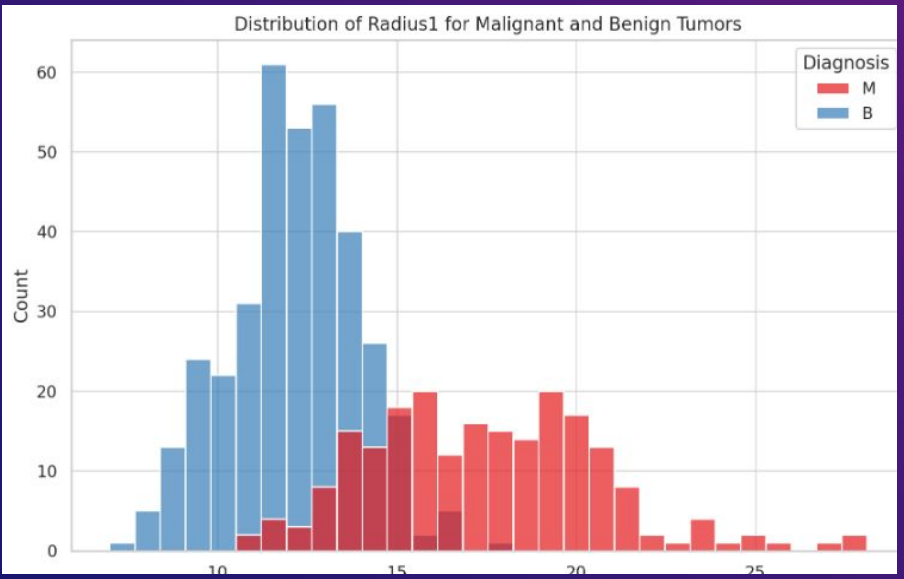
GROUP 1

UCI Breast Cancer

DATA ANALYSIS

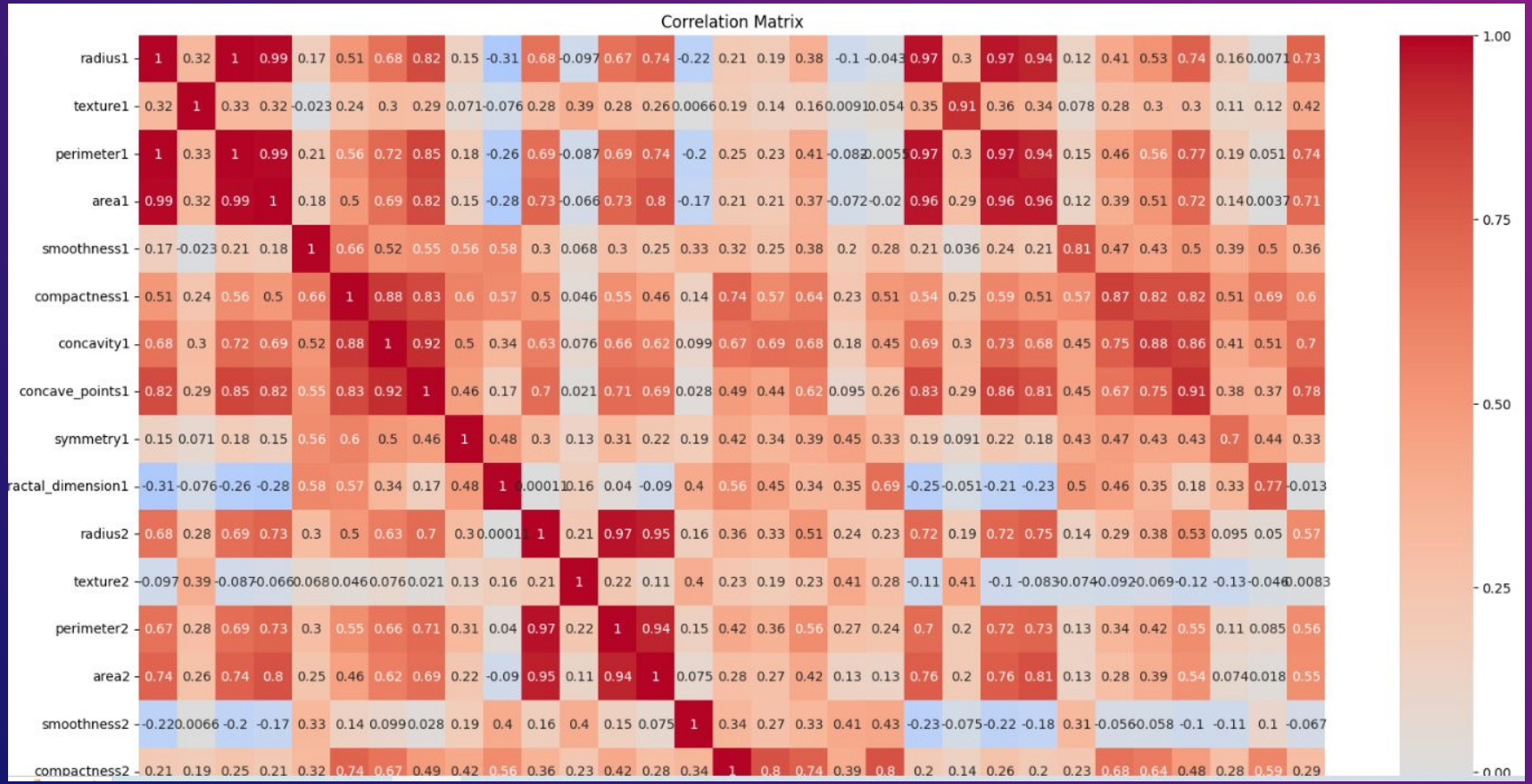
Dhruval Patel-drp164
Adarsh Savani - ass213
Dhruvita Patel-dbp123
Victoria Land-vml58
Kaitlyn Koenig-kmk366

TRENDS IN THE DATA~

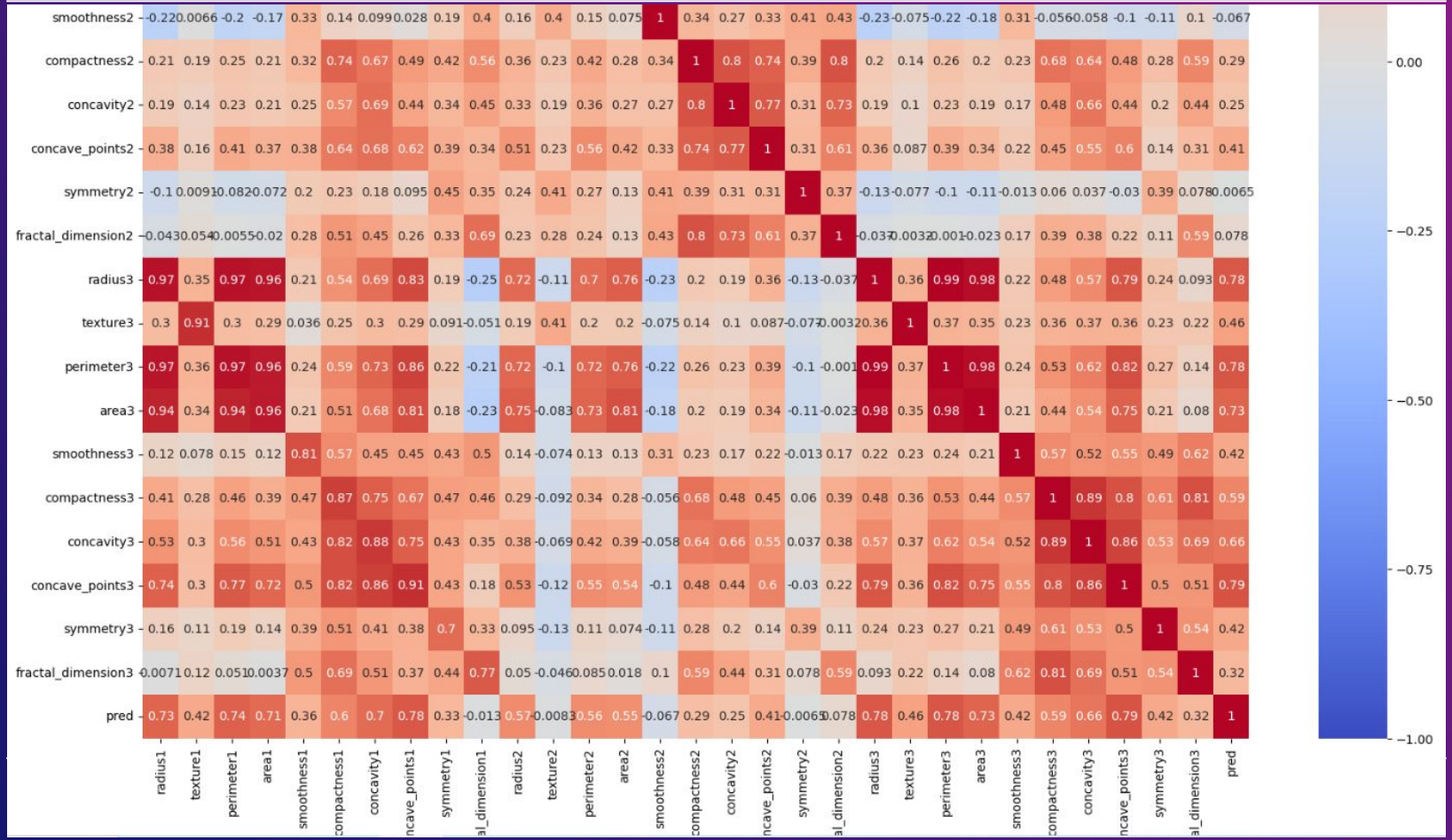


The histogram visualizing the distribution of `radius1` (mean radius of cell nuclei) in breast tissue reveals key insights into the characteristics of benign and malignant tumors. Benign tumors predominantly exhibit smaller mean radii, suggesting the presence of smaller cell nuclei, while malignant tumors tend to display larger mean radii, indicative of larger cell nuclei. This clear differentiation in nuclear size is crucial in medical diagnosis and research, as it aids in distinguishing between benign and malignant breast tissues. Such exploratory data analysis is essential for understanding the underlying patterns and characteristics of breast cancer, highlighting the significance of nuclear features in the context of cancer diagnosis and treatment planning.

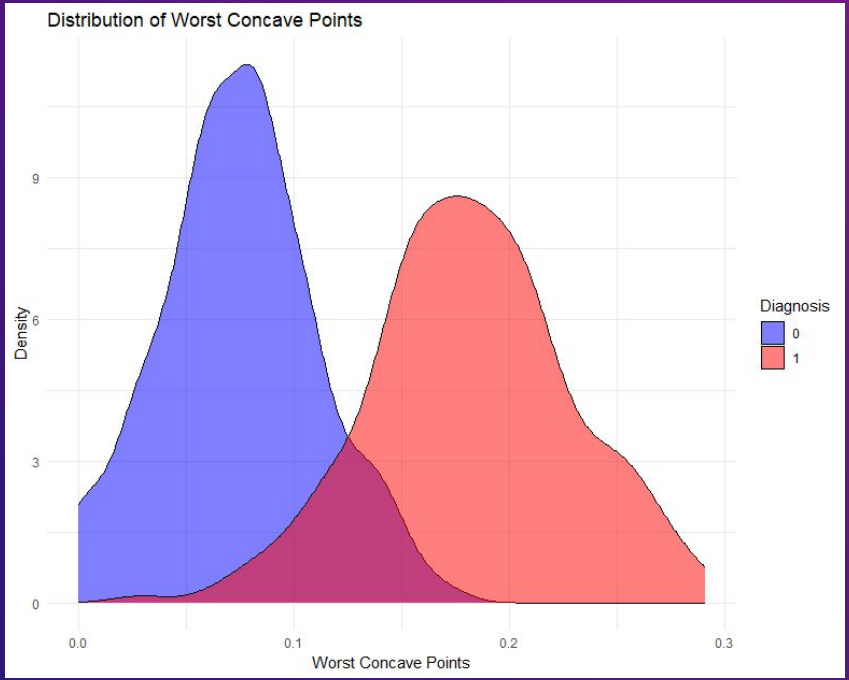
CORRELATION PLOT FOR ALL FEATURES WITH RESPECT TO DIAGNOSIS



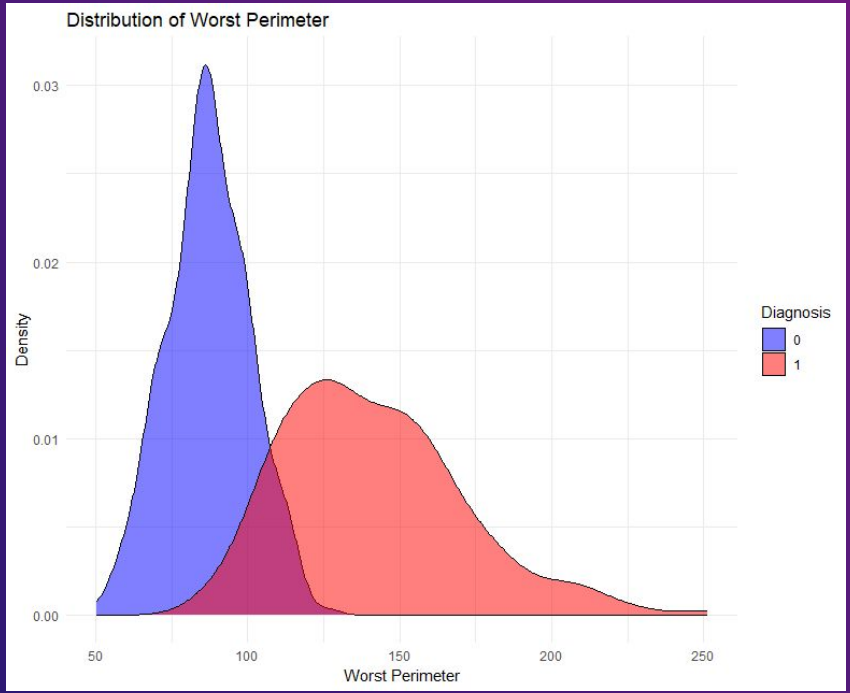
CORRELATION PLOT FOR ALL FEATURES WITH RESPECT TO DIAGNOSIS



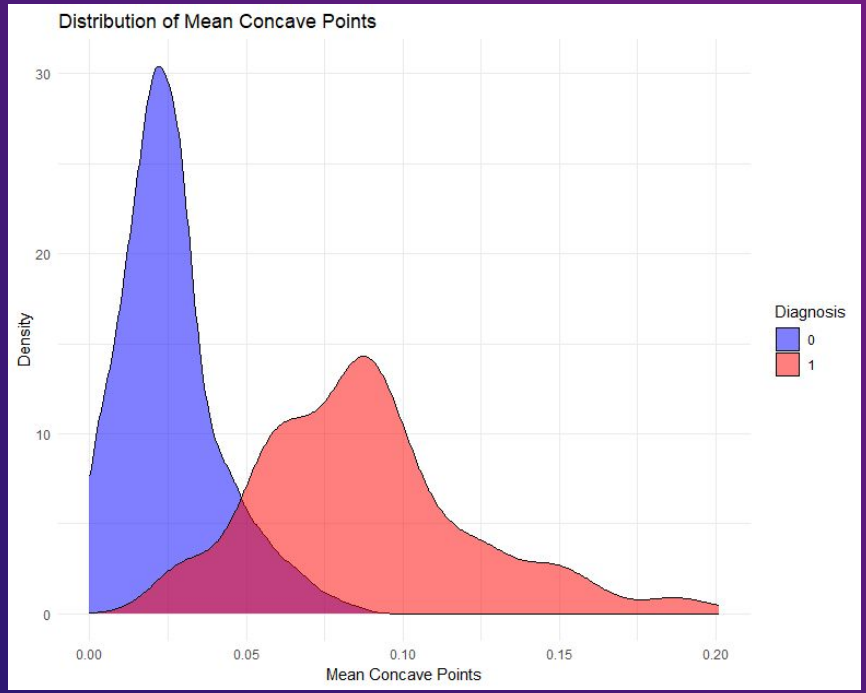
CONCAVE POINTS (WORST)



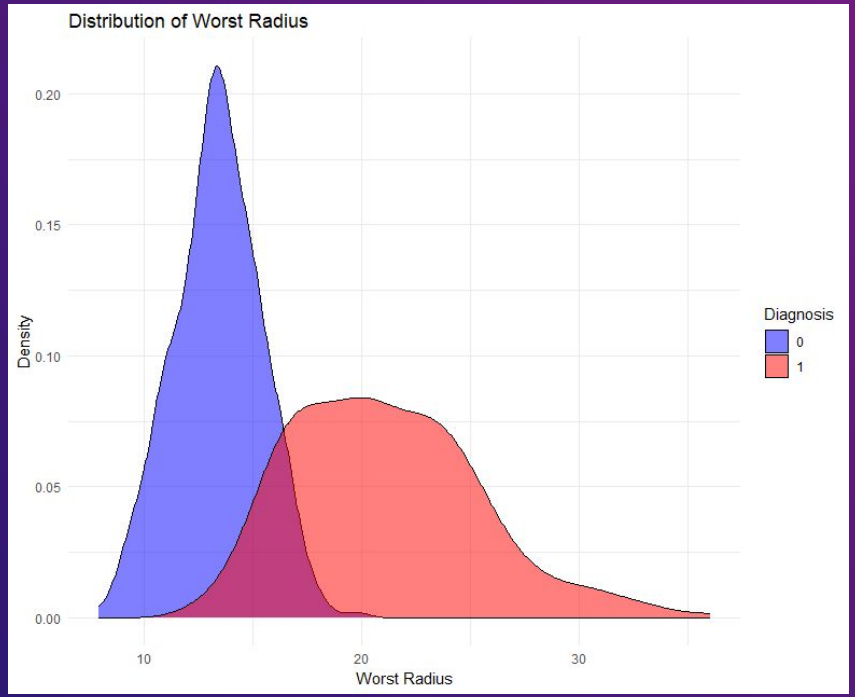
PERIMETER (WORST)



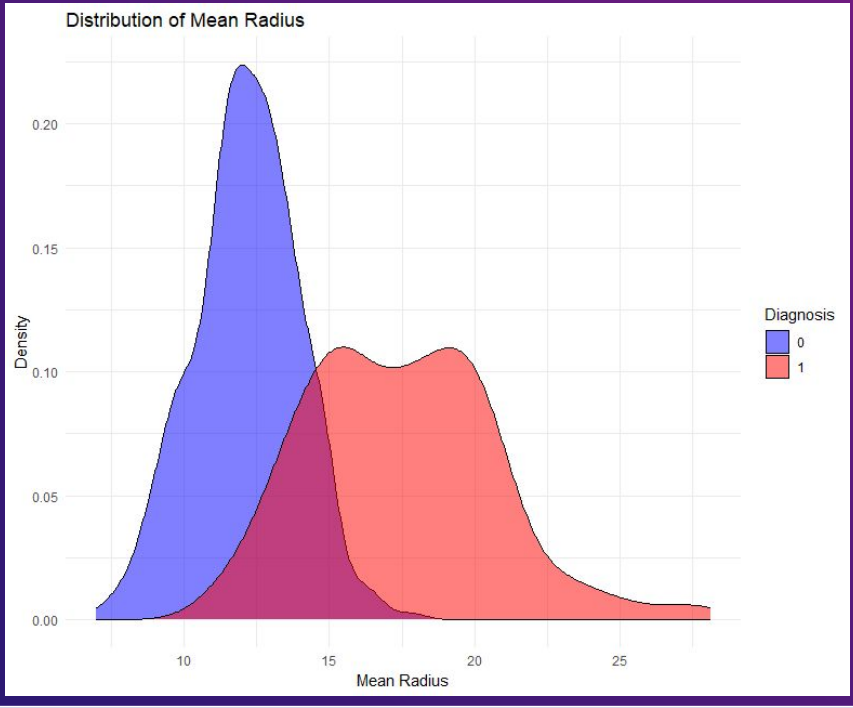
CONCAVE POINTS (MEAN)

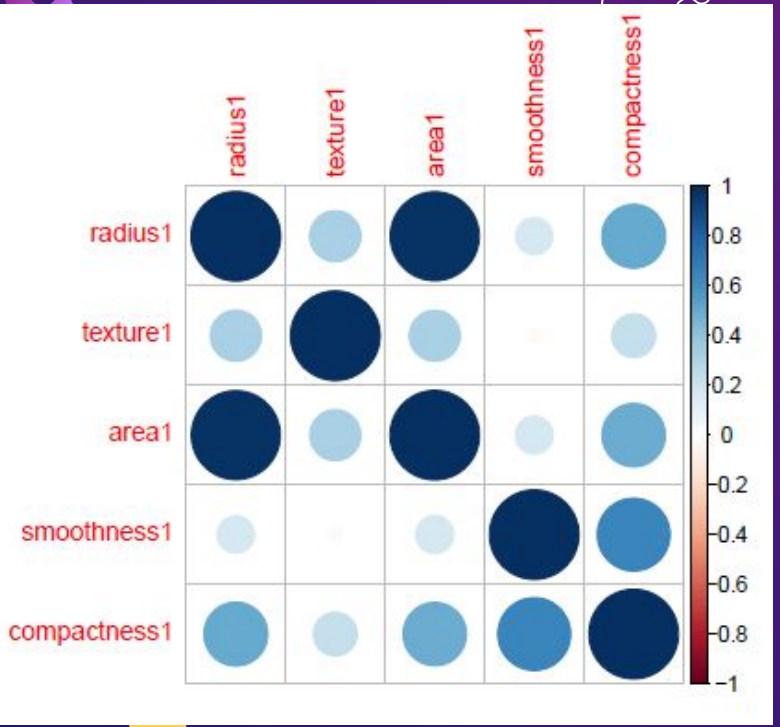


RADIUS (WORST)



RADIUS (MEAN)


Distribution of Mean Radius
Density
0.20
0.15
0.10
0.05
0.00
10
15
20
25
Mean Radius
Diagnosis
0
1



The Correlation plot visualizes the correlation matrix of selected features from the breast cancer dataset. Each cell in the map shows the correlation coefficient between two features, ranging from -1 to 1. Here are the key trends and insights from this plot:There are strong positive correlations between some features. For instance, radius1 and area1 show a high positive correlation. This indicates that as the mean radius of cell nuclei increases, the mean area also tends to increase, which is logical given the geometric relationship between radius and area.

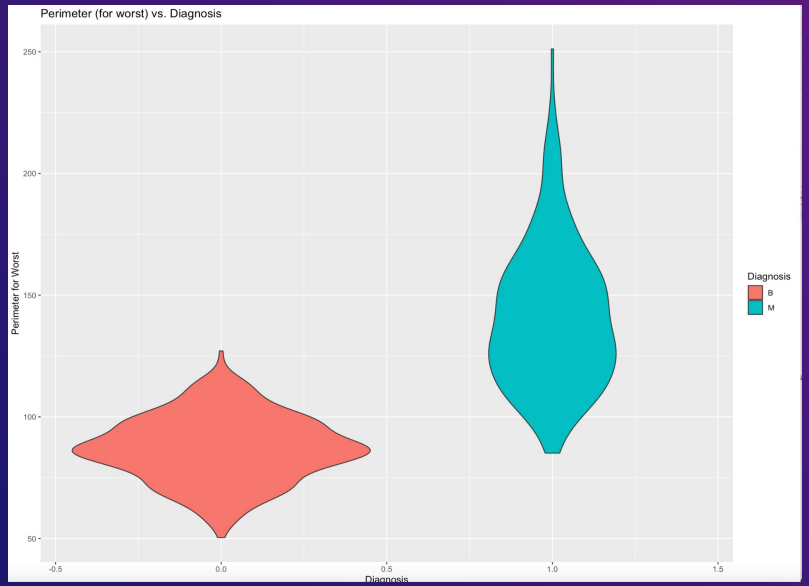
VIOLIN PLOTS (TOP 10 VARIABLES)

Quick info:

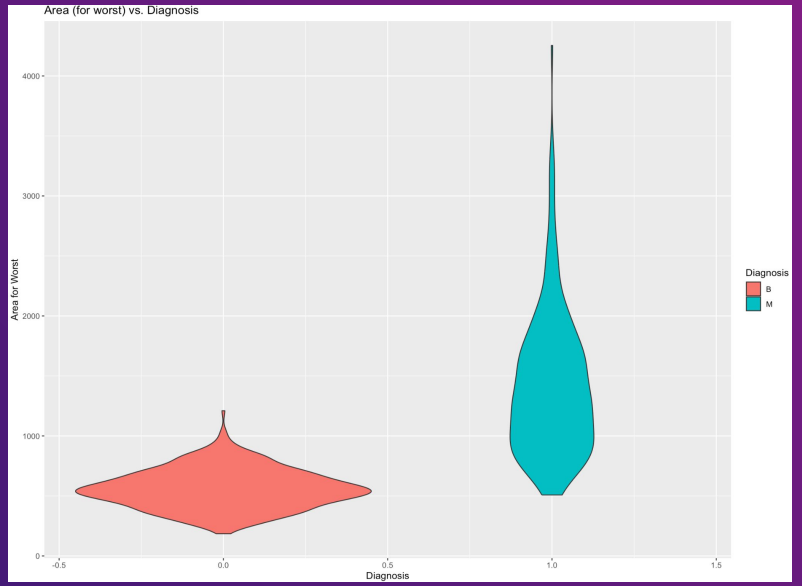
- Width of the Violin:
  - Violin plots use the density of the data as a factor in how it looks. The wider it is, the higher the density of data points.
- Symmetry:
  - The more symmetrical the violin plot is, the more symmetrical the data is.
- Multiple Violins:
  - The multiple violins represent the two groups — Malignant and Benign tumors.

Used for the top 10 variables from the Random Forest Model…

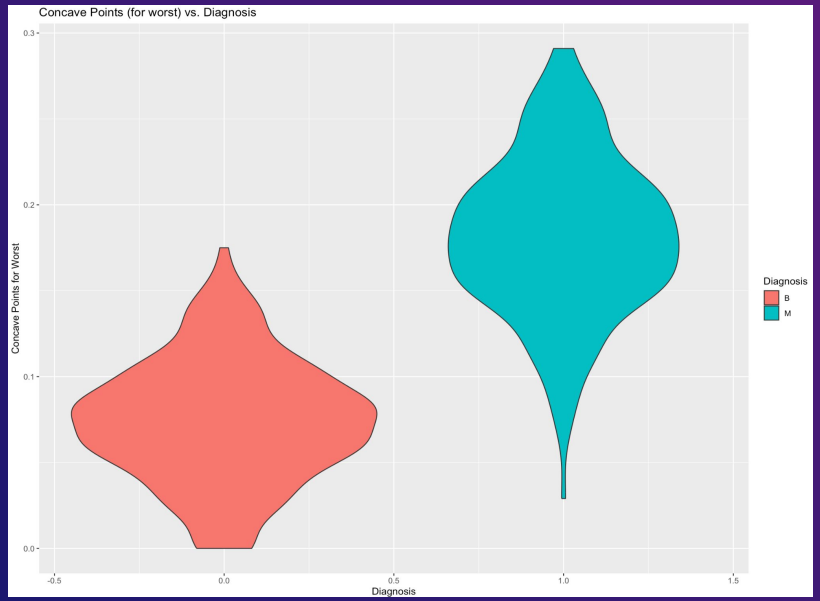
1) Perimeter (Worst)



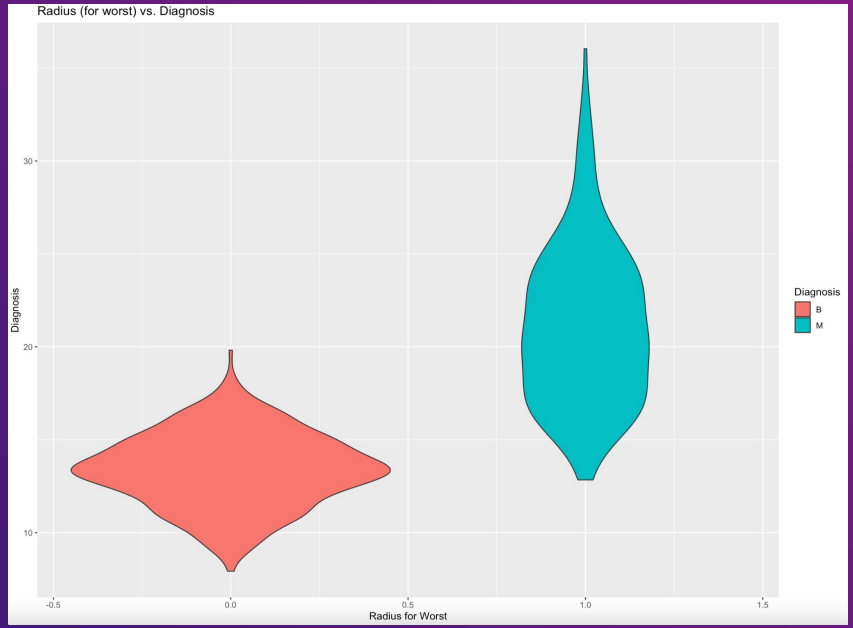
2) Area (Worst)



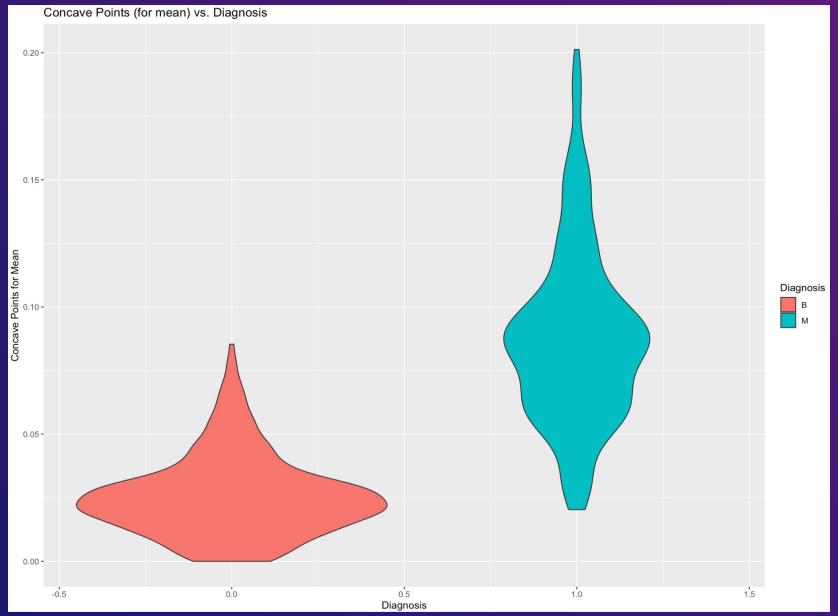
3) Concave Points (Worst)



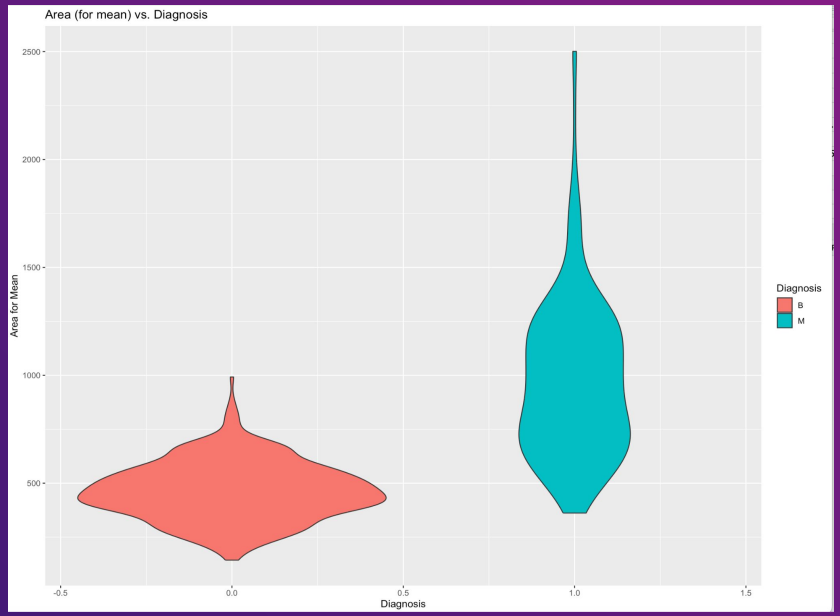
4) Radius (Worst)



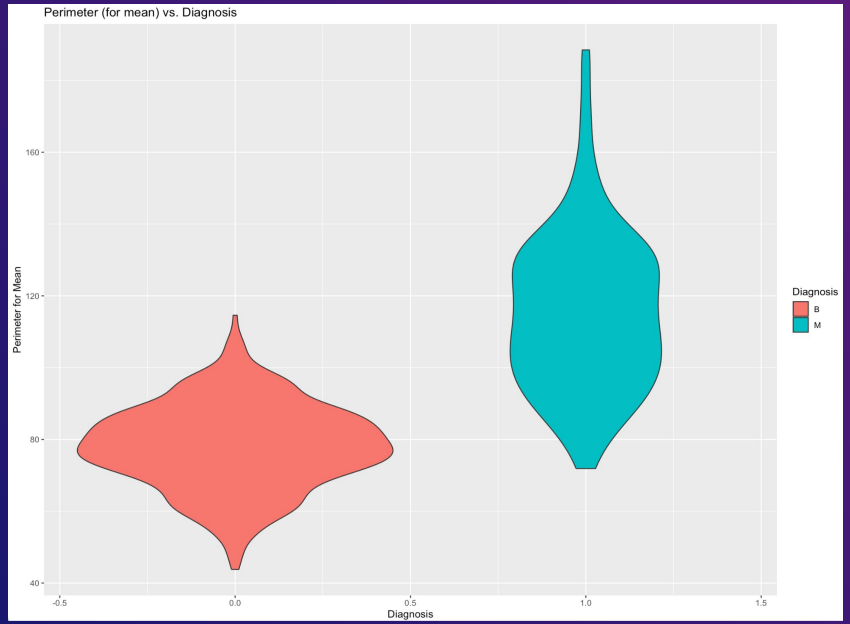
5) Concave Points (Mean)



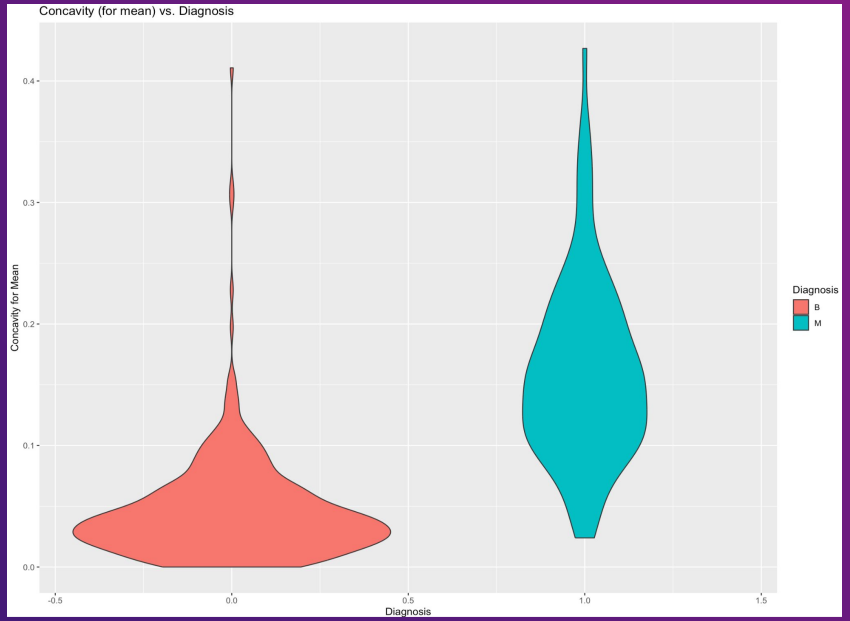
6) Area (Mean)



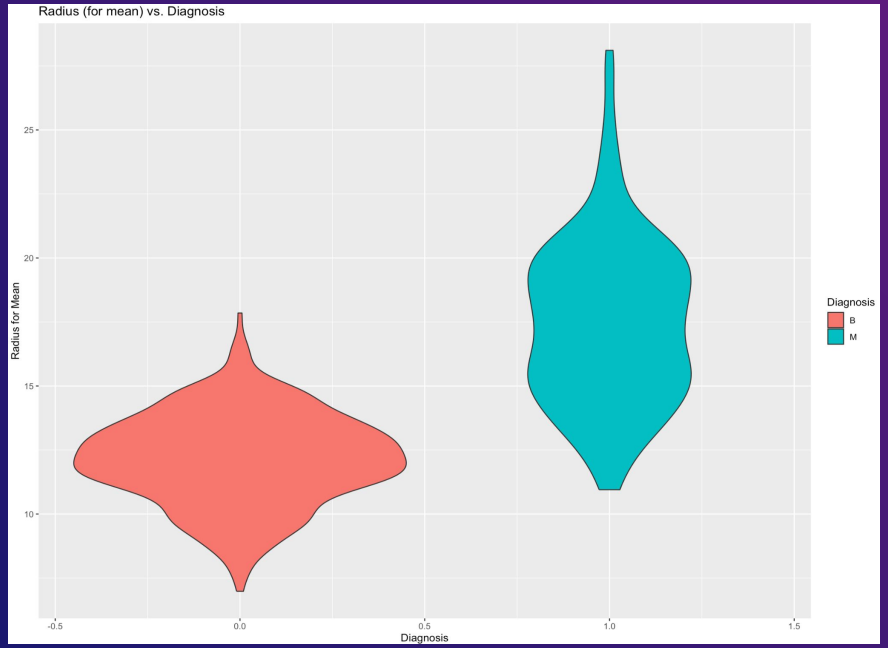
7) Perimeter (Mean)



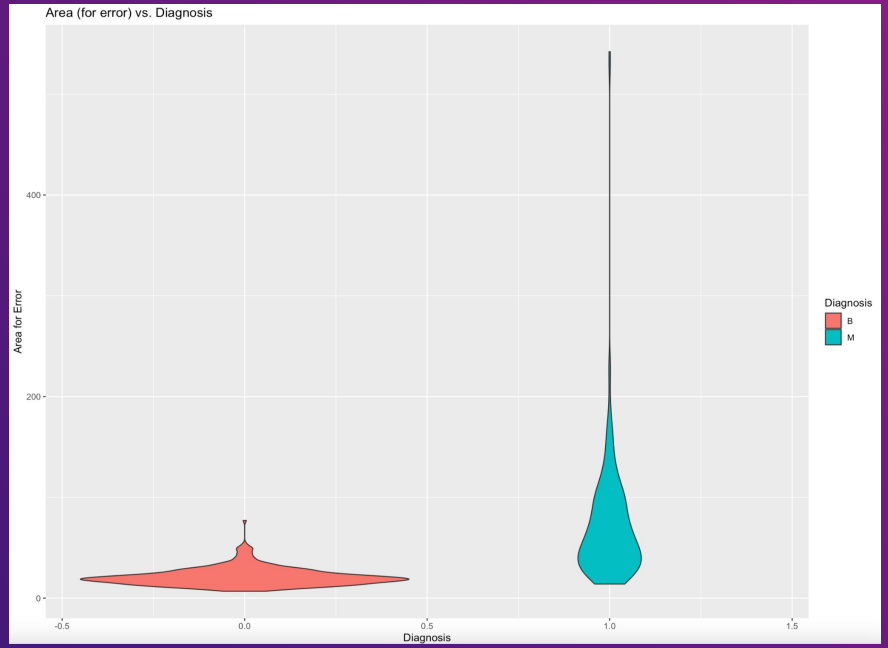
8) Concavity (Mean)



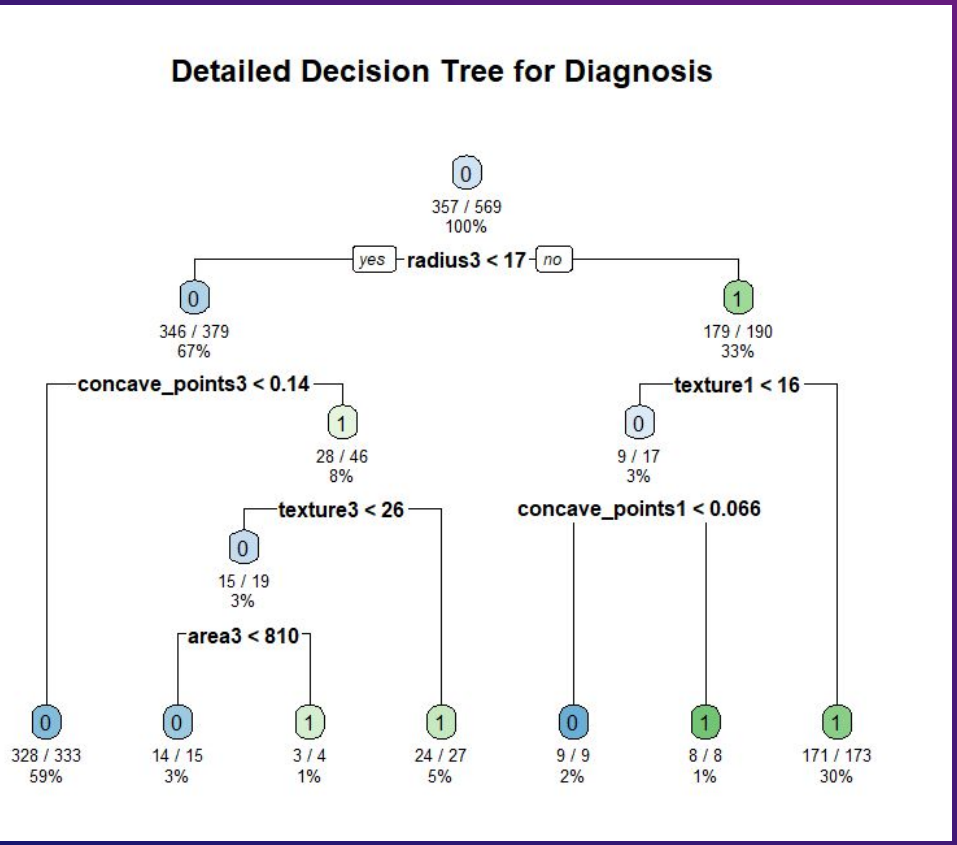
9) Radius (Mean)



10) Area (Error)

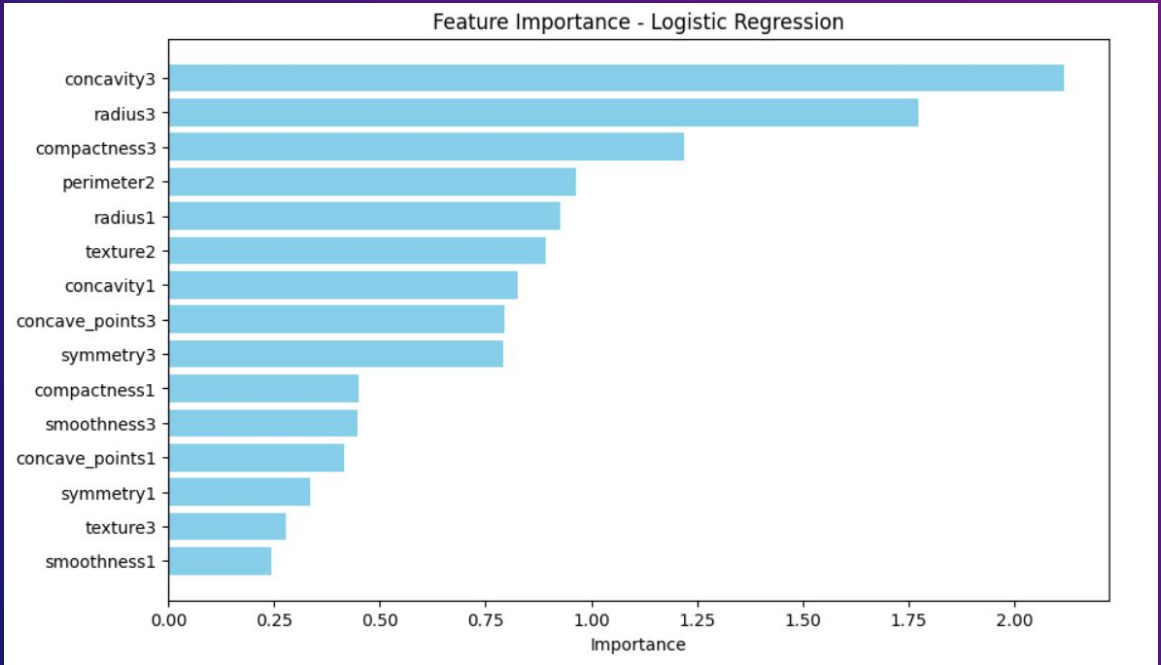


Diagnosis : "Benign (0)", "Malignant (1)"



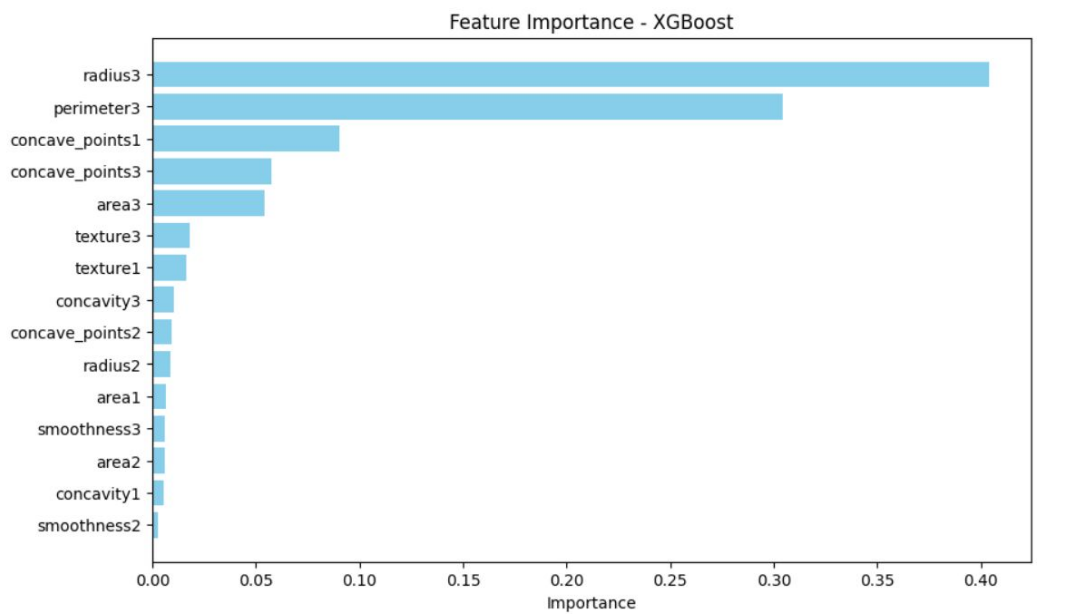
The decision tree graphically represents a series of decision points used to classify observations as either malignant or benign based on specific features from the dataset. Starting from the top, the tree first splits the data based on the 'radius3' feature. If 'radius3' is less than 17, it further examines 'concave_points3' and tends to classify the tumor as benign if 'concave_points3' is also low. Conversely, for larger values of 'radius3', it looks at 'texture1' and is more likely to classify the tumor as malignant for smaller values of 'texture1'. Each node in the tree provides a probability estimate for the classification, with the final nodes (leaves) offering the definitive classification based on the preceding feature thresholds, thereby giving a clear, rule-based approach to diagnosis classification.

LOGISTIC REGRESSION



When creating the logistic model there are multiple ways of picking features. One way is using all features (control) which had an error avg of .0562807018. Another way is looking at a correlation map and picking the darkest spot that gave an error avg of .0624298246. The final way is using Recursive Feature Elimination (RFE) to pick the top 15 features which gave an error avg of .0481666667.

XGBOOST CLASSIFICATION




```
The mean XG_all accuracy : 0.965359649122807
The mean XG_cmm accuracy : 0.9425614035087719
The mean XG_rfe accuracy : 0.9651491228070175
```

→ All Features

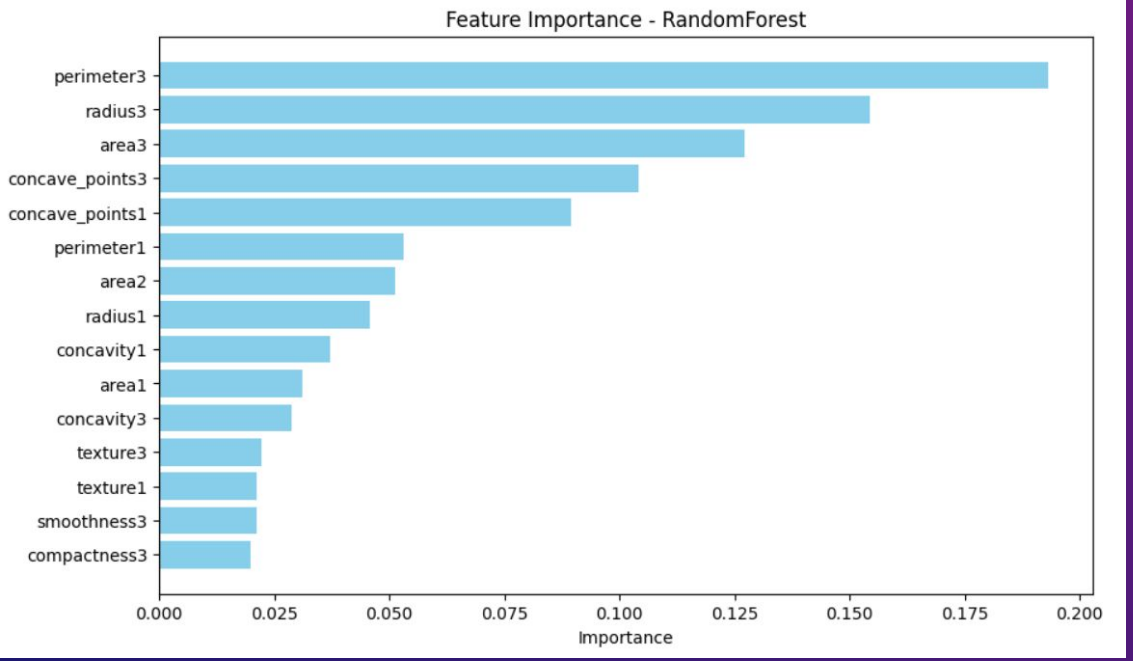
→ Correlation Matrix (top 15)

→ RFE (15)

The XGBoost model demonstrates high predictive accuracy on the test set, achieving low error rates across multiple evaluation metrics

The model's performance on the test set suggests the effectiveness of the model

RANDOM FOREST CLASSIFICATION



The Random Forest model is doing a very good job across all types, the all model uses all the features, cmm uses the top 15 features with highest correlation as shown in previous plots and the RFE model uses top 15 features from RFE.

I ran 1000 simulations, and for each simulation we split data into 80-20 and we train a new model, save its accuracy, in the end I take the mean of all the accuracies.

```
The mean RF_all accuracy : 0.9596929824561404
The mean RF_cmm accuracy : 0.945359649122807
The mean RF_rfe accuracy : 0.9647894736842105
```

→ All Features
→ Correlation Matrix (15 features)
→ RFE (15)

THANK YOU