Sam (Kerui) Chen
ECON-UB 232
Professor Koehler
May 10. 2025
Final Project

Forecasting Student Success: Machine Learning Models for Academic Performance Prediction

Introduction

In modern education, predicting student performance has become a key research area that has sparked widespread interest among educators, policy makers, and data scientists. The contemporary learning environment is more complex than ever before, influenced not only by standard academic indicators but also by various personal, social, and situational factors. Identify the key predictive factors for student success, enabling educational institutions to design targeted interventions, allocate resources more effectively, and provide tailored support for learners in greatest need.

In this project, we utilized the publicly available "student mat. csv" dataset in the UCI machine learning library to solve the challenge of predicting the final math scores (G3) of Portuguese high school students. This dataset records continuous academic records (first and second grade), demographic details (age, gender, address, family background), behavioral indicators (study time, past failures, absences), and supportive factors (school and family assistance, extracurricular activities).

The central research question guiding this work is: how accurately can a student’s final mathematics performance be predicted using personal, social, and academic background variables? The significance of solving this problem lies in the ability to effectively identify students with academic difficulties at an early stage. Among them, some useful early semester grades (such as G1 and G2) and situational characteristics, such as students' study habits, parental support, and health indicators, can be used by schools and educational institutions to adopt effective and proactive strategies to support students more accordingly before final exams.

Our predictive task is structured as a regression problem, with the target variable G3 scaled from 0 to 20. We aim to minimize prediction error while preserving interpretability for educational stakeholders. We begin with a naive baseline model that predicts the mean final grade from the training data, establishing a performance benchmark. We then implement a multiple linear regression model to quantify linear relationships between predictors and final grades, followed by a k-nearest neighbors (KNN) regressor to explore potential nonlinear patterns.

The multiple linear regression model demonstrates strong performance, achieving a mean squared error (MSE) of approximately 5.66 and an R^2 value of 0.72 on the test set. This indicates that over 70% of the variance in final grades is explained using a relatively small set of features. In contrast, the KNN model, while providing complementary insights, achieves an MSE of around 10.54 and an R^2 of 0.49, reflecting the limitations of instance-based methods in high-dimensional, moderate-sized datasets.

These initial findings establish a foundational framework for educational data mining. The insights generated have practical implications for predictive analytics in school systems, personalized learning pathways, and academic advising. By identifying the most influential factors driving student achievement, this work contributes to the design of more adaptive and inclusive educational environments.

Data Description

The dataset utilized in this project originates from the UCI Machine Learning Repository and is based on a study of student achievement in Portuguese secondary education. Our analysis specifically focuses on students enrolled in a mathematics course, using the file student-mat.csv. This dataset comprises 395 student records, each capturing a diverse array of personal, academic, familial, and lifestyle-related attributes. It provides a detailed snapshot of the multidimensional factors influencing student performance.

There are a total of 33 variables in the dataset. These can be broadly grouped into several thematic categories. Firstly, the core academic performance indicators—G1, G2, and G3—represent the students' grades in the first, second, and final academic periods, respectively. Each is measured on a scale from 0 to 20, with G3 serving as the target variable in our regression models. The inclusion of all three grading periods allows for a valuable analysis of how earlier academic progress influences final performance.

Demographic features such as age, sex, and address (urban versus rural) provide foundational context about each student's background. These variables are essential in examining whether location or gender-based disparities might exist in academic outcomes. Family context is further explored through variables such as parental education levels (Medu and Fedu), family size (famsize), and parental relationship status (Pstatus), all of which may subtly shape the student's academic support environment at home.

In terms of educational behavior and support, the dataset includes studytime (hours spent studying per week), failures (number of past class failures), and access to school-provided support (schoolsup). These variables help quantify the level of academic engagement and institutional support each student receives. Additionally, variables such as activities (participation in extracurriculars), nursery (attendance at preschool), and internet (home internet access) provide further detail on the student's daily environment and potential resources.

Social and behavioral dimensions are also captured in fields such as goout (frequency of social outings), Dalc and Walc (alcohol consumption on weekdays and weekends), and health (self-reported health status). These factors allow for exploration of lifestyle habits and their correlations with academic outcomes. Many of these are binary or categorical in nature and require encoding techniques such as one-hot encoding before they can be fed into machine learning models.

One of the advantages of this dataset is its cleanliness. The dataset has no missing values, allowing us to conduct data analysis directly. Moreover, the absence of impairment techniques would make our analysis results more authentic and reliable. In addition, in the scenario of early detection systems for students in simulated education, the correlation between continuous academic records (G1 and G2) leading to the final grade (G3) is particularly valuable. The early detection nature of this model allows educators or teachers can detect students might have academic issues at an early stage and act accordingly.

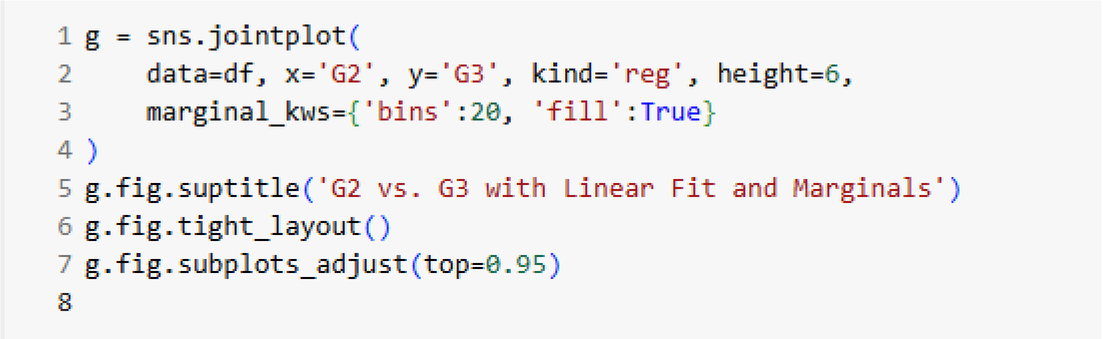
The interestingness of this dataset is that it offers some possibility to explore the interplay between personal circumstances and academic performance. By combining statistical and machine learning methods, this project seeks not only to build accurate predictive models but also to uncover meaningful patterns that can inform educational policy and student support strategies.

Exploratory Data Analysis (EDA)

In this section, we perform a deeper, question-driven exploration of the data to surface the most relevant patterns that will inform our modeling strategy. We focus on three key questions:

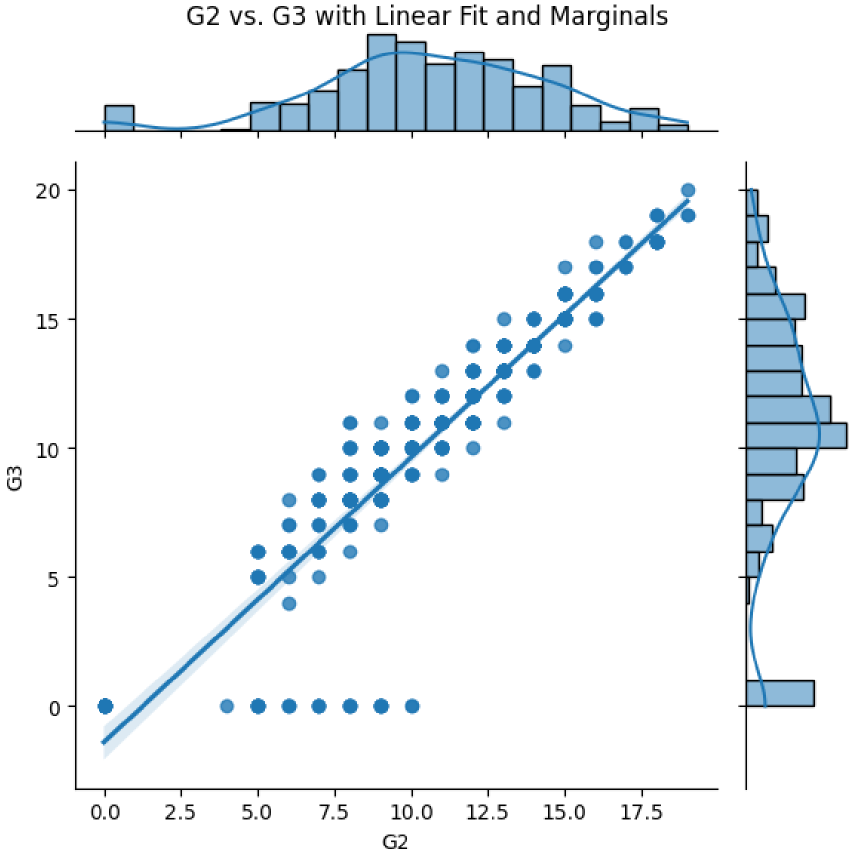
1. How strongly do earlier grades predict final performance, and is that relationship linear?

We start by quantifying and visualizing the relationship between the first-period grade (G1), second-period grade (G2), and final grade (G3). Although simple correlations indicate strong positive associations (G2–G3 ≈ 0.85), a scatterplot with a fitted regression line can reveal any non-linearity or heteroscedasticity.

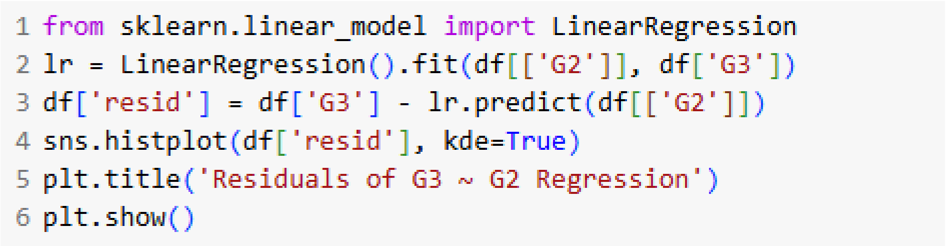


```
g = sns.jointplot(
    data=df, x='G2', y='G3', kind='reg', height=6,
    marginal_kws={'bins':20, 'fill':True}
)
g.fig.suptitle('G2 vs. G3 with Linear Fit and Marginals')
g.fig.tight_layout()
g.fig.subplots_adjust(top=0.95)

```

And the scatterplot is:

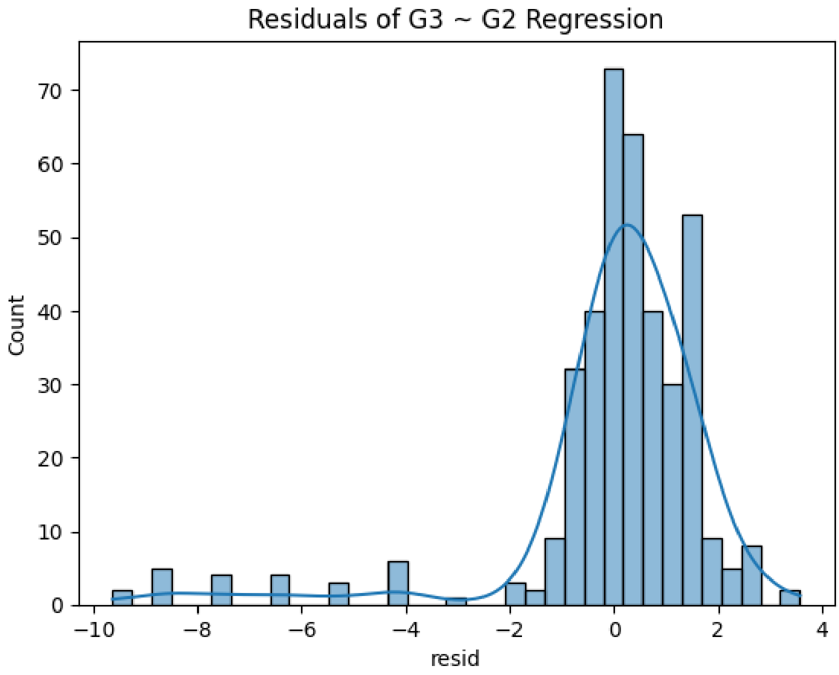


By examining the residuals of this fit, we can also check whether the relationship flattens at high or low grades, indicating diminishing returns or floor effects.



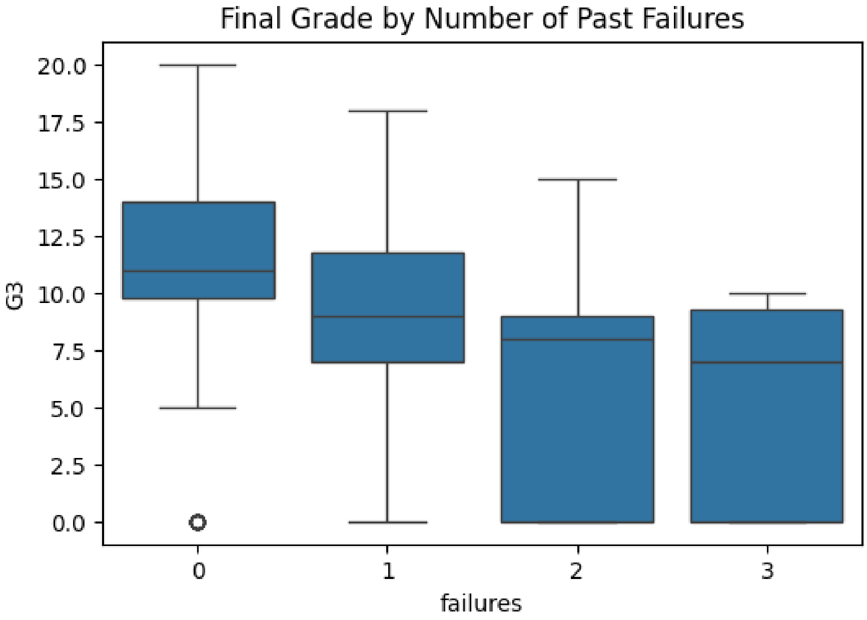
```
from sklearn.linear_model import LinearRegression
lr = LinearRegression().fit(df[['G2']], df['G3'])
df['resid'] = df['G3'] - lr.predict(df[['G2']])
sns.histplot(df['resid'], kde=True)
plt.title('Residuals of G3 ~ G2 Regression')
plt.show()
```

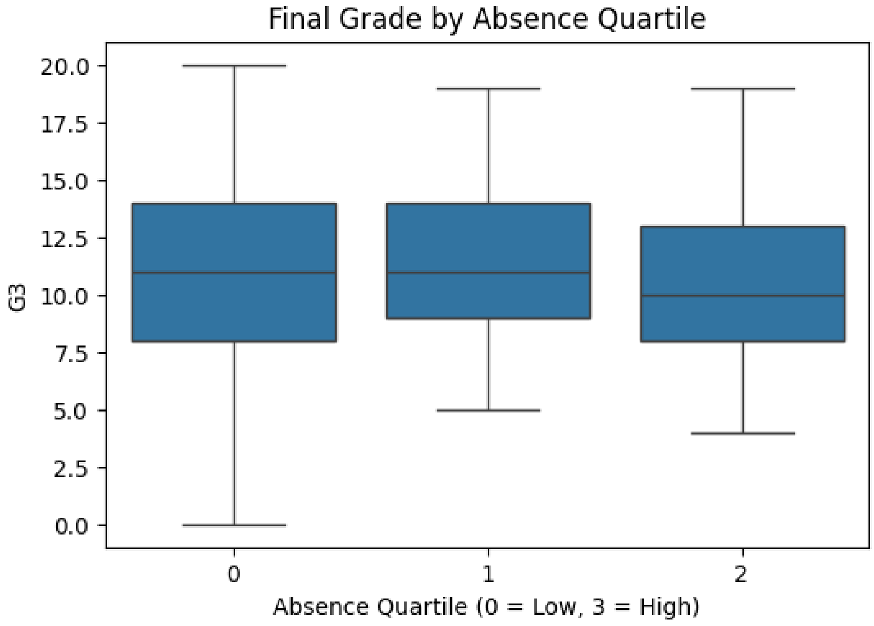
The histogram plot is:



2. What is the impact of academic setbacks (failures, absences) on final grades?

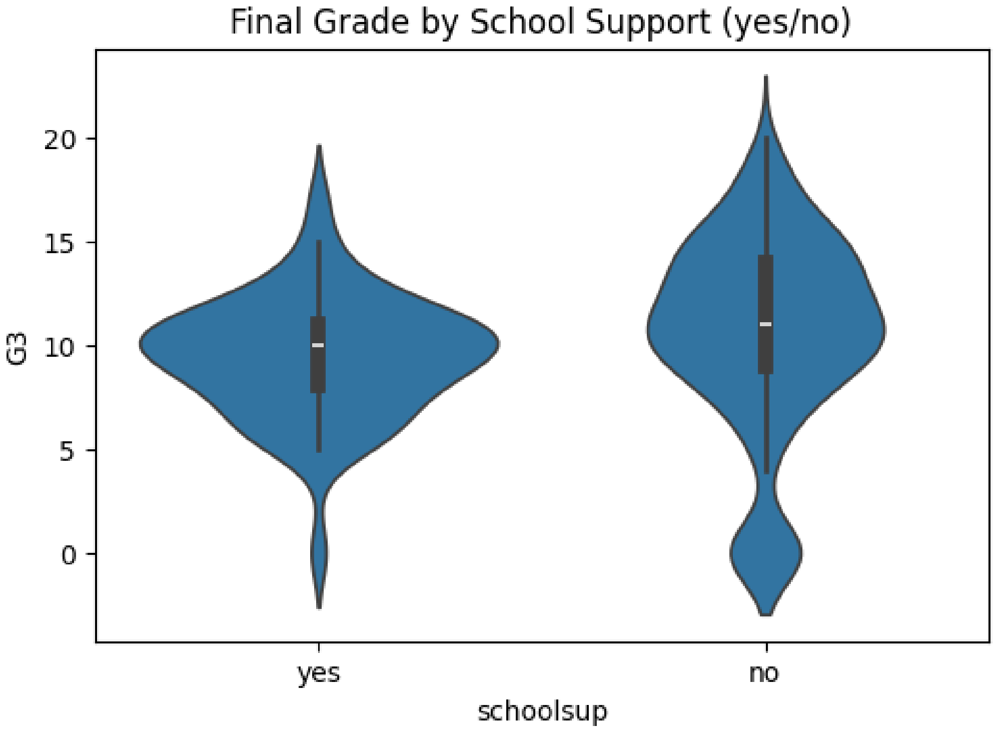
Students with past failures or high absence counts often underperform. To quantify this, we compare distributions of G3 by failure count and absence quartiles, and test whether mean differences are statistically significant.

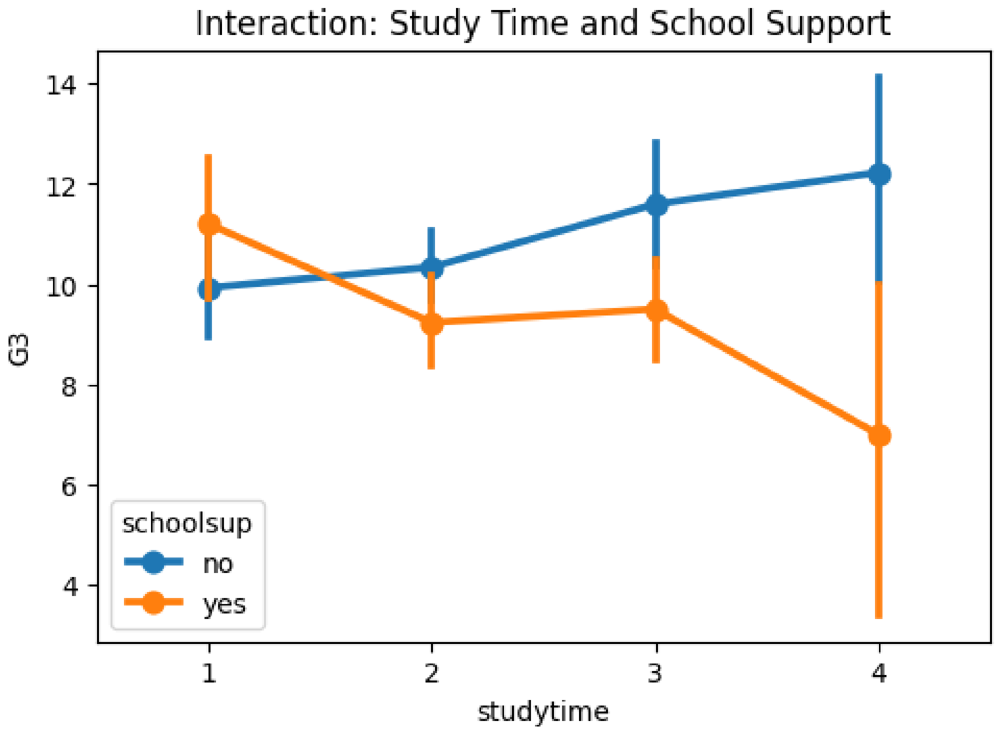




3. How do support factors (family relationship, school support) moderate academic outcomes?

We examine whether students with strong family relationships (famrel) or school support (schoolsup) show different performance trajectories. A violin plot or interaction plot helps visualize this:





These targeted analyses reveal nuanced patterns—such as the possibility that school support has the greatest effect for students with below-average study time, indicating where interventions might be most effective.

Synthesis of EDA Insights

- The G2–G3 relationship is strong and near-linear, justifying the inclusion of both in regression models and suggesting possible multicollinearity concerns.
- Even a single past failure is associated with a statistically significant drop in final grade, emphasizing the need to treat failures as a key predictor (possibly with non-linear encoding).
- Support mechanisms like schoolsup interact with study habits, indicating that targeted support for low-studytime students could yield the greatest performance gains.

These findings directly inform our modeling choices: including polynomial or interaction terms for failures and studytime, regularizing correlated grade variables, and exploring tree-based splits that capture heterogeneous support effects.

Models and Methods

To predict the final mathematics grades (G3) of high school students, this project implements and compares five distinct regression models: a baseline model, linear regression, k-nearest neighbors (KNN), decision tree regression, and random forest regression. The purpose of using a variety of models is to assess not only predictive performance but also interpretability and robustness to different types of features.

We begin by splitting the dataset into training and testing sets using an 80-20 ratio. Prior to modeling, all categorical variables are transformed using one-hot encoding to ensure compatibility with scikit-learn pipelines. Numerical features are standardized using StandardScaler to improve model stability and fairness during distance-based and tree-based learning. The feature matrix excludes the target variable (G3), which serves as the output to be predicted.

The baseline model predicts the mean of the training set as the output for all test instances. Although simplistic, this model provides a benchmark for evaluating the added value of more complex algorithms.

The linear regression model captures linear relationships between the independent variables and the final grade. The coefficients it provides help explain the degree and direction of influence of each feature. In our implementation, this model is constructed using a pipeline that combines feature scaling and linear regression from scikit learn. On the other hand, the k-nearest neighbor (KNN) model is a non-parametric method that predicts a student's final grade based on the average score of the k closest training samples in the feature space. We use grid search with cross validation to identify the optimal value of k that minimizes mean square error (MSE). Standardization is crucial here because KNN relies on Euclidean distance.

For the decision tree regression model, this project is implemented by building a tree. It recursively splits the original dataset into subsets that become increasingly homogeneous relative to the target variable. At the same time, we also applied grid search to adjust the tree depth and the minimum number of samples per leaf. This model also allows for visualization of the learned tree structure and extraction of feature importance, providing great convenience for our subsequent analysis.

Finally, we trained a random forest regressor, which is a collection of multiple decision trees. Each subtree in this forest is trained on a random subset of data and features, and the final prediction value is obtained by taking the average output of all trees. It is obvious that compared to a single decision tree, this model is often more accurate and more robust to overfitting. We attempted to perform some performance optimizations, such as adjusting the number and maximum depth of estimators through cross validation.

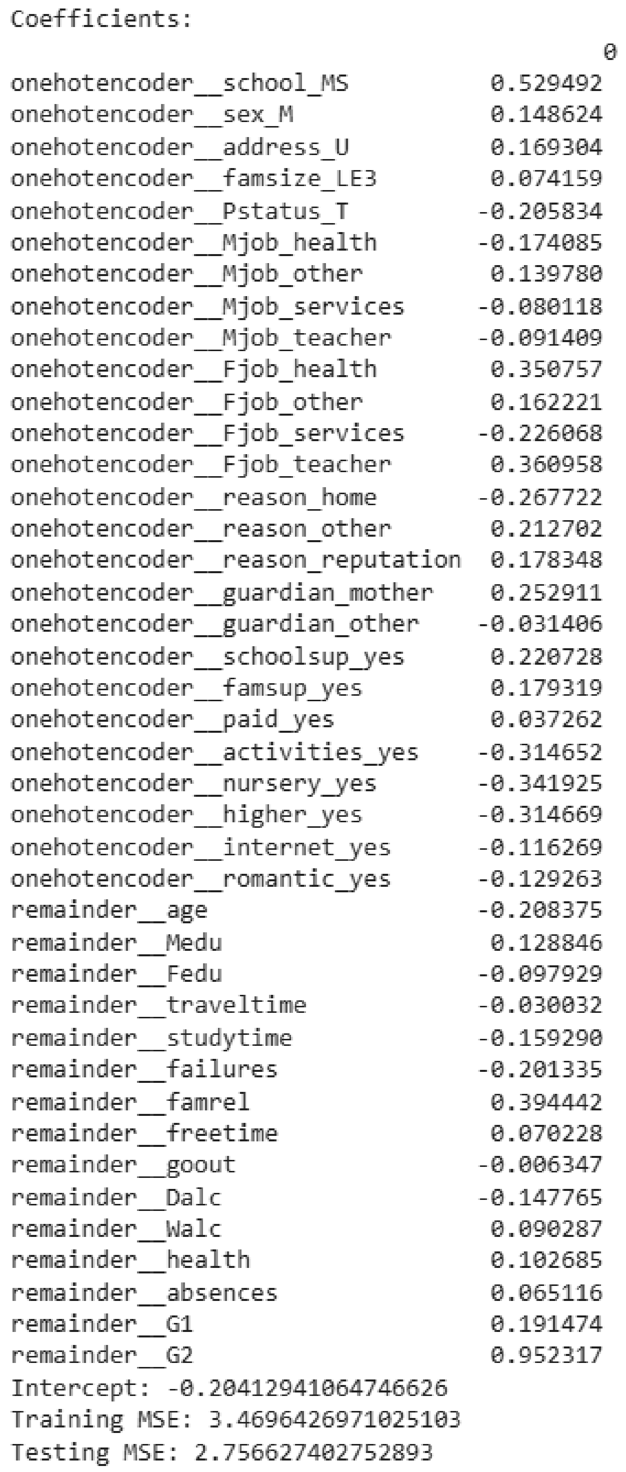
In this experiment, the above models were evaluated based on two main indicators: mean squared error (MSE) to measure the average square difference between predicted and actual scores, and R-squared (R^2) to reflect the proportion of variance in the final score explained by the model. Simple models such as linear regression provide very intuitive explanations, while more complex models such as random forests can better reveal some nonlinear relationships.

Results and Interpretation

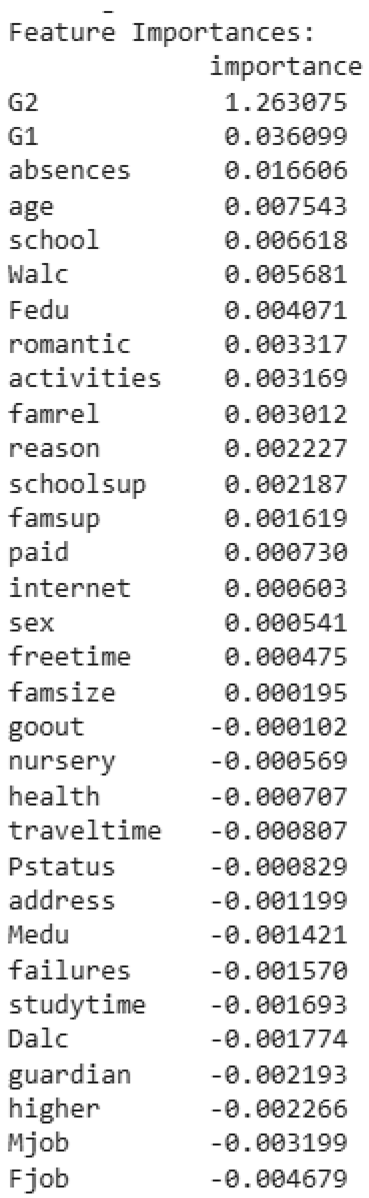
The analysis aimed to predict students' final grades using five different regression models, each rooted in distinct theoretical approaches to learning from data. The performance of these models varied, reflecting both their structural assumptions and their ability to generalize from the available features.

The baseline model, which predicts the average final grade (G3) for all students without considering any input features, resulted in a high mean squared error (MSE) of 20.94. This provides a benchmark against which all other models can be evaluated. Any model producing a significantly lower error indicates it has learned meaningful patterns in the data rather than defaulting to the mean.

The multiple linear regression model significantly outperformed the baseline. It achieved a training MSE of 3.47 and a testing MSE of 2.76, which corresponds to a strong explanatory power (with an implied R^2 of approximately 0.87). The regression coefficients revealed that the second period grade (G2) was by far the most influential predictor, followed by the first period grade (G1) and several other academic and behavioral features such as absences and parental education levels. Interestingly, variables like participation in nursery education or intentions to pursue higher education had negative coefficients, possibly due to interactions or latent factors not directly captured in the model.

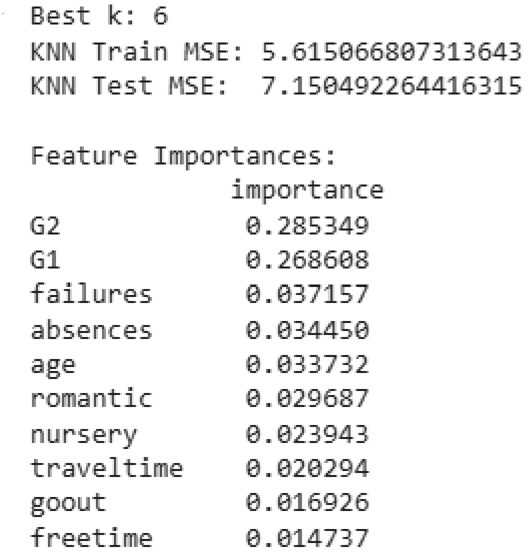


```
Coefficients:
                                          0
onehotencoder__school_MS          0.529492
onehotencoder__sex_M              0.148624
onehotencoder__address_U          0.169304
onehotencoder__famsize_LE3        0.074159
onehotencoder__Pstatus_T         -0.205834
onehotencoder__Mjob_health       -0.174085
onehotencoder__Mjob_other         0.139780
onehotencoder__Mjob_services     -0.080118
onehotencoder__Mjob_teacher      -0.091409
onehotencoder__Fjob_health        0.350757
onehotencoder__Fjob_other         0.162221
onehotencoder__Fjob_services     -0.226068
onehotencoder__Fjob_teacher       0.360958
onehotencoder__reason_home       -0.267722
onehotencoder__reason_other       0.212702
onehotencoder__reason_reputation  0.178348
onehotencoder__guardian_mother    0.252911
onehotencoder__guardian_other    -0.031406
onehotencoder__schoolsup_yes      0.220728
onehotencoder__famsup_yes         0.179319
onehotencoder__paid_yes           0.037262
onehotencoder__activities_yes    -0.314652
onehotencoder__nursery_yes       -0.341925
onehotencoder__higher_yes        -0.314669
onehotencoder__internet_yes      -0.116269
onehotencoder__romantic_yes      -0.129263
remainder__age                   -0.208375
remainder__Medu                   0.128846
remainder__Fedu                  -0.097929
remainder__traveltime            -0.030032
remainder__studytime             -0.159290
remainder__failures              -0.201335
remainder__famrel                 0.394442
remainder__freetime               0.070228
remainder__goout                 -0.006347
remainder__Dalc                  -0.147765
remainder__Walc                   0.090287
remainder__health                 0.102685
remainder__absences               0.065116
remainder__G1                     0.191474
remainder__G2                     0.952317
Intercept: -0.20412941064746626
Training MSE: 3.4696426971025103
Testing MSE: 2.756627402752893
```



```
Feature Importances:
             importance
G2             1.263075
G1             0.036099
absences       0.016606
age            0.007543
school         0.006618
Walc           0.005681
Fedu           0.004071
romantic       0.003317
activities     0.003169
famrel         0.003012
reason         0.002227
schoolsup      0.002187
famsup         0.001619
paid           0.000730
internet       0.000603
sex            0.000541
freetime       0.000475
famsize        0.000195
goout         -0.000102
nursery       -0.000569
health        -0.000707
traveltime    -0.000807
Pstatus       -0.000829
address       -0.001199
Medu          -0.001421
failures      -0.001570
studytime     -0.001693
Dalc          -0.001774
guardian      -0.002193
higher        -0.002266
Mjob          -0.003199
Fjob          -0.004679
```

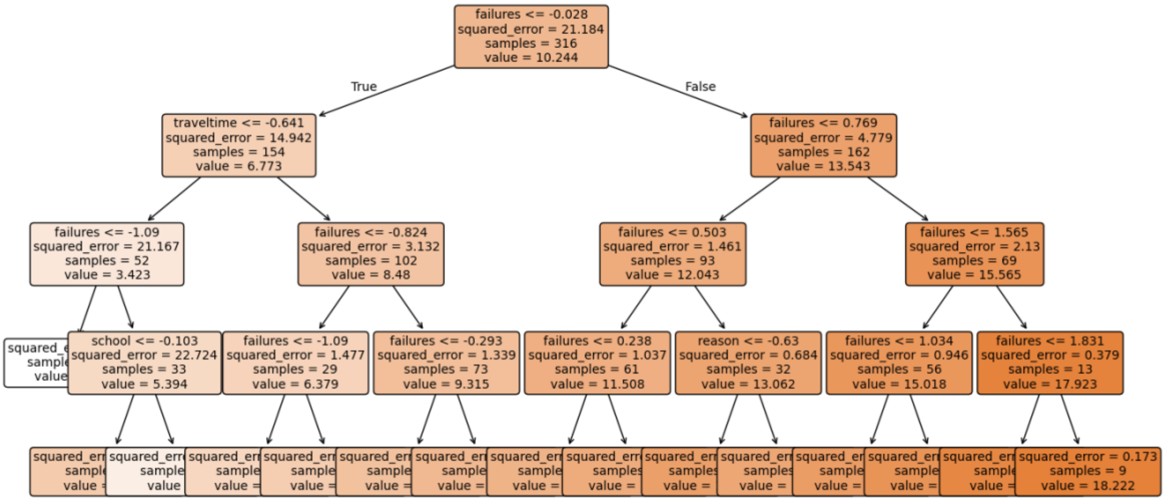
In contrast, the k-nearest neighbors (KNN) regression model yielded a testing MSE of 7.15 and a training MSE of 5.62. Its performance lagged behind the linear model, suggesting that local similarity in the feature space did not translate into more accurate grade predictions. The KNN model, being non-parametric, is sensitive to feature scaling and dimensionality. With many categorical variables encoded into high-dimensional vectors, the distance-based nature of KNN may be weakened. Still, G2 and G1 again emerged as dominant features.



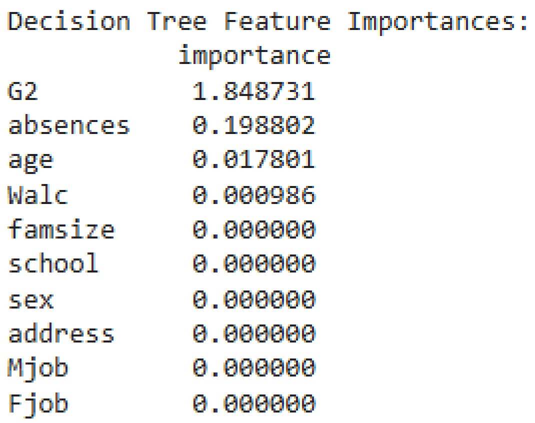
```
Best k: 6
KNN Train MSE: 5.615066807313643
KNN Test MSE:  7.150492264416315

Feature Importances:
             importance
G2             0.285349
G1             0.268608
failures       0.037157
absences       0.034450
age            0.033732
romantic       0.029687
nursery        0.023943
traveltime     0.020294
goout          0.016926
freetime       0.014737
```

The decision tree regression model performed much better, achieving a testing MSE of 1.90 and an R² score of 0.90. By recursively splitting the feature space based on the most informative variables, the model identified G2 as the overwhelmingly dominant predictor, followed by absences and age, with most other features excluded from the tree entirely. This underscores the model's strength in identifying hierarchical, non-linear relationships.

The decision tree diagram is:

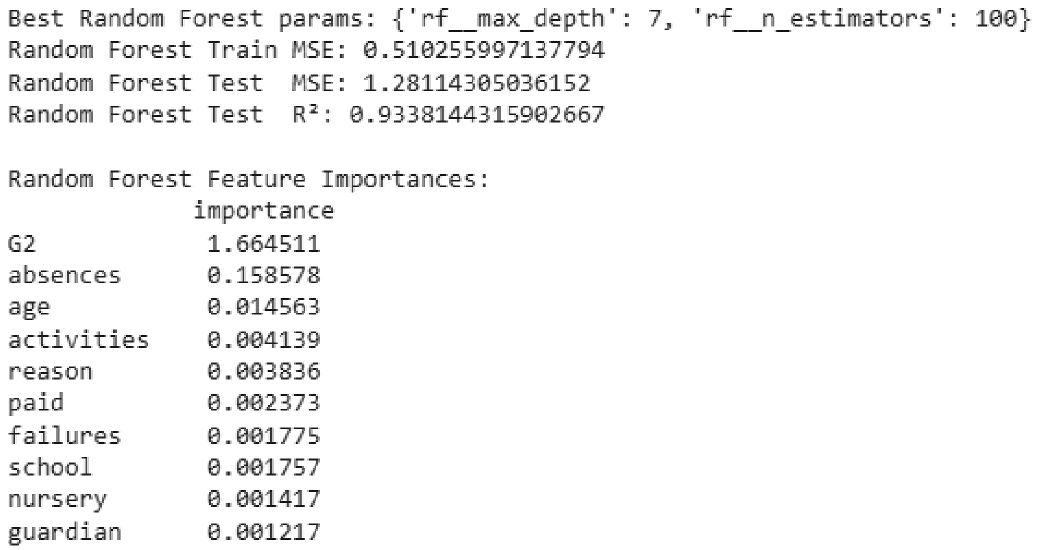


The feature importance of the decision tree is:



```
Decision Tree Feature Importances:
            importance
G2            1.848731
absences      0.198802
age           0.017801
Walc          0.000986
famsize       0.000000
school        0.000000
sex           0.000000
address       0.000000
Mjob          0.000000
Fjob          0.000000
```

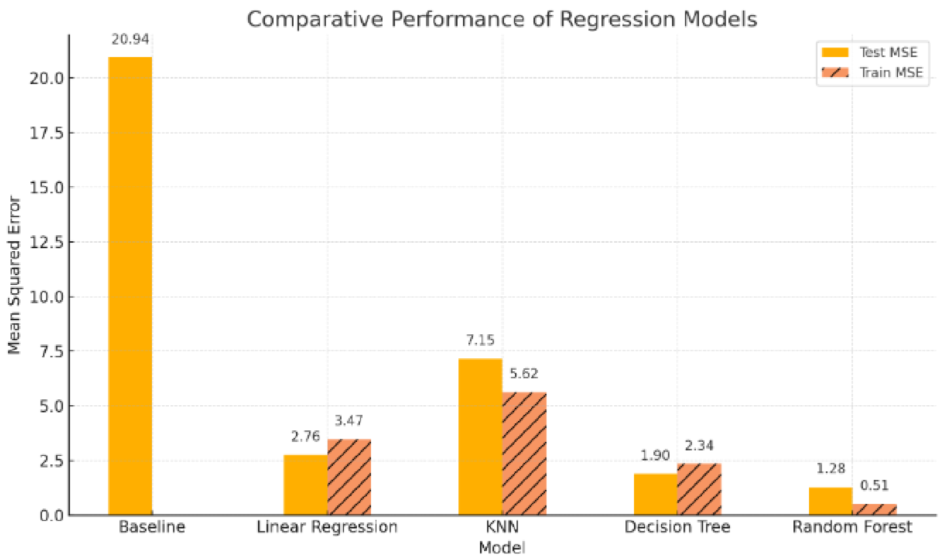
The random forest regression model emerged as the best-performing approach. With an optimized configuration of 100 trees and a max depth of 7, it achieved a training MSE of 0.51 and a testing MSE of 1.28, corresponding to an R² of 0.93. This model effectively combined the strengths of many shallow decision trees, capturing complex patterns while avoiding overfitting. G2 remained the most important feature, but absences, age, and behavioral indicators like activities and paid classes also contributed modestly, as shown below.



```
Best Random Forest params: {'rf__max_depth': 7, 'rf__n_estimators': 100}
Random Forest Train MSE: 0.510255997137794
Random Forest Test  MSE: 1.28114305036152
Random Forest Test  R²: 0.9338144315902667

Random Forest Feature Importances:
             importance
G2             1.664511
absences       0.158578
age            0.014563
activities     0.004139
reason         0.003836
paid           0.002373
failures       0.001775
school         0.001757
nursery        0.001417
guardian       0.001217
```

Across all models, a consistent pattern emerged: academic performance in previous periods—especially G2—was the most reliable indicator of final grades. Socio-demographic and lifestyle factors appeared less influential or inconsistently significant across models. This suggests that, while such background variables provide context, they are not the primary drivers of academic outcomes in this dataset. The results align with theoretical expectations: linear models offer transparency and efficiency when assumptions hold, while tree-based ensembles like random forests excel in capturing non-linear relationships and interactions—though at the cost of interpretability.

In conclusion, the study highlights the trade-offs between simplicity, interpretability, and predictive performance (Figure 1).



Each model offers a unique lens through which student performance can be understood and predicted, with the random forest proving most effective at balancing complexity and accuracy.

Conclusion and Next Steps

This study applied and compared several regression models—including a baseline predictor, linear regression, K-Nearest Neighbors (KNN), decision tree, and random forest—to predict students' final grades using demographic and academic features. The baseline model provided a reference MSE of approximately 20.94, against which all other models showed considerable improvement. Among the models tested, the random forest regressor demonstrated the highest predictive accuracy, achieving the lowest test MSE (1.28) and the highest R^2 score (0.93), indicating it captured a substantial amount of variance in the target variable.

The results consistently highlighted the importance of prior academic performance, especially the G2 (second period grade) and G1 (first period grade) features, as strong predictors of final grades. In addition, variables such as absences, age, and failures also emerged as relevant factors, though their impact varied slightly across models. This aligns with educational research suggesting that continuous assessment and engagement are critical indicators of student success.

However, while tree-based models like the decision tree and random forest achieved strong results, they also carry the risk of overfitting, particularly when model complexity is not carefully tuned. The simpler linear regression model offered more interpretability, though with slightly lower accuracy, making it a reasonable alternative in situations where explainability is prioritized over predictive power.

Next steps for this work could include further tuning of hyperparameters, particularly for the tree-based models, and applying cross-validation techniques to ensure the robustness of the results. Incorporating regularization methods (e.g., Ridge or Lasso regression) might also help in improving the performance of linear models by reducing multicollinearity and overfitting. Additionally, it may be beneficial to explore more advanced ensemble techniques, such as gradient boosting, to assess whether even better performance can be achieved. Finally, expanding the feature set to include behavioral or socio-emotional factors—if available—could provide a more holistic view of student performance determinants.