AKADEMIA GÓRNICZO-HUTNICZA IM. STANISŁAWA STASZICA W KRAKOWIE
Wydział Inżynierii Mechanicznej i Robotyki

Regression Analysis of Student Performance in Secondary Education

Author: Marwan Abdalfadeel

Subject: Python in Machine Learning and Data Science

Supervisor : dr hab. Inż. Ziemowit Dworakowski

1. Introduction

1.1 Project Objective

This project investigates factors influencing student academic performance using the **Student Performance** dataset from the UCI Machine Learning Repository. The dataset contains demographic, social, and educational attributes of secondary school students together with their academic results.

The primary objective of the project is not only to predict final academic outcomes, but to examine how selected variables interact and jointly influence student performance. In particular, the study focuses on identifying **moderation effects** between behavioral variables such as study time and absences.

The central research hypothesis is:

Higher study time weakens the negative effect of absences on the final grade (G3).

To test this hypothesis, both a Linear Regression model (with an explicit interaction term) and a Multilayer Perceptron (MLP) Regressor were implemented following the experimental workflow used in the AGH Galaxy laboratory exercises. The comparison between linear and non-linear models allows assessment of whether the observed interaction is robust across modeling approaches.

2. Dataset Description

2.1 Data Source

The dataset used in this project is the **Student Performance** dataset obtained from the UCI Machine Learning Repository. It consists of two separate subsets:

- `student-mat.csv` – Mathematics course
- `student-por.csv` – Portuguese language course

Each dataset contains 33 variables describing student demographic characteristics, family background, social behavior, and academic history.

2.2 Dataset Structure

In this project, the Mathematics and Portuguese datasets were analyzed separately to allow subject-specific comparison of model behavior.

Each dataset includes:

- **Demographic variables** (e.g., age, sex)
- **Family background variables** (e.g., parental education)
- **Behavioral variables** (e.g., study time, alcohol consumption, social activities)
- **Academic variables** (e.g., number of absences, first and second period grades G1 and G2)

The primary target variable is:

- **G3** – Final grade (integer scale from 0 to 20)

Grades from earlier evaluation periods:

- **G1** – First period grade
- **G2** – Second period grade

are included as explanatory variables because they contain prior academic information.

2.3 Selected Features for the Experiment

Although the dataset contains 33 variables, this study focuses on a hypothesis-driven subset of features:

- **absences**
- **studytime**
- **absences × studytime** (interaction term)
- **G2** (control variable)

The interaction term was constructed to directly test whether study time moderates the relationship between absences and final grade.

Because the dataset contains both numerical and categorical variables, preprocessing steps were applied where necessary; however, the final experimental setup used only numerical variables, which allowed direct implementation of Linear Regression and MLP models without additional encoding procedures.

3. Related Work

The Student Performance dataset from the UCI Machine Learning Repository has been widely used as a benchmark dataset for educational data mining and predictive modeling tasks. Most previous studies focus on predicting the final grade (G3) using supervised learning approaches such as Linear Regression, Decision Trees, Random Forests, Support Vector Machines, and Neural Networks.

In many cases, demographic, behavioral, and academic variables are treated as independent predictors, and model performance is evaluated using standard metrics such as accuracy (for classification), Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and the coefficient of determination (R^2).

A well-known characteristic of this dataset is the strong correlation between earlier grades (G1 and G2) and the final grade (G3). As a result, many predictive models achieve high performance when including G2 as a feature, since it contains information very close in time to the final evaluation.

However, fewer studies explicitly examine interaction or moderation effects between variables. In particular, limited attention has been given to whether increased study effort can compensate for disadvantages related to high absenteeism or unfavorable learning conditions. Rather than focusing exclusively on predictive accuracy, this project investigates whether study time moderates the negative relationship between absences and final academic performance.

This interaction-based perspective distinguishes the present study from purely accuracy-driven modeling approaches.

4. Initial Data Analysis

4.1 Descriptive Statistics

Descriptive statistics were computed separately for the Mathematics and Portuguese datasets to obtain an overview of variable distributions before model training.

The target variable G3 (final grade) exhibits moderate variability across students, with most values concentrated in the mid-range of the 0–20 grading scale. This suggests that prediction is meaningful but not trivial.

Behavioral variables such as **studytime**, **absences**, and alcohol consumption show substantial dispersion across observations. In particular:

- Study time is recorded on an ordinal scale (1–4), representing increasing levels of weekly study effort.
- Absences show a right-skewed distribution, with most students having relatively low absence counts and a smaller group exhibiting high absenteeism.
- Alcohol consumption variables demonstrate variability across students, indicating potential behavioral differences.

The observed variability in these features suggests that they may contribute to explaining differences in academic outcomes and supports further investigation using regression models.

4.2 Correlation Analysis

To investigate linear relationships between numerical variables, a Pearson correlation analysis was conducted separately for the Mathematics and Portuguese datasets. The resulting correlation matrices are presented in **Figure 1**.

Strong positive correlations were observed between earlier academic grades (G1 and G2) and the final grade (G3). This confirms temporal consistency in academic performance, as students who perform well in earlier evaluation periods tend to achieve higher final grades. In particular, G2 exhibits the strongest correlation with G3, indicating that recent academic performance is a highly informative predictor of final outcomes.

In contrast, behavioral and contextual variables such as **studytime**, **absences**, alcohol consumption, and travel time show weak individual correlations with G3. This suggests that these variables do not exert strong independent linear effects on final grade when considered in isolation.

The relatively weak marginal correlations of studytime and absences support the motivation for introducing an interaction term. It is possible that their effects are conditional — for example, increased study time may reduce the negative impact of absences rather than directly increasing grades on its own.

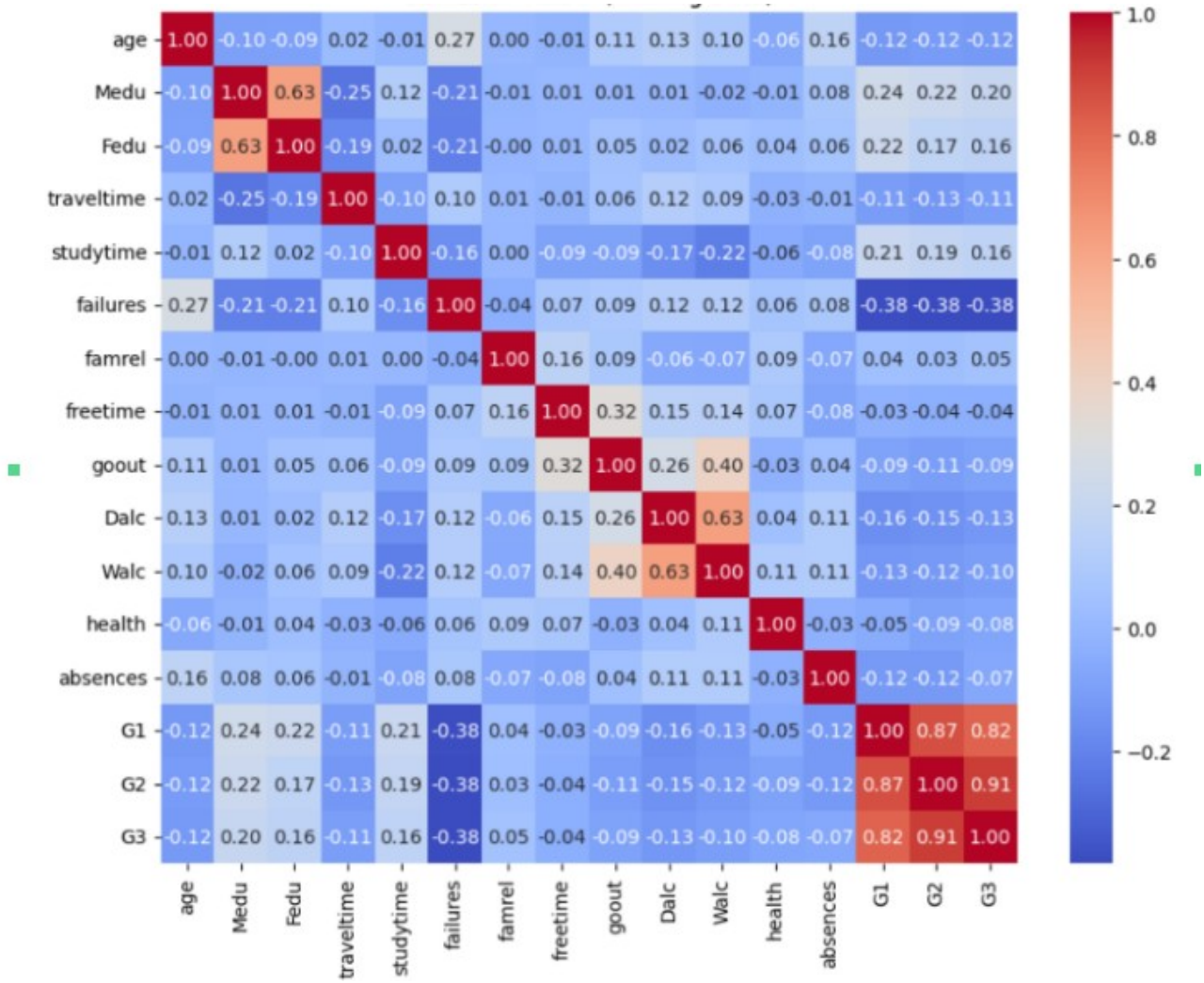


Figure 1. Correlation matrix of numerical variables for the Student Performance dataset (Mathematics and Portuguese subsets analyzed separately)

4.3 Visual Analysis

To complement the correlation analysis, visual exploration was performed using scatter plots and box plots to examine the relationship between selected variables and the final grade (G3). The results are presented in **Figure 2**.

The distribution of final grades shows moderate variability, with most students scoring between approximately 8 and 15 on the 0–20 scale. This indicates that while extreme outcomes exist, the majority of observations are concentrated in the middle range.

The relationship between **absences** and G3 exhibits high dispersion. Although very high absence counts are occasionally associated with lower grades, the overall spread suggests that absences alone do not fully determine academic performance. Students with similar absence levels may still achieve substantially different final results.

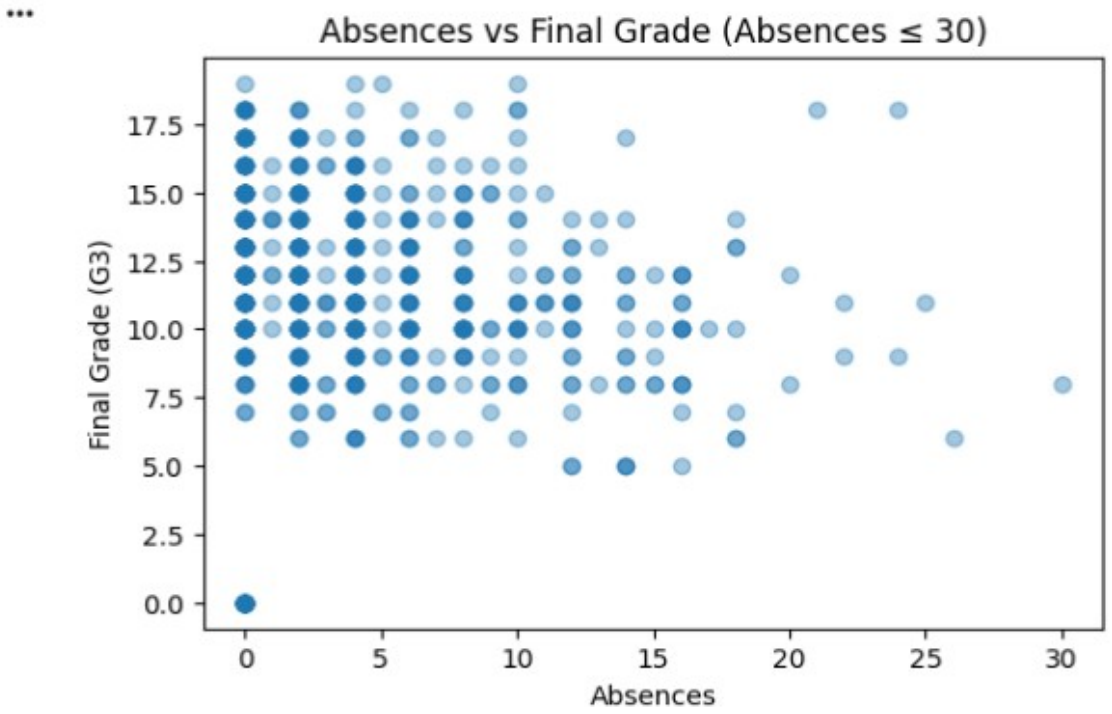
For **studytime**, higher levels are associated with slightly higher median grades. However, there is considerable overlap between studytime groups (1–4), indicating that increased study effort does not guarantee higher performance when considered independently.

Variables such as **internet access** and **travel time** show only minor differences in grade distributions. In contrast, higher levels of alcohol consumption are associated with slightly lower median grades, although variability remains substantial within each group.

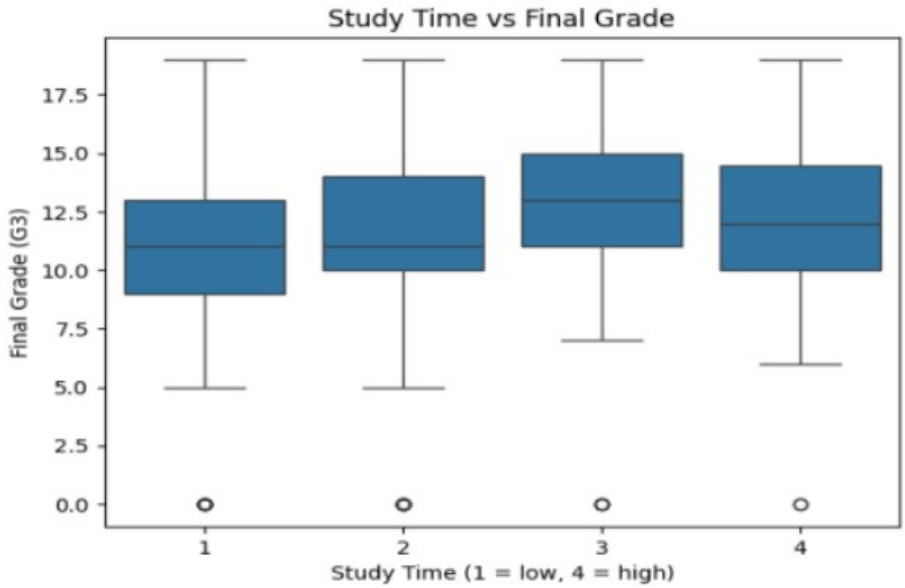
Overall, the visual analysis confirms that individual behavioral and contextual variables exhibit weak standalone effects on final grade. This observation further motivates the investigation of interaction-based relationships, particularly whether study time moderates the effect of absences

Figure 2. Visual exploration of selected relationships with final grade (G3):

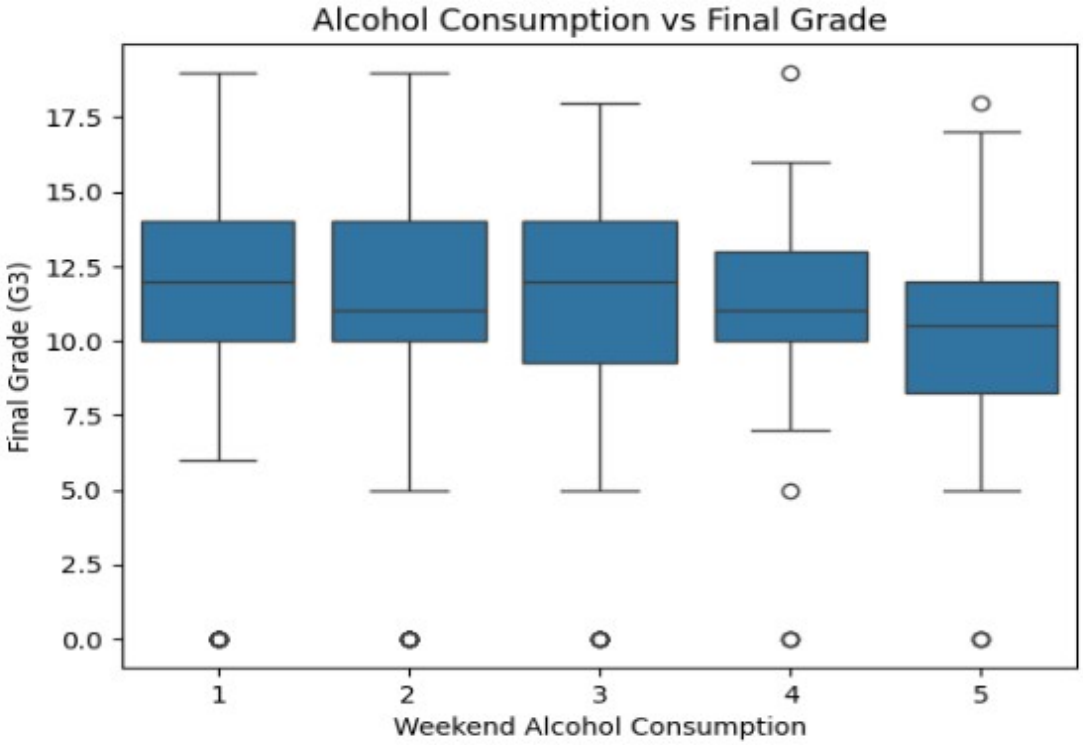
(a) Absences vs. G3



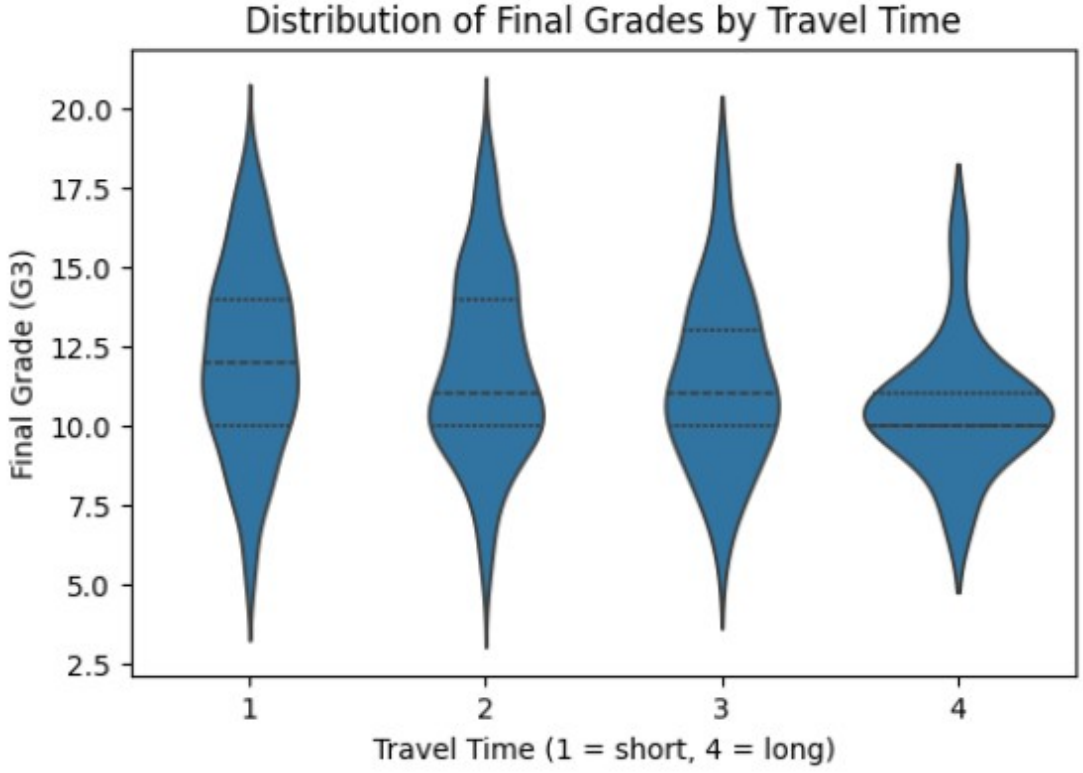
(b) Studytime vs. G3



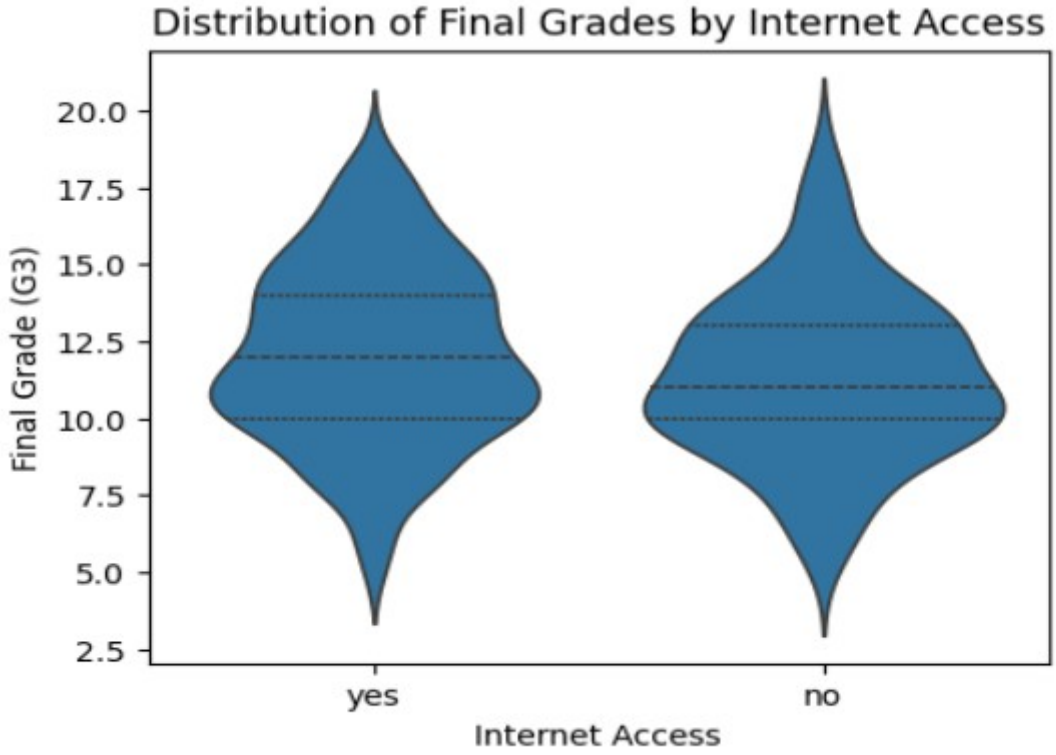
(c) Alcohol consumption vs. G3



(d) Travel time vs. G3



(d) Internet access vs. G3



5. Biases and Assumptions

Before formulating the main research hypothesis, several domain-based assumptions were considered based on educational theory and common expectations regarding student performance:

- Increased study time is expected to positively influence academic performance.
- A higher number of absences may negatively affect learning outcomes due to missed instructional time.
- Alcohol consumption may reduce academic focus and negatively impact performance.
- Longer travel time may contribute to fatigue and indirectly affect attendance or study efficiency.
- Internet access may support independent learning and access to educational resources.

These assumptions serve as contextual motivation for hypothesis development. However, they do not imply causality, as the dataset is observational and does not allow controlled experimentation. Therefore, the analysis focuses on statistical associations rather than causal claims.

6. Hypothesis Formulation

Based on the exploratory data analysis and observed correlation patterns, several potential interaction-based research hypotheses were considered. These hypotheses examine whether the effect of one variable depends on the level of another variable (moderation effects).

Candidate Interaction Hypotheses

- Higher study time weakens the negative effect of absences on final grades.
- Parental education moderates the effect of study time on final grades.
- Lack of internet access amplifies the negative effect of low study time on final grades.
- High alcohol consumption combined with frequent social outings reduces final grades more than either factor alone.
- Longer travel time intensifies the negative effect of absences on final grades.
- Higher alcohol consumption strengthens the negative effect of absences on final grades.

Although multiple interaction hypotheses were considered, the primary focus of this project is the first hypothesis:

> **H_1:** Higher study time weakens the negative effect of absences on final grades (G3).

This hypothesis was selected for detailed investigation because both **studytime** and **absences** exhibit relatively weak individual correlations with G3, as shown in the exploratory analysis. This suggests that their relationship with academic performance may not be purely additive, but instead conditional.

To test this moderation hypothesis explicitly, an interaction term (absences × studytime) is incorporated into the regression model.

7. Methodology

This study evaluates whether increased study time compensates for student absences by modeling the final grade (G3) using supervised regression techniques. Two regression approaches were implemented following the workflow used in the AGH laboratory exercises:

1. Linear Regression (interpretable baseline model)
2. Multilayer Perceptron (MLP) Regressor (non-linear model)

Both models were implemented in Python using the `scikit-learn` library.

7.1 Feature Engineering

The experiment focuses on a hypothesis-driven subset of variables:

- **absences**
- **studytime**
- **G2** (second period grade, included as a control variable)

To explicitly test the moderation hypothesis, an interaction feature was constructed:

abs_x_study=absences×studytime

The final feature vector used for modeling was:

$X = [absences, studytime, |x_{study}|, G2]$

The target variable was:

$y = G3$

Both Mathematics and Portuguese datasets were analyzed separately.

7.2 Data Splitting Strategy

To ensure proper model evaluation and prevent data leakage, fixed random splits were used with `random_state=1`.

For **Linear Regression**, a three-way split was applied:

- 60% training set
- 20% validation set
- 20% test set

This was implemented using two sequential `train_test_split` calls:

- First: 80% temporary training + 20% test
- Second: 75% training + 25% validation from the temporary training set

For the **MLP Regressor**, the dataset was split into:

- 80% training set
- 20% test set

Within the training set, 5-fold cross-validation was used for hyperparameter tuning.

7.3 Linear Regression Model

An Ordinary Least Squares (OLS) Linear Regression model was trained using `sklearn.linear_model.LinearRegression`.

The model estimates coefficients by minimizing the Mean Squared Error (MSE) using a closed-form least-squares solution.

The primary parameter of interest is the coefficient associated with the interaction term:

$\beta_{abs\_x\_study}$

Interpretation:

- If $\beta_{abs\_x\_study} > 0$: study time weakens the negative effect of absences
- If $\beta_{abs\_x\_study} < 0$: study time does not compensate for absences

Model performance was evaluated using:

- Root Mean Squared Error (RMSE)
- Coefficient of determination (R^2)

These metrics were computed separately for training, validation, and test sets.

Additionally, residual-vs-predicted plots were generated using `PredictionErrorDisplay` to visually assess prediction behavior.

7.4 Multilayer Perceptron (MLP) Regressor

To examine whether the interaction effect remains under a non-linear modeling framework, an MLP Regressor (`sklearn.neural_network.MLPRegressor`) was implemented.

The MLP model optimizes parameters using backpropagation and stochastic gradient-based optimization. The objective function minimized is Mean Squared Error (MSE).

7.4.1 Cross-Validation

Before hyperparameter tuning, 5-fold cross-validation was performed on the training set using:

- R^2
- Mean Absolute Error (MAE)
- Root Mean Squared Error (RMSE)

This provides an estimate of generalization performance.

7.4.2 Hyperparameter Tuning (Grid Search)

Hyperparameters were optimized using `GridSearchCV` with 5-fold cross-validation, minimizing negative RMSE.

The search grid included:

- Hidden layer sizes: (20), (50), (50, 20)
- Activation functions: ReLU, tanh
- L2 regularization parameter (alpha): 0.0001, 0.001, 0.01
- Learning rate initialization: 0.001, 0.01
- Maximum iterations: 2000
- Random state: 1

The best-performing configuration was selected and evaluated once on the held-out test set.

Test metrics reported:

- R^2
- MAE
- RMSE

7.5 Counterfactual (Ceteris Paribus) Analysis

To interpret model behavior beyond aggregate error metrics, a controlled counterfactual experiment was conducted.

Procedure:

- G2 fixed at its median value
- Absences fixed at 20
- Studytime varied across levels 1–4
- Interaction term recomputed accordingly

Predicted G3 values were computed for each studytime level.

The change:

ΔG3=G3^(studytime=4)−G3^(studytime=1)

was used as an interpretable measure of compensation effect.

This procedure was applied to both the Linear Regression and MLP models to ensure comparability.

8. Results

8.1 Linear Regression Results

The Linear Regression model provides a direct test of the moderation hypothesis through the coefficient associated with the interaction term:

βabs_x_study

Model performance was evaluated using RMSE and R^2 on training, validation, and test sets. Overall, both datasets exhibited reasonable predictive performance, primarily driven by the inclusion of G2 as a control variable.

For the **Mathematics dataset**, the estimated interaction coefficient was:

βabs_x_study=−0.0036

The negative sign indicates that increasing study time does not reduce the negative association between absences and final grade. In other words, the interaction effect does not support the hypothesis for Mathematics. The magnitude of the coefficient is small, suggesting a limited moderation effect.

For the **Portuguese dataset**, the estimated interaction coefficient was:

βabs_x_study=0.000087

The positive sign indicates that higher study time is associated with a slight reduction in the negative relationship between absences and final performance. Although the magnitude of this coefficient is very small, its direction is consistent with the proposed hypothesis.

Overall, the Linear Regression results suggest subject-dependent behavior: evidence consistent with a compensatory effect appears in Portuguese but not in Mathematics. However, the small magnitude of the interaction coefficients indicates that the effect size is limited.

8.2 MLP Regression Results

To examine whether the observed interaction pattern persists under a more flexible modeling framework, a Multilayer Perceptron (MLP) Regressor was trained with hyperparameter tuning via 5-fold cross-validation.

Since the MLP model does not provide directly interpretable interaction coefficients, the moderation hypothesis was evaluated using controlled counterfactual predictions (Section 7.5).

For the **Portuguese dataset**, counterfactual predictions showed an increase in predicted final grade when study time was increased from level 1 to level 4 while keeping absences and G2 fixed. This directional change is consistent with the compensatory pattern observed in the Linear Regression model.

For the **Mathematics dataset**, counterfactual predictions did not show a consistent increase in predicted grade when study time was increased under fixed absence conditions. This suggests that additional study time does not systematically offset the effect of absences in the Mathematics dataset.

Taken together, the MLP results align with the general directional pattern observed in the Linear Regression analysis: a compensatory trend is present in Portuguese but not clearly observable in Mathematics. However, as with the linear model, the magnitude of the predicted changes remains relatively small.

8.3 Counterfactual Analysis

To further evaluate the moderation hypothesis, a controlled counterfactual experiment was conducted for both models.

For the Linear Regression model, the interaction coefficient provides the primary evidence regarding moderation. The counterfactual analysis therefore serves as an illustrative confirmation of the coefficient’s direction.

For the MLP model, which does not provide interpretable coefficients, counterfactual predictions play a central interpretative role. By fixing selected input variables and varying others, the effect of study time can be examined under controlled conditions.

In this experiment:

- G2 was fixed at its median value
- Absences were fixed at 20
- Studytime was varied across its observed levels (1–4)
- The interaction term was recomputed accordingly

This setup isolates the effect of increasing study time while keeping other inputs constant.

Table 1 presents the predicted G3 values under this controlled configuration.

Table 1. Counterfactual predictions for fixed absences (20) and median G2

Dataset	Model	Studytime 1	Studytime 2	Studytime 3	Studytime 4	ΔG3 (4−1)
Math	Linear	11.21	11.07	10.92	10.77	−0.44
Math	MLP	10.88	10.74	11.08	11.14	+0.26
Portuguese	Linear	11.47	11.59	11.70	11.82	+0.35
Portuguese	MLP	10.82	11.30	11.56	12.25	+1.43

ΔG3 represents the change in predicted final grade when studytime increases from level 1 to level 4.

For the **Mathematics dataset**, the Linear Regression model predicts a decrease in G3 as studytime increases under fixed absences (ΔG3 = −0.44), consistent with the negative interaction coefficient. The MLP model shows only a small positive change (ΔG3 = +0.26), indicating no strong compensatory effect.

For the **Portuguese dataset**, both models predict an increase in G3 as studytime increases. The Linear Regression model shows a modest increase (ΔG3 = +0.35), while the MLP model shows a larger increase (ΔG3 = +1.43). These results are directionally consistent with the hypothesis that studytime mitigates the negative association between absences and final grade in Portuguese.

8.4 Prediction Error Analysis

Residual analysis was performed to assess model behavior beyond aggregate error metrics. The residual-versus-predicted plots for the Mathematics and Portuguese datasets are shown in **Figure 3** and **Figure 4**, respectively.

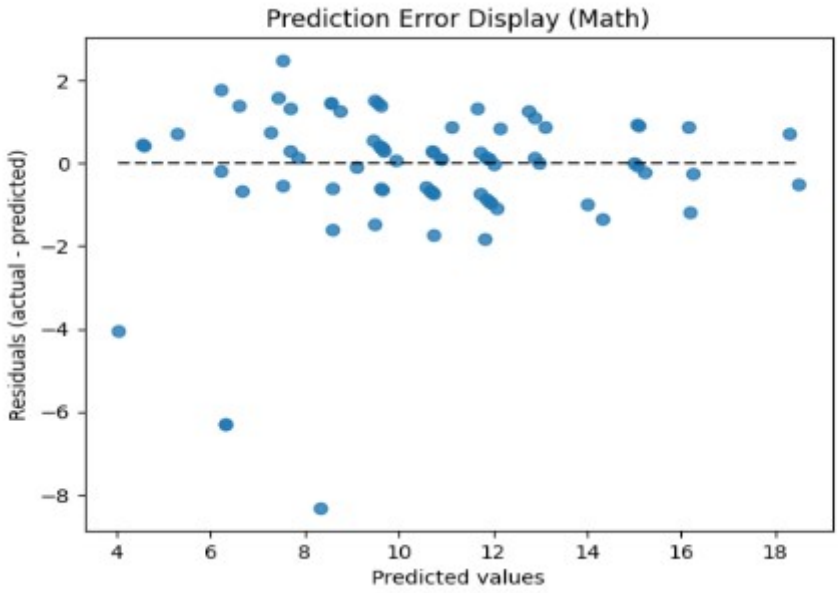


Figure 3. Residual vs. predicted values for the Mathematics dataset (Linear Regression).

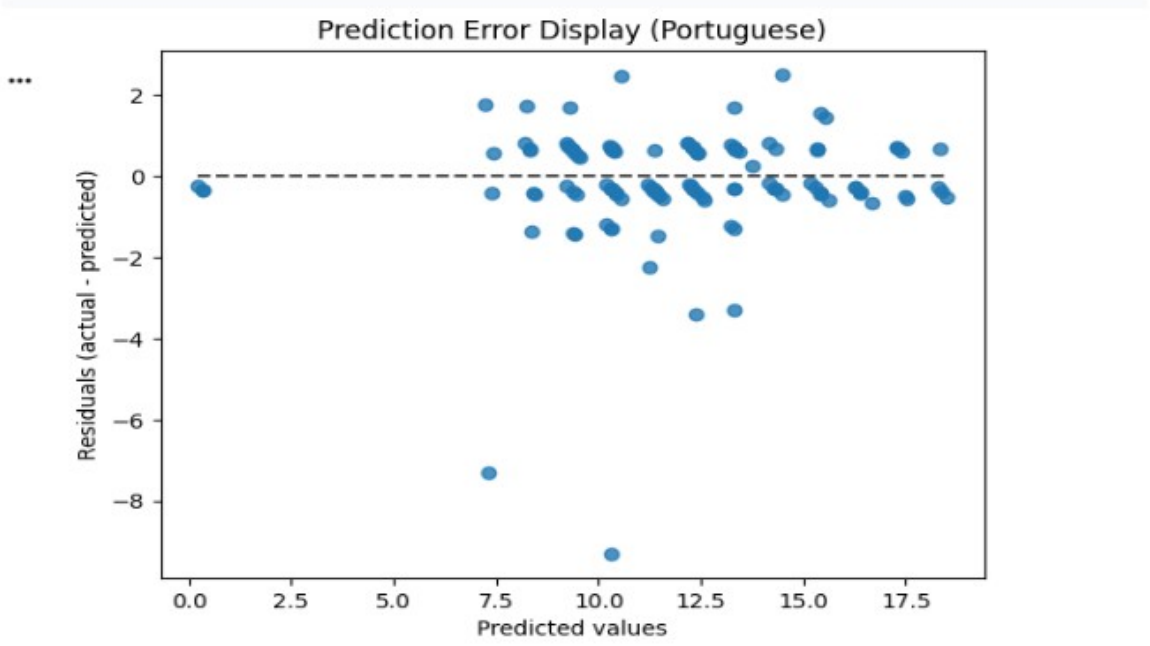


Figure 4. Residual vs. predicted values for the Portuguese dataset (Linear Regression).

The residuals in both datasets are generally centered around zero, indicating the absence of systematic bias in predictions. However, several extreme negative residuals are visible, suggesting that some students performed substantially below model expectations.

No clear non-linear pattern is observed in the residual distributions, which supports the adequacy of the linear specification for capturing the main trends in the data.

9. Discussion

The results suggest that the compensatory role of studytime is subject-dependent.

For Portuguese, both linear and non-linear models show a positive directional effect when studytime increases under fixed absence conditions. In contrast, for Mathematics, no consistent compensatory pattern is observed.

One possible interpretation is that academic performance in Mathematics may depend more strongly on continuous classroom participation, whereas Portuguese may allow greater flexibility for independent study. However, this interpretation remains speculative and cannot be confirmed without additional domain-specific investigation.

The comparison between Linear Regression and MLP indicates that allowing non-linear relationships does not fundamentally alter the qualitative conclusions. While the MLP model produces slightly larger predicted changes in some cases, the overall directional pattern remains similar.

Given the small magnitude of the interaction coefficients and predicted changes, the observed moderation effects should be interpreted cautiously.

10. Conclusion

This study examined whether increased studytime compensates for student absences in predicting final academic performance.

Using both Linear Regression (with an explicit interaction term) and a Multilayer Perceptron Regressor, evidence of a compensatory effect was found for the Portuguese dataset but not for the Mathematics dataset.

The results suggest that the hypothesis is partially supported. However, the magnitude of the interaction effect is small, indicating that increased studytime does not strongly offset the negative association between absences and final grade.

The counterfactual analysis proved valuable in interpreting model behavior beyond aggregate performance metrics and provided an interpretable framework for evaluating interaction effects.

Future work could extend this analysis by including additional behavioral variables, testing statistical significance of interaction terms, or exploring alternative modeling approaches.

11. AI Usage Statement

Artificial intelligence tools (ChatGPT by OpenAI) and (Gemini by Google)were used for language refinement, structural improvements, and clarification of methodological explanations. All data preprocessing, modeling, experimentation, and interpretation were performed independently by the author in accordance with the AGH laboratory workflow