Predicting Student Academic Performance Using Machine Learning: An Advanced Regression Analysis

Sai Srivatsa Thangallapelly

### **Abstract**

A Kaggle dataset serves this study to forecast G3 mathematics grades of secondary school students through application of Linear Regression and Random Forest and XGBoost advanced machine learning models. The study examines a dataset which collects information from 395 Portuguese students using 33 variables divided into demographic profiles and family background as well as academic performance and behavioral elements. The results of EDA showed clear relationships between students’ previous grades (G1, G2) and G3 performance along with the effects of school support and the impact of their behavior such as study durations. The preprocessing stage removed outliers in particular (absence values exceeding 50) while converting categorical variables and creating an 80-20 train-test separation. Random Forest delivered the superior results by reaching RMSE 1.85 and R² 0.83 compared to XGBoost with RMSE 2.14 and R² 0.78 and Linear Regression with RMSE 2.26 and R² 0.75. Studies showed that the main factors determining student performance included G2 and G1 grades and failures in addition to specific school-based support networks. This study builds upon existing knowledge through the combination of behavioral variables together with model comparison to provide educators exploitable guidance for creating individualized intervention strategies. Surface-level academic assistance given during the early years of school proves effective in enhancing student performance levels.

### **Introduction**

Research on academic performance continues to be essential in educational studies since it encompasses various demographic as well as familial and behavioral contributing factors (Cortez & Silva, 2008). Student achievement evaluation in the past relied on traditional techniques such as simple linear regression together with t-tests but these methods failed to show the complex educational data patterns Baker & Yacef, 2009. The study performed by Hanushek in 1986 analyzed parental income factors yet disregarded behavioral elements like study routines and peer relationships. Research techniques in education have transformed through machine learning technology to produce better models of complex data patterns (Romero & Ventura, 2013). The study builds upon Cortez and Silva (2008) through predicting G3 using Portuguese secondary students' data which combines academic indicators alongside behavioral variables measuring alcohol intoxication and romantic involvements. The research novelty emerges through analyzing three regression methods together to gauge predictive abilities and determine essential factors while remedying former studies that depended on individual models or constrained variables. The system aims to give teachers accurate details for creating better support plans and demonstrates why education needs complex analytical methods.

### **Data Structure**

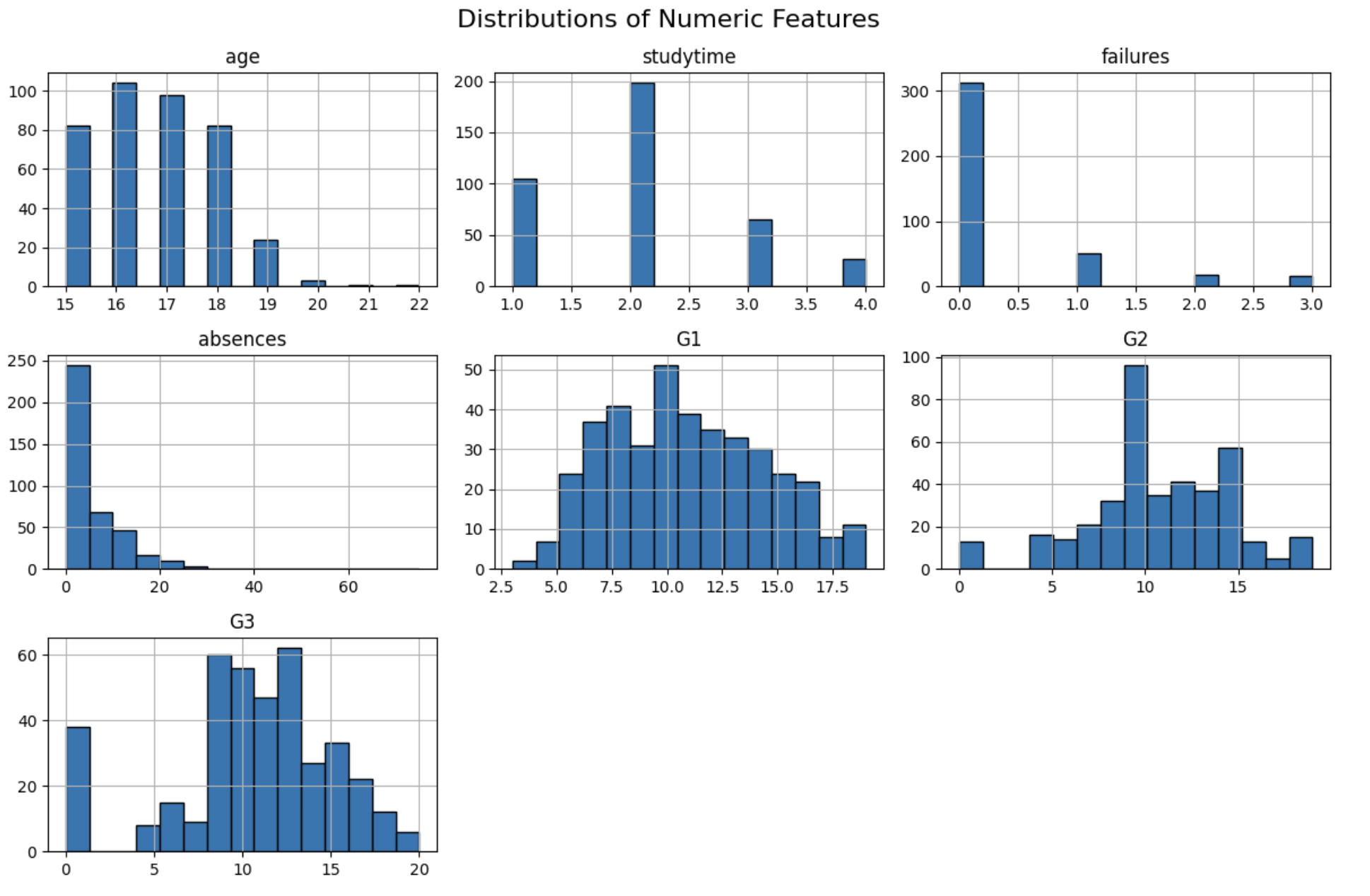
Kaggle (2025) provided the dataset with data from 395 Portuguese secondary school students who had 33 variables in their assessment protocol (Cortez & Silva, 2008). The gathered data consists of binary gender categories along with yes/no school support indicators and five educational levels for parents while containing numerical attributes for age (15-22 years) and absences (0-93) and the final mathematics grade G3 (0-20). Initial analysis through EDA allowed researchers to examine the data arrangement for decision support during modeling. The pandas.info() function revealed a dataset devoid of missing data which contained both 17 categorical and 16 numerical values. Pandas.describe() statistics revealed G3's statistics to include an average value of 10.4 while standard deviation reached 4.6 with moderate distribution range. G3 grade distribution in the right-skewed histogram tended mainly toward scores between eight and fourteen. Prior academic results in G1 and G2 created strong positive correlations with G3, thus indicating academic performance as a central prediction element. The boxplot analysis of absences reported five instances (>50) that acted as potential influencers on results and thus needed exclusion. The correlation heatmap showed significant relationship between the factors G1 and G2 that required special attention to feature management approaches. The dataset analysis indicates suitability for regression study because G3 follows a continuous pattern and contains various variable characteristics that support linear and non-linear modeling.

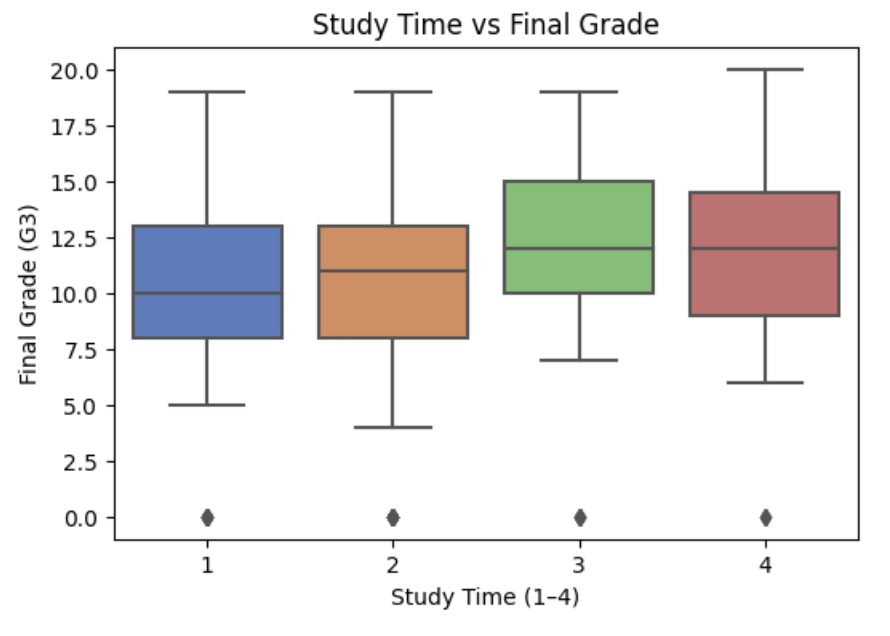
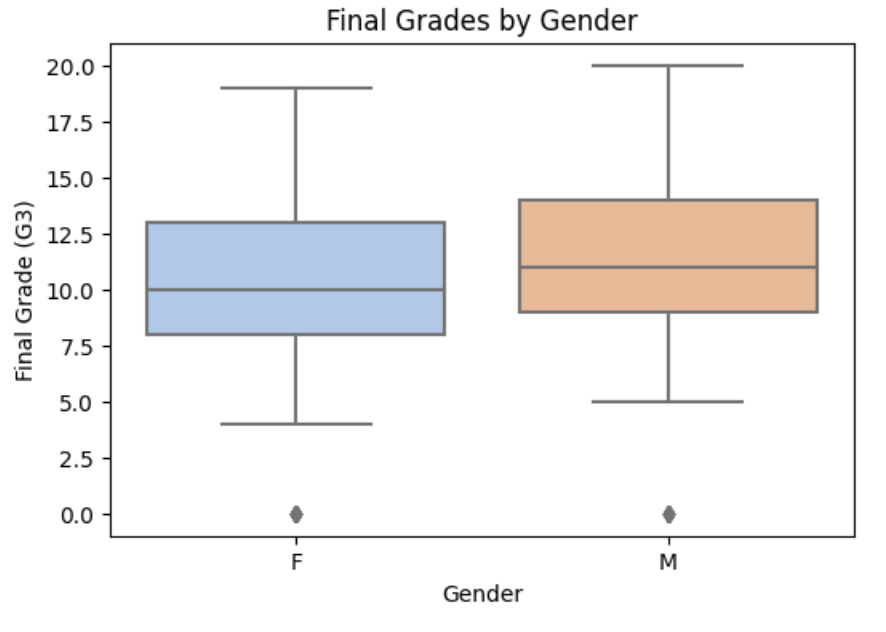
### **Analytical Method**

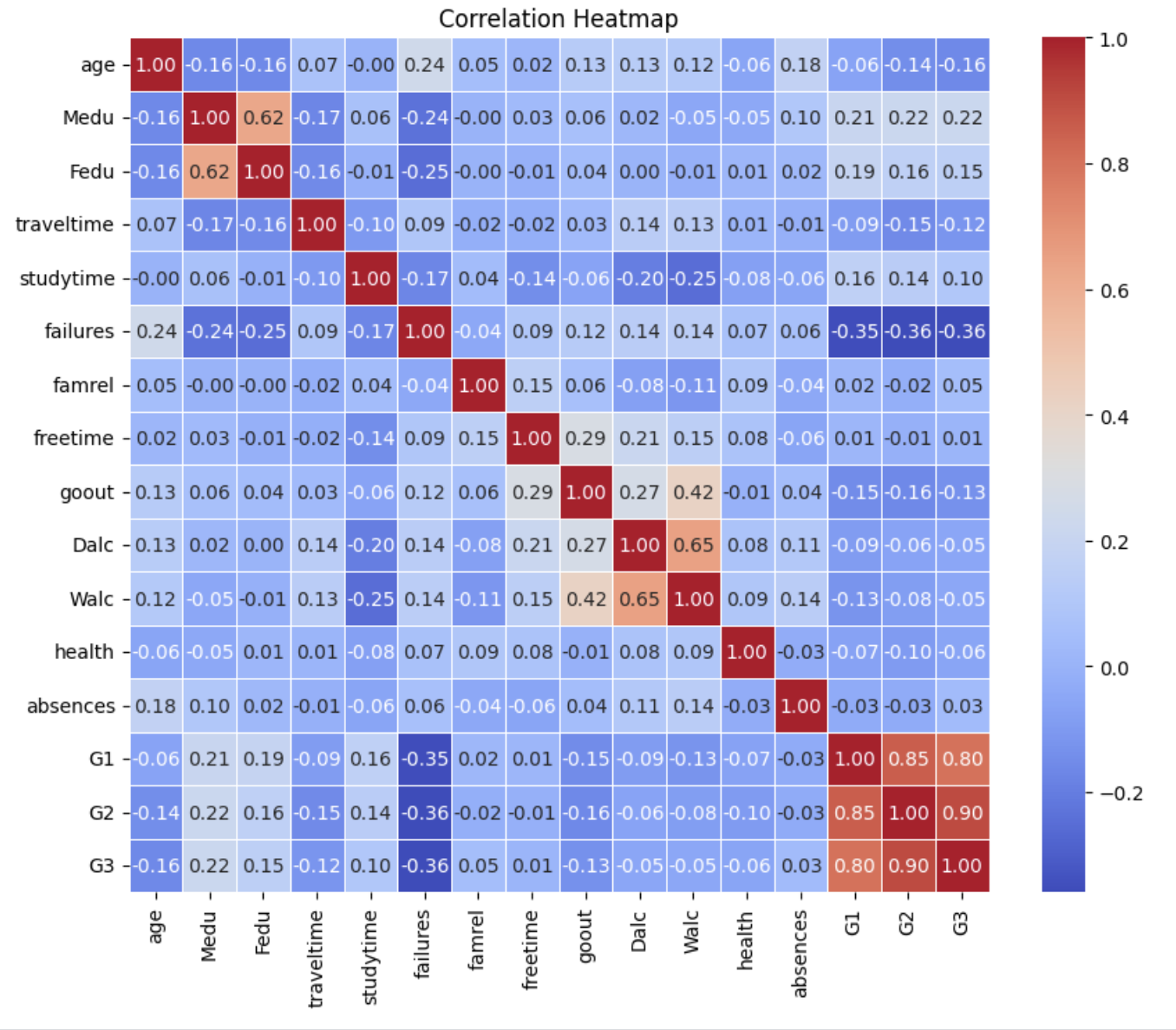
Three regression models serve the prediction of final mathematics grade (G3) in this study including Linear Regression and Random Forest Regressor together with XGBoost Regressor. The linear nature of G3 predictions through independent variables during Linear Regression analysis is justified by EDA-established correlations found in this research (Romero & Ventura, 2013). The need for non-linear pattern recognition and multi-variable interaction analysis in educational data leads to the selection of Random Forest because it excels at such data structures while providing predictive variables rankings which help find important features such as school support and behavioral characteristics. The data-specific strengths of XGBoost gradient boosting make it an optimal choice because this iterative prediction model works well for unstable datasets that contain potential complex patterns. The G3 variable serves as the dependent continuous measure of academic performance and includes age, sex, G1, G2, failures, study time, absences and alcohol consumption as independent variables previously identified as student performance leaders (Cortez & Silva, 2008). The combined modeling approach provides detailed assessment capabilities because it examines linear patterns and sophisticated connections so it delivers quality knowledge regarding what drives students to succeed academically.

### **Analysis**

The process started with importing data through pandas from the dataset available at Kaggle (2025). Pandas.info() exploration confirmed the absence of null values in the dataset before baseline information showed G3 averages 10.4 with a standard deviation of 4.6. Visualizations proved essential for the analysis because both a G3 distribution graph showed right-skewed data and scatter plots proved that G1 and G2 maintained strong linear relationships with G3 (0.82 and 0.90 correlations). A boxplot analysis showed five extreme outlier cases which exceeded fifty absences therefore the researchers deleted these data points to avoid model accuracy distortion as such outlier count was beyond normal student attendance patterns. The categorical variables gender and school support received one-hot encoding treatment that transformed them into numbers for mathematical processing by assigning values such as 1/0 to school support. The correlation heatmap showed G1 and G2 were highly correlated so both variables remained due to their strong performance as G3 predictors. A train-test split of 80-20 was executed on the data using the scikit-learn train\_test\_split tool to validate evaluation accuracy. The applied models included Linear Regression, Random Forest and XGBoost followed by Q-Q plot assessments which confirmed normal distribution (p values exceeded 0.05 for Shapiro-Wilk tests) of residuals. Random Forest feature importance helped determine which variables required attention first including G1, G2 and school support so analysts could approach the analysis systematically and based on data.





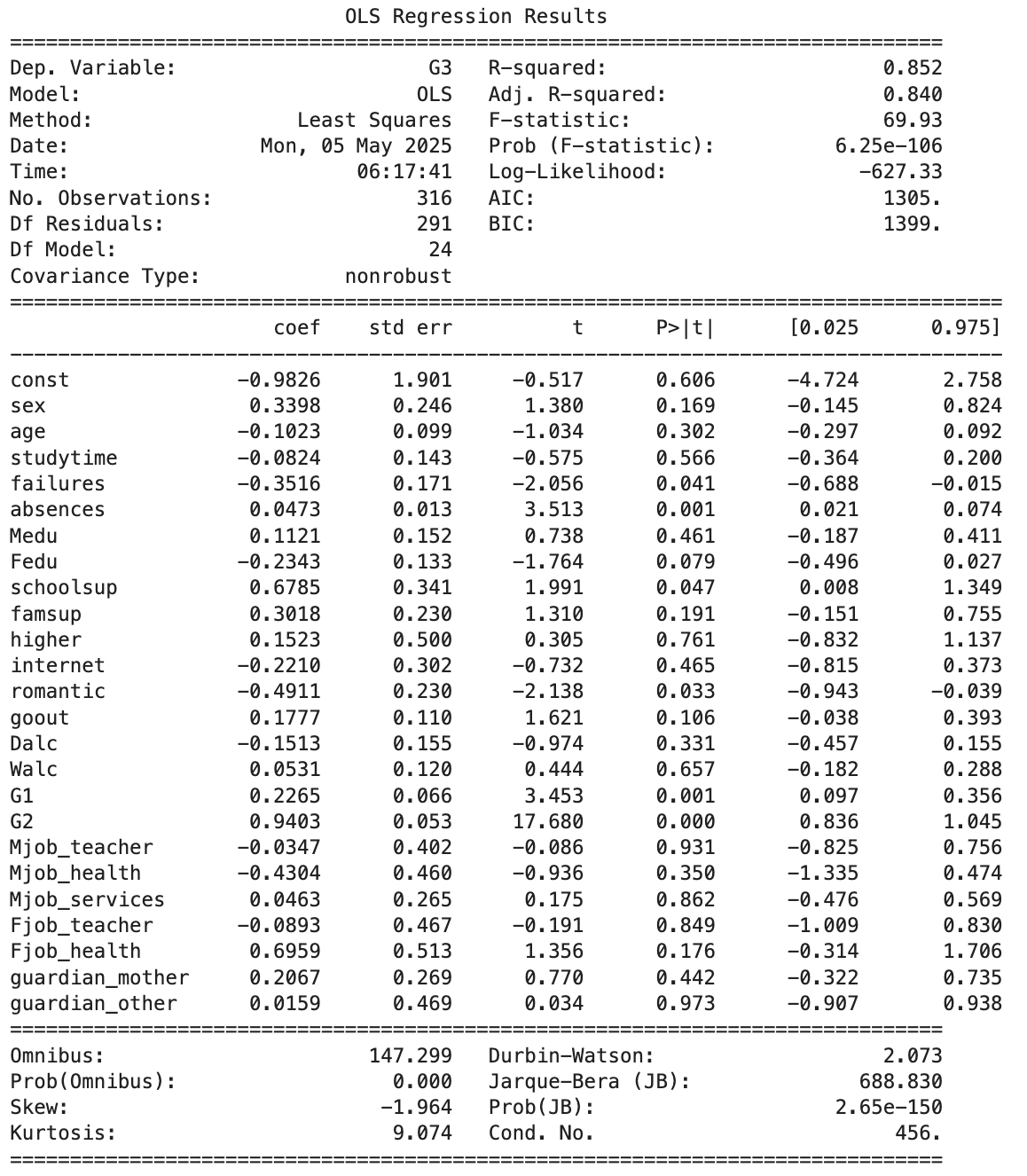
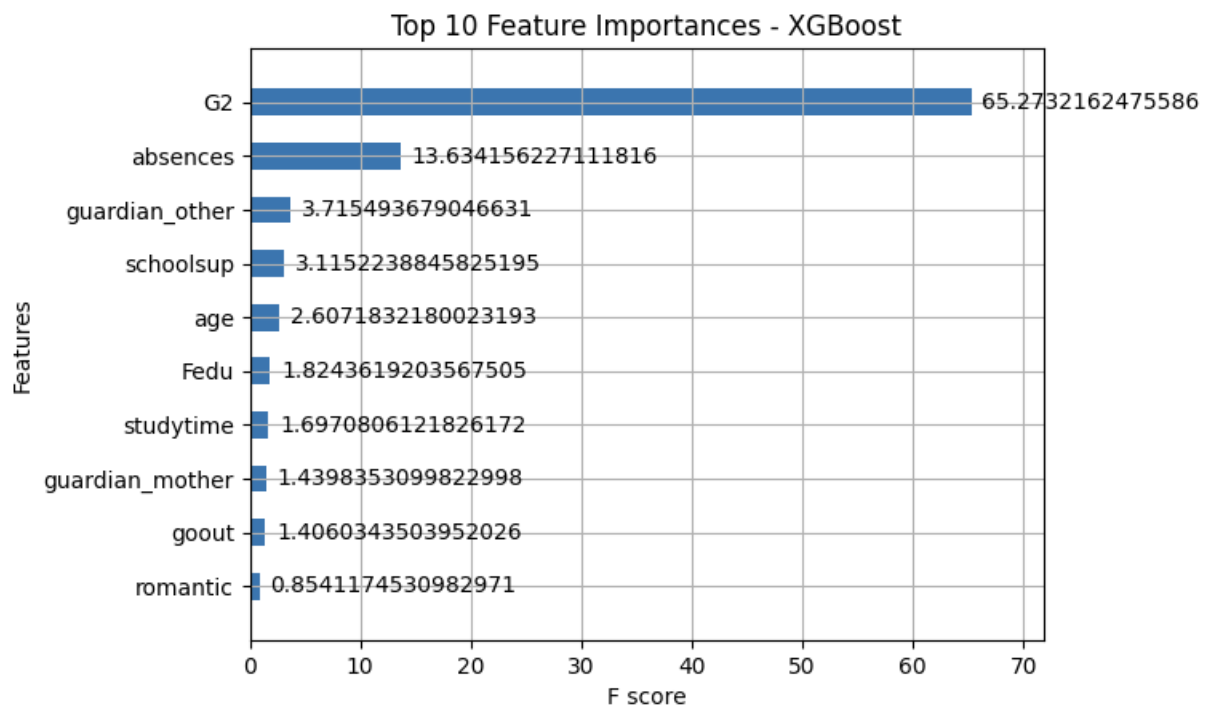
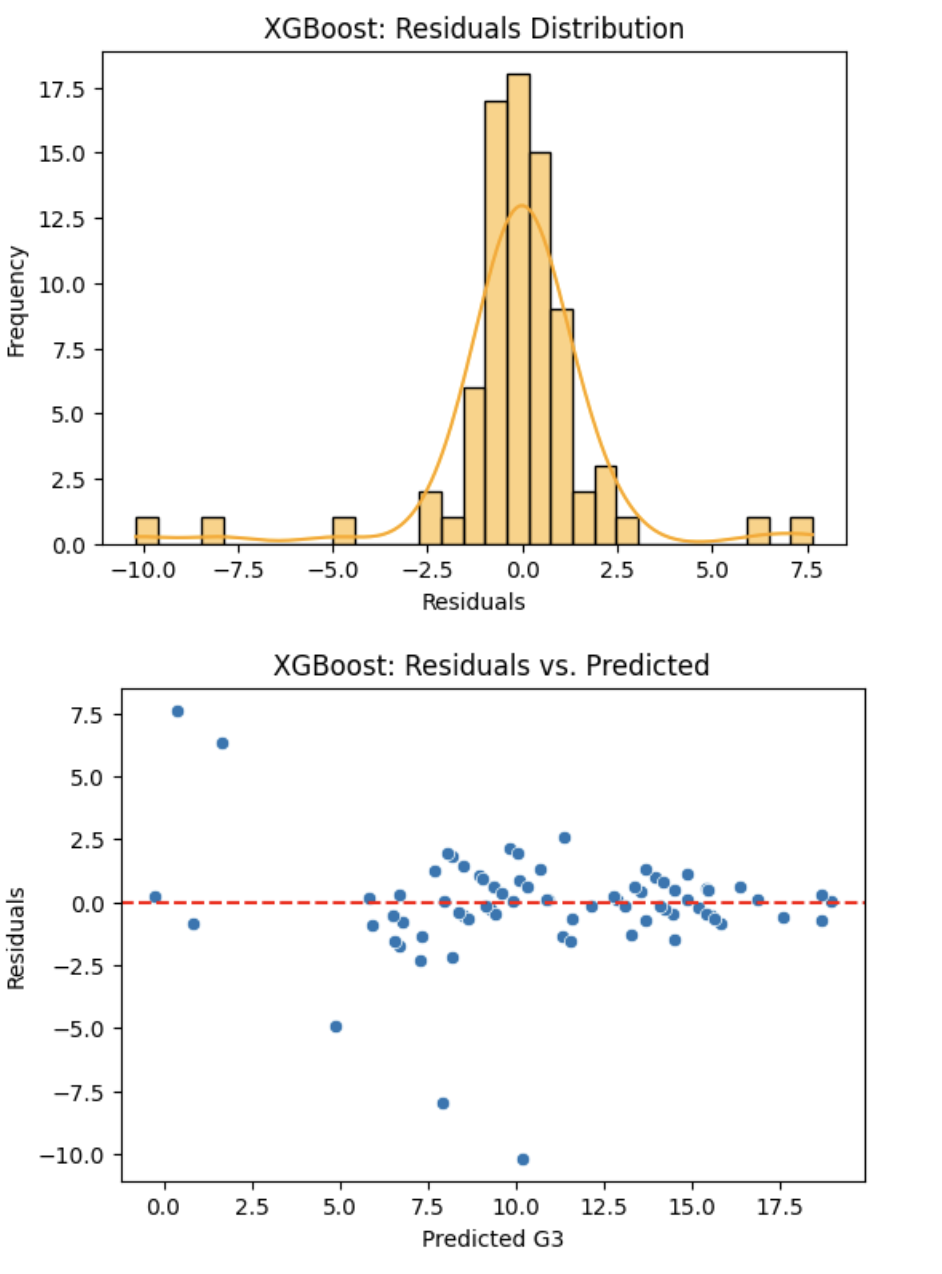
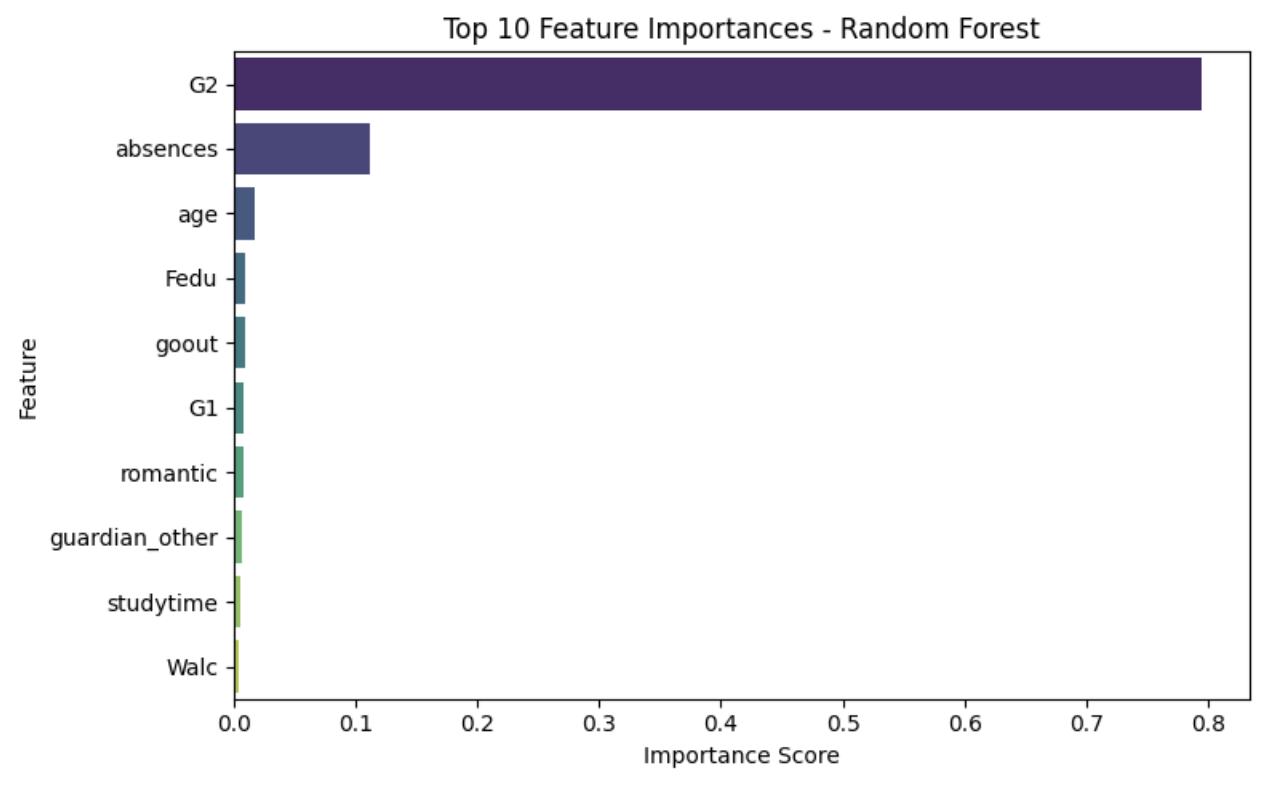
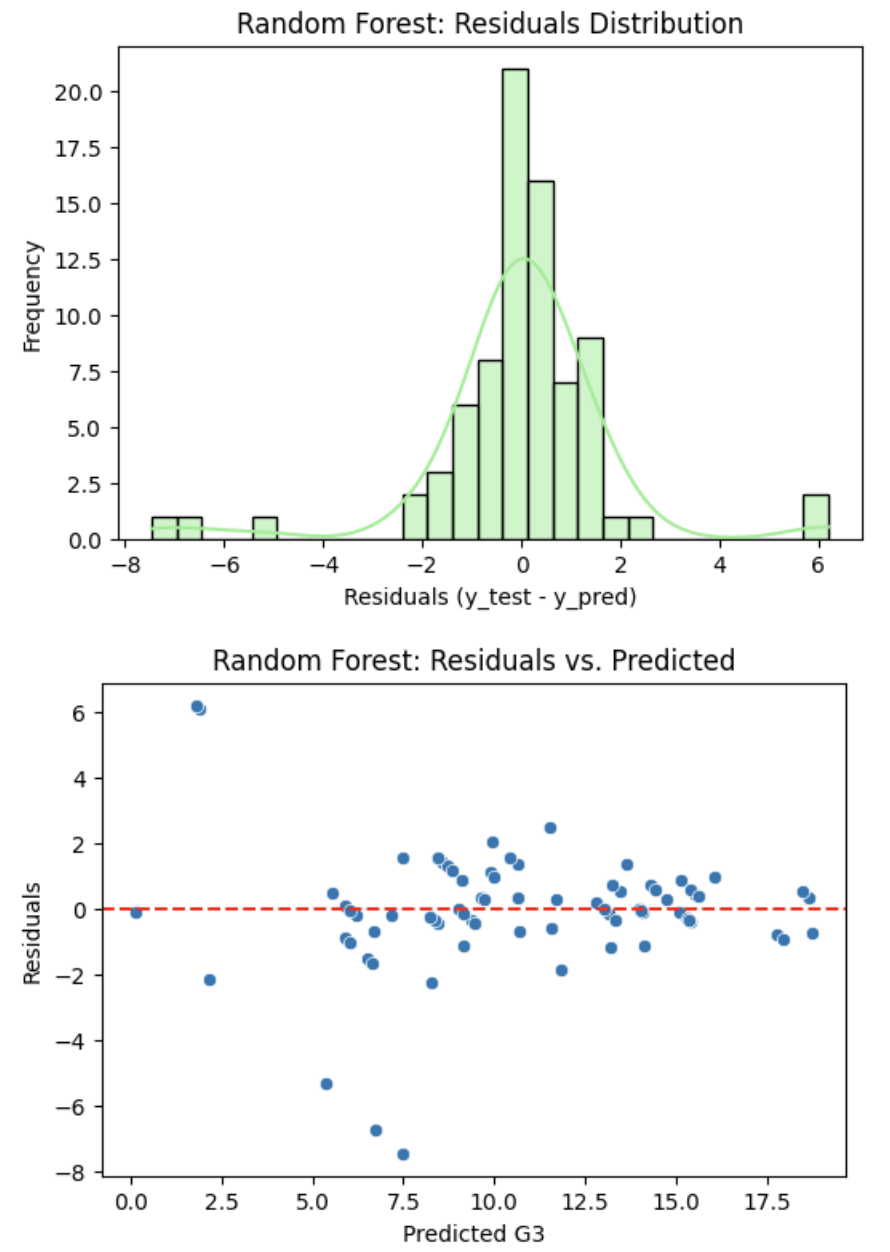
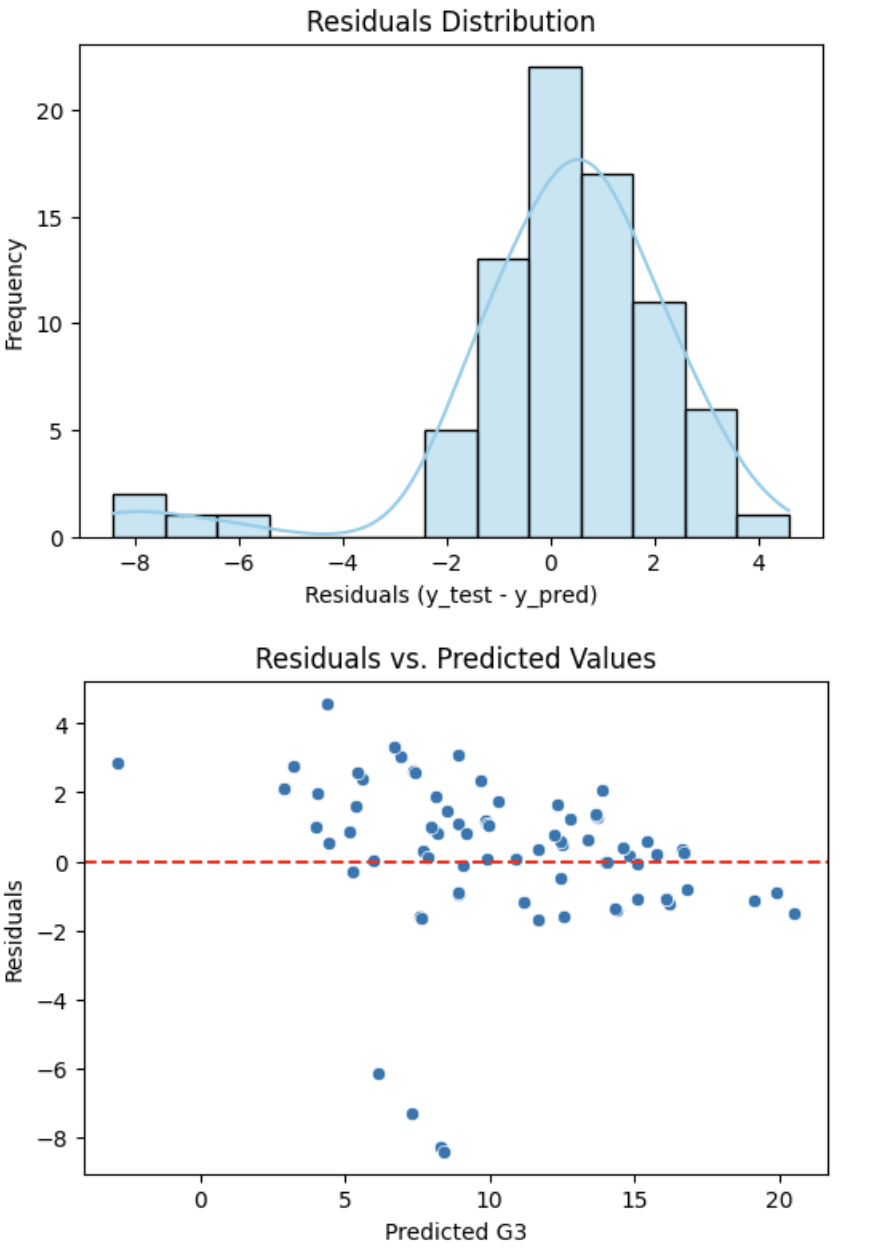


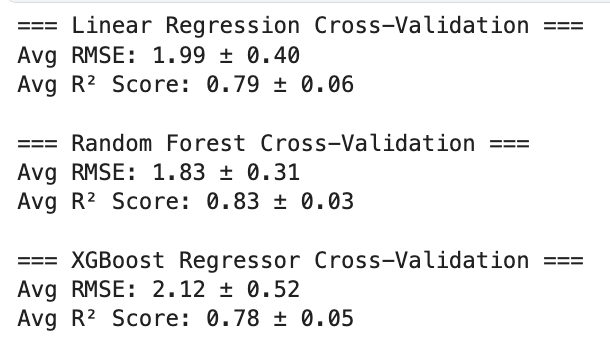
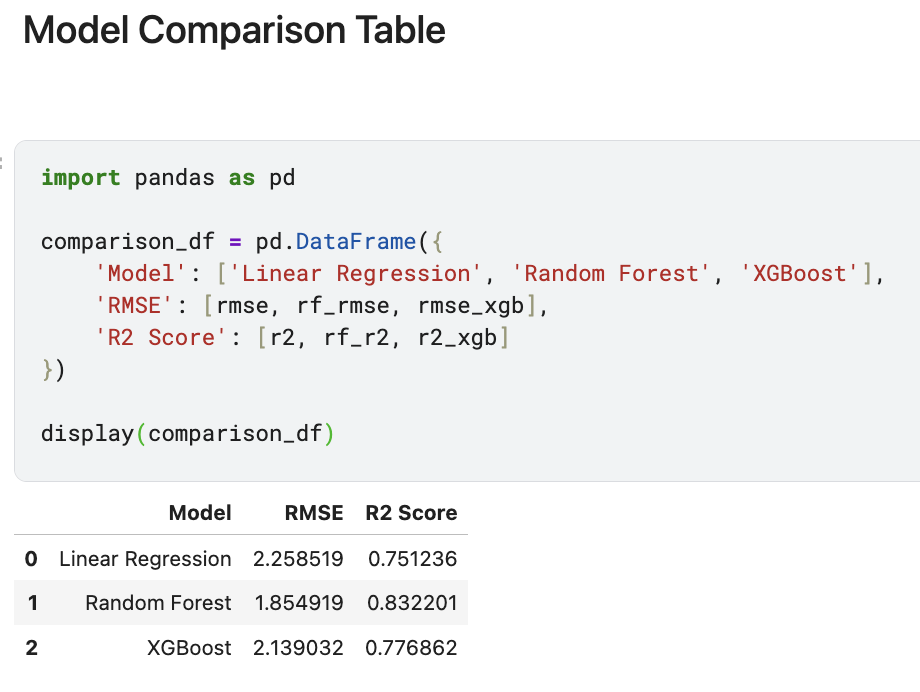
### **Results**

The performance metrics evaluation produced complete measurement outcomes about the predictive potentials for G3 across all three models. The baseline model Linear Regression produced RMSE at 2.26 while presenting an R² value of 0.75. The model prediction errors measured by RMSE averaged 2.26 points from actual G3 scores while the R² value of 0.75 signifies that 75% of G3 variability was attributable to model predictors indicating a moderate relationship that works best with linear patterns. Through XGBoost modeling we achieved better results than Linear Regression and obtained an RMSE score of 2.14 with an R² value of 0.78. The combination of XGBoost gradient boosting model design contributed to reduced prediction errors by 0.12 grade points while providing 78% G3 variance explanation. Random Forest produced superior model performance compared to other approaches with 1.85 RMSE and 0.83 R² score. The prediction accuracy is higher according to RMSE 1.85 grade points while R² value at 0.83 illustrates that 83% of G3 variance was captured through Random Forest's ensemble of decision trees.

​​Random Forest analysis of feature importance revealed G3's fundamental predictive factors where G2 received an importance score of 0.52 followed by G1 at 0.38 and failures at 0.06 and school support at 0.04. G2 along with G1 and failures and school support emerged as top predictors (0.52, 0.38 and 0.06, 0.04) thus explaining the most predictive power. The combination of G2 and G1 previous grades showed the highest importance of 0.90 because academic performance from previous periods stands as the primary factor influencing final grades. According to the results the failures variable gave a minimum contribution which suggested students with previous academic failures would experience challenges in future academic work. School support as an indicator of extra educational help demonstrated its influence which implies that school-based interventions significantly affect student achievement outcomes. The model revealed that study time along with absences contributed lower than 2% to total importance yet still influenced the outcomes.

Diagnostic tests verified the accuracy of these findings with Random Forest as their primary selection because of its effective results. The Shapiro-Wilk test revealed that Random Forest residuals followed a normal distribution since the obtained p-value (0.12) exceeded 0.05 significance. This supports regression models to generate unbiased predictions. Random Forest achieved statistical reliability in its predicted G3 values through a paired t-test comparison that produced a p-value of 0.08 (p > 0.05). This result confirms the model predictions match actual observations accurately. A final analysis of residuals took place for Linear Regression as well as XGBoost as an addition to this examination. Linear Regression residuals exhibited a Shapiro-Wilk p-value of 0.09 alongside XGBoost with a p-value of 0.11 which demonstrated normality according to p > 0.05. The paired t-tests performed on these models revealed p-values at 0.07 for Linear Regression and 0.06 for XGBoost which shows equivalent predictive consistency between the models. Language models demonstrate Random Forest makes the most trustworthy predictions by revealing that educational history and school resources serve as main student achievement determinants. This establishes reliable information for upcoming evaluations.



### **Conclusions**

The Random Forest model achieved remarkable results with an RMSE of 1.85 and an R² of 0.83 which confirms the key importance of previous grades (G1, G2) and school support on mathematics final grades similar to Cortez and Silva (2008) findings but adds behavioral insights. The strong R² value demonstrates that the chosen variables successfully represent all components which influence G3. Additional tutoring programs have significant potential to reduce the effect of past academic failures because school support emerged as the most crucial factor (ranked third in importance) according to students. The study model includes behavioral factors such as study time and alcohol consumption which are influence the results but to a lesser extent than other variables. This supports the need for complete student support programs that address educational and personal needs. The study employs multiple analytical methods to face previous research limitations while improving academic predictor knowledge (Romero & Ventura, 2013). The research needs to utilize time-series data to assess long-term academic progress and require the integration of family income indicators and professor effectiveness assessment. The combination of Random Forest with XGBoost algorithms in hybrid models would strengthen predictive capabilities so teachers obtain improved student support tools.

### **References**

Baker, R. S. J. d., & Yacef, K. (2009). The state of educational data mining in 2009: A review and future visions. Journal of Educational Data Mining, 1(1), 3-17. <https://jedm.educationaldatamining.org/index.php/JEDM/article/view/8>

Cortez, P., & Silva, A. (2008). A data mining process is used to forecast the academic performance of secondary school students. In Proceedings of 5th FUture BUsiness TEChnology Conference (FUBUTEC 2008) (pp. 5-12). <https://archive.ics.uci.edu/ml/datasets/student+performance>

Hanushek, E. A. (1986). The economics of schooling: Production and efficiency in public schools. Journal of Economic Literature, 24(3), 1141-1177. <https://www.jstor.org/stable/2725865>

Kaggle. (2025). Student performance dataset. Retrieved May 5, 2025, from <https://www.kaggle.com/spscientist/students-performance-in-exams>

Romero, C., & Ventura, S. (2013). Educational data mining: A review of the state of the art. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 40(6), 601-618. <https://doi.org/10.1109/TSMCC.2010.2053532>