Breast Cancer Data Analysis Using PySpark

Name: Bhargav Marre
Roll No.: 2211CS010363
Academic Year: 2025 - 2026

1. Dataset Description

The dataset consists of breast cancer diagnostic measurements collected from biopsy samples. Each record includes multiple numeric features describing cell nuclei characteristics such as radius, texture, perimeter, area, and smoothness. The target variable, “Diagnosis,” classifies tumors as either malignant (M) or benign (B). Missing values and duplicates were checked and appropriately handled. The dataset was standardized for column names and numeric formats to ensure compatibility with PySpark operations.

2. Operations Performed

- Data loading and schema inspection using PySpark DataFrame API.
- Normalization of column names and casting of numeric fields.
- Detection and handling of missing or duplicate entries.
- Summary statistics computed for each numeric column.
- Correlation analysis between numeric features.
- Visualization of distributions and diagnosis-wise comparisons using Seaborn and Matplotlib.

3. Key Insights

- Malignant tumors exhibit higher mean radius, perimeter, and area values than benign ones.
- Strong correlation observed among geometric features (radius, area, perimeter).
- Compactness, concavity, and smoothness features show distinct group distributions by diagnosis.
- The correlation matrix reveals redundant relationships among certain size-based attributes.
- Boxplots show clear feature separation between malignant and benign tumors.

4. Visualizations

Graph Type	Description
Feature Distributions	Histograms of key numeric features showing their value spread across samples.
Diagnosis-wise Boxplots	Comparison of feature distributions between malignant and benign tumors.
Correlation Heatmap	Matrix visualization showing inter-feature relationships and strength of correlations.
Pairplot (Optional)	Pairwise scatter plots highlighting separability between tumor classes.

5. Recommendations

- Perform dimensionality reduction (e.g., PCA) to minimize redundant features.
- Use PySpark MLlib for classification modeling to predict diagnosis.
- Integrate model evaluation metrics (precision, recall, F1-score) for accuracy assessment.
- Develop dashboards for medical professionals to visualize feature relationships in real time.
- Maintain data privacy and compliance with healthcare regulations during deployment.

6. Future Work & Conclusion

This project demonstrates how PySpark can be effectively used for large-scale medical data analysis. The exploratory study of breast cancer features highlights patterns that differentiate malignant and benign tumors. Future work may involve building machine learning classification pipelines using PySpark MLlib and automating data visualization dashboards for diagnostic support. PySpark’s scalability and efficiency make it an excellent tool for processing healthcare datasets and generating actionable insights that can support early detection and better treatment planning.