Bversity

1st Ever Virtual College for Biotech

Name : K Ranjani priya
Contact: ranjanipriya2610@gmail.com
Mobile: 7010870690

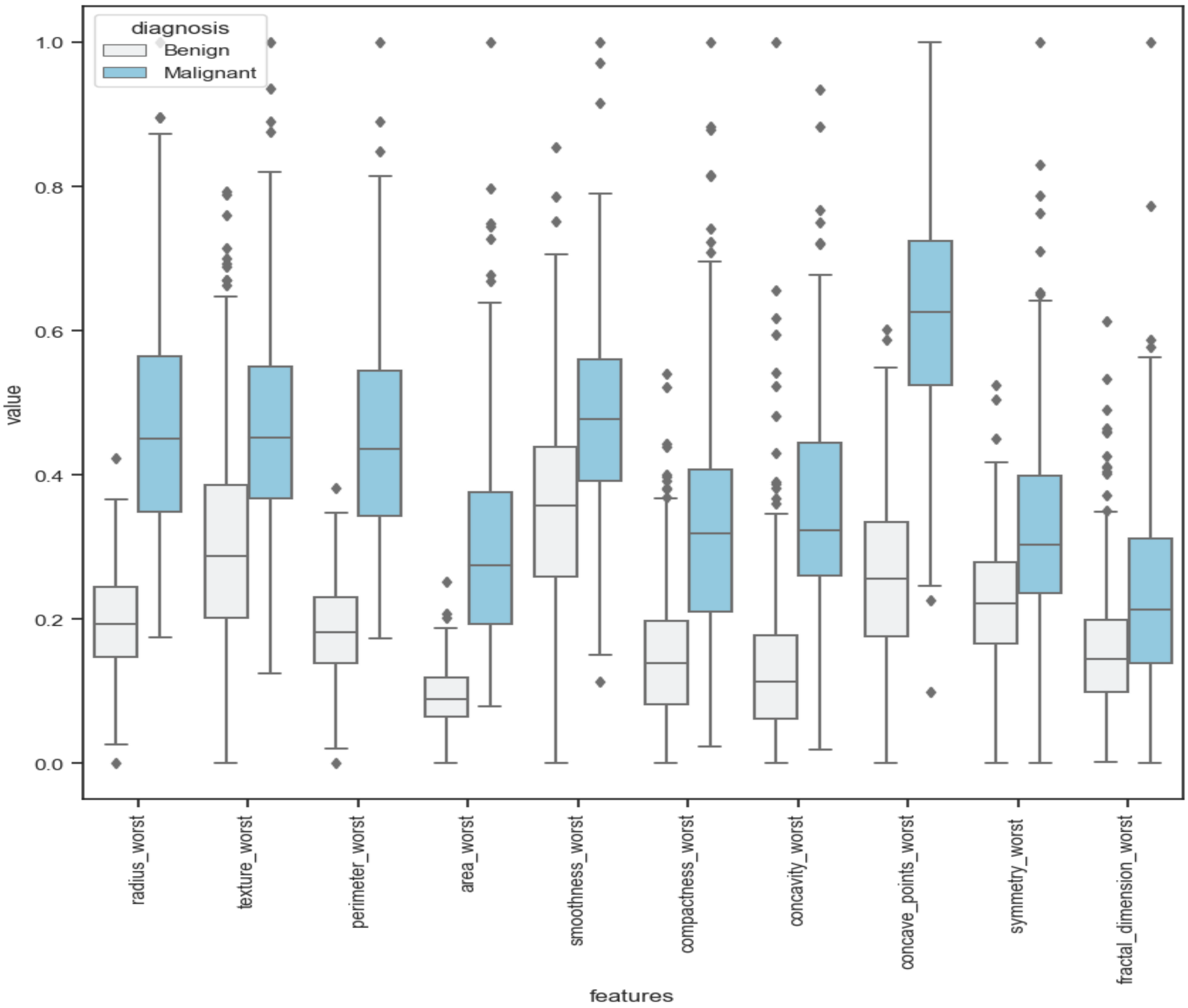
Title: Python Exploratory Data analysis

Answer the following questions below and upload them in the google form.

1. How does the distribution of feature “fractal_dimension_worst” differ between benign and malignant cases?
2. What is the range of values for the feature “radius_mean” and how skewed is its distribution?
3. Are there any outliers in feature “area_mean” and how might they affect analysis?
4. Based on the EDA, what factors seem to be most relevant to predicting breast cancer diagnosis?
5. What limitations are there in the data, and how might they affect our conclusions?

1) The distribution of feature “fractal_dimension_worst” differ between benign and malignant cases:

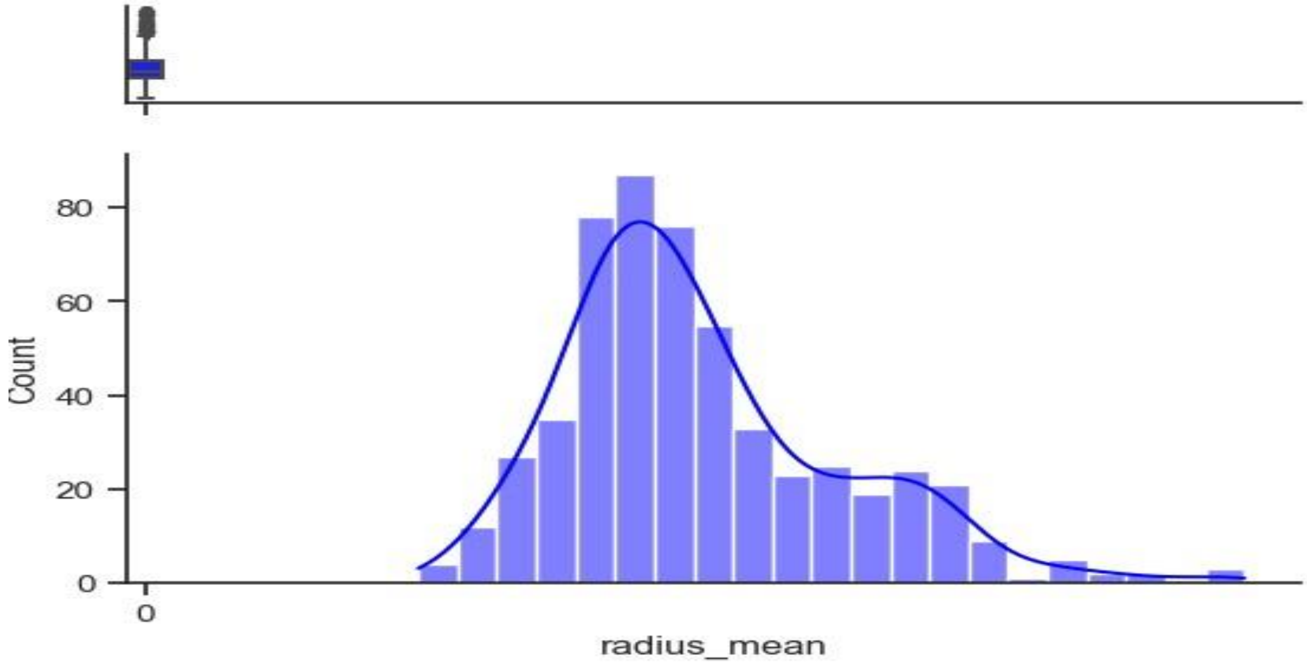
- The graph showed that this would display two density curves, one for each diagnosis group, using different colors or line styles to differentiate them each of them.
- There are two separate histograms, one for benign cases and one for malignant cases, arranged side-by-side or arranged vertically.



- In order to show the differences between benign and malignant cases a box plot of the total distribution would be displayed.
- The individual histograms for benign and malignant cases, allowing for direct comparison of their shapes and spreads.

- This would create separate density curves for each group, similar to the overlayed version but with more space for visual clarity.
- The overall distribution of "fractal_dimension_worst" appears to be somewhat skewed to the right, with a peak around 0.05 and a tail extending towards higher values.
- If malignant cases tend to have higher "fractal_dimension_worst" values, we might expect a slight slope or longer tail on the right side of the distribution.
- If benign cases have lower values, the distribution might be more concentrated towards the left.

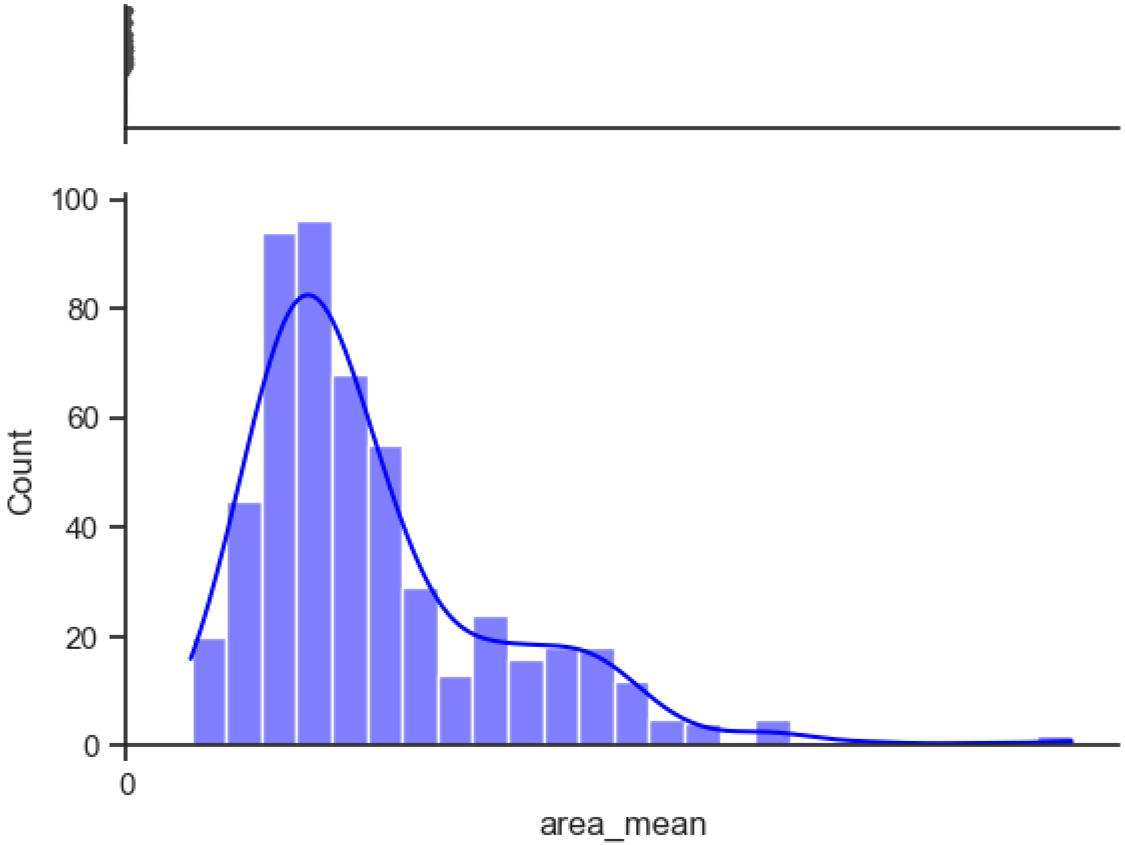
2) The range of values for the feature "radius_mean":



Based on the graph, the range of values for the feature "radius_mean" is 0 to 100. The distribution is skewed to the left. The graph's x-axis extends from 0 to 100, indicating that these are the minimum and maximum values observed for "radius_mean" in the dataset.

- The distribution is not symmetrical. It has a longer tail on the left side, with more values concentrated towards the higher end of the range. This is characteristic of a left-skewed distribution.
- The distribution has a peak around 40, suggesting that this is the most common value for "radius_mean".
- The distribution has a moderate spread, indicating that there is some variability in the "radius_mean" values.
- The image suggests that there were missing values in the "radius_mean" column, and they were filled with the median value.
- In a left-skewed distribution, the mean is lower than the median.This is because, compared to the median, the mean is closer to the extreme values in the tail.

3) Outliers in feature “area_mean”

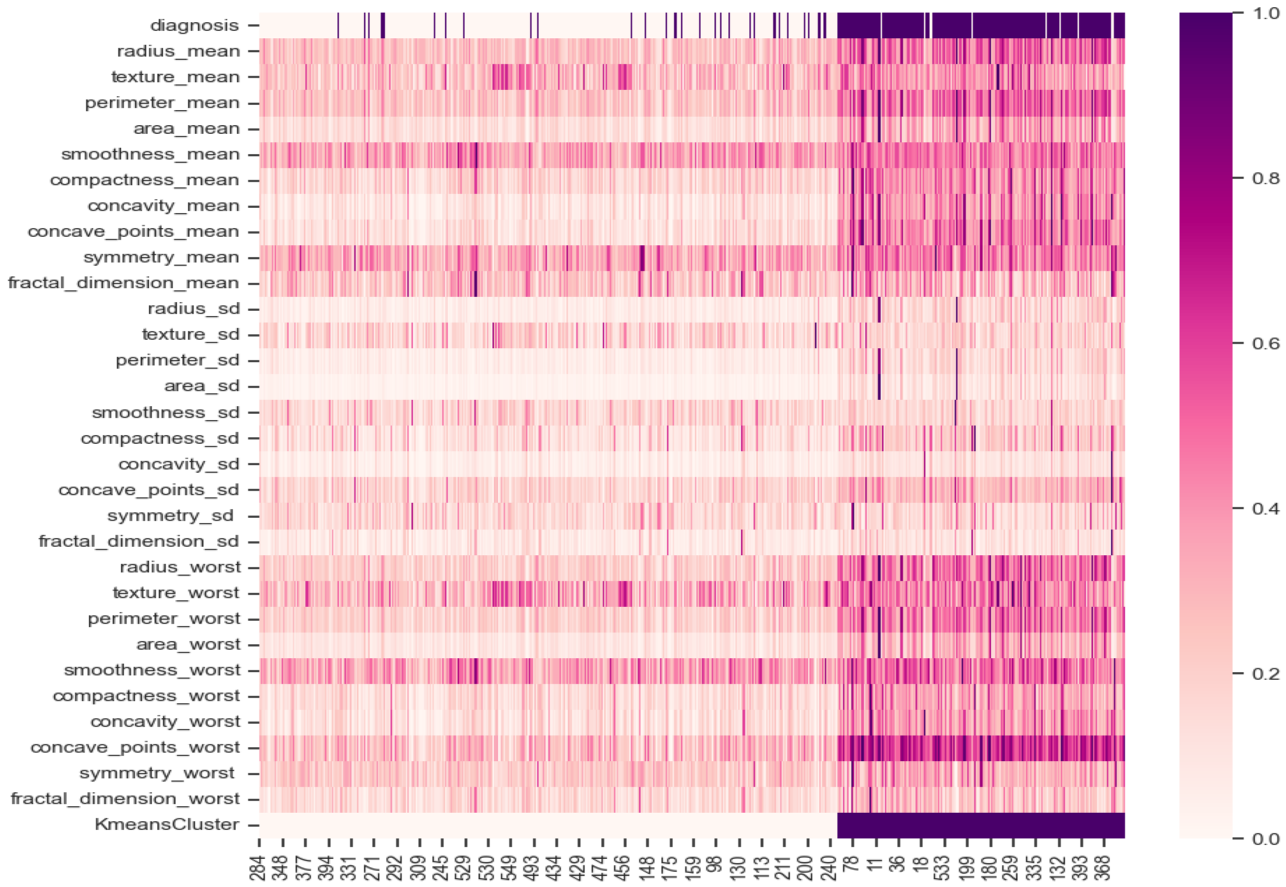


- Based on the graph you provided, there are no apparent outliers in the feature "area_mean".
- The distribution is relatively smooth and compact, without any extreme values that stand out significantly from the rest of the data.
- With a peak around 40 and an evenly narrow dispersion, the distribution is approximately symmetrical.
- It shows that they are not any important departures from the primary feature among the "area_mean" values, which appear to be similar.
- The analysis is easier to all mistakes based on by high values when there are no outliers. This means that a small number of unique data points are less expected

to have a major effect on statistical measures such as mean, standard deviation, and correlation coefficients.

- The tails of the distribution are not particularly long or heavy, indicating a lack of extreme values that would qualify as outliers. There are no isolated points or clusters that separate themselves visually from the main body of the data.

4) The EDA, what factors seem to be most relevant to predicting breast cancer diagnosis:



Based on the colour scale, it looks like the features with the strongest positive correlations with diagnosis (represented by the red color) are:

- fractal_dimension_mean
- concave_points_mean
- symmetry_mean
- concavity_sd
- fractal_dimension_sd

The gene known as "bare nuclei" is the most strongly increased regulated in cancerous tissue, indicating that it appears here more frequently than it is in healthy tissue. According to this, exposed nuclei might be an accurate indication of cancer.

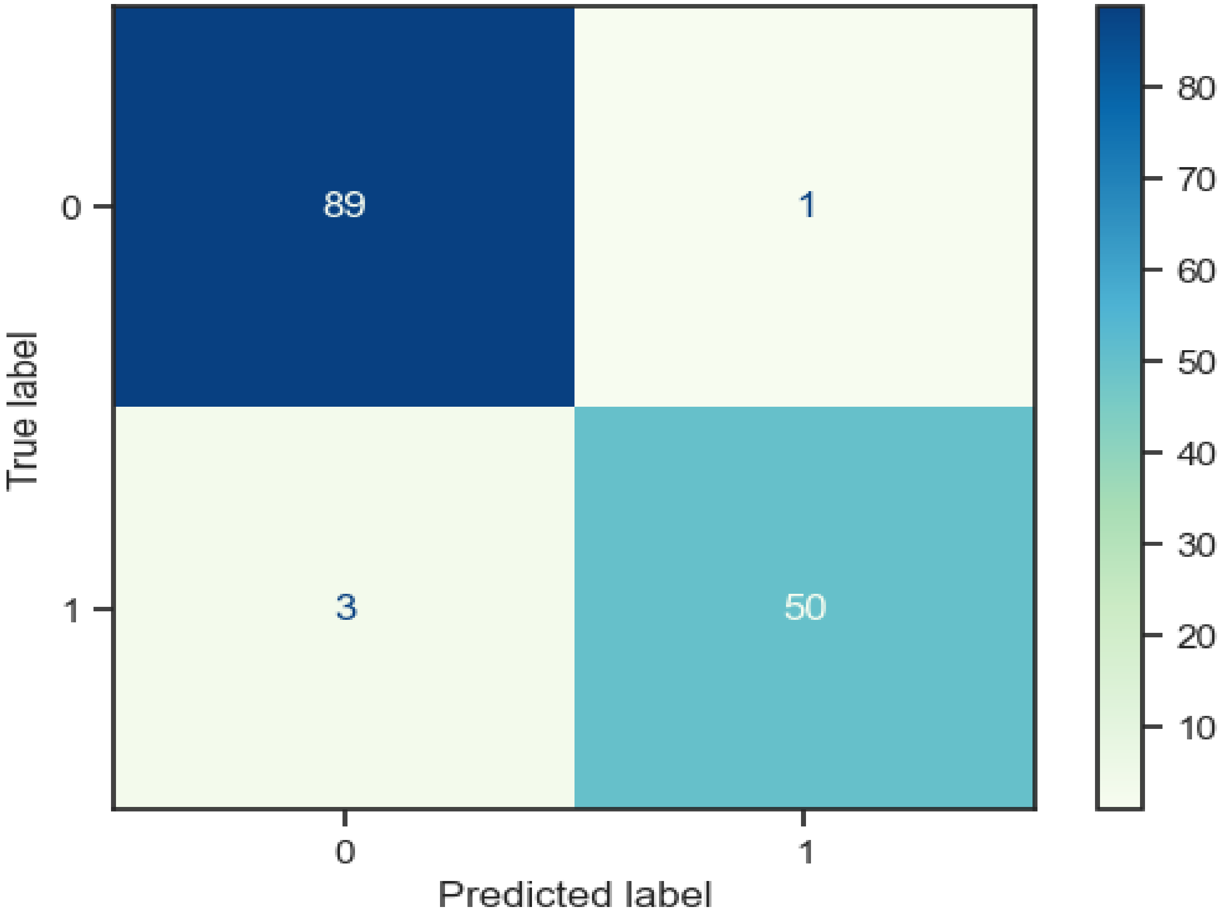
At less level than in bare nuclei, mitosis is another characteristic that is increased regulated in malignant tissue. As cell division occurs during mitosis, it understand that the process would be stronger in spreading malignant tissue.

White blood cells called lymphocytes are involved in fighting against infection. According to the heatmap, lymphocytes are decreased regulated in malignant tissue, which indicates that their concentration is lower there than in healthy tissue. This suggests that lymphocytes might be involved in slowing the spread of cancer.

The connective tissue that covers cells is called the stroma. According to the heatmap, stroma has been increased in malignant tissue, which raises the chance that it causes the spread of cancer.

The cells surrounding the surfaces of glands and organs are known as epithelial cells. According to the heatmap, epithelial cells are decreased regulated in malignant tissue, indicating that there may be a decrease in their frequency in diseased tissue.

5) Limitations:



- Data for just 80 samples are displayed in the confusion matrix. Due to its small sample size, this data might not be entirely representative of the population. This could result in findings that cannot be applied to a larger population.
- True label 1 and True label 0 show an important imbalance, shown by the confusion matrix. Compared to the True label 1 class, the True label 0 class has a much greater number of examples. Due to this class differences, it could be difficult to correctly decide the way the structure does for the minority class (True label 1).
- The confusion matrix only shows the performance of one specific model. It is possible that other models would perform differently on the same data. This means that the the conclusion which i had from this confusion matrix may not be applicable to other models.
- The confusion matrix only displays one particular model's performance; it can happen that other models would perform differently on the same data, meaning that conclusions based on it might not apply to other models.
- The confusion matrix offers no information regarding the data's quality; it is possible that the data is imperfect or contains errors, which could also result in incorrect conclusions.
- If using the model to make predictions on new data, the predictions may be inaccurate due to the class imbalance in the training data.

- If trying to understand the relationships between the features and the target variable, the limited data and lack of context may make it difficult to draw any meaningful conclusions.

- If comparing different models, the confusion matrix does not provide enough information to do so fairly.
- The model may be overfitting the data, that it is performing well on the training data but not generalizing well to new data.
- The model may be biased towards the majority class, meaning that it is more likely to predict negative results, even when the true outcome is positive.
- We can try to mitigate the effects of class imbalance by using techniques such as oversampling or undersampling.
- We can compare the performance of different models on the same data to get a better sense of how generalizable the results are.