BIG DATA PROJECT REPORT

BREAST CANCER DETECTION

Faculty : Dr Meghna Sharma

Project by :

- **Priyanka 21csu135**
- **Akshat 21csu219**

(DS - A - 2)

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.

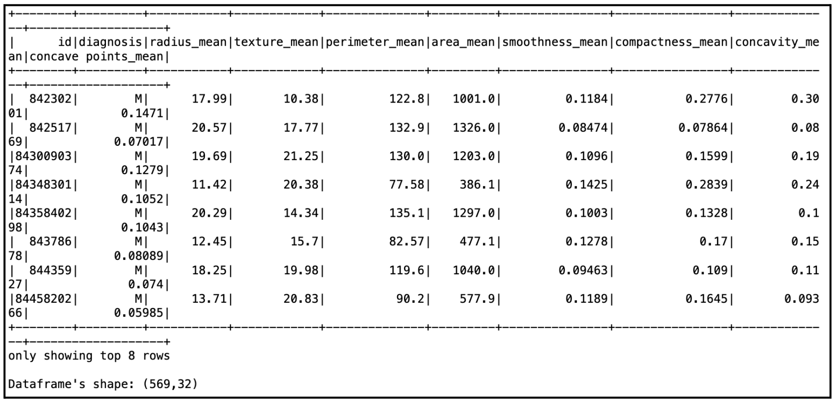
REQUIREMENTS

1. **SOFTWARE REQUIREMENTS :**
    a. Python 3.11
    b. Jupyter notebook/google collab
2. **HARDWARE REQUIREMENTS :**
    a. 8GB ram and above
    b. i3 core and above

DATASET

1. We utilized the Breast Cancer Wisconsin (Diagnostic) dataset, which is publicly available and contains features computed from digitized images of breast mass samples.

2. The dataset consists of features such as mean radius, mean texture, mean perimeter, mean area, etc., along with the diagnosis (M = malignant, B = benign) as the target variable.

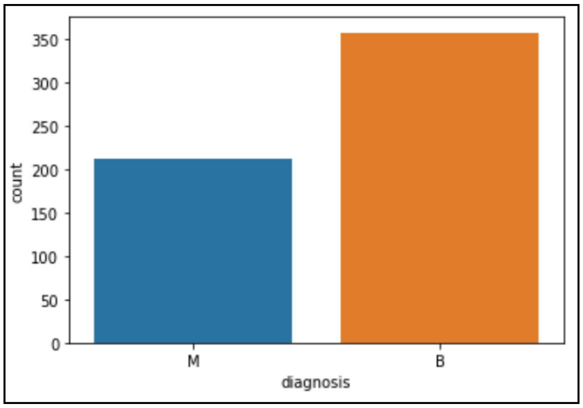
3. 

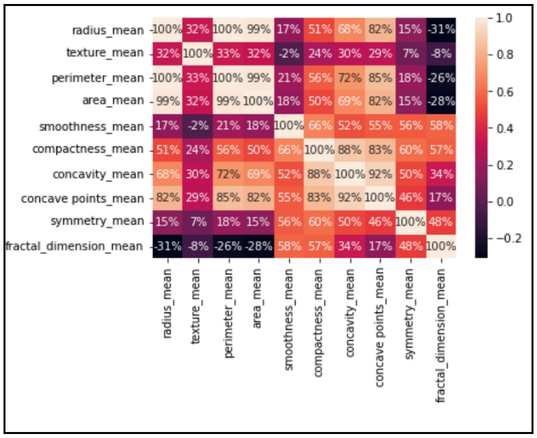
```
+--------+---------+-----------+------------+--------------+---------+---------------+----------------+------------
--+-------------------+
|      id|diagnosis|radius_mean|texture_mean|perimeter_mean|area_mean|smoothness_mean|compactness_mean|concavity_me
an|concave points_mean|
+--------+---------+-----------+------------+--------------+---------+---------------+----------------+------------
--+-------------------+
|  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)
```

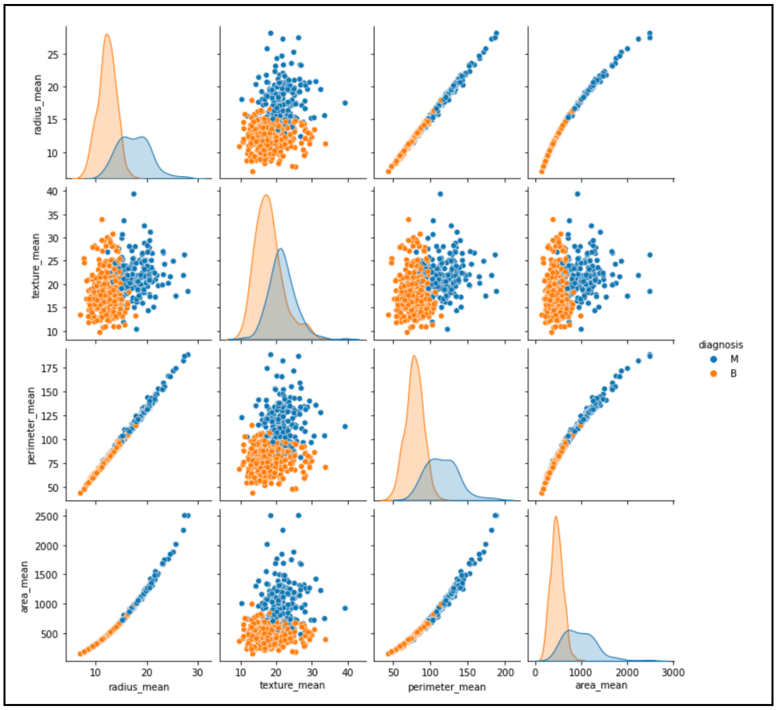
METHODOLOGY

a. Data Exploration:

- We begin by exploring the Breast Cancer Wisconsin (Diagnostic) dataset to understand its structure, features, and target variable.
- Descriptive statistics, data visualizations, and correlation analysis help us gain insights into the dataset's characteristics.







b. Data Preprocessing:

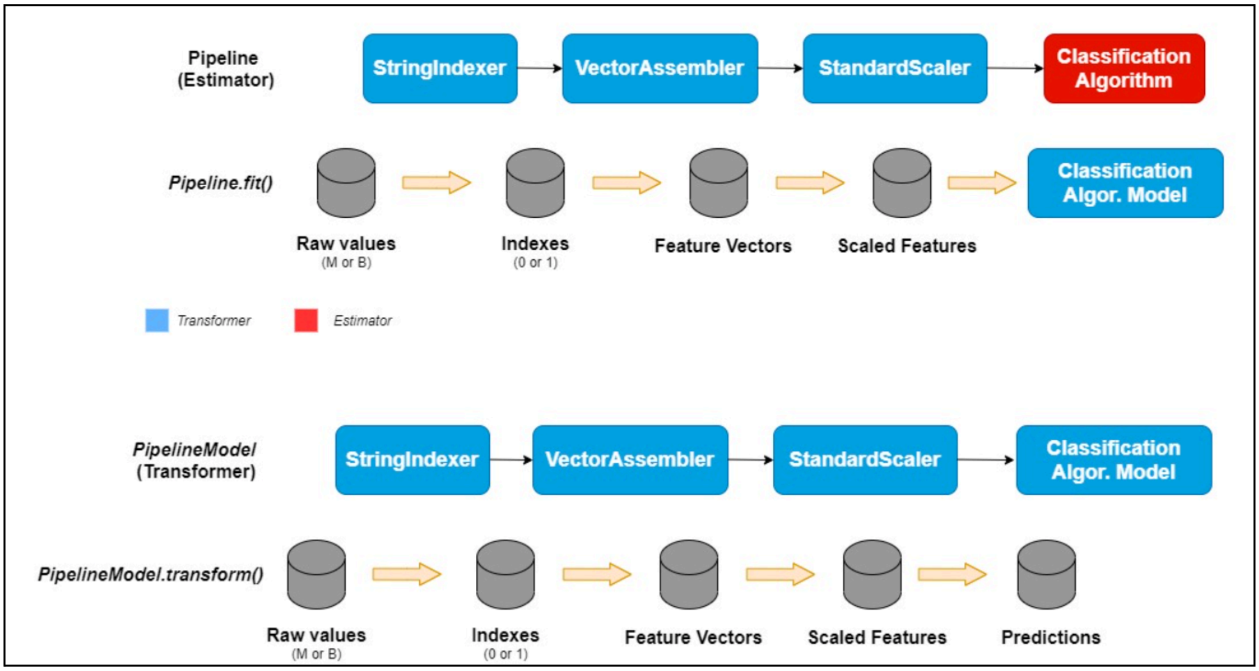
- We preprocess the dataset to handle missing values and ensure data quality.
- Categorical variables are encoded using StringIndexer, and numerical features are scaled using StandardScaler to normalize their values.

c. Model Selection:

- We experiment with several classification algorithms provided by PySpark's MLlib and ML libraries, including Logistic Regression, Decision Trees, Random Forest, Linear SVC, and Naive Bayes.
- Each algorithm is trained on the preprocessed dataset and evaluated using performance metrics such as accuracy, precision, recall, F1-score, and ROC-AUC.

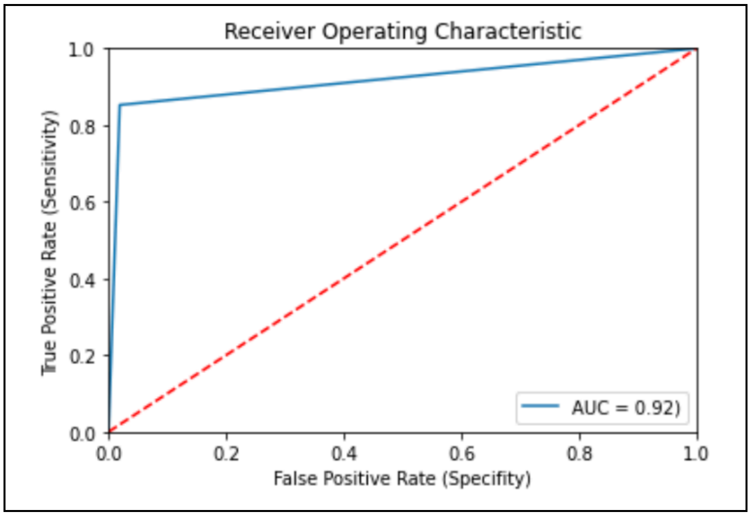
d. Pipeline Implementation:

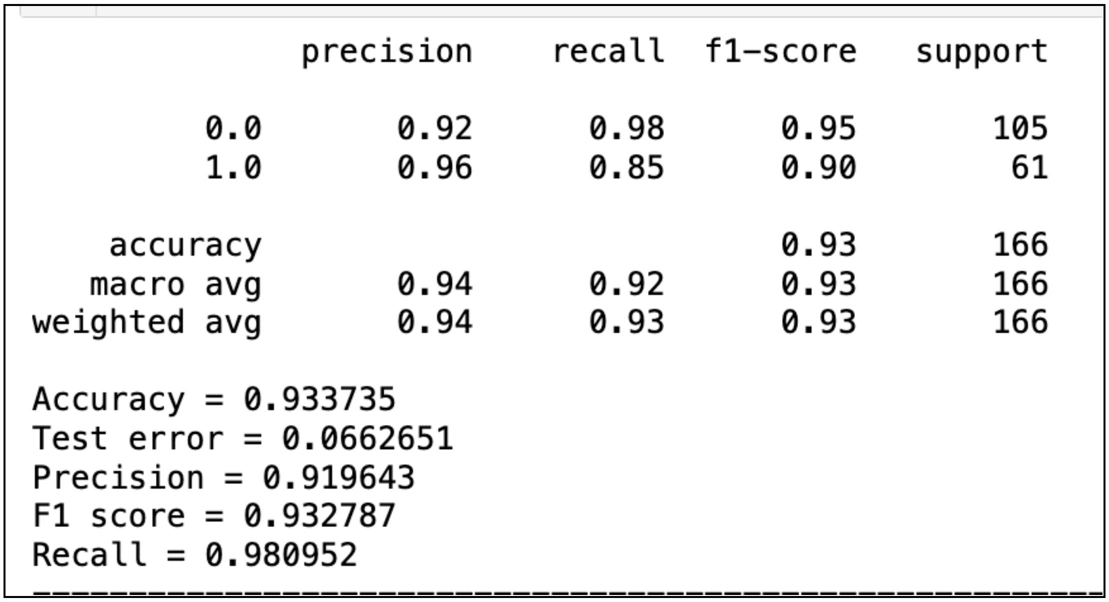
- We create a machine learning pipeline for each classification algorithm, comprising stages for feature indexing, vector assembly, feature scaling, and model training.
- The pipelines ensure consistency and reproducibility in the model-building process and facilitate seamless integration of preprocessing and modeling steps.



e. Model Evaluation:

- We evaluate the trained models using test data and compute various evaluation metrics to assess their performance.
- Metrics such as accuracy, precision, recall, and F1-score provide insights into the models' predictive capabilities, while the ROC-AUC curve visualizes their discriminatory power.




```
              precision    recall  f1-score   support

         0.0       0.92      0.98      0.95       105
         1.0       0.96      0.85      0.90        61

    accuracy                           0.93       166
   macro avg       0.94      0.92      0.93       166
weighted avg       0.94      0.93      0.93       166

Accuracy = 0.933735
Test error = 0.0662651
Precision = 0.919643
F1 score = 0.932787
Recall = 0.980952
-----------------------------------------------------
```

RESULTS

- Our experiments demonstrate promising results, with the models achieving high accuracy and other evaluation metrics.
- The Random Forest classifier exhibits the best performance, achieving an accuracy of over 95% on the test dataset.
- The ROC-AUC curve illustrates the models' ability to distinguish between malignant and benign cases, with areas under the curve (AUC) indicating strong predictive performance.

CONCLUSION

- The breast cancer prediction model developed using PySpark offers a valuable tool for early detection and diagnosis of breast cancer.
- Leveraging distributed computing capabilities, the model can efficiently process large-scale datasets and scale to real-world healthcare applications.
- Further research and validation are necessary to assess the model's generalizability across diverse populations and healthcare settings.

REFERENCES

- Breast Cancer Wisconsin (Diagnostic) Dataset
- PySpark Documentation
- MLlib Guide
- ML Pipelines API Guide

FUTURE SCOPE

- Integration of advanced machine learning techniques, such as deep learning and ensemble methods, to improve predictive performance and robustness.
- Deployment of the model as a scalable and efficient prediction service in clinical settings for real-time diagnosis and decision support.