Table of Contents

[Abstract 2](#_Toc174401351)

[Acknowledgements 2](#_Toc174401352)

[Chapter 1: Introduction 2](#_Toc174401353)

[I. Overview 2](#_Toc174401354)

[II. Problem Statement 2](#_Toc174401355)

[III. Research Questions 2](#_Toc174401356)

[Chapter 2: Background 2](#_Toc174401357)

[I. Dataset 2](#_Toc174401358)

[II. Literature Review 2](#_Toc174401359)

[III. Research Gap 2](#_Toc174401360)

[IV. Algorithms 2](#_Toc174401361)

[a. TFIDF 2](#_Toc174401362)

[b. Count Vectorizer 2](#_Toc174401363)

[c. Logistic Regression 2](#_Toc174401364)

[d. SVM 2](#_Toc174401365)

[e. XGBoost 2](#_Toc174401366)

[f. Decision Tree 2](#_Toc174401367)

[g. Artificial Neural Network 2](#_Toc174401368)

[Chapter 3: Methodology 2](#_Toc174401369)

[I. Tools and Techniques 2](#_Toc174401370)

[II. EDA and Visualization 2](#_Toc174401371)

[Chapter 4: Results and Conclusion 2](#_Toc174401372)

[I. Critical Discussion 2](#_Toc174401373)

[II. Conclusion 2](#_Toc174401374)

[III. Future Work 2](#_Toc174401375)

[Chapter 5: Legal, Ethical and Professional Issues 2](#_Toc174401376)

[References 2](#_Toc174401377)

[Appendices 2](#_Toc174401378)

# Abstract

# Acknowledgements

# Chapter 1: Introduction

## Overview

Breast cancer is one of the most common types of diseases in women. Early diagnosis is one of the major factors that help in avoiding fatal outcomes. The diagnosis of breast cancer is traditionally done using histopathological analysis of the biopsied tissue. This process takes a lot of time, effort and requires a lot of specialized medical expertise. With the exponential growth in the performance of Machine Learning and Artificial Intelligence this process can be sped up to provide a lot of support to the medical practitioners in this field (Barth, 2024).

In this research a carefully curated dataset called “Wisconsin Breast Cancer” data is taken from UCI Machine Learning repository (Karanam, 2022). ML and AI can be used to analyse this data and understand which feature is affecting the predictions from these algorithms. Understanding the importance of each feature is very important as it can help to enhance model accuracy and provide further insights into the behaviour of the tumour itself and also about the tumour’s progression. Another approach is to use unsupervised learning techniques like clustering that have the capability to reveal the inherent structures in the data that may not be very apparent to a human viewing the data. By grouping patients together, it is possible to understand if a group of patients have a feature that all have same type of target variable classification.

In this way the application of ML in the diagnosis can help to understand various aspect of the disease and help to provide more insights into the behaviour of the tumour.

## Research Questions

1. What are the most important features for classifying malignant tumours from benign tumours using Wisconsin breast cancer data?
2. Can unsupervised learning algorithms be used to detect clusters among the data that have an association with breast cancer?

# Chapter 2: Background

## Dataset

As mentioned earlier the dataset used is this research is the Wisconsin Breast Cancer data. This dataset has information regarding the breast tissue of various patients with tumours. Overall, this dataset has around 550 observations and 33 features with various attributes, represent multiple attributes of the tumour in patients. This dataset contains a target variable that has two classes – benign and malignant. Some of the key features in the dataset include perimeter, smoothness, and concavity. The data also includes missing values which need to addressed going further (*Breast cancer Wisconsin (diagnostic) dataset*, 2024).

## Literature Review

The paper will also seek to establish the efficiency of each of the said clustering algorithms in the diagnosis of breast cancer particularly in differentiating between benign and malignant tumors. It uses the Wisconsin Breast Cancer dataset, which contains 699 instances with nine integer-valued attributes and is split into two classes: Benign and malignant. The study employs five different clustering algorithms: , there are DBSCAN, Farthest First, Canopy, LVQ, and Hierarchical clustering, with the experimentation done using Weka. The major objective of the study is to assess and compare the reliability of these clustering algorithms for the prognosis of outcomes of breast cancer.

The use of the Weka, a data mining tool in implementing and testing the clustering algorithms forms the basic methodology of the study. The data used in this paper is the Wisconsin Breast Cancer dataset, on which each of the above algorithms is run as a way of assessing their efficiency. The studies of each clustering technique are quantified and assessed to identify the potential approach to the diagnosis of breast cancer. The outcomes indicated that there is significant improvement on the accuracy of the Farthest First algorithm which gets an accuracy of 72% compared to other clustering methods.

The study is presented from a procedural perspective, starting with the extraction of the data from the source which in this case was the UCI Machine Learning Repository; preprocessing and last but not least the clustering algorithms. DBSCAN is another clustering technique, which aims at clustering on the basis of the density, which is available in the given region. The strength of DBSCAN is in the detection of non-spherical clusters and the indifference to noise, however, sets containing a large difference between cluster densities of areas are problematic for DBSCAN. Farthest First clustering is a simple modification of k-means whereby the initial cluster centers are located at the greatest distance from existing centers and is faster in clustering than standard k-means because reassignment and adjustment are infrequent. The kind of clustering that is used prior to k-means or hierarchy for example is the canopy clustering which involves placing data in clusters based on two distance parameters, T1 and T2. LVQ which is a competitive learning algorithm, group data items where the clusters are gained by learning the center of the clusters, and associating new data items to the nearest cluster. Hierarchical clustering is a type of clustering that forms a tree structure where each node represents a cluster, its techniques can combine clusters to form new clusters, or it can split larger clusters into smaller clusters leading to a formation of dendrogram to represent the clusters.

Farthest First method has a better average Euclidean distance compared to Near Neighbors method, it produces a correct prediction of 72%. This is an important result because it can be seen that Farthest First is more appropriate for clustering of the data set namely Wisconsin Breast Cancer. Still, the paper gives little regard to why Farthest First is superior to the other approaches and when Farthest First is likely to be more or less potent. Further, the study does not assess how different clustering techniques are affected by parameters in order to get better results, that may be significant.

The paper seems to fail to consider other critical measures that would give a better comparison of the effectiveness of the clustering algorithms. For instance, measures like precision, recall, the F-measure, or the F1 score, which are important in diagnosing the efficiency of a diagnostic tool in a medical setting, are omitted. The lack of such measurements makes the work insufficient to give a proper evaluation of the clustering algorithms. In addition, the question of how to deal with missing values that are typical for real-life medical datasets is beyond the scope of the given study. The paper also briefly notes that the further research would focus on the missing data, but it does not offer any concrete approaches and methods regarding such data.

The analysis is relevant to the use of clustering algorithms for diagnosing breast cancer, including the proposed Farthest First algorithm. Nevertheless, it would be better if the factors that influenced the performance of the clustering algorithm have been discussed more expansively together with having more evaluation criteria to sustain the assessment. Moreover, the problem of deletion of some data and understanding the impact of various parameters of the algorithms to be used would strengthen the study and make the results useful for real-life application. They also reveal the directions for choosing the suitable clustering methodologies in medical diagnosis and the potential for the further investigations in this sphere.

Uniquely, this paper addresses the problem of breast cancer which has become the second cause of death among women through a predictive model employing machine learning. It uses K-means clustering to classify breast cancer data and analyses the probability of cells being cancerous and Decision Trees classification. This model leverages the Wisconsin Breast Cancer dataset from the UC Irvine Machine Learning Repository, which includes 699 instances with 11 distinct attributes, 9 of which describe the physical characteristics of cell nuclei present in a digitized image of a fine needle aspirate (FNA) of a breast mass. The dataset is categorized into two classes: The benign and the malignant types of the tumor where 458 benign patients and 241 patients with malignant tumor reported.

The following is the breakdown of the theoretical framework used to develop the methodology: Clustering is one of the unsupervised learning techniques that aim at categorizing data into clusters that are most similar. The K-means algorithm adopts a process of clustering of data that is by minimizing the variance between the clusters. In the present work, the given algorithm is applied with the help of orange tool which is the environment for data analysis and visualization. The data is then fed into the orange environment where the K-mean clustering procedure is applied so as to cluster the instances. The clustering process is illustrated through scatter plots, which makes it possible for example to analyze how one or another attribute correlates with the presence of cancerous cells. For example, there are certain features like bland chromatin and normal nucleoli and these form the basis of how clusters of malignancy of the cells are determined.

Supervised learning technique is utilized and the classifier chosen is the Decision Tree. Decision Trees sort data points by using a process of dividing dataset into subsets according to the input features value. The concept of the algorithm is that the initial split will be made on a feature which has the highest information gain. Information gain is calculated using entropy which is also a measure of the amount of ‘impurity’ in a set of data. The basis on which the decision is made is the information gain, with the node that yields the highest value selected for the splitting of the data; this process is a recursive one and is continued right until the so-called terminal nodes or leaves of the tree have been generated or all of the data points in the data set have been classified into the same class. In the present work, the Decision Tree model is developed in Python and the dataset is divided into train and test datasets for the validation of the model. The training set contains 60% of the data set and on the other hand the testing set having 40% data of the data set.

The performance of the Decision Tree model is therefore measured as the number of correct predictions as against the overall number of predictions. The employed model reaches the accuracy of 94 %. As much as 16% of patients were successfully predicted to contain cancerous cells, which stipulates a high reliability of the approach. This accuracy is quite a milestone because if applied clinically, it indicates that the Decision Tree algorithm has a near perfect capability of classifying breast cancer instances from the data set provided. The high accuracy rate also supports the utilization of machine learning algorithms in early identification and diagnosis of breast cancer as performed by the medical experts.

This paper also describes the steps taken to perform the data preprocessing in order to get ready for the analysis. All values that are disregarded as missing are archived; therefore, the dataset has only 683 rows. This step is very important because missing and incomplete data greatly affect the performance of the machine learning models. this helps the study to have a clean record of the inputs to be analyzed hence making the results more accurate and reliable.

That is why, visualization of the data is crucial for the analysis of the clustering and classification outcomes. The distribution of attributes and their relationship with the cancerous cells are also depicted in the study, employ graphical displays including scatter diagram, bar chart, and pie chart. Such visualizations allow defining which attributes are the most important for the breast cancer prediction.

The last proposition of the paper expresses the necessity of early diagnosis and subsequent treatment of breast cancer. The model that has been proposed in this paper involves the use of K-means clustering and Decision Trees classification; the proposed model has the benefit of being comprehensive in assessing the likelihood of cancerous cells in tissues taken from the breast. Hence, the high accuracy rate of the model confirms the usefulness of the model for diagnosing the breast cancer disease and its clinical applicability by medical researchers and practitioners. There is also a further research implication in terms of exploring more machine learning algorithms and data sets for improving the model’s prediction capability. The conclusions are therefore leading to the understanding of data mining techniques and their application are very useful in the development of medical diagnostic as well as in the enhancement of the health of patients.

The paper aims at assessing how K-means clustering algorithm could be used to examine breast cancer data with an aim of determining genes that help in prognosis of breast cancer. Breast cancer is one of the main diseases affecting the female population throughout the world; it is especially essential to diagnose it in the early stages. This research is centered on employing data mining approach to analyze the gene expression values of breast cancer patients, with special interest on the biomarker for classification of patients using some characteristics. The work also discusses the results of K-means clustering analysis applied to the dataset of breast cancer gene expression with similar studies with the goal to increase the knowledge about the bre clash multiplex and its molecular basis.

Information about breast cancer is given in the introduction of the work; breast cancer is an illness that displays symptoms of uncontrolled cell division. They point that for cancer to be formed the genes that control cell growth as well as differentiation must be changed. The study also underscores the contribution of GWAS in the discovery of genetic markers including SNPs that put the patient at risk for breast cancer. Nonetheless, it points out that the findings of GWAS give only a boundary framework of these genetic variants and their roles especially concerning the emergence of breast cancer subtypes. The introduction also gives a background on the microarray technology that has allowed the discovery of molecular markers and the classification of the different classes of breast cancer through mRNA expression patterns. However, the study establishes that most of the postulated biomarkers remain ill-defined mostly regarding causality to breast cancer.

The features of the applied methodology are described in the methodology section, where the authors reported about K-means clustering – the most used clustering method because of its simple and fast application. In K-means clustering the data points are divided into a number of clusters where each data point can surely be a part of only one cluster. Dataset which the study uses comes from M. Zwitter and M. Soklic from University Medical Centre, Institute of Oncology in Ljubljana, Yugoslavia.

The study uses the Orange open source, Python as well as the K-means clustering methodology. The dataset consists of 286 tuples with nine attributes, including the classification of menopausal status into three categories: There were also subgroups named “premeno,” “ge40,” and “it40,” depending on the tumor size. It is done in three stages where the distance measure and the procedures of initialization are different in each stage with aim to obtain the best clustering.

The random initialization only applies to the formation of the clusters in the first place and the Manhattan distance measure is used in the first phase. They use proximity between the data points in order to cluster the data and the solutions are then judged on the basis of the likelihood of classifying the data correctly. The study gives a mean classification accuracy of 66. 99% in this phase. In Phase II, the distance to centroids serves as the evaluation criterion, the distance being the Euclidean one, and the initialization is random. This phase increases the proposed average classification accuracy by an average of 74. 98%. Last but not least, in the third comparison of the same parameters, distance measure in CLARANS program is used with Pearson Correlation while initialization is with agglomerative clustering. This phase yields the perfect score of 1 or 100% in classification, which confirms that the used features Pearson Correlation and agglomerative clustering is useful in the separation of the different clusters with respect to tumor size and menopausal status.

In the results data analysis part, the authors describe the outcomes of clustering in all the phases in detail. In this phase, the clusters are described according to the tumor size ‘ge40’ with large tumors and ‘premeno’, ‘it40’ with small tumors. In Phase II new cluster patterns are driven by size and large tumor sizes are clubbed under ‘premeno’ and small ones under ‘ge40’ and ‘it40’. The results achieved in Phase III are considerably better than in the Phase II: the clusters are clearly distinguishable; therefore, 100% of the objects are classified correctly. In this study, the clustering results are presented in the form of scatter plot diagrams since they assist in the comparison of the tumor size and the phase and or the menopausal status.

The conclusion of the study restates the use of K-means clustering in assessment and evaluation of attributes of breast cancer. This paves way for the fact that by varying different measures like distance measures, scoring methods and inauguration techniques, the process of clustering can be exploited to generate highly accurate classification. Three methods that were tested were found out to be efficient in this case, namely, agglomerative clustering-Pearson Correlation. It is also shown that the results of clustering algorithms can be and should be combined with radiologists’ decisions to increase the possibility of a correct diagnosis of breast cancer. As for the future directions, it points to probabilistic clustering algorithms and other similar sophisticated approaches towards the improvement of the examination of the breast cancer data. This approach may help in the identification of the biomarkers and also in classifying the type of breast cancer correctly, ultimately helping the patients and their prognosis.

## Algorithms

### Logistic Regression

Logistic regression is a statistical approach used commonly to predict the output of the logistic function for binary classification. Multiple forms of linear regression it portrays the probability of cause or a dependent variable and one or more predicting variables. Compared to linear regression that estimates a measure, that is a continuous variable, logistic regression is planned to estimate a dichotomous variable. The concept is centered on prediction percentage of an attribute of being in a specific class. This probability is then used to classify the input based on a sort of threshold of likely occurrence.

In the case of logistic regression, the dependent variable is qualitative, though they have given it a value of 0 and 1 whereby 1 may represent the existence of a trait or occurrence while 0 represents the lack of the occurrence or the absence of the trait. The algorithm works with the purpose of identifying the most suitable model that will capture the chances of having a binary outcome, given the independent variables. It does this through the use of a logistic function or sigmoid function which have values ranging from 0 to 1. This output is the post estimation probability that the dependent variable belongs to a certain category.

Another model that is commonly used in medical research is logistic regression because it is suitable for use when the dependent variable is binary and categorical dependent variable can also be used when the outcomes of interest are multiple Logistic regression has an added advantage of being an analytically explicable model that describes the connection of the dependent variable to the probability of a specific outcome of the logistic regression model. When the research is of a quantitative nature, each independent variable is given a coefficient which indicates the degree and direction of the tested variable on the outcome. When the coefficient obtained is a positive value then this is an indication that more probability of the stated outcome is realized with an increase in the independent variable while the converse is true when the coefficient obtained is negative.

The model is adjusted in a way which allows to maximize the probability of the observed data to be predicted by the model. This means that changes will be made to the coefficients cyclically for the optimal result to be obtained. Once the model is built, then it is possible to estimate, for a given new data, the probability of occurrence of the outcome. Thus, with regard to the predicted probability, if it goes beyond a certain specified value, which in most cases is 0. 5, the model categorizes the input into that of being positive; otherwise, it categorizes it as negative.

Because it is not complicated or time-consuming and because it is certain that the relationship between the independent and dependent variables to be tested is linear, the logistic regression is preferred for the binary classification. The model also offers insight into the extent of the contribution of each independent predictor to the prediction which may helpful in essentially all domains inclusive of medicine / healthcare, finance, and social sciences.

Nevertheless, logistic regression is a powerful tool because of its ability to model the crude log odds of the outcome in terms of the set of independent variables; it does however, make the basic assumption of a linear relationship between the set of independent variables and the log odds, which can be a limitation in the more complex datasets. It also assumes that the independent variables are not closely related or ‘collinear’, as this will complicate the coefficients and affect reliability of the projections.

For example, some techniques like the L1 and L2 regularization techniques can be applied on the logistic regression approach in a view to solve some problems like overfitting especially when the model has many features. These techniques punish large coefficient values, they hence acting by default as a method to simplify the model while enhancing the capacity of a similar model to generalize on unknown data.

### Random Forest Classifier

Random Forest Classifier is a rather effective and universal algorithm of machine learning comprising the main goal of classification but is also applicable for regression tasks. It is an instance of the ensemble method of learning where many individual models are compiled together into a single meaningful model. The Random Forest Classifier is created using decision trees which are known as base models of the developed model. On the other hand, a decision tree is a straightforward model that ‘divides the given data into subsets using the feature values and comes out with a decision concerning the class of the given input data’.

Random Forest Classifier reduces a number of weaknesses of decision trees, including high variance and overfitting, through the combination of multiple decision trees into an ensemble. The basic concept is to architect a great many decision trees during training and then during testing, using all the trees constructed during the learning phase and applied to the test instance, sum up the results yielded by each of the trees to arrive at the final result. This is due to the fact that the data and features are randomly selected and as a result, a variety of trees is developed and these enhance the model.

The process starts first with generating Bootstrap samples from the original training data. Bootstrap sampling entails taking of data points in equal opportunities but in this case, one sample may feature in the data set many occasions as compared to other samples not featuring in the data set. To illustrate, one of the bootstrap samples is utilized, to develop a different decision tree. During the construction of the tree, it is assumed that at each node, one randomly selects a part of features; then, among all these features, the most optimal one is used to split the data. This feature randomness, also termed as feature bagging comes as an extra measure adding diversity among the trees.

After all trees are grown, the Random Forest Classifier sums up the predictions of all the trees. In the case of classification, each tree makes its particular decision for classification, the result is sum of trees decisions for particular class and the class with the highest sum is the final decision of the forest. This majority-vote mechanism reduces the variance and plies an ensemble-learning process to each individual, which is a common problem for decision trees. This is because the predictions of the Random Forest Classifier are an average of multiple decision trees which reduced variance on unseen data.

An added advantage of the Random Forest Classifier is that is has the capacity to handle the high dimensionality of the features and the data points. It also accepts data with null values and can run as accurately when most of the data is null. Moreover, feature interactions can be modelled with the help of the algorithm without a need to construct new features for this purpose.

Another advantage of Random Forest classifier is that it does not dependence a lot on outliers and noise in the given data. The so-called bagging of the trees means that many decision trees are created and the error rate of the algorithm is significantly less sensitive to the outliers in comparison with the algorithm based on one decision tree. Moreover, the model is not very sensitive to over fitting and this especially depends on the number of trees in the forest. In general, increasing the number of trees results in better and more stable models, but at the same hand, model complexity may become a bottleneck.

Yet another feature worth noting is that the Random Forest Classifier is able to give out estimate of the importance of the features used. The algorithm can then order the features according to how much the accuracy reduces when the model didn’t include it when predicting a target variable. It is highly beneficial when working in high dimensions where it provides an insight into which features should be used to decrease its dimension.

However, similar to most of the other classifiers, Random Forest Classifier is not without its drawbacks. One of the difficulties that are associated with the use of MBT is that it is not very interpretable. However, while a single decision tree is simple to portray and understand, a RF with hundreds or even thousands of trees in it is a ‘black box’ that is hard to explain the decision-making process of the model. Also, the use of this algorithm may be slow, particularly when the dataset is large and a number of trees are used, in terms of both time for training and memory space needed.

Another factor to understand about Random Forest is that it has hyperparameters the adjustment of which is always an important aspect to look at. Number of trees setting, the maximum depth of trees and the minimum number of samples which must be in the node so that it can be split are very influential parameters in the model. Still, Random Forest is less sensitive to hyperparameters than other models, which makes it quite convenient for many classification tasks.

when it comes to implementation, simulation and usage, the Random Forest Classifier can be applied to a variety of fields and disciplines such as finance, health, and bioinformatics among others because of its flexibility, and high accuracy, and, robustness. It remains the most popular method of classification when there is a need to get a strong and stable model for classification of various and heterogeneous data.

Therefore, the Random Forest Classifier can be thoroughly considered as an outstanding algorithm of ensemble learning that has no significant drawbacks in comparison with the decision trees, it extends and develops the effectiveness of which it is based. Since it can blend multiple trees into one reliable model, it is suitable for data classification, with high speed as well as accuracy on any type of data set.

### Support Vector Machine (SVM)

Under supervised machine learning, SVC is a prominent recommendation for users to employ due to its efficacy in classification. The tool is based on one of the most powerful methods of machine learning that have become popular in the 1990 s and that remained very popular even now because of their effectiveness in different scenarios, called Support Vector Machines.

Thus, in its essence, SVC tries to determine which hyperplane best separates data points of various classes in feature space. Hyperplane can be recognized as the dividing plane that splits the data into various classes. Whereas in a two-dimension world this hyperplane is a line and in a three dimensions world it is a plane. In higher dimensions it is a hyperplane. Essentially, the idea chosen is to identify that hyperplane that gives the largest margin between the two classes of data and the margin between the hyperplane and the nearest points belonging to any of the two classes is referred to as support vectors.

This is so because SVC is capable of handling data that is Liner separable as well as the one that is not Liner separable. For linearly separable data SVC has an easy time in arriving at the correct hyperplane with the maximum margin. Nevertheless, the given real-world cases are not entirely separable by a linear boundary. To counter this, SVC defines the so-called soft margin, in which some of the training instances are on the wrong side of the margin and even outside the hyperplane. It further allows the model to learn better from the new data by not being overfitting to its classification.

The soft margin approach is regulated by means of a single hyperparameter named C and is used to control the trade-off of classifying new observations with the maximum margin and minimum amount of classification error. A small value of C is conducive to a large margin while a small number of misclassifications may occur while a large value of C makes an attempt to classify all training examples correctly by minimizing the margin which leads to overfitting of the model. Choosing an appropriate C is far more decisive in fine tuning to perfection the model’s complexity/accuracy trade of.

When the data is not separable by a hyperplane in the original space, the SVC uses what is called the kernel trick. Kernel trick is a technique where the original data is transformed into a higher dimensional space, so that it becomes possible to draw a linear hyperplane to separate the data. This transformation is carried out implicitly and does not require the computation of coordinates of the data in the higher-dimensional space, thus making the computation much simpler.

Several kernel functions can be used, each serving different types of data:

* Linear Kernel: Good when the data is almost linearly separable in the original space or almost fully pre-classified.
* Polynomial Kernel: Appropriate for the situations where the emergence of complex and nonlinear boundaries between classes is possible, which generally can be described by polynomial equations.
* Radial Basis Function (RBF) Kernel: Often referred to as the Gaussian kernel this is the most commonly used kernel in KNN. It is used when data is non-separable in the original feature space. The RBF kernel transforms the data to the infinite-dimensional space; in this space, the classes can be separated by a linear hyperplane.
* Sigmoid Kernel: Comes closer to the behaviour of a Neural Network and may be used in specific situations, while not as popular as the RBF kernel.

There is no theory that bounds the choice of the kernel and it depends on the nature of the problem, as well as the data. SVC with an RBF kernel should be especially used when the shape of the hyperplane is nonlinear and it cannot be described using linear or polynomial equation.

SVC is also capable of handling high dimensionality and hence is useful wherever p > N as it is for microarray data. This is commonly used in text classification, work in bioinformatics, and image recognition problems. SVC is also not very sensitive to the overfitting problem, especially when applying the RBF kernel since it strikes a nice balance between the margins of the data and the amount of error made on it.

Still, SVC has limitations. That is, it may be quite resource demanding, particularly when dealing with huge data feeds, as the number of iterations depends on the amount of data fed into the method. Staying in the area of the very large data, training an SVC is time-consuming and might consume a lot of memory. Furthermore, SVC cannot handle ‘‘multi-class’’ classification problems; however, they can be expanded to the multi-class problems by ‘‘one-to-one’ or ‘one-to-(number of -class - one)’’ approach where a set of binary classifiers is trained to distinguish one class from the others.

In addition, there are some disadvantages we ought to know: first, it is difficult to decide the right kernel and to attend on the corresponding parameters (norms for the units of measure, C for the soft margin, gamma for the RBF Kernel etc). Al these affect the performance of the final model and if wrong hyperparameter values are selected, results in either underfitting or overfitting.

SVC has been successfully used in concrete practical applications in numerous fields of application. For example, in text classification which comprises of spam detection, SVC has capability to deal with large feature space of text data. In the past, it has been used in bioinformatics for instance in gene expression analysis genomics where the number of features (genes) are normally very many as compared to the number of samples. In image classification, especially for objects and faces, SVC – with RBF kernel option – gives a very high accuracy.

Thus, the Support Vector Classifier can be stated as valuable predictor in binary classification problems and effective means for combating the ‘curse of dimensionality’. Thus, by making the biggest possible distance between classes and applying kernel tricks to deal with high nonlinearity SVC offers rather efficient solution to classification problems. Despite the fact that it has some computation issues and depends on many hyperparameters, because of its ability to consider complex patterns of data and low tendency to overfit, it can be considered useful algorithm in the arsenal of machine learning algorithms.

### Principal Component Analysis (PCA)

This statistical tool commonly used in data analysis and machine learning is also well known with the name of dimension accent reduction. The technique is very useful when one has a large set of data to analyse by cutting down the number of variables as much as possible while retaining as much information as can be. This makes it easier to visualise, define, and manipulate the data in a number of ways in applications ranging from pattern recognition, data compression and noise elimination.

PCA operates on the basis where by the data variability is maximized in certain directions referred to as the principal components. They are therefore linear combinations of the original variables and more importantly they are uncorrelated and hence be referred to as principal components. The first principal component accounts for the largest variability in the data and each of the following components in turn takes the maximum variability that is perpendicular to the previous components.

This is done on the basis of the initial check of the data and if the variables are measured at sets of intervals, the results are standardized. Standardization makes all the variables in a set to have similar importance in the analysis. After standardization, PCA computes the covariance matrix of the data which helps to know the relation between the variables. The next step is to in find the eigenvalues and eigenvectors of this covariance matrix. The eigenvectors again denote the direction of the principal components and the eigenvalues are related to the direction’s variance.

After determination of the eigenvectors for the various eigenvalues, the eigenvectors are arranged in decreasing order of the corresponding eigenvalues. The top eigenvectors are called the principal components of the original feature space and they are basis to the new feature space. The original data is then mapped into this new lower-dimensional space while preserving most of the variability of the data in the process.

The number of the chief components specified depends on metrics of dimensionality reduction and the amount of data that still has to be retained. It is worth noting that as a measure of work done, one often uses the explained variance ratio to decide on how many components should be retained. The explained total variance is a measure which tells of how much of the total Variance in the data is accounted for by each principal component. The key idea of dimensionality reduction is the choice of such components that jointly account for most of the total variance, which, as a rule, amounts to 90–95%.

PCA finds more application where datasets have many variables which are inter-correlated since it condenses all the information in a dataset into a new set of un-correlated variables. This can result in the high performance and the interpretability of the machine learning solutions. In such areas as genetics, finance, image processing, PCA is employed to capitalize on the most important aspects of large datasets.

For instance, in image compression PCA can be applied for the compression of pixel information by creating a new space with fewer dimensions than the original space, completing the reduction process with minimal loss of image information. Likewise, in finance, PCA aids in the determination of variables that dictate the prices of the assets since a large number of correlated market variables are compressed into a few principal components that portray the market influences.

Still, as with many statistical methods, PCA has some drawbacks. Its primary disadvantage is the assumption of linearity: it can analyse only linear dependencies of variables and cannot capture more intricate interactions between the variables that can arise in datasets with other characteristics. When this occurs, kernel PCA can be used to describe non-linear patterns because it is an extension of PCA. The other disadvantage is that the PCA method is dependent on the scale of measurement. When the data is not standardized, it is still possible that the variables having a large scale will have a higher weight in the calculation of principal components which means that they will have a distorting effect on the results.

In addition, PCA does not incorporate the class labels of the concerned data and therefore can be viewed as an unsupervised technique. This means that this loss function may not always lead to a componentisation that results in the most useful features from classification points of view. For the supervised learning issues, it is wiser to use other dimensionality reduction techniques, such as for instance LDA because it incorporates class separability.

### KMeans Clustering

K-means clustering is an example of the unsupervised learning technique for partitioning of the data into clusters. The purpose of K-means is to partition the n data points into K clusters in the way that the data in one cluster have higher mean similarity compared to data in other clusters. This method is popular with data mining, pattern recognition, image segmentation and it is simple to process and efficient.

Stating with K initial centroids the points that are chosen for each cluster. These centroids can be chosen at random or according to certain strategies that include the use of the K-means++ which is a better choice since it enhances the quality of the clusters formed. For each data point, the corresponding nearest centroid is determined by distance, usually Euclidean distance, is determined. This assignment procedure creates K clusters where each cluster coming from the nearest possible data points to its centroid.

At the last step once all data points have been allocated to the clusters the algorithm computes new centroids by taking the average of the distance of nearest data points from each cluster. The centroids are recalculated, and all the features are reallocated according to the closest new calculated points which represent the centroids. In this process, the data points are re–assigned to a different cluster and the centroids are recalculated, and this process is reiterated until the centroids no longer shift or the number of iterations is met.

The first outstanding advantage of K-means clustering is its simplicity and high efficiency in terms of the amount of time it requires to work with large data sets. The algorithm is fast giving the results and is easy to formulate and to propel especially when dealing with high dimension data. However, one of the disadvantages of K-means is also present here. This brings the problem of deciding the right number of clusters, K. The parameter K needs to be defined in advance and is usually not clear; and often, methods such as the ‘elbow method,’ silhouette analysis, or cross-validation are applied.

Besides, the K-means algorithm presupposes that clusters are spherical and of equal size, which may or may not be true in principle. It can also be sensitive to outliers, because outliers may shift the position of centroids and thus the clusters will be non-optimal. The selection of the features to form the initial centroids can also affect the subsequent steps and give different solutions at different trial runs.

### DBSCAN (Density-Based Spatial Clustering of Applications with Noise)

The proposed algorithm DBSCAN: Density-Based Spatial Clustering of Applications with Noise is an unsupervised machine learning algorithm used in the clustering of data points density model. Unlike other clustering algorithms such as K –Means DBSCAN does not have any pre-requisites on the number of clusters to be formed. Rather than that, its near smaller clusters of densely packed points and marks points lying in the low-density regions as outliers or noise points. This makes DBSCAN especially appropriate to the type of clusters which are irregular in shape and to datasets which contain noise.

The fundamental notion that is underpinning DBSCAN is density. The algorithm uses two main parameters: epsilon (ε) and min\_samples are two parameters; Every point in a dense area is a neighbour of every other point, and epsilon is the maximum distance between two points before they are considered as not in a dense region and min\_samples is the minimum number of points to form a dense region.

DBSCAN then retrieves a random point from the dataset in the process of the algorithm. In case the number of points within the ε-radius of this point is at least min\_samples, it is classified as core point and a new cluster is initiated around it. After that, the algorithm goes through a process of checking each individual point in this neighbourhood. If such points are also the core points, their neighbours are included into the cluster evaluation. It goes until no further points can be allocated to the constituted cluster.

If a point does not have at least min\_samples within its ε-radius, it will be a border point, if the point is extremely close to the cluster that is in fact very distinguishable. General points which aren’t part of the core points or border point are called noise, meaning that they don’t fall in any of the clusters.

The discovery of clusters of any shapes is one of the biggest strengths of this algorithm commonly referred to as DBSCAN. It is also able to handle clusters that are irregular shapes or have stretched out shapes unlike the K-means algorithms that look for circular shapes. This makes it especially relevant when used with geographic data, with images, or with any situation where clusters may be unbalanced.

DBSCAN also has good ability to handle noise. It does this by subsuming the outliers under the label noise so that they do not affect the formation of clusters. This feature is rather helpful in the real-life data sets since such data can always prove to have some noises that distorts the pattern when other clustering models are used.

However, some weaknesses can be identified with regards to DBSCAN. It is also important to note that different ε and min\_samples will have a direct impact of the algorithm that is used in the system. If these parameters are initialized with inadequate values, then there can be poor clustering either merging clusters that should be different or splitting a cluster into further parts that it should not be. Determination of some of these parameters maybe difficult and normally needs little knowledge or trial-and-testing. In the same regards, DBSCAN can experience problems with datasets of nonuniform density since there could be just one ε all over the data field.

The final weakness of DBSCAN is that it has high time complexity especially when working with large dataset. The algorithm needs the calculation of distances of between every two points, which becomes cumbersome with the large number of points in the set.

### Gaussian Mixture Model (GMM)

A Gaussian Mixture Model (GMM) is a probabilistic model used for modelling the distribution of the different populations in a population when the particular data point does not have to be labelled as belonging to such population. Mostly, it forms the basis of clustering, density estimation as well as pattern recognition. On its operation, GMM supposes that the data has been generated by several Gaussian distributions where every singulate is a cluster of the data. One should remind that each Gaussian distribution is described by a mean value, variance, and the mixing coefficient as well as indicates a certain percentage of data set.

GMM extremely explicit in its operating, it means that data points conform likely to be belonging one and the other to a number of distributions, where every distribution, can be considered as a cluster. Unlike other clustering which involve assignments such as K-means where each instance is only assigned to be in one cluster, GMM gives a probability of each instance to be in to which cluster. This probabilistic nature enables GMM to achieve better sensitivity on the cluster shapes and the overlaps between them.

In a GMM, the parameters of the model are estimated with the use of the Expectation-Maximization (EM) algorithm, which is iterative and two-step. The Expectation (E) step of the EM approach involves computing the probability of each of the data points in belonging to each of the clusters that are currently postulated to be generated from the varying Gaussian distributions. In the Maximization (M) step, these probabilities are employed to refine the parameters of Gaussian densities wherein the updates are in the means, variances and mixing coefficients. These two steps are then repeated until the parameters used are stabilised to an optimal level.

A major advantage of GMM is the capability of fitting programs that involve clusters shaped as ellipses and not balls in contrast to K-means. Another advantage of GMM is that it can work under clusters that have different size and density which gives it a competitive edge of handling different types of data. Furthermore, with respect to the idea of overlapping clusters, GMM has an advantage over K-Means since it is probabilistic in respect to the assignment of data points to clusters.

Still, GMM has its disadvantage. It assumes that the respective distributions are Gaussian and this may not in fact be the case in many research datasets. Similarly, as with many other clustering algorithms, GMM has its disadvantage, namely that the number of clusters must be prespecified by the user. The question of how many components one should employ is normally resolved by the numbers like the Bayesian Information Criterion (BIC) or the Akaike Information Criterion (AIC) which penalize for model complexity in terms of the number of components as well as high residual variance.

## Metrics

### Confusion Matrix:

A confusion matrix is a type of table used in order to determine the accuracy of the result of choosing one class by a classification algorithm. It shows the outcomes of true positive (TP), true negative (TN), false positive (FP), false negative (FN) models. These counts can be useful in identifying the kind of errors, which the model is committing. Specifically:

**True Positives (TP):** Made correct predictions of positive observation.

**True Negatives (TN):** Made the actualization of negative observation predictions right.

**False Positives (FP):** Type I error – It is when the hypothesis H1 is rejected when in actual sense it is false or, cases that were predicted to be positive when in real they are negative.

**False Negatives (FN):** These include, the failure to reject the null hypothesis when in actual sense it should have been rejected (this one is also called type II error);.

### Classification Report:

The classification report also gives detailed information about the accuracy, specificity, and the F1-Score about each class in a classification task. These metrics are defined as

**Precision**: A measure of the degree the numerical output of positive values is accurate when compared to the total number of actual positive observations (Precision = TP / TP + FP). Small percentage of positives also mean a smaller number of false positives.

**Recall (Sensitivity or True Positive Rate)**: The ratio of correct prediction of all positive observations actually in the class to all the observations included in the actual class (Recall = TP / (TP + FN)). High recall implies low false negative rate.

**F1-Score**: Averages of the precision and recall where more of both is a good thing (F1-Score = 2 \* (Precision x Recall) / (Precision + Recall)). It becomes particularly advantageous when the class distribution is say 1:999.

### Accuracy

Accuracy is calculated as the number of observations that were classified correctly to the total number of observations. This is one of the easily understandable and widely used measures for the efficiency of the classification model (Accuracy = (True Positive + True Negative) / (Total Positive + Total Negative + False Positive + False Negative)). If we recall the definition of accuracy, we see that it is quite straightforward, though in the case of imbalanced classes it can be rather misleading.

### Silhouette Score:

The silhouette score is used as a measure in order to assess the quality of a clustering algorithm. It determines the degree of similarity of an object to own cluster to other clusters. The silhouette score varies between -1 and 1:

* A value in the range of 0 and 1 mean that the points in one cluster are closer to each other and also farther from points belonging to another cluster.
* Where the score is at 0 it paints a picture that the point is just on the boundary or extremely close to the boundary of two clusters.
* A negative value to ‘s’ can be interpreted as that the data points could be classified into the wrong clusters.

Silhouette score is most beneficial when used for comparing the various results of different clustering algorithms or as the results of the same algorithm with various parameters.

### Feature Importance:

Feature importance was applied in the context of the Random Forest classifier to determine the importance of the features in the making of the prediction. The measure of the importance of a feature is the mean and standard deviation of the accumulation by how much the carryover of the impurity has been reduced by each tree in the forest. An attribute with a high number on importance is deemed to have a higher impact on the prediction of the target feature.

### Cross-tabulation (crosstab):

For assessment, cross-tabulation is employed in order to compare the clustering outcomes with the actual class definition. It comes with what is known as contingency table that displays the cluster, the clustered points against the actual diagnosis labels (e. g., benign or malignant). This assist in determining the level of accuracy of the identified clusters to the existing classification.

# Chapter 3: Methodology

## Tools and Techniques

Pandas: Parameters used for data handling and data analysis which include loading datasets, dropping columns and check for missing values.

Matplotlib: Pyplot is a plotting library that can be used for creating statics/interacting/animated plots using Python.

Seaborn: An extension of the Matplotlib for creating qualitative statistical graphics and informative displays.

NumPy: A library that may be used for mathematical computations as well as for functions of handling arrays.

Logistic Regression: A classifier or a model employed in binary classification to predict the probabilities of a categorical nature.

Random Forest Classifier: A type of classification technique, similar to the Random Forest, only that it creates multiple Decision Trees and combines them into a single model to have a better and more stable forecast.

Support Vector Classifier (SVC): The supervised learning approach used in binary classification that reveal the best hyperplane that differentiates different classes.

Principal Component Analysis (PCA): A technique employed in an effort to bring about a decrease in the number of variables used but which will preserve a great deal of variance in the data.

K-means Clustering: Controlled clustering: An unsupervised learning algorithm employed in partitioning of the dataset into KK clusters according to features.

DBSCAN (Density-Based Spatial Clustering of Applications with Noise): An algorithm that can form a number of meaningful clusters out of the data points that are most condensed and is able to identify points that are density minorities.

Gaussian Mixture Model (GMM): The K means clustering model that is a probabilistic model working under the assumption that the data at hand has come from a more than one Gaussian distribution.

Train-Test Split (from sklearn. model\_selection): A procedure employed in the separation of the overall dataset into sub-samples of training and testing data with the purpose of model assessment.

StandardScaler (from sklearn. preprocessing): A kind of feature scaling in which all features in a dataset have zero mean and unit variance; used before applying a machine learning algorithm.

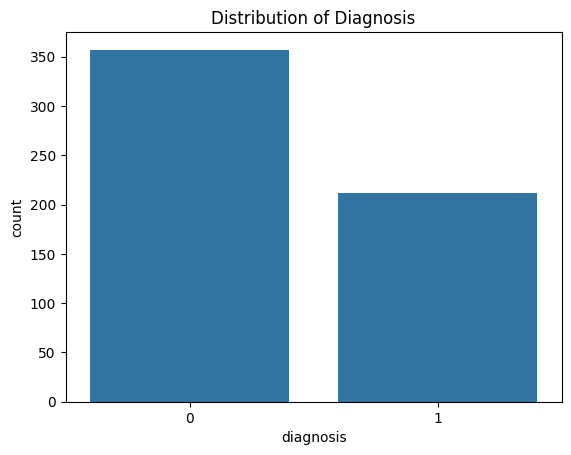
Confusion Matrix (from sklearn. metrics): A model of assessment that gives a brief description of the outcome of the classification process through the number of good and bad forecasts.

Classification Report (from sklearn. metrics): gives the precision, recall, the F1 score, and the accuracy, for each class of a classification model.

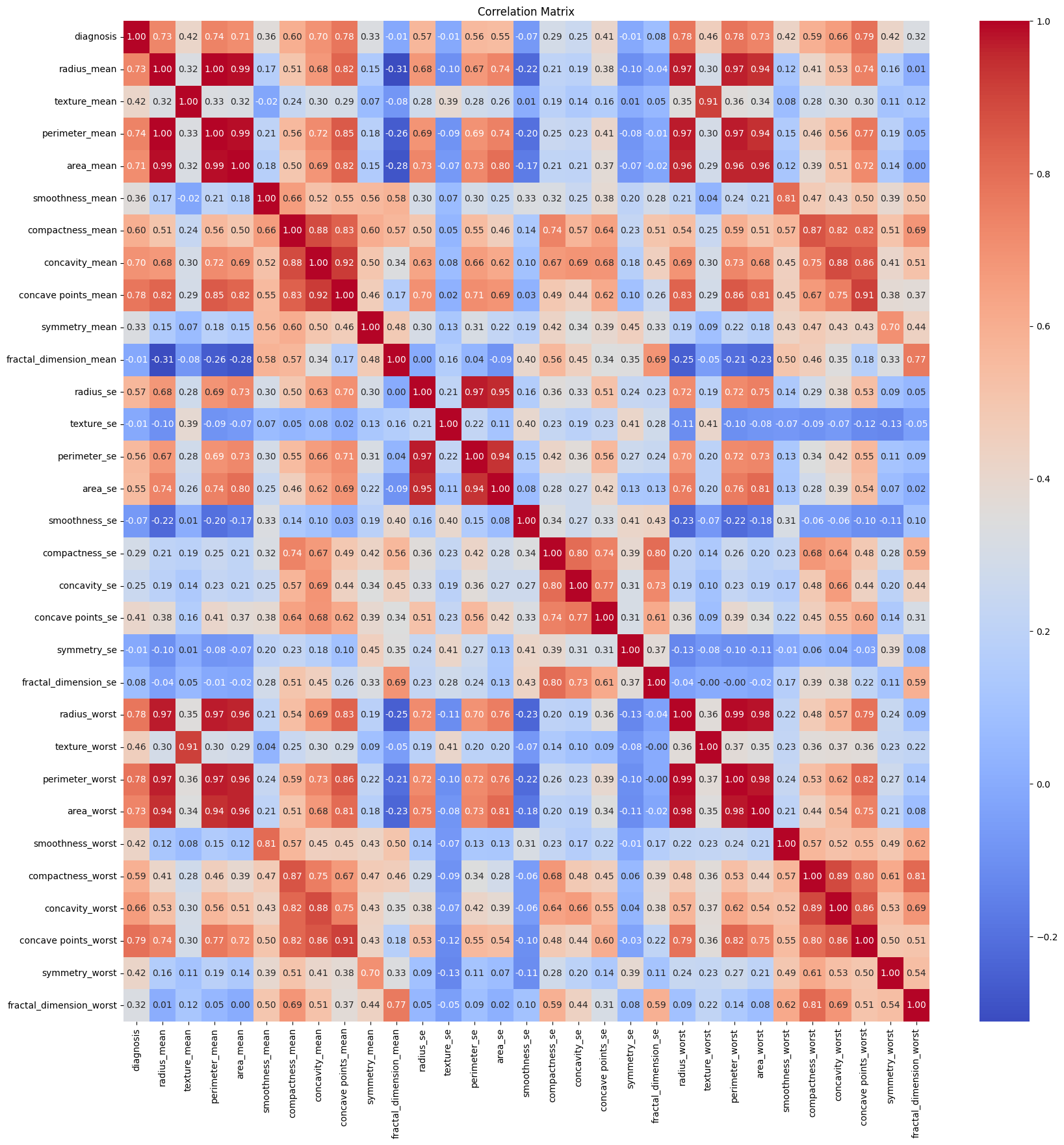
Silhouette Score (from sklearn. metrics): A measure applied to assess the quality of clusters looking at how much a point is close to its own cluster rather than to others.

Feature Importance: A procedure to compute the worth of features in Random Forest to predict results of future occurrences.

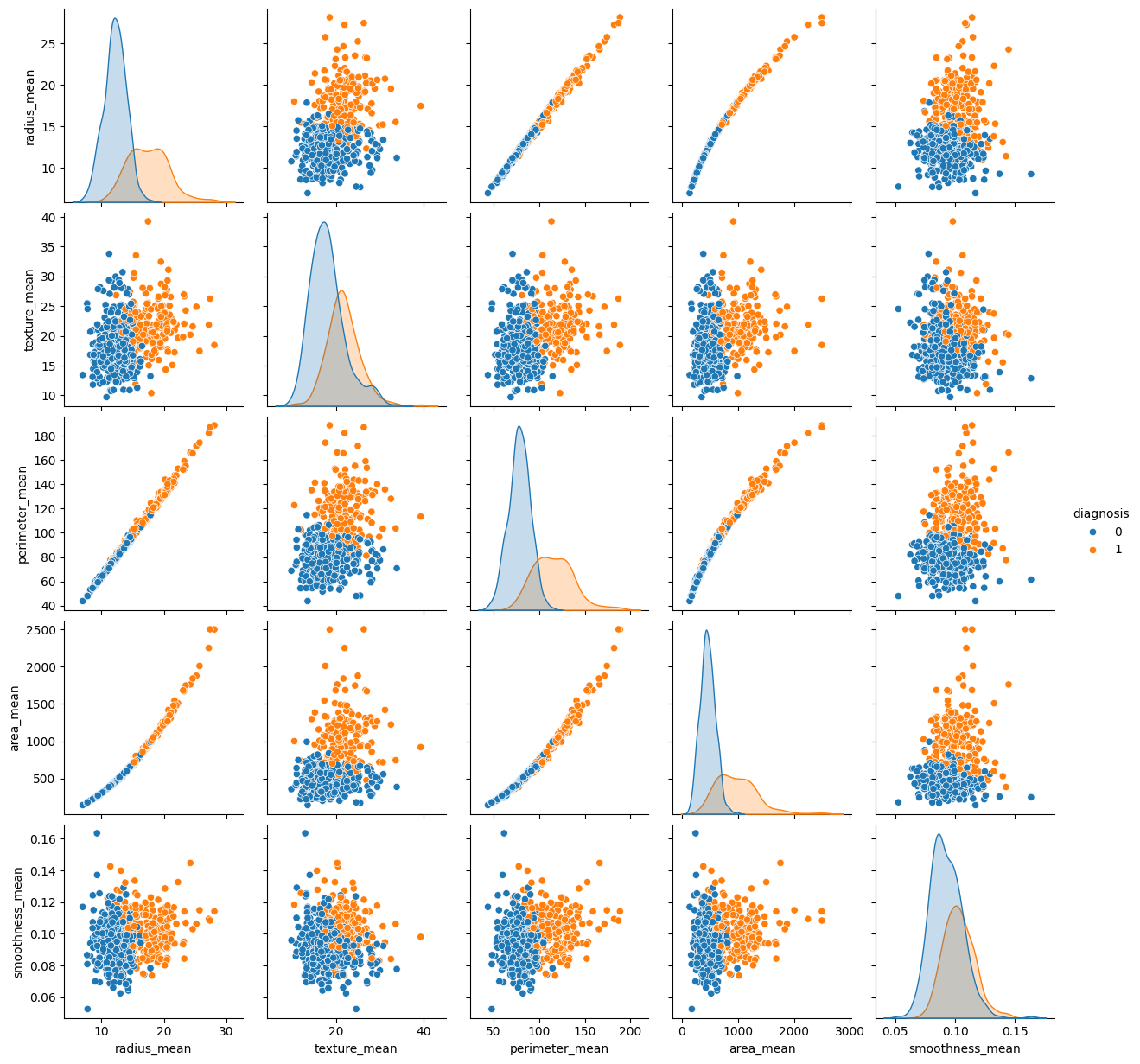
## EDA and Visualization



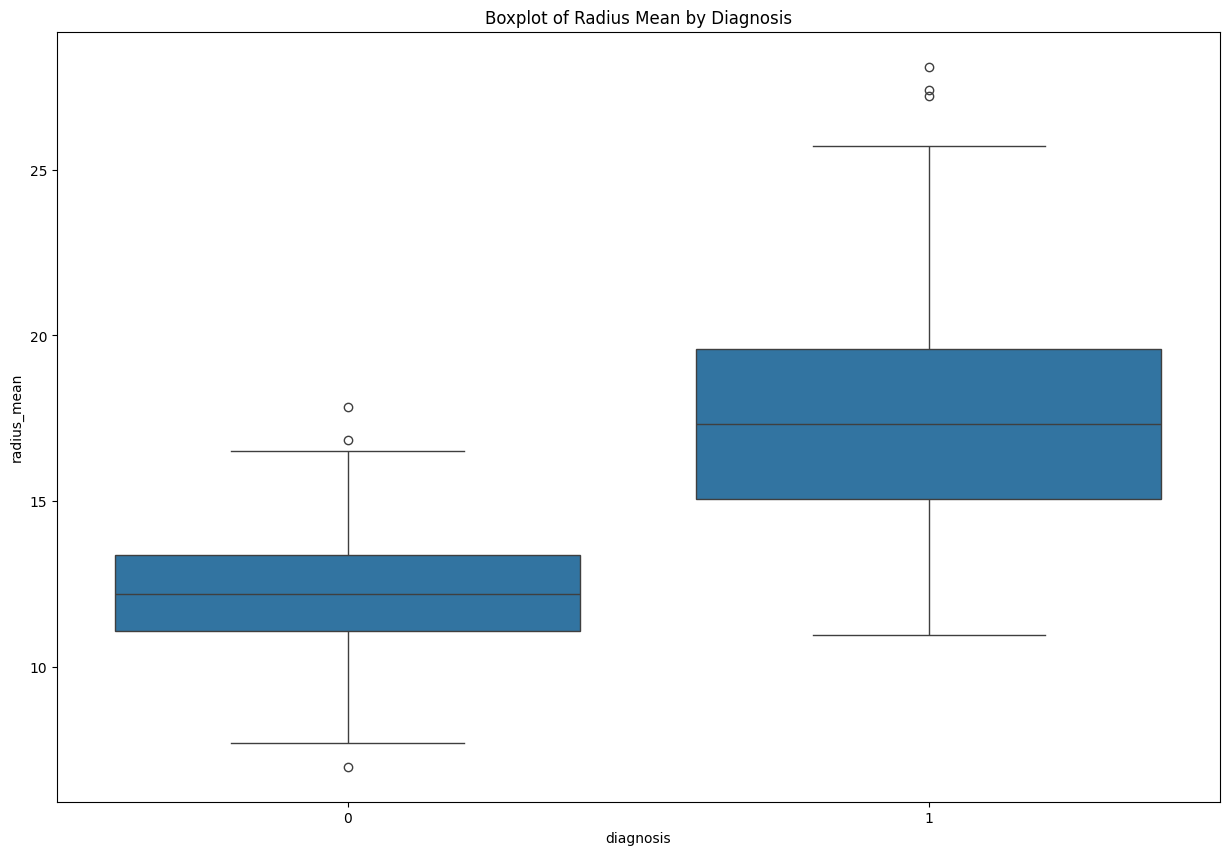
Distribution of Diagnosis: The first is a bar plot of the distribution of diagnoses in the dataset. There are two categories: For this study, two categories of breast cancer are used which are; the benign which is assigned the label of 0 and the malignant which is assigned the label of 1. As seen from the plot, there are clearly more observations in the ‘benign’ class than in the ‘malignant’ class, a phenomenon that is observable with many medical datasets that aim at the early detection of diseases.



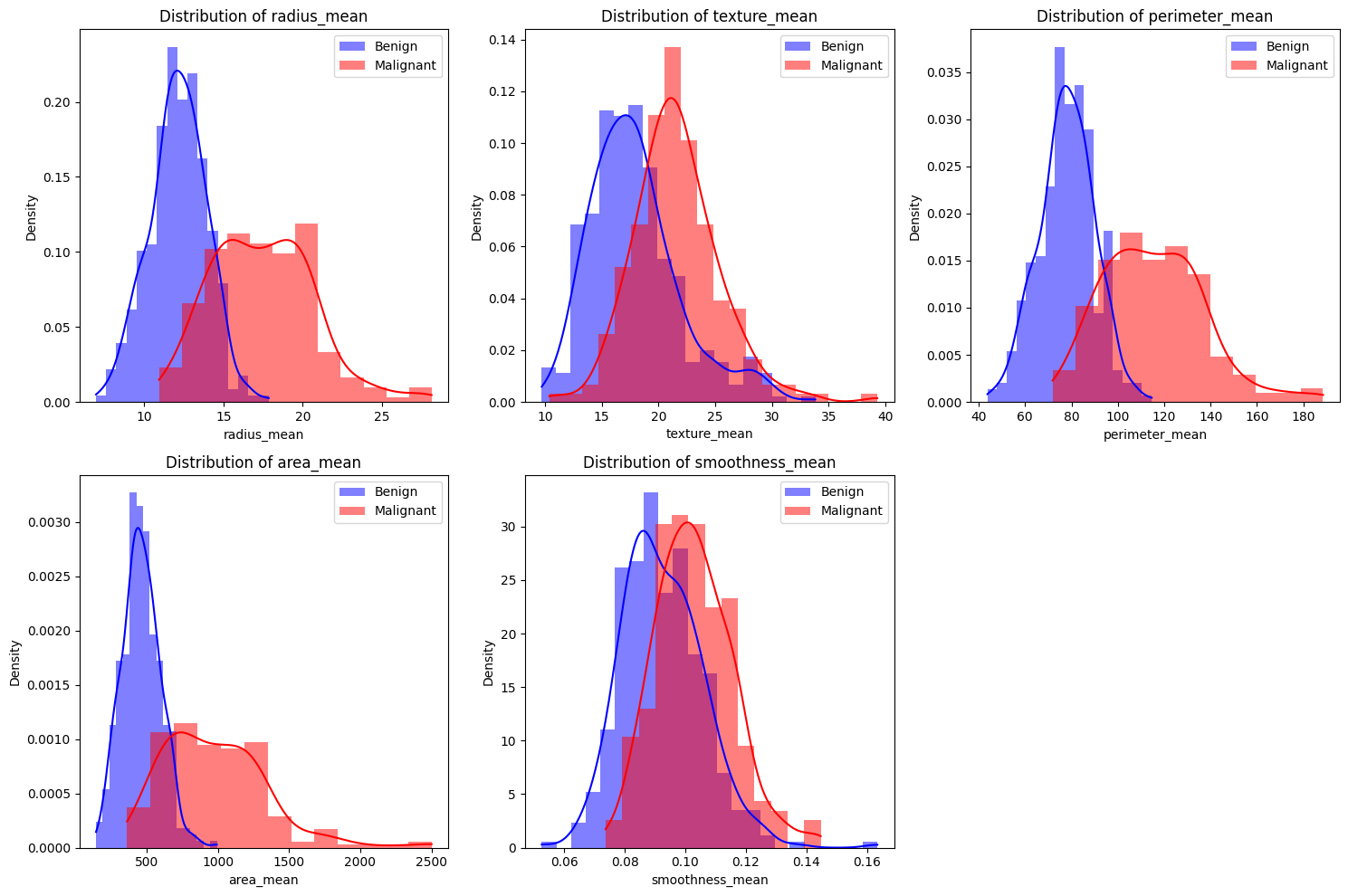
Correlation Matrix: The second image is the joint plot of the correlation where the participant is exposed to correlation coefficients of different features in the dataset. The degree of the colour reflects the level of the correlation – the denser the colour is, the stronger the correlation. It is very useful for determining which features are most interconnected, to better comprehend the relationships and, in some cases, feature selection or engineering.



Pairplot of Selected Features: The third image is example of pairplot which is the scatter plot for the selected features as well as histograms at the diagonal. This plot aids in placing the various features with regard to the diagnosis in a graphical view. Scatter plots and trends of point organization show how effectively one or another characteristic can differentiate between the malignant and benign ones.



The last boxplot is of radius\_mean by diagnosis and it shows the distribution of tumor radius for the cases diagnosed as bening (0) and malignant (1). From the boxplot it is evident that the median of the radius is larger for malignant tumors and the spread of values is wider in this case hence we can conclude that the larger radii are more associated with malignancy. Based on the positive significance, the radius can be regarded as a discriminative feature in diagnosing breast cancer inasmuch as the visualization suggests.

Fig. 5: Distribution of the averaged values of selected features (radius\_mean, texture\_mean, perimeter\_mean, area\_mean, and smoothness\_mean) depending on the diagnosis – benign or malignant. This information is conveyed by these density plots depicting that for most of the features, the densities are shifted up for malignant cases than in benign cases. It appears that these features can be used to differentiate between benign and malignant cases, and that tumors of a malignant nature are usually associated with higher values.

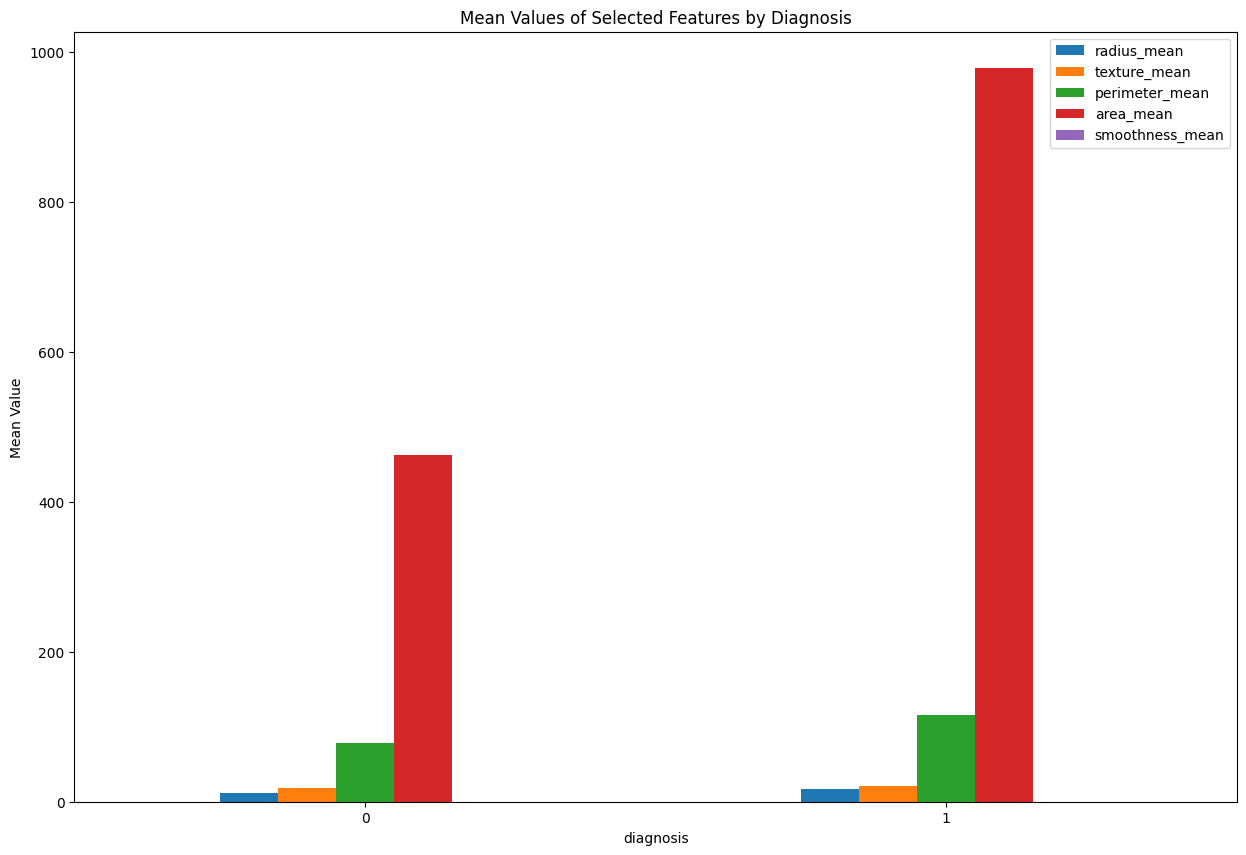


Image 6: The bar plot of mean values for selected features by diagnosis presents the bar plot of the mean of the same features with regard to cases that have benign (0) and malignant (1) disease. Mean values for the feature set include; radius\_mean which is 18. 58 for malignant cases as compared to 13. 38 for benign cases, perimeter\_mean stands at 149. 3 for malignant cases as opposed to 117. 4 for benign cases, mean area stands at 604. 2 for malignant cases and 520. 2 for benign cases, texture\_mean is slightly more at 12. 2 for malignant cases as compared to 10 This has the effect of underlining that, the elements listed above, are pivotal in the diagnosis process, since higher mean values are related to malignant tumours.

# Chapter 4: Results and Conclusion

## Results

**Logistic Regression Results:**

From the confusion matrix and the classification report, one realizes that compared to other classifiers, Logistic Regression produced an accuracy of about 98%. 25%. It is clear that presented model has high precision, recall and F1-scores for both classes, benign and malignant, although a little better scores for the benign class: precision of 0. 99 and F1-score of 0. 99. The small disparity of the performance measures that the different classes presented also indicates that Logistic Regression has good stability for this dataset.

**Random Forest Results:**

Out of the algorithms used the accuracy found in the Random Forest model was found to be around 97 percent. The result reveals that the mean accuracy of Random Forest Classifier is 30. 044%, the mean area under the curve of ROC is 0. 890 that is slightly lower from the Logistic Regression. The confusion matrix likewise shows a relatively higher misclassification for the malignant class which gave four false negatives and thus the lower recall of 0. 94 for this class. For all the above stated reasons the best model still provides high accuracy for the malignant cases the precision being at 0 and 98%. This plot reveals that the features the model uses most are concave points\_mean, concave points\_worst and area\_worst amongst others; a significant fact which underscores the ability of the specified tumor details to classify the tumor.

**Support Vector Machine (SVM) Results**

The SVM model obtained a level of accuracy that variably ranged from 94 percent to as high as 97 percent. About 66%, equal to Logistic Regression. In the class report it can be seen that in both classes, the performance of the SVM was good with a precision, recall and F1-score close to 0. 98. The model is slightly more in balance than Random Forest in terms of the differential between the two classes; the recall is almost the same for each class.

**KMeans Clustering:**

To begin with, only the first image has been selected and feature vectors are extracted from the images through the extraction of different features with size of 1000 x 1000 and size of 182 x 182 for applied clustering technique of KMeans with K=2clusters. The location of this plot is such that there is a division between the two clusters, the yellow points refer to and the purple points refer to. The silhouette score is approximately at 0. The clustering quality is fairly reasonable by referential comparison to be around 697 which imply that the clusters are well reconstructed. Cross-tabulation at the bottom shows that the clustering is somewhat accurate in terms of ‘real’ diagnosis labels: purple cluster definitely contains more benign cases (labeled as 0) than malignant ones while other clusters do not differ much from actual 0/1 distribution.

On the whole KMeans Clustering delivered an accuracy of around 85%. of 4% which shows that the clusters derived out of the model are quite close to the true labels. This makes it easy to conclude that the model known as KMeans was somewhat successful in the clustering of similar data points albeit its slight inefficiency.

**DBSCAN Clustering:**

The second image depicts the use of DBSCAN clustering management technique. As can be observed, DBSCAN also failed in the process of identifying clear clusters and instead put all points into a single cluster and this is evident from the uniform color of the points. At the bottom of the message there would be an indication that the algorithm could not locate other clusters, possibly because of the epsilon and min\_samples values. It suggests that the current setting of DBSCAN is not appropriate for this particular data or the data contains no density-based clusters in the PCA reduce space.

DBSCAN returned more than two clusters hence the measures of performance such as accuracy cannot be determined. This result also shows that DBSCAN with the chosen parameters is not appropriate for density-based clusters in complex datasets which may not have density-based clusters or where certain parameter settings may be required to find such clusters.

**Gaussian Mixture Model (GMM) Clustering:**

The third image depicts a GMM clustering of this vector, which will be explained later, that was successful in clustering relevant images. Similarly to KMeans, GMM assigned two clusters marked with yellow and purple points. Yet, the clusters are not as well separated as in KMeans, which is traceable to the silhouette score of 0. 585. This means that GMM can recognize the structures of cluster while the separation is not very clear as there is some overlapping of clusters especially in the PCA space.

While Random Forest gave an accuracy of roughly 90 percent, K Nearest Neighbor along with Gaussian Mixture Model (GMM) gave an accuracy of around 93 percent. 3%, outperforming KMeans. This may indicate that GMM was better at capturing the underlying distribution of the data because of the ability of the method in dealing with overlapping clusters and relationship in the data.

## Conclusion

Logistic Regression, Random Forest and SVM were all shown to classify benign and malignant tumors with high accuracy, although Logistic Regression was shown to be the best of the three by a small margin. The high levels of precision, recall, and F1-score on Logistic Regression for both classes underlines the capability and reliability for this dataset. Random Forest has moderate sensitivity but is better in terms of interpretable feature importance, some of which are profoundly important to distinguish benign tumors from malignant ones, for instance, concave points\_mean and area\_worst. The SVM model performed nearly at par with the Logistic Regression with respect to both the classes and had better times on the complex data structure; therefore, it can be trusted as an alternative.

Surprisingly, the unsupervised models gave a much more diverse result than the supervised ones with a number of them outperforming the supervised models. KMeans clustering displayed acceptable results giving out an accuracy of about 85% to suggest that it created decent clusters of similar data. And the silhouette score for KMeans also indicated that although the clusters were not perfectly separable from each other, they were reasonably well defined. On the other hand, DBSCAN did not perform very well having problem with the current dataset to identify the clusters and they grouped all the points into a single cluster indicating that with the current setting it has the drawback of not identifying clusters well.

The Gaussian Mixture Model (GMM), on average, gives about 93% correctness more than the KMeans, outweighing the later. Comparing with the percentages of 3%, the proposed model still received results that were not as high as the supervised models. Indeed, GMM was capable of finding overlapping clusters and it also provided a probability distribution of the data but the clusters were not clearly distinguishable as with the case of KMeans.

However, it is evident from the above experiment, that in the normal case of the given numeral data-set, the supervised is much more efficient and accurate than the unsupervised models for this classification job. Logistic Regression and SVM provide good and comparable accuracy, and Random Forest provides additional interpretability using features importance. Despite the reasonable results in clustering, the KMeans and GMM are significantly lower in accuracy to the supervised methods and DBSCAN is not able to identify the clusters and thus might not be suitable until further adjustment of parameters has been made or used in this current dataset.

Therefore, in conclusion, although nonlinear analysis based on unsupervised learning such as KMeans and GMM gives a better picture of the distribution of the data available, classification of breast cancer efficiently happens with supervised models like Logistic Regression and SVM, and so are preferred in this case. If the particular characteristics of the application such as the need for interpretability of the model or when the non-linear relationships exist the choice between these models can be more specific.

## Future Work

# Chapter 5: Legal, Ethical and Professional Issues

# References

# Appendices