|  |  |
| --- | --- |
| Breast Cancer Detection Report  Machine Learning | Abstract  Report covers breast cancer classification (logistic regression, random forest) and clustering (K-means) using machine learning.  Syed Ali Hamza Shah  22102368 |

Table of Contents

[Introduction: 2](#_Toc156140615)

[Problem Statement in this report 2](#_Toc156140616)

[Logistic Regression Problem Statement: 2](#_Toc156140617)

[Random Forest Problem Statement: 2](#_Toc156140618)

[K-means Clustering Problem Statement: 2](#_Toc156140619)

[Domain and Data description 2](#_Toc156140620)

[Malignant (M) Cancerous Cells: 3](#_Toc156140621)

[Benign (B) Non-Cancerous Cells: 3](#_Toc156140622)

[y (Outcomes): 3](#_Toc156140623)

[x (Predictors): 3](#_Toc156140624)

[Exploratory Data Analysis (EDA) 3](#_Toc156140625)

[Dataset shape 4](#_Toc156140626)

[Data Cleaning 4](#_Toc156140627)

[Descriptive statistics 5](#_Toc156140628)

[Correlation Analysis 5](#_Toc156140629)

[Splitting the Dataset 8](#_Toc156140630)

[Methodology and Model Development 8](#_Toc156140631)

[Logistic Regression Definition: 8](#_Toc156140632)

[Explanation in the Context of the Code: 9](#_Toc156140633)

[Random Forest Definition: 9](#_Toc156140634)

[Explanation of the Code: 9](#_Toc156140635)

[K-Means: 9](#_Toc156140636)

[Explanation of the Code: 10](#_Toc156140637)

[Model Evaluation and Results 11](#_Toc156140638)

[Logistic Regression and Random Forest accuracy: 11](#_Toc156140639)

[Accuracy by Confusion Matrix 12](#_Toc156140640)

[Logistic Regression: 12](#_Toc156140641)

[Random Forest: 13](#_Toc156140642)

[Metrics on the model 13](#_Toc156140643)

[Hierarchical Clustering and Dendrogram Visualization for k-means: 14](#_Toc156140644)

[Silhouette Score in K-Means: 15](#_Toc156140645)

[Conclusion 16](#_Toc156140646)

[Appendix A 17](#_Toc156140647)

[Works Cited 20](#_Toc156140648)

[Code Link 21](#_Toc156140649)

# Introduction:

This report addresses the classification and clustering of the data used in breast cancer dataset using various methods of machine learning. Logistic regression model aims to predict whether a tumour is malignant or benign based on clinical features, emphasizing interpretability and accuracy (Tingting Li, 2022). The random forest model, leveraging an ensemble of decision trees, seeks robustness and accuracy in predicting malignancy (Sutong Wang, 2020). Additionally, K-means clustering explores patterns in patient demographics and clinical characteristics, identifying distinct subgroups for potential personalized treatment strategies (Ashutosh Kumar Dubey, 2016). The report provides insights into data exploration, cleaning, and visualization, emphasizing key features such as mean radius, texture, and perimeter. Exploratory data analysis, correlation analysis, and the elbow method for optimal cluster identification contribute to a comprehensive understanding of the dataset. Finally, the evaluation of machine learning models includes training accuracy, confusion matrices, and key metrics like precision, recall, and F1-score, enabling a thorough comparison of logistic regression and random forest models. The report concludes with an analysis of hierarchical clustering and silhouette scores, demonstrating K-means' effectiveness in revealing well-defined clusters within the breast cancer dataset.

Problem Statement in this report:

Logistic Regression Problem Statement:

Developing a predictive model using logistic regression to classify breast tumours as either Malignant (cancerous) or Benign (non-cancerous) based on features derived from clinical and diagnostic data. (M Kalaiyarasi, 2020 ) The goal is to create an accurate and interpretable model that aids in the early and reliable identification of malignant tumours, facilitating timely medical intervention and improving overall patient outcomes. This analysis will leverage a dataset containing information on tumour characteristics, patient demographics, and other relevant factors to train a logistic regression model capable of distinguishing between the two tumour types with high sensitivity and specificity.

Random Forest Problem Statement:

Building a random forest classification model to predict the likelihood of malignancy in breast tumours using a dataset comprising features like Mean radius, perimeter, texture of the tumour cells, Standard error of the radius, perimeter, tumour texture of the cells, Worst (major) radius, perimeter, texture of the tumour cells and many other factors. The goal is to create an accurate and robust model that distinguishes between benign and malignant tumours, assisting in effective early-stage diagnosis.

## K-means Clustering Problem Statement:

Exploring the segmentation of breast cancer patients based on their demographic and clinical characteristics using K-means clustering. (Ashutosh Kumar Dubey, 2016) Identify distinct groups or clusters among patients, considering factors like Mean radius, perimeter, texture of the tumour cells and Standard error of the radius, perimeter, tumour cells texture, Worst radius, tumour texture cells and many other features. This analysis aims to uncover potential patterns or subgroups within the patient population that might aid in personalized treatment strategies or risk assessment.

# Domain and Data description

Breast cancer is a common and potentially fatal disease affecting around 20% of women globally. (Muhammad Yasin Mohsin, 2021) Early detection is crucial for effective treatment, as delayed diagnosis can lead to severe consequences. The dataset used in this study is from [kaggle](https://www.kaggle.com/datasets/utkarshx27/breast-cancer-wisconsin-diagnostic-dataset) (SINGH, 2023), which includes information about 1707 breast lumps, categorizing them as either cancerous (malignant) or not (benign).

In the context of cancer, particularly breast cancer, "Malignant" (M) and "Benign" (B) refer to the two main categories that describe the nature and behaviour of tumour cells:

## Malignant (M) Cancerous Cells:

- These cells are characterized by their ability to grow uncontrollably and invade nearby tissues or organs.

- Malignant cells can potentially metastasize, spreading to other body parts through the bloodstream, forming secondary tumours.

- They are considered cancerous and pose a more serious health threat due to their invasive nature and potential to cause significant harm.

## Benign (B) Non-Cancerous Cells:

- Benign cells, on the other hand, cannot attack close tissues of body’s further portions.

- They typically grow slower and tend to have a well-defined boundary or encapsulation.

- While they are abnormal in terms of growth or appearance, they are not considered cancerous and generally do not pose a significant threat to health compared to malignant cells.

The dataset reports the average, standard deviation, and the worst value, making up 30 features. There is detailed description of each property use in the table 1 (Appendix A).

## y (Outcomes):

This represents the target variable or the outcome we want to predict. It has the factor with two levels:

- **M** means malignant (cancerous).

- **B** means benign (not cancerous).

## x (Predictors):

This is the collection of variables that were utilised to forecast the result. A feature matrix taken from biopsy samples is included in the dataset. The mean, standard error, and worst value of ten nuclear measurements make up the thirty features that each biopsy has (SINGH, 2023). These nuclear measurements describe the following characteristics of cell nuclei: radius (the average distance between the centre and points on the nucleus' perimeter), texture (the nucleus' standard deviation of grayscale values), perimeter, area, smoothness (the narrow variation in the nucleus' radius lengths), compactness (the measurement of the nucleus' compactness), concavity (the degree to which the nucleus's contour is concave), concave pts (the quantity of concave portions in the nucleus' contour), symmetry, and fractal dim (the nucleus' fractal dimension, approximating the "coastline.")

Conclusively, each biopsy consist of mean, standard error, and worst value for each of the ten nuclear measures of total 30 feature, and main objective is to use them for prediction of the breast cancer.

# Exploratory Data Analysis (EDA)

EDA is vital for understanding this dataset and uncovering patterns or insights. Here are some key components considered for this dataset:

## Dataset shape

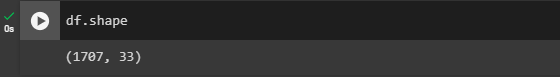


Figure 1

`df shape` helps to find the sizes of a Data Frame, where `df` is assumed to be the variable name representing the dataset. The result 1707, 33 indicates that the Data Frame has 1707 rows and 33 columns.

## Data Cleaning



Figure 2

`df isnull ().sum () helps to sum the amount of missing values in all column of the dataset (Vikash Kumar Mishra, 2023). It provides the count of missing values for each column (figure 2). For most columns (e.g., 'id', 'diagnosis', 'radius mean', etc.), the count is 0. This means that these columns do not have any missing values. The column 'unnamed: 32' has 1707 missing values. This suggests that all values in this column are missing, as the total count matches the number of rows in your dataset (1707). The column doesn't contribute valuable information and consists entirely of missing values. It has been removed from the dataset by using df.dropna (figure 3) to simplify further analysis.

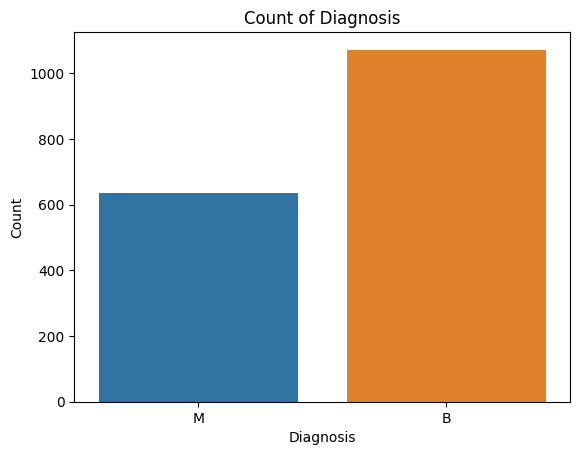


Figure 3

## Descriptive statistics

According to the dataset Benign (B) instance, projected in the Diagnosis count are higher than Malignant (M) as projected in figure 4.

B (Benign): There are 1071 instances in the dataset classified as benign, indicating non-cancerous breast masses.

M (Malignant): Are 636 instances classified as malignant, indicating cancerous breast masses.

The dataset has most benign cases (1071) and fewer malignant cases (636). Understanding the distribution of the target variable ('diagnosis') is crucial for assessing the balance of classes in this dataset, which can impact the training and evaluation of machine learning models.

To make this M and B understand for the code, use the `Label Encoder` from the scikit-learn library to encode the categorical values (M and B) in the 'diagnosis' column of the Data Frame. The code snippet performs label encoding, replacing categorical values with numerical equivalents (0 for benign and 1 for malignant).

### Correlation Analysis

This analysis measures the strength and relationship among 2 variables. It provides the comprehensive effects of one variable over other. Some analysis are mentioned below:

#### Pair plot analysis

Pair plot analysis shows the graphical relationship of one feature over another feature. As there are more than 30 variables in this dataset so for simplicity this report uses 5 variables to understand the correlation.

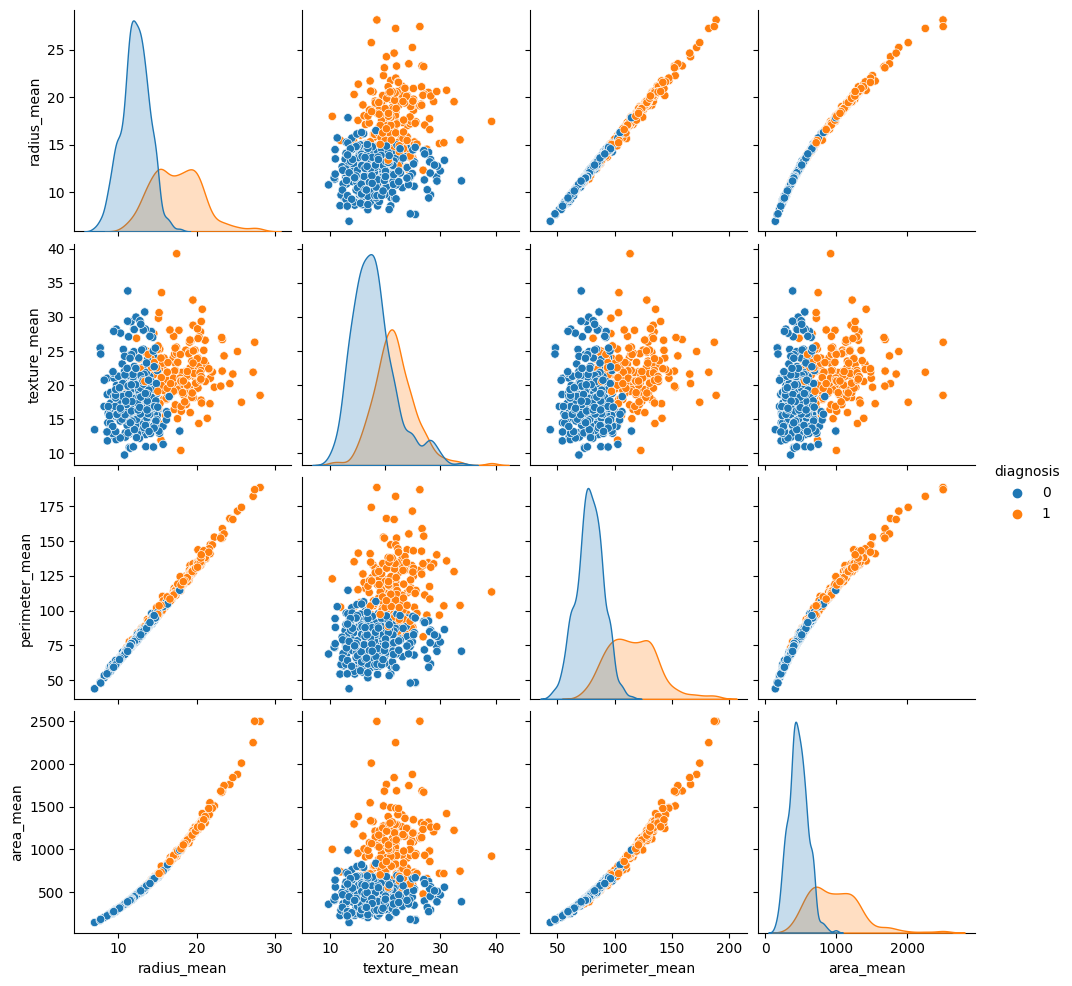


Figure 4

Here are some of the things that we can see from the pair plot:

A positive correlation exists between radius mean and texture mean, perimeter mean, and area mean. This means that as the values of these variables increase, the value of radius mean also tends to increase.

There is a negative correlation between radius mean and smoothness mean. This means that as radius mean increases, smoothness mean tends to decrease.

The two clusters of points in the scatter plots for radius mean to texture mean and mean radius to perimeter mean suggest that there may be two distinct groups of data points in the dataset. This is because the "diagnosis" variable is binary.

The scatter plots for the other variables (texture mean vs. perimeter mean, etc.) do not show a clear pattern.

Conclusively, the pair plot suggests that there may be some relationships between the variables in the dataset.

#### Heat Map Analysis

Visualize the correlation among variables and effectively displays the correlation of multiple variable in a grid. (Navigation, 2016)

#### 

Figure 5

There are distinct blocks of positive and negative correlations. The upper left corner displays strong positive correlations, with darker reds indicating the strongest correlations (around 0.8 or higher). These appear to be between radius, perimeter, area, and texture features. In contrast, the lower right corner shows strong negative correlations (dark blue cells) between smoothness and other variables, particularly radius and compactness.

The diagonal, as expected, shows perfect positive correlations (1.0) for each variable with itself.

##### Strongest correlations:

The strongest positive correlations (darkest red) are observed between:

##### Means of the perimeter and radius (0.92)

##### The area mean and radius mean (0.90)

##### Area mean and perimeter mean (0.89)

##### The following pairs have the strongest negative association (darkest blue):

##### The mean of smoothness and compactness (-0.78)

##### Variable relationships:

Radius, perimeter, area, and texture features seem highly correlated, suggesting they might capture similar information about the tumours.

Smoothness mean has distinct negative correlations with other features, particularly compactness mean, potentially indicating rougher textures associated with denser tumours. Other correlations, while not as strong.

# Splitting the Dataset

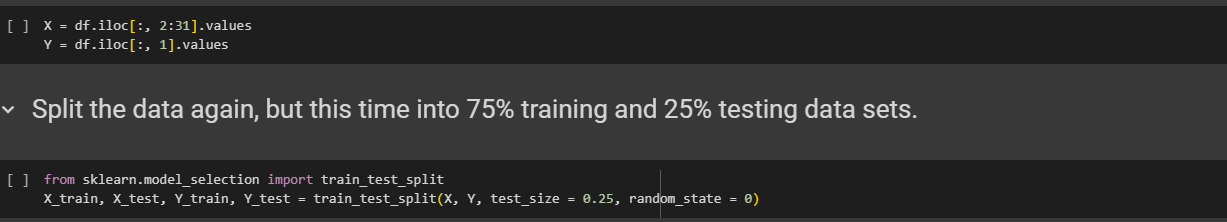
The dataset is segmented into 2 categories. Prepares the breast cancer dataset for machine learning by separating it into feature sets (X) and target sets (Y) and train test split function used to split these sets into testing and training subsets. The machine learning models are trained on the training set (X train, Y train), and their performance on unseen data is assessed on the testing set (X test, Y test) (Vikash Singh, 2021). As seen in figure 6, divide the dataset into two parts: 75% of the training data set model and 25% of the testing data set model.

Figure 6

# Methodology and Model Development

Experiment and evaluation of machine learning techniques involve applying and assessing the performance of different algorithms on a breast cancer dataset. This study utilized two supervised learning techniques, Logistic Regression and Random Forest, along with one unsupervised learning technique, K-Means. The goal is to analyse and compare the effectiveness of these methods in handling the specific benign and malignant classification of the breast cancer data.

## Logistic Regression Definition:

The goal variable is categorical and has two alternative outcomes, supervised learning techniques like logistic regression is best to utilize because it used to solve binary classification problems using statistical methods (Tingting Li, 2022). It maps the output to the range [0, 1] by using a logistic function (sigmoid) to forecast the likelihood that an instance would belong to a specific class. The threshold value, usually 0.5, is used to determine the decision limit. The instance is categorised as one class if the anticipated probability is higher than this threshold; if not, it is classified as the other class.

### Explanation in the Context of the Code:

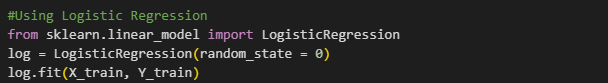


Figure 7

In the figure 7 it is showing how logistic regression has been implemented with the code that imports the Logistic Regression class from scikit learns `linear model` module. Then a logistic regression model is instantiated and assigned to the variable log. By fixing the random seed, the ‘random state=0’ parameter guarantees reproducibility. Using the fit method, trained on the input training data X train - feature set, Y train - target set.

Conclusively, Logistic Regression in this context is employed as a binary classification model to predict whether a cancer is M or B.

## Random Forest Definition:

A supervised learning technique called Random Forest uses an ensemble learning algorithm to build several decision trees during training (Sutong Wang, 2020). The system then outputs a class that is either the mean prediction (regression) of the individual trees or the mode of the classes (classification). A random subset of the features is used to build each tree in the forest, and the combined forecasts of all the individual trees determine the final prediction.

Random Forest is renowned for its resilience, adaptability, and capacity to manage datasets with multiple dimensions. Combining the strengths of several decision trees helps to lessen over fitting and increase overall accuracy.

### Explanation of the Code:

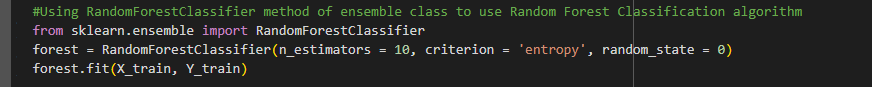


Figure 8

Figure 8 shows how Random Forest has been implemented with the code that imports the `Random Forest Classifier` class from scikit learns `ensemble` module. ‘Random forest classifier’ instance created. `n estimators=10`: States the amount of decision trees in the forest (10).`criterion='entropy'`: The criterion for splitting nodes is set to 'entropy', which measures the information gain. Using the ' fit ' method, on the input training data this model is trained.

The Random Forest model is applied to forecast either a breast mass is M or B based on the features such as radius, texture, perimeter, and others. The number of tree are controlled by ‘n estimator’ in the code, and entropy criterion is used to guide the tree-building process based on information gain. The `random state` parameter ensures consistent results when re-running the code.

## K-Means:

An unsupervised machine learning approach called K-Means is used for clustering, or putting comparable data points in one group. The purpose of K-Means is to divide a dataset into 'k' clusters, each of which contains a data point that belongs to the cluster with the closest mean (Ashutosh Kumar Dubey, 2016).

The algorithm is iterative in nature. In the feature space, it initializes 'k' centroids, or representative points, at random. Each data point is then allocated to the cluster with the nearest centroid. Following assignment, the centroids are updated by calculating the average of the data points for each cluster. Assignment and updating are done again until convergence, which is typically reached when the centroids hardly vary at all.

### Explanation of the Code:

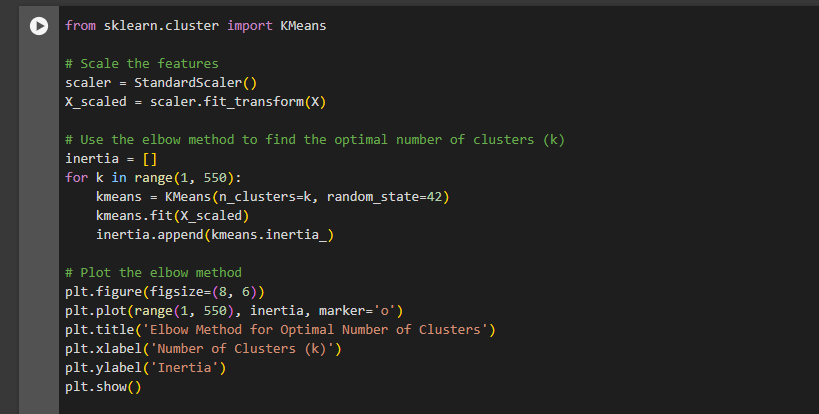


Figure 9

Imports the `K Means` class from scikit learn's `cluster` module. Imports the `Standard Scaler` class for feature scaling. Instantiates a `Standard Scaler` to scale the features. Scales the feature set `X` using the `fit transform` method.

#### Elbow Method:

The Elbow Method determines a clustering algorithm's optimal amount of clusters (k). The point on the plot on which the decline in inertia begins to slow down (forming an elbow-like shape) is considered the optimal k value, indicating a balanced trade-off between model complexities and clustering quality. (M A Syakur1, 2018) This study mentioned in figure (9) uses a loop to run the K-Means algorithm with diverse values of 'k' and shows inertia (sum of squared distances) for each iteration. Iterates over a range of possible 'k' values.

.

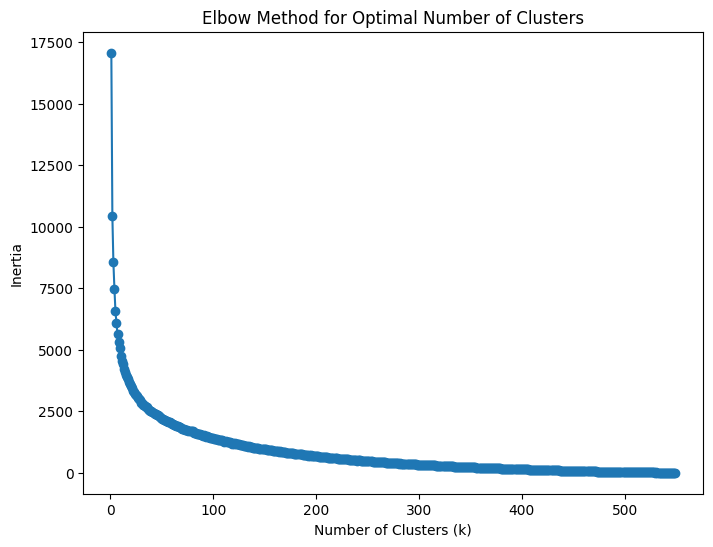


Figure 10

As shown in figure 10, the result is a list (`inertia`) that can be used to plot an elbow curve to determine the best number of clusters on breast cancer dataset. The point on the curve where the inertia starts to level off indicates a suitable value for 'k'.

# Model Evaluation and Results

Model accuracy on the training data represents the proportion of correctly predicted instances by the model among all instances in the training dataset. It is computed by dividing the total number of cases in the training set by the number of successfully classified instances.

## Logistic Regression and Random Forest accuracy:

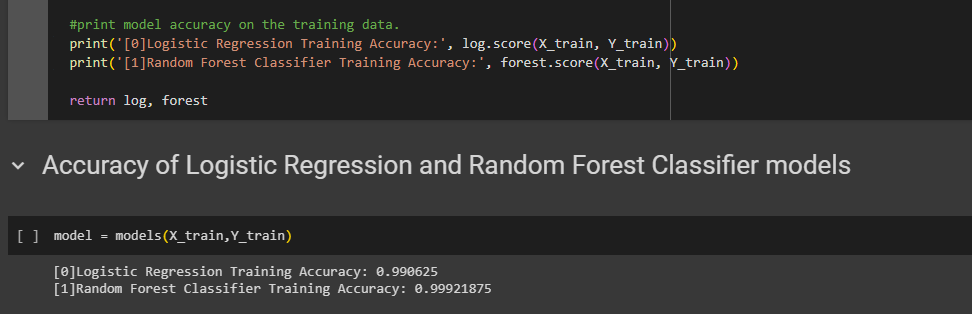


Figure 11

* Logistic Regression, the training accuracy is 99.06% (0.990625), indicating that the Logistic Regression model properly predicted the target variable for about 99.06% of the instances in the training data.
* Random Forest Classifier, the training accuracy is 99.92% (0.99921875), indicating that the Random Forest model correctly predicted the target variable for approximately 99.92% of the instances in the training data.

As it is seen from Figure 11, High training accuracy suggests that the models perform well on the data they were trained on.

## Accuracy by Confusion Matrix

Model accuracy on test data indicates how well the machine learning model functions on never-before-seen data from the test set. It is quantified using a confusion matrix (Ratih Prasetya, 2019). We can assess the model's efficacy because the confusion matrix offers a thorough analysis of its predictions.

### Logistic Regression:

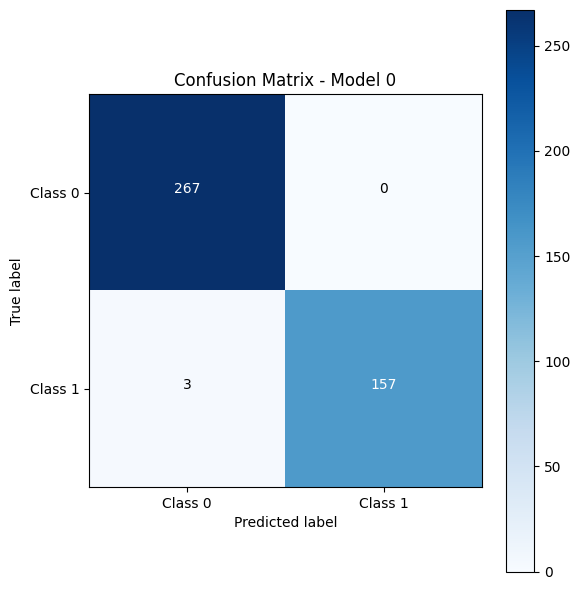


Figure 12

The confusion matrix shows in figure (12) describes performance of a logistic regression model classifying breast cancer tumours from a dataset.

The model excels at identifying true positives and true negatives, suggesting it can effectively distinguish between malignant and benign tumours in most cases.

The main weakness lies in the false negatives (2 cases).

The confusion matrix suggests that the logistic regression model performs well in classifying breast cancer tumours with high accuracy, precision, and recall. However, there's always room for improvement, particularly in minimizing false negatives.

### Random Forest:

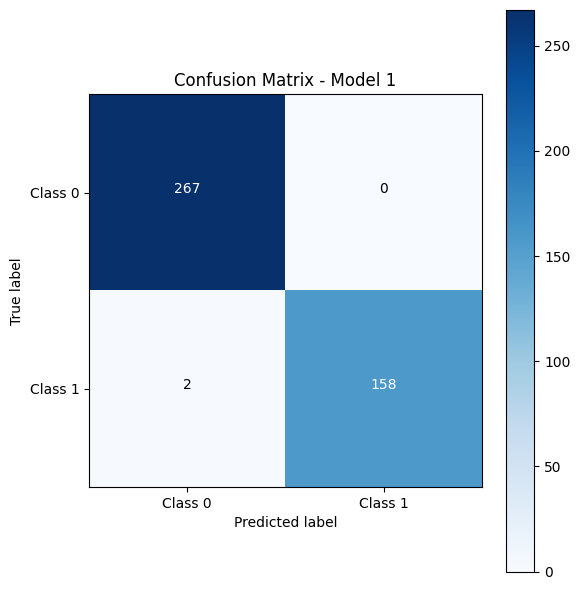


Figure 13

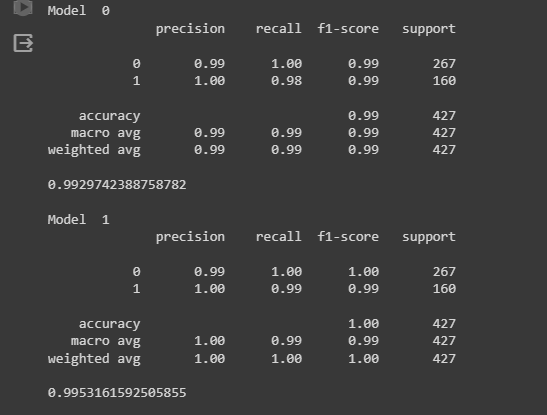
The confusion matrix presents in figure 13 the performance of a Random Forest model classifying breast cancer tumours from a dataset. It is performing well, Similar to the logistic regression model, this Random Forest model excels at identifying true positives and true negatives, demonstrating strong differentiation between malignant and benign tumours in most cases. Again, the main weakness is only the false negatives (2 cases).

The Logistic Regression and Random Forest models demonstrate promising performance in classifying breast cancer tumours, with high accuracy, precision, and recall. However, minimizing false negatives remains a key area for improvement in both models.

## Metrics on the model

Evaluates and reports performance metrics for each model on the test data, the model used are Logistic regression (0) and Randon Forest(1), aiding in the comparison and selection of the most suitable model for the accuracy evaluation of each task as available on table 2.

Table 2



• Precision: The fraction of correct positive forecasts.

• Recall: The percentage of real positives that are appropriately recognized.

• F1-score: Indicates overall performance by combining recall and precision into a single metric.

• Accuracy: The percentage of total accurate forecasts.

• Support: The quantity of each class's occurrences in the test dataset.

Both models demonstrate excellent performance on the test dataset. Random Forest (Model 1) appears to edge out Logistic Regression (Model 0) with marginally better metrics.

It's essential to consider the specific application and the relative importance of precision and recall when selecting the "best" model.

## Hierarchical Clustering and Dendrogram Visualization for k-means:

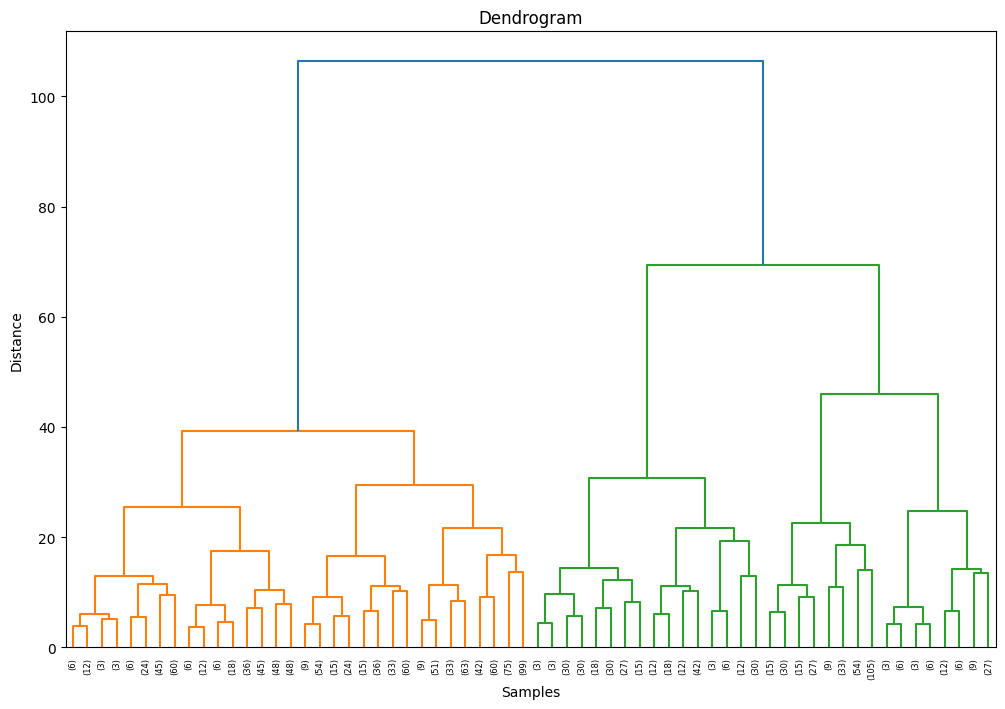
An approach for clustering data called hierarchical clustering creates a hierarchy of clusters (Anna D. Peterson, 2018). Hierarchical clustering begins with individual data points and iteratively merges or separates clusters depending on a similarity metric, without predetermining the number of clusters. As seen in figure 14, this procedure produces a dendrogram, or tree-like structure, which depicts the links and hierarchy between groups.

Figure 14

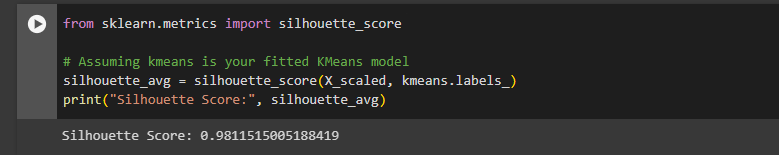
The dendrogram is particularly useful for understanding the hierarchy and structure within the data. It allows users to identify optimal cluster divisions or determine the appropriate number of clusters based on the desired granularity.

Look for major branch mergers occurring closer to the bottom of the dendrogram. These likely represent initial, well-defined clusters. Observe how smaller clusters merge into larger ones. This provides insights into how different sub-groups relate and share some similarities.

## Silhouette Score in K-Means:

An indicator of a clustering technique's quality, such as K-Means in unsupervised learning, is the Silhouette Score (Ratih Prasetya, 2019). It measures the degree of cluster separation and the coherence of the data points inside the clusters to which they have been assigned. A high positive score (around 1) denotes that the data points are poorly matched to nearby clusters and well-matched to their own clusters. The Silhouette Score is a number between -1 and 1.

Overlapping clusters are indicated by a score close to 0. A low score suggests that the data points may have been mislabelled to the incorrect cluster. Better-defined clusters are typically indicated by a higher Silhouette Score.



A high **silhouette score of 0.981**, which is very near to 1, is indicated. Demonstrates that the K-Means method created well-defined clusters, and that the data points inside each cluster are more comparable to one another than they are to those outside of it.

# Conclusion

In conclusion, this comprehensive analysis of breast cancer data showcases machine learning techniques' efficacy in classification and clustering tasks. The logistic regression and random forest models exhibit high training accuracy, demonstrating their ability to predict tumour malignancy accurately based on clinical features. Confusion matrices and metrics like precision and recall further emphasize the models' proficiency distinguishing between benign and malignant tumours. The K-means clustering approach provides valuable insights into patient segmentation, potentially contributing to personalized treatment strategies. Exploratory data analysis, correlation analysis, and visualization techniques enhance our understanding of critical features in the dataset. Overall, combining supervised and unsupervised learning techniques contributes to a robust understanding of breast cancer characteristics, facilitating accurate prediction and insightful patient subgroup identification.

# Appendix A

Table 1

| **Column** | **Description** |
| --- | --- |
| x.radius\_mean | Mean radius of the tumor cells |
| x.radius\_mean | Mean radius of the tumor cells |
| x.texture\_mean | Mean texture of the tumor cells |
| x.perimeter\_mean | Mean perimeter of the tumor cells |
| x.area\_mean | Mean area of the tumor cells |
| x.smoothness\_mean | Mean smoothness of the tumor cells |
| x.compactness\_mean | Mean compactness of the tumor cells |
| x.concavity\_mean | Mean concavity of the tumor cells |
| x.concave\_points\_mean | Mean number of concave portions of the contour of the tumor cells |
| x.symmetry\_mean | Mean symmetry of the tumor cells |
| x.fractal\_dimension\_mean | Mean "coastline approximation" of the tumor cells |
| x.radius\_se | Standard error of the radius of the tumor cells |
| x.texture\_se | Standard error of the texture of the tumor cells |
| x.perimeter\_se | Standard error of the perimeter of the tumor cells |
| x.area\_se | Standard error of the area of the tumor cells |
| x.smoothness\_se | Standard error of the smoothness of the tumor cells |
| x.compactness\_se | Standard error of the compactness of the tumor cells |
| x.concavity\_se | Standard error of the concavity of the tumor cells |
| x.concave\_points\_se | Standard error of the number of concave portions of the contour of the tumor cells |
| x.symmetry\_se | Standard error of the symmetry of the tumor cells |
| x.fractal\_dimension\_se | Standard error of the "coastline approximation" of the tumor cells |
| x.radius\_worst | Worst (largest) radius of the tumor cells |
| x.texture\_worst | Worst (most severe) texture of the tumor cells |
| x.perimeter\_worst | Worst (largest) perimeter of the tumor cells |
| x.area\_worst | Worst (largest) area of the tumor cells |
| x.smoothness\_worst | Worst (most severe) smoothness of the tumor cells |
| x.compactness\_worst | Worst (most severe) compactness of the tumor cells |
| x.concavity\_worst | Worst (most severe) concavity of the tumor cells |
| x.concave\_points\_worst | Worst (most severe) number of concave portions of the contour of the tumor cells |
| x.symmetry\_worst | Worst (most severe) symmetry of the tumor cells |
| x.fractal\_dimension\_worst | Worst (most severe) "coastline approximation" of the tumor cells |
| Y | target |

# Works Cited

Anna D. Peterson, A. P. (2018). Merging K-means with hierarchical clustering for identifying general-shaped groups. *WILEY online library*.

Ashutosh Kumar Dubey, U. G. (2016). *Analysis of k-means clustering approach on the breast cancer Wisconsin dataset.* Rajasthan: International Journal of Computer Assisted Radiology and Surgery.

M A Syakur1, B. K. (2018). Integration K-Means Clustering Method and Elbow Method For Identification of The Best Customer Profile Cluster. *IOP Publishing Ltd*.

M Kalaiyarasi, R. D. (2020 ). *Classification of Benign or Malignant Tumor Using Machine Learning.* Erode: IOP Publishing Ltd.

Muhammad Yasin Mohsin, M. R. (2021). *Accuracy Improvement for the Diagnosis of Breast.* Islamabad: IEEE.

Navigation, A. (2016). Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Oxford University Press*, 2847-2849.

Ratih Prasetya, A. R. (2019). *Data Mining Application on Weather Prediction Using Classification Tree, Naïve Bayes and K-Nearest Neighbor Algorithm With Model Testing of Supervised Learning Probabilistic Brier Score,Confusion Matrix and ROC .* Jakarta: JAICT Journal of Applied Communication and Information Technologies.

SINGH, U. (2023, May). *Breast Cancer Wisconsin Diagnostic Dataset*. Retrieved from Kaggle: https://www.kaggle.com/datasets/utkarshx27/breast-cancer-wisconsin-diagnostic-dataset

Sutong Wang, Y. W. (2020). *An improved random forest-based rule extraction method for breast cancer diagnosis.* ELSEVIER.

Tingting Li, Y. L. (2022). *Logistic regression analysis of ultrasound findings in predicting the malignant and benign phyllodes tumor of breast.* Chengdu: PLOS ONE.

Vikash Kumar Mishra, S. S. (2023). *Dealing with Missing Values in a Relation Dataset Using the DROPNA Function in Python.* Uttar Pradish: Wiley .

Vikash Singh, M. P. (2021). *Impact of train/test sample regimen on performance estimate stability of machine learning in cardiovascular imaging.* Los Angeles: Scientific Reports volume.

# Code Link

<https://colab.research.google.com/drive/1Xbq74utSB854gKuksXAWqSS7_3Ux56i2?usp=sharing>