DSC 505 Assignment-1

1. Introduction to Data Exploration

The first lesson from this course is proper data exploration should precede analysis. I was able to find out how to access information about the structure and content of a dataset using pandas, which can quickly fetch insights into a dataset's structure and content. This is quite necessary in establishing a stable basis of decision when deeper into the project.

2. Managing Missing Data:

Real-world data is mostly incomplete, so the project was able to drive home that identifying and dealing with missing values were important parts of any analysis. By practicing how missing data may affect the integrity of a dataset, I learned how filling or removing such data are properly critical for accurate analysis. This skill ensures that the insights derived from any dataset are reliable.

3. Class Distribution Analysis:

The reason I evaluated the target variable distribution, whether the cancer detected was malignant or benign was to find how checking for class imbalances is so critical. This has some impacts not only during analysis but also on how future machine learning models should be built given such an imbalance.

4. Importance of Data Preprocessing

One thing that proved very precious was data preprocessing, especially cleaning and encoding. The lesson of learning how to drop columns that are irrelevant and transform the type of categorical variables (like converting malignant/benign into 1s and 0s) means how preprocessing actually preprends the actual modeling.

This lesson reminds me that no matter how complex the analysis might seem to be, it will still be wrong if the data isn't prepared well.

5.Feature Scaling and Normalization:

Working with feature scaling was a big takeaway. I learned that features may be misinterpreted by the model for being of different ranges if feature scaling is not taken into consideration. Using the StandardScaler() to normalize features, I see how this brings about all variables being treated equally for analysis or feeding data into machine learning algorithms.

6. The Power of Data Visualization:

Rendering data into visualization using matplotlib and seaborn improved the ability to communicate insights and instilled a newfound importance of visual storytelling in data science. Visualizing the distribution of diagnoses allowed me to grasp my data much better instantaneously with what is normally obscure information now lay bare before my eyes.

7 . Ethics and Human Aspect of Data Analysis

This project used great ethics in its considerations as data science becomes necessary when working with sensitive data like health records. Each row of data represented a person's life, so that realization gave me proper responsibility that data scientists carry in analysis. This made me understand that one should hence approach the analysis of data with care and respect for the human context behind it.

R. CHANDU BADRINATH MANIKANTA
AP24122060018

DSC 505 Assignment-1

8. Preparation for Machine Learning:

Preprocessing and Scaling Features-was fundamental training regarding practical application of machine learning models: So, I understood why a properly preprocessed and scaled dataset would make such a huge difference in model performance; an otherwise improperly prepared data set could produce awkward or biased results.

After performing the pre-processing techniques on the provided dataset i.e Cancer detection,below are the findings from the exploratory data analysis and visualizations.

1.Out of the all observations, 37.3% of them were diagnosed to be malignant (risky)and 62.7% of them were benign. (almost out of danger)

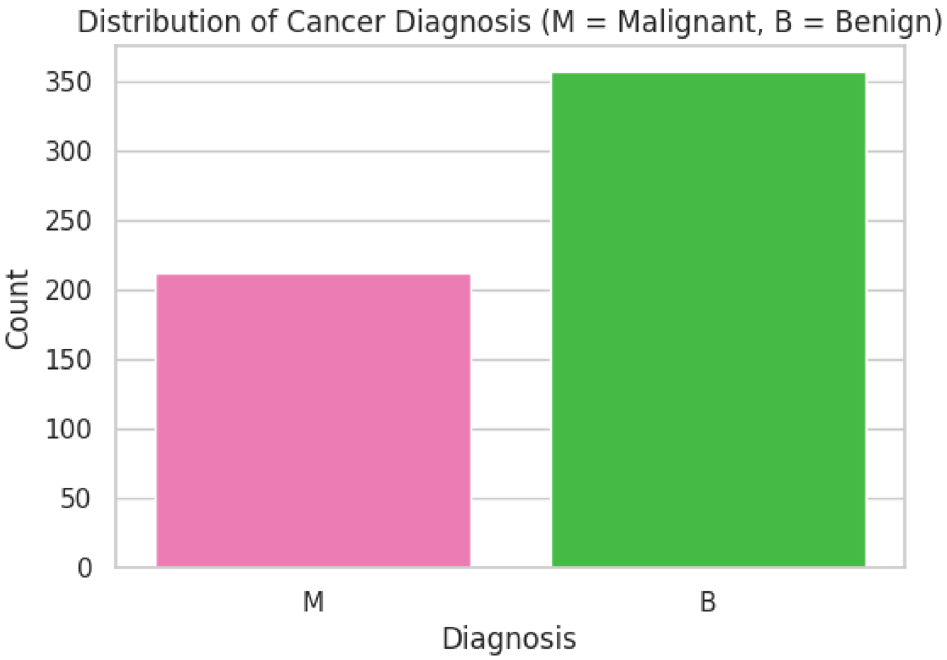
Approximately 212 of them were in the Risky stage.

2. Out of all, 25th percentile of them in Malignant are having the area mean just below 1000, whereas for benign it falls under 500.
3. While observing scatter plot between the Radius mean and texture mean by diagnosis, we can understand that ‘M’ observations are more widely spread out and most of the observations are even falling into the cluster of ‘B’ , while having the capacity to become potential outliers.

4.All the radius,texture,perimeter and area mean were more incase of ‘B’ rather than ‘M’ by observing area plots.

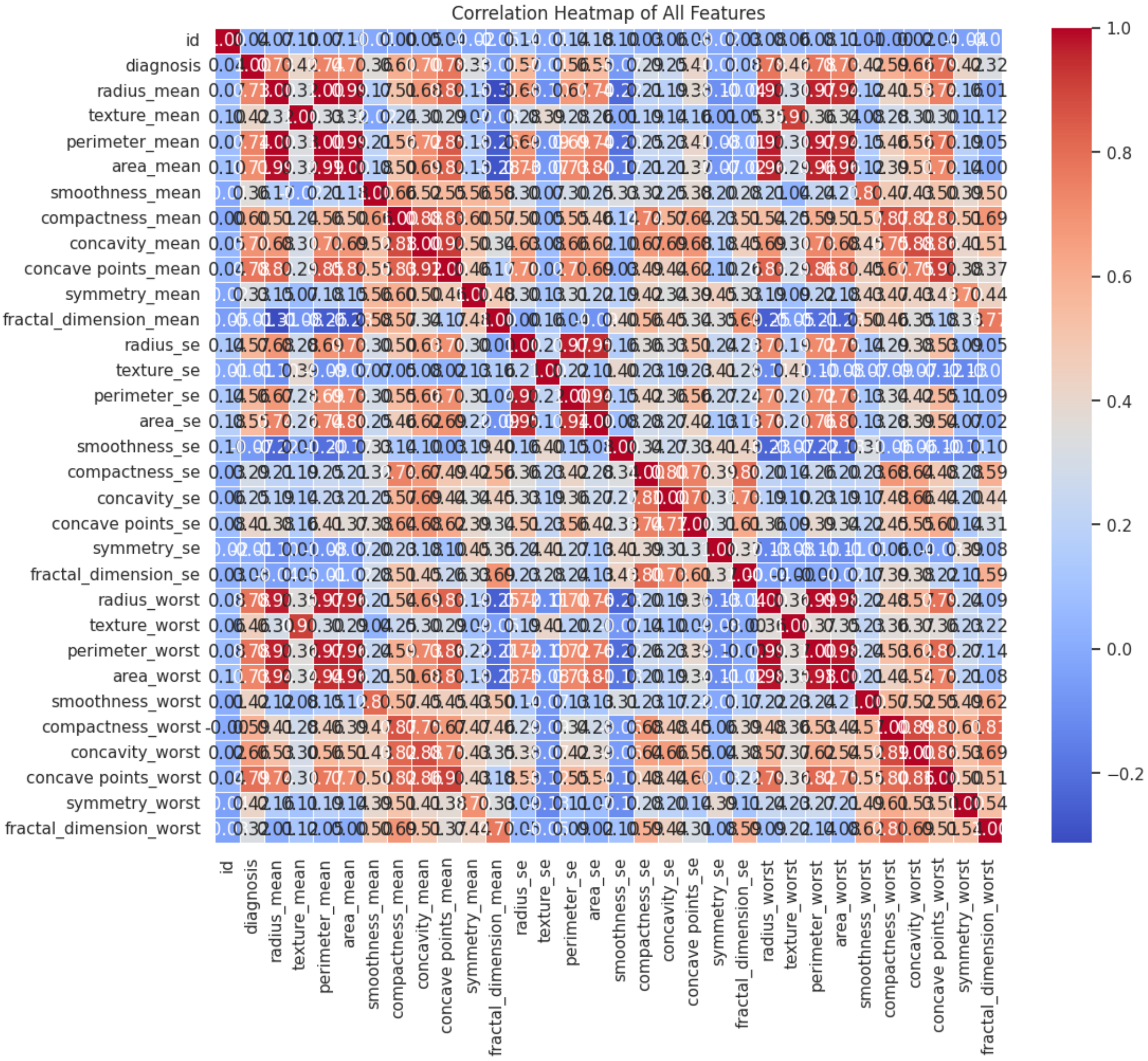
5. In the cluster map, ‘B’ points are more scattered in the perimeter_mean v/s texture than ‘M’.
6. Incase of texture_mean v/s area_mean the scattering is less for ‘B’ thus by reducing the scope of outliers

Visual Representations



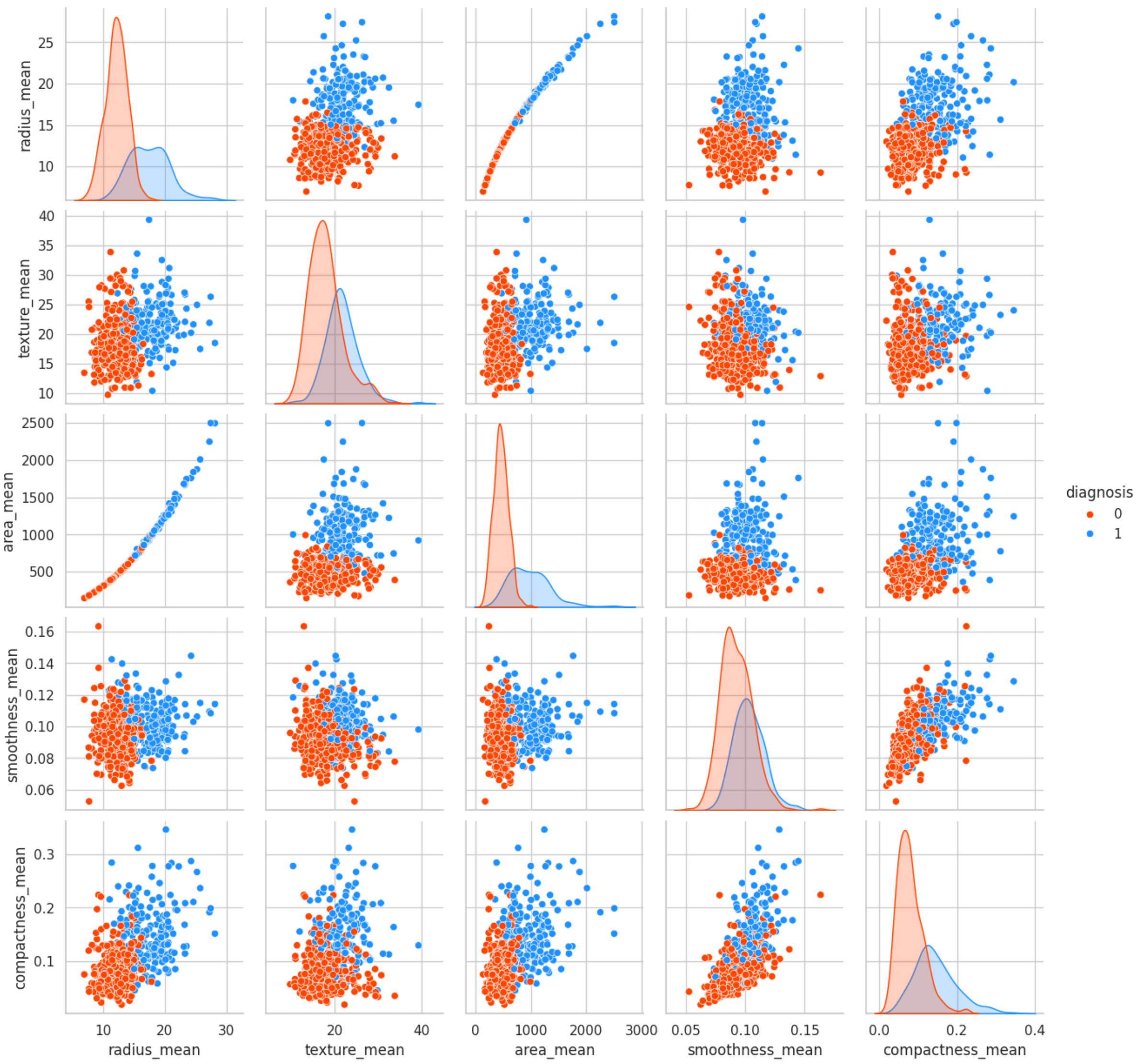
R. CHANDU BADRINATH MANIKANTA
AP24122060018

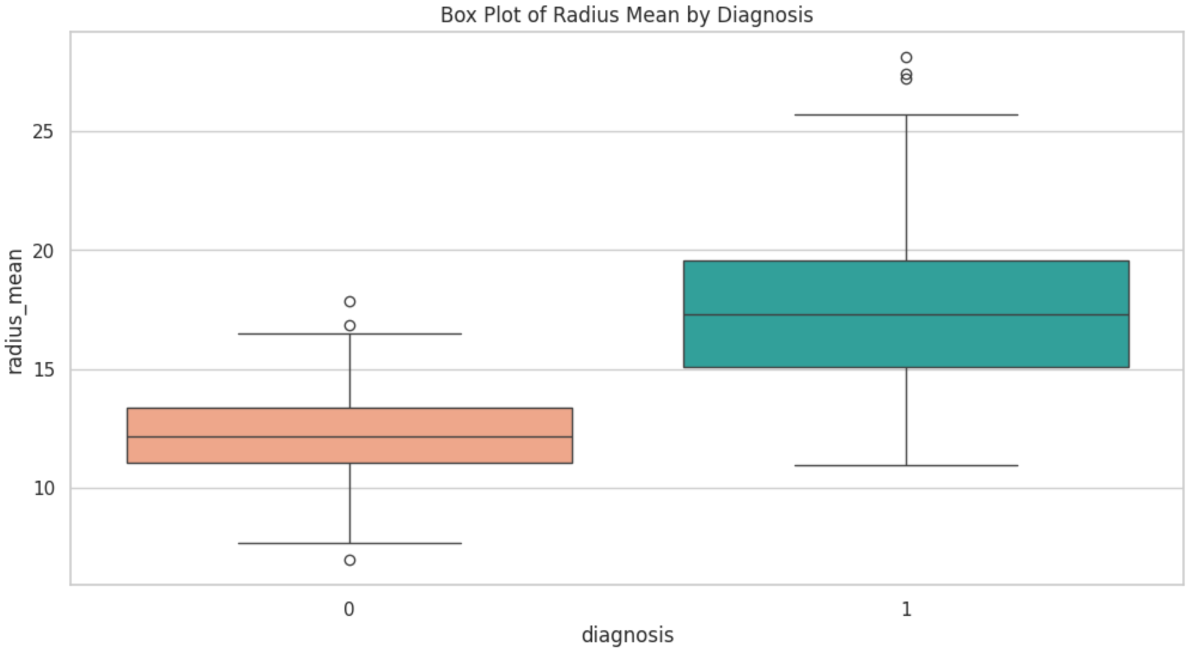
DSC 505 Assignment-1



R. CHANDU BADRINATH MANIKANTA
AP24122060018

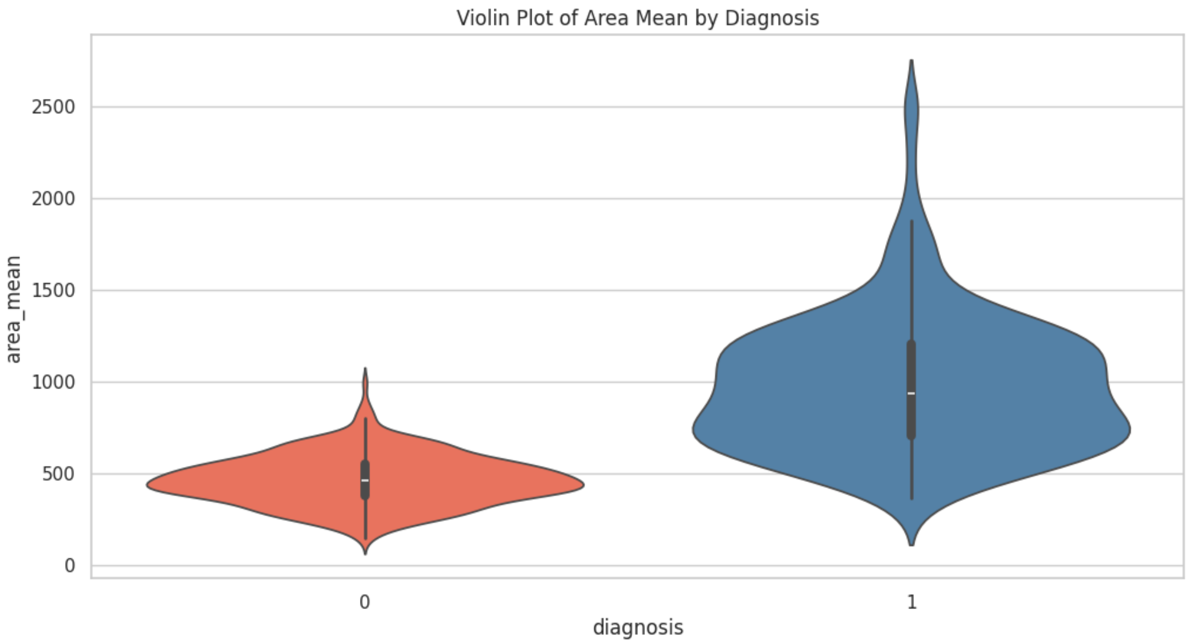
DSC 505 Assignment-1


Pair Plot of Selected Features by Diagnosis


Box Plot of Radius Mean by Diagnosis

R. CHANDU BADRINATH MANIKANTA
AP24122060018

DSC 505 Assignment-1



R. CHANDU BADRINATH MANIKANTA
AP24122060018