DOCUMENTATION FOR DATA SCIENCE TASK IMPLEMENTATION

BREAST CANCER PREDICTION PROJECT

Task Description

Task: Classification
Target Variable : Diagnosis (‘malignant or ‘benign’)
Features: Various measurements of breast mass, such as radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry, and fractal dimension, measured at mean, standard error, and 'worst' or largest (mean of the three largest values).

1. Chart

Chart Used: Bar chart, Histogram, Heat map, Scatter plot
Reason: Bar chart was used to visualize the count of diagnosis (i.e the target variable). The Histogram for each feature was plotted to visualize the distribution (it reveals if there is either a normal distribution or not) of values for malignant and benign diagnosis. The heat map reveals the correlation matrix to understand the relationship between features.

2. Algorithm

Logistics Regression: it is a suitable algorithm for binary classification tasks. Given that the dataset seems to have a classification nature with ‘Diagnosis’ (malignant or benign), Logistics Regression performs effectively in predicting the diagnosis. It has a good accuracy score.

Other Algorithms were used such as Support Vector Machine and XGBoost Classifier, which is a gradient boosting machine. It is an ensemble that can capture complex relationships in data, and also provides a good predictive accuracy score.

3. Feature Scaling

Feature Scaling Method: Robust Scaler

Robust Scaler helps to maintains the shape of the distribution and is less sensitive to outliers.
It helps to ensure effective model performance in predicting the target variable after training the model.

4. Feature Engineering

No feature engineering was applied on the dataset.

5. Data Cleaning

A column was dropped, because it was filled with null values.