Project Report: Breast Cancer Prediction Model

1. Introduction

Breast cancer is a significant health issue, with early detection being critical for improving survival rates. This project aims to develop a machine learning model to predict whether a breast tumor is malignant or benign based on a set of features extracted from digital images of breast mass. The prediction model is intended to assist in early diagnosis, potentially leading to better outcomes for patients.

2. Dataset

The dataset used in this project is sourced from the Kaggle Breast Cancer Wisconsin Dataset, which is also available on the UCI Machine Learning Repository. The dataset contains features computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. These features describe the characteristics of the cell nuclei present in the image.

2.1. Attribute Information

The dataset comprises the following attributes:

1. **ID number**: A unique identifier for each patient.
2. **Diagnosis**: The classification of the tumor, either malignant (M) or benign (B).
3. **Ten real-valued features computed for each cell nucleus**:
    - **Radius**: Mean of distances from center to points on the perimeter.
    - **Texture**: Standard deviation of gray-scale values.
    - **Perimeter**: The perimeter of the tumor.
    - **Area**: The area of the tumor.
    - **Smoothness**: Local variation in radius lengths.
    - **Compactness**: Calculated as (perimeter^2 / area - 1.0).
    - **Concavity**: Severity of concave portions of the contour.
    - **Concave points**: Number of concave portions of the contour.
    - **Symmetry**: Symmetry of the tumor.
    - **Fractal dimension**: "Coastline approximation" minus 1.

Each feature is recorded with four significant digits and the dataset contains no missing values.

2.2. Class Distribution

The dataset consists of 569 instances, with the following class distribution:

- 357 benign cases
- 212 malignant cases

3. Methodology

3.1. Data Preprocessing

The dataset was loaded and preprocessed using the pandas library to handle data manipulation. Missing values were not an issue as the dataset contains complete records. The features were then scaled and normalized using scikit-learn to improve model performance.

3.2. Feature Selection

Given the large number of features (30), a feature selection process was conducted to identify the most relevant features for predicting the diagnosis. This was done using techniques such as correlation analysis and recursive feature elimination (RFE).

3.3. Model Training

The following machine learning algorithms were employed:

- **Logistic Regression**: A simple and interpretable model that estimates the probability of a binary outcome.
- **Decision Tree**: A non-linear model that splits data based on feature values to make predictions.
- **Random Forest**: An ensemble method that builds multiple decision trees and combines their outputs for a more robust prediction.

3.4. Model Evaluation

The models were trained and evaluated using scikit-learn. Performance metrics such as accuracy, precision, recall, and F1-score were used to assess the effectiveness of each model. Cross-validation was employed to ensure that the models generalize well to unseen data.

3.5. Deployment

The best-performing model was deployed using Flask, allowing for real-time predictions based on new input data. The deployed model accepts the coordinates of the aforementioned attributes and returns a prediction of whether the breast mass is "cancerous" (malignant) or "non-cancerous" (benign).

4. Results and Discussion

The Random Forest model outperformed Logistic Regression and Decision Tree models, achieving the highest accuracy and balanced performance across all metrics. The model demonstrated strong predictive power, particularly in correctly identifying malignant tumors, which is crucial for early intervention.

5. Conclusion

The project successfully developed a machine learning model capable of predicting breast cancer with high accuracy using features derived from digitized images of breast masses. The deployment of the model allows for practical application in clinical settings, potentially aiding in early diagnosis and improving patient outcomes.

Future work may involve further refining the model, exploring additional features, and testing the model on more diverse datasets to ensure its robustness and applicability across different populations.

6. References

- Kaggle Breast Cancer Wisconsin Dataset