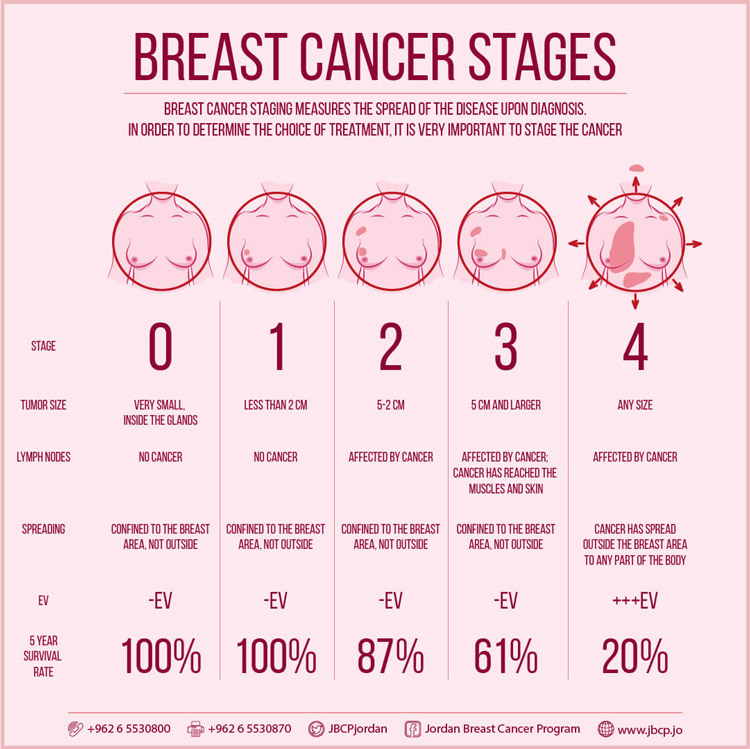
**Prediction of Breast Cancer Using Machine Learning Algorithms**

Sahaya Jefrin M - 2148112

**1. Introduction**

Breast cancer is a type of cancer that starts in the breast. It can start in one or both breasts. Breast cancer can spread when the cancer cells get into the blood or lymph system and then are carried to other parts of the body. The lymph (or lymphatic) system is a part of your body's immune system. It is a network of lymph nodes (small, bean-sized glands), ducts or vessels, and organs that work together to collect and carry clear lymph fluid through the body tissues to the blood. The clear lymph fluid inside the lymph vessels contains tissue by-products and waste material, as well as immune system cells.



**2. Problem Description**

Finding breast cancer early and getting state-of-the-art cancer treatment are two of the most important strategies for preventing deaths from breast cancer. Breast cancer that’s found early, when it’s small and has not spread, is easier to treat successfully. In this work, we will apply a breast cancer classification to predict whether the cancer is Benign or Malignant.

**3. Data Description**

Features are computed from a digitized image of a Fine Needle Aspirate (FNA) of a breast mass. They describe the characteristics of the cell nuclei present in the image.

The data set was taken from UCI Machine Learning Repository which is Breast Cancer data.

**3.1 Attribute Information:**

1) ID number

2) Diagnosis (M = malignant, B = benign)

Ten real-valued features are computed for each cell nucleus:

a) radius (mean of distances from the centre to points on the perimeter)

b) texture (standard deviation of grey-scale values)

c) perimeter

d) area

e) smoothness (local variation in radius lengths)

f) compactness (perimeter^2 / area - 1.0)

g) concavity (severity of concave portions of the contour)

h) concave points (number of concave portions of the contour)

i) symmetry

j) fractal dimension ("coastline approximation" - 1)

**4. Methodology**

**4.1 Supervised Learning:**

Supervised learning is the type of machine learning in which the training is done based on labelled training data that is the target or response variables are clearly defined. In supervised learning, the training data provided to the machines work as the supervisor that teaches the machines to predict the output correctly. A supervised learning algorithm aims to find a mapping function to map the input variable with the output variable.

**4.2 Principal Component Analysis:**

As the number of features is very high the dimension is reduced using principal component analysis. The Principal Component Analysis is an unsupervised learning technique for reducing the dimensionality of data. It increases interpretability and minimizes information loss. The data is standardized before performing PCA this ensures that each feature has a mean = 0 and variance = 1. Eigenvalues and eigenvectors of the covariance matrix are computed. The

percentage of variation is calculated using the eigenvalues and the components with variation greater than 90% are considered for the analysis.

**4.3 Logistic Regression:**

The logistic regression model is fitted on these selected features and predictions are made. Logistic regression is a Supervised Learning technique. It is used for predicting the categorical dependent variable using a given set of independent variables. In Logistic regression, an “S” is a shaped logistic function, which predicts two maximum values (0 or 1). Logistic Regression is a significant machine learning algorithm because it can provide probabilities and classify new data using continuous and discrete datasets. Logistic Regression can be used to classify the observations using different types of data and can easily determine the most effective variables used for the classification.

**4.4 Linear Discriminant Analysis:**

A statistical method called linear discriminant analysis (LDA) is employed to classify data. LDA determines the optimum way to divide the classes using a linear combination of the predictor variables. A threshold can be established to categorise new data after the linear combination has been discovered. As a dimensionality reduction method, supervised classification issues are frequently solved using linear discriminant analysis also known as normal discriminant analysis or discriminant function analysis.

**4.5 Quadratic Discriminant Analysis:**

Data can be divided into two categories using the statistical method known as quadratic discriminant analysis (QDA) (e.g. male or female). While QDA and linear discriminant analysis (LDA) are comparable techniques, LDA cannot handle data that cannot be separated linearly. In QDA each group's mean and covariance are calculated. It then computes a discriminant function for each group using these statistics.

Each data point is given a score using the discriminant function, which is a quadratic function. After that, the data point is assigned to the group with the highest score. While QDA is more adaptable than LDA, it also has a higher risk of overfitting. When the groups are divided, QDA typically performs better than LDA. LDA often performs better than QDA when the groups are not sufficiently separated.

**4.6 K Nearest Neighbour’s Classifier:**

K Nearest neighbours (KNN) algorithm is a kind of supervised machine learning algorithm which can be utilized for both classification problems when the conditional distribution of (Y|X) is not specified. KNN is a non-parametric algorithm. A test observation is chosen and the k nearest points to the test observation are taken. To determine which data points are closest to a given query point, the distance between the test point and the other data points will need to be calculated. These distance metrics help to form decision boundaries, which partition query points into different regions. The common distance measures used are Euclidean distance, Minkowski distance etc. The test observation is assigned to a class based on the most frequent.

**4.7 Naive Bayes:**

The Naive Bayes algorithm is a supervised learning algorithm, which is based on the Bayes theorem and used for solving classification problems. It is mainly used in text classification that includes a high-dimensional training dataset. Naïve Bayes Classifier is one of the simple and most effective Classification algorithms that help in building fast machine learning models that can make quick predictions. It is a classification technique based on Bayes’ Theorem with an assumption of independence among predictors.

Where, P(Y|X) = P(X|Y) \* P(Y)/P(X)

* P(Y|X) is Posterior probability: Probability of hypothesis Y on the observed event X.
* P(X|Y) is Likelihood probability: The probability of the evidence given that the
* probability of a hypothesis is true.
* P(Y) is Prior Probability: The probability of the hypothesis before observing the evidence.
* P(X) is Marginal Probability: Probability of Evidence.

Types of Naive Bayes model:

* Gaussian: The Gaussian model assumes that features follow a normal distribution.
* Multinomial: The Multinomial Naïve Bayes classifier is used when the data is multinomial distributed.
* Bernoulli: The Bernoulli classifier works similarly to the Multinomial classifier, here the features will have only 2 categories 0 and 1.
* Hybrid: This model is used when a feature set contains both categorical and continuous variables.

**4.8 Support Vector Machine:**

The supervised learning algorithm known as Support Vector Machines (SVMs) may be applied to classification or regression problems. SVMs are adaptable machine learning algorithms that work with many different types of data sets.

SVMs are a popular classification method because they have a high degree of accuracy while being comparatively easy to use. Finding a hyperplane that optimises the margin between the two classes is the basic goal of SVMs. Support vectors refer to the data points that are closest to the hyperplane. The margin is the separation between the support vectors and the hyperplane. Finding a hyperplane with the biggest margin is the objective.

**4.9 Decision Tree:**

Decision trees are a type of machine learning algorithm that is used to create models that can predict outcomes based on certain inputs. Decision trees are used in a variety of applications, such as classification and regression. A decision tree is a graph that displays various choice-related outcomes. It consists of nodes, which represent the decisions that must be taken, and branches, which stand in for the potential results of those choices. Your decision-making process may be visually organised with the aid of a decision tree. You may use it to map out every option you could have to make as well as the various results of each choice.

**5. Explorative Data Analysis:**

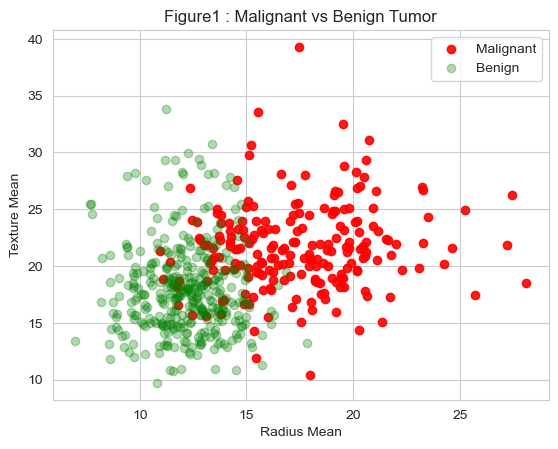
Exploratory data analysis is used to identify the important characteristics of the data and to understand the data better. To make the data suitable for analysis, pre-processing is performed on the data. The fundamental analysis is carried out, including checking for null values, determining the data types of each variable to determine whether data formatting is necessary or not, producing a summary of the data set, and calculating the correlation between the variables. Since this data set contains no null values and all variables are stored as either integers or floats, data type conversion for the features is not necessary.

However, the study variable is transformed from a categorical data type to an integer data type (i.e., 1= B and 0 = M). Since all categorical variables are kept as integers, data binning is not necessary. The statistical overview of the data collection would be obtained using the describe function. All the variables in the data set have their means, maximums, minimums, counts, and median values calculated. Using value counts functions, the summary of categorical variables is generated. Categories B and M are for the response variable. The frequency of class B is determined to be 357, and that of class M is to be 212.

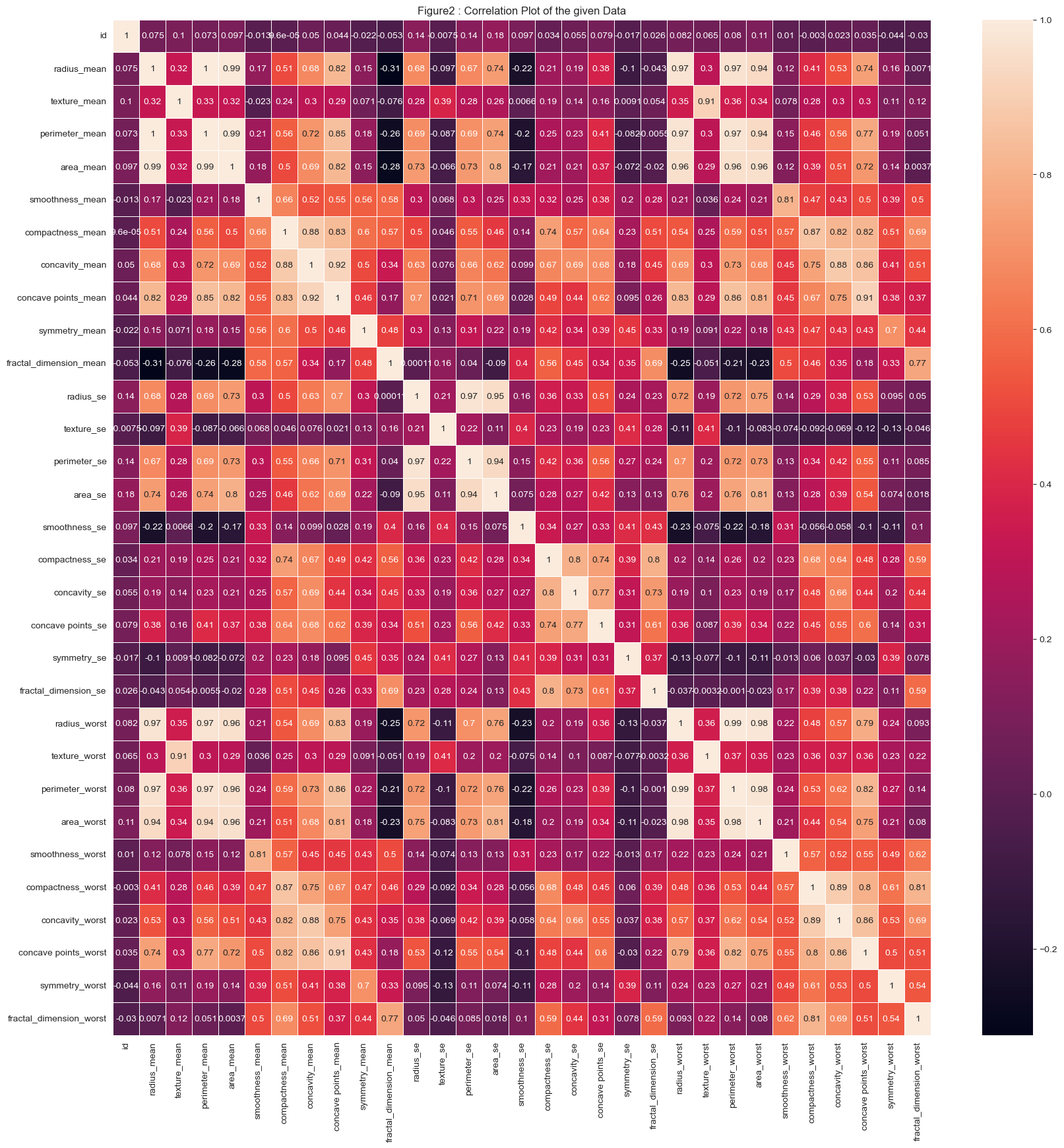
**5.1 Data Visualization:**

5.1.1 Scatter Plot

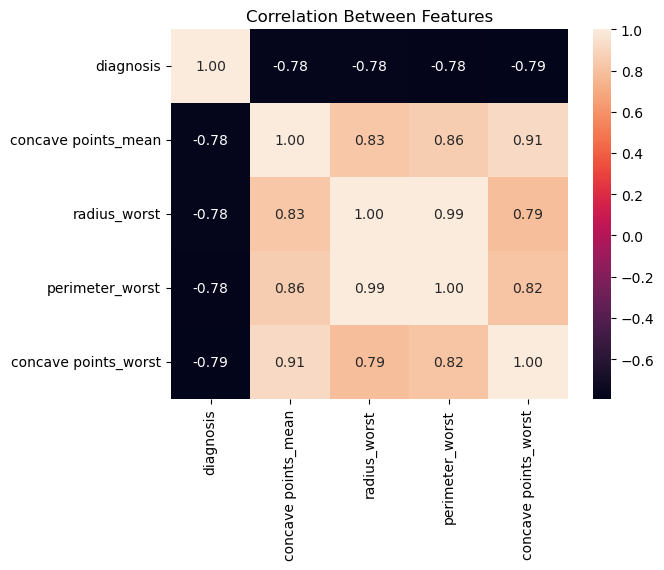
There are several outlier numbers in the scatterplot below that are not uncommon. The two values together, however, clearly do not suit the overall relation. Because two variables are scattered. It may be inferred from the plot that there may be a positive relationship between the major variables.

****

5.1.2 Heat Map

****

From the heatmap, it can be seen that the relationship between "diagnostic," "concave points mean," "radius worst," "perimeter worst," and "concave points worst"  is more than 0.75. It shows that it has a strong correlation. Heat Map for those variables are plotted below.



The heatmap clearly shows that all variables are multicollinear in nature.

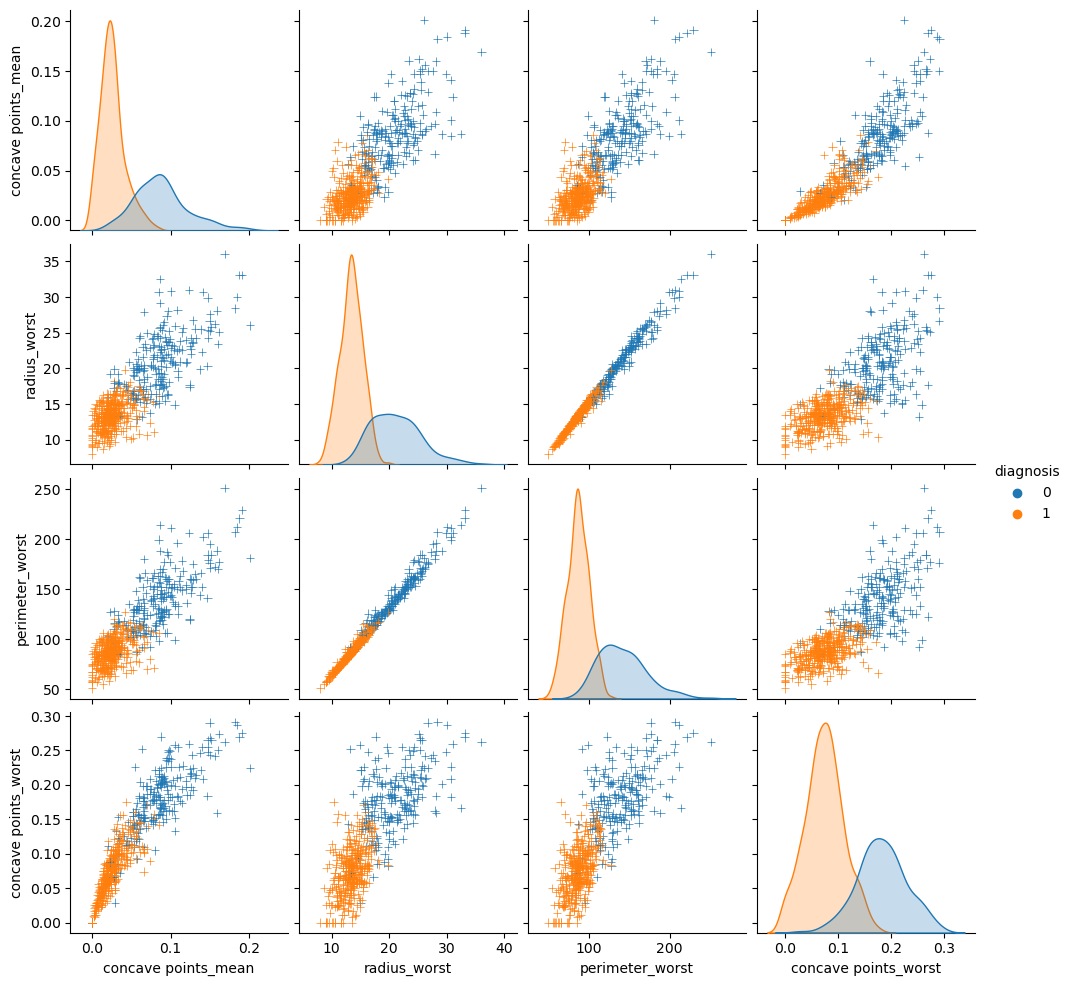
5.1.3 Count Plot:



Count of label 'B': 285

Count of label 'M': 170

**5.1.4 Pair Plot**

****

Here the variables have a positive relationship with each other and the concave points\_worst variable have bell shaped curve, it shows that follows a normal distribution.

**6. Analysis:**

6.1 Synthetic Minority Oversampling Technique

By the count plot of the study variable diagnosis, it's seen that it’s an imbalanced classification.

The percentage of Benign diagnosis patients is 62.741652021089635

The percentage of Malignant diagnosis patients is 37.258347978910365

The synthetic minority oversampling technique is used to obtain a balanced classification.

Before Oversampling, counts of label 'B': 285

Before Oversampling, counts of label 'M': 170

After Oversampling, the shape of train\_X: (570, 31)

After Oversampling, the shape of train\_y: (570,)

After Oversampling, counts of label 'B': 285

After Oversampling, counts of label 'M': 285

Principle component analysis is used as Dimensionality Reduction Technique. PCA technique is used with a percentage variation set as 95%. 11 features are obtained which have a significant effect on the study variable.

**6.2 Model Building:**

The selected components along with the target variable are split into train and test data. 20% of the data is used for testing and the remaining portion of data is used for training the model.

**6.2.1. Logistic Regression Model**

The logistic regression model is fitted on these selected features and predictions are made.

**Accuracy Measures:**

**a). Accuracy Score: 0.98**

The accuracy measure is 0.98. Accuracy refers to the closeness of a measured value to a standard or true value. A value greater than 70% is a great model performance. The accuracy measure of the given data set is 0.98 thus the model is a good model.

**b). Classification Report:**

precision recall f1-score support

B 0.98 0.98 0.98 64

M 0.98 0.98 0.98 50

accuracy 0.98 114

macro avg 0.98 0.98 0.98 114

weighted avg 0.98 0.98 0.98 114

Out of all the positive predictions: -

* Only 98% of Benign diagnoses are correctly expected to be patients with breast cancer.
* Only 98% of Malignant diagnoses are correctly expected to be patients with breast cancer.

Out of all the actual positives:

* 98% of Benign diagnoses are correctly expected to be patients with breast cancer.
* 98% of Malignant diagnoses are correctly expected to be patients with breast cancer.

The harmonic mean of precision and recall, with equal weights for each, is used to calculate the F-measure. This feature makes it possible to compare models and evaluate a model's performance by accounting for both precisions and recall with a single score.

A lower beta value favours less recall and favours precision more. Here, the weighted average is the percentage of observations in each class, and the macro average is the AM of the scores from the two classes. Since the categorisation is balanced, the macro average and weighted average are the same.

**c). Confusion Matrix:**

[[63 1]

[ 1 49]]

In this data, we have 2 categories B and M. So we have a 2x2 confusion matrix. Here rows are test data, columns are classes obtained by the logistic model. There are 2 categories so 63 represents several observed values classified as **B**  belonging to 1st class, and 1 in the first row is several observed values classified as **B** but which is from class **M**. Similarly, for the next row 1 is the number of observations classified as **M** but belonging to 1st class, 49 is the number of observations classified as **M** belonging to the 2nd class.

**6.2.2. Logistic Regression with K-Fold Cross Validation**

The logistic regression model is fitted with K-Fold cross-validation on these selected features and predictions are made. The number of splits is kept at 30 and shuffle is considered True. The scoring is based on accuracy score.

**Accuracy Measures:**

Accuracy Score: 0.97

The accuracy measure is 0.97. Accuracy refers to the closeness of a measured value to a standard or true value. A value greater than 70% is a great model performance. The accuracy measure of the given data set is 0.97 thus the model is a good model.

**6.2.3 Linear Discriminant Analysis**

**Accuracy Measures:**

a). Accuracy Score: 0.97

The accuracy measure is 0.97. Accuracy refers to the closeness of a measured value to a standard or true value. A value greater than 70% is a great model performance. The accuracy measure of the given data set is 0.97 thus the model is good.

b) Confusion matrix:

[[48, 2],

[ 1, 63]]

In this data, we have 2 categories B and M. So we have a 2x2 confusion matrix. Here rows are test data, columns are classes obtained by the logistic model. There are 2 categories so 48 represents several observed values classified as **B**  belonging to 1st class, and 2 in the first row is several observed values classified as **B** but which is from class **M**. Similarly, for the next row 1 is the number of observations classified as **M** but belonging to 1st class, 63 is the number of observations classified as **M** belonging to the 2nd class.

c) Multilabel Confusion Matrix:

[[[63, 1],

[ 2, 48]],

[[48, 2],

[ 1, 63]]]

In this data, we have 2 categories B and M. So we have a 2x2 multilabel confusion matrix. Here rows are test data, columns are classes obtained by the logistic model. There are 2 categories so 63 and 48 represent several observed values classified as **B** belonging to 1st class, and (1 and 2) in the first row are several observed values classified as **B** but which is from class **M**. Similarly, for the next row (2 and 1) is the number of observations classified as **M** but belonging to 1st class, (48 and 63 ) is the number of observations classified as **M** belonging to the 2nd class.

**6.2.4. Implementing Linear Discriminant Analysis with K- Fold Cross Validation**[**¶**](http://localhost:8888/notebooks/2148112_ML_CAC_3_Mini_Project.ipynb#Implementing--Linear-Discriminant-Analysis-with-KFold-Cross-Validation)

**Accuracy Measures:**

Accuracy Score: 0.96

The accuracy measure is 0.96. Accuracy refers to the closeness of a measured value to a standard or true value. A value greater than 70% is a great model performance. The accuracy measure of the given data set is 0.96 thus the model is a good model.

**6.2.5.  Quadratic Discriminant Analysis :**

**Accuracy Measures:**

a). Accuracy Score: 0.947

The accuracy measure is 0.947. Accuracy refers to the closeness of a measured value to a standard or true value. A value greater than 70% is a great model performance. The accuracy measure of the given data set is 0.947 thus the model is good.

**6.2.6. Implementing Quadratic Discriminant Analysis with KFold Cross Validation**

**Accuracy Measures:**

Accuracy Score: 0.965

The accuracy measure is 0.965. Accuracy refers to the closeness of a measured value to a standard or true value. A value greater than 70% is a great model performance. The accuracy measure of the given data set is 0.965 thus the model is good.

**6.2.7. K-Neighbour’s Classifier Model**

K Neighbour’s classifier model is fitted on the given dataset considering the distance measure as Euclidean distance with a p-value of 2.

**Accuracy Measures:**

a). Accuracy Score: 0.99

The accuracy measure is 0.99. Accuracy refers to the closeness of a measured value to a standard or true value. A value greater than 70% is a great model performance. The accuracy measure of the given data set is 0.99 thus the model is good.

b). Classification Report:

precision recall f1-score support

0 1.00 0.98 0.99 50

1 0.98 1.00 0.99 64

accuracy 0.99 114

macro avg 0.99 0.99 0.99 114

weighted avg 0.99 0.99 0.99 114

Out of all the positive predictions: -

* Only 100% of Benign diagnoses are correctly expected to be patients with breast cancer.
* Only 98% of Malignant diagnoses are correctly expected to be patients with breast cancer.

Out of all the actual positives:

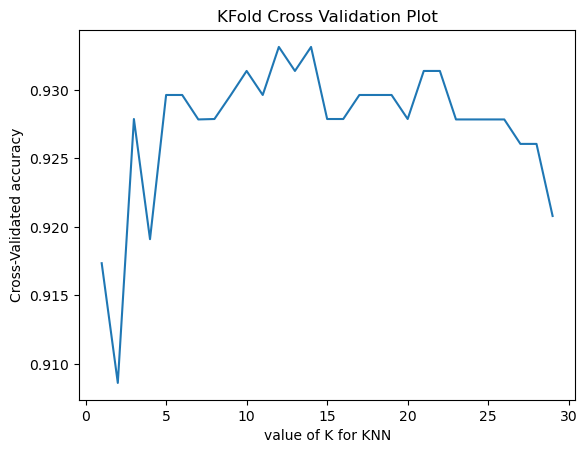
* 98% of Benign diagnoses are correctly expected to be patients with breast cancer.
* 100% of Malignant diagnoses are correctly expected to be patients with breast cancer.

The harmonic mean of precision and recall, with equal weights for each, is used to calculate the F-measure. This feature makes it possible to compare models and evaluate a model's performance by accounting for both precisions and recall with a single score.

A lower beta value favours less recall and favours precision more. Here, the weighted average is the percentage of observations in each class, and the macro average is the AM of the scores from the two classes. Since the categorisation is balanced, the macro average and weighted average are the same.

**6.2.8. K Neighbour’s Classification using K-Fold Cross Validation**

K Neighbour’s classifier model is fitted on the given dataset with K-Fold. cross-validation on these selected features and predictions are made. we choose the k value for which accuracy is max. Therefore accuracy in this case is maximum for k=11 and 14. After fixing k we fit the model again and we estimate the test accuracy.



**Accuracy Measures:**

Accuracy Score: 0.96

The accuracy measure is 0.96. Accuracy refers to the closeness of a measured value to a standard or true value. A value greater than 70% is a great model performance. The accuracy measure of the given data set is 0.96 thus the model is a good model.

**6.2.9. Gaussian Naive Bayes model**

Here the model consists of variables which are continuous in nature.

**Accuracy Measures:**

a). Accuracy Score: 0.89

The accuracy measure is 0.89. Accuracy refers to the closeness of a measured value to a standard or true value. A value greater than 70% is a great model performance. The accuracy measure of the given data set is 0.89 which is closer to 70% thus the model is suitable.

b). Classification Report:

precision recall f1-score support

0 0.83 0.88 0.85 50

1 0.90 0.86 0.88 64

accuracy 0.87 114

macro avg 0.87 0.87 0.87 114

weighted avg 0.87 0.87 0.87 114

Out of all the positive predictions: -

* Only 83% of Benign diagnoses are correctly expected to be patients with breast cancer.
* Only 90% of Malignant diagnoses are correctly expected to be patients with breast cancer.

Out of all the actual positives:

* 88% of Benign diagnoses are correctly expected to be patients with breast cancer.
* 86% of Malignant diagnoses are correctly expected to be patients with breast cancer.

The harmonic mean of precision and recall, with equal weights for each, is used to calculate the F-measure. This feature makes it possible to compare models and evaluate a model's performance by accounting for both precisions and recall with a single score.

A lower beta value favours less recall and favours precision more. Here, the weighted average is the percentage of observations in each class, and the macro average is the AM of the scores from the two classes. Since the categorisation is balanced, the macro average and weighted average are the same.

**6.2.10. Gaussian Naive Bayes model using K-Fold Cross Validation**

**Accuracy Measures:**

Accuracy Score: 0.94

The accuracy measure is 0.94. Accuracy refers to the closeness of a measured value to a standard or true value. A value greater than 70% is a great model performance. The accuracy measure of the given data set is 0.94 thus the model is a good model.

**6.2.11. Support vector Machine.**

In the code, we have used kernel='linear', as here we are creating SVM for linearly separable data. However, we can change it for non-linear data. And then we fitted the classifier to the training dataset(x\_train, y\_train). The model performance can be altered by changing the value of C(Regularization factor), gamma, and kernel.

**Accuracy Measures:**

a). Accuracy Score: 0.98

The accuracy measure is 0.98. Accuracy refers to the closeness of a measured value to a standard or true value. A value greater than 70% is a great model performance. The accuracy measure of the given data set is 0.98 which is closer to 1% thus the model has very good accuracy

b). Confusiuon Matrix:

[[49 1]

[ 1 63]]

In this data, we have 2 categories B and M. So we have a 2x2 confusion matrix. Here rows are test data, columns are classes obtained by the logistic model. There are 2 categories so 49 represents several observed values classified as **B**  belonging to 1st class, and 1 in the first row is several observed values classified as **B** but which is from class **M**. Similarly, for the next row 1 is the number of observations classified as **M** but belonging to 1st class, 63 is the number of observations classified as **M** belonging to the 2nd class.

**6.2.12. Support vector Machine using KFold Cross Validation**

**Accuracy Measures:**

Accuracy Score: 0.97

The accuracy measure is 0.97. Accuracy refers to the closeness of a measured value to a standard or true value. A value greater than 70% is a great model performance. The accuracy measure of the given data set is 0.97 thus the model is a good model.

# 6.2.13. Decision Tree

# 

**Accuracy Measures:**

a). Accuracy Score: 0.92

The accuracy measure is 0.92. Accuracy refers to the closeness of a measured value to a standard or true value. A value greater than 70% is a great model performance. The accuracy measure of the given data set is 0.92 thus the model has very good accuracy

b). Confusion Matrix:

precision recall f1-score support

0 0.89 0.94 0.91 50

1 0.95 0.91 0.93 64

accuracy 0.92 114

macro avg 0.92 0.92 0.92 114

weighted avg 0.92 0.92 0.92 114

Out of all the positive predictions: -

* Only 89% of Benign diagnoses are correctly expected to be patients with breast cancer.
* Only 95% of Malignant diagnoses are correctly expected to be patients with breast cancer.

Out of all the actual positives:

* 94% of Benign diagnoses are correctly expected to be patients with breast cancer.
* 91% of Malignant diagnoses are correctly expected to be patients with breast cancer.

The harmonic mean of precision and recall, with equal weights for each, is used to calculate the F-measure. This feature makes it possible to compare models and evaluate a model's performance by accounting for both precisions and recall with a single score.

A lower beta value favours less recall and favours precision more. Here, the weighted average is the percentage of observations in each class, and the macro average is the AM of the scores from the two classes. Since the categorisation is balanced, the macro average and weighted average are the same.

c) Confusion Matrix:

[[47 3]

[ 6 58]]

In this data, we have 2 categories B and M. So we have a 2x2 confusion matrix. Here rows are test data, columns are classes obtained by the logistic model. There are 2 categories so 47 represents several observed values classified as **B**  belonging to 1st class, and 1 in the first row is several observed values classified as **B** but which is from class **M**. Similarly, for the next row 1 is the number of observations classified as **M** but belonging to 1st class, 58 is the number of observations classified as **M** belonging to the 2nd class.

**6.2.14. Decision Tree using K-Fold Cross Validation:**

**Accuracy Measures:**

Accuracy Score: 0.94

The accuracy measure is 0.94. Accuracy refers to the closeness of a measured value to a standard or true value. A value greater than 70% is a great model performance. The accuracy measure of the given data set is 0.94 thus the model is a good model.

**7. Conclusion:**

The data is pre-processed and explorative data analysis is carried out to uncover the relationships between the variables and to extract important variables. Data visualization is performed to graphically interpret the various features and their significant relationships with the study variable which is the class of diagnosis. As the dataset is imbalanced Synthetic Minority Over Sampling analysis is executed to achieve a balanced data set. The number of features in this data set is 30 which is very high. So, principal component analysis is carried out to reduce the dimension of the feature set to 15 features which have a significant effect on the study variables. The features with a low effect on the response variable are eliminated.

Various models are fitted on this transformed data set including the

* Logistic regression model with K-Fold cross-validation and without cross-validation,
* Linear Discriminant Analysis with K-Fold cross-validation and without cross-validation,
* Quadratic Discriminant Analysis with K-Fold cross-validation and without cross-validation,
* Gaussian Naive Bayes model with K-Fold cross-validation and without cross-validation,
* K Nearest Neighbour’s model with K-Fold cross-validation and without cross-validation,
* Support Vector Machine with K-Fold cross-validation and without cross-validation,
* Decision Tree with K-Fold cross-validation and without cross-validation fitted to the given data set.

Among all these models K Nearest Neighbour’s model is the best suitable model for the given dataset as the accuracy score obtained is highest in this model which is 0.99. Thus, predictions based on K Nearest Neighbour’s model are considered for future forecasting as the model adequacy is high.

**8. References**

* K. P. Bennett and O. L. Mangasarian: "Robust Linear Programming Discrimination of Two Linearly Inseparable Sets", Optimization Methods and Software 1, 1992, 23-34
* <https://www.kaggle.com/datasets/uciml/breast-cancer-wisconsin-data>
* <https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29>