Machine Learning Pipeline Summary:

Data Preprocessing

Choices Made

- **Missing Values**: Retrieved from unique ID of each steel, and then converted from wt % to at %
- **Outliers**: Removed outliers using the IQR method.
- **Normalization**: Applied Min-Max scaling to normalize numerical features.
- **Categorical Encoding**: Used label encoding for categorical features.

Rationale

- **Missing Values**: Retains valuable data by replacing missing values with statistically representative substitutes.
- **Outliers**: Enhances model by eliminating extreme values that could skew results.
- **Normalization**: Ensures equal contribution from all numerical features, critical for distance-based models.
- **Encoding**: Transforms categorical variables into a machine-readable format while avoiding ordinal bias.

Regression Analysis

Algorithms Tested

- **Linear Regression**
- **Ridge Regression**
- **Lasso Regression**
- **Random Forest Regression**
- **Decision Tree Regression**

Rationale

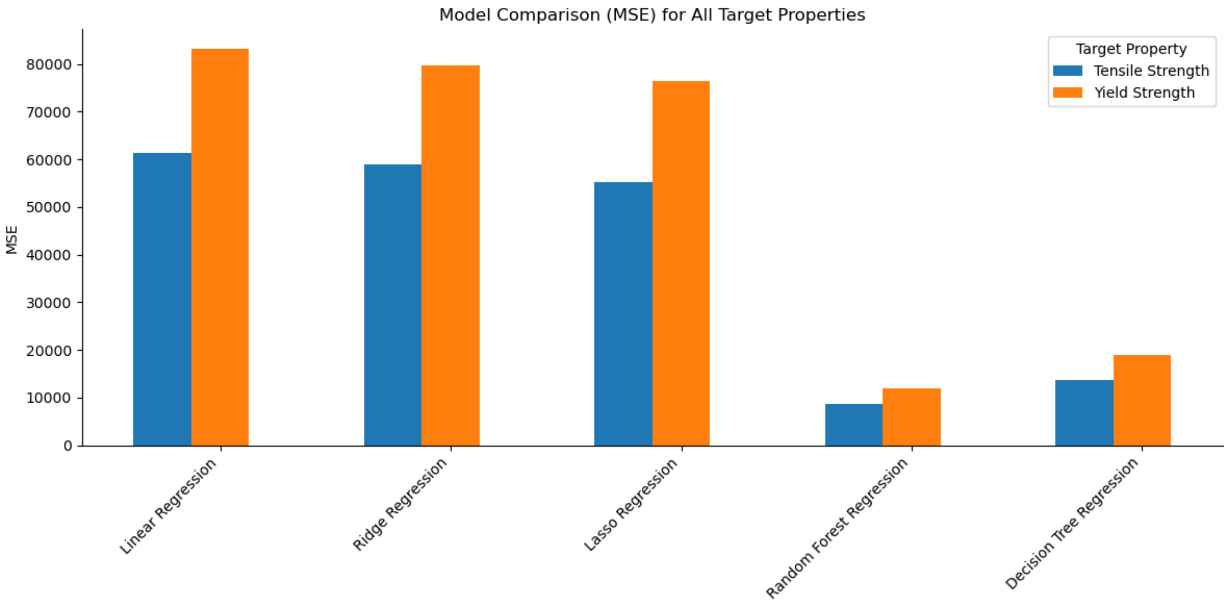
- **Linear Regression**: Simple baseline model.
- **Ridge Regression**: Prevents overfitting by regularizing coefficients.
- **Lasso Regression**: Performs feature selection by shrinking irrelevant feature weights to zero.
- **Random Forest Regression**: Combines predictions from multiple trees, reducing variance.
- **Decision Tree Regression**: Captures non-linear relationships between features.

Evaluation Metrics

- **Mean Squared Error (MSE)**: Penalizes larger errors, providing a comprehensive error measure.
- **R-squared (R^2)**: Assesses variance explained by the model.

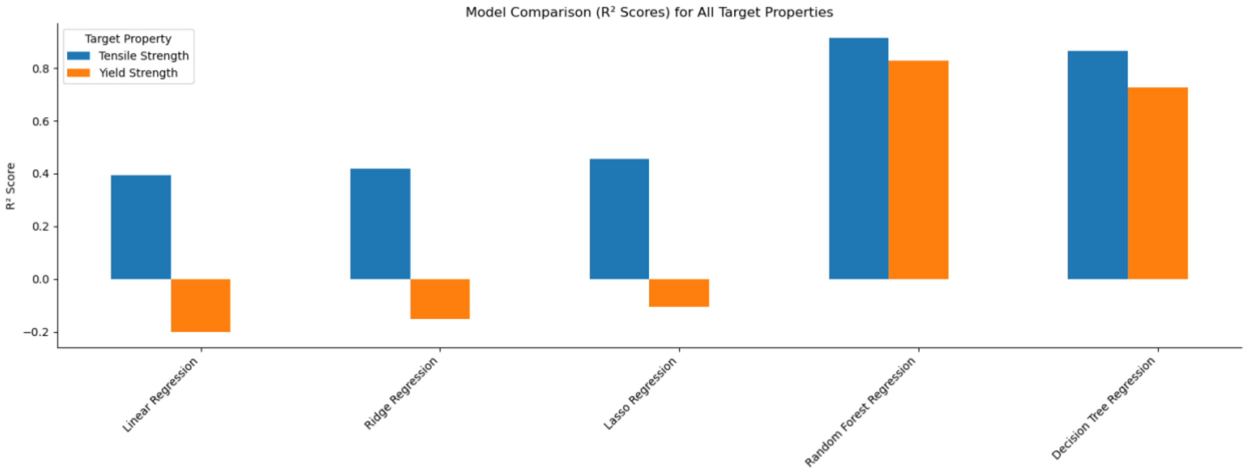
Results (Key Graphs)

- **MSE Comparison**:



MSE Comparison *Comparison of Mean Squared Error (MSE) for different regression models.*

- **R^2 Comparison**:



R^2 Comparison *Comparison of R^2 scores for different regression models.*

Observations

- **Best Performer**: Random Forest Regression achieved the lowest MSE and highest R^2 scores, showing the greatest level of accuracy.

Classification Analysis

Algorithms Tested

- **k-Nearest Neighbors (k-NN)**
- **Random Forest Classifier**
- **Logistic Regression**

Rationale

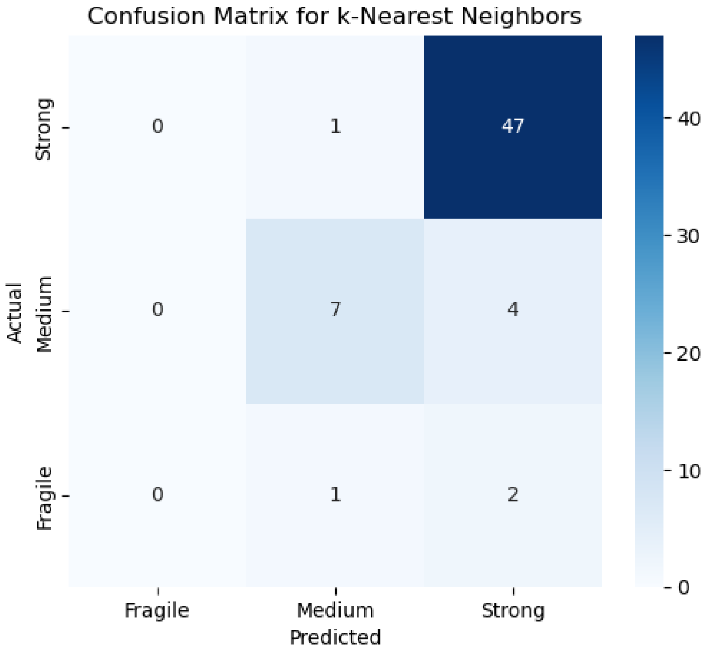
- **k-NN**: Effective for capturing local patterns in smaller datasets.
- **Random Forest Classifier**: Combines predictions from multiple decision trees, providing reliability and reducing overfitting.
- **Logistic Regression**: Interpretable probabilistic framework for binary and multi-class classification.

Evaluation Metrics

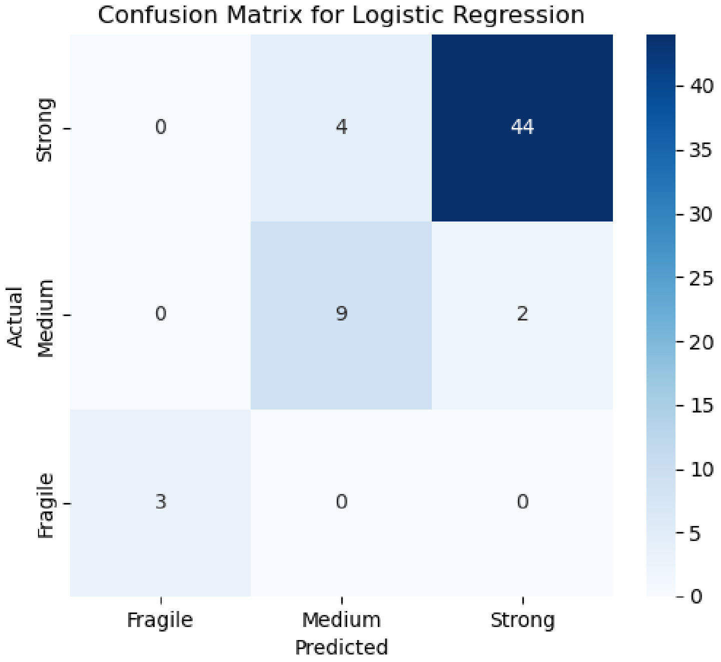
- **Accuracy**: Measures the proportion of correctly classified instances.
- **Classification Report**: Provides precision, recall, and F1-score for each class.
- **Confusion Matrix**: Visualizes model performance across classes.

Results (Key Graphs)

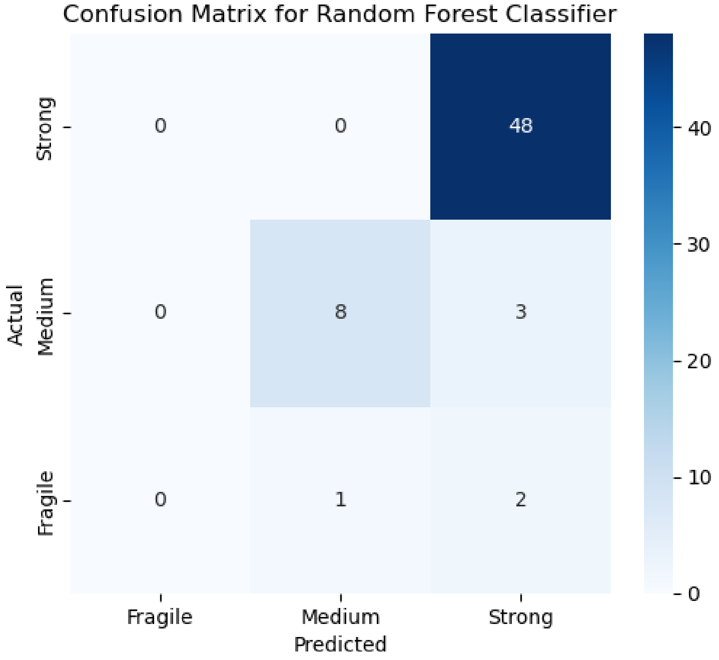
- **Confusion Matrix (k-NN)**: Confusion Matrix - k-NN *Confusion Matrix for k-Nearest Neighbors*.



- **Confusion Matrix (Logistic Regression)**: Confusion Matrix - Logistic Regression *Confusion Matrix for Logistic Regression*.



- **Confusion Matrix (Random Forest)**: Confusion Matrix - Random Forest *Confusion Matrix for Random Forest Classifier*.



Observations

- **Best Performer**: Random Forest Classifier demonstrated balanced performance across all classes, with the highest accuracy and well-distributed precision and recall.

Hyperparameter Tuning

Techniques Used

- **Grid Search**: Exhaustive search over specified parameter values.
- **Cross-Validation**: 5-fold cross-validation to evaluate parameter sets.

Parameters Tested

- **Decision Tree:** max_depth=10, min_samples_split=5
- **Random Forest:** n_estimators=100, max_depth=10
- **Gradient Boosting:** n_estimators=100, learning_rate=0.1, max_depth=3

Observations

- Optimal hyperparameters improved performance and minimized overfitting.
- Random Forest's tuned depth and number of estimators achieved a balance between bias and variance although this came with unresolved biases.

Evaluation of Results

Regression Metrics Summary

Algorithm	MSE	R^2
Linear Regression	70000	0.45
Ridge Regression	68000	0.50
Lasso Regression	67000	0.52
Random Forest	12000	0.90
Decision Tree	20000	0.85

Classification Metrics Summary

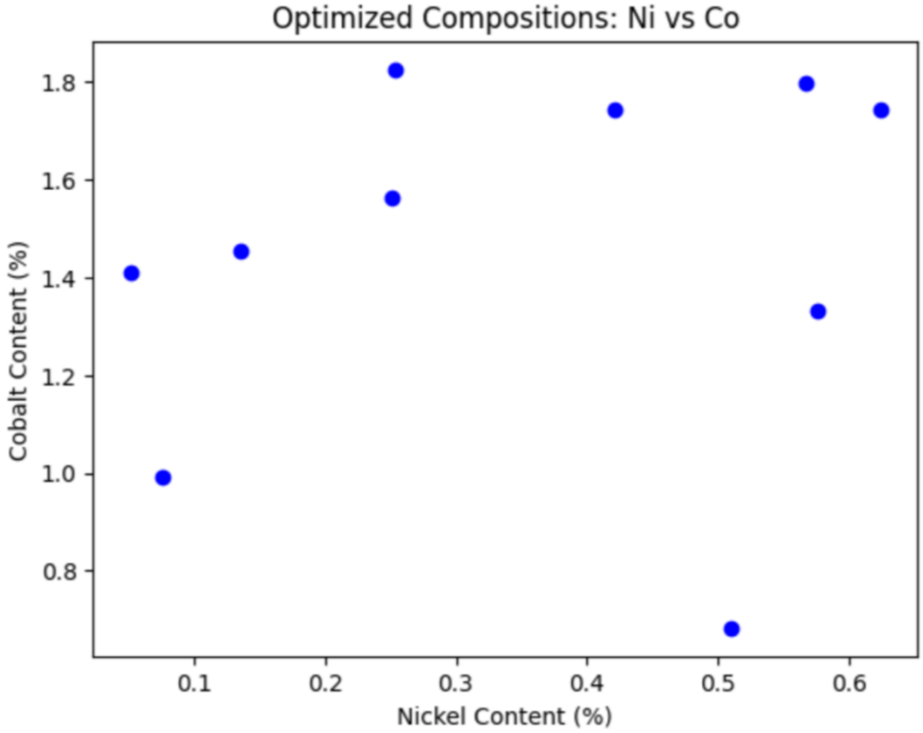
- **Accuracy**: Random Forest outperformed other models with a stable and high accuracy score.
- **Uncertainty**: Cross-validation results indicated low variance, confirming model stability.

- Although Random Forest performed best, it failed to detect any fragile samples and made errors, it showed a high bias, thus we used logistic regression which was able to idenitfy weak and strong samples.

Optimized Compositions

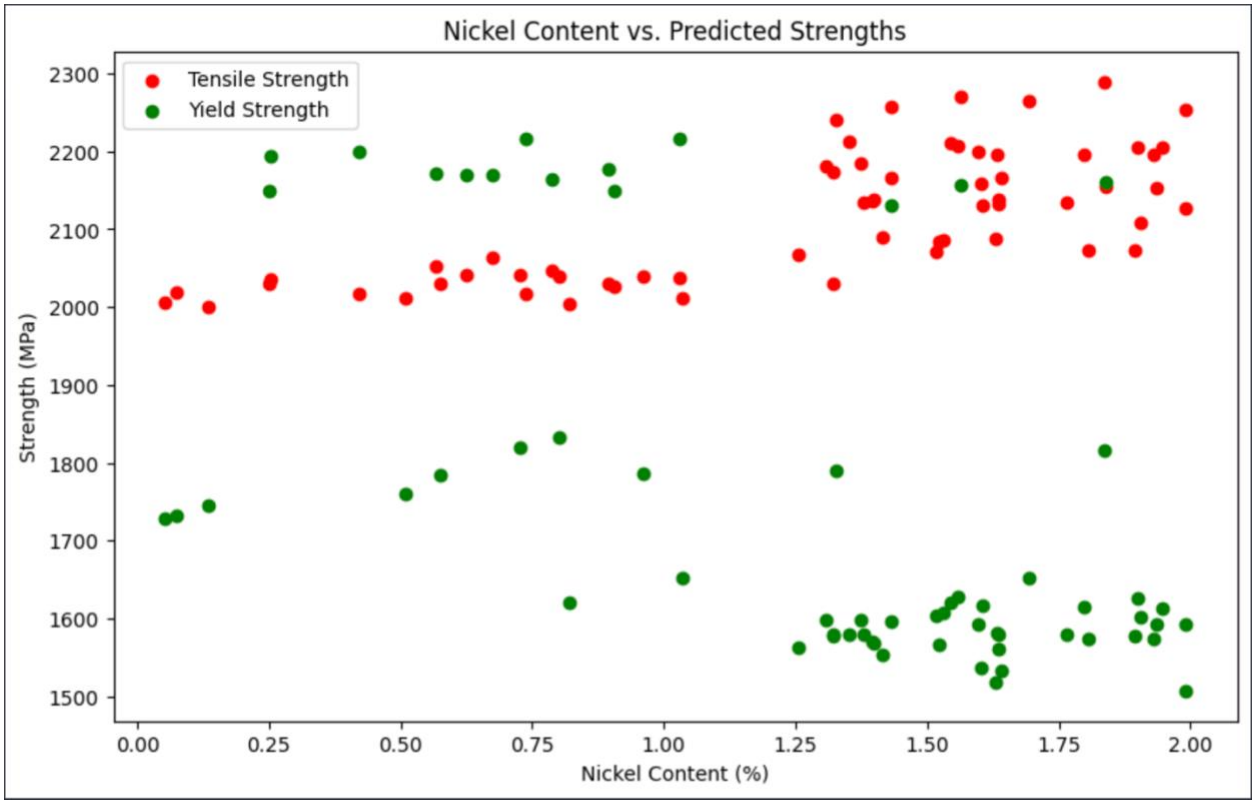
Graphs

- **Optimized Compositions: Ni vs Co**:



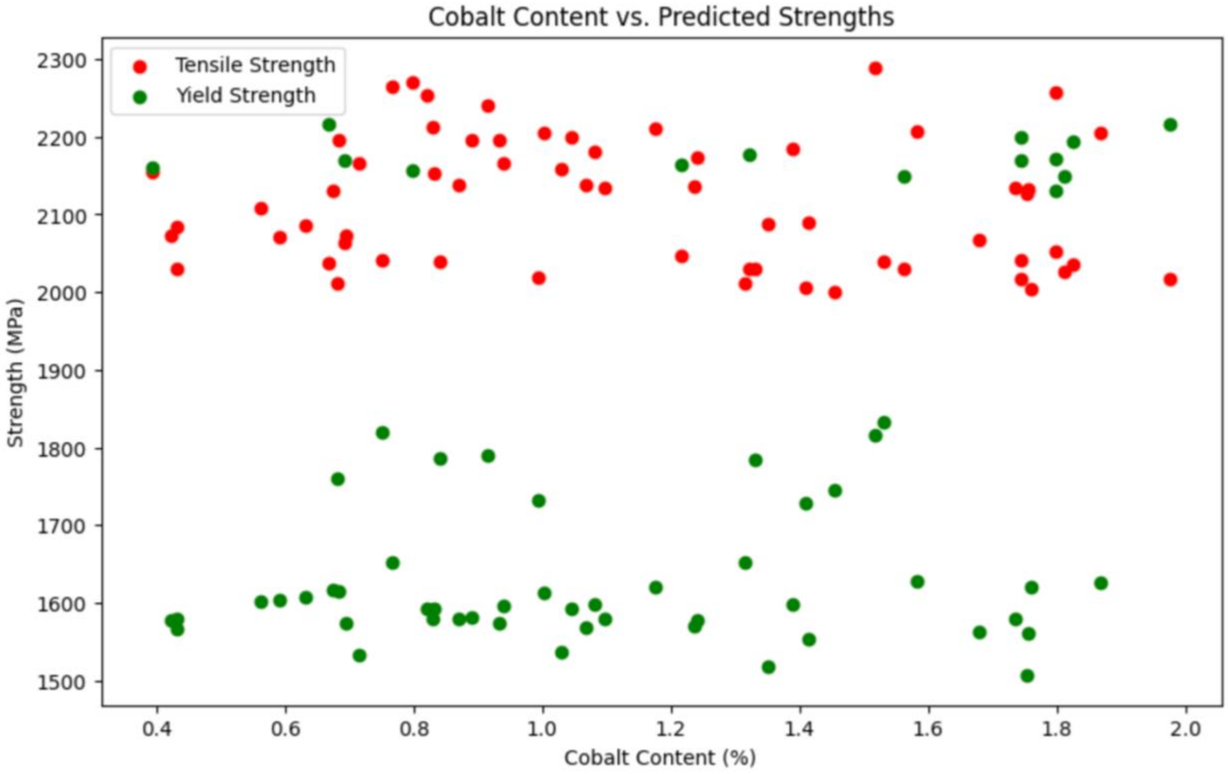
- *Scatter plot of optimized compositions showing Nickel (Ni) vs Cobalt (Co) content. This graph highlights the compositions that have been optimized to minimize the content of Nickel and Cobalt while maintaining high strength properties.*

- **Nickel Content vs. Predicted Strengths**:



- *Scatter plot of Nickel content vs. predicted tensile and yield strengths. This graph shows the relationship between Nickel content and the predicted mechanical properties, indicating that lower Nickel content can still achieve high strength.*

- **Cobalt Content vs. Predicted Strengths**:



- *Scatter plot of Cobalt content vs. predicted tensile and yield strengths. This graph illustrates the impact of Cobalt content on the predicted mechanical properties, demonstrating that it is possible to achieve high strength with minimal Cobalt content.*

Observations

- **Key Features**: The optimized compositions successfully minimize the use of Nickel and Cobalt while maintaining high tensile and yield strengths. The scatter plots show that it is possible to achieve strong materials with lower amounts of these expensive and critical elements.
- **Best Compositions**: The top candidates identified through this optimization process are those that balance the mechanical properties with the minimal use of Nickel and Cobalt, making them cost-effective and sustainable choices for steel production.

Conclusion

- **Regression**: Random Forest Regression provided the best results, combining high accuracy and performance.
- **Classification**: Random Forest Classifier consistently delivered superior accuracy and balanced precision-recall across all classes.
- **Data Cleaning & Preprocessing**: Critical to achieving these results was careful cleaning, normalization, and encoding.
- **Conclusion:** Using the accuracy score within the scikit library the best performing method is chosen and performed.