Breast Cancer Prediction Project Report

1. Introduction

Breast cancer is one of the most common cancers among women worldwide. Early detection and diagnosis are crucial for effective treatment and improved survival rates. This project aims to predict breast cancer using machine learning models, leveraging a dataset containing various features extracted from breast cancer biopsies. The objective is to compare the performance of different machine learning algorithms and identify the most accurate model for breast cancer prediction.

2. Dataset Description

The dataset used for this project is the Breast Cancer Wisconsin (Diagnostic) Dataset, which includes 569 samples and 6 features:

- **mean_radius**
- **mean_texture**
- **mean_perimeter**
- **mean_area**
- **mean_smoothness**
- **diagnosis (target variable: 0 = benign, 1 = malignant)**

Data Inspection

- No Missing Values: All columns have 569 non-null values.
- No Duplicates: There are no duplicated rows in the dataset.
- Diagnosis Distribution: 357 malignant (1) and 212 benign (0) cases.

3. Data Preprocessing

To ensure that the machine learning models perform optimally, the following preprocessing steps were performed:

Standardization

The features were standardized using `StandardScaler` to ensure they are on the same scale, which is crucial for algorithms like SVM.

Train-Test Split

The dataset was split into training (80%) and testing (20%) sets to evaluate the model's performance on unseen data.

4. Machine Learning Models

Three different machine learning models were used for this project:

1. **Random Forest Classifier**
2. **Support Vector Machine (SVM)**
3. **Decision Tree Classifier**

Model Descriptions

- **Random Forest:** An ensemble method that uses multiple decision trees to improve the overall performance.
- **SVM:** A classifier that finds the optimal hyperplane for separating classes.
- **Decision Tree:** A tree-structured model used for classification and regression.

5. Model Training and Evaluation

Each model was trained on the standardized training set and evaluated on the testing set. The evaluation metrics used were:

- Accuracy
- Confusion Matrix
- Classification Report (Precision, Recall, F1-Score)

Results

Random Forest

- Accuracy: 94.74%
- Confusion Matrix:

```
[[41, 2],
 [4, 67]]
```

Classification Report:

Precision: 0.91 (Benign), 0.97 (Malignant)

Recall: 0.95 (Benign), 0.94 (Malignant)

F1-Score: 0.93 (Benign), 0.96 (Malignant)

SVM

- Accuracy: 94.74%
- Confusion Matrix:

```
[[40, 3],
 [3, 68]]
```

Classification Report:

Precision: 0.93 (Benign), 0.96 (Malignant)

Recall: 0.93 (Benign), 0.96 (Malignant)

F1-Score: 0.93 (Benign), 0.96 (Malignant)

Decision Tree

- Accuracy: 90.35%
- Confusion Matrix:

[[40, 3],

[8, 63]]

Classification Report:

Precision: 0.83 (Benign), 0.95 (Malignant)

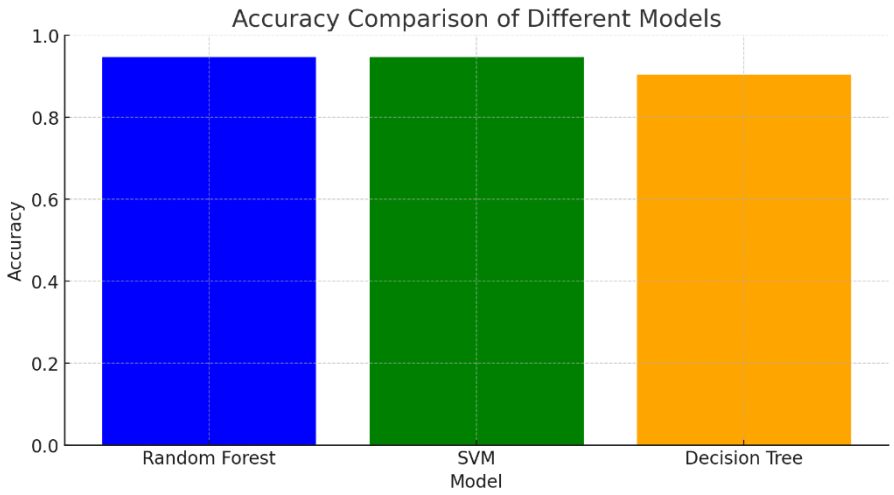
Recall: 0.93 (Benign), 0.89 (Malignant)

F1-Score: 0.88 (Benign), 0.92 (Malignant)

6. Accuracy Comparison

A bar plot was created to visually compare the accuracy of the different models:

- **Random Forest** and **SVM** both achieved the highest accuracy at 94.74%.
- **Decision Tree** had a slightly lower accuracy of 90.35%.



7. Conclusion

- Both Random Forest and SVM were the most accurate models for predicting breast cancer, with an accuracy of 94.74%.
- The Decision Tree model, while slightly less accurate at 90.35%, still performed well.