Data Analysis Report

GitHub Repository: Link

Report: **Breast Cancer Dataset Analysis**

Prepared by: G.ARJUN Student Number: 23096918

Introduction:

Breast cancer: 569 samples, 30 numeric features to describe cell nuclei obtained from biopsies that can be used to classify tumors as malignant or benign. This study will provide an investigation in terms of distributions, and relations and clustering to improve understanding around features describing breast cancer characteristics which could help to viable detection/detection models.

Exploratory Analysis & Statistical Summary:

EDA digs into the data to find first signs of patterns: Summary Statistics: Mean, median, standard deviation and ranges of each feature provide a basic overview of the nature of the data. Check Balance of the Dataset — Target Variable Distribution: Make Sure You Have Almost Same Number Of Malignant and Benign Samples Before Moving Forward. A heatmap can expose correlations between features which may assist in feature selection or engineering. Visualizing Feature Distributions: Use histograms, box plots and scatter plots to look at feature distributions, outliers and potential separability of the classes.

Statistical Summary Dataset has mean, minimum and maximum values in a features like radius etc and the target variable (benign = 0, malignant =1) Summaries demonstrate variance of these features across benign and malignant tumors, with the introduction of ”cluster” column (0-1–2…) supports clustering analysis. This dataset contains 569 instances of breast cancer data (malignant and benign), features are computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. Summary statistics describe the ranges and distributions of these features The target labels are 0 (benign) and 1 (malignant), with an average label value of around 0.63 Data received as a cluster column and clustered by the factor into three categories (0, 1, 2) — samples. This data classification purpose should be for predicting based on the quantitative features whether a patient has breast cancer or not.

Correlation Analysis

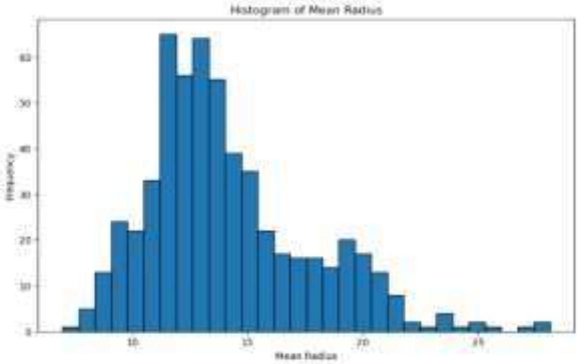
Key correlations: There are positive correlations between mean radius, perimeter and area which means that they all increase or decrease together.

So the negative correlations for these features (mean radius, perimeter and area) w.r.t. target variable means they can be predictive as per to Tumor type which helps in classifying.

In this, this matrix denotes relationships between features in the breast cancer dataset. Mean radius, perimeter and area have strongly positive correlations with one another but similarly negatively linearly dependent on their apoptotic equivalents:. Stat. Mean texture on the other hand demonstrates a lower correlation with all of these variables. It also shows strong negative correlation of target variable with mean radius, perimeter and area columns meaning that the caner type (malignant/ benign) are highly predicted as Malignant based on these measurements.

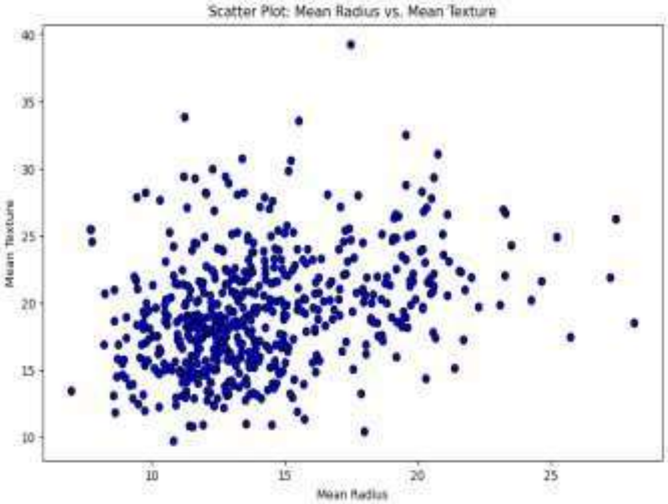
Exploratory Data Analysis:

Histogram PLOT:



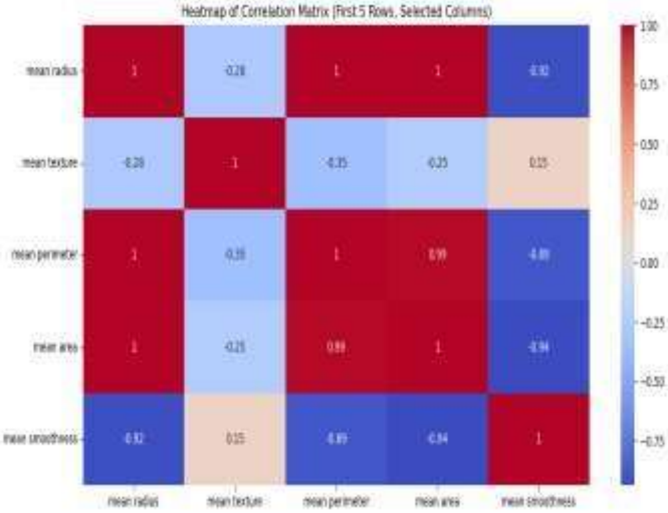
Visually it indicates that most data The x is bunched mostly between 15 and 20 to sink in the axis, where it shows a max on this range. On the Histogram, the frequency of a y-axis of data points flying in to every bin highlighting how many values are in several mean intervals radius. This focus on on range indicates a size that is most probably expected; nature of the objects or samples under investigation More detailed dataset background and what maters in the mean radius values enhance understanding.

Scatter Plot :



The scatter plot details a K-Means clustering (two clusters) using the mean radius and texture, two features used to define different tumor types. While these show valuable insights, interpreting cluster separations must be performed carefully since unsupervised clustering does not explicitly represent medical labels.This is the scatter plot showing kmeans clustering of mean radius and Mean texture to group data points into distinct clusters based on similarity. This hints at 2 separate clusters like so:

Heatmap plot:



Counts of relationships in a correlation heatmapModerate to high linear relationship (e.g. mean radius with perimeter and area) Those negative correlations are probably due to mean radius and mean smoothness —exchange them. Feature correlations may contain implications of feature dependencies, which can be beneficial in eliminating unnecessary features and enhancing model explainability. The provided heatmap shows the correlation matrix of numerical variables related to cell or biological sample measurements. It shows the relationships i.e., mean radius and means smoothness have a strong negative correlation (-0.92) which means that as mean radius is growing, it decreases usually tend to Withee crease in scenario.

Insights and Observations:

Breast Cancer dataset exploration has shown machine learning models have good opportunity to classify tumors as Malignant or Benign on characteristics such as cell size, shape and texture. Clustering methods and correlation analysis give an understanding of data dimensions, separability that helps in feature selection and model iteration. In that they will be crucial for the development of clinically relevant diagnostic tools to facilitate earlier detection and patient-specific intervention in breast cancer.

Conclusion:

The most correlated features in the Breast Cancer dataset are derived from tumor size and can be instrumental to differentiate between malignant tumors (cancer) and benignity ones. Furthermore, clustering techniques can provide more of the dynamic division in tumor samples on a biological level. These findings will contribute to continued research in the development of diagnostics for breast cancer and more optimized treatment routes.