Helping women detect breast cancer earlier.

PRESENTED BY:

Priyanka Nagpal 21csu135

Akshat Rakheja 21csu219

CONTENT

INTRODUCTION

Breast cancer is one of the most common cancers among women worldwide. Early detection and accurate prediction of breast cancer are crucial for timely treatment and improved survival rates. In this project, we leverage PySpark, a powerful distributed computing framework, to develop a breast cancer prediction model using machine learning algorithms.

DATASET

We utilized the Breast Cancer Wisconsin (Diagnostic) Dataset which is publicly available and contains featured computed from digitzed images of breast mass samples .

The dataset consists of features mean radius , mean perimeter , mean texture, mean area along with diagnoses (M = Malignant , B= Benign) as the target variable

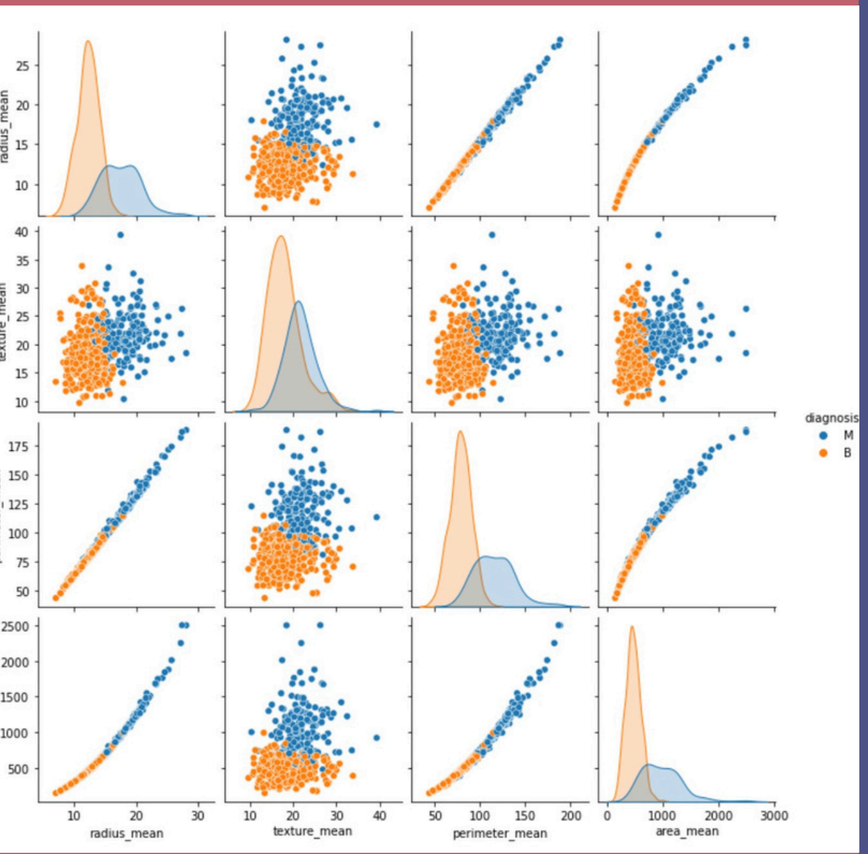
```
#Show some rows and columns of the dataframe to understand it better:
df.select(df.columns[:10]).show(8)

#Dataframe shape
print("Dataframe's shape: (%s,%d)" %(df.count(), len(df.columns)))
print("")
```

```
+--------+---------+-----------+------------+-------------+---------+--------------+---------------+------------
--+------------------+
|      id|diagnosis|radius_mean|texture_mean|perimeter_mean|area_mean|smoothness_mean|compactness_mean|concavity_me
an|concave points_mean|
+--------+---------+-----------+------------+-------------+---------+--------------+---------------+------------
--+------------------+
|  842302|        M|      17.99|       10.38|        122.8|   1001.0|        0.1184|         0.2776|         0.30
01|            0.1471|
|  842517|        M|      20.57|       17.77|        132.9|   1326.0|       0.08474|        0.07864|         0.08
69|           0.07017|
|84300903|        M|      19.69|       21.25|        130.0|   1203.0|        0.1096|         0.1599|         0.19
74|            0.1279|
|84348301|        M|      11.42|       20.38|        77.58|    386.1|        0.1425|         0.2839|         0.24
14|            0.1052|
|84358402|        M|      20.29|       14.34|        135.1|   1297.0|        0.1003|         0.1328|          0.1
98|            0.1043|
|  843786|        M|      12.45|        15.7|        82.57|    477.1|        0.1278|           0.17|         0.15
78|           0.08089|
|  844359|        M|      18.25|       19.98|        119.6|   1040.0|       0.09463|          0.109|         0.11
27|            0.074|
|84458202|        M|      13.71|       20.83|         90.2|    577.9|        0.1189|         0.1645|         0.093
66|           0.05985|
+--------+---------+-----------+------------+-------------+---------+--------------+---------------+------------
--+------------------+
only showing top 8 rows

Dataframe's shape: (569,32)
```

DATA EXPLORATION



DATA PRE PROCESSING

- handle missing values
- scale numerical features
- encode categorical variables

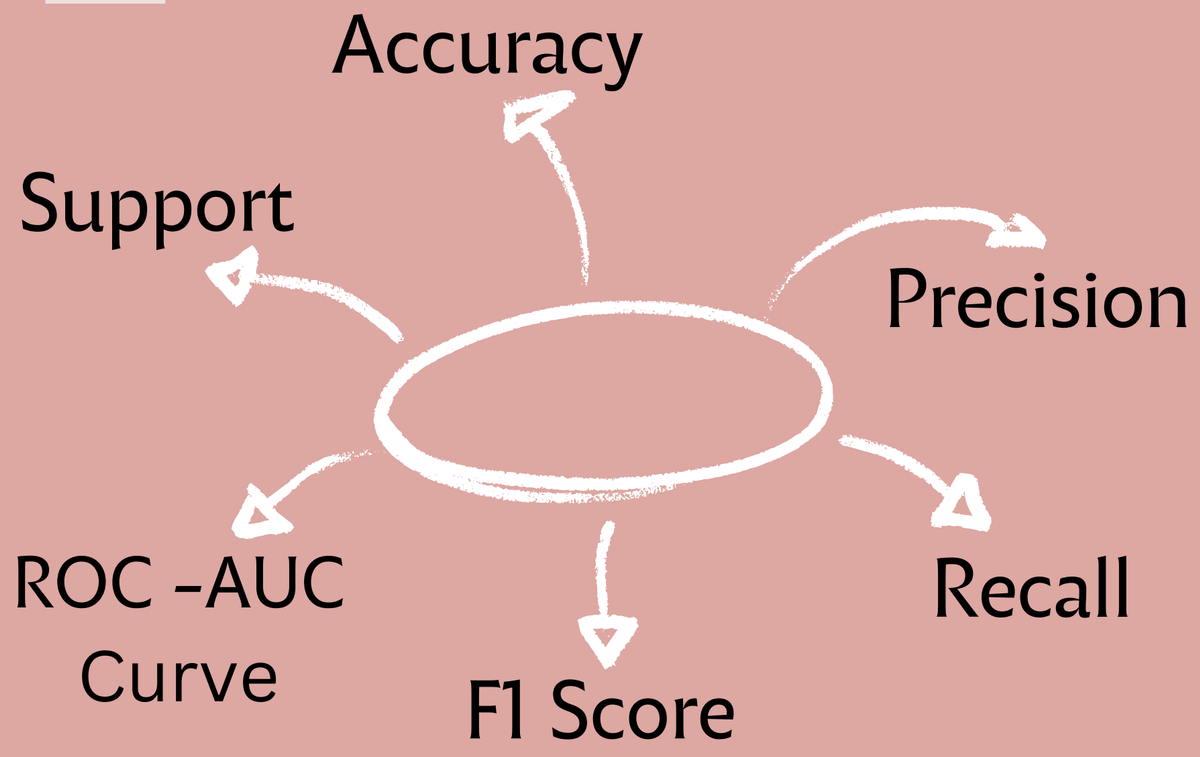
MODEL SELECTION

1. Logistic Regression
2. Decision Tree
3. Random Forest
4. Linear SVC
5. Naive Bayes

PIPELINE

Create a machine learning pipeline for each classification algorithm, comprising stages for feature indexing, vector assembly, feature scaling, and model training.

MODEL EVALUATION



RESULT AND CONCLUSION

01. Our experiment demonstrates promising results, with the models achieving high accuracy and other evaluation metrics

02. The Linear SVC classifier exhibits the best performance, achieving an accuracy of over 95% on the test dataset.

03. The breast cancer prediction model developed using PySpark offers a valuable tool for early detection and diagnosis of breast cancer.

04. Future: Explore deep learning & ensembles for better predictions.

THANKYOU