**Breast Cancer Data Analysis Report**

* **Introduction**

This report presents an analysis of breast cancer data using machine learning techniques. The goal is to build models that can accurately classify breast cancer cases based on provided features. The dataset used is the Breast Cancer Wisconsin (Diagnostic) Data, obtained from the scikit-learn library.

* **Dataset Overview**
* Source: Breast Cancer Data [PROVIDED]
* Features: Various attributes computed from a digitized image of a fine needle aspirate (FNA) of a breast mass.
* Target Variable: Diagnosis (Malignant or Benign)

• **Data Pre-processing**

* Data Loading: Loaded the dataset using scikit-learn.
* Exploratory Data Analysis (EDA): Explored the features, checked for missing values, and analysed the distribution of classes (Malignant and Benign).

• **Model Building**

✓ Model Selection: Three different models were chosen for the breast cancer classification task:

1. Logistic Regression
2. Support Vector Machines (SVM)
3. Random Forest

* **Model Training and Evaluation**

Each model was trained on the dataset and evaluated using the following metrics:

* 1. Accuracy: Overall correct predictions.
  2. Precision: The ratio of correctly predicted positive observations to the total predicted positives.
  3. Recall: The ratio of correctly predicted positive observations to the all observations in the actual class.
  4. F1 Score: The Harmonic Mean of precision and recall.
* **Model Tuning**

For the Random Forest model, a Grid Search was employed to find the best hyperparameters that maximize the model's performance in terms of accuracy.

* **Results**

✓ Model Performance: The models achieved the following performance metrics on the test set:

**1. Logistic Regression: Accuracy: 0.956140350877193**

**Precision: 0.9459459459459459**

**Recall: 0.9859154929577465 F1 Score: 0.9655172413793103** **2. Support Vector Machines (SVM): Accuracy: 0.956140350877193**

**Precision: 0.9459459459459459**

**Recall: 0.9859154929577465 F1 Score: 0.9655172413793103**

**3. Random Forest:**

**Accuracy: 0.9649122807017544**

**Precision: 0.958904109589041**

**Recall: 0.9859154929577465**

**F1 Score: 0.9722222222222222**

✓ Best Model after Tuning

After hyperparameter tuning, the Random Forest model achieved an accuracy of % on the test set using the following best parameters: (parameter values).

Best Parameters: {'max\_depth': None, 'min\_samples\_leaf': 1, 'min\_samples\_split': 2, 'n\_estimators': 150}

Accuracy with Best Parameters: 0.9649122807017544

• **Conclusion**

The models performed reasonably well in classifying breast cancer, with the Random Forest model showing the highest accuracy after hyperparameter tuning. The results indicate the potential effectiveness of machine learning in aiding breast cancer diagnosis.

* **DOWNLOAD THE JUPYTER FILE:**
* [**C:\Users\hp\Downloads\BREAST CANCER DATASET.py**](file:///C:\Users\hp\Downloads\BREAST%20CANCER%20DATASET.py)
* [**C:\Users\hp\Downloads\BREAST CANCER DATASET.docx**](file:///C:\Users\hp\Downloads\BREAST%20CANCER%20DATASET.docx)