BREAST CANCER DIAGNOSIS ANALYSIS

GROUP 6
15332-MALEESHA NILUMINDA
15378-SHANALIE RANASINGHE
15669-GANESHI UMAYANGANA

Analysis of Wisconsin Breast Cancer

ABSTRACT

Globally, breast cancer is a major public health issue that requires prompt diagnosis and treatment to enhance patient outcomes. Leveraging machine learning techniques, this study aims to analyze the Wisconsin Breast Cancer dataset to develop a robust classification model capable of distinguishing between benign and malignant tumors.

The dataset includes a variety of features that were taken from digital images of breast mass fine needle aspirates (FNAs), providing important information about the nature of the tumors. Our investigation aims to build an efficient prediction model to help medical practitioners make well-informed decisions about diagnosis and treatment through rigorous data preparation, model selection, and performance evaluation. Our project seeks to improve patient care by using contemporary computational techniques to advance breast cancer detection.

TABLE OF CONTENT

Contents

LIST OF FIGURES

LIST OF TABLES

1. INTRODUCTION

Breast cancer is still a major global health concern that requires precise diagnosis methods and effective treatment plans to lessen its effects on people and society.With the advent of machine learning and the availability of comprehensive datasets such as the Wisconsin Breast Cancer dataset, opportunities have emerged to employ computational methods in improving breast cancer diagnosis.

This dataset encompasses diverse features derived from digitized images of fine needle aspirate (FNA) of breast masses, including quantitative measurements like radius, texture, perimeter, area, smoothness, compactness, concavity, symmetry, and fractal dimension.By analyzing these features, machine learning algorithms can be trained to differentiate between benign and malignant tumors with high accuracy.

In order to create a reliable classification model that can correctly categorize breast tumors, we thoroughly analyze the Wisconsin Breast Cancer dataset in this study.

> *"Cancer survivors are blessed with two lives. There is your life before cancer, and your life after. I am here to tell you your second life is going to be so much better than the first." —Hoda Kotb*

2. DESCRIPTION OF THE QUESTION

Our main goal in this exploratory analysis is to develop a robust classification model capable of accurately classifying breast tumors and make the product to identify the cancer patient. So that we want to get a better understanding about the difference between benign and malignant tumors, which is presented by the qualitative variable "Diagnosis".And also we aim to accomplish the following main goals:

1. Help medical practitioners make well-informed judgments about patient diagnosis and treatment by doing thorough data preparation
2. Model selection, and performance evaluation

Our research harnesses computational methods to advance oncology and enhance breast cancer screening, aiming to improve patient outcomes and contribute to ongoing efforts in the field.

3. DESCRIBE THE DATASET

Variable	Description	Data Type
Diagnosis	M- malignant , B- benign	Categorical
ID number	ID number of the patient	-
Radius	Mean of distance from center to point on the perimeter	Quantitative
Texture	The standard deviation of grayscale values.	Quantitative

Perimeter	The perimeter of the cancer area	Quantitative
Area	Part of the cancer	Quantitative
Smoothness	Local variation in radius lengths	Quantitative
Compactness	perimeter^2 / area-1.0	Quantitative
Concavity	Severity of concave portions of the contour	Quantitative
Concave points	Number of concave	Quantitative
Symmetry		Quantitative
Fractal dimension	Coastile approximation-1	Quantitative

Table 1 - Variable Description

4. DATA PREPROCESSING

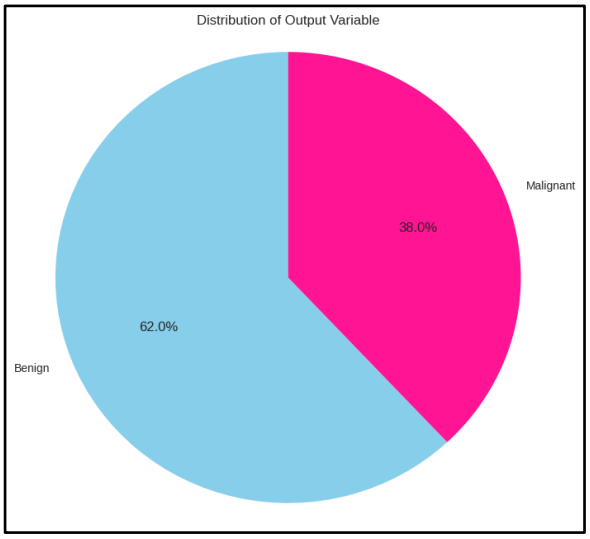
At the start of the project, our main objective was to develop a robust classification model capable of distinguishing between benign and malignant tumors.
We began the feature engineering process by removing the ID and unnamed:32 variable columns. And also we confirmed there were no missing values and no duplicates. After that, we assign Y(predictor variable) as a Diagnosis variable and the remaining 30 variables are the X(explanatory variables).
After that we used this reprocessed dataset for our analysis.

5. IMPORTANT RESULTS IN DESCRIPTIVE ANALYSIS

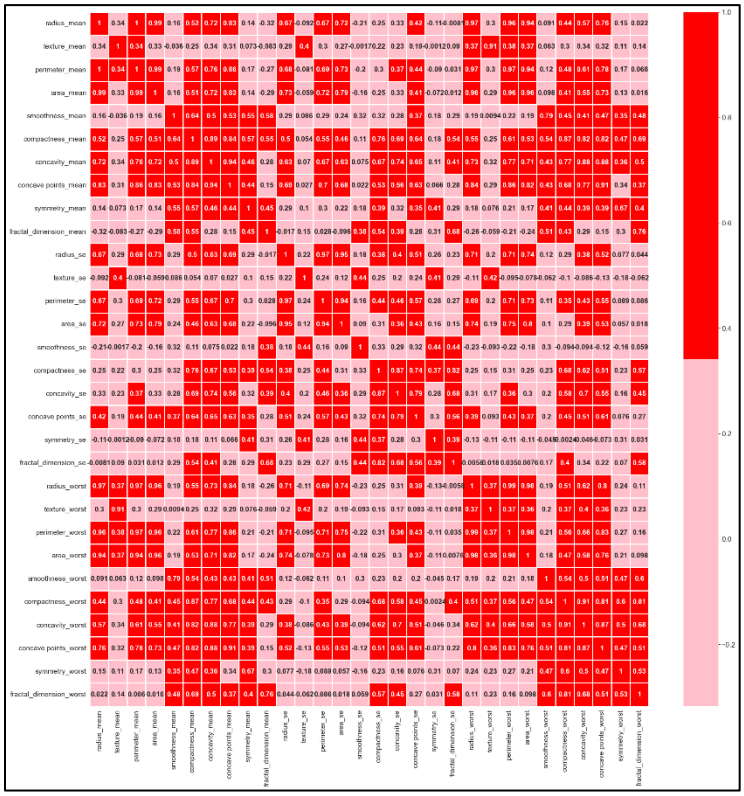
- **Response variable - Diagnosis**

The response variable Diagnosis has two categories. These are alignant and benign.
From the pie chart, the majority of the patients are in the bening category. The percentage of that is 62%. And the malignant category has 38% .
Then also this shows the somewhat imbalanced dataset.

Figure 1- Pie Chart of Diagnosis



- **Correlation plot of variables**



This plot shows the correlation between the explanatory variables. Then it shows some multicollinearity between the variables.

Figure 2- Correlation plot of all the variables

- **Outlier Analysis, Multicollinearity, Class imbalance, Large number of variables**

  - We use the Mahalnobius technique to identify the outliers. But we got a very low outlier percentage.
  - Our explanatory variables have the mean, se, and worst measurement of each variable. Then these variables can be related to each other. However, we can't remove the multicollinearity variables before our analysis part.
  - Due to the seriousness and sensitivity of breast cancer, we didn't remove the outliers and the multicollinearity variable. Because we haven'y good domain knowledge about the breast cancer.
  - The malignant tumor category has 62% and the benign tumor category has 38%. Then the two categories are somewhat imbalanced. Therefore we used sampling techniques(SMOTE) and did further analysis parallel with the SMOTH dataset also.
  - Dataset has 30 variables then we parallelly analyze with all the variables and with selected variables using tree-based methods.

- **Univariate Analysis**

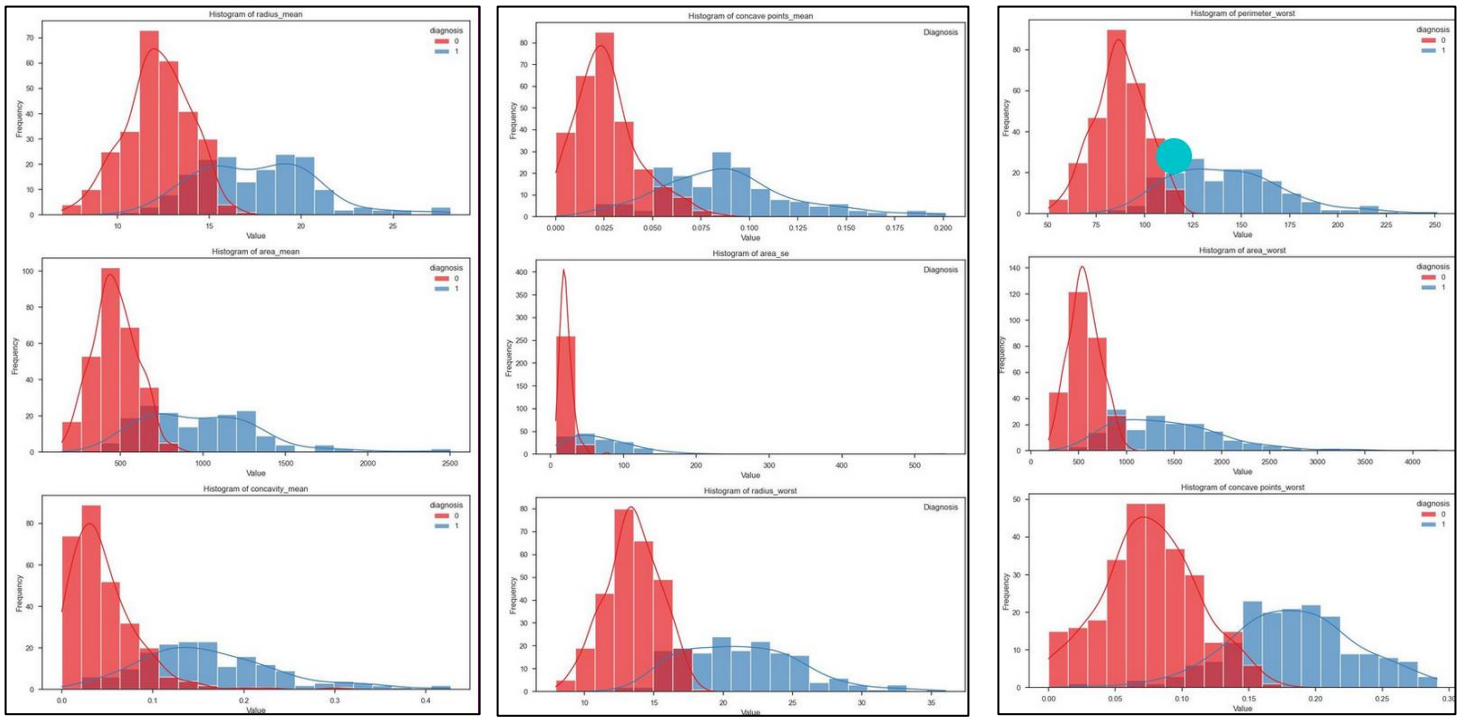


Figure 3- Histograms of the variables

- Most histograms related to benign tumors exhibit a positive skew, indicating data concentration towards lower values.
- The "area standard error" histogram is highly positively skewed, with data clustering towards lower values
- In contrast, the distribution of "concave points worst" is less skewed, resembling a bell shape and suggesting a more balanced distribution.
- As the radius and area worsen, the frequency decreases, possibly due to challenges in isolating larger masses from surrounding tissue on mammogram images.
- A slight positive skew in the radius_mean distribution implies smaller masses are more prevalent in breast tissues

- **Bivariate Analysis**

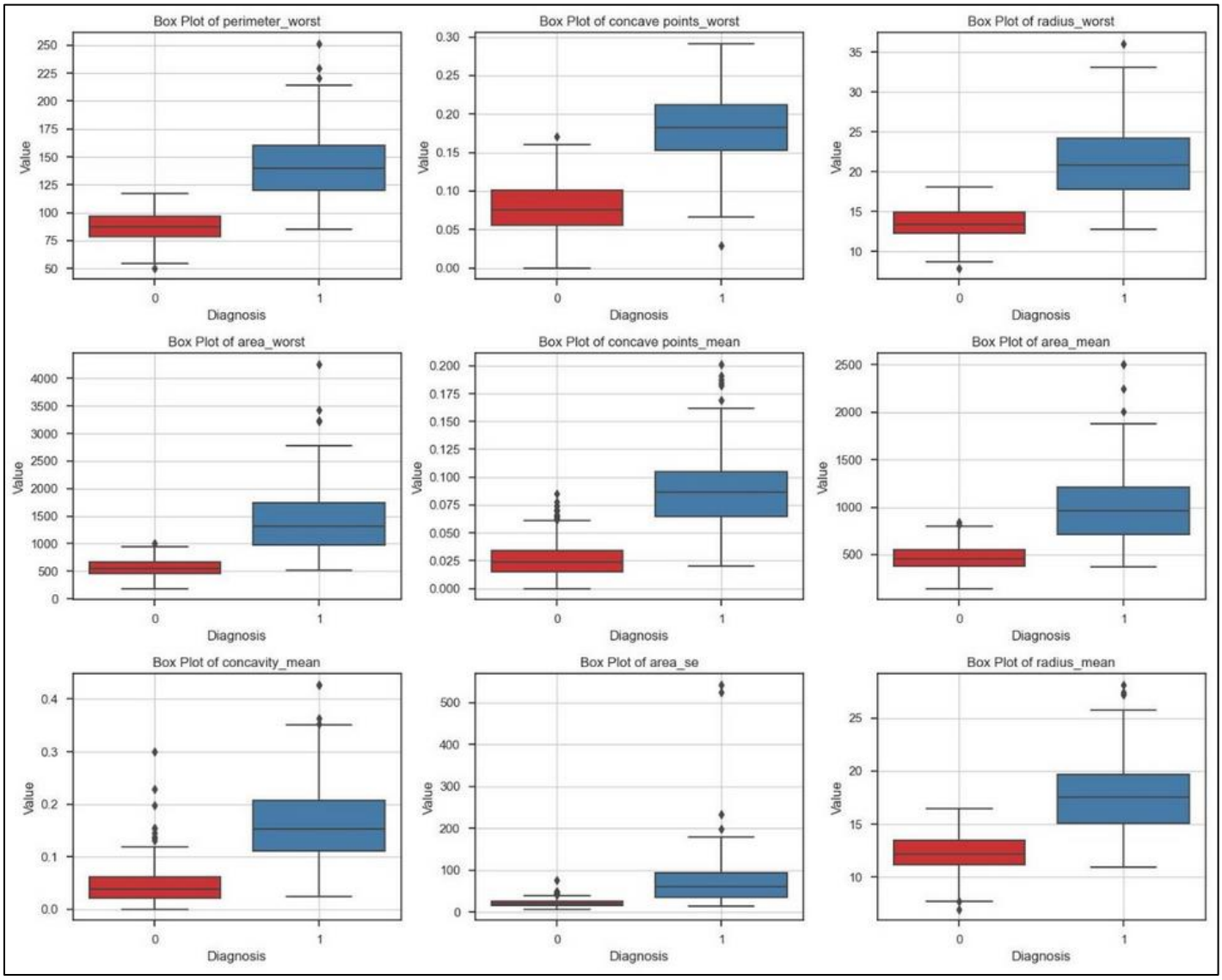


Figure 4- Boxplots of the variables

- All the histogram's identical patterns are the same.
- In this histograms 1-Malignant tumor category and 0-Benign tumor category.
- All the boxplots Malignant tumor category have a higher median than the Benign tumor maximum value. Also the malignant category spread is higher than the Benign category spread.
- Also some boxplots have some outliers.

- **Relationship between the explanatory variables**

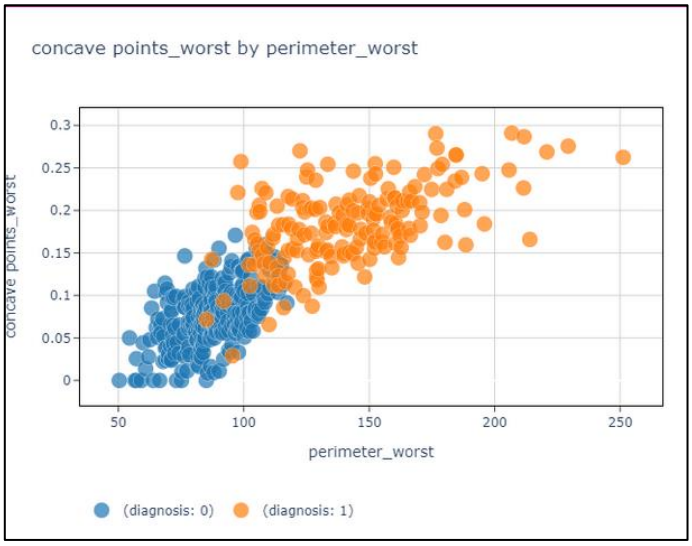


Figure 6 - Scatterplot of perimeter_worst Vs concave point_worst

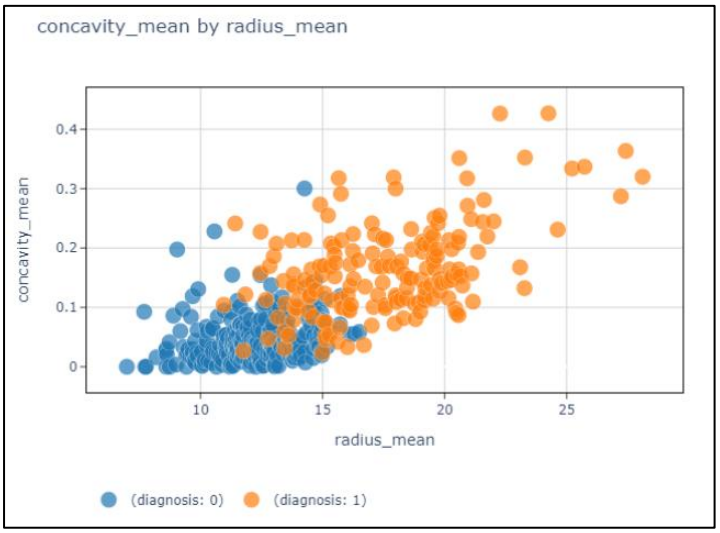


Figure 5- Scatterplot of radius_mean vs concavity_mean

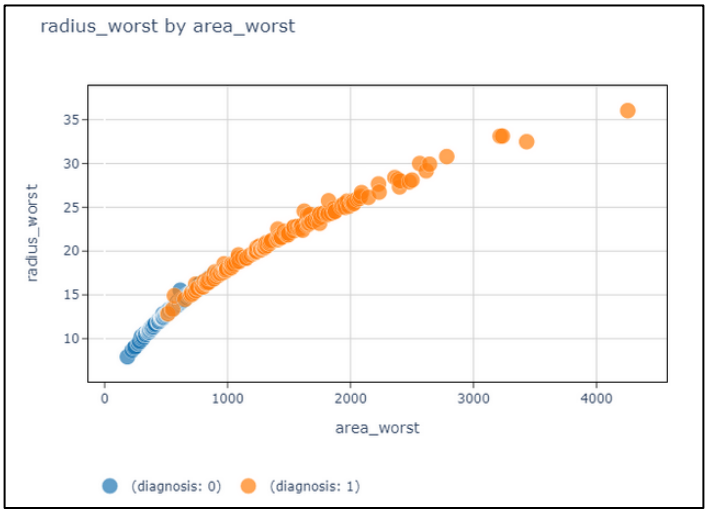


Figure 8- Scatterplot of area_worst Vs radius_worst

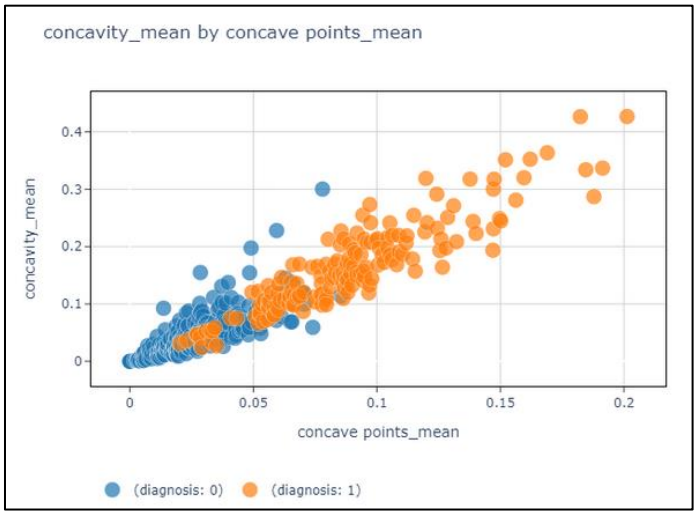


Figure 7- Scatterplot of concave point_mean Vs concavity_mean

- Concave Point worst and Perimeter worst are positively correlated, demonstrating a positive association(Figure 5). Also Concave points and concavity demonstrate a positive correlation(Figure 8).
- Radius worst and area worst are strongly connected, almost like they follow a straight line together. So Radius worst has the highest value of the center for the estimated range(Figure 7).
- And also Concavity and radius both show a positive relationship, but it's not as precise as the previous plot. It shows a moderate positive relationship(Figure 6).
- In all scatterplots, yellow points represent the malignant tumor category and blue points represent the benign tumor category. Here the malignant tumor category has higher performance than the benign tumor category.

- **PLSR**

This partial least square regression score plot has two components. And also observations represent the 2 separate groups. Therefore our dataset may have clusters.

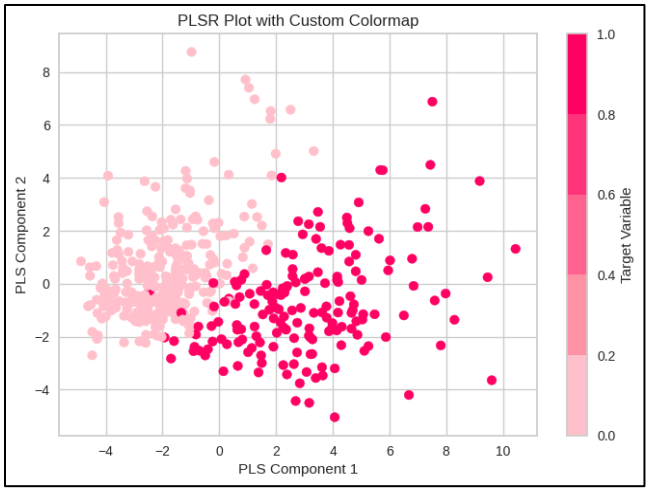


Figure 9- PLSR Score Plot

- **K-means Clustering**

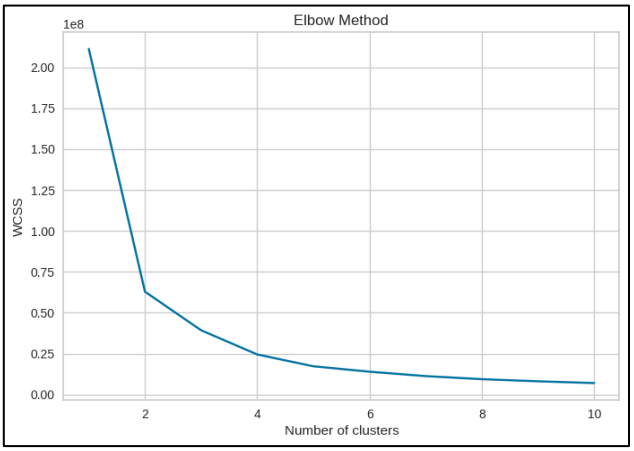


Figure 11- Elbow Plot

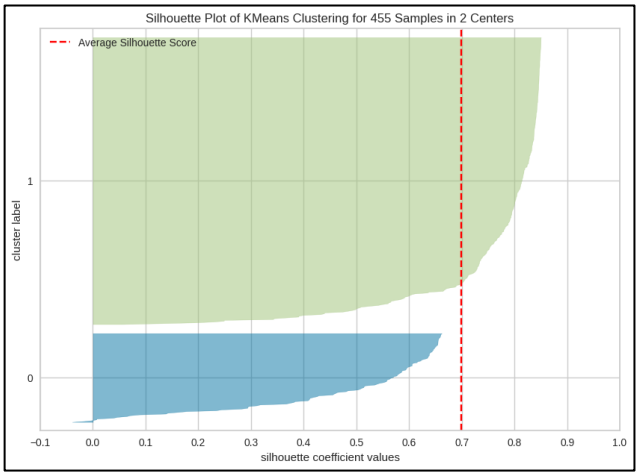


Figure 10- Shilhoutte Plot

Using the elbow method we get the maximum number of clusters are 2. Then we calculate the silhouette value using the two clusters. Then here the average silhouette value is 0.7. It is of high value. Then we can clearly cluster our dataset under two clusters. But here one cluster has more observations and the other cluster has very few observations. Then also these two clusters are imbalanced.

But using the two clusters using the training test we calculate the number of observations under each category.

Then cluster_1 has 347 observations and cluster_0 has 108 observations. Therefore two clusters are already imbalanced.

In cluster_0 malignant tumor category has 108 observations but the benign category hasn't any observations. Also cluster_1 malignant tumor category has 65 observations and the benign category has 282 observations. Both cluster's responses are heavily imbalanced.

If a patient is classified into Cluster _0 in new data, they will be labeled as a cancer patient (malignant) regardless of whether they have cancer.the reason of that Cluster _0 hasen't any observations under benign tumor category.

Therefore we did not consider the clustering technique for our further analysis and then we ignored the clusters and used a full training and testing dataset.

- **Final Findings of Descriptive Analysis.**

- Malignant tumors have higher mean, worst, and se measurements than benign tumors
- In uni-variate result, the frequency decreases When the radius worst and the area worst increases. After bivariate analysis, we can predict the area worst and the radius worst have a strong positive relation with each other.
- There are outliers in each of these features as shown in the boxplots
- Certain level of separation in the values for features in the benign (diagnosis=0) and malignant (diagnosis=1)data points

6. IMPORTANT RESULTS IN ADVANCED ANALYSIS

Our diagnosis breast cancer dataset response variable is a binary variable with malignant and benign two categories.
Compare the model performances we used below keywords.

- **Classification Accuracy:** It defines how often the model predicts the correct output. It is an important parameter to determine the classification problems' accuracy.
- **Precision:** It can be defined as the number of correct outputs supplied by the model or the percentage of all positive classes
- **Recall:** It is defined as the percentage of positive classes that our model accurately predicted.
- **F1-score:** The F-score allows us to assess both recall and precision simultaneously.

Then we used accuracy and recall values to compare the model performances.
Here we fit all the models with the SMOTE dataset, with selected variables using the decision tree method and with all the variables. But fitting models with all variables yields the highest accuracy compared to models fitted with the other two techniques. Then we use fitting models with all the variables for our further comparisons.

1. Binary logistic regression

Our predictor variable has two categories, malignant and benign then we used a binary logistic regression model.

Accuracy		F1_Score	
Training Set	**Test Set**	**Training Set**	**Test Set**
0.9868	**0.9649**	**0.9825**	**0.9487**

Table 2 - Results of Binary Logistic Regression

2. KNN (K nearest neighbor)

K-nearest neighbors (KNN) algorithm is a type of supervised ML Algorithm, one that relies on labeled input data to learn a function that produces an appropriate output when given new unlabeled data. The following values were found on test data the value of k=3 to get the best results.

Accuracy		F1_Score	
Training Set	**Test Set**	**Training Set**	**Test Set**
0.9560	**0.9211**	**0.9401**	**0.8831**

Table 3- Results of KNN

3. Random Forest

Decision trees are the basic learning model used by the categorization method Random Forest. Since each tree will make a different mistake, aggregating the findings of several trees ought to produce results that are more accurate than those of a single tree, according to the basic premise of Random Forest. After parameter tuning, we got 10 max_splits and 1000 tress in our final random forest model.

Accuracy		F1_Score	
Training Set	**Test Set**	**Training Set**	**Test Set**
0.9999	**0.9825**	**0.9999**	**0.9750**

Table 4- Results of Random Forest with hyperparameter tunning

4. XGBoost

XGBoost is a high-performance machine learning algorithm for predictive modeling, delivering exceptional accuracy and efficiency through advanced gradient-boosting techniques

Accuracy		F1_Score	
Training Set	**Test Set**	**Training Set**	**Test Set**
0.9998	**0.9649**	**0.9999**	**0.9487**

Table 5- Results of XGBoost

5. Model Performance Comparisons

After using the accuracy and the F1-Score values, We want to identify what is the best method for for predict breast cancer.

	Accuracy		F1_Score	
MODEL	**Test Set**	**Training Set**	**Training Set**	**Test Set**
Binary Logistic	**0.9868**	**0.9649**	**0.9825**	**0.9487**
KNN	**0.9560**	**0.9211**	**0.9401**	**0.8831**
Random Forest	**0.9999**	**0.9825**	**0.9999**	**0.9750**
XG Boost	**0.9998**	**0.9649**	**0.9999**	**0.9487**

Table 6- Compare the model permormances with all the variables

Above various types of Machine Learning Models, we decided that the Random Forest Classifier performs best with the following variables most importantly model building.

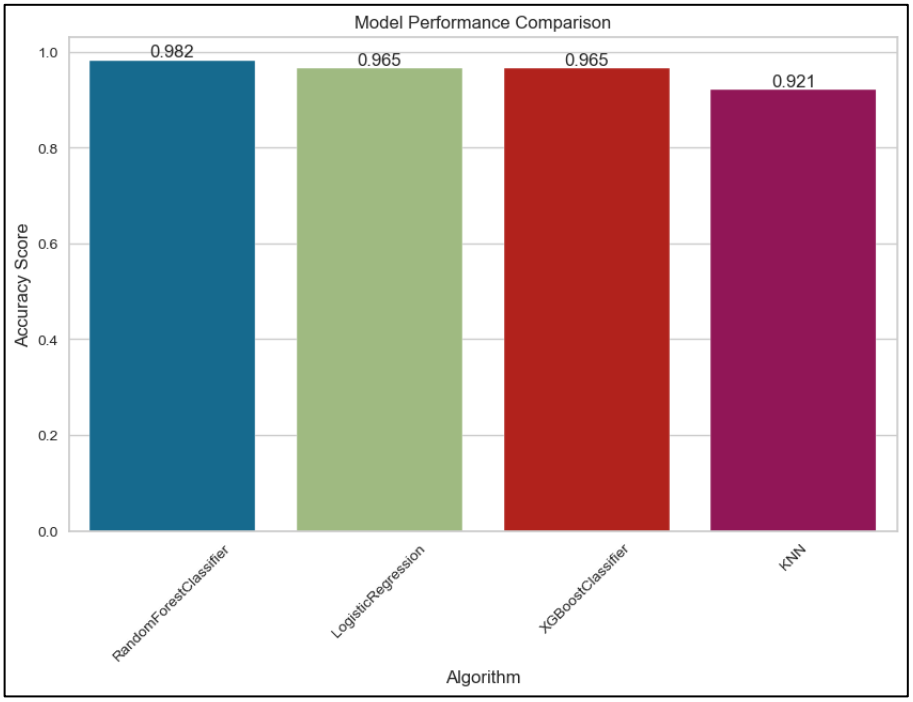


Figure 12- Model Performance Comparison Plot

6. Variable Importance and select the important variables.

After selecting the random forest model as the best model we want to select the important variables under all 30 variables.

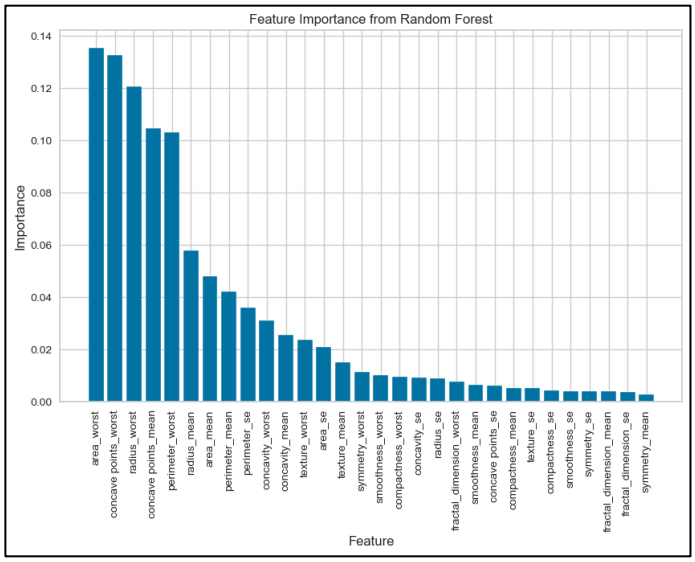


Figure 14- Feature Importance from Random Forest

FEATURES	IMPORTANCE
PERIMETER_WORST	0.2566
CONCAVE POINT_WORST	0.1796
RADIUS_WORST	0.1553
AREA_WORST	0.1507
CONCAVE POINT_MEAN	0.1168
AREA_MEAN	0.0479
CONCAVITY_MEAN	0.0338
AREA_SE	0.0307
RADIUS_MEAN	0.0287

Figure 13- Importances of Selected Varibles

After selecting the 9 important variables we fit the random forest model again then we got **0.9649** accuracy value.

7. ISSUES ENCOUNTED AND SOLUTIONS.

- The dataset was unbalanced since there were more observations under the "Benign" (62%) category : By combining SMOTE with domain knowledge, we can effectively address class imbalance in the dataset while ensuring that the synthetic samples generated are meaningful and relevant for breast cancer classification.
- Multicollinearity exists between variables: Get help from domain expertise to identify which features are most relevant for predicting breast cancer outcomes and focus on those while excluding redundant or less informative features.
- Dataset has a large number of variables : Get help from domain knowledge and expertise to prioritize variables that are most likely to be clinically relevant or biologically significant in the context of breast cancer. Focus on variables that are known to be associated with breast cancer risk, diagnosis, or prognosis.
- The distribution of responses within the cluster is highly imbalanced : Take into account any domain-specific knowledge or constraints that may influence the distribution of observations within clusters.

8. LIMITATION OF OUR ANALYSIS

- The dataset only considers tumor measurement of the patient, omitting information on the patient's other disease history.
- Dataset include small number of observations.
- The breast cancer dataset could be its lack of genetic or molecular information about the tumors. Understanding the genetic mutations or molecular subtypes of the tumors could provide valuable insights into their behavior and potential treatment options.
- ***Without these information, the dataset may not fully capture the heterogeneity of breast cancer and limit the depth of analysis and conclusions that can be drawn from it.***

9. CONCLUSIONS

- The random forest approach, which had the highest test accuracy and a high F1 score, turned out to be the best model overall in our search to correctly classify breast cancer diagnoses.
- The final random forest model identified the following key variables: perimeter_worst, concave point_worst,radius_worst, area_worst, concave point_mean, area_mean, concavity_mean, area_se and radius_mean. These factors are essential for correctly identifying between cases that are benign and those that are malignant.
- Our analysis resulted in a notable reduction in the number of features utilized for classification, from an initial 31 variables down to a refined set of 9. Process efficiency is increased by this reduction, which also may lead to an improvement in model interpretability.

10. REFERENCES

1. **https://www.researchgate.net/publication/337486825_Breast_Cancer_Wisconsin_Diagnostic_Data_Set**
2. **https://www.kaggle.com/code/mennaafifyy/breast-cancer-dataset-eda-knn-classification**
3. **https://www.analyticsvidhya.com/blog/2020/10/overcoming-class-imbalance-using-smote-techniques/**
4. **https://medium.com/@shashmikaranam/exploratory-data-analysis-breast-cancer-wisconsin-diagnostic-dataset-6a3be9525cd**
5. **https://www.mayoclinic.org/diseases-conditions/breast-cancer/symptoms-causes/syc-20352470**
6. **https://www.diva-portal.org/smash/get/diva2:1679145/FULLTEXT02**
7. **https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7349542/**
8. **https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7349542/**
9. **https://bmcmedinformdecismak.biomedcentral.com/articles**

11. APPENDIXES

https://github.com/GaneshiUmayangana/BreastCancer-Prediction