FINAL PROJECT CHECK-IN

Problem Statement

- Breast Cancer is a disease commonly found in women, an abnormally grown cells in the breast tissue referred as tumor. These tumors are not always cancer cells, these are sometimes benign, pre-malignant or malignant. To identify and diagnose this various tests such as MRI, ultrasound, biopsy and mammograms are used.

- The data set we considered has details regarding the breast cancer results from breast fine-needle aspiration test, which involves collecting some fluid or cells from a breast cyst and results are classified and reported as ‘1’ and ‘0’ which refers to Malignant(presence of cancer cells) and Benign(absence of cancer cells).

- This is considered as a classification problem in machine learning. The goal or the objective of this project is to classify whether the breast cancer is malignant or benign and also predict the recurrence and non recurrence of the cases using Kmeans, Tree classifier and Logistic Regression.

Data Details

- The data set is considered from GitHub: https://github.com/milaan9/93_Python_Data_Analytics_Projects/tree/main/007_Breast_Cancer_Prediction_with_ML
- I have tried to perform the initial data cleaning by identifying the number of rows and columns in the dataset, datatypes of all the columns, null values if any.(attached few screenshots)
- We will be using heatmap, scatter matrix to understand the correlation of the data and then go ahead with the machine learning models using spark.

```
In [14]: len(df_cancer.index)

Out[14]: 569
```

```
In [6]: len(df_cancer.columns)

Out[6]: 32
```

```
In [7]: df_cancer.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 569 entries, 0 to 568
Data columns (total 32 columns):
 #   Column                   Non-Null Count  Dtype
---  ------                   --------------  -----
 0   id                       569 non-null    int64
 1   diagnosis                569 non-null    object
 2   radius_mean              569 non-null    float64
 3   texture_mean             569 non-null    float64
 4   perimeter_mean           569 non-null    float64
 5   area_mean                569 non-null    float64
 6   smoothness_mean          569 non-null    float64
 7   compactness_mean         569 non-null    float64
 8   concavity_mean           569 non-null    float64
 9   concave points_mean      569 non-null    float64
 10  symmetry_mean            569 non-null    float64
 11  fractal_dimension_mean   569 non-null    float64
 12  radius_se                569 non-null    float64
 13  texture_se               569 non-null    float64
 14  perimeter_se             569 non-null    float64
 15  area_se                  569 non-null    float64
```

```
In [9]: df_cancer.columns

Out[9]: Index(['id', 'diagnosis', 'radius_mean', 'texture_mean', 'perimeter_mean',
       'area_mean', 'smoothness_mean', 'compactness_mean', 'concavity_mean',
       'concave points_mean', 'symmetry_mean', 'fractal_dimension_mean',
       'radius_se', 'texture_se', 'perimeter_se', 'area_se', 'smoothness_se',
       'compactness_se', 'concavity_se', 'concave points_se', 'symmetry_se',
       'fractal_dimension_se', 'radius_worst', 'texture_worst',
       'perimeter_worst', 'area_worst', 'smoothness_worst',
       'compactness_worst', 'concavity_worst', 'concave points_worst',
       'symmetry_worst', 'fractal_dimension_worst'],
      dtype='object')
```

Proposed Solution

- Identify the relationship between different variable combination and then reduce the number of data visualized while preserving the relevant information.
- Implement the machine learning models like K means, Tree classifier and Logistic Regression.
- The project will include the initial exploratory data analysis using pandas, seaborn, matplotlib which provides us with the useful knowledge about data pre-processing and then we will be considering the models of ML to address the classification problem identified in our data set(referring to benign and malignant). Our data set has two outcomes, and these models will help us in generating predictions.
- I hope to derive at the conclusion where these algorithms help in improvising the diagnoses and identifying the risk and outcome predictions.

Thank You