Random Forest Analysis Report

Date: 2025-04-06

Introduction

This report presents the results of a Random Forest analysis for both classification and regression tasks. Random Forest is an ensemble learning method that constructs multiple decision trees during training and outputs the class that is the mode of the classes (for classification) or the mean prediction (for regression) of the individual trees. The analysis includes data exploration, model training, hyperparameter tuning, and advanced interpretability techniques.

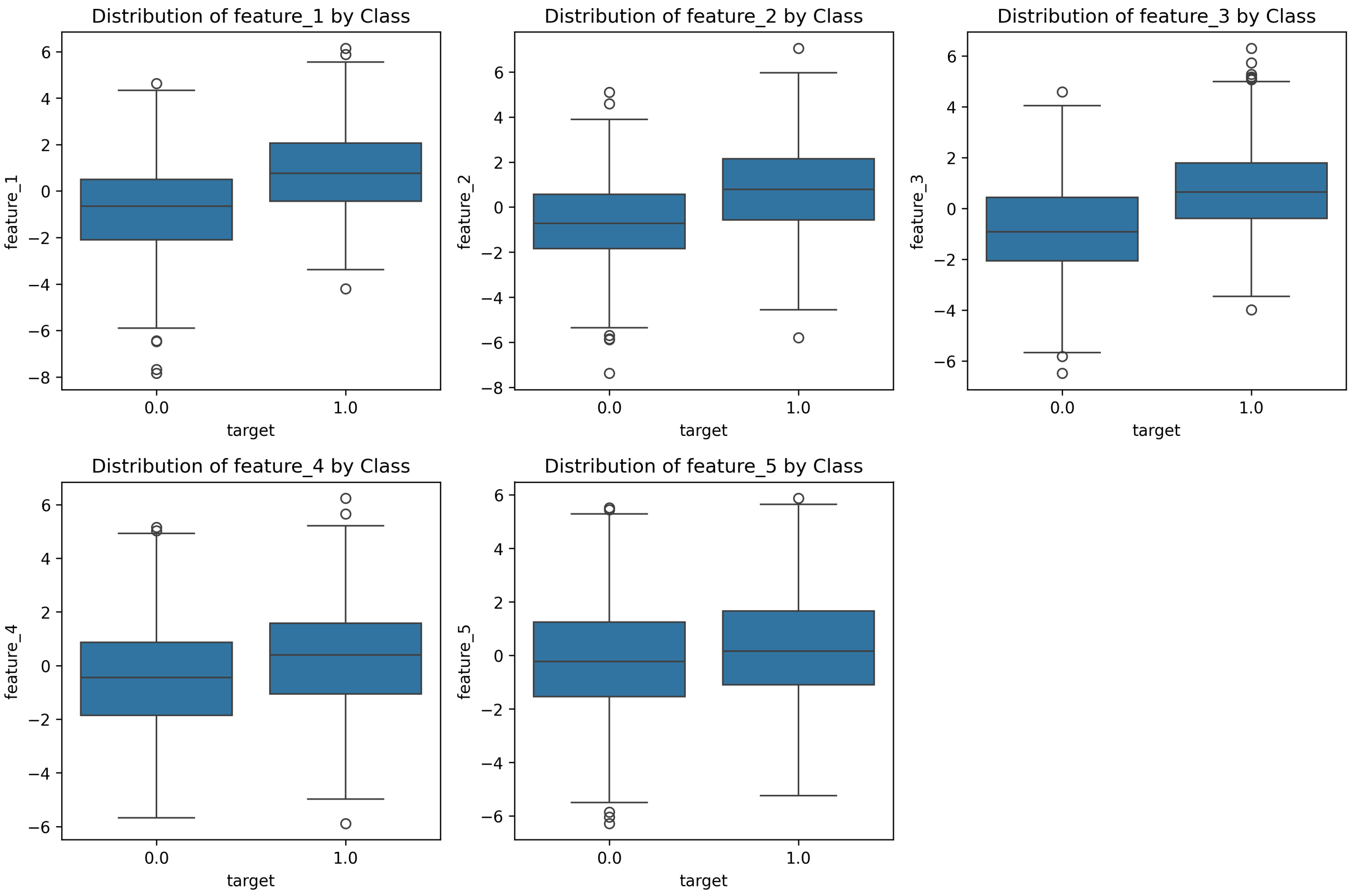
Data Description

The analysis includes two synthetic datasets: one for classification and one for regression. Both datasets contain 1000 samples with 10 features each. For the classification task, the target is a binary variable. For the regression task, the target is a continuous variable. The features include informative, redundant, and repeated features to simulate real-world data complexity.

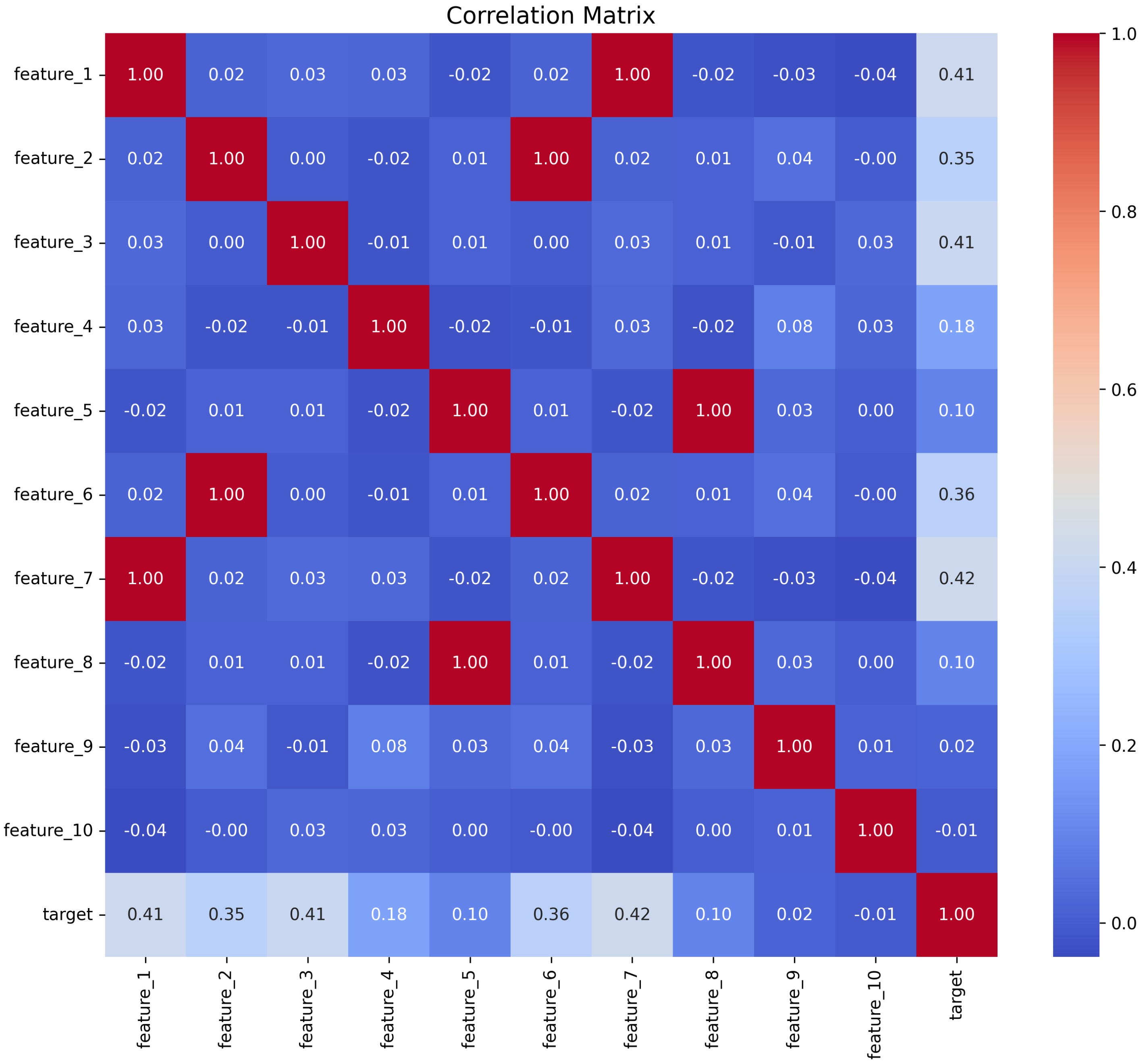
Classification Data Exploration

The classification dataset was explored using various visualization techniques. Feature distributions by class, correlation matrices, and other exploratory data analysis methods were used to understand the data structure and relationships between features and the target variable.

Random Forest Analysis Report



Random Forest Analysis Report

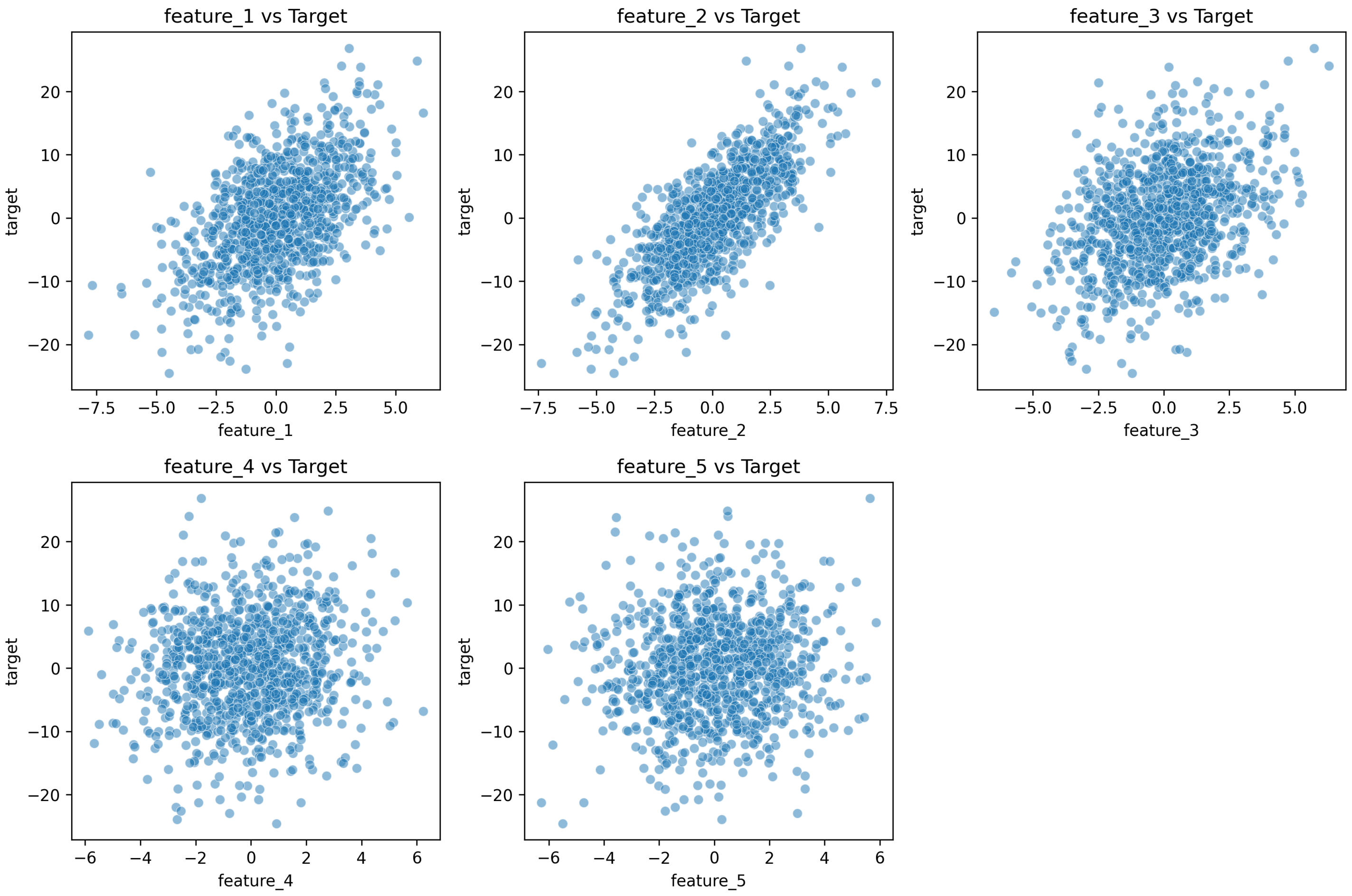


	feature_1	feature_2	feature_3	feature_4	feature_5	feature_6	feature_7	feature_8	feature_9	feature_10	target
feature_1	1.00	0.02	0.03	0.03	-0.02	0.02	1.00	-0.02	-0.03	-0.04	0.41
feature_2	0.02	1.00	0.00	-0.02	0.01	1.00	0.02	0.01	0.04	-0.00	0.35
feature_3	0.03	0.00	1.00	-0.01	0.01	0.00	0.03	0.01	-0.01	0.03	0.41
feature_4	0.03	-0.02	-0.01	1.00	-0.02	-0.01	0.03	-0.02	0.08	0.03	0.18
feature_5	-0.02	0.01	0.01	-0.02	1.00	0.01	-0.02	1.00	0.03	0.00	0.10
feature_6	0.02	1.00	0.00	-0.01	0.01	1.00	0.02	0.01	0.04	-0.00	0.36
feature_7	1.00	0.02	0.03	0.03	-0.02	0.02	1.00	-0.02	-0.03	-0.04	0.42
feature_8	-0.02	0.01	0.01	-0.02	1.00	0.01	-0.02	1.00	0.03	0.00	0.10
feature_9	-0.03	0.04	-0.01	0.08	0.03	0.04	-0.03	0.03	1.00	0.01	0.02
feature_10	-0.04	-0.00	0.03	0.03	0.00	-0.00	-0.04	0.00	0.01	1.00	-0.01
target	0.41	0.35	0.41	0.18	0.10	0.36	0.42	0.10	0.02	-0.01	1.00

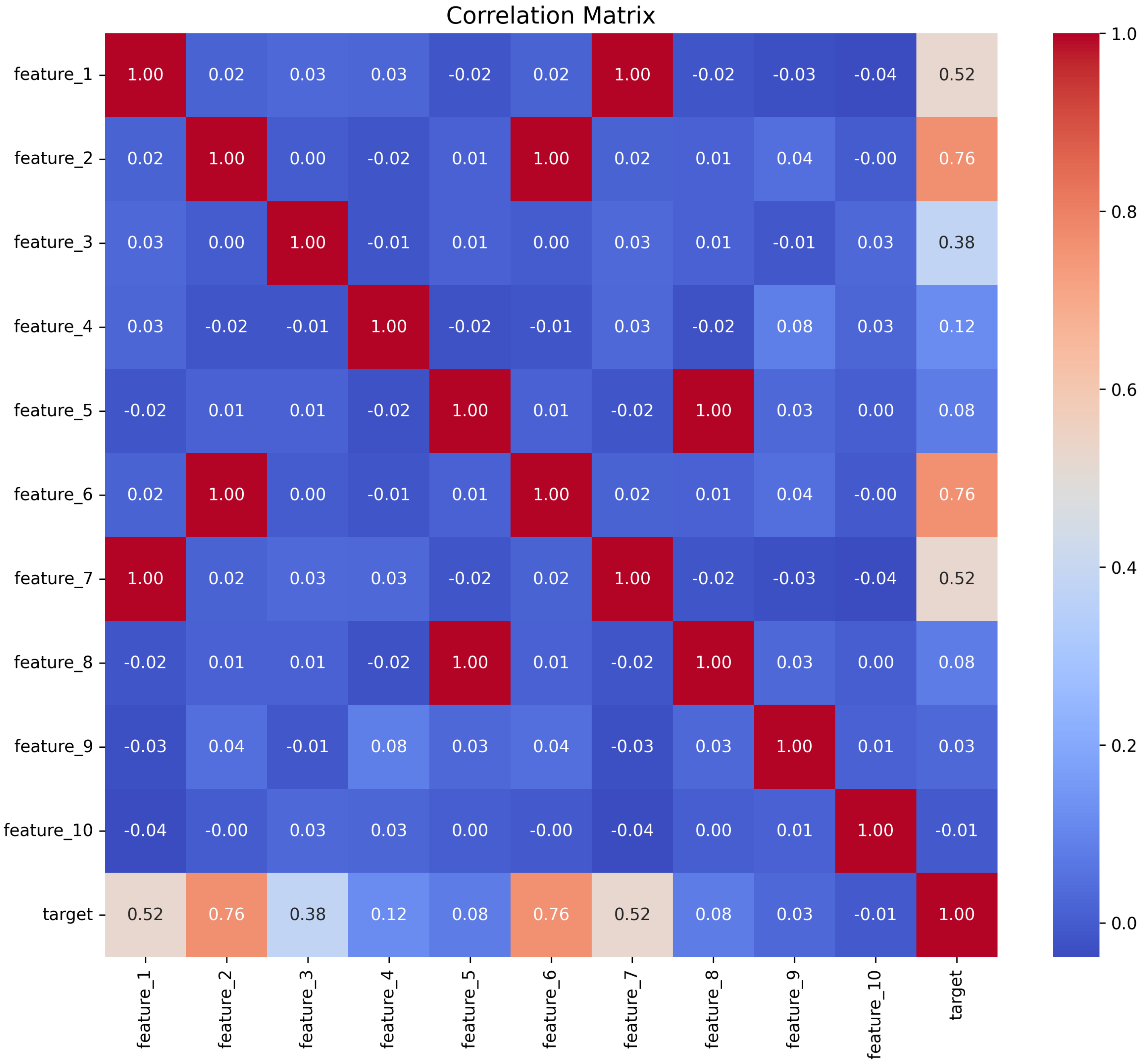
Regression Data Exploration

The regression dataset was explored using scatter plots, correlation matrices, and other visualization techniques to understand the relationships between features and the target variable.

Random Forest Analysis Report



Random Forest Analysis Report



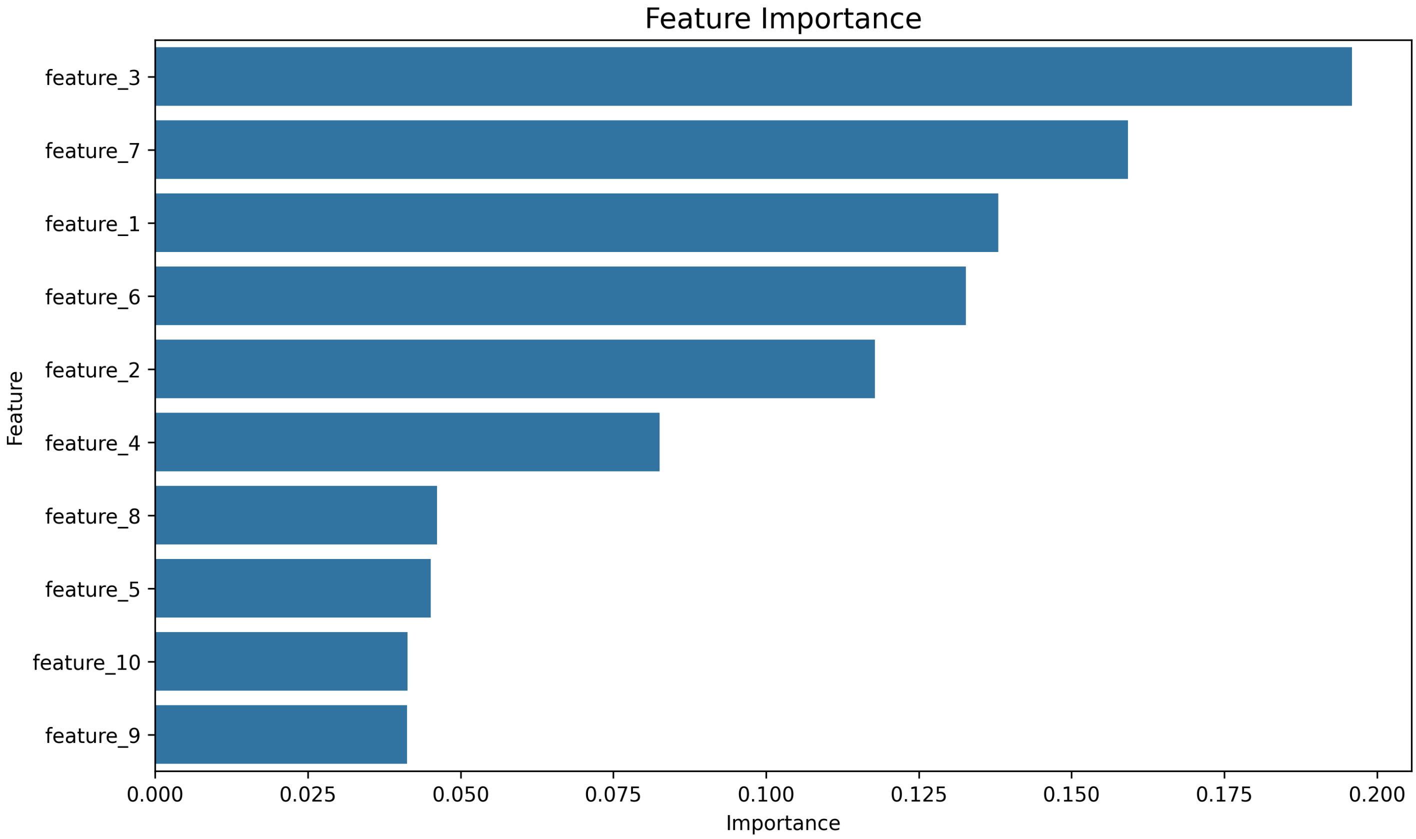
Correlation Matrix

	feature_1	feature_2	feature_3	feature_4	feature_5	feature_6	feature_7	feature_8	feature_9	feature_10	target
feature_1	1.00	0.02	0.03	0.03	-0.02	0.02	1.00	-0.02	-0.03	-0.04	0.52
feature_2	0.02	1.00	0.00	-0.02	0.01	1.00	0.02	0.01	0.04	-0.00	0.76
feature_3	0.03	0.00	1.00	-0.01	0.01	0.00	0.03	0.01	-0.01	0.03	0.38
feature_4	0.03	-0.02	-0.01	1.00	-0.02	-0.01	0.03	-0.02	0.08	0.03	0.12
feature_5	-0.02	0.01	0.01	-0.02	1.00	0.01	-0.02	1.00	0.03	0.00	0.08
feature_6	0.02	1.00	0.00	-0.01	0.01	1.00	0.02	0.01	0.04	-0.00	0.76
feature_7	1.00	0.02	0.03	0.03	-0.02	0.02	1.00	-0.02	-0.03	-0.04	0.52
feature_8	-0.02	0.01	0.01	-0.02	1.00	0.01	-0.02	1.00	0.03	0.00	0.08
feature_9	-0.03	0.04	-0.01	0.08	0.03	0.04	-0.03	0.03	1.00	0.01	0.03
feature_10	-0.04	-0.00	0.03	0.03	0.00	-0.00	-0.04	0.00	0.01	1.00	-0.01
target	0.52	0.76	0.38	0.12	0.08	0.76	0.52	0.08	0.03	-0.01	1.00

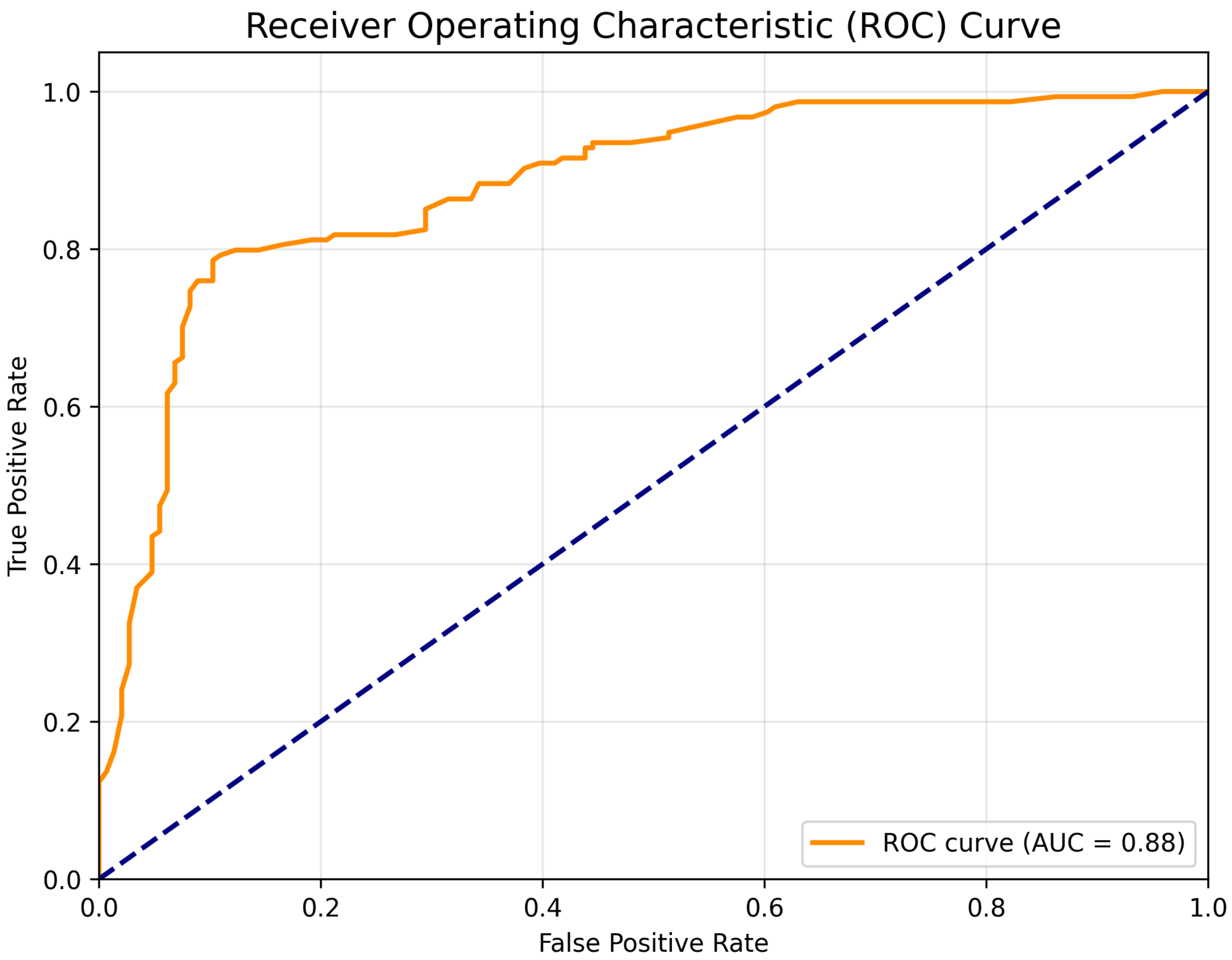
Random Forest Analysis Report

Classification Results

The Random Forest classifier was trained on the classification dataset and achieved an accuracy of 0.7900 on the test set. The model was evaluated using precision, recall, F1 score, and ROC curve analysis. Cross-validation was performed to assess the model's generalization performance.



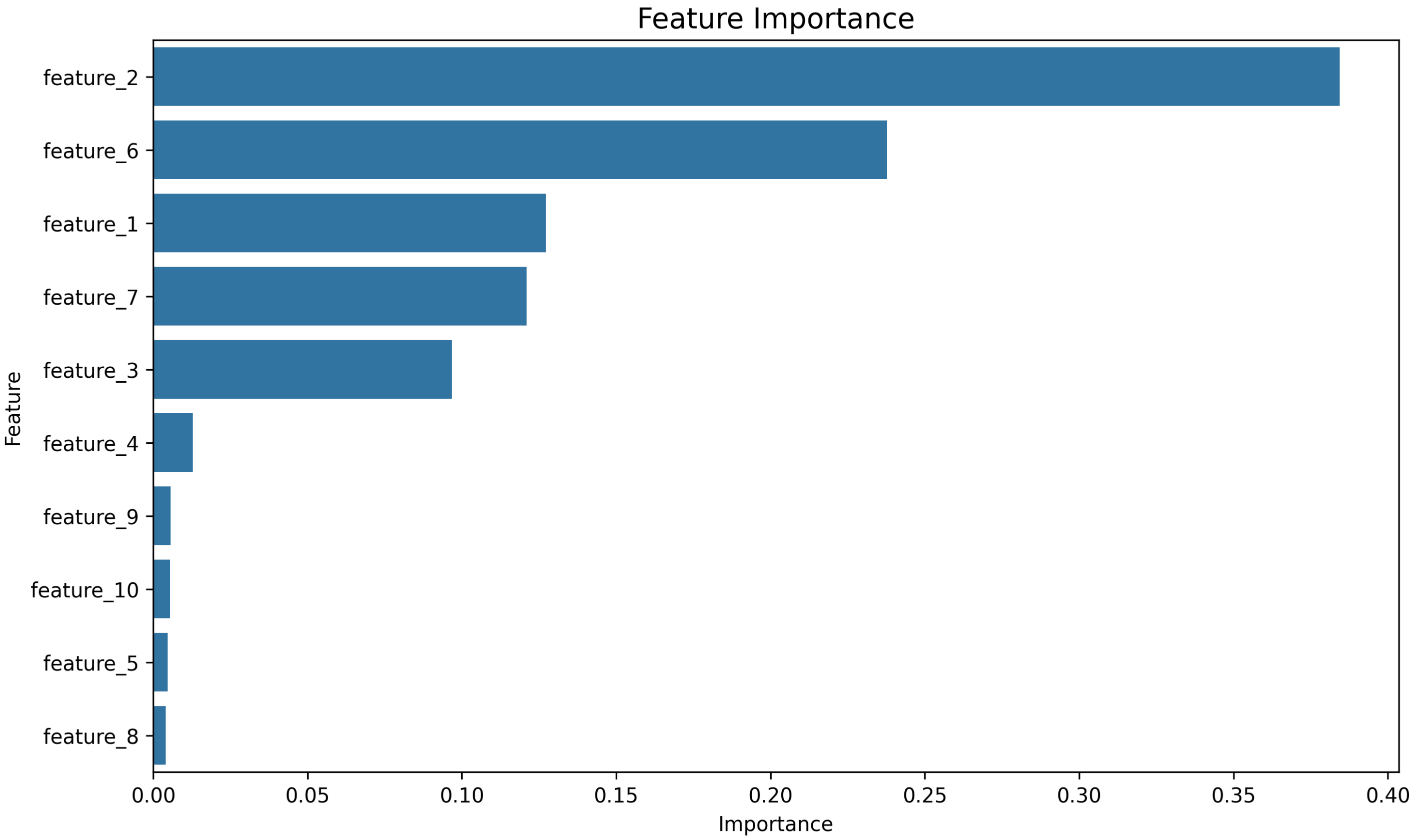
Random Forest Analysis Report



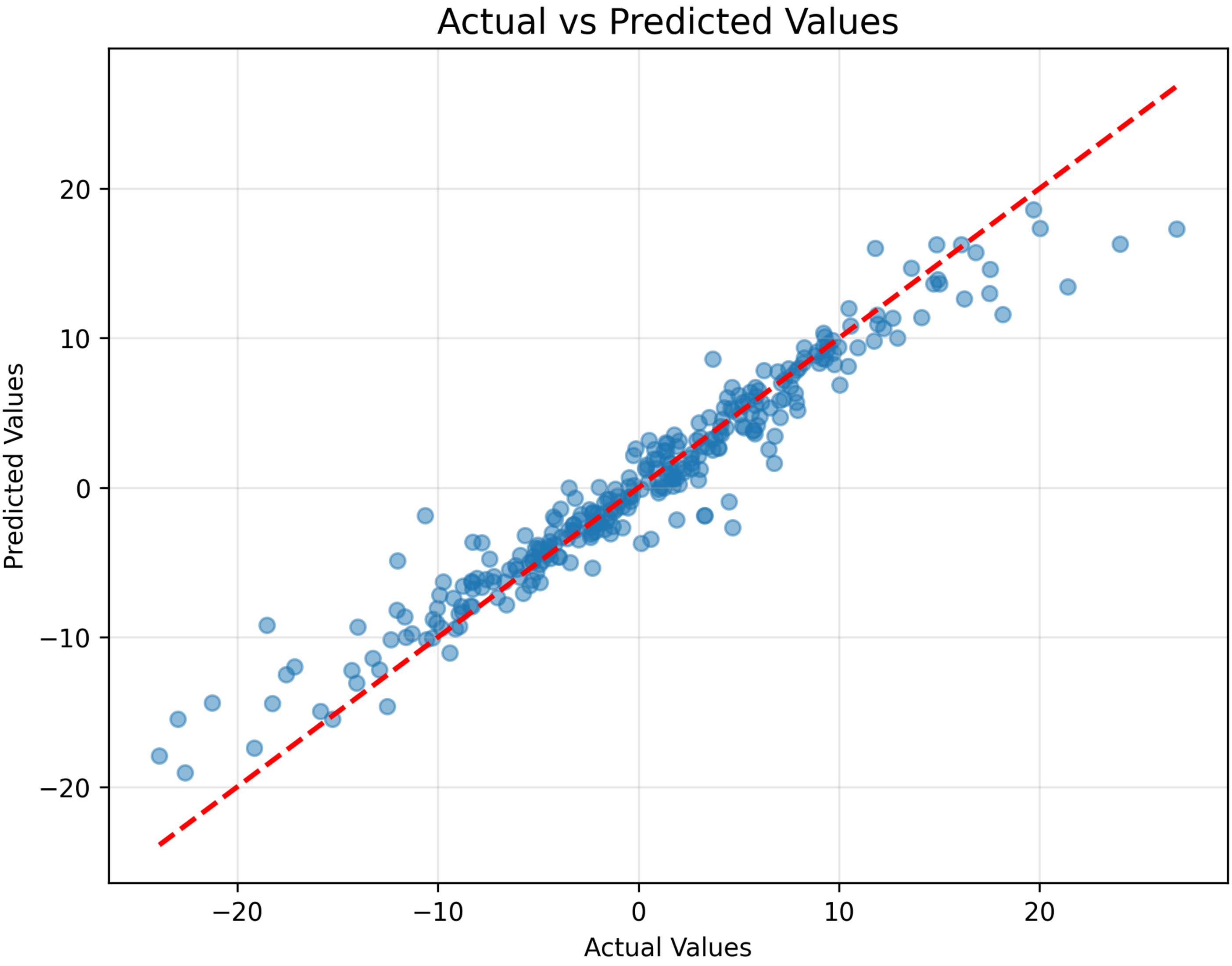
Regression Results

The Random Forest regressor was trained on the regression dataset and achieved an R² score of 0.9262 on the test set. The model was evaluated using mean squared error, root mean squared error, mean absolute error, and R² score. Cross-validation was performed to assess the model's generalization performance.

Random Forest Analysis Report



Random Forest Analysis Report



Random Forest Analysis Report

Hyperparameter Tuning

Grid search with cross-validation was performed to find the optimal hyperparameters for both the classification and regression models. The hyperparameters tuned included the number of trees, maximum depth, minimum samples split, and minimum samples leaf.

For classification, the best parameters were: {'max_depth': None, 'min_samples_leaf': 2, 'min_samples_split': 10, 'n_estimators': 50}

For regression, the best parameters were: {'max_depth': None, 'min_samples_leaf': 1, 'min_samples_split': 2, 'n_estimators': 100}

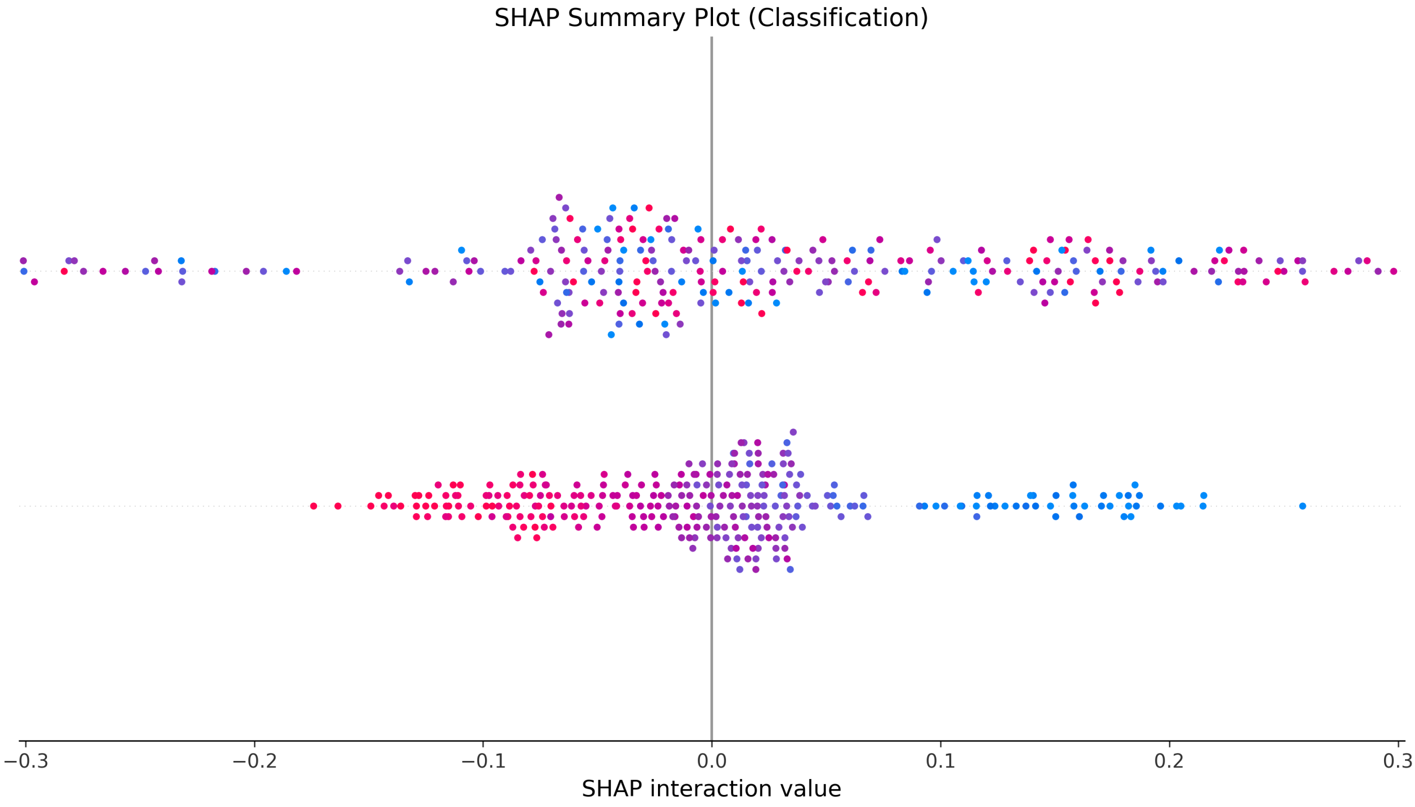
Advanced Interpretability

Several advanced interpretability techniques were applied to understand the models better:

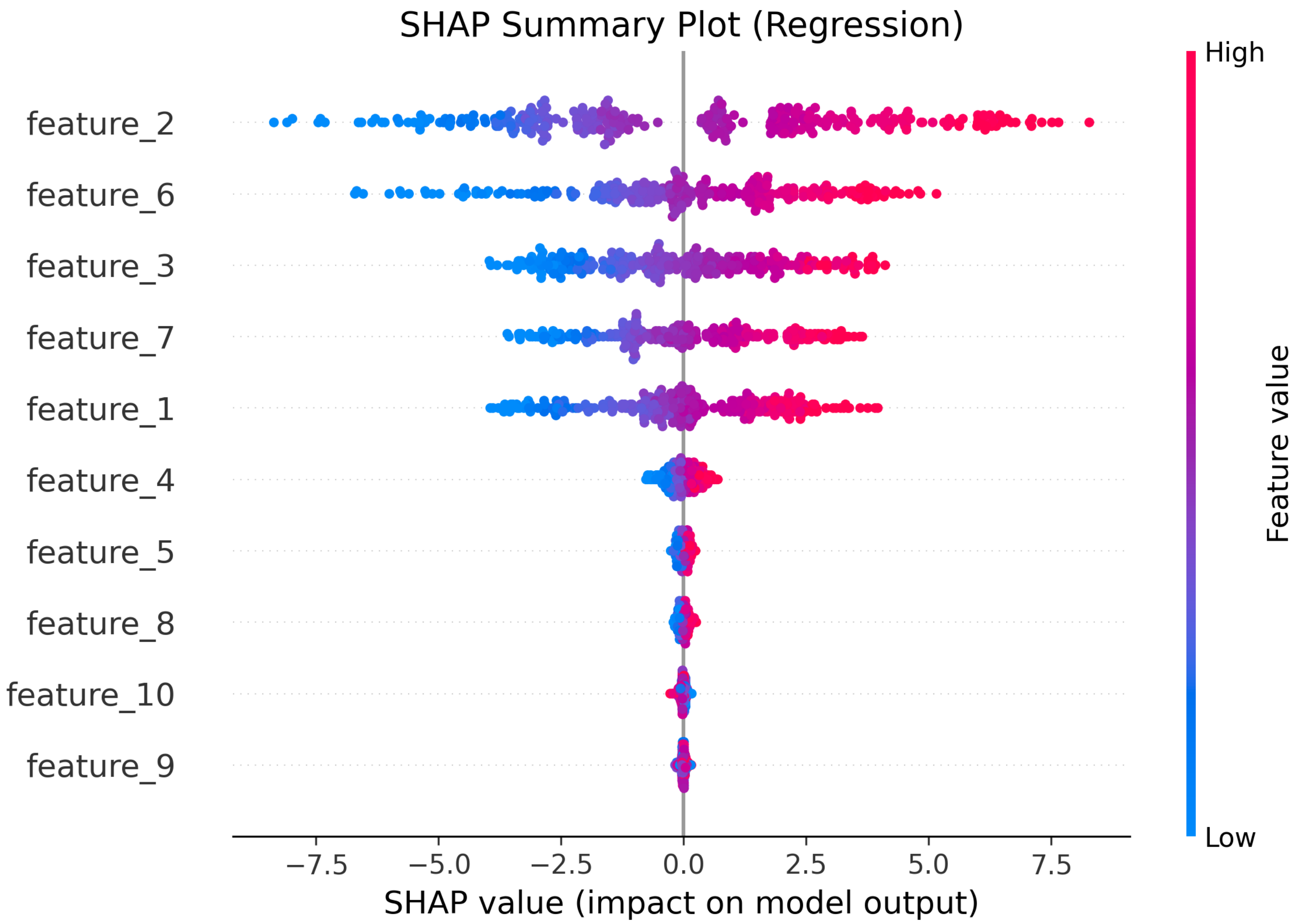
1. SHAP (SHapley Additive exPlanations) values were calculated to explain the output of the models.

2. Permutation importance was calculated to assess the importance of each feature.

3. Partial dependence plots were generated to visualize the relationship between the most important features and the target variable.



Random Forest Analysis Report



Random Forest Analysis Report

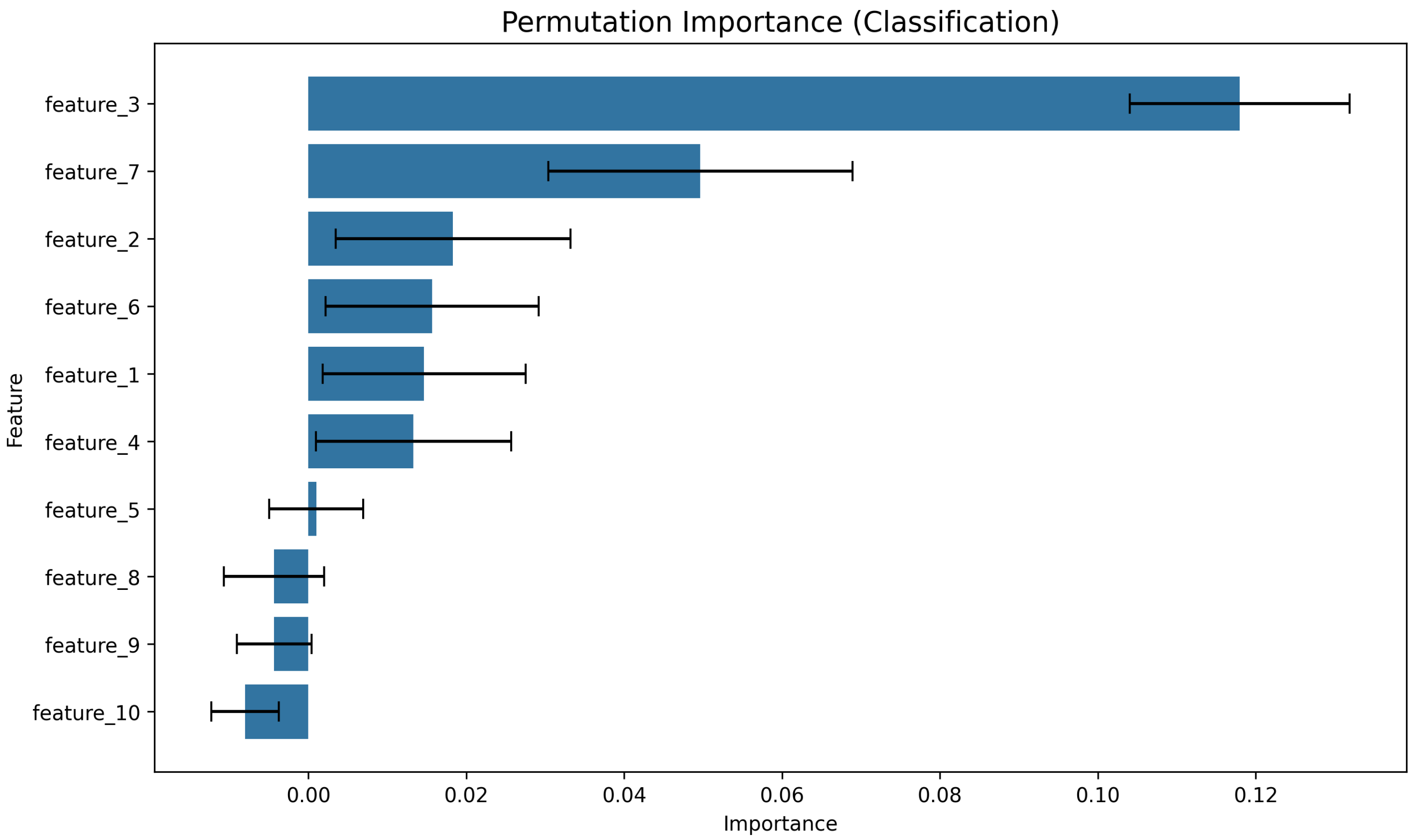
Feature Importance Comparison

Different feature importance measures were compared to ensure robustness of the feature importance rankings:

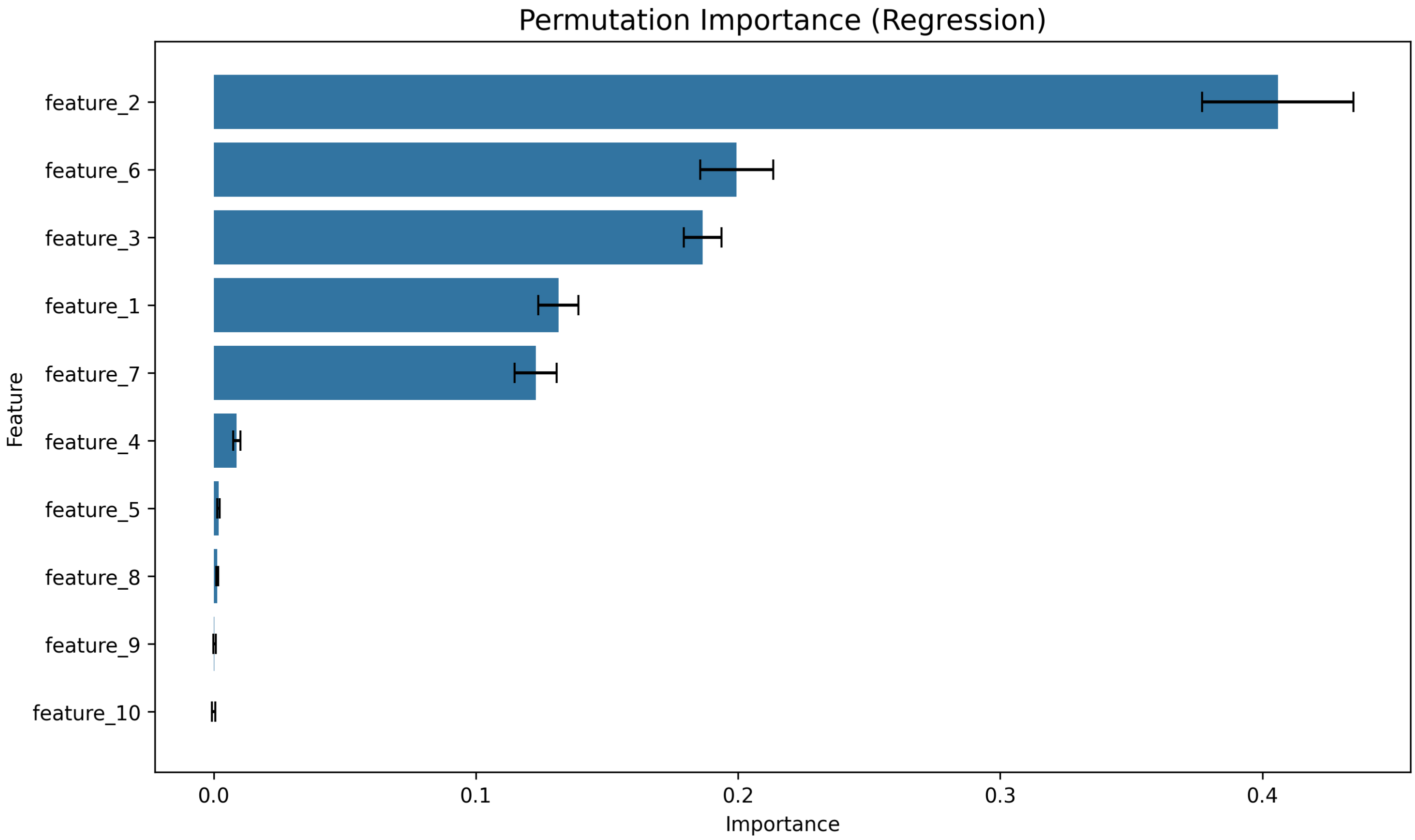
1. Built-in feature importance from Random Forest

2. Permutation importance

3. SHAP values



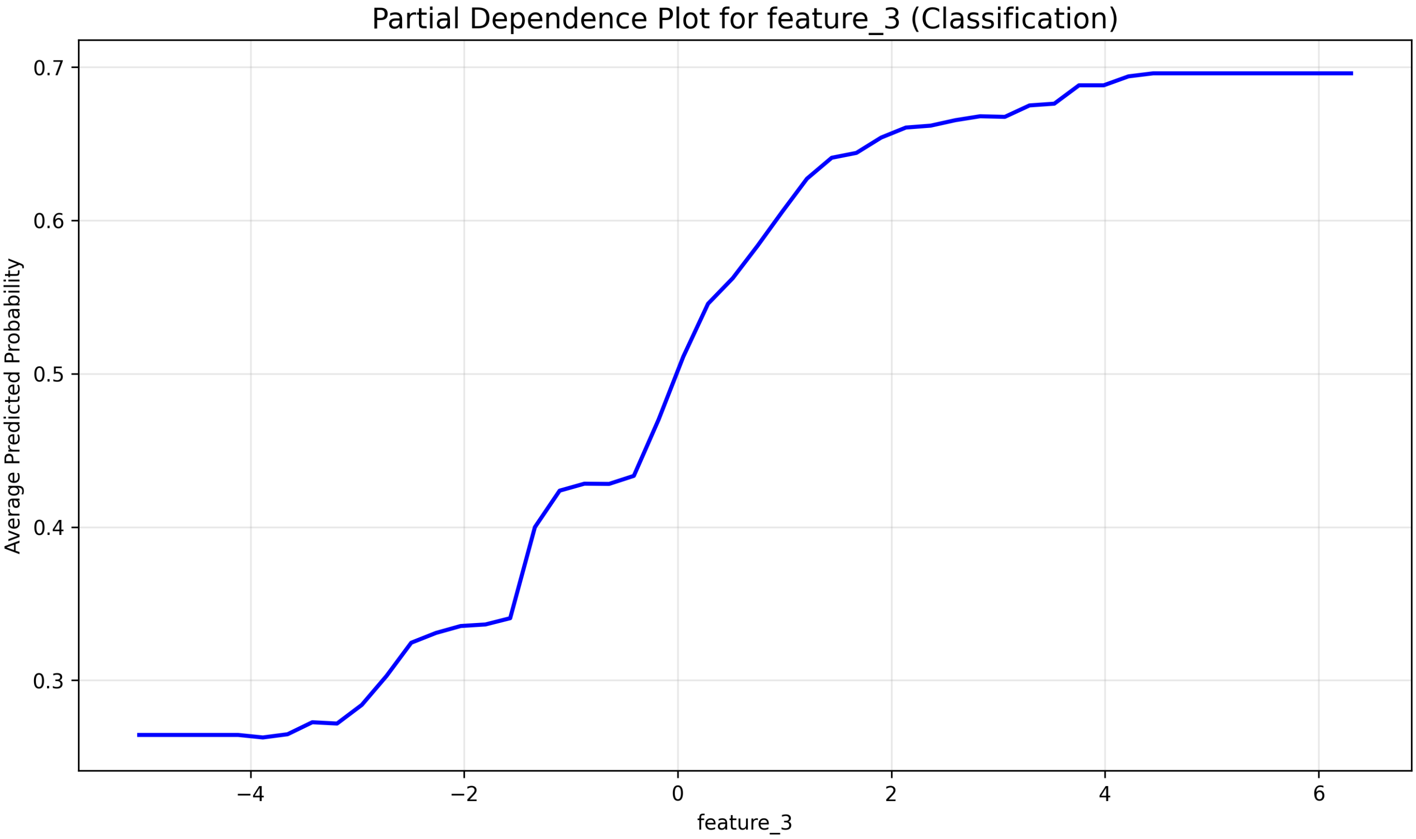
Random Forest Analysis Report

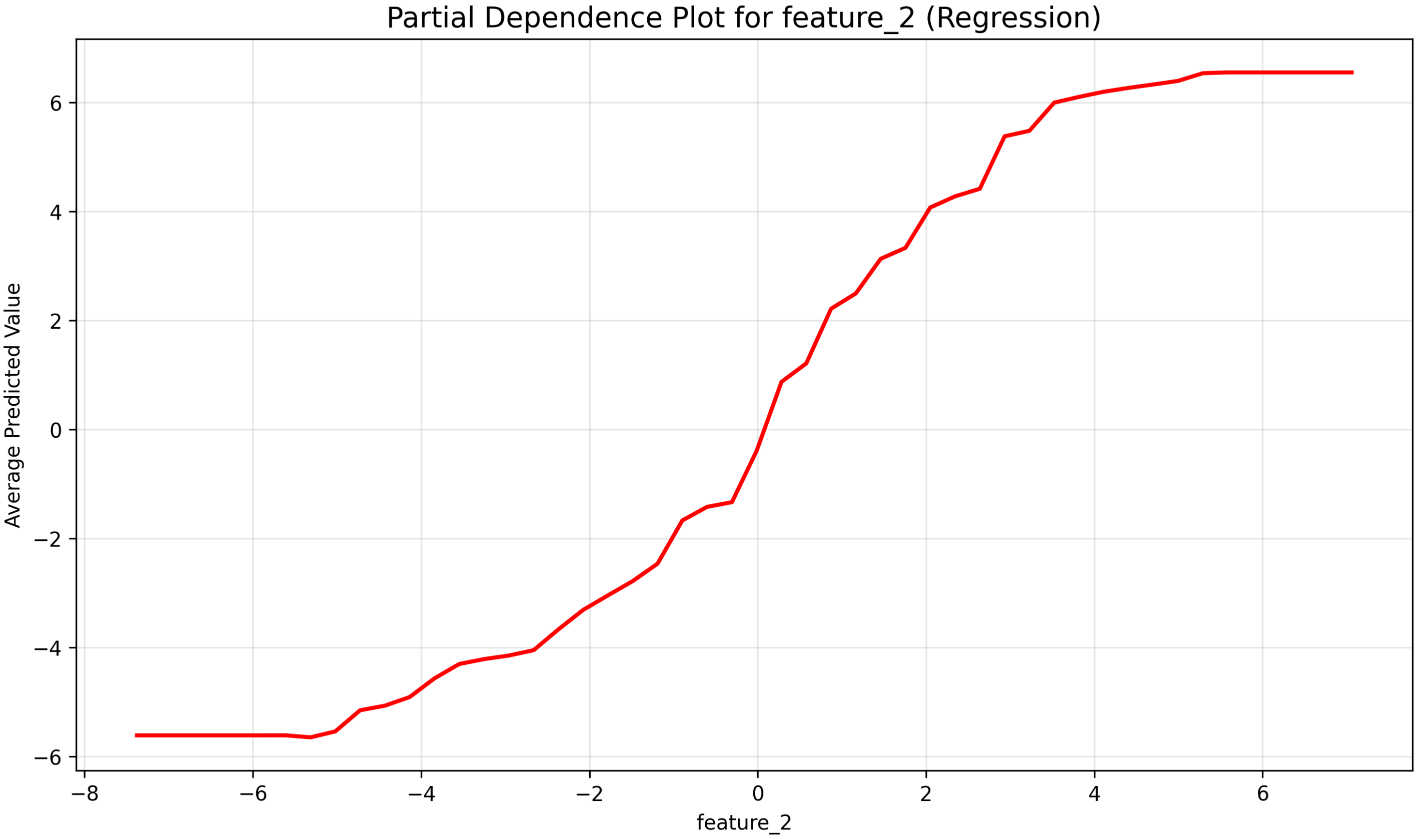


Partial Dependence Plots

Partial dependence plots were generated to visualize the relationship between the most important features and the target variable, while accounting for the average effect of other features.

Random Forest Analysis Report





Conclusion

Random Forest Analysis Report

Random Forest performed well on both classification and regression tasks. For classification, the best model achieved an accuracy of 0.7900. For regression, the best model achieved an R^2 score of 0.9262.

The most important features for classification were: feature_3, feature_7, feature_1.
The most important features for regression were: feature_2, feature_6, feature_1.

SHAP values and permutation importance provided consistent feature importance rankings, confirming the robustness of the feature importance analysis. Hyperparameter tuning improved model performance for both tasks.

Random Forest is a powerful and versatile algorithm that can handle both classification and regression tasks effectively. Its ability to capture non-linear relationships, handle high-dimensional data, and provide feature importance rankings makes it a valuable tool for predictive modeling.