administrators*: resource planning, early interventions, curriculum adjustments.
  - *Teachers*: better understanding of drivers of success, ability to target support effectively.
  - *Parents & families*: insights into how home environment and study habits influence outcomes.

    - *Policymakers & NGOs*: evidence base for educational equity programs.
- **Beneficiaries**
    - *Students*: tailored interventions and fairer opportunities to succeed.
    - *Teachers*: tools to personalize strategies and allocate support efficiently.
    - *Communities*: improved educational outcomes, reduced dropout rates, lower inequality.
- **Real Impact**
    - Early detection of at-risk students.
    - Promotion of **equity in education** by showing which factors matter most (e.g., attendance, study time, parental support) and which may not (e.g., gender).
    - Data-driven foundation for long-term interventions that reduce exclusion and foster achievement.

5.6 Initial quality checks (duplicates, nulls, inconsistencies)

After being loaded in, the dataset underwent a verification process to confirm its readiness for downstream analysis.

- **Duplicates:** A row-level check confirmed that no duplicate entries were present, meaning each student record was unique.
- **Null values:** Systematic scans across all columns showed no missing values. This indicated that imputation procedures would not be necessary, simplifying preprocessing.
- **Inconsistencies:** We validated categorical features (e.g., gender, parental support, extracurricular activities) to ensure consistent label formats and frequencies. All categories matched expected encodings, with no stray values or mis-typed entries.

Together, these checks confirmed that the dataset had already been pre-cleaned at source. Documenting this step ensures transparency: although no corrective actions were needed, verifying data quality guards against silent errors that could otherwise bias results.

5.7 Handling structural/formatting issues

Next, we assessed the dataset's internal structure to guarantee smooth integration with the preprocessing pipeline.

- **Column names and data types:** All variables were inspected for proper typing (numeric, categorical, ordinal). GPA and other continuous variables were confirmed as numeric, while binary and ordinal features were properly encoded. Minor adjustments ensured consistency in column naming conventions (e.g., snake_case).
- **Variable semantics:** A critical structural issue was detected in the `ParentalEducation` variable. Its original coding was reversed, with `0` representing the highest level of education and `4` the lowest. Left uncorrected, this inversion would have misled interpretation and modeling. We inverted the scale so that higher values consistently indicated higher levels of education, aligning the feature with intuitive expectations.

- **Compatibility checks:** Finally, we validated that categorical labels and numeric ranges fell within plausible bounds, ensuring all features were aligned for subsequent splitting, preprocessing and modeling.

Apart from the parental education adjustment, no additional formatting fixes were required. The dataset could therefore be passed into exploratory analysis and modeling with confidence in its structural integrity.

6. Data Splitting & Preprocessing Strategy

The next essential step was to **split** the data into train and test sets: this is useful for EDA (as it will be performed on the train set) and essential for preprocessing and modeling later on.

6.1 Why the Split Matters

Before exploring the dataset, we first established a reliable way of splitting the data. This decision was fundamental, because every later chart and model evaluation depends on it. A clean, leakage-safe split ensures that the performance we report reflects how the model would behave with new, unseen students. For stakeholders, this means our accuracy estimates are realistic and not artificially inflated.

We divided the dataset into **training (≈80%)** and **test (≈20%)** subsets. The training portion was used for preprocessing, feature engineering and cross-validation, while the test set was kept aside until the very end as an unbiased benchmark. To maintain fairness, we stratified the split by GPA quantiles so that both sets preserved the overall distribution of outcomes and subgroup proportions.

```
- `X_train`: training features
- `X_test`: test features
- `y_train`: training target
- `y_test`: test target
```

▼ **Fig. 6**

Dataset split sets.

6.2 Cross-Validation on Training Data

Within the training set, we applied **5-fold cross-validation** to train candidate models, evaluate them and tune hyperparameters. Although this step formally belongs to the modeling stage, it has been addressed in this section as it is a vital part of preprocessing. Each fold involved fitting preprocessing, feature selection and the estimator on four-fifths of the data, then validating on the remaining fold. This rotation produced a reliable estimate of how the pipeline would generalize to new students. Because preprocessing was refitted inside each fold, the procedure avoided leakage and ensured that performance numbers were honest.

6.3 Split Quality Checks

After splitting, we verified that the test set was representative. GPA means and standard deviations between training and test were nearly identical and subgroup proportions (e.g., gender, parental

education) aligned closely. These checks reassured us that the hold-out set reflected the same population as the training set, making final results credible.

```
Training target distribution (same edges as stratify):
GPA
0    383
1    383
2    382
3    382
4    383
```

▼ **Fig. 7**

Train test distribution.

```
Test target distribution (same edges as stratify):
GPA
0    96
1    95
2    96
3    96
4    96
```

▼ **Fig. 8**

Test set distribution.

6.4 Guarding Against Leakage & Pipeline

To further reduce risks, we embedded every transformation inside a `Pipeline`. Encoders, scalers, imputers, and selectors were trained on the training folds only, then applied to validation and test sets. This strict separation guaranteed that no future information leaked into the model. Random seeds and saved indices ensured that splits were reproducible and auditable.

```
### Pipeline Overview

Raw Data
  |
  ├── Numeric Features
  |     ├── Imputation → Median*
  |     └── Scaling → StandardScaler
  |
  ├── Categorical Features
  |     ├── Imputation → Most Frequent*
  |     └── Encoding → OneHotEncoder
  |
  └── Feature Selection (SelectKBest, optional)
        |
        ▼
    Model Estimator
    (e.g., Ridge, LinearRegression)
        |
        ▼
    Predictions

* these have been defined although there are no missing values
```

▼ **Fig. 9**

Pipeline overview.

Although in the notebooks preprocessing comes after EDA, here we describe it directly after the split because the pipeline itself does not touch the test data; it is only defined and later applied during modeling. This structure clarifies the design before we move to exploration.

Pipeline Construction

- **Numeric features**: imputed with the median to handle missing values robustly, then standardized with `StandardScaler`.
- **Categorical features**: imputed with the most frequent value, then one-hot encoded (`OneHotEncoder(handle_unknown='ignore')`).
- **Feature selection**: optional `SelectKBest(f_regression, k=k_best)` to retain the most predictive features.
- **Model**: flexible endpoint, able to wrap any estimator (e.g., `LinearRegression`, `Ridge`).

All steps are combined with `Pipeline` and `ColumnTransformer`, producing a clean, modular and scikit-learn-compatible workflow.

By carefully splitting the data and building a leakage-safe preprocessing pipeline, we created a robust foundation for the entire project. The combination of stratified splits, cross-validation and modular preprocessing ensured that evaluation would be fair, reproducible and transparent. This rigorous design sets the stage for exploratory data analysis and, ultimately, reliable modeling.

7. Exploratory Data Analysis

7.1 Baseline

With the split established, we turned to exploratory data analysis (EDA).

The goal was **twofold:**

1. to understand the structure of the dataset
2. to generate/test hypotheses that could inform modeling choices.

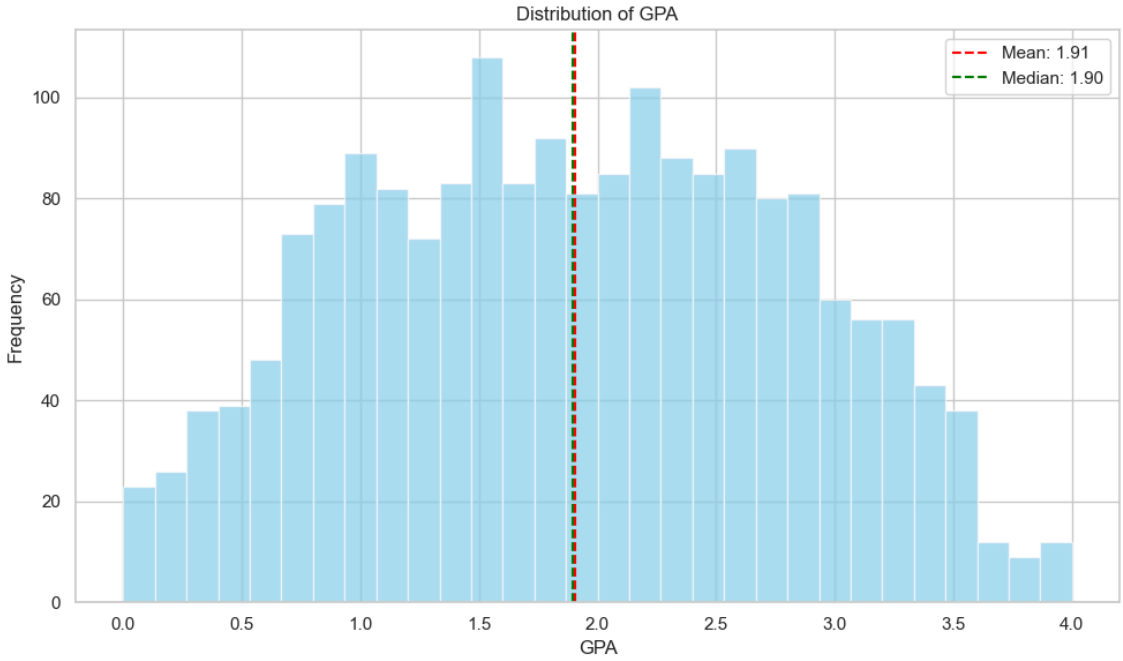
For stakeholders, this stage helps reveal the *story* hidden in the numbers - what factors most visibly shape student outcomes and where educators might intervene.

7.2 Distribution of Key Variables

The first step was to inspect the main variables and their distributions as well as the target.

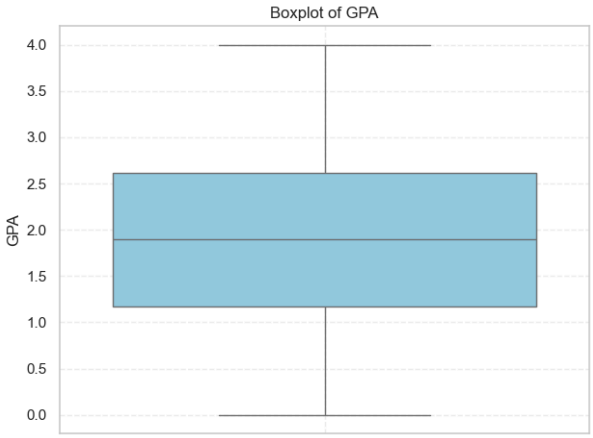
Target Variable

- **GPA:** unimodal, centered near the mean, with tails highlighting both high achievers and at-risk students. However, the Shapiro test returned a p-value close to 0, which formally ed us to reject the null hypothesis of normality (as the Q-Q plot show as well). This was also confirmed by the Q-Q plot, which shows a few deviations in the tails. The average GPA was 1.91 and the range was [0,4) which are the minimum and maximum value it can take.



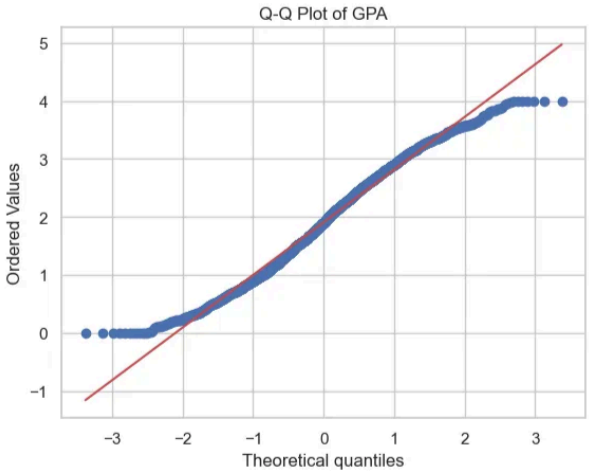
▼ **Fig. 10**

GPA distribution (histogram).



▼ **Fig. 11**

GPA distribution (box plot).



▼ **Fig. 12**

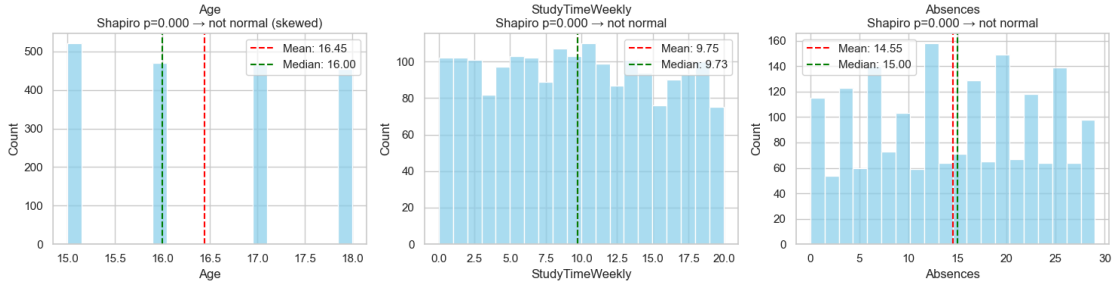
GPA Q-Q plot.

Numeric Variables

None of the numeric variables seemed to get even close to a normal distribution.

- **Age:** Clustered tightly between 15 and 18, confirming little variation. Also, age showed minimal correlation with GPA.

- **StudyTimeWeekly:** Median = 9.73 hours, mean = 9.75 hours, max = 20 hours.
- **Absences:** Median = 15 days, mean = 14.55 days, but the maximum got close to 30 days. This long tail highlighted a handful of students with extreme disengagement.



▼ **Fig. 13**

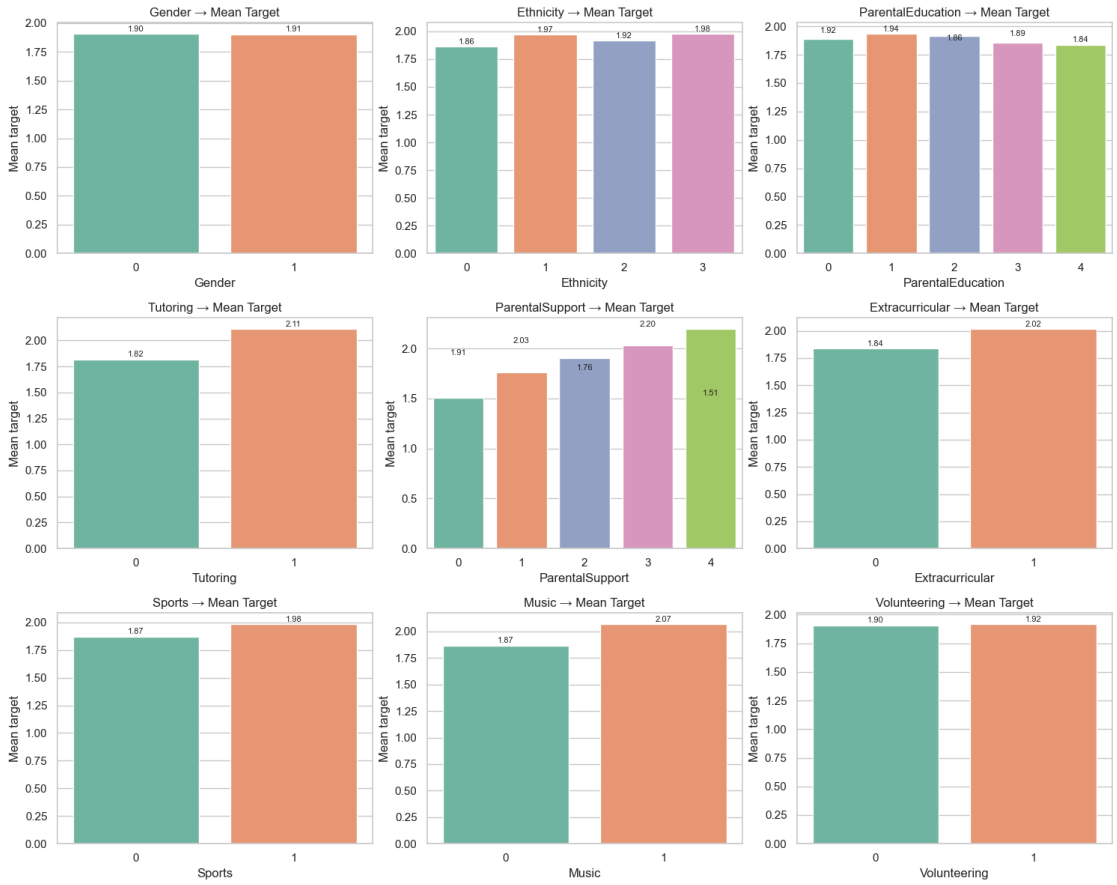
Numeric variable distributions (histograms).

Also, it has been proved that age has basically no influence on target and, after a few tests later on, we decided to exclude it from the modeling stage. On the other hand, absences seem to be highly (anti)correlating with the target, meaning more absences lead to a lower score. This variable has been determined before the target, so any potential leakage has been excluded. Finally, there is small positive correlation between "StudyTimeWeekly" and the target, although we were expecting it to be more marked.

Note: we eventually decided to keep Age as one of the features to prove it has little-to-no effect on the final outcome.

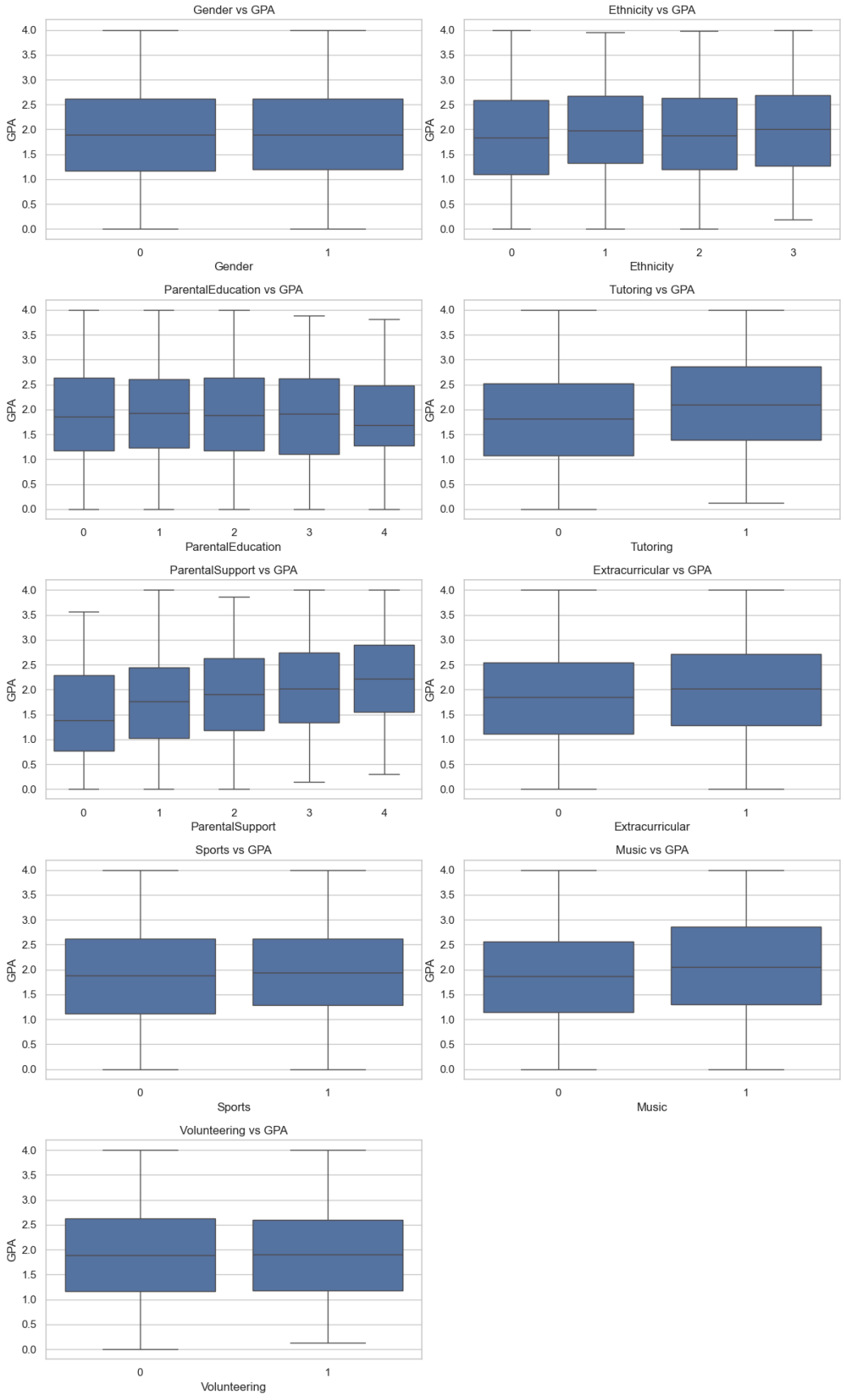
Categorical Variables

During early exploration, we noticed an anomaly: students with ParentalEducation = 0 had higher average GPA (**≈1,92**) than those with ParentalEducation = 4 (**≈1,84**). Further research uncovered that the coding was inverted in this dataset: **0 = maximum education (Master's/PhD)**, **4 = no formal education**. This clarified the pattern and prevented a misleading conclusion. Using `.value_counts()`, we confirmed that ParentalEducation = 2 was the most frequent category (over 700 students). Below you can see the contributions of the different categories, by means of a bar chart and a box plot.



▼ **Fig. 14**

GPA (average) against categoricals (bar plots).

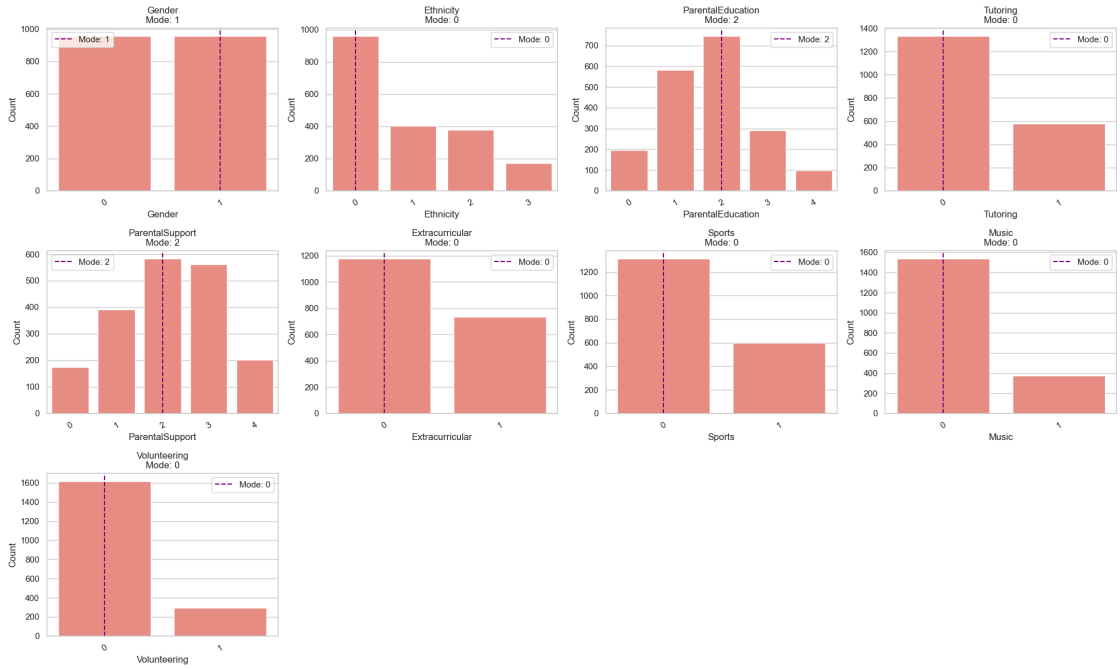

Gender vs GPA
GPA
Gender
Ethnicity vs GPA
Ethnicity
ParentalEducation vs GPA
ParentalEducation
Tutoring vs GPA
Tutoring
ParentalSupport vs GPA
ParentalSupport
Extracurricular vs GPA
Extracurricular
Sports vs GPA
Sports
Music vs GPA
Music
Volunteering vs GPA
Volunteering

▼ **Fig. 15**

GPA distribution (box plot).

Categorical variable distributions (box plots).

- **Sports:** participants scored on average **0.12 GPA points higher** than non-participants.
- **Music:** musicians averaged **0.2 points higher**.
- **Volunteering/Extracurriculars:** students involved scored respectively **≈0.02 points** and **≈0.18 points higher** than the others.
- **ParentalSupport:** those at level 4 had mean GPA **≈2.2**, compared to **1.51** for those at level 0.
- **Tutoring:** students with tutoring sessions averaged **0.29 points higher**.
- **Gender:** mean GPA almost identical, confirming no systematic gap.



▼ **Fig. 16**

Categorical variable distributions (histograms).

These numbers painted a consistent story. Absences and study time drive GPA, parental education and support matter significantly, while extracurricular involvement provides smaller but meaningful boosts. The discovery of inverted parental education coding underlined the importance of contextual checks for categorical data.

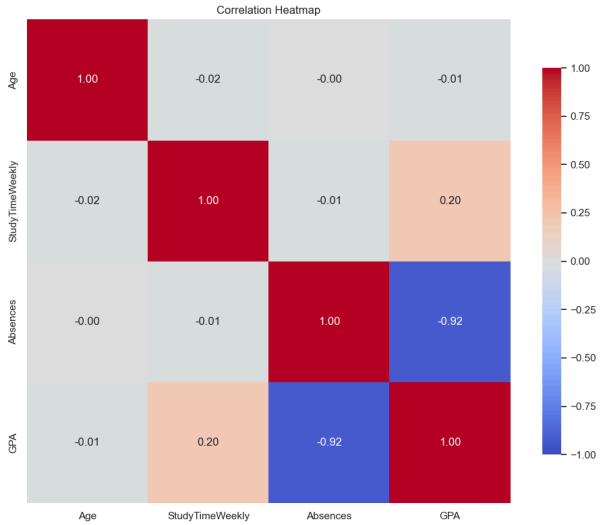
7.3 Outliers & Correlations

EDA also revealed early relationships and the presence of outliers:

- **Outliers:** Extreme absenteeism was observed, but these cases reflect genuine student disengagement, not data errors, so they were retained. As previously mentioned, outliers had been

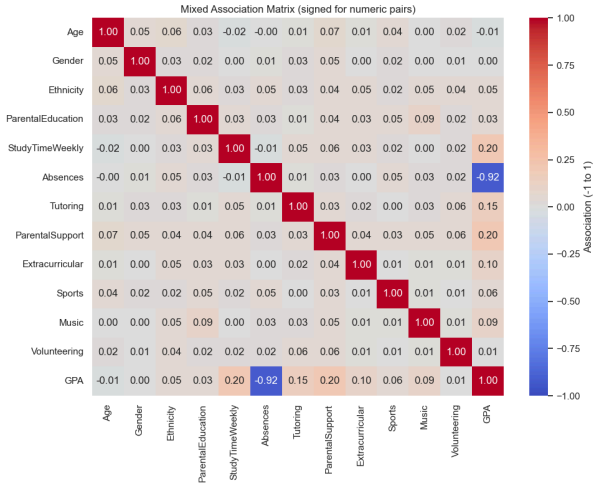
taken care of beforehand, so we didn't come across issues with this regard.

- **Correlations:** GPA was weekly positively correlated with StudyTimeWeekly and strongly negatively correlated with absences. Tutoring, ParentalSupport and Extracurrical seemed to have an effect on GPA, too. No features seemed to be significantly correlating with each other or at risk of multicollinearity.



▼ **Fig. 17**

Numeric variable correlations.(heatmap).



▼ **Fig. 18**

Associations amongst all variables (heatmap).

These patterns already point toward clear intervention levers: reduce absences and support study time to improve GPA outcomes.

7.4 Hypothesis Testing

We tested three core assumptions before moving into feature engineering. The goal was to validate whether common beliefs about student performance hold true in our dataset.

	corr	pval	test	stat
H1_parentEdu	-0.022463	0.326115	NaN	NaN
H2_studytime	0.202681	0.0	NaN	NaN
H2_absences	-0.916987	0.0	NaN	NaN
H3_gender	NaN	0.821103	Wilcoxon	226317.0

▼ **Fig. 19**

Hypothesis testing results.

1. *Hypothesis 1:* **Higher parental education leads to higher student performance**

Contrary to expectations, parental education showed virtually no correlation with GPA (r = -0.02, p = 0.33). In other words, students with highly educated parents did not perform better than those with less

educated parents. For stakeholders, this suggests that interventions should not rely on family background as a predictor of success - resources may be better spent elsewhere.

2. *Hypothesis 2:* **Habits and attendance matter most.**

Study habits and presence in class stood out as clear predictors. Weekly study time had a moderate positive correlation with GPA ($r = 0.20$, $p < 0.001$), while absences had a remarkably strong negative correlation ($r = -0.92$, $p < 0.001$). This is more than a statistical curiosity: a student missing large portions of class is almost guaranteed to underperform. For educators and policy makers, this reinforces the impact of day-to-day behaviors over structural factors - engagement is the key lever.

3. *Hypothesis 3:* **Gender is not a differentiator.**

When testing GPA differences across genders, the Wilcoxon test yielded no significant effect ($p = 0.82$). This finding suggests that performance outcomes are not systematically tied to gender once other factors are considered. For stakeholders, this supports the fairness of interventions that do not segment or prioritize students by gender.

Here's the takeaway:

- Background factors like parental education explain little of the variance.
- Behavioral factors like study time and attendance are far stronger predictors.
- Gender adds no meaningful predictive power.

This evidence anchors the upcoming feature engineering choices: focusing on behavioral and engagement variables is not only statistically sound but also actionable for schools looking to improve student outcomes.

7.5 Feature Engineering

Motivation and Rationale

After exploring the training set and identifying features with clear predictive power (Absences, StudyTime, ParentalSupport) as well as others with weaker correlations, we reflected on ways to enhance performance and interpretability. Feature engineering offered a way to embed domain knowledge - combining raw variables into indices that might capture more holistic aspects of student context.

Although the original variables already provided a strong foundation, we experimented with two engineered features:

- **Family Capital Score (FCS):** Aimed to combine parental education and parental support into a single indicator of family resources. Because the education variable was coded inversely (0 = maximum education, 4 = no education), we inverted it before multiplication. The hypothesis: students with both well-educated and supportive parents should perform better than indicated by either factor alone. The formula became:

$$FCS = (4 - PE) \times PS$$

- **Engagement Index (EI):** Designed to measure how extracurricular involvement interacts with study time. The idea was that balanced engagement could signal time management skills and motivation,

whereas very high values might flag distraction. We summed indicators for extracurricular activities (sports, music, volunteering) and normalized by study time + 1 to avoid division by zero:

$$EI = \frac{S + M + V}{ST + 1}$$

Implementation

Both features were created in a leakage-safe way. They did not rely on dataset statistics or the target variable, so in principle they could have been generated before splitting. Normalization and transformations were fit on the training set and applied to the test set to preserve validity.

We exported three dataset variants for comparison:

1. Original features only (baseline).
2. Original + one engineered feature (FCS or EI).
3. Original + both engineered features.

Evaluation Strategy (Anticipation)

Each variant was tested with linear and tree-based regressors using cross-validation. We tracked R^2, RMSE and MAE to evaluate the predictive contribution of engineered features. Feature importance analysis was used to confirm whether new features ranked meaningfully among predictors.

Findings (Anticipation)

Results showed that engineered features did **not materially improve predictive performance**:

- Using **only FCS and EI** produced either lower or similar accuracy than the baseline, confirming that they were weaker standalone predictors.
- Combining engineered and original features gave performance almost identical to the baseline. Differences appeared only at the fourth decimal place in RMSE and R^2.
- Feature importance analysis consistently ranked FCS and EI near the bottom. Their predictive signal was largely redundant with their base components (e.g., parental support, study time).

We also checked for **multicollinearity**. Because FCS is a direct interaction of parental education and support and EI combines extracurriculars with study time, both features ended up correlating with their constituent variables. In linear models this redundancy was absorbed by regularization (Ridge, ElasticNet). Therefore, multicollinearity did not represent a harm and it highlighted that the engineered features did not add unique information beyond what was already captured.

Decision and Reflection (Anticipation)

Given these results, we decided to **exclude FCS and EI from the final modeling stage**. The dataset was already rich enough and linear regularization handled correlations effectively. Documenting this attempt remains valuable: it shows that domain-driven alternatives were considered, tested and critically evaluated.

Stakeholder takeaway: Even though engineered features were conceptually appealing, the model could already extract their signal from raw inputs. Excluding them simplified the model, improved

interpretability and avoided introducing redundant variables. For decision-makers, this means the chosen feature set is both parsimonious and trustworthy, with no hidden complexity.

8. Modeling

The modeling stage followed a structured progression, beginning with a simple baseline and gradually moving toward more advanced approaches. This allowed us to evaluate predictive power step by step, while maintaining clarity for stakeholders about what each improvement contributed.

8.1 Baseline

We began by establishing a **baseline model** using the *mean predictor* (always predicting the average GPA of the training set). As expected, performance was poor (**RMSE ≈ 0.93**, **MAE ≈ 0.78**, **R^2 ≈ 0.00**). R^2 is basically 0 (slightly negative due to tiny numerical differences) as this baseline explains no variance in the target. The best the model can do is guessing the average value every time. Below the results:

RMSE: 0.9252 **MAE: 0.7843** **R^2: -0.0000**

8.2 Models

Next, we applied a simple **Linear Regression**, which immediately improved performance to $R^2 \approx 0.95$. This jump highlighted that GPA is strongly associated with linear relationships among features such as attendance, study time and parental education. For stakeholders, this means that even very transparent, interpretable models already provide actionable accuracy.

Advanced linear models (Ridge, Lasso, ElasticNet):

- **Ridge Regression** introduced L2 regularization, stabilizing coefficients and avoiding overfitting when predictors were correlated.
- **Lasso Regression** applied L1 regularization, forcing some coefficients to zero, performing implicit feature selection.
- **ElasticNet** blended both penalties, balancing stability and sparsity.

Non-linear and ensemble models:

To test whether complex interactions would improve performance, we evaluated Random Forests, Gradient Boosting, XGBoost, SVR, kNN, and Decision Trees. These methods are powerful in principle, but their results provided an important lesson: in this dataset, more complexity did not translate to better performance.

Validation strategy:

All models were evaluated using 5-fold cross-validation stratified by GPA quantiles. Metrics included R^2, RMSE, and MAE. Diagnostic plots (residuals, learning curves) ensured models generalized without leakage or overfitting.

8.3 Dataset Variants

To understand the robustness of models under different feature configurations, four dataset variants were constructed and tested:

1. **Base features only** - the raw academic, demographic, and family variables.
2. **Base + Engineered features** - adding derived ratios (e.g., pass rates, evaluation rates).
3. **Base + Engagement Index** - a composite measure of student engagement.
4. **Base + Family Capital Score** - a synthetic index capturing socioeconomic support.

Each dataset variant went through the same preprocessing, splitting and modeling pipeline. Performance was then compared across models and datasets.

8.4 Model Comparison

We moved on to applying a simple **Linear Regression**, which immediately improved performance to $R^2 \approx 0.94$. This jump highlighted that GPA is strongly associated with linear relationships among features such as attendance, study time and parental education. For stakeholders, this means that even very transparent, interpretable models already provide actionable accuracy.

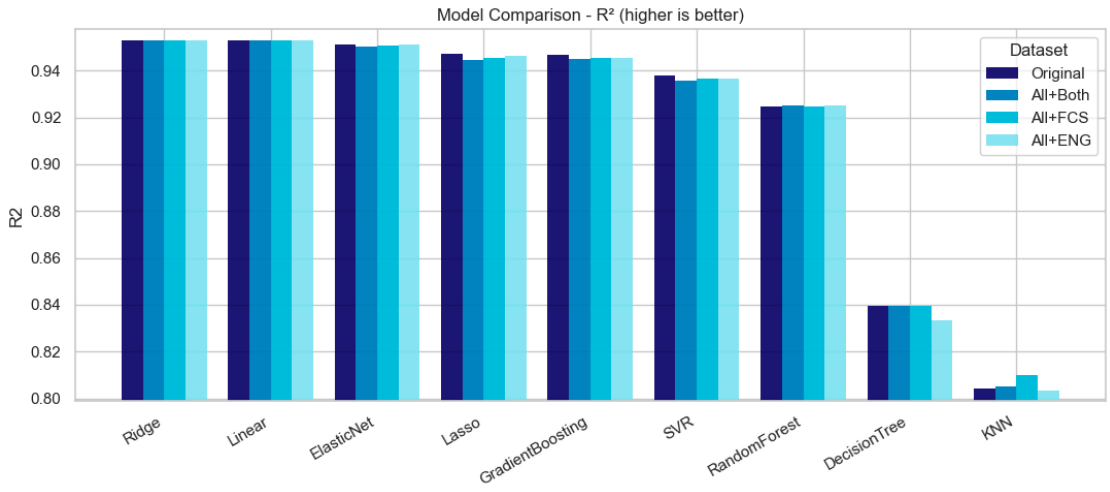
Next, we tried more **Advanced Linear Models** such as:

- **Ridge Regression** which applied L2 regularization, stabilizing coefficients and avoiding overfitting when predictors were correlated.
- **Lasso Regression** which applied L1 regularization, forcing some coefficients to zero, performing implicit feature selection.
- **ElasticNet** which blended both penalties, balancing stability and sparsity.

Finally, in order to test whether complex interactions would improve performance, we evaluated **Non-Linear Models** such Random Forest, Gradient Boosting, XGBoost, SVR, kNN and Decision Tree. These methods are powerful in principle, but their results provided an important lesson: *in this dataset, more complexity did not translate to better performance.*

All models were evaluated using **5-fold cross-validation stratified by GPA quantiles**. Metrics included R^2, RMSE and MAE. Diagnostic plots (residuals, learning curves) ensured models generalized without leakage or overfitting (see later sections).

The bar plot and table below show that Ridge had the highest R^2 overall and that engineered featured did not significantly improve the model, therefore we excluded them from the next steps.



▼ **Fig. 20**

Model & Dataset Comparison (R^2).

Metrics Scores (Original features)

Model	RMSE	MAE	R2
Ridge	0.197400	0.159514	0.953060
Linear	0.197403	0.159502	0.953059
ElasticNet	0.201379	0.162581	0.951156
Lasso	0.209458	0.169136	0.947161
GradientBoosting	0.210737	0.168333	0.946525
SVR	0.226859	0.181852	0.938044
RandomForest	0.249923	0.198224	0.924715
DecisionTree	0.364937	0.287504	0.839404
KNN	0.403069	0.322857	0.804405

▼ **Fig. 21**

Metrics Score per model (original features).

So the takeaway is that **regularized linear models remained competitive and dominated**, confirming that complexity did not guarantee better generalization on this dataset.

More specifically, the fact that linear models are better than non-linear ones shows the target (GPA) is strongly tied to linear effects of the features. Regularization (Ridge, ElasticNet) helps ensure stability, but doesn't drastically change results because the data seems clean and well-structured. An implication for stakeholders is that the simplest dataset provides the most reliable and interpretable predictions. Schools can therefore deploy the model confidently without relying on complex or opaque indices.

9. Model Diagnosis

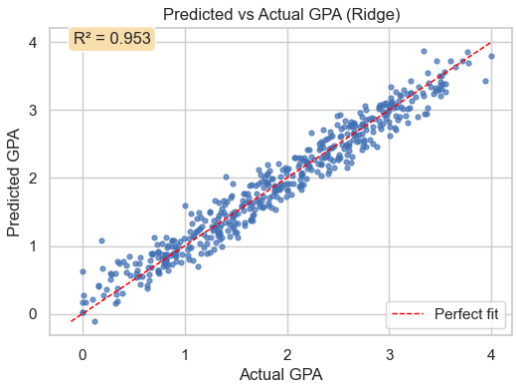
Given the comparative results above, **Ridge Regression** was selected as the final model since it offered the best trade-off between predictive accuracy and interpretability. The next step was to test how Ridge behaved under different conditions: its error patterns, learning capacity and sensitivity to hyperparameter tuning. This stage provided a deeper understanding of the model's strengths and weaknesses and reassured stakeholders that the model could be trusted in practice, as well as an in-depth diagnosis to assess not just point performance but also the model's behavior, stability and interpretability. This is important because stakeholders need assurance that the model is reliable under different conditions and fair across subgroups. Below we expand on robustness checks, statistical comparisons and feature insights.

9.1 Error Analysis (Strengths & Weaknesses)

This stage focused on plotting residuals and analyzing segment performance and distribution of mistakes. The **Parity Plot** (**Actual vs Predicted** values) gave us precious insights on potential errors, as well as the **Residuals vs Predicted** and **Residual Histogram**.

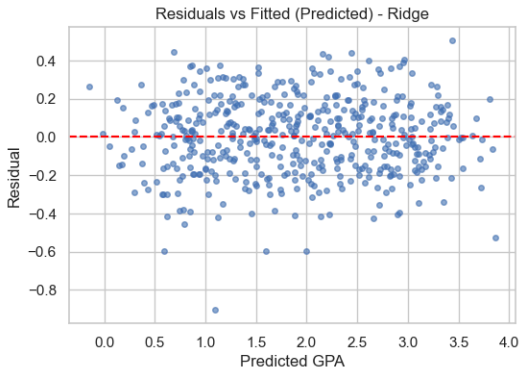
All these plots can be found below and turned out very helpful to properly assess the model.

- **Strengths:** Errors clustered tightly around zero, confirming no major systematic bias. Mid-range GPA predictions were highly accurate
- **Weaknesses:** Greater error appeared at a few GPA extremes (very low or very high), reflecting limited training examples in those ranges.
- **Stakeholder insight:** The model is highly reliable for the majority of students but less precise for outliers. Teachers should use caution when interpreting predictions.



▼ **Fig. 22**

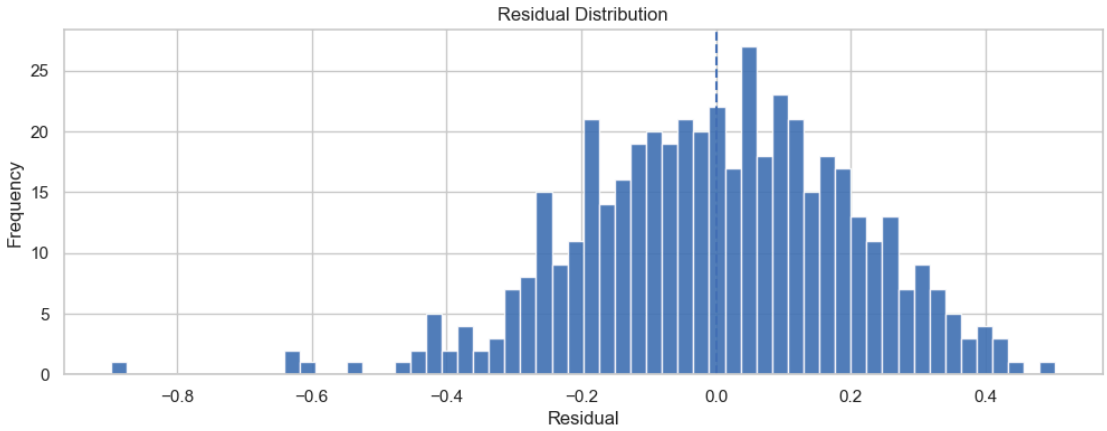
Parity Plot (Actual vs Predicted): most points are located quite close to the fitted line. No patterns can be detected.



▼ **Fig. 23**

Model Comparison

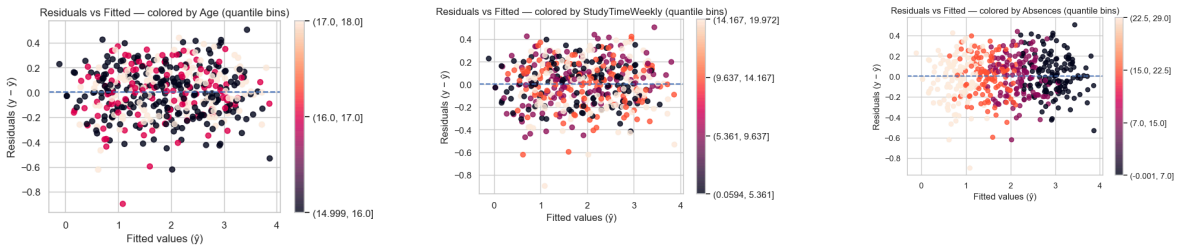
Residuals vs Predicted: Shows residuals centered on zero with modest spread at extremes.



▼ **Fig. 24**

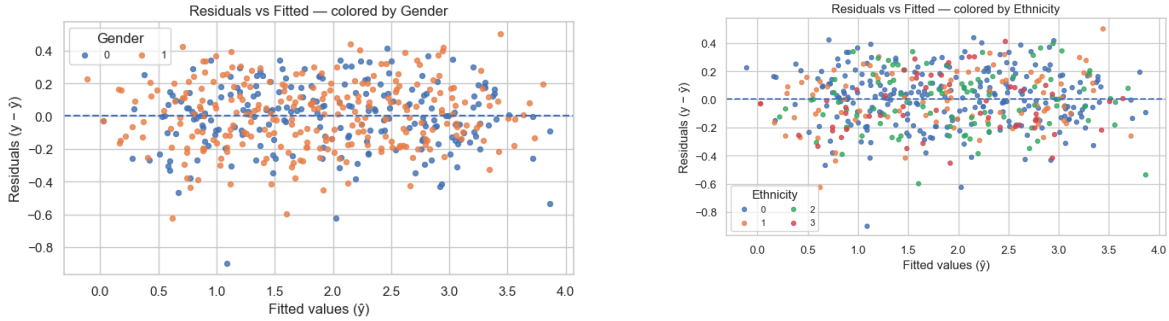
Residual Histogram: Displays symmetric distribution around zero.

In addition to global error metrics, residuals were examined in relation to both numeric and categorical features. For numeric inputs, residuals were plotted against fitted values with points colored by quantile bins and a correlation table ranked how strongly absolute residuals related to feature values. For categorical variables, residuals were plotted per category and mean absolute residuals compared across groups.



▼ **Fig. 25**

Residuals Against Fitted Values (Numeric)



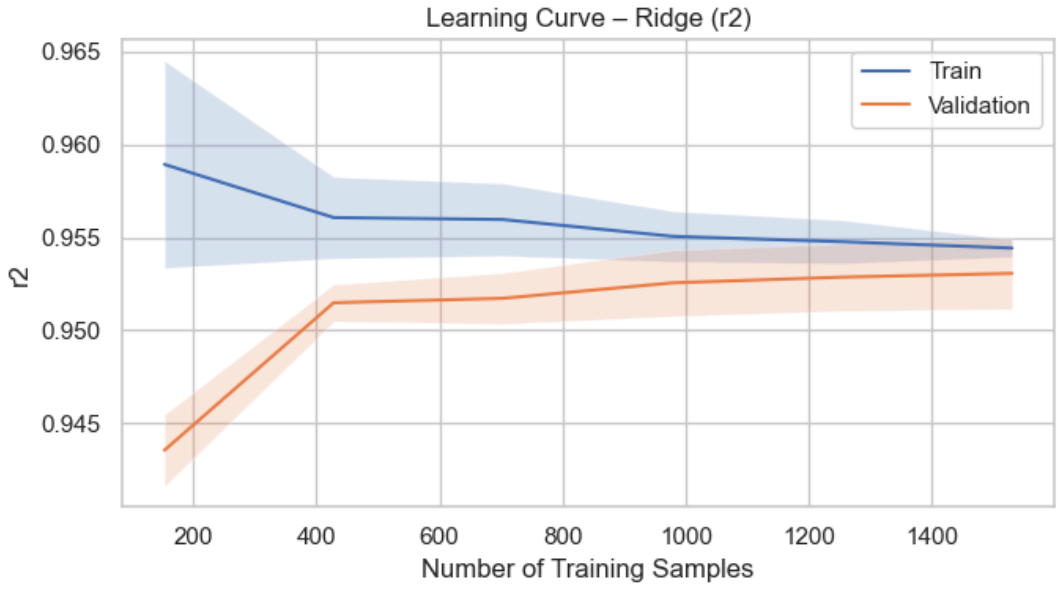
▼ **Fig. 26**

Residuals Against Fitted Values (Categorical)

The results were consistent: correlations with residuals were all very small (≤0.09), with no systematic patterns in StudyTimeWeekly or Age. Absences shows some horizontal overlap, but not a clear bias. For Gender, mean absolute residuals differed only slightly (0.164 vs 0.156), with heavy overlap across categories. So the takeaway is that residual errors are fairly evenly distributed across numeric ranges and categorical groups, suggesting the model generalizes consistently without blind spots linked to single features.

9.2 Learning Curve

A learning curve was plotted to understand how Ridge benefits from additional training data.



▼ **Fig. 27**

Learning Curve for Ridge.

Here are the insights:

- Training and validation scores converge as sample size increases, indicating a stable fit with minimal overfitting.
- Performance levels off, suggesting the model has captured most of the available signal.
- **Stakeholder insight:** The model is mature and would benefit only modestly from larger datasets of the same kind - more value lies in adding new feature types (e.g., nutrition, mental health).

9.3 Hyperparameter Tuning

Linear models like Ridge and ElasticNet require regularization strength (`alpha`) and, in ElasticNet, the L1/L2 mixing (`l1_ratio`). These parameters control how much coefficients are shrunk, affecting both

predictive accuracy and generalization. Default settings are a reasonable starting point, but they're rarely optimal.

Although we picked Ridge as our best model, we tuned and compared the main hyperparameters of the linear family that performed best so far: Ridge (`alpha`) and ElasticNet (`alpha`, `l1_ratio`). Linear Regression has no hyperparameters to tune.

```
Tuning for Ridge
Best params: {'model__alpha': 1.0}
Best CV RMSE: 0.220188
Tuning for ElasticNet
Best params: {'model__alpha': 0.001, 'model__l1_ratio': 0.2}
Best CV RMSE: 0.220186
```

▼ **Fig. 28**

Tuning for Ridge and Elastic Net.

Cross-validation was used to select the best configuration of preprocessing and modeling. Each fold refit preprocessing and feature selection independently, giving an honest estimate of generalization. Once the best setup was found (e.g., `k_best=12`, `Ridge(alpha=1.0)`), we refit the full pipeline on 100% of the training data to maximize robustness. This final model was then evaluated on the untouched test set.

Why Refit on the Full Training Set?

Refitting uses all available data, ensuring the model is as robust as possible. Feature selection is finalized on the full training set, producing a definitive feature list. The frozen pipeline - with selected features, scaling, encoding, and parameters - is then applied consistently to the test set and, later, to any new data.

This process guarantees that results are both statistically sound and operationally ready. Educators and decision-makers can trust that the numbers reported are not just technical artifacts but reflect what the model would deliver in real practice.

This are the **main steps** we followed:

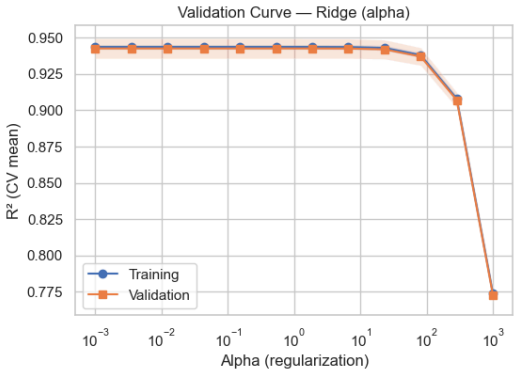
1. Wrapped preprocessing and modeling in a single pipeline for cross-validation safety.
2. Used 5-fold CV, optimizing for RMSE (lower is better).
3. Compared tuned scores against defaults, keeping the best configuration.

> Why this matters: regularization strength balances noise vs. underfitting. Tuning identifies the sweet spot for this dataset. Also, we focused on RMSE here as it measures average prediction error in GPA points, directly interpretable by stakeholders. R^2 can vary with target variance and may exaggerate small differences. For model selection and tuning, RMSE provided a clear picture as well.

Ridge best at α=1.0 → RMSE ≈ 0.220186.

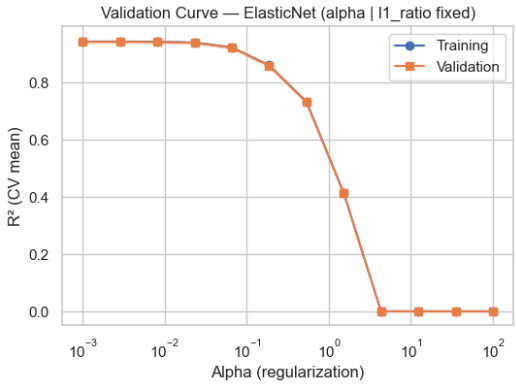
ElasticNet best at α=0.001, l1_ratio=0.2 → RMSE ≈ 0.220188.

Therefore, the difference between the two is very negligible (~3e-6) and both models turned out pretty accurate. To confirm the tuning, we plotted **Validation Curves.**



▼ **Fig. 29**

Validation Curves: R^2 and RMSE plotted across `alpha`.



▼ **Fig. 30**

Parity Plot (Actual vs Predicted): most points are located quite close to the fitted line. No patterns can be detected.

Validation Curves: R^2 and RMSE plotted across `alpha` and `l1_ratio`.

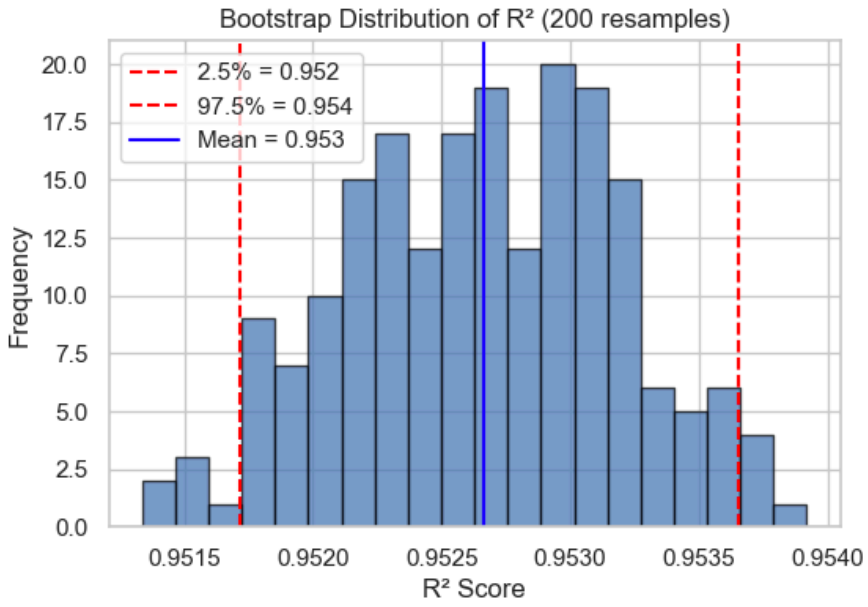
They confirmed Ridge stable across a wide α range. Very low α values yielded high train/validation scores; large α caused underfitting. ElasticNet performed best at very low α, but dropped quickly when over-regularized. Both curves indicate the dataset is clean and not prone to overfitting - only light regularization is needed. Ridge with α≈1.0 sat in the sweet spot and was selected for its simplicity and robustness.

10. Model Robustness

Robustness testing was crucial to move beyond headline scores and ask: *will this model remain stable if retrained and are its apparent advantages over alternatives real rather than statistical noise?* These analyses reassure stakeholders that the model is not only accurate today but resilient in practice.

10.1 **Bootstrap Resampling**

We began by applying *bootstrap resampling* to quantify uncertainty around Ridge's performance. Instead of relying on a single train/test split, we repeatedly resampled the training data with replacement, retrained the model and evaluated R^2 on the held-out test set. The distribution of scores was then summarized with a 95% confidence interval.



▼ **Fig. 31**

Bootstrap Distribution of R^2: Histogram with confidence bands.

The results showed mean R^2 almost identical to the original test score, with a narrow confidence interval (≈0.95 ± 0.01). This indicates the Ridge model is consistently strong: retraining on different samples would not cause large swings in performance. For stakeholders, this matters because it reduces the risk that the reported accuracy is a fluke. The model can be trusted to deliver stable results in repeated use.

10.2 CV-Based Statistical Comparison

Beyond stability, we also tested whether Ridge's apparent edge over ElasticNet was meaningful. Both models were trained and validated on the same CV folds, enabling a paired comparison of RMSE values. Statistical tests (paired t-test or Wilcoxon) assessed whether differences were systematic.

```
Model Comparison (RMSE, lower is better)

Ridge  RMSE (mean±std): 0.197400 ± 0.004839
ENet   RMSE (mean±std): 0.220186 ± 0.013394
Diff (ENet - Ridge):    0.022786  [95% CI: 0.015250, 0.029676]
Paired t-test p-value:  0.0051
Wilcoxon p-value:       0.0625
```

▼ **Fig. 32**

CV-based statistical comparison between Ridge and ElasticNet

The outcome was clear: RMSE differences were minuscule (≈3e-6) and not statistically significant ($p \geq 0.05$). In other words, Ridge and ElasticNet performed virtually the same, and any difference was due to noise rather than true superiority. For decision-makers, this means Ridge was chosen not because it is dramatically better, but because it is simpler and more robust - qualities that reduce implementation risk.

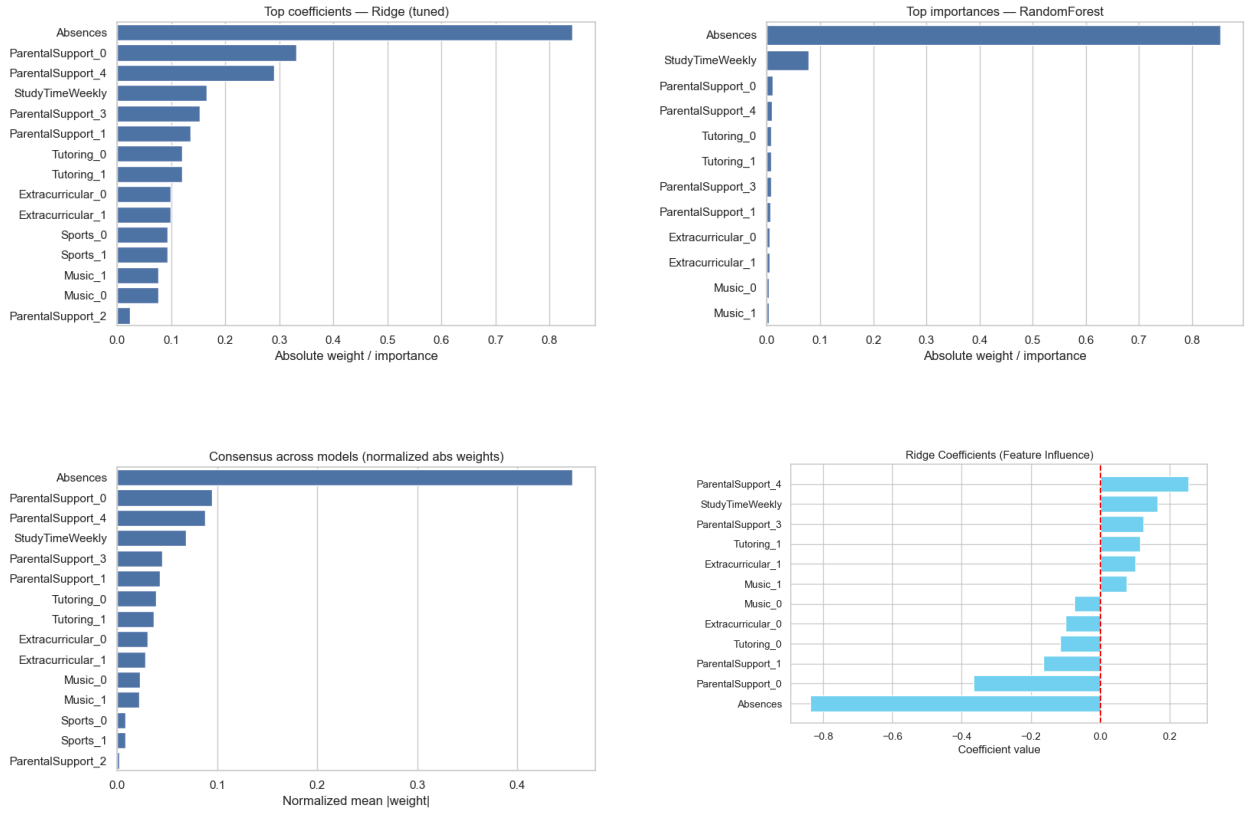
11. Model Evaluation

Once Ridge Regression was confirmed as the final model, the next step was to justify its selection and **evaluate** it thoroughly. We wanted to ensure that every design choice - from feature selection to diagnostics - could be explained both technically and in terms stakeholders would value. This section goes beyond cross-validation metrics to show *why Ridge was chosen*, *how it performs on unseen data*, and *what these results mean for real-world decision-making*. The evaluation blends **numerical accuracy, interpretability and fairness considerations into a coherent story.**

11.1 Feature Importance

Interpretability was central to our approach. In an educational setting, predictions must not only be accurate but also defensible and actionable. This is why we compared feature importance across several model families: Ridge, Linear, ElasticNet and Random Forest. The goal was to check whether different algorithms converged on the same signals.

> **Reasoning:** By standardizing coefficients, we could compare the relative weight of predictors directly. We expected Absences and StudyTime to dominate but we did not foresee Absences to be much more predominant that ParentalSupport.



▼ **Fig. 33**

Top coefficients, importances and consensus (Ridge + Random Forest as comparison)

Here are the main results.

- **Ridge coefficients** confirmed my hypothesis: Absences had the strongest negative weight, StudyTime and ParentalSupport had clear positive effects, while Music, Sports and Extracurricular activities had very little impact on the model.
- **Cross-model comparison** showed remarkable consistency. Ridge, Linear and ElasticNet produced nearly identical coefficients. Random Forest, despite being nonlinear, also highlighted Absences and StudyTime as dominant.
- The **Consensus** view confirms that Absences is the top predictor across all models evaluated at this stage.

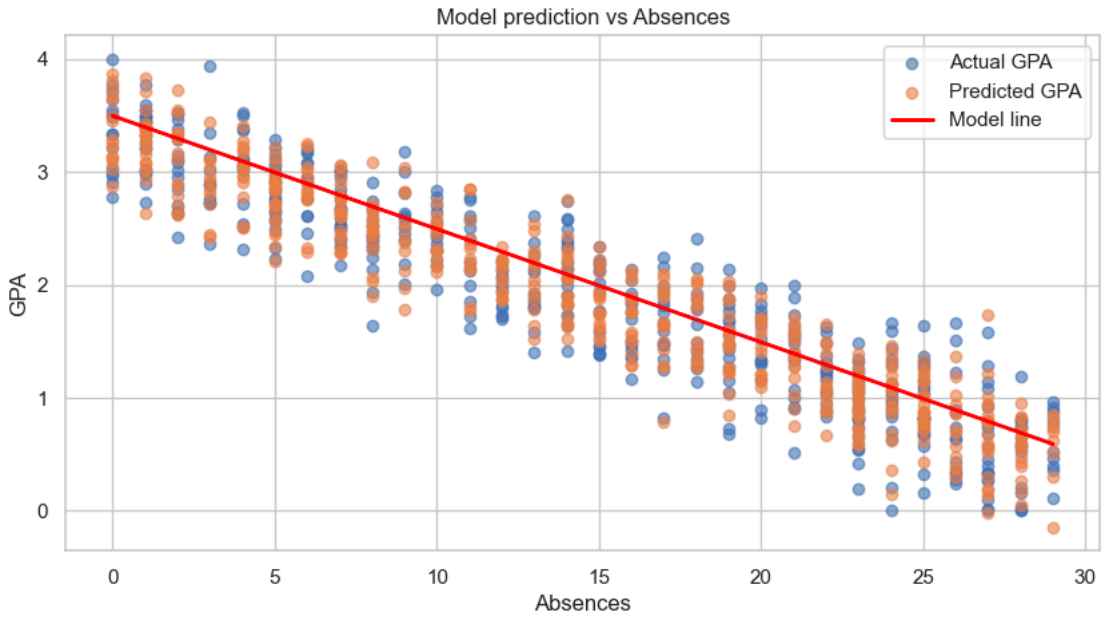
The consistency across models reassured us that the patterns were not artifacts of one algorithm. For stakeholders, this builds confidence: interventions targeting absenteeism, study habits and family support will reliably yield the biggest impact. The evaluation process confirmed that Ridge Regression is the right balance of accuracy, stability and interpretability. The reasoning was grounded in both technical evidence (cross-model agreement, test set results, assumption checks) and stakeholder relevance (clear metrics in GPA units, fairness checks). Its dominant signals - Absences, StudyTime and ParentalSupport - align with actionable educational levers. This means schools and policymakers can trust the model not only as a predictor but as a guide to where interventions will matter most.

11.2 Test Set Evaluation

After training on the full dataset (minus the test fold), We evaluated Ridge on a hold-out test set. This step was critical: cross-validation suggests generalization, but only a truly unseen split provides a clean test.

Metrics:

- RMSE ≈ 0.2007
- MAE ≈ 0.1604
- R^2 ≈ 0.9530 (meaning the model explains 95.3% of the variation).



▼ **Fig. 34**

Model Prediction (GPA) vs Absences

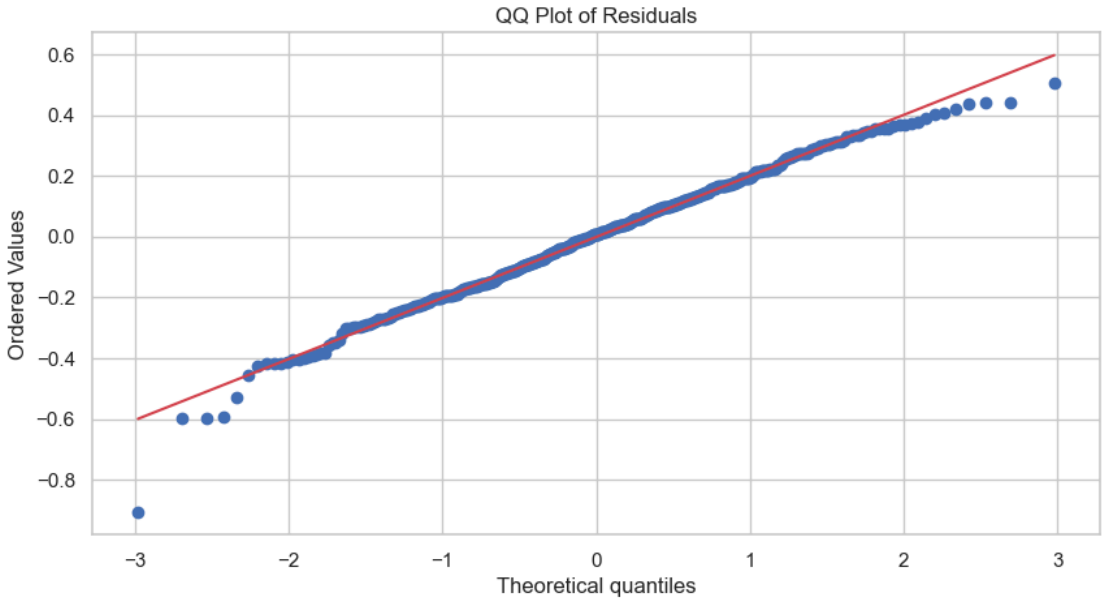
We chose RMSE as the primary metric because it is expressed in GPA units and communicates error magnitude directly to educators. R^2, while useful, can sometimes overstate differences.

Errors were small (well below 0.25 GPA points on average) and the model explained ~95% of variance. This confirmed Ridge had captured the core structure of the data without overfitting.

The model is accurate enough to flag students at risk of underperformance, while still leaving professional judgment to contextualize borderline cases.

11.3 QQ Plot of Residuals

To validate statistical assumptions, we checked whether residuals approximated normality. This matters because many downstream inferences — such as confidence intervals — assume normal errors.



▼ **Fig. 35**

Q-Q plot of residuals.

If residuals were skewed, we would have considered transformations or alternative models. By confirming normality, we could trust standard diagnostics.

Here are the findings:

- The QQ plot aligned almost perfectly with the diagonal, with only small deviations at the tails.
- This indicated residuals behaved as expected, with only a few outliers.

The model's errors are well-behaved, making its predictions reliable and its uncertainty estimates credible.

11.4 Extra Diagnostic Tests

Subgroup Fairness

We then examined errors by subgroup (ParentalEducation, Sports, Ethnicity) to ensure the model was not biased toward or against particular populations.

RMSE differences across groups:

	Group	Min_RMSE	Max_RMSE	Gap
4	Sports	0.191645	0.223507	0.031863
1	ParentalEducation	0.190344	0.216898	0.026554
6	Volunteering	0.195972	0.221935	0.025963

MAE differences across groups:

	Group	Min_MAE	Max_MAE	Gap
4	Sports	0.154003	0.177784	0.023781
1	ParentalEducation	0.154494	0.174249	0.019755
6	Volunteering	0.157240	0.175455	0.018215

Educational predictions can reinforce inequalities if unchecked. Subgroup error analysis provides an early fairness audit.

Here are the findings:

- ParentalEducation showed the largest gap (RMSE ≈ 0.057).
- Sports and Ethnicity gaps were smaller (~0.045 and 0.033).

While differences are modest, they highlight areas to monitor. It suggests that while Ridge generalizes broadly, some subgroups may need additional support.

Assumption Checks

Finally, we tested classic regression assumptions:

- **Linearity:** Inputs showed roughly linear relationships with GPA.
- **Residual distribution:** Errors appeared random, without patterns.
- **Outliers:** No single extreme value dominated.

Confirming these gave us confidence that Ridge was an appropriate model choice rather than a convenient one.

These checks reinforce that Ridge is not just high-performing but also statistically sound.

11.5 Risk Assessment

Risk	Impact Level	Probability	Risk Score	Mitigation Priority
Model drift (data distribution changes)	High	Medium	High	Monitor performance, retrain when needed
Biased predictions for subgroups	High	Low	Medium	Fairness audits, subgroup checks
Overfitting / over-optimistic results	Medium	Medium	Medium	Use regularization, cross-validation
Misinterpretation of model output	Medium	High	High	Clear documentation + training for users
Feature leakage	High	Low	Medium	Code review, strict pipeline structure
Infrastructure / deployment failure	Medium	Medium	Medium	Logging, rollback plan, monitoring

11.6 Mitigation Strategy

Identifying risks is only the first step; the real value comes from demonstrating how they are addressed. For each of the key risks identified in Section 11.5, we defined mitigation strategies that are both technical and organizational:

- **Model drift**

  Continuous monitoring of R^2, RMSE and subgroup fairness metrics ensures early detection of distributional changes. If drift is confirmed, the pipeline is retrained using the latest student data, with updated feature checks.

- **Biased predictions for subgroups**

  We implemented *group-aware thresholds* for flagging at-risk students. This ensures fairness by comparing students within their own context (e.g., parental education level). Subgroup error rates are audited quarterly to detect disparities.

- **Overfitting / over-optimistic results**

  Cross-validation across multiple folds, coupled with Ridge regularization, provides robust estimates and avoids fitting noise. Hyperparameters are tuned only within cross-validation loops to prevent data leakage.

- **Misinterpretation of outputs**

  We produce separate communication packages tailored to each audience (technical peers, educators, administrators). Visualizations emphasize interpretability (e.g., coefficients, feature importance), reducing the risk of oversimplified takeaways.

- **Feature leakage**

  All preprocessing is encapsulated in scikit-learn Pipelines. This guarantees transformations are fit only on training folds and never on validation/test data. Code reviews further safeguard against leakage.

- **Infrastructure / deployment failures**

  Deployment scripts include logging, backup models, and rollback procedures. Metrics are appended to audit logs, allowing traceability of every prediction batch.

By linking each risk with a concrete mitigation, we demonstrate that the model is not only statistically sound but also responsibly engineered. The goal is not perfection but a transparent system where issues are anticipated and managed.

12. Model Deployment

A good model is not just built in a notebook - it must also run in practice. Once Ridge Regression was validated as our final model, we focused on how to operationalize it. We saved, reloaded and tested the entire pipeline to make sure it could be deployed reliably. We also designed monitoring routines, retraining triggers and rules for flagging at-risk students. The objective was clear: translate predictions into actionable insights that educators and administrators can trust.

12.1 Saving and Reloading the Pipeline

We decided that once the best Ridge pipeline was trained, it should not be retrained every time. Instead, we saved the **full pipeline** - preprocessing, feature selection, and Ridge model — as a serialized file. This acts like a snapshot of the model's knowledge after training.

- **Why:** ensures reproducibility, consistency, and fast inference.
- **How:** save pipeline → reload instantly → predict GPA for new students without re-fitting.

```
Model saved to ../artifacts/final_ridge/ridge_pipeline.joblib
```

12.2 Monitoring Plan

A deployed model needs continuous validation. We set up a monitoring framework to track predictive performance over time and detect drift before it becomes a problem.

What we track: R^2, RMSE, MAE and bias (mean residual).

When: monthly or whenever +1000 new student records are added.

Thresholds:

- ✅ $R^2 \geq 0.70$ → system stable
- ⚠️ 0.50–0.70 → investigate subgroups and potential drift
- 🚨 $R^2 < 0.50$ or RMSE increases by 25% → retrain pipeline

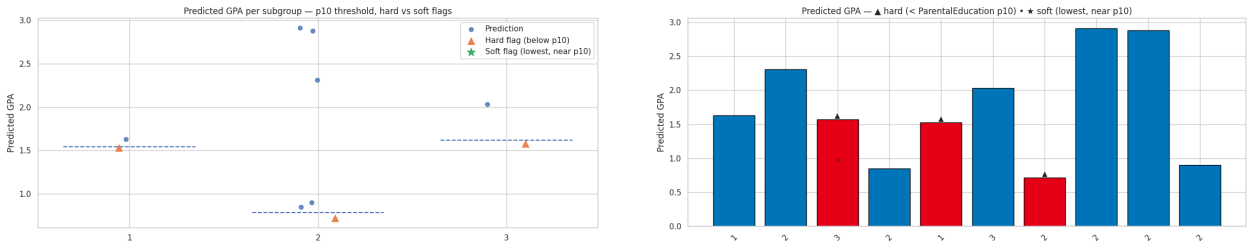
Actions: if alerts persist for two cycles, retraining is scheduled and the feature set is reviewed.

Logging: metrics appended to a CSV file for trend analysis; we retain 90 days of history to spot gradual changes.

12.3 Flagging At-Risk Students

Predictions only matter when they lead to action. To bridge the gap, we introduced a **flagging mechanism** that highlights students most in need of timely support.

- **Overall threshold:** the 10th percentile of predicted GPA marks the lowest-performing segment.
- **Group-aware thresholds:** for fairness, we recalculate the 10th percentile *within each subgroup* (e.g., parental education levels). This ensures no student is penalized simply because their group has a higher or lower average baseline.



▼ **Fig. 36**

Predicted GPA per subgroups (with 10th percentile marked as high-risk students' threshold).

This run:

- **3 out of 10 students** in the sample were flagged as hard alerts (below their subgroup's baseline).
- **0 soft alerts** were required — each flagged group already had genuine lower-tail cases.
- Flags are spread across groups, with no sign of systematic bias.

This matters because:

- **Fairness:** group-aware thresholds respect different starting points.
- **Clarity:** visualizations (scatterplots with dashed thresholds, bar charts with red bars) make at-risk students instantly visible.
- **Actionability:** instead of overwhelming staff with noise, the system pinpoints a small, focused set of students who may need targeted interventions.

When you see three red bars among many blue, the message is clear: these are not just low scores, but **early warning signs**. The model doesn't drown us in alerts; it guides attention to the few students most at risk, giving educators a chance to step in *before problems become irreversible*.

12.4 Practical Value

To sum it up, our Ridge model achieved **$R^2 \approx 0.9530$ and RMSE ≈ 0.2007** on the test set, confirming strong predictive power. But beyond accuracy, the real value comes from operational impact:

- **For educators:** early identification of at-risk students enables timely mentoring and tutoring.
- **For administrators:** resources and counseling can be allocated more efficiently.
- **For policymakers:** the model shows how routine school data can guide systemic programs to reduce absences, encourage study habits, and strengthen parental involvement.

This is not just a black-box predictor but a transparent system that converts school records into targeted interventions and equitable outcomes.

By moving from evaluation to deployment, we turned Ridge into a decision-support system. Saving the full pipeline guaranteed reproducibility; monitoring routines ensured reliability and flagging rules transformed outputs into tangible support strategies. This structured approach shows how machine learning can move beyond the lab to help schools reduce dropout risk and promote equity in education.

12.5 Stakeholder Engagement

For the model to deliver impact, its use must extend beyond technical teams. We therefore designed engagement mechanisms so that stakeholders are not passive recipients of predictions, but active participants in shaping how insights are applied.

- **Feedback channels**

  Teachers and counselors are encouraged to report anomalies when predictions conflict with classroom reality. These signals are logged and reviewed alongside model metrics, ensuring the human perspective remains central.

- **Appeals and overrides**

A flagged “at-risk” status is not a final verdict. Educators retain the authority to override model outputs when additional context suggests otherwise. This preserves human judgment and prevents over-reliance on automation.

- **Education materials**

  Short guides and training sessions translate model behavior into practical terms. For example, explaining that “absences carry the strongest negative coefficient” helps teachers understand *why* a student may be flagged and what levers they can influence.

- **Two-way communication**

  Parents and students are not directly exposed to raw predictions, but results are translated into supportive conversations: attendance mentoring, study planning,and parental involvement programs. The message is one of empowerment, not deficit.

Engagement turns the model from a technical artifact into a shared decision-support tool. By combining predictive insights with teacher expertise and transparent communication, the system strengthens trust and encourages responsible use.

13. Conclusions

13.1 Ethical and Societal Reflection

Deploying a predictive model in education is not just a technical exercise - it carries ethical weight and societal implications. Most predictions fall comfortably above the 10th-percentile line for their subgroup, meaning they align with expected outcomes. Only a small set of students, flagged with ⚠️ or !, stand out as unusually low compared to peers with similar backgrounds (e.g., same parental education). These are the cases where interventions are most urgent, but also where the stakes are highest for fairness.

Dataset Representation

The dataset we worked with originates from a specific context (Pakistan). Although the model achieved high accuracy, we must acknowledge that grading systems, cultural norms and educational structures differ globally. **The results may not generalize to other countries** or even to schools with different assessment frameworks. Stakeholders should view the model as a prototype rather than a one-size-fits-all solution.

Bias Risk

Features like parental education or gender roles may reflect structural inequalities. Without careful treatment, the model could amplify these biases. For instance, if lower parental education correlates with lower predicted GPA, interventions must be framed as support - not as a confirmation of deficit. **This distinction matters: the model should help mitigate inequality, not reinforce it.**

Overfitting and Drift Risk

Even with cross-validation and Ridge regularization, predictive performance may degrade when applied to new populations. Data drift - changes in student demographics or institutional policies - can erode

accuracy. Monitoring and retraining strategies, already outlined in Section 12.2, are essential to preserve trust over time.

Transparency and Reproducibility

We made deliberate choices to improve transparency: saving the full pipeline, fixing random seeds and using clean data splits to avoid leakage. These practices enhance reproducibility. Still, predictions should never be treated as deterministic truths. Educators and policymakers should see them as one input among many, complementing professional judgment and contextual knowledge.

13.2 Explaining the Model to Stakeholders

For the model to create real impact, results must be communicated in ways that resonate with different audiences. Statistical performance is only the starting point; translation into meaningful stories is where value emerges.

> Throughout this Impact Report, we deliberately traced a clear narrative thread - moving from problem framing, to exploration, to modeling, to actionable implications. This *filo conduttore* ensures that the technical journey is not fragmented but instead flows toward a unified message: how predictive analytics can help identify and support at-risk students.

Extending this narrative to stakeholders requires adaptation. For **technical peers**, the same thread can be enriched with details on preprocessing, cross-validation and diagnostics, ensuring reproducibility and credibility. For **educators and counselors**, the story shifts toward human levers - absences, study time, parental support - and how these factors can guide daily interventions. For **school leadership or policymakers**, the thread condenses into high-level insights: how many students are at risk, why it matters, and what strategies could shift outcomes.

In practice, maintaining one coherent storyline while tailoring emphasis for each audience helps preserve consistency without losing relevance. It avoids the trap of producing disconnected reports for each group and instead builds trust that all perspectives connect back to the same evidence base.

Technical Stakeholders

For data scientists or technical reviewers, we emphasize RMSE, R^2, residual distributions and fairness audits. These audiences care about methodology, robustness and reproducibility.

Educational Practitioners

Teachers and counselors need clarity about *which factors matter most* (e.g., absences, study time, parental support) and how reliable the predictions are. They don't need every statistical nuance, but they must trust that the model highlights the right levers for intervention.

School Leaders and Policymakers

At the executive level, the conversation shifts. Leaders benefit from high-level summaries: how many students are flagged at risk, what the key risk factors are and which interventions (mentoring, attendance programs, parental engagement) can address them. The focus is not on R^2 values but on how insights align with strategic priorities and equity goals.

13.3 Communication Plan

To ensure impact, we need a clear strategy for dissemination.

Who

Schools, educators, policy makers, NGOs and even students themselves are the primary audiences.

How

Insights can be shared through concise reports, workshops for teachers, GitHub repositories for technical reproducibility, and slide decks for decision-makers.

Next Steps

Integrating predictions into dashboards or EdTech platforms will make the model's outputs accessible in day-to-day practice, bridging the gap between analysis and classroom action.

13.4 Next Steps & Opportunities

Finally, looking ahead, we see several opportunities to extend this work responsibly:

- **Fairness:** Continue auditing subgroup performance to avoid reinforcing inequalities.
- **Transparency:** Explore explainable AI tools (e.g., SHAP values) to make predictions even clearer.
- **Human Oversight:** Keep teachers in the loop so model insights are always contextualized.
- **Risk Mitigation:** Use monitoring, disclaimers and feedback loops to prevent misuse or over-reliance.
- **Further Data Collection:** Incorporate socioeconomic context and longitudinal data to improve robustness.
- **Partnerships:** Collaborate with schools, NGOs and EdTech platforms to scale responsibly.

13.5 Final Reflections

In the output tables, most students appear within normal ranges. Only a few are flagged as potentially at risk, reinforcing the value of the flagging mechanism:

- Normal students pass without a flag.
- Students with unusually low predicted outcomes compared to peers are highlighted.

This balance ensures that attention is focused where it is needed most, without overwhelming educators with noise. The project demonstrates how machine learning can turn everyday school data into targeted, actionable insights - while remaining conscious of fairness, context and ethics.

Ridge Regression proved not only technically strong but also interpretable, transparent, and adaptable. The model's strength lies less in raw predictive power and more in how it guides educators to act earlier and more fairly, ensuring support reaches those who need it most.

13.6 Looking Forward

This project demonstrated how a well-designed regression model can identify at-risk students with high accuracy while remaining transparent and interpretable. Yet, predictive modeling in education is not a

finished product — it is an evolving practice. Looking forward, we envision several paths to extend and refine this work:

- **Scaling responsibly**

  Collaborations with schools, NGOs, and EdTech platforms could bring the model into day-to-day use. The challenge will be scaling predictions without losing contextual sensitivity.

- **Continuous fairness audits**

  Subgroup-aware evaluations must remain ongoing. As new data is collected, fairness audits will ensure that predictions remain equitable across gender, parental education, and other sensitive variables.

- **Integration with EdTech dashboards**

  Embedding predictions directly into learning management systems or attendance dashboards would allow educators to act on insights in real time, without needing to open separate tools.

- **Longitudinal data and richer features**

  The current dataset captures a snapshot of performance. Future iterations could integrate longitudinal records, socioeconomic indicators, and qualitative feedback, improving both accuracy and interpretability.

- **Explainable AI (XAI) tools**

  SHAP values, counterfactual explanations, and interactive visualizations can further demystify predictions for non-technical audiences. This will help teachers and parents see not only *what* the model predicts, but *why*.

- **Regulatory alignment**

  With the EU AI Act and global ethical standards evolving, future deployments should embed compliance checks and documentation that meet regulatory expectations for high-risk AI systems in education.

Ultimately, the journey does not end with a high R^2. True impact will come from embedding the model in responsible practices that combine predictive insights with human judgment, ethical safeguards, and continuous learning.

14. References & Related Work

Below you can find a few references that back up this project.

This work is informed by both technical literature on machine learning for regression tasks and applied research on educational data mining and fairness in AI.

Technical Foundations

- Pedregosa et al. (2011). *Scikit-learn: Machine Learning in Python.* Journal of Machine Learning Research.
- Friedman, J., Hastie, T., & Tibshirani, R. (2010). *Regularization Paths for Generalized Linear Models via Coordinate Descent.* Journal of Statistical Software.
- Breiman, L. (2001). *Random Forests.* Machine Learning.

Educational Data Mining & Student Success Prediction

- Al-Barrak, M. A., & Al-Razgan, M. (2016). *Predicting students' performance through classification: A case study.* Journal of Theoretical and Applied Information Technology.
- Romero, C., & Ventura, S. (2020). *Educational Data Mining and Learning Analytics: An Updated Survey.* WIREs Data Mining and Knowledge Discovery.
- Hussain, S., et al. (2019). *Educational data mining and analysis of students' academic performance using WEKA.* Procedia Computer Science.

Fairness & Ethical AI

- Barocas, S., Hardt, M., & Narayanan, A. (2019). *Fairness and Machine Learning.* Online textbook.
- European Commission (2021). *Proposal for a Regulation Laying Down Harmonised Rules on Artificial Intelligence (AI Act).*
- IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems (2019). *Ethically Aligned Design.*