|  |  |
| --- | --- |
| A person holding pink ribbons  Description automatically generated  big data project | Abstract  The Breast Cancer Wisconsin (Diagnostic) Data Set contains measurements from fine needle aspirates of breast masses. It was donated to the UCI Machine Learning Repository on October 31, 1995. The dataset is multivariate and consists of 569 instances with 32 features. Each instance corresponds to a breast mass, and the goal is to predict whether the mass is benign (non-cancerous) or malignant (cancerous) based on the provided features. |

**Team Members   
  
Ahmed Tarek 20201701702**

**Adham Khaled 20201701704**

**Abd El Rahman Maged 20201701722**

**Mark Yousri 20201701731**

**Abd El Aziz Wael 20191703020**

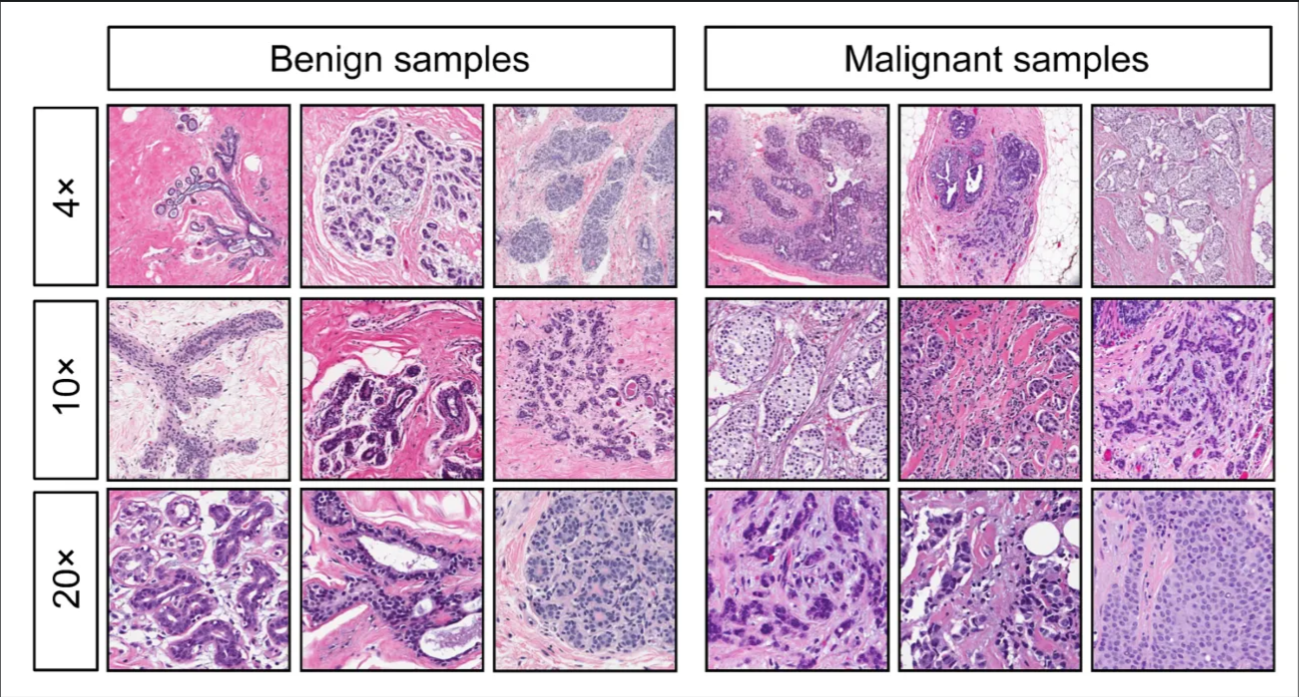
**Ali Rafeeq 20201701705**

Info about the data set

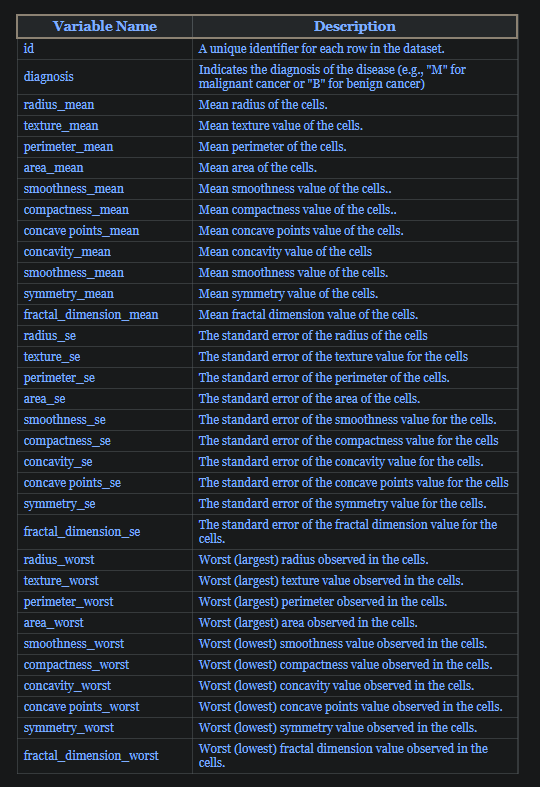
The **Breast Cancer Wisconsin (Diagnostic) Data Set** is a well-known dataset used for breast cancer diagnosis.

Key points about the dataset:

* **Classes**: There are two classes: benign and malignant.
* **Features**: The 32 features include various measurements such as radius, texture, smoothness, compactness, concavity, symmetry, and fractal dimension.
* **Task**: The primary task is **classification**, where the goal is to predict the class label (benign or malignant) based on the feature values.



Data Card



Initial Data Findings

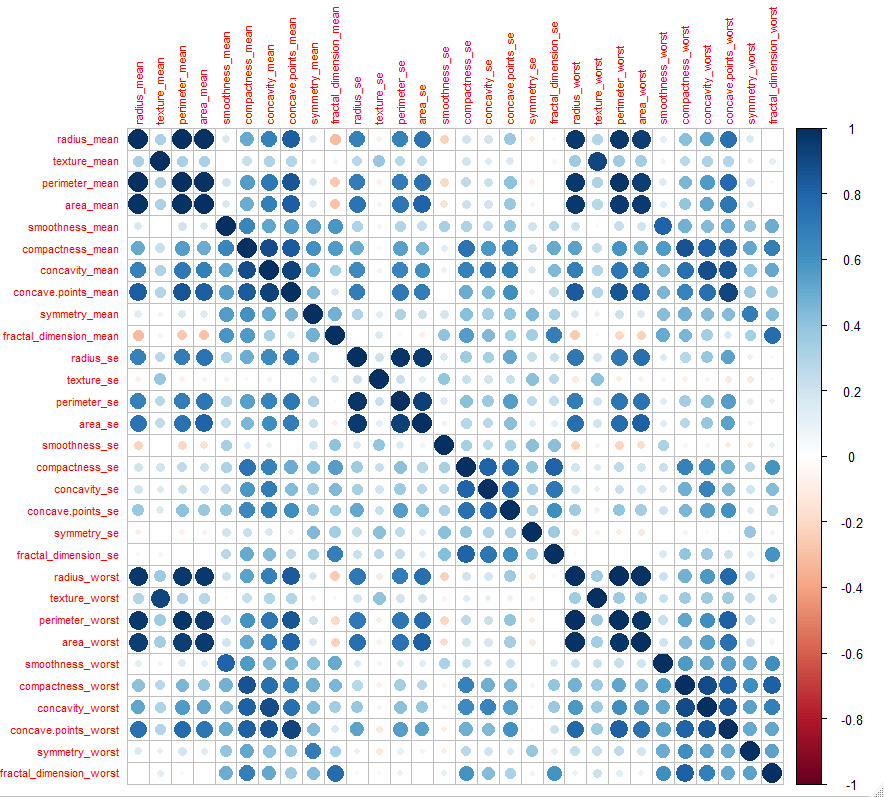
* Zero Null values were Found.
* The number of Benign to Malignant is 357 to 212 respectively 62% to 37.3%.

A blue and red pie chart

Description automatically generated

**Correlation Matrix:**

Showing Variance in correlations of continuous columns [3-32] with mostly all values above 0.5 and very few negative correlations

**Comparing** radius Mean column with respect to diagnosis column:

A graph of a comparison of a number of objects

Description automatically generated with medium confidence

-We conclude from the shape of the data, when the radius

mean exceed 17, there is a high likeliness that it is malignant.

A graph with yellow squares and black lines

Description automatically generated

- Due to the presence of outliers, it’s difficult to conclude results.

A graph with yellow squares and black dots

Description automatically generated

- We conclude from the shape of the data, when the Area

mean exceed 1000, there is a high likeliness that it is malignant.

Histograms

A screenshot of a graph

Description automatically generated

**radius\_mean Columns:**

For the radius\_mean column, positive skewness indicates that most radii are larger than the mean 10-15, meaning that the data is skewed toward numbers larger than the mean. This may be because cancer cells have the greatest ability to grow and spread in larger dimensions, so their radius is also larger than average.

**texture\_mean Columns:**

 The histogram shows that the most common range for ‘texture\_mean’ is around 15-20, with the frequency of occurrences tapering off for lower and higher values. This suggests that the majority of the cell nuclei have a texture mean value within this range.Regarding skewness, the histogram suggests a slight right skewness, meaning there are more observations with ‘texture\_mean’ values higher than the average compared to those with lower values. Skewness is a measure of the asymmetry of the probability distribution of a real-valued random variable about its mean. In this case, a right skew indicates that the tail on the right side of the distribution is longer or fatter than the left side.

**perimeter\_mean Columns:**

The histogram in the image you provided is titled “Histogram of perimeter\_mean,” which indicates the distribution of the ‘perimeter\_mean’ variable from the Wisconsin Breast Cancer dataset. The ‘perimeter\_mean’ represents the average size of the nucleus perimeter in the cell images.The histogram shows a right-skewed distribution, with the majority of the data points falling between 50 and 100 on the x-axis. This suggests that most of the cell nuclei have a perimeter mean value within this range, with fewer nuclei having very small or very large perimeter mean values.The right skewness of the histogram indicates that there are more observations with ‘perimeter\_mean’ values higher than the average, which could suggest a presence of outliers or a long tail of higher values.

**area\_mean Columns:**

From the histogram, we can observe that the majority of the data points are concentrated in the range of approximately 250 to 750 on the x-axis. This indicates that most of the cell nuclei have an area mean value within this interval. The bars decrease in height as the value on the x-axis increases, which suggests that there are fewer occurrences of higher area mean values and thus, a lower frequency of larger cell nuclei areas.The distribution of the histogram appears to be right-skewed, with a longer tail extending towards the higher values of ‘area\_mean’. This skewness implies that while the majority of the cell nuclei are of a moderate size, there are a number of nuclei that are significantly larger, which could be indicative of abnormal growth patterns associated with malignant tumors.

**smoothness\_mean Columns:**

From the histogram, it’s evident that the data points are spread across the range from 0.00 to 0.15 on the x-axis, with varying frequencies as indicated by the heights of the bars. The y-axis, ranging from 0 to 100, likely represents the count of observations within each bin.The distribution of the histogram appears to be somewhat symmetrical, suggesting that the ‘smoothness\_mean’ values are evenly distributed around the mean. This implies that there is a balance in the dataset regarding the smoothness of the cell nuclei, with no significant skewness toward higher or lower values.

**compactness\_mean Columns:**

From the histogram, we can see that the ‘compactness\_mean’ values range from 0.00 to 0.30. The majority of the data points are concentrated in the lower end of the range, with the frequency of occurrences decreasing as the ‘compactness\_mean’ value increases. This suggests that most of the cell nuclei in the dataset are relatively less compact.The distribution appears to be right-skewed, with a longer tail extending towards the higher ‘compactness\_mean’ values. This skewness indicates that there are fewer instances of highly compact cell nuclei, which could be characteristic of certain types of breast cancer tumors.

**concavity\_mean Columns:**

From the histogram, it’s clear that the majority of the data points fall within the lower range of ‘concavity\_mean’ values, specifically between 0.0 and 0.1. The frequency of occurrences decreases as the ‘concavity\_mean’ value increases, which is indicated by the decreasing height of the bars from left to right.The distribution appears to be right-skewed, with a longer tail extending towards the higher values of ‘concavity\_mean’. This skewness suggests that while most cell nuclei have lower concavity, there are a few with significantly higher concavity, which could be indicative of more aggressive tumor characteristics.

**concave points\_mean Columns:**

From the histogram, we can observe that the majority of the data points are concentrated around the lower end of the range, specifically between 0.00 to 0.05. The frequency of occurrences decreases as the ‘concave.points\_mean’ value increases, which is indicated by the decreasing height of the bars towards the value of 0.20.The distribution appears to be right-skewed, suggesting that there are more observations with lower ‘concave.points\_mean’ values and fewer observations with higher values. This skewness could be indicative of the presence of a smaller number of cell nuclei with more pronounced concave points, which may be characteristic of certain types of breast cancer tumors.

**symmetry\_mean Columns:**

From the graph ,the x-axis represents the ‘symmetry\_mean’ values ranging from 0.00 to 0.30, and the y-axis represents the frequency of these values.it appears that the distribution of ‘symmetry\_mean’ is slightly skewed to the right as the mean of the values is in range 16-20. This indicates that there are a small number of cases with higher ‘symmetry\_mean’ values. In other words, most of the data points have ‘symmetry\_mean’ values that are lower than the average, but there are a few cases with significantly higher values.

**fractal\_dimension\_mean Columns:**

From the histogram, it appears that the distribution of ‘fractal\_dimension\_mean’ is unimodal, with the highest frequency between 0.05 and 0.06. This suggests that most of the data points have ‘fractal\_dimension\_mean’ values around this range.the distribution also appears to be positively skewed, meaning that the tail on the right side of the distribution is longer or fatter than the left side. This indicates that there are a small number of cases with higher ‘fractal\_dimension\_mean’ values.

**texture\_se Columns:**

From the histogram, it appears that the distribution of ‘texture\_se’ is unimodal, with the highest frequency just under the value of 1.0 on the x-axis. This suggests that most of the data points have values around this range.the distribution also appears to be positively skewed. This indicates that there are a small number of cases with higher values.

**perimeter\_se Columns:**

From the histogram, it appears that the distribution of ‘perimeter\_se’ is unimodal, with the highest frequency just under the value of 1.0 on the x-axis. This suggests that most of the data points have ‘perimeter\_se’ values around this range. the distribution also appears to be positively skewed, meaning that the tail on the right side of the distribution is longer or fatter than the left ,The skewness in the ‘perimeter\_se’ distribution could indicate that there are a few instances where the perimeter measurements vary more than usual. These could be outliers or could indicate more aggressive or different types of cancer.

**Radius\_se columns:**

From the histogram, it appears that the distribution of ‘radius\_se’ is unimodal, with the highest frequency just under the value of 1.0 on the x-axis. This suggests that most of the data points have ‘radius\_se’ values around this range. the distribution also appears to be positively skewed, meaning that the tail on the right side of the distribution is longer or fatter than the left side. This indicates that there are a small number of cases with higher ‘radius\_se’ values.

**area\_se Columns:**

From the histogram, it appears that the distribution of ‘area\_se’ is unimodal, with the highest frequency just under the value of 1.0 on the x-axis. This suggests that most of the data points have ‘area\_se’ values around this range. the distribution also appears to be positively skewed, meaning that the tail on the right side of the distribution is longer or fatter than the left side. This indicates that there are a small number of cases with higher ‘area\_se’ values. The skewness could indicate that there are a few instances where the tumor sizes vary more than usual.

**smoothness\_se Columns:**

From the histogram, it appears that the distribution of ‘smoothness\_se’ is unimodal, with the highest frequency just under the value of 0.010 on the x-axis. This suggests that most of the data points have ‘smoothness\_se’ values around this range.the distribution also appears to be positively skewed.

**compactness\_se Columns:**

From the histogram, it appears that the distribution of ‘compactness\_se’ is unimodal, with the highest frequency just under the value of 0.010 on the x-axis. This suggests that most of the data points have ‘compactness\_se’ values around this range.the distribution also appears to be positively skewed.

**concavity\_se Columns:**

From the histogram, it appears that the distribution of ‘concavity\_se’ is unimodal, with the highest frequency just under the value of 0.010 on the x-axis. This suggests that most of the data points have ‘concavity\_se’ values around this range.The distribution also appears to be positively skewed.

**concave points\_se Columns:**

The histogram of ‘concave\_points\_se’ values shows a right-skewed distribution. This is indicated by the fact that the majority of the data points are concentrated on the left side of the graph (towards the lower ‘concave\_points\_se’ values), with fewer data points as you move to the right (towards the higher ‘concave\_points\_se’ values).The histogram suggests that most of the cell nuclei have a small standard error for the number of concave points. This could potentially indicate that the variation in the number of concave points across different cell nuclei images is generally low. However, the presence of some data points with higher ‘concave\_points\_se’ values could indicate the presence of images with a higher variation in the number of concave points.

**symmetry\_se Columns:**

The histogram shows a right-skewed distribution. This is indicated by the longer tail on the right side of the graph and most of the data concentrated on the left side. In other words, while most samples have lower ‘symmetry\_se’ values, there are a few samples with relatively high ‘symmetry\_se’ values.The skewness is positive or right-skewed, meaning the tail on the right side of the distribution is longer or fatter. The mean is typically greater than the median in a positively skewed distribution, which is likely the case here.this could indicate that if  measurement high, it could potentially indicate malignancy.

**fractal\_dimension\_se Columns:**

The histogram shows a right-skewed distribution. This is indicated by the longer tail on the right side of the graph and most of the data concentrated on the left side. In other words, while most samples have lower ‘fractal\_dimension\_se’ values, there are a few samples with relatively high ‘fractal\_dimension\_se’ values.The skewness is positive or right-skewed, meaning the tail on the right side of the distribution is longer or fatter. The mean is typically greater than the median in a positively skewed distribution, which is likely the case here.

**radius\_worst Columns:**

Regarding skewness, the histogram does show a right skewness. while most data points are gathered on the left side. In other words, while most tumors have a smaller ‘worst radius,’ there is a significant number that have larger ‘worst radii.’

**texture\_worst Columns:**

his histogram also shows a right skewness. This is evident from the long tail extending towards higher values on the right side, while most data points are gathered on the left side. In other words, while most tumors have a smaller ‘worst texture,’ there is a significant number that have larger ‘worst texture.’ This skewness could be relevant for medical diagnosis or treatment decisions as it reflects on the variability and distribution of tumor textures within this dataset

**perimeter\_worst Columns:**

This histogram shows a right skewness. This is evident from the long tail extending towards higher values on the right side, while most data points are gathered on the left side. In other words, while most tumors have a smaller ‘worst perimeter,’ there is a significant number that have larger ‘worst perimeter.’ This skewness could be relevant for medical diagnosis or treatment decisions as it reflects on the variability and distribution of tumor perimeters within this dataset.

**area\_worst Columns:**

This histogram shows a strong right skewness. This is evident from the long tail extending towards higher values on the right side, while most data points are gathered on the left side. In other words, while most tumors have a smaller ‘worst area,’ there is a significant number that have larger ‘worst area.’ This skewness could be relevant for medical diagnosis or treatment decisions as it reflects on the variability and distribution of tumor areas within this dataset.

**smoothness\_worst Columns:**

This histogram shows a slight right skewness. This is evident from the long tail extending towards higher values on the right side, while most data points are gathered on the left side. In other words, while most tumors have a smaller ‘worst smoothness,’ there is a significant number that have larger ‘worst smoothness.’ This skewness could be relevant for medical diagnosis or treatment decisions as it reflects on the variability and distribution of tumor smoothness within this dataset.

**compactness\_worst Columns:**

This histogram shows a strong right skewness. This is evident from the long tail extending towards higher values on the right side, while most data points are gathered on the left side. In other words, while most tumors have a smaller ‘worst compactness,’ there is a significant number that have larger ‘worst compactness.’ This skewness could be relevant for medical diagnosis or treatment decisions as it reflects on the variability and distribution of tumor compactness within this dataset.

**concavity\_worst Columns:**

This histogram shows a strong right skewness. This is evident from the long tail extending towards higher values on the right side, while most data points are gathered on the left side. In other words, while most tumors have a smaller ‘worst concavity,’ there is a significant number that have larger ‘worst concavity.’ This skewness could be relevant for medical diagnosis or treatment decisions as it reflects on the variability and distribution of tumor concavities within this dataset.

**concave points\_worst Columns:**

This histogram shows a slight right skewness. This is evident from the long tail extending towards higher values on the right side, while most data points are gathered on the left side. In other words, while most tumors have a smaller ‘worst concave points,’ there is a significant number that have larger ‘worst concave points.’ This skewness could be relevant for medical diagnosis or treatment decisions as it reflects on the variability and distribution of tumor concave points within this dataset.

**symmetry\_worst Columns:**

In the next column, `symmetry\_worst`, positive skewness indicates that the distribution of worst symmetry values is skewed towards higher values. This suggests that there may be more observations with higher symmetry in certain cases compared to the mean, which could potentially indicate certain characteristics of the cells being observed.

**fractal\_dimension\_worst Columns:**

This histogram shows a strong right skewness. This is evident from the long tail extending towards higher values on the right side, while most data points are gathered on the left side. In other words, while most tumors have a smaller ‘worst fractal dimension,’ there is a significant number that have larger ‘worst fractal dimension.’ This skewness could be relevant for medical diagnosis or treatment decisions as it reflects on the variability and distribution of tumor fractal dimensions within this dataset.

Boxplots

A comparison of a number and a number

Description automatically generated with medium confidenceA group of graphs with text

Description automatically generated with medium confidence

If the "radius\_mean" column contains outliers and still has positive skewness, this may mean that the outliers do not significantly affect the overall data distribution and the distribution remains the same as the original distribution.

If the "texture\_mean" column contains outliers and still has a slight positive skewness, this may mean that the outliers do not significantly affect the overall distribution of the data and the distribution remains the same as the original distribution.

If the "perimeter\_mean" column contains outliers and still has positive skewness, this may mean that the outliers do not significantly affect the overall distribution of the data and the distribution remains the same as the original distribution.

If the "area\_mean" column contains outliers and still has positive skewness, this may mean that the outliers do not significantly affect the overall distribution of the data and the distribution remains the same as the original distribution.

If the "smoothness\_mean" column contains outliers and still has a normal distribution, this may mean that the outliers do not significantly affect the overall distribution of the data and the distribution remains the same as the original distribution.

If the "compactness\_mean" column contains outliers and still has positive skewness, this may mean that the outliers do not significantly affect the overall distribution of the data and the distribution remains the same as the original distribution.

If the "concavity\_mean" column contains outliers and still has positive skewness, this may mean that the outliers do not significantly affect the overall distribution of the data and the distribution remains the same as the original distribution.

If the "concave points\_mean" column contains outliers and still has positive skewness, this may mean that the outliers do not significantly affect the overall distribution of the data and the distribution remains the same as the original distribution.

If the "symmetry\_mean" column contains outliers and still has slight positive skewness, this may mean that the outliers do not significantly affect the overall distribution of the data and the distribution remains the same as the original distribution.

If the "fractal\_dimension\_mean" column contains outliers and still has positive skewness, this may mean that the outliers do not significantly affect the overall distribution of the data and the distribution remains the same as the original distribution.

A group of graphs with black and yellow lines

Description automatically generated

A comparison of a graph

Description automatically generated with medium confidence

If the "radius\_se" column contains outliers and still has positive skewness, this may mean that the outliers do not significantly affect the overall data distribution and the distribution remains the same as the original distribution.

If the "texture\_se" column contains outliers and still has positive skewness, this may mean that the outliers do not significantly affect the overall distribution of the data and the distribution remains the same as the original distribution

If the "perimeter\_se" column contains outliers and still has positive skewness, this may mean that the outliers do not significantly affect the overall distribution of the data and the distribution remains the same as the original distribution.

If the "area\_se" column contains outliers and still has positive skewness, this may mean that the outliers do not significantly affect the overall distribution of the data and the distribution remains the same as the original distribution.

If the "smoothness\_se" column contains outliers and still has a positive skewness, this may mean that the outliers do not significantly affect the overall distribution of the data and the distribution remains the same as the original distribution.

If the "compactness\_se" column contains outliers and still has positive skewness, this may mean that the outliers do not significantly affect the overall distribution of the data and the distribution remains the same as the original distribution.

If the "concavity\_se" column contains outliers and still has positive skewness, this may mean that the outliers do not significantly affect the overall distribution of the data and the distribution remains the same as the original distribution.

If the "concave points\_se" column contains outliers and still has positive skewness, this may mean that the outliers do not significantly affect the overall distribution of the data and the distribution remains the same as the original distribution.

If the "symmetry\_se" column contains outliers and still has positive skewness, this may mean that the outliers do not significantly affect the overall distribution of the data and the distribution remains the same as the original distribution.

If the "fractal\_dimension\_se" column contains outliers and still has positive skewness, this may mean that the outliers do not significantly affect the overall distribution of the data and the distribution remains the same as the original distribution.

A comparison of a graph

Description automatically generated with medium confidenceA group of graphs with red lines

Description automatically generated

If the "radius\_worst" column contains outliers and still has positive skewness, this may mean that the outliers do not significantly affect the overall data distribution and the distribution remains the same as the original distribution.

If the "texture\_worst" column contains outliers and still right skewed, this may mean that the outliers do not significantly affect the overall distribution of the data and the distribution remains the same as the original distribution.

If the "perimeter\_worst" column contains outliers and still has positive skewness, this may mean that the outliers do not significantly affect the overall distribution of the data and the distribution remains the same as the original distribution.

If the "area\_worst" column contains outliers and still has strong right skewness, this may mean that the outliers do not significantly affect the overall distribution of the data and the distribution remains the same as the original distribution.

If the "smoothness\_worst" column contains outliers and still has slight positive skewness, this may mean that the outliers do not significantly affect the overall distribution of the data and the distribution remains the same as the original distribution.

If the "compactness\_worst" column contains outliers and still has strong positive skewness, this may mean that the outliers do not significantly affect the overall distribution of the data and the distribution remains the same as the original distribution.

If the "concavity\_worst" column contains outliers and still has a strong positive skewness, this may mean that the outliers do not significantly affect the overall distribution of the data and the distribution remains the same as the original distribution.

If the "symmetry\_worst" column contains outliers and still a positive skewness, this may mean that the outliers do not significantly affect the overall distribution of the data and the distribution remains the same as the original distribution.

If the "fractal\_dimension\_se" column contains outliers and still has strong positive skewness, this may mean that the outliers do not significantly affect the overall distribution of the data and the distribution remains the same as the original distribution.

If the "concave.points\_worst" column contains outliers and still a slight positive skewness, this may mean that the outliers do not significantly affect the overall distribution of the data and the distribution remains the same as the original distribution.

**Anova test (Hypothesis):**

For each subcategory in our dataset w created an ANOVA test

**Mean Columns:**

The F-statistic is 85.09 with a p-value much less than 0.05 (p < 2e-16), indicating that there is a significant difference in means across different diagnoses for the mean columns. This suggests that the mean values of these features differ significantly depending on the diagnosis.

A screenshot of a computer code

Description automatically generated

**Standard Error (SE) Columns:**

The F-statistic is 113.7 with a p-value much less than 0.05 (p < 2e-16), indicating a significant difference in means across different diagnoses for the SE columns.

Similar to the mean columns, this suggests that the SE values of these features also differ significantly depending on the diagnosis.

A screenshot of a computer code

Description automatically generated

**Worst Columns:**

The F-statistic is 117.1 with a p-value much less than 0.05 (p < 2e-16), indicating a significant difference in means across different diagnoses for the worst columns.

Again, like the previous two sets of columns, this suggests that the worst values of these features also differ significantly depending on the diagnosis.

Conclusion, we reject the Null hypothesis in all 3 tests

A computer screen shot of a computer code

Description automatically generated

**Post-hoc Testing (Tukey’s Test)**

**Mean Columns:**

The difference in means between diagnosis 1 and diagnosis 0 is 56.22, with a 95% confidence interval ranging from 44.27 to 68.16.

The p-value is reported as 0, indicating that the difference is statistically significant.

A screenshot of a computer code

Description automatically generated

**Standard Error (SE) Columns:**

The difference in means between diagnosis 1 and diagnosis 0 is 5.42, with a 95% confidence interval ranging from 4.42 to 6.42.

The p-value is reported as 0, indicating that the difference is statistically significant.

A computer code with numbers and symbols

Description automatically generated

**Worst Columns:**

The difference in means between diagnosis 1 and diagnosis 0 is 93.20, with a 95% confidence interval ranging from 76.32 to 110.08.

The p-value is reported as 0, indicating that the difference is statistically significant.

A computer code with blue text

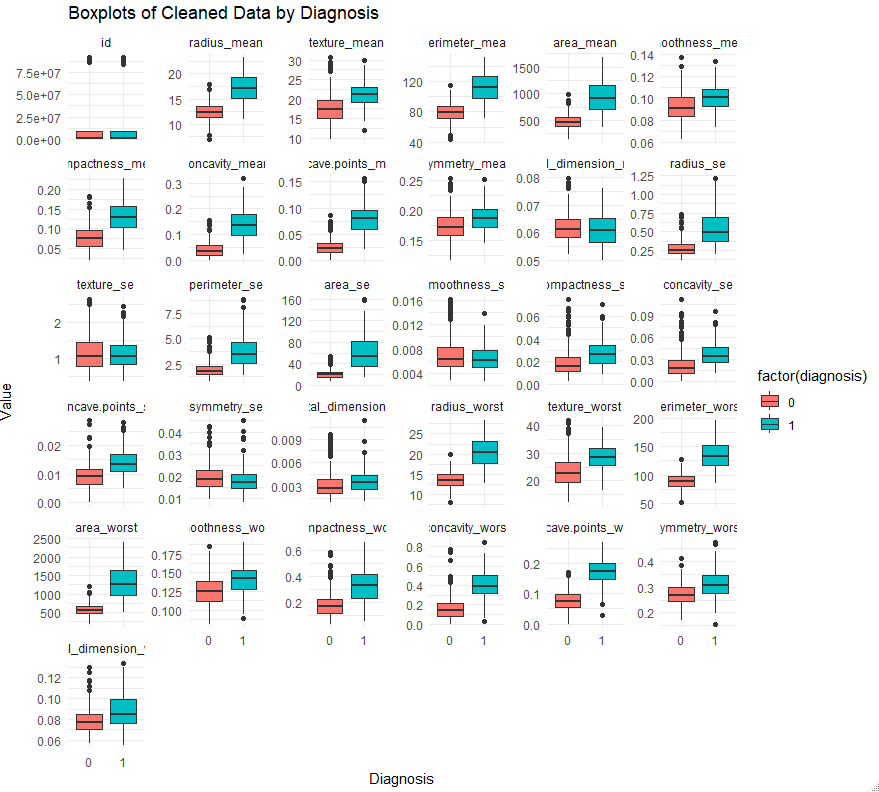
Description automatically generated

**Using Z-scores to remove outliers:**

Shown in the difference of thresholds between the above boxplots and the ones here. e.g the radius\_mean column had max values near 30 (suspected outlies) and the highest value below is nearly 20.

A screenshot of a computer code

Description automatically generated



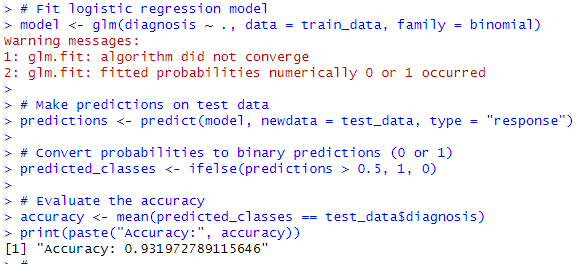
**Shuffling the data then splitting them to train and test data.**

**Splitting the Data into train and test**

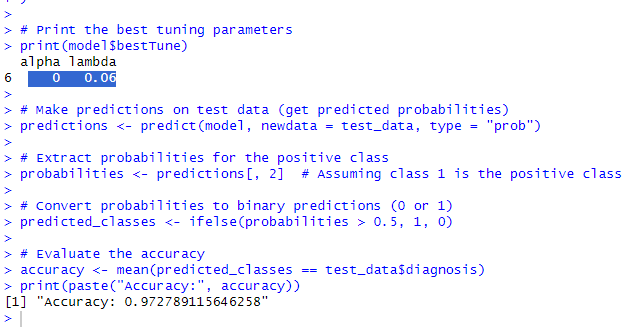
A computer code with blue text

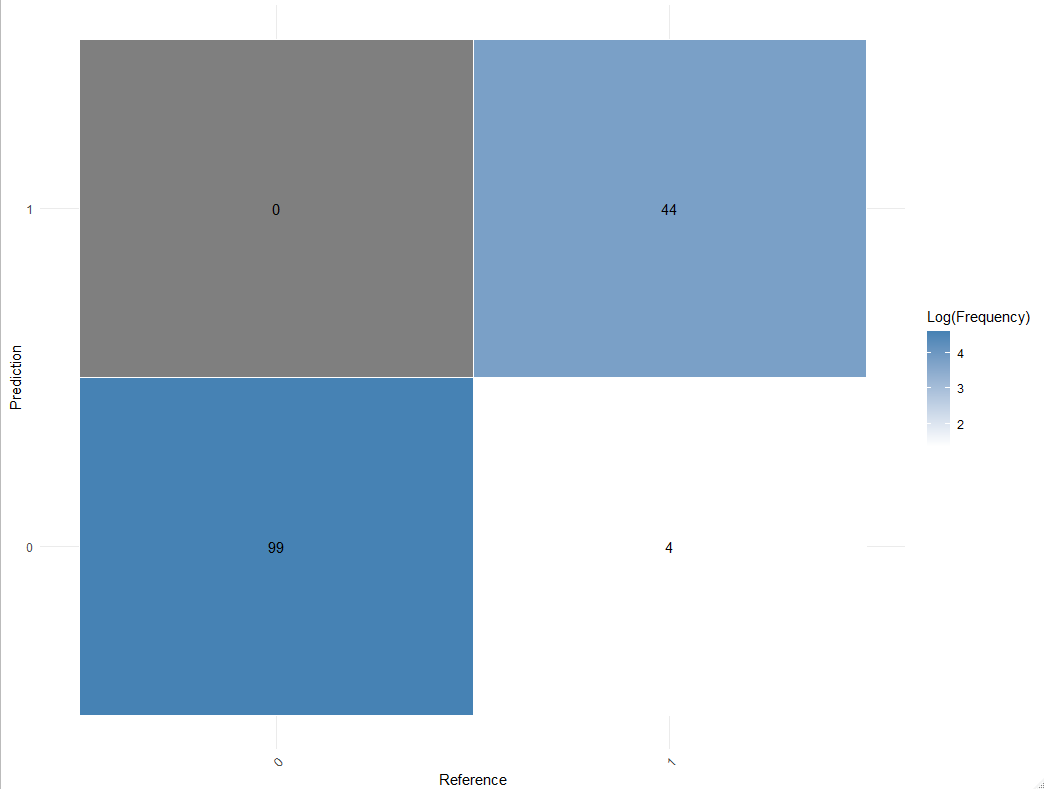
Description automatically generated

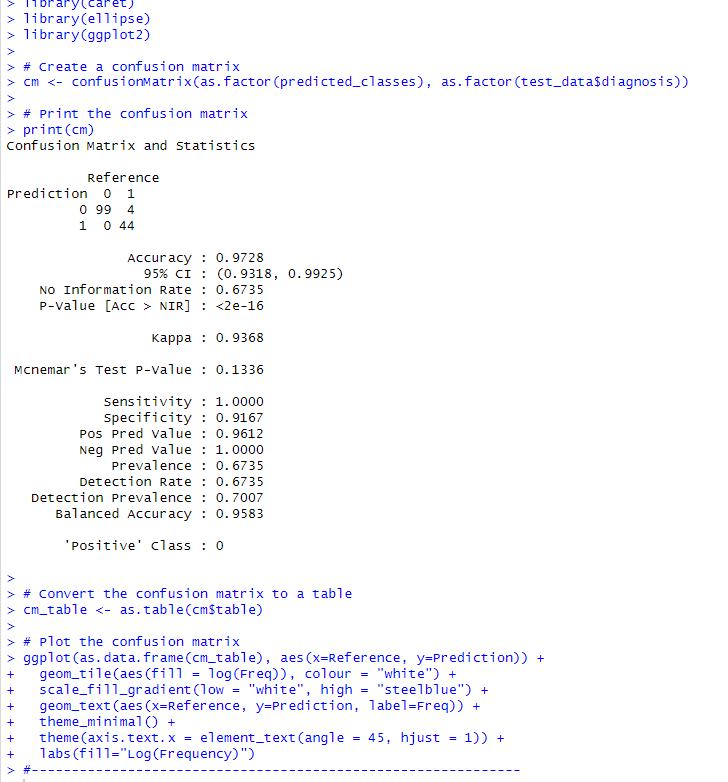
**Applying the plain model on the data:**



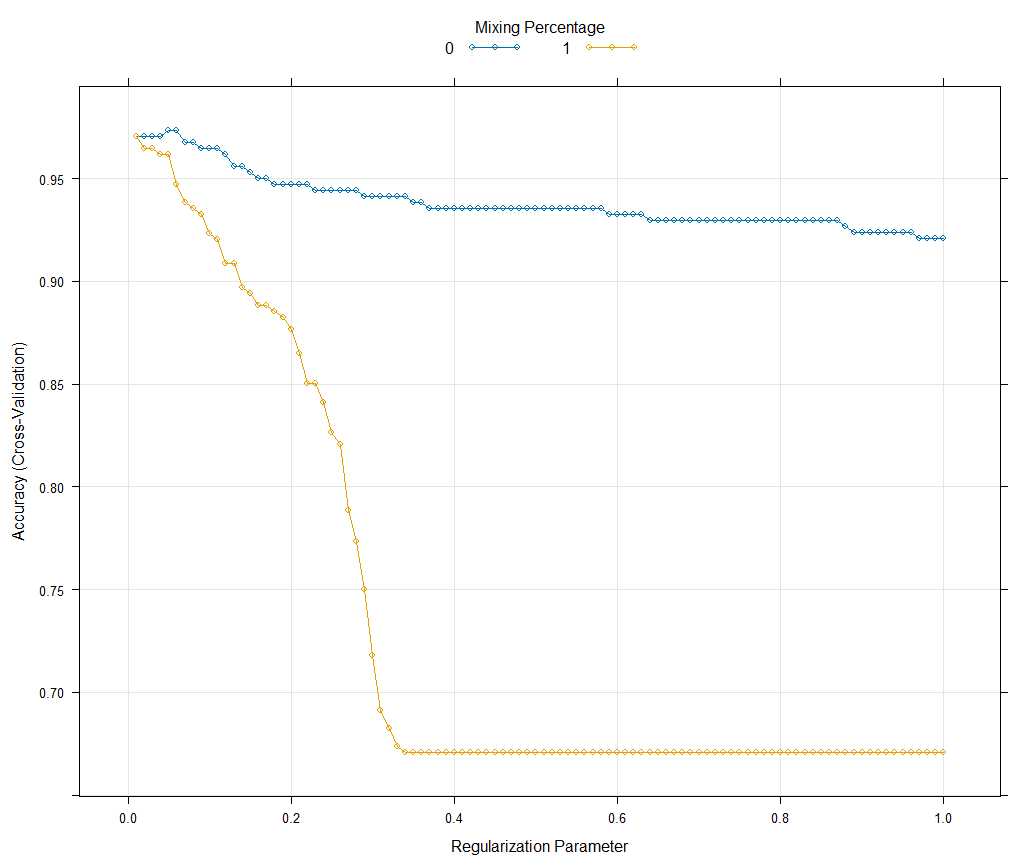
**Using Grid search to optimize the model:**







**Regularization parameters Graph:**

****