Statistical Analysis of Student Grades: Predictive Modeling Approaches

Abstract

This project aimed to analyze and predict student final grades using machine learning techniques, specifically Naive Bayes and Ordinary Least Squares (OLS) Regression. The research focused on understanding the relationship between various student characteristics, including alcohol consumption, study time, social habits, and academic performance.

Data Characteristics

Dataset Details (https://www.kaggle.com/datasets/uciml/student-alcohol-consumption)

- **Source**: Survey of secondary school students in Math and Portuguese language courses
- **Total Observations**: 581 unique records
- **Number of Predictors**: 32
- **Key Variables**: Final grades, alcohol consumption, study time, social time, absences

Descriptive Statistics Highlights

- **Final Grades**
  - Mean: 10.46
  - Standard Deviation: 5.20
  - Median: 11.00
- **Alcohol Consumption**
  - Daily Alcohol Consumption (Dalc)
    - Mean: 1.081
    - Standard Deviation: 0.31
  - Weekend Alcohol Consumption (Walc)
    - Mean: 1.785
    - Standard Deviation: 0.97

Methodology

Data Preprocessing

1. Merged two datasets (student-mat and student-por)
2. Removed duplicate columns

3. Excluded rows without final grades (G3)
4. Split data into 70% training and 30% testing sets
5. Set random seed (42) for reproducibility

Modeling Approaches

1. Ordinary Least Squares (OLS) Regression

Model Performance:

- Adjusted R-squared: 0.8206 (82.06% variance explained)
- Residual Standard Error: 2.179
- F-statistic: 155.3 (p < 2.2e^-16)
- Test RMSE: 2.21

Most Influential Predictors:

1. G2 (Second Period Grades)
   - Coefficient: 0.852
   - Significance: p < 2e^-16
2. G1 (First Period Grades)
   - Coefficient: 0.316
   - Significance: p = 1.78e^-7
3. Absences
   - Coefficient: 0.213
   - Significance: p = 0.0002
4. Social Time (goout)
   - Coefficient: -0.406
   - Significance: p = 0.0037
5. Weekend Alcohol Consumption (Walc)
   - Coefficient: 0.373
   - Significance: p = 0.0049

2. Naive Bayes Classification

Model Performance:

- Overall Accuracy: 72.03%

Prediction Breakdown:

- Low Grades (0-10): 73% correctly predicted
- Medium Grades (11-15): 48% correctly predicted
- High Grades (16-20): 74% correctly predicted

Challenges:

- Difficulty in accurately predicting medium-grade cases
- Misclassification of medium grades:
    - 23 low-grade cases classified as medium
    - 29 high-grade cases classified as medium

Key Insights

1. Previous academic performance (G1 and G2) is the strongest predictor of final grades.
2. Social time negatively correlates with academic performance.
3. Absences have a slight positive correlation with grades (potentially due to other underlying factors).
4. Weekend alcohol consumption shows a modest positive correlation with grades.

Recommendations for Future Research

Model Improvement

- Experiment with alternative machine learning models:
    - Random Forests
    - Support Vector Machines
    - Neural Networks

Feature Expansion

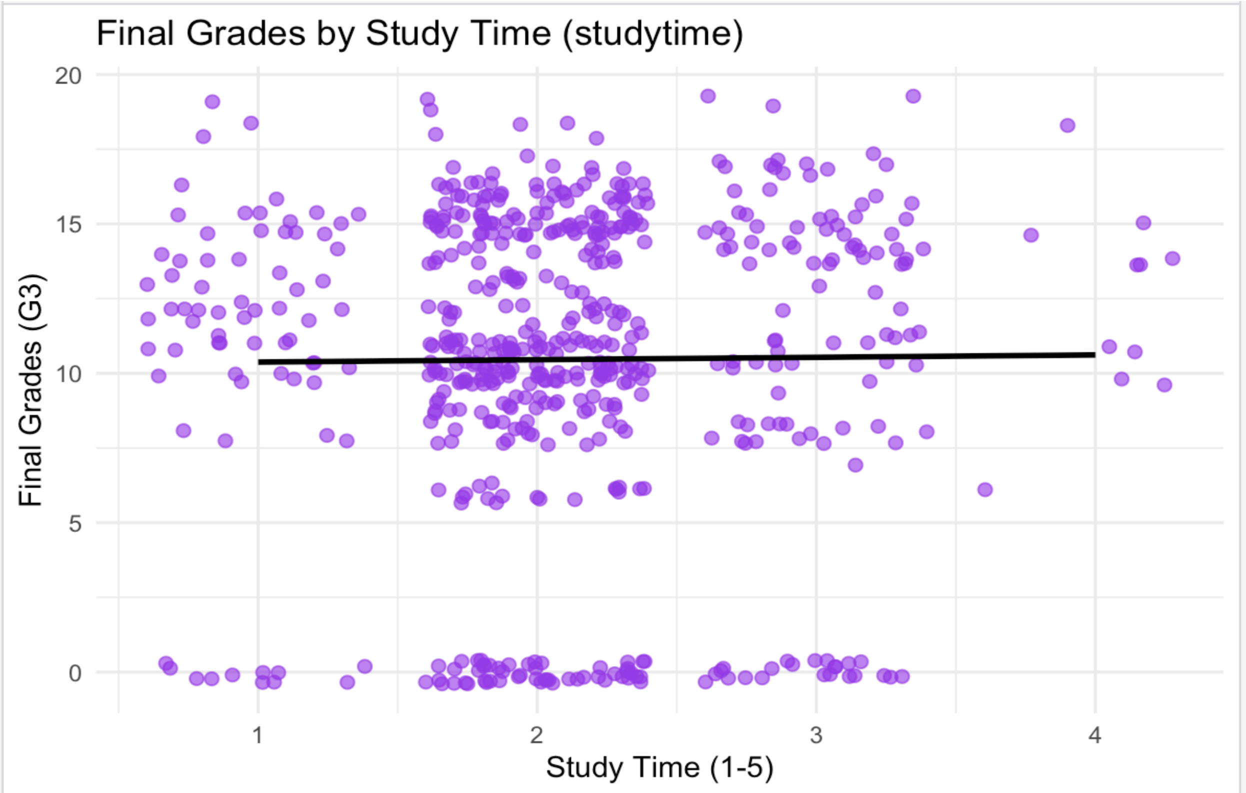
- Investigate additional predictive variables:
    - Extracurricular activities
    - Mental health factors
    - Educational resource access

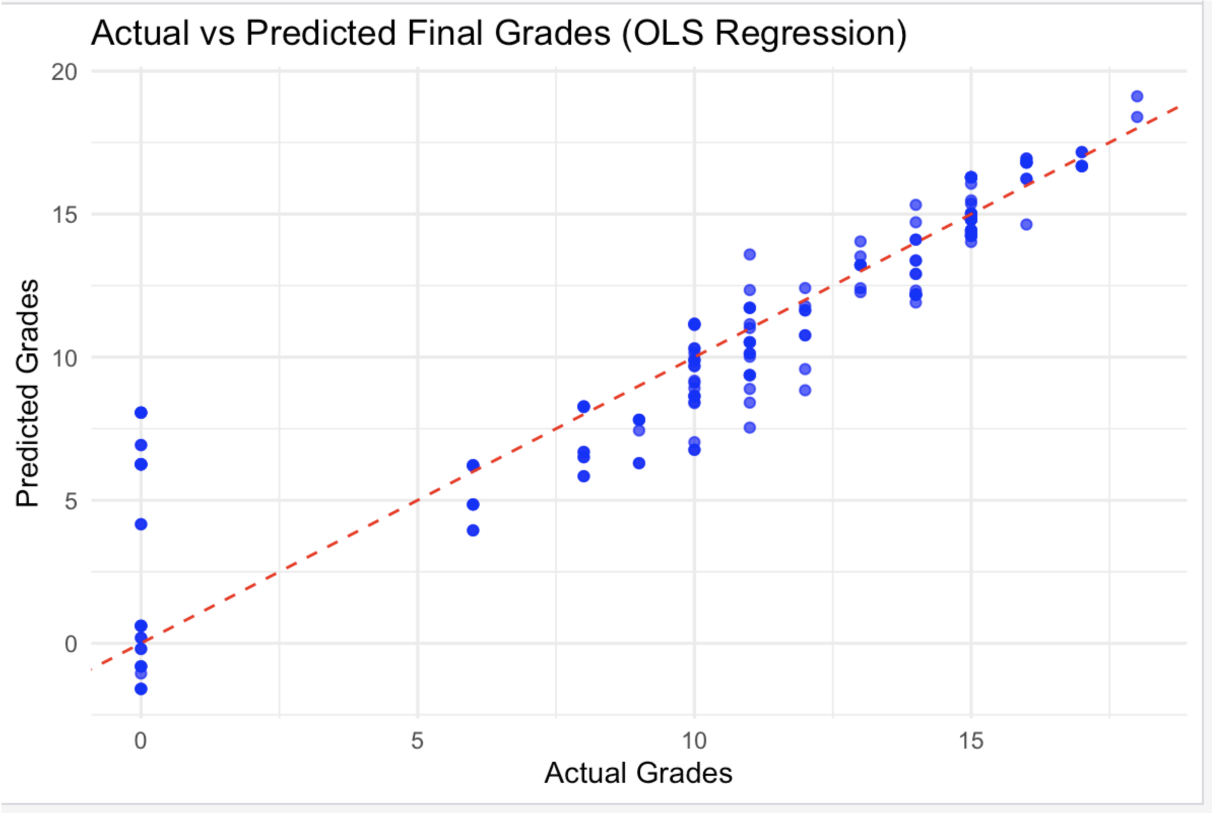
Methodological Enhancements

- Perform hyperparameter tuning
- Explore more sophisticated feature selection techniques
- Develop more granular grade categorization methods

Conclusion

The study demonstrates the potential of predictive modeling in understanding student academic performance. While both OLS Regression and Naive Bayes provided valuable insights, there remains significant room for improvement in predictive accuracy, particularly for medium-grade classifications.


Final Grades by Study Time (studytime)
Final Grades (G3)
20
15
10
5
0
1
2
3
4
Study Time (1-5)


Actual vs Predicted Final Grades (OLS Regression)
Predicted Grades
20
15
10
5
0
0
5
10
15
Actual Grades