Select a page:

Final Submission

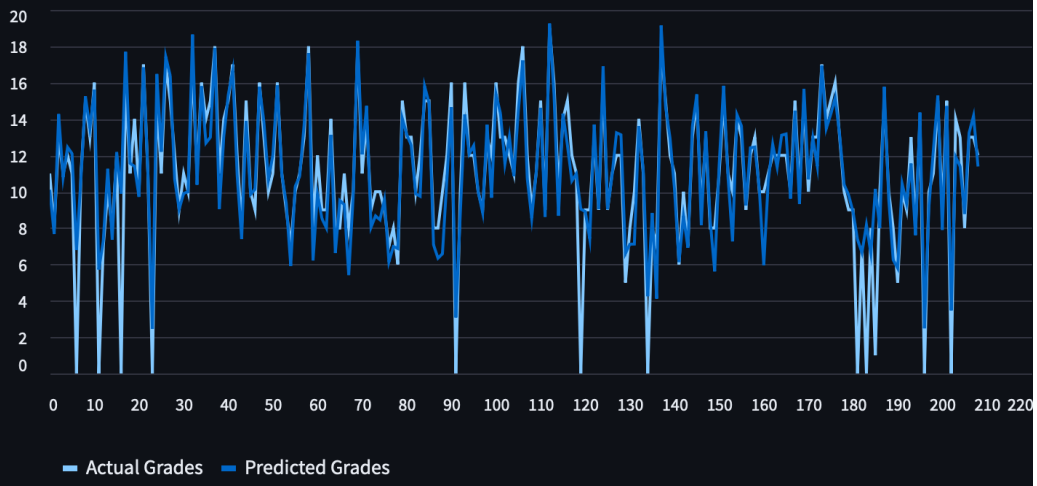
Machine Learning Project

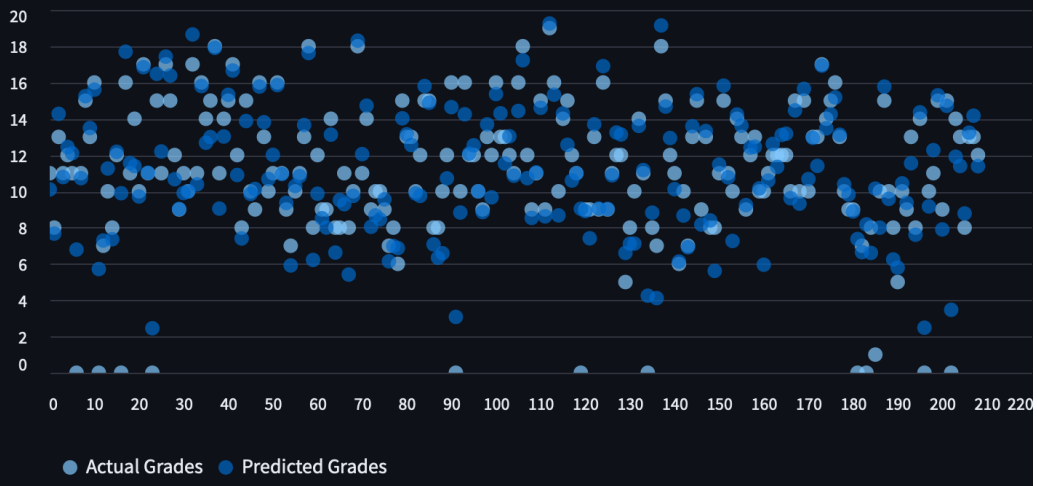
Final Submission

Please refer to the Proposal tab for further details on background and insights to our research problem.

Visualizations

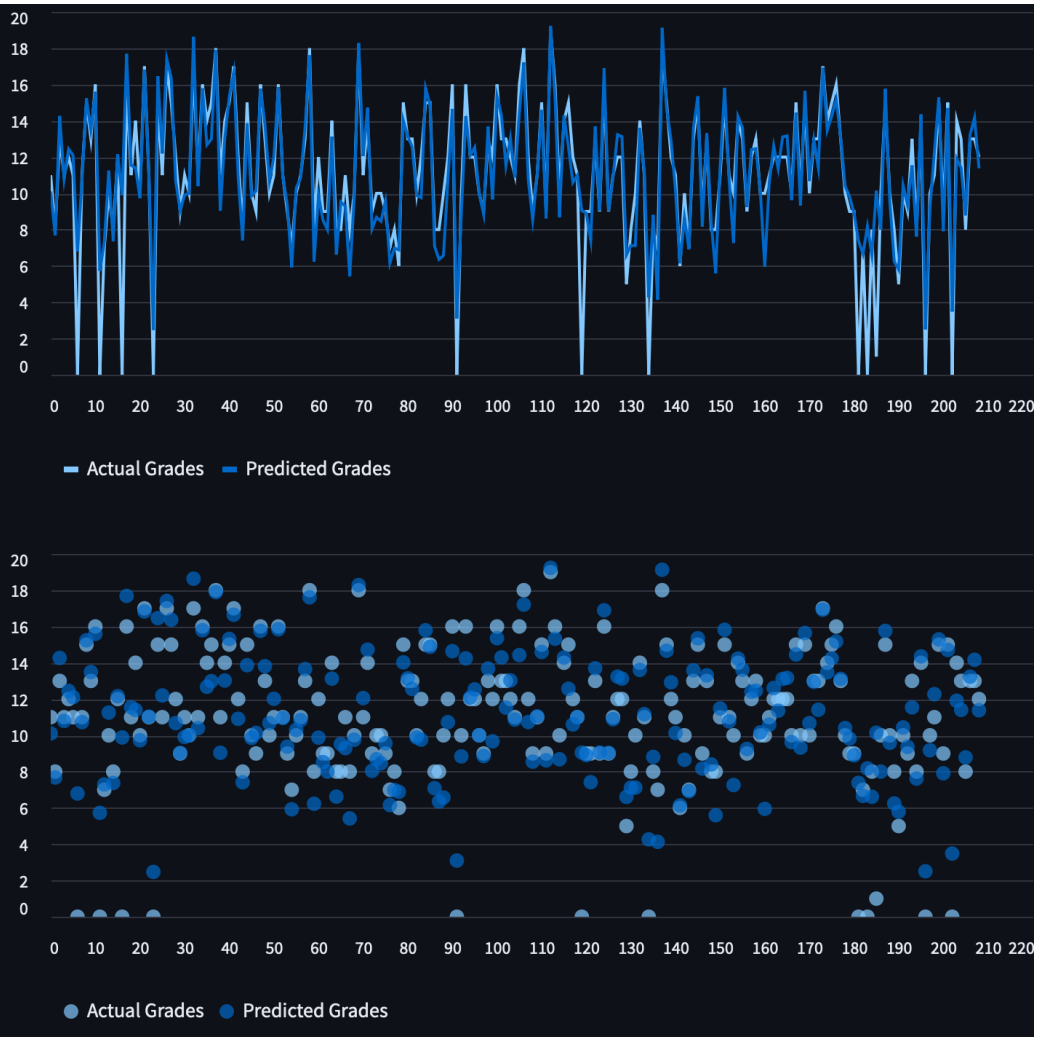
Linear Regression Scatterplot



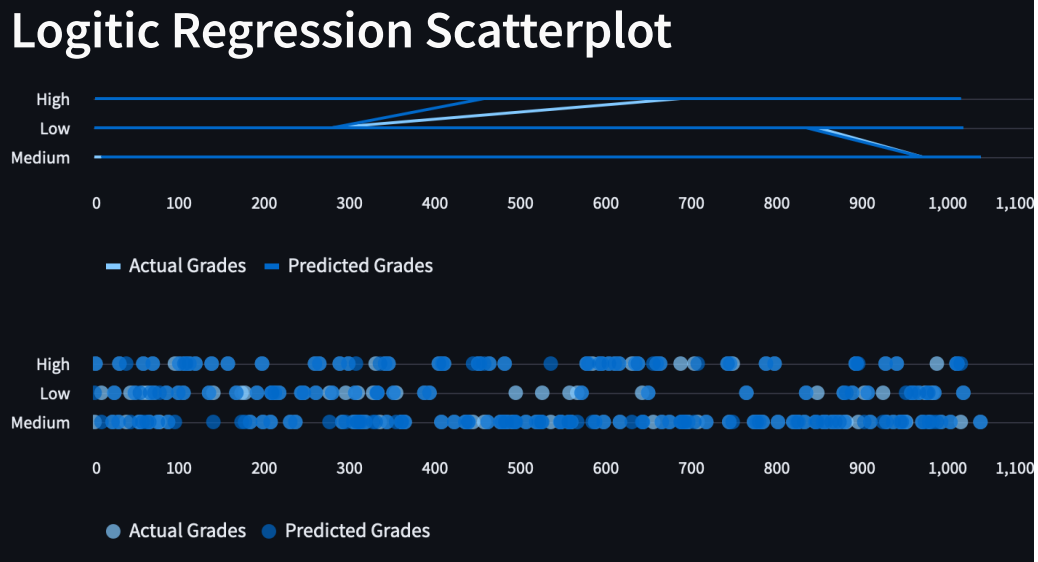


Scatter Plot

Lasso Scatterplot



Scatter Plot



Logistic Regression Scatter Plot

Results and Discussion

Linear Regression Analysis:

These metrics provide a quantitative summary of the model's performance. The R-squared score of 0.81 demonstrates that 81% of the variance in the actual grades is explained by the model, which reflects strong predictive power. The Mean Squared Error (MSE) and Mean Absolute Error (MAE) indicate that the model's predictions are generally close to the actual values, with minimal large deviations. The relatively low MAE, in particular, highlights that the model's errors are consistently small, signifying a high level of accuracy in its predictions.

Analysis of Model Performance: The model performs well overall, as evidenced by the high R-squared score and the alignment between predicted and actual grades. The low MSE suggests that significant outliers are rare, while the low MAE confirms that errors are minor for most data points. The model appears to effectively capture the linear relationships in the dataset. However, discrepancies in the predictions, particularly in instances where grades deviate significantly, may indicate the presence of influential outliers or non-linear patterns that the model cannot adequately address. These limitations could hinder its ability to generalize to more complex relationships in the data.

Why Linear Regression Was Chosen: Linear Regression was selected due to its simplicity and interpretability. It is particularly suited for predicting continuous outcomes, making it an ideal choice for forecasting student grades. The model allows for straightforward interpretation of coefficients, enabling an understanding of how input features impact the predicted grades. Additionally, Linear Regression is computationally efficient and works well when the relationship between features and the target variable is linear. This combination of simplicity, efficiency, and interpretability makes it a reliable tool for analyzing and predicting grade trends.

Logistic Regression Analysis: These metrics provide a quantitative summary of the model's performance. The classification metrics indicate that the model achieves an overall accuracy of 81%, which is a strong result for this multi-class classification problem. The precision, recall, and F1-score vary across the three classes ("High," "Medium," and "Low"), with the model performing best in predicting the "High" and "Medium" categories. The "Low" category shows lower precision and recall, suggesting room for improvement in handling this class.

Analysis of Model Performance: The model performs well overall, achieving balanced scores across most categories. It performs particularly well in predicting "High" grades, with a precision of 90% and an F1-score of 87%. However, it struggles somewhat with the "Low" category, where recall is only 67%, indicating that some low grades are being misclassified. This could be due to the class imbalance or overlapping feature distributions. Logistic Regression's ability to capture linear relationships between features and the probability of a particular grade range makes it a suitable choice for this problem. However, it may struggle if the decision boundaries are not strictly linear.

Why Logistic Regression was Chosen: Logistic Regression was chosen for its simplicity and interpretability. It provides a direct probabilistic approach to predicting grade categories, making it easy to understand the impact of different features on classification. The model's coefficients further help identify which features most strongly influence grade predictions, such as `G2` and `G1`. This interpretability makes it a valuable tool for analyzing the relationships between features and categorical outcomes.

Lasso Regression Analysis: These metrics provide a quantitative summary of the model's performance. An R-squared score of 0.81 indicates that 81% of the variance in the actual grades is explained by the model, which aligns closely with the performance of Linear Regression. The Mean Squared Error (MSE) of 2.89 and Mean Absolute Error (MAE) of 0.89 suggest that the model's predictions deviate minimally from the actual values on average.

Analysis of Model Performance: Lasso Regression performs similarly to Linear Regression, with nearly identical R-squared scores and slightly improved error metrics (lower MSE and MAE). The use of L1 regularization in Lasso helped shrink some coefficients toward zero, potentially mitigating the impact of less important features and improving generalization. However, since the underlying relationship appears mostly linear, the improvements are marginal. Like Linear Regression, Lasso may struggle with capturing any non-linear patterns in the data.

Why Lasso Regression was Chosen: Lasso Regression was selected to incorporate regularization and address potential issues of overfitting by penalizing large coefficients. This method also aids in feature selection by driving insignificant coefficients to zero, which can improve interpretability and model robustness. This makes Lasso an attractive alternative to Linear Regression, particularly when dealing with datasets that may contain irrelevant or redundant features.

Comparison of Models: All three models—Linear Regression, Logistic Regression, and Lasso Regression—demonstrate strong performance, but they have distinct strengths and limitations:

- Linear Regression explains 81% of the variance in grades with minimal errors, making it a strong choice for continuous prediction tasks. Its simplicity and interpretability are major advantages, though it may struggle with non-linear patterns.
- Logistic Regression excels in categorizing grades into "High," "Medium," and "Low," achieving balanced accuracy across categories, especially for "High" grades. However, it struggles slightly with the "Low" category, suggesting potential room for improvement with non-linear classifiers.
- Lasso Regression slightly improves on Linear Regression in terms of error metrics while adding regularization to enhance generalization. Its R-squared score is nearly identical, showing comparable explanatory power while reducing potential overfitting.

Next Steps:

1. Feature Engineering: Incorporate additional features or transform existing ones to better capture potential non-linear relationships.
2. Model Exploration: Test non-linear models such as Random Forests or Gradient Boosted Trees to address any non-linear dependencies.
3. Hyperparameter Tuning: Optimize Logistic Regression and Lasso hyperparameters to further improve performance.
4. Address Class Imbalance: Consider techniques such as SMOTE for Logistic Regression to handle imbalanced classes.

Final Analysis: The analysis shows that student performance is closely tied to factors like prior grades (`G1` , `G2`), parental education, and student activities, which are highly predictive of final grades. Models like Linear Regression and Logistic Regression performed well, suggesting that the dataset captures strong

linear relationships. However, the difficulty in predicting "Low" grades points to challenges like class imbalance or subtle patterns that linear models miss. Addressing these gaps with techniques like SMOTE or more flexible models like Random Forests could better account for underrepresented groups and students with unique circumstances.

This project highlights how machine learning can be a valuable tool for educators, helping identify struggling students early so they can get the support they need. By leveraging data-driven insights, schools could better personalize teaching and improve outcomes. Moving forward, exploring more advanced models and additional features, like attendance trends, could uncover deeper insights. The work here is a step toward creating smarter, more equitable systems that give every student a better chance to succeed.

Project Contributions

Below is a table showing each contributor and their tasks for the project.

	Contributor	Tasks
0	Sreehitha	Lasso Regression Model, Visualizations, Metrics, Streamlit Page Setup
1	Anushya	Final Script, Video
2	Areeba	Power Point, Future Steps, Conclusion, Final Submission
3	Saanvi	Logisitc Regression Model + Visualizations + Metrics
4	Tanvi	Results, Analysis

Gantt Chart