Disease Diagnosis Project Report

1. Introduction

This project aims to predict the diagnosis of breast cancer using various machine learning algorithms and compare their performance. The dataset used for this project is from the UCI Machine Learning Repository.

2. Data Retrieval

The dataset was retrieved from the UCI repository and saved locally. The data includes features relevant to breast cancer diagnosis.

3. Data Exploration and Preprocessing

Exploratory Data Analysis (EDA) was conducted to understand the data distribution and relationships between features. Preprocessing steps included handling missing values, feature scaling, and feature engineering.

4. Algorithm Selection

Four machine learning algorithms were selected:

- Decision Tree
- Logistic Regression
- Random Forest
- Support Vector Machine (SVM)

The selection was based on their suitability for classification tasks and their different characteristics.

Disease Diagnosis Project Report

5. Model Training and Evaluation

The models were trained on the preprocessed data, and their performance was evaluated using metrics like accuracy, precision, recall, and F1 score.

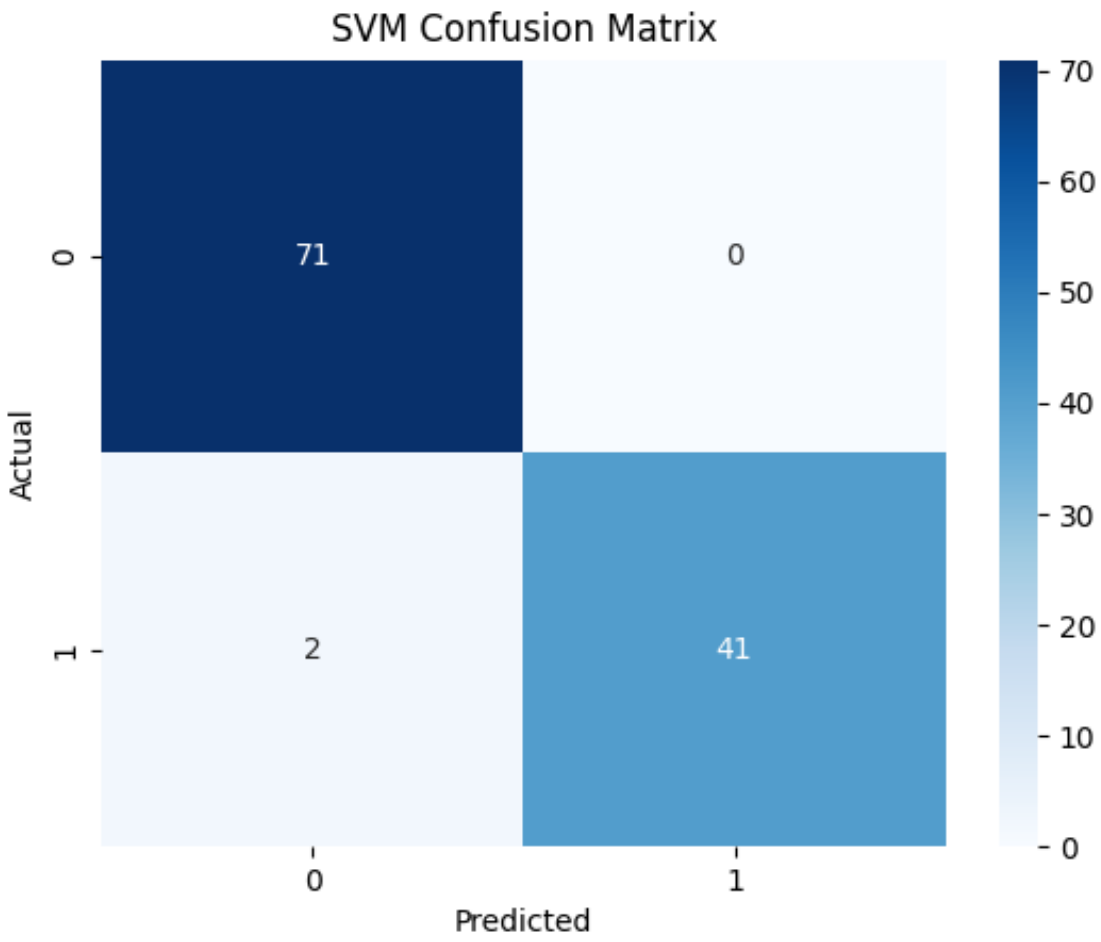
6. Hyperparameter Tuning

Hyperparameter tuning was performed to optimize the performance of the models. The best parameters were identified and the models were re-evaluated.

7. Visualization

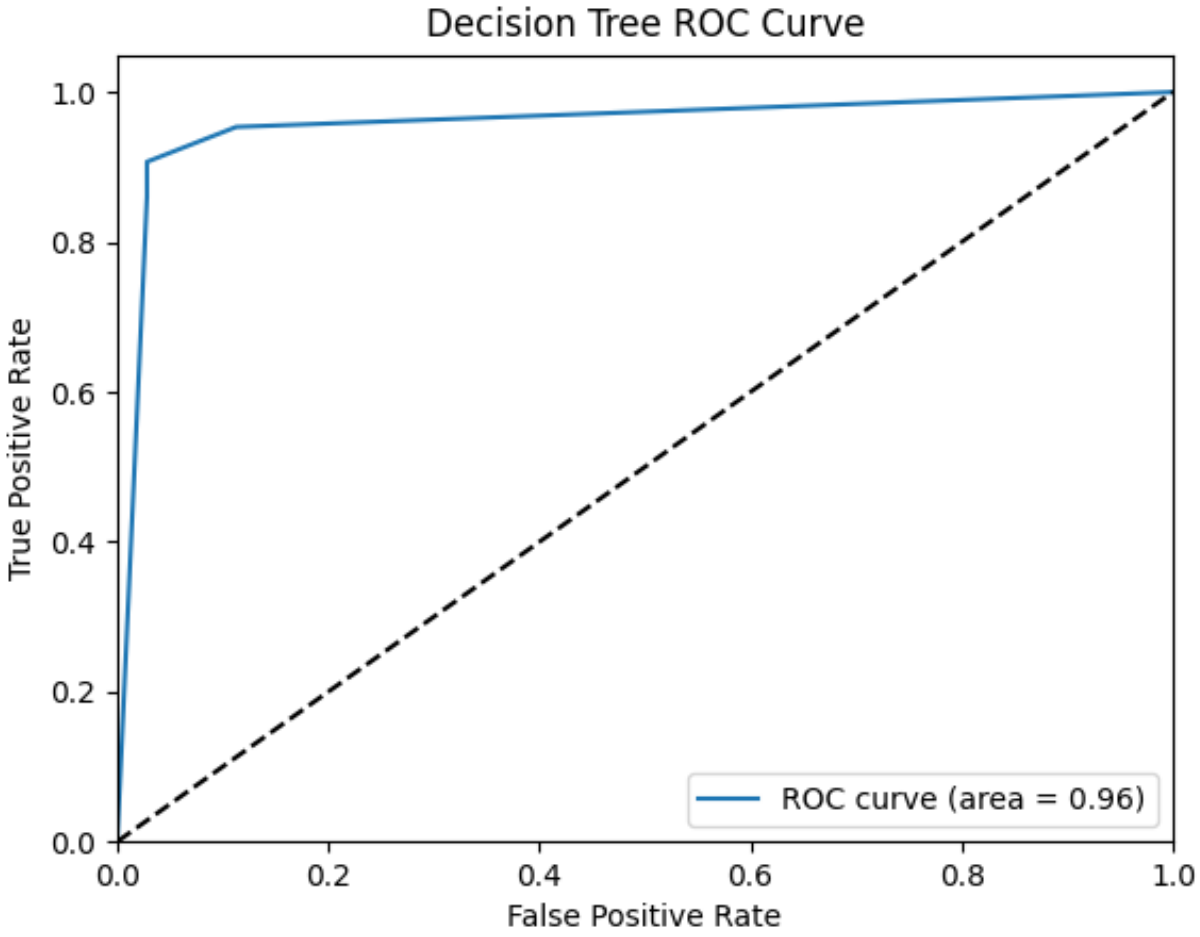
Confusion Matrix

Disease Diagnosis Project Report



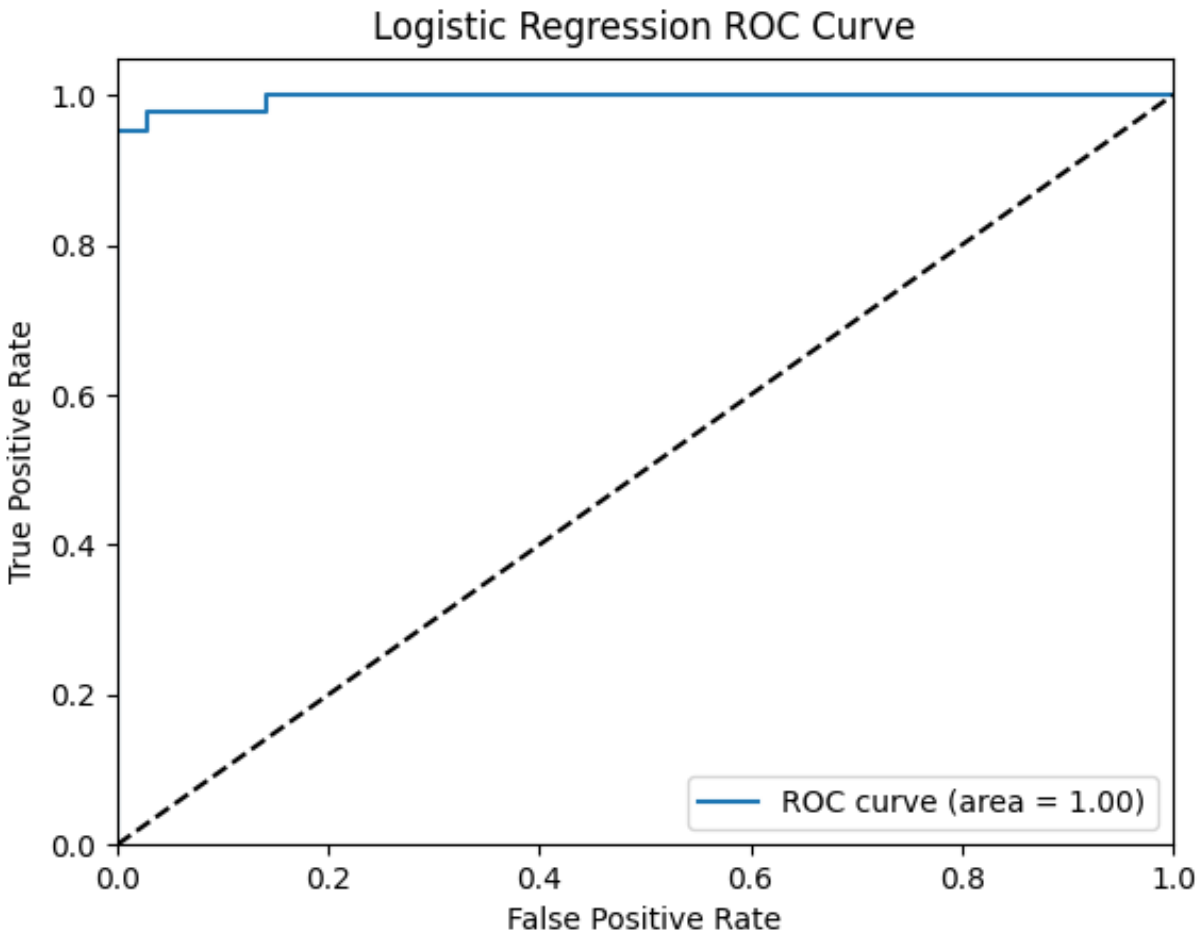
ROC Curve

Disease Diagnosis Project Report



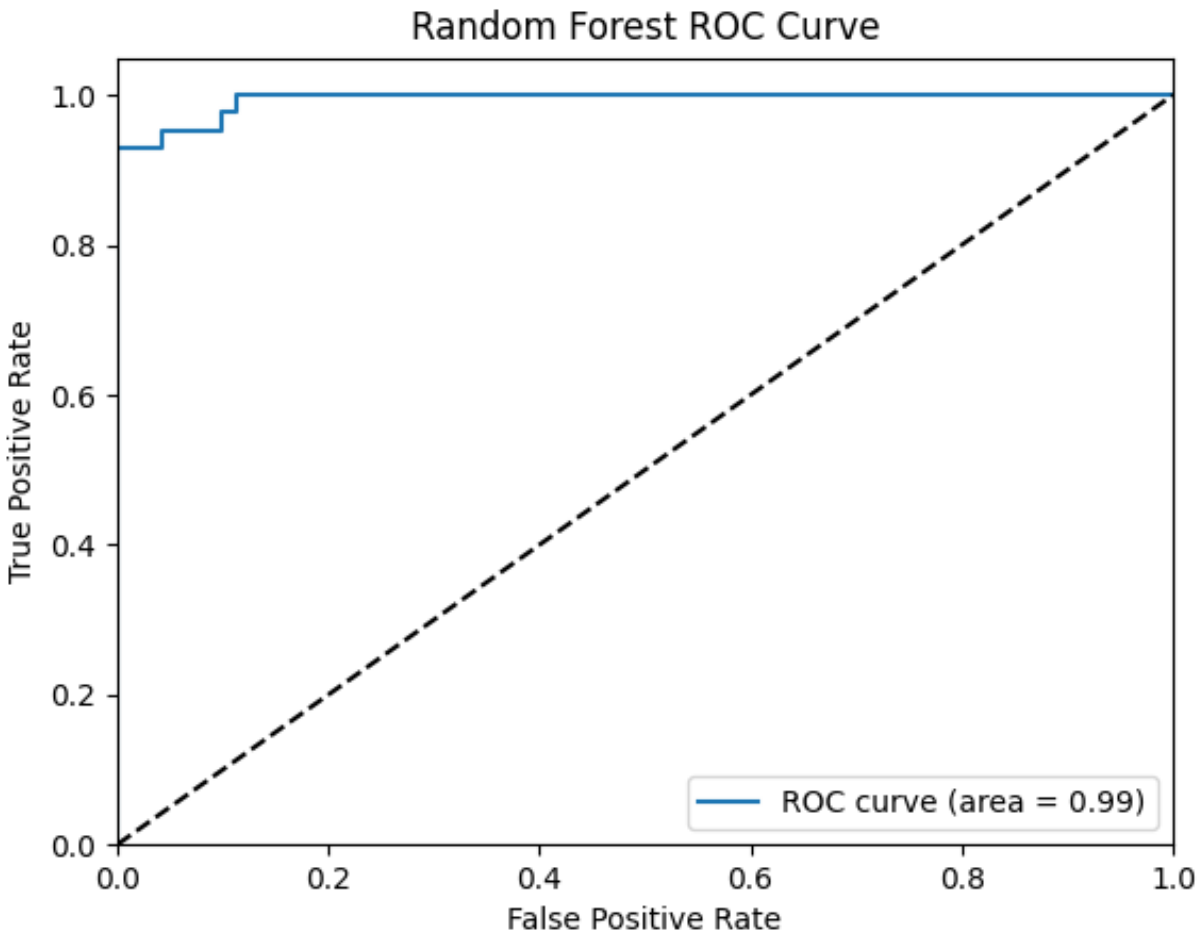
ROC Curve

Disease Diagnosis Project Report



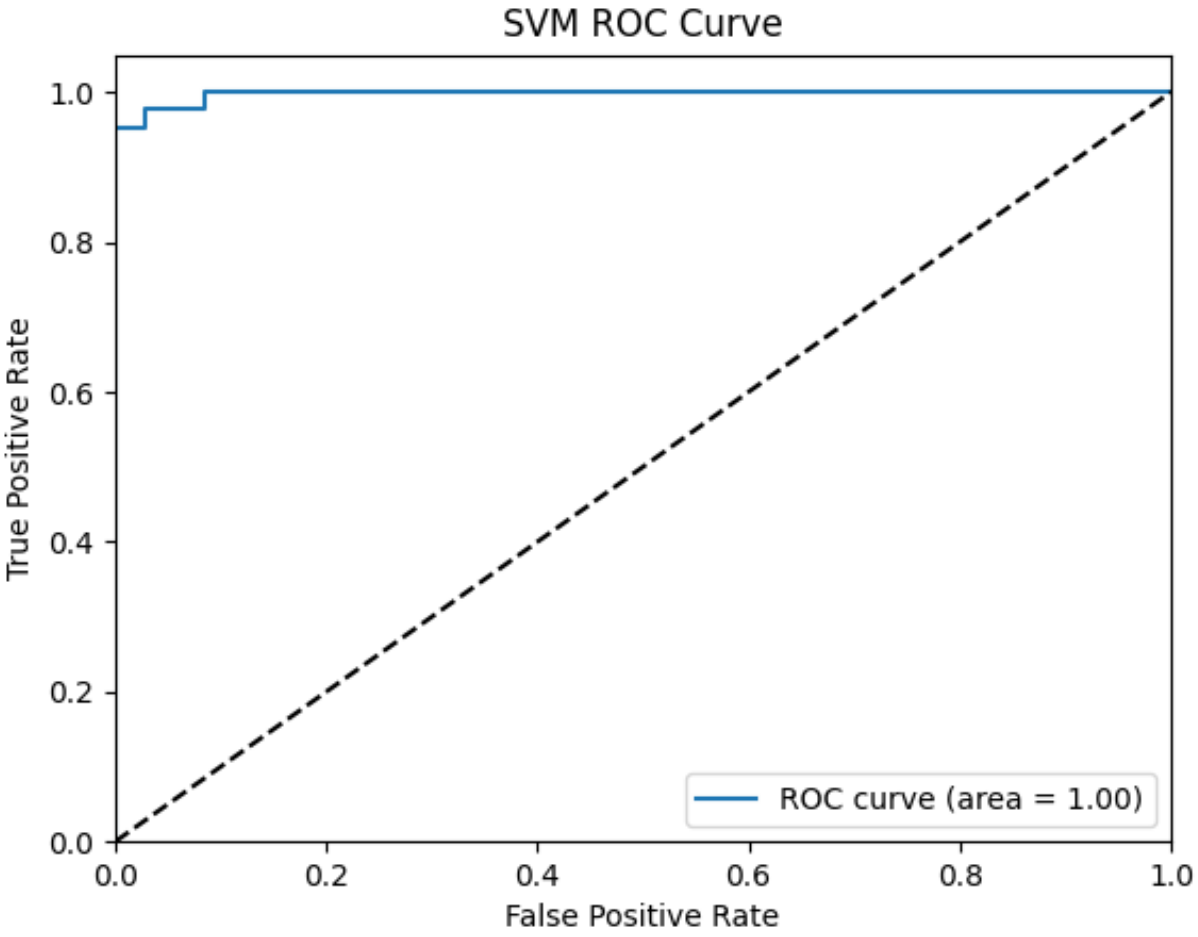
ROC Curve

Disease Diagnosis Project Report

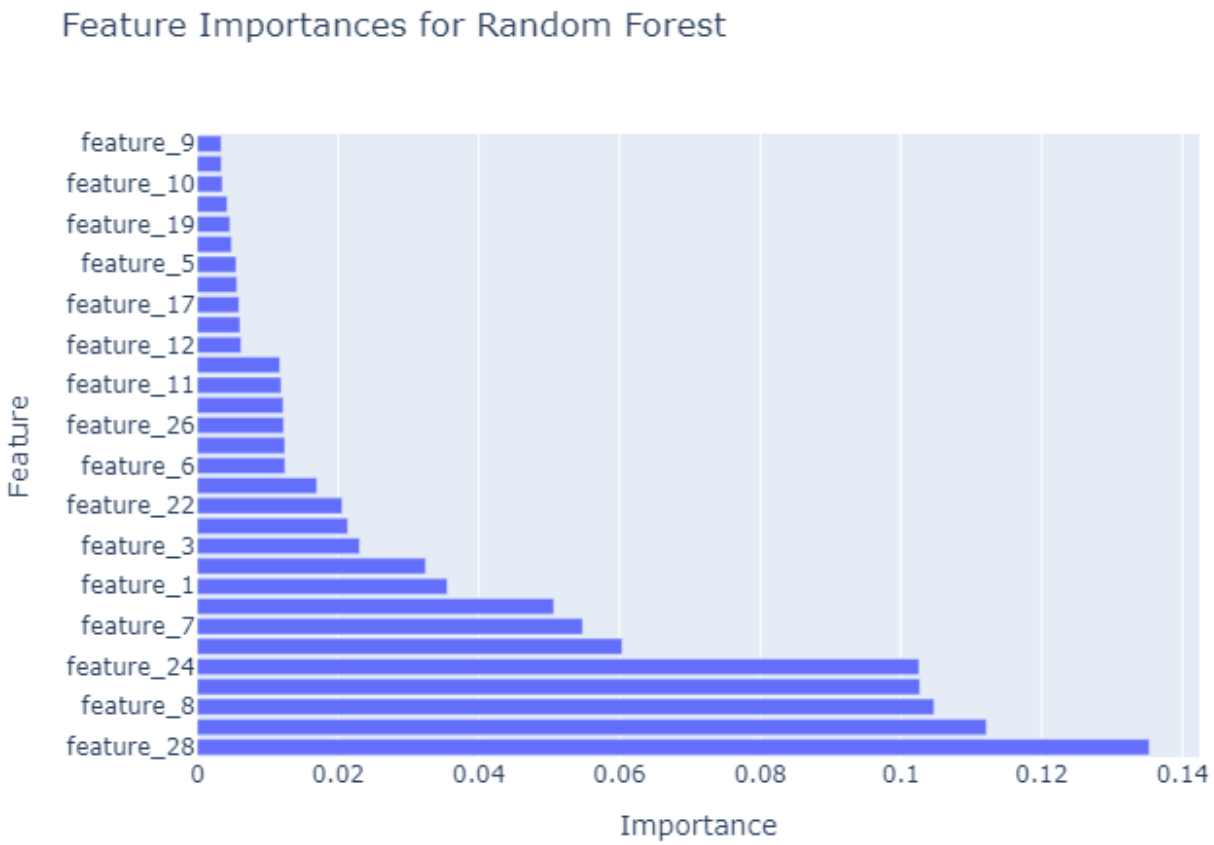


ROC Curve

Disease Diagnosis Project Report



images/Feature Importance for Random Forest



8. Conclusion

The project demonstrated the application of various machine learning techniques to a real-world classification task. Logistic Regression and SVM showed the best performance for this dataset.

9. Future Work

Future work could include exploring more advanced algorithms, incorporating additional data sources, and deploying the model for real-time diagnosis.