

# **BREAST CANCER PREDICTION USING MACHINE LEARNING MODELS**

**A Thesis submitted  
in Partial Fulfillment of the Requirements for the  
Degree of**

**MASTER OF SCIENCE  
In  
Applied Mathematics**

**by  
Nidhi Bhati  
(2K22/MSCMAT/26)**

**Under the supervision of**

**Prof. Anjana Gupta**



**Department of Applied Mathematics**

**DELHI TECHNOLOGICAL UNIVERSITY  
(Formerly Delhi College of Engineering)  
Shahbad Daultapur, Main Bawana Road, Delhi-42**

**May, 2024**



## **DELHI TECHNOLOGICAL UNIVERSITY**

(Formerly Delhi College of Engineering)  
Shahbad Daulatpur, Main Bawana Road, Delhi-42

### **CANDIDATE'S DECLARATION**

I, (Nidhi Bhati) 2K22/MSCMAT/26 hereby certify that the work which is being presented in the thesis entitled "Prediction of Breast Cancer using Machine Learning Models" in partial fulfillment of the requirement for the award of the Degree of Master of Science, submitted in the Department of Applied Mathematics, Delhi Technological University is an authentic record of my own work carried out during the period from August 2023 to April 2024 under the supervision of Professor Anjana Gupta.

The matter presented in the thesis has not been submitted by me for the award of any other degree of this or any other Institute.

**Candidate's Signature**

This is to certify that the student has incorporated all the corrections suggested by the examiners in the thesis and the statement made by the candidate is correct to the best of our knowledge.

**Signature of Supervisor**

**Signature of External Examiner**



## **DELHI TECHNOLOGICAL UNIVERSITY**

(Formerly Delhi College of Engineering)  
Shahbad Daulatpur, Main Bawana Road, Delhi-42

### **CERTIFICATE BY THE SUPERVISOR**

Certified that Nidhi Bhati (2K22/MSCMAT/26) has carried out their search work presented in this thesis entitled “Prediction of Breast Cancer using Machine Learning Models” for the award of Master of Science from Department of Applied Mathematics, Delhi Technological University, Delhi, under my supervision. The thesis embodies results of original work, and studies are carried out by student themselves and content of the thesis do not form the basis for the award of any other degree to the candidates or to anybody else from this or any other University/Institution.

Place: Delhi

Date: June, 2024

Prof. ANJANA GUPTA

SUPERVISOR

DEPARTMENT OF APPLIED MATHEMATICS

DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)

Bawana Road, Delhi-

110042

# **BREAST CANCER PREDICTION USING MACHINE LEARNING MODELS**

Nidhi Bhati

## **ABSTRACT**

Predictive modeling using machine learning techniques plays an important role in various fields. We can explore the use of machine learning algorithms to predict outcomes such as disease progression, stock prices, or customer behavior. Using large data sets and advanced algorithms, we aim to increase accuracy and provide useful information for decision makers.

Worldwide, breast cancer is the most frequent cancer diagnosed in women, and its incidence is rising yearly. Early diagnosis and accurate detection of relapse are essential to improve prognosis. In this study, To determine which machine learning (ML) method was most effective in predicting the recurrence of breast cancer, we examined several models.

Eleven machine learning algorithms can be used to create a prediction model: logistic regression (LR), random forest (RF), support vector classification (SVC), decision tree, multi-layer perceptron (MLP), extreme gradient boosting (XGBoost), gradient boosting decision tree (GBDT), Adaptive Boosting (AdaBoost), Linear Discriminant Analysis (LDA), Gaussian Naive Bayes (GaussianNB), and Light Gradient Boosting Machine (LightGBM). Metrics like area under the curve (AUC), accuracy, sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), and F1 score were used to assess each algorithm's performance.

## **ACKNOWLEDGEMENTS**

At the outset of this report, we extend our heartfelt appreciation to all individuals who have supported us in completing this dissertation. Without their proactive direction, assistance, collaboration, and support, we could not have advanced toward achieving the desired outcomes. Prof. Anjana Gupta provided diligent assistance and support that enabled us to complete our dissertation, for which we are eternally grateful. We express our sincere appreciation to each other for working together to complete this project while preserving our individuality. We are thankful that Delhi Technological University provided this opportunity to us. We additionally express our sincere gratitude and respect to our parents as well as other family members, who have always provided us with both material and moral support. Finally, but just as importantly, we would like to express our heartfelt gratitude to all of our friends who supported us in any way during this effort. This quick acknowledgement does not imply a lack of gratitude for anything.

Thanking You

**NIDHI BHATI**

## TABLE OF CONTENTS

<b>Candidate's Declaration</b>	<b>i</b>
<b>Certificate by Supervisor</b>	<b>ii</b>
<b>Abstract</b>	<b>iii</b>
<b>Acknowledgements</b>	<b>iv</b>
<b>Content</b>	<b>vi</b>
<b>References</b>	<b>vii</b>
<b>1 INTRODUCTION</b>	<b>1</b>
1.1 Machine Learning .....	1
1.2 Types of Machine Learning .....	2
1.2.1 Supervised Learning .....	2
1.2.2 Unsupervised Learning .....	2
<b>2 LITREATURE REVIEW</b>	<b>18</b>
2.1 Brief About Breast Cancer.....	
<b>3 MODEL BUILDING</b>	<b>23</b>
3.1 Histogram.....	24
3.2 Correlation Matrix.....	29
3.3 Scatter Plot .....	30
3.4 Boxplot.....	32
3.5 Logistic Regression.....	34
3.6 Decision Tree .....	37
3.7 KNN Model.....	40
3.8 XGBoost.....	42
<b>4 CONCLUSION</b>	<b>43</b>

## List of Figures

Fig 1. -Types of machine learning techniques.....	08
Fig 2 -Difference between supervised/unsupervised/reinforcement.....	09
Fig 3. -Supervised learning working.....	09
Fig 4 -Type of supervised.....	10
Fig 5- Graph of Linear regression.....	11
Fig 6- Graph of multiple linear regression.....	12
Fig -Sigmoid function graph.....	12
Fig 8 -KNN working.....	14
Fig 9- KNN graph.....	15
Fig 10- Working of unsupervised.....	15
Fig 11-Type of unsupervised learning.....	17
Fig 12-Types of clustering.....	19
Fig 13-PCA Figure.....	20
Fig 14-Graph of PCA.....	21
Fig 15-K means clustering.....	23
Fig 16-Anatomy of breast.....	26
Fig 17-Benign v/s Malignant tumor.....	27
Fig 18-Graph between Benign v/s Malignant.....	29
Fig 19-Histogram between Benign v/s Malignant 1.....	32
Fig 20- Histogram between Benign v/s Malignant 2.....	33
Fig 21- Histogram between Benign v/s Malignant 3 .....	34
Fig 22-Correlation Matrix.....	36
Fig 23-Scatter Plot.....	37
Fig 24-Box Plot.....	40
Fig 25-Logistic Regression Code.....	41
Fig 26- Logistic Regression Graph.....	42
Fig 27-Feature importance in logistic regression.....	43
Fig 28-Decision tree code.....	44
Fig 29- Feature importance Decision tree code .....	45
Fig 30-KNN code.....	46
Fig 31-KNN Graph.....	47
Fig 32-XGBoost Code.....	48

# CHAPTER 1

## INTRODUCTION TO MACHINE LEARNING

The capability pertaining to a system to instantly receive, participate, and subsequently derive insights from extensive datasets —without having to be deliberately programmed to do so— and then extend that knowledge to discover new information on its own is known as machine learning.

In a nutshell, the following are applications where the machine learning algorithm can be used:

- (1) gaining a deeper understanding of the cyber domain incident that generated the data for the study,
- (2) modeling the occurrences underpinnings,
- (3) using the built model to anticipate the values that the event will produce in the future; and
- (4) proactively identifying any anomalous behaviour of the phenomena so that suitable corrective action may be done in advance.

Machine learning is a forward-thinking field that has benefited from recent technological developments, especially the innovation of clever methods and enhancements in storage and hardware. Many of the jobs that were predictable a few decades ago can now be completed more accurately and efficiently. A wide range of commercial domains, including medical and human resources, media and entertainment, financial management and investments, marketing and sales, supply chain management, and operational processes so on have seen the emergence or evolution of several machine learning applications. There are certain notable trends being seen in the industry's applied machine learning systems. These developments will further harness the potential of artificial intelligence and machine learning to help businesses and societies in general

Some of these leanings are following:

- (1) reduced code capacity and quicker machine learning system installation
- (2) expanding the usage of lightweight systems that can function on Internet of Things (IoT) devices with limited resources
- (3) automatically generated codes for constructing ML models
- (4) creating innovative procedures for reliable management of machine learning system growth to boost efficiency and dependability



(5) deep learning technologies are being increasingly used in goods for all markets and uses

(6) increasing application of generative adversarial network-based models for various image processing applications, including image enhancement and searching

(7) increased use of unsupervised learning-based systems, which operate with little to no human intervention

(8) use of reinforcement learning based systems,

(9) emergence of learning-based systems based on few shots, if even zero shots.

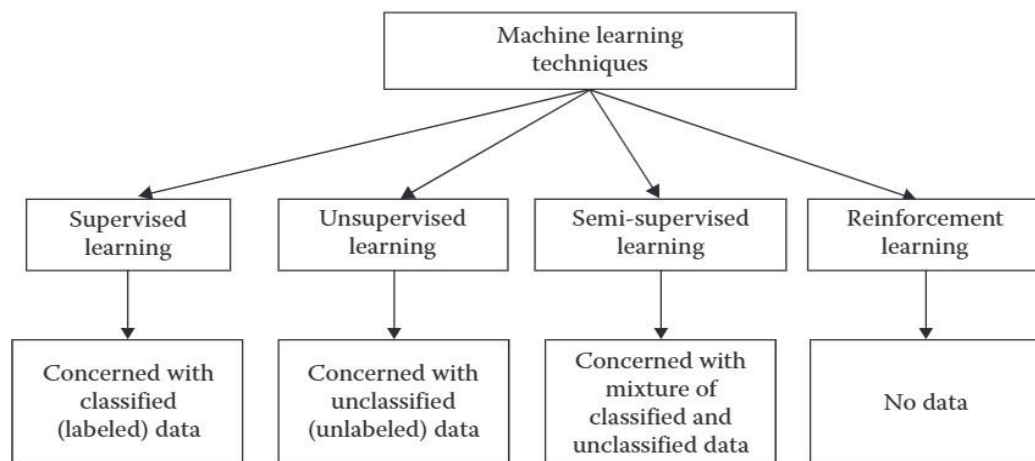


Fig 1

## TYPES OF LEARNING

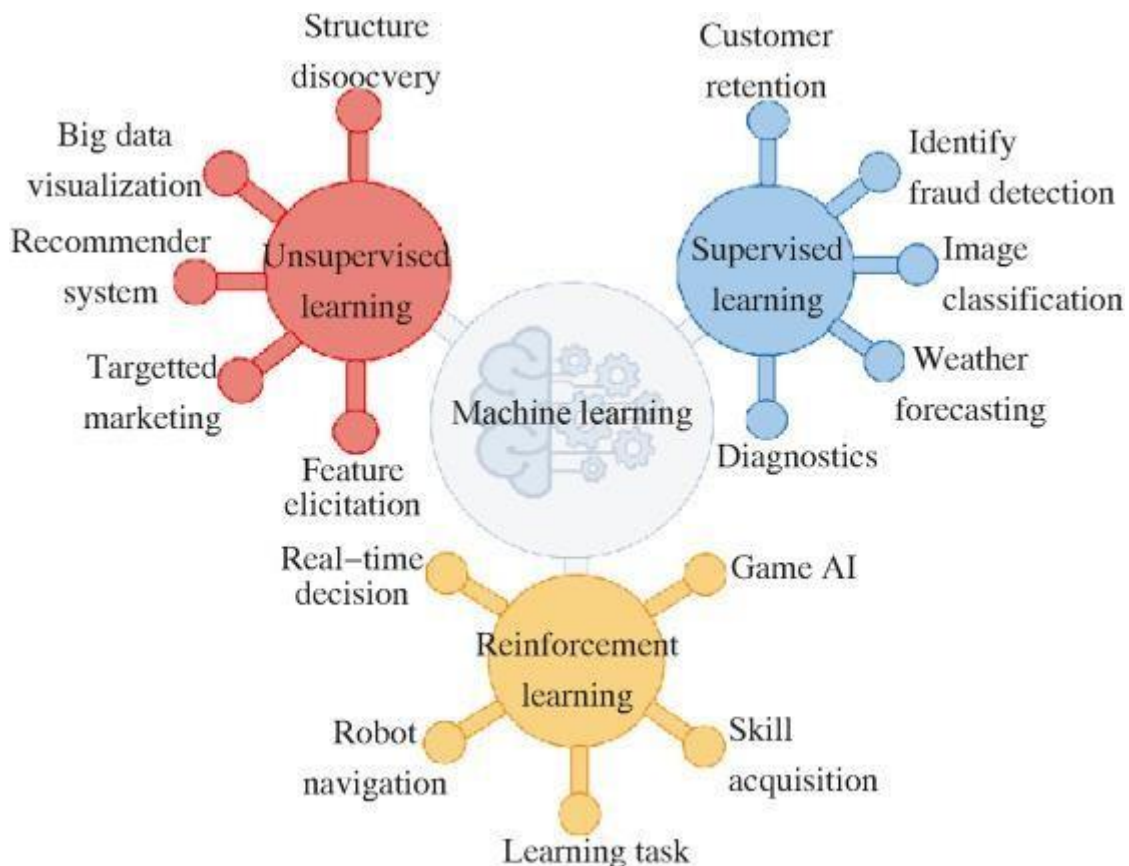
Machine learning is the subdivision of Artificial Intelligence (AI) that concentrates on developing Techniques and algorithms enabling computers to acquire knowledge from datasets advance drawing from previous experiences without requiring complex programming for every task. In short, machine learning (ML) trains systems to think and understand like humans by using the given data set.

Now, let's look at the various types of machine learning algorithms that will be essential for future requirements. ML is commonly employed as a training system to learn from past mistakes and improve performance in the future. Machine learning can be used to make predictions for large data sets. To get profitable and beneficial opportunities, it is beneficial to generate quick and correct outcomes.

We would like to learn a function in two main settings. For the  $m$  samples in the training set,  $E$ , we are aware of the values of  $f$  in one method known as supervised learning. We presume that a hypothesis,  $h$ , will be a good approximation for  $f$  if it can be found for the members of  $E$  that closely agrees with  $f$ . This is especially true if  $E$  is big.

The other is that unsupervised learning techniques can be used to solve taxonomy problems where meaningful categories need to be created for data classification.

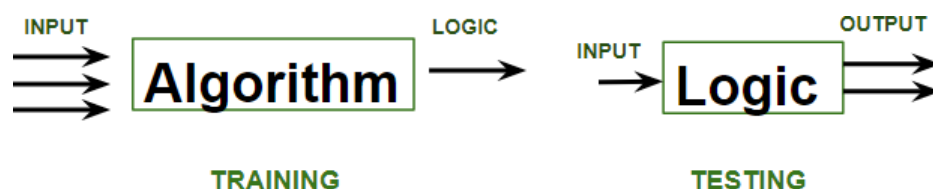
Fig. 2



## Supervised learning

Numerous fields, including business, healthcare, and finance, use this machine learning technology extensively. This kind of machine learning involves training algorithms on datasets so they can predict or decide based on input data. A data list with input and output data is used to teach the map. In order to generate precise predictions about incoming data, the algorithm attempts to understand the relationship between input and output data.

Fig 3

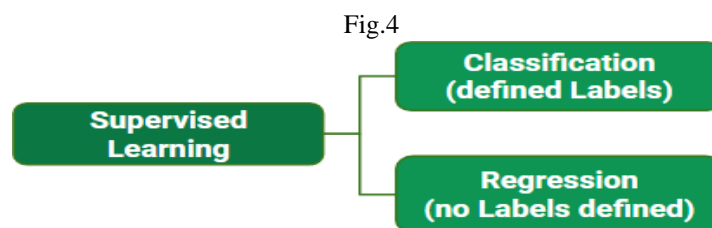


Supervised learning is where a model is trained on an assemblage of information. A data list is a file that contains both input and output parameters.

Registered materials used in educational supervision are in the nature of ideas and written materials. Inputs are the features or characteristics of the data used to make predictions, while outputs are the expected outcomes or goals that the algorithm is trying to predict

## Types of Supervised Learning Algorithm

Regression and classification are the two basic divisions of supervised learning. Regression teaches algorithms to forecast normal values, such home prices or city temperatures. In distribution, algorithms are trained to forecast how various items or catalogs will be distributed, including whether or not a client would buy a particular product. But a lot of educational resources are required for educational supervision to be successful. Furthermore, the representativeness and caliber of the training data may have an impact on the model's accuracy.



## LINEAR REGRESSION

A popular statistical technique for simulating the link between a few "explanatory" variables and a genuine, valuable outcome is linear regression. When seen As part of a learning task, the domain set  $X$  represents a subset of  $\mathbb{R}^d$ , where  $d$  is a positive integer, and the label set  $Y$  consists of real numbers. The best approximation for the relationship between variables, such as forecasting a baby's weight based on her age and birth weight, is a linear function  $h: \mathbb{R}^d \rightarrow \mathbb{R}$ , which we would like to learn. This is an illustration of a predictor for linear regression with  $d = 1$ .

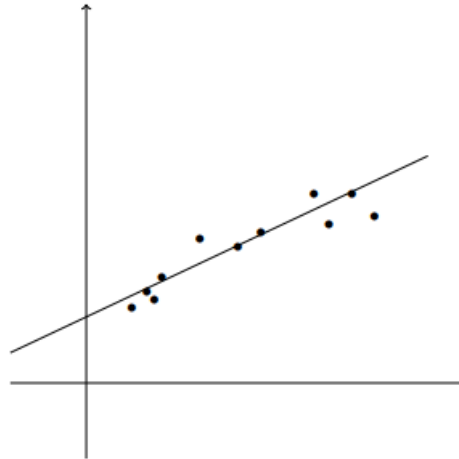


Fig 5

The definition of a loss function for regression is the next step. While the meaning of the loss in classification is simple— $h(x, y)$  only reflects whether or not  $h(x)$  predict correctly  $y$ —in regression, if a baby weighs 2.90 kg, both the 2.90 kg and the 5 kg forecasts are "wrong," albeit it is obvious that we would prefer the 2.90 kg prediction. For this reason, we must specify the extent to which we will be "penalized" for difference between  $h(x)$  and  $y$ . One typical method which is apply the squared loss function, which is,

$$\ell(h, (x, y)) = (h(x) - y)^2.$$

## TYPES OF LINEAR REGRESSION

### Simple Linear Regression

There is just and one dependent variable and one independent variable in this kind of linear regression, which is the most basic kind. The following is equation for basic linear regression:  $y = \beta_0 + \beta_1 X$ , where:

The dependent variable is  $Y$ .

The independent variable is  $X$ .

Slope is  $\beta_1$ , And intercept is  $\beta_0$ .

### Multiple linear regression

This involves one dependent variable and more than one independent variable. The equation for multiple linear regression is:

$$y = \beta_0 + \beta_1 X + \beta_2 X + \dots \dots \beta_n X$$

where:

- $Y$  is dependent variable
- $X_1, X_2, \dots, X_p$  are independent variables
- $\beta_0$  is intercept
- $\beta_1, \beta_2, \dots, \beta_n$  are slopes [by the constraint [24]]

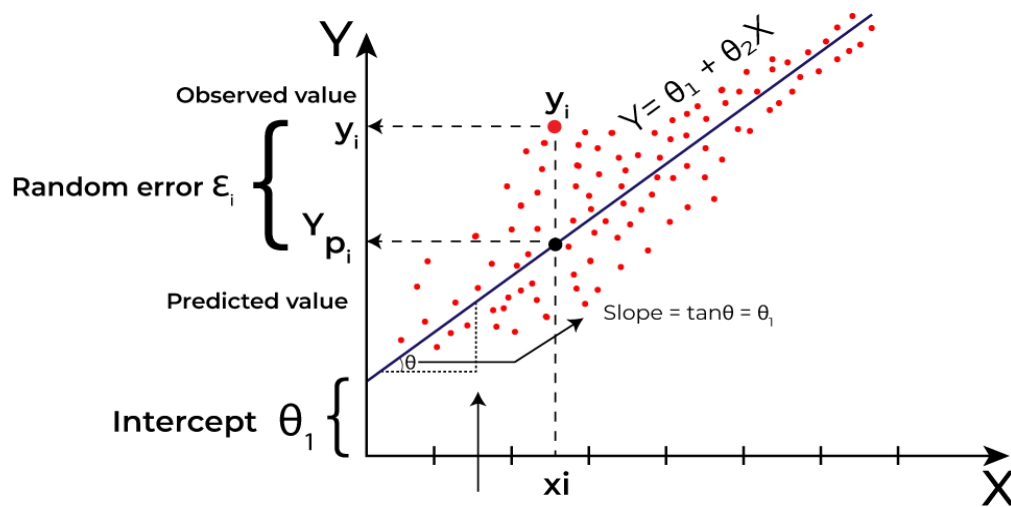


Fig 6

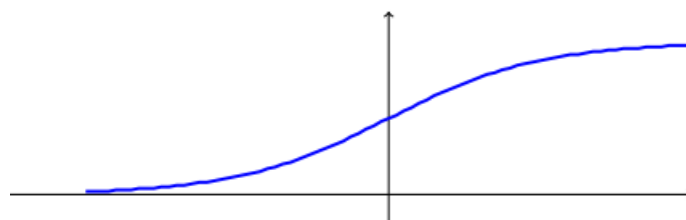
## LOGISTIC REGRESSION

We discover A set of functions  $h$  mapping from  $\mathbb{R}^d$  to the interval  $[0, 1]$  by logistic regression. Nonetheless, classification problems employ logistic regression.  $h(x)$  can be seen as the likelihood that  $x$ 's label is 1. The sigmoid function  $\phi_{\text{sig}}$ , which maps real numbers to the interval  $[0, 1]$ , is composed with the set of linear functions  $L_d$ . is the hypothesis class linked to logistic regression. Specifically, the logistic function—which is defined as—is the sigmoid function utilized in logistic regression.

$$\phi_{\text{sig}}(z) = \frac{1}{1 + \exp(-z)}.$$

The name “sigmoid” means here is “S-shaped,” referred to a plot of function, shown in the figure given below

Fig 7



## KNN (K nearest neighbour)

KNN represents a non-parametric learning method characterized by its slower pace of learning. Typically, a database with data point divided into many classes is used to forecast the categorization of a fresh sample point.

A machine learning approach called k-Nearest Neighbors (kNN) uses the proximity of a new data point to its neighboring data points in the training set to determine its class or value. It is a part of the supervised learning class of algorithms, in which predictions are made using labeled data. The algorithm locates the "k" neighbors who are closest to the new data point, determines which class is in the majority, and computes the average label value as a forecast. Since KNN is well-known for being simple to comprehend and apply, it is frequently used for straightforward classification and regression applications.

KNN has applications in classification. In machine learning, The k-Nearest Neighbors (KNN) algorithm is a yet effective categorization technique. The idea that comparable variables often belong to the same class is adhered to. Based on the label of its closest neighbors in the feature space, the KNN algorithm predicts.

**Non-parametric:** It doesn't assume anything about the distribution of the underlying data. Most data in the real world defies the commonly held theoretical presumptions (as we see in linear regression models, for instance).

**Lazy:** It makes no generalizations using the training data points. Alternatively, we could say that there is either very little or no explicit training period. Unlike so-called eager learning algorithms, which can discard training examples after learning them and carry on learning without knowing the test example,

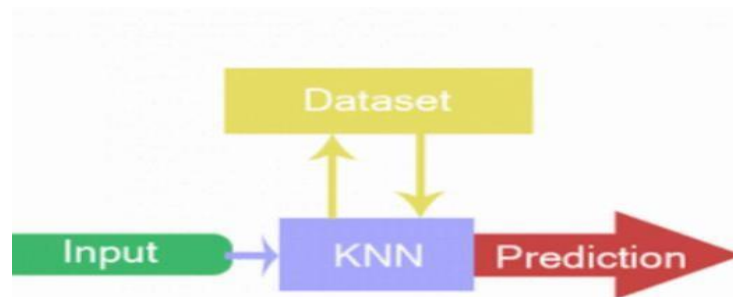


Fig 8

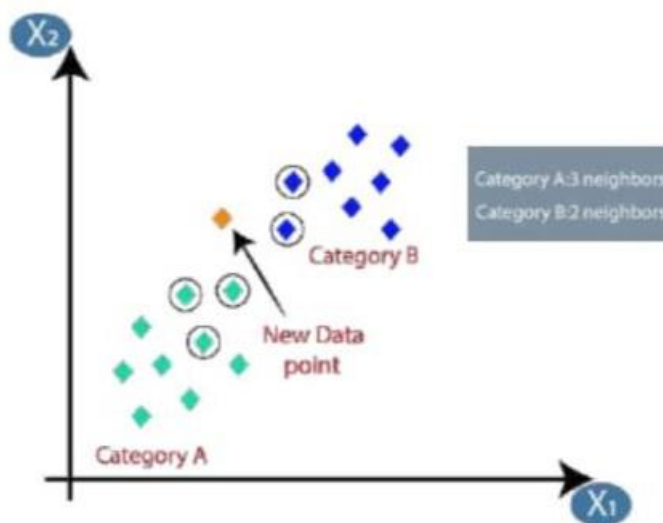


Fig 9

The measurement of data point similarity is the fundamental concept of KNN. The distance between the feature vectors of the variables can be used to accomplish this. The type of data and the problem we need to solve will determine the distance metrics, like Manhattan distance or Euclidean distance, to use. The concept of "nearest" between two data points is defined by the distance measure.

In the KNN algorithm, the value of parameter 'k' determines the quantity of neighboring data points taken into account for generating a prediction in a model. For example, if k is set to 4, the algorithm will find the four nearest neighbours to a given data set. The class label which occur most frequently among these neighbour is assigned to the data point which is being classified.

The main advantage of KNN is its simplicity and interpretability. The algorithm doesn't make strong assumption about underlying data distribution and can be easily implemented.

Furthermore, KNN's versatility in selecting the distance metric and K value makes it adjustable to many scenarios.

When utilizing KNN, there are few thing to keep in mind and limitations. The distance metric and the choice of k can have an impact on how well a KNN performs. Additionally, the approach could be computationally demanding, particularly when handling a lot of datasets. When dealing with issues involving continuous features and balanced class distributions, KNN is more appropriate.

Finding the Value of k: Determine the value of k: Decide on the number of nearest neighbours (k) when making forecasts, it's important to take into account. This can be based on domain knowledge or using techniques like cross-validation to find the optimal value

Finding Neighbours: Identify the k nearest neighbours: Arrange the distances in ascending order and choose the k instances with the smallest distances to the test instance.

## DECISION TREE

For supervised classification, a decision tree is still a straightforward paradigm. Classifying a single distinct targeted feature is its primary application. Every internal node conducts a boolean test on a feature that is input (a test typically consists of more than two options, although these can be transformed into a sequence of Boolean tests). The input feature values are labeled on the edges. A value is specified for the target feature by every leaf node.

### The Decision Tree Learning Algorithm

With the data at hand, how can a decision tree be constructed? Prior to anything else, we must choose which input feature tests to do first. Then, by dividing the example each time an input feature is tested, we may create a decision tree given the order in which the input features are tested. Let's examine the Jeeves training set once more. Every row serves as an illustration. Ten instances are provided. We have five feature values for each example: day, outlook, temperature, humidity, and wind. Since Day varies depending on the example, it is actually not a useful feature. We shall hence concentrate on the remaining four input features. The decision made by Jay over whether or not to play cricket will be our main focus.

The decision tree is a versatile and strong model. We are able to locate and produce several decision trees inside a given data set. Consequently, when choosing which tree to create, there are a few things to consider. Initially, distinct decision trees will result from testing the input features in different orders. Which sequence then ought should we follow? Finding the ideal order in the search space is difficult because there are a lot of different orders.

Considering, we'll take a greedy stance. We shall select the most avaricious option at each stage rather than figuring out the ideal sequence for testing the features. You might be asking yourself, what distinguishes choosing the best option from the best one?

The primary distinction is that the myopic approach just considers the current action, whereas the optimal strategy takes the future into account. In order to produce an ideal tree, we must consider the potential effects of our feature selection at each stage on our final decision. The myopic strategy, on the other hand, simply considers what feature to test at the current stage and has no regard for the future. Let's now assume that we have selected a myopic strategy for the feature testing order. We now have an additional option. Should we continue to develop the tree all the way to maturity or should we halt at some point?

Remember how we talked about over-fitting? A smaller tree may perform better since it may generalize better to test data that has not yet been seen, as the larger tree may overfit the training set. We need to assume something, or have a prejudice, regarding which tree to answer this question. The Occam's razor principle, which holds that the simplest model or hypothesis is most often the best model, can serve as the foundation for one possible assumption. We would advise choosing a tree that is smaller than the entire tree based on Occam's razor. We still have a ton of alternatives in this regard. For example, we can expand the tree until it reaches a given depth or number of nodes.

Decision trees essentially use a sequence of true/false questions to determine which class a given sample belongs to. Here is an illustration of a decision tree that someone may utilize in the real world to choose what to do on a certain day.



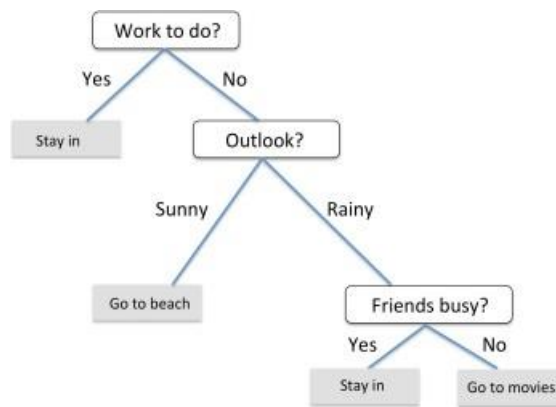


Fig 10

While the query in this picture is based on a categorical variable, when the features are continuous, we can pose questions based on numbers, such as " $x_2 < 6$ ?" Starting at the root node, we pose a query that divides the data according to a feature to determine whether or not the information gain is maximized in order to construct the decision tree. This is what we do for every single node until the decision tree is able to classify all of the training data.

## XG BOOSTING

XGBoost an abbreviation for eXtreme Gradient Boosting is a highly efficient machine learning algorithm known for its speed and precision. It's part of the boosting algorithms family, which are type of ensemble method that amalgamate the predictions of several weak learners.

### Characteristics of XGBoost:

1. **Ensemble Method:** XGBoost concepts a predictive model by participating the forecasts of many individual model typically decision trees.
2. **Gradient Boosting:** XGBoost employs a boosting technique to create an extremely accurate ensemble model. Each succeeding weak learner aims to correct the errors made by its precursors.
3. **Handling Missing Values:** Because of its effective method for handling missing values, XGBoost can handle real-world data without requiring a lot of pre-processing.
4. **Parallel Processing:** XGBoost enables concurrent processing, enabling it to train model on huge datasets within a practical timeframe. [by [25]]
5. **Regularization:** XGBoost enhances outdated gradient boosting by joining regulation components into the objective function, which improves simplification and prevents overfitting.
6. **Tree Pruning:** XGBoost builds trees level by level, evaluating at each level whether adding a new node (split) improves the overall objective function. If not, the split is node.

XGBoost is commonly used in various machine learning tasks, counting regression and classification. It is popular due to its ability to handle large data sets and deliver top presentation.

## **UNSUPERVISED LEARNING**

**Unsupervised learning** is a type of machine learning that learn from unlabelled data. This mean that object has no prefix or no group. The goal of unsupervised learning is to find pattern and relationship in data without any specific direction or information given. Information will take action without guidance. The instrument used here is to group different data based on resemblances, patterns, and differences without requiring preceding training on the data.

Unlike inspection, there is no trainer, so the machine is not trained. The machines are thus left to find patterns which are hidden in nameless files on their own.

We can use unsupervised learning to analyze human data and categorize different groups based on human typical and behavior. This group can affect many species and allows you to categorize diseases without following to existing forms.

As the name suggesting, unsupervised learning is type of machine learning technique in which the model is not supervised using training dataset. Instead, the model itself discovers hidden patterns and insights in the given object. This can be likened to the learning that occurs when the human brain learns something new.

Unsupervised learning is a sort of machine learning in which the model is trained([by 26]) using nameless data and is allowed to run on the data disadvantaged of supervision. Trace learning cannot be directly applied to recovery or classification problems because, different trace learning, we have input data but no output data. The persistence of unsupervised learning is to find the basic structure of the data set, group the data according to likenesses, and represent the data set in a crushed format

## **WORKING OF UNSUPERVISED LEARNING**

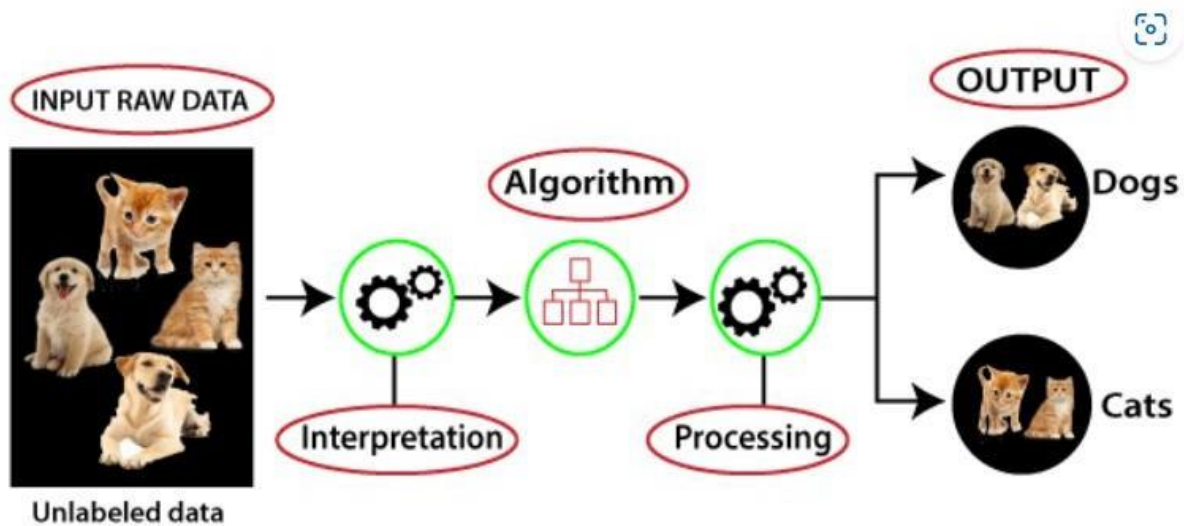


Fig 11

## TYPE OF UNSUPERVISED LEARNINGS .

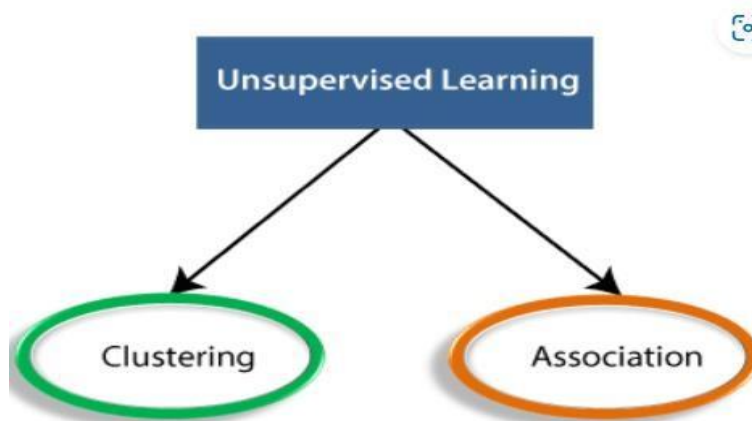


Fig 12

### Clustering

An unsupervised machine learning technique called clustering organizes and classifies various items, data points, or descriptions according to similarities or patterns. Clustering functions with unlabelled data, in contrast to supervised learning, which uses labelled data with a target variable.

### Objective

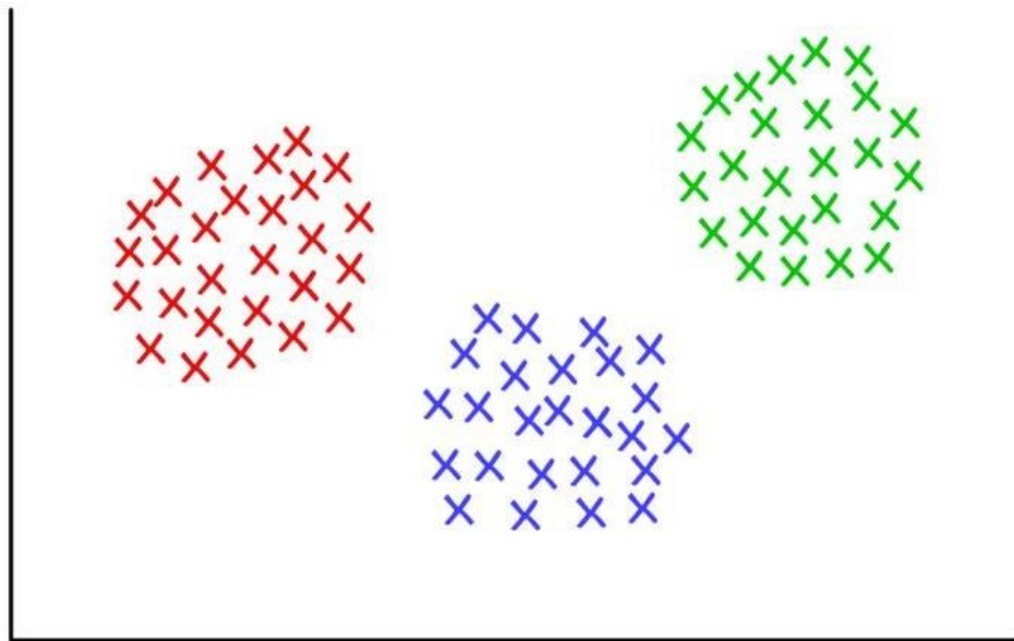
The primary objective of clustering is to combine similar data points that have been gathered, which enables us to find inherent structure in the data. Finding relationships, patterns, and hidden insights is aided by it.

## Types of Clustering:

**Hard Clustering:** Every data point belongs entirely—or partially—to a cluster. Each data point will either belong to cluster 1 or cluster 2 if, for example, we need to cluster four data points into two clusters.

**Soft Clustering:** Soft clustering calculates the probability or possibility that a point would fit into many clusters as opposed to assigning every data point to a single cluster. It offers a more complex perspective on cluster engagement.

Fig 13



## ASSOCIATION

The goal of association analysis, a potent technique in unverified learning, is to find patterns or relationships among the variables or items in a dataset. Let's examine the important details.

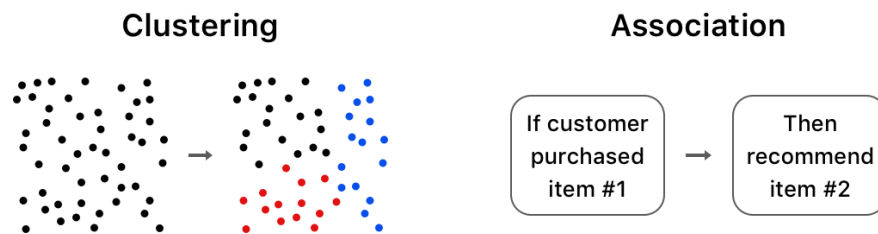
### Objective:

- **Descriptive, Not Predictive:** Association analysis is descriptive rather than predictive. It focuses on discovering interesting relationships hidden within large datasets.
- **Rule-Based Approach:** It represents relationships in the form of rules or recurrent item sets.

## Applications:

- **Retail Market Basket Analysis:** One common application is analysing retail baskets or transactional datasets. By classifying which items are frequently purchased together, retailers can enhance product placement, cross-selling, and raises.
- **Web Usage Mining:** Association examination helps uncover patterns in user behaviour on websites, such as identifying pages frequently visited together or common map reading paths.
- **Healthcare:** In healthcare, it can disclose associations between symptoms, diseases, or drug prescriptions.

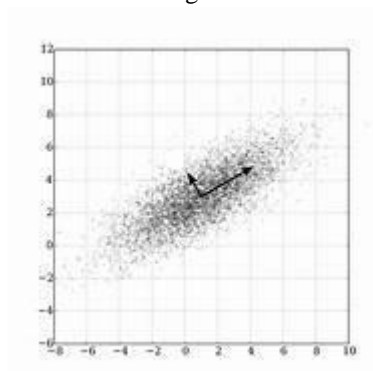
## UNSUPERVISED LEARNING



## Principal Components Analysis (PCA)

Principal Component Analysis is a potent technique for reducing the dimensionality of large datasets. Many issues arise when the number of features or dimensions increases, such as overfitting, longer computation times, and decreased model accuracy. The "curse of dimensionality" is the culmination of these problems.

Fig 14



## Objective:

- **Dimensionality Reduction:** PCA aims to decrease the number of input features though retaining as much unique data as possible.
- **Orthogonal Transformation:** It transforming correlated variables into a group of uncorrelated variables, known as principal components.

### How PCA Works:

- **Variance Maximization:** PCA identifies a fresh set of axes (principal components) that efficiently capture the highest variability present within the dataset
- **Eigenvalues and Eigenvectors:** PCA computes eigenvalues and eigenvectors from the covariance matrix of the original features.
- **Projection:** Data points are projected onto the principal components, effectively reducing the dimensionality.

### Applications:

- **Feature Engineering:** PCA is widely used for exploratory data analysis and analytical modeling.
- **Visualization:** It simplifies data visualization by reducing dimensions.
- **Noise Reduction:** By directing on high-variance components, PCA filters out noise.

### Steps in PCA:

- Compute the covariance matrix of the original features.
- Calculate eigenvalues and eigenvectors.
- Sort eigenvalues in descendant order.
- Choose the k most significant eigenvectors, where k represents the desired dimensionality reduction
- Perform a projection of the project data onto the specified eigenvectors

An exponential growth in the number of characteristics or dimensions in the data collection is needed to provide statistically significant results. When working with large amounts of data, this can result in issues including machine learning models being overfitted, faster computation times, and decreased precision—a phenomenon known as the "curse of the problem." The number of possible combinations rises with the number of features, which makes it more difficult to sample data and perform costly operations like sorting or grouping. Furthermore, certain machine learning algorithms are able to comprehend size and need more data in order to reach the same accuracy level with less data. Extraction of features. A removal technique e called "dimension reduction" seeks to minimize the number of entries while retaining as much of the original data as feasible.

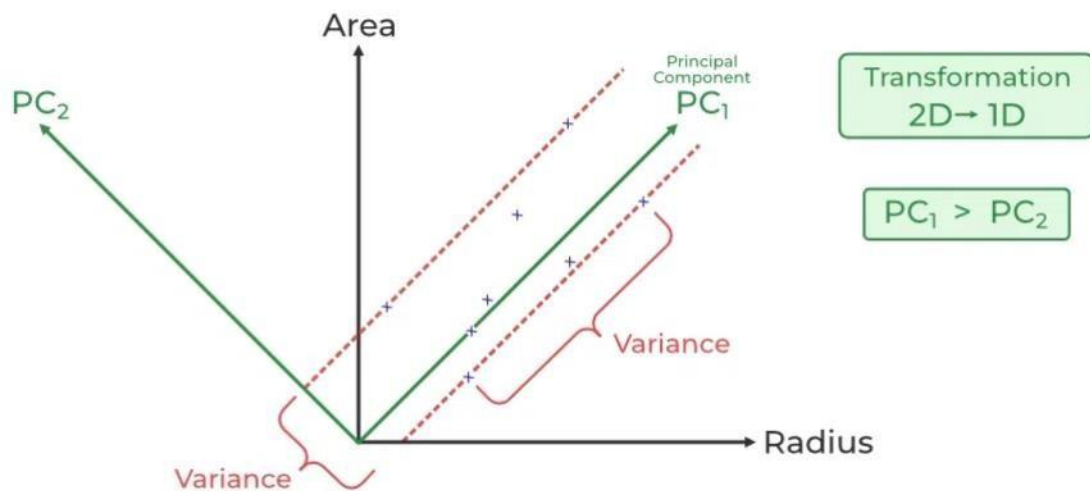


Fig 15

## K MEANS CLUSTERING

The cluster is denoted by  $K$ , and data points are assigned to one of the  $K$  clusters according to how far they are from the cluster. The group centers are first placed in random locations in space. Based on how far away each data point is from the cluster center, it is allocated to one of the clusters. A new cluster center is assigned following the payment of each point owed to a cluster. Until the proper group is identified, this process is repeated. We begin the analysis assuming that there are a certain number of groups and that the elements must be placed in one of the groups.

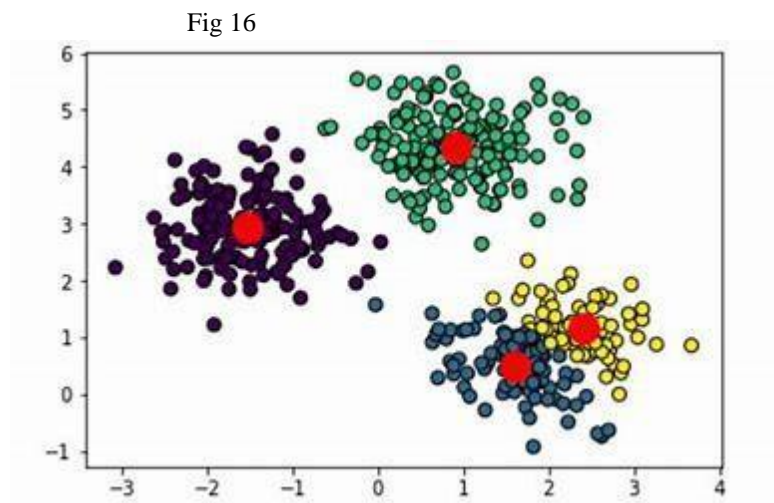


Fig 16

### How k-means clustering works?

We have a dataset consisting of items characterized by specific features and their corresponding values, similar to vectors. Our objective is to group these items together effectively. To accomplish this task, we employ an unsupervised learning technique known as the K-means algorithm. Here, the variable ' $K$ ' represents the desired number of clusters or groups into which we aim to classify our objects. Conceptually, we can visualize the items as

points residing within an  $n$ -dimensional space. The algorithm partitions these objects into ' $K$ ' distinct groups or clusters based on their similarity. To calculate that similarity, we shall use the Euclidean distance as a unit of measurement.

The algorithm operates in the following manner:

Initially,  $k$  points, which are also known as cluster centroids or means,—were reset at random.

2. After classifying every item to the near mean, we update and Identify the average coordinates medians of all the object that have been classified in that cluster thus so far.

3. After a set quantity of repetitions, we repeat the procedure to obtain our clusters.

Given that the "points" mentioned above represent the average values of the items they encompass, they are termed as means. There are various options available for establishing these means. Resetting the mean at random points in the approaching the dataset logically is essential. Initializing the means at random values within the constraints of the data set is an additional technique.



## CHAPTER 2

### BRIEF ABOUT BREAST CANCER

Machine learning has vast request We can apply it wherever and to any kind of data Here we'll its application over the data of breast cancer patients . Before applying different models to data set we'll brief about breast cancer .

### INTRODUCTION

Compared to other known malignancies, the statistics on breast cancer are horrifying. An increasing number of patients in India are being diagnosed with breast cancer. Breast cancer affects everyone, regardless of age, gender, or religion. For Indian women, it is a type of cancer that occurs most frequently.. In India, it has been noted that one in every 22 women will develop breast cancer. One of the two people who get it passes away. India's breast cancer situation differs from that of the West. In this case, it is detrimental to younger women, and over half of them give in at an advanced stage. It is predicted that by 2030, breast cancer would kill more Indian women than all other cancers combined. Pre-emptive detection is the most effective strategy for managing this scourge.

. Mammography is now the most effective early detection procedure and it has been shown to work. Due to earlier detection and treatment, a decline has been noted in both fatalities from critical breast cancer and routine mammography patients. Routine examination programs, however, offer the radiologist access to ton of exams, which increases the likelihood of a mistaken diagnosis.by the constraint[23] Mammograms can sometimes be quite challenging to read and understand in their original form. Mammograms are breast exams using low-energy X-rays that aid in the early diagnosis of cancer. Electronic images of the breasts are provided via digital mammography. The device assists in identifying the presence of malignant cells and has the ability to identify even the tiniest tumors.

Our goal is to use mammography images to produce a dependable tool for the detection of breast tumors or any other type of cancer. This thesis uses various image processing techniques as its foundation. An overview of breast cancer the problems and obstacles and challenges faced during the detection of breast cancer opens this chapter. The rationale for the study and its aims then discussed.

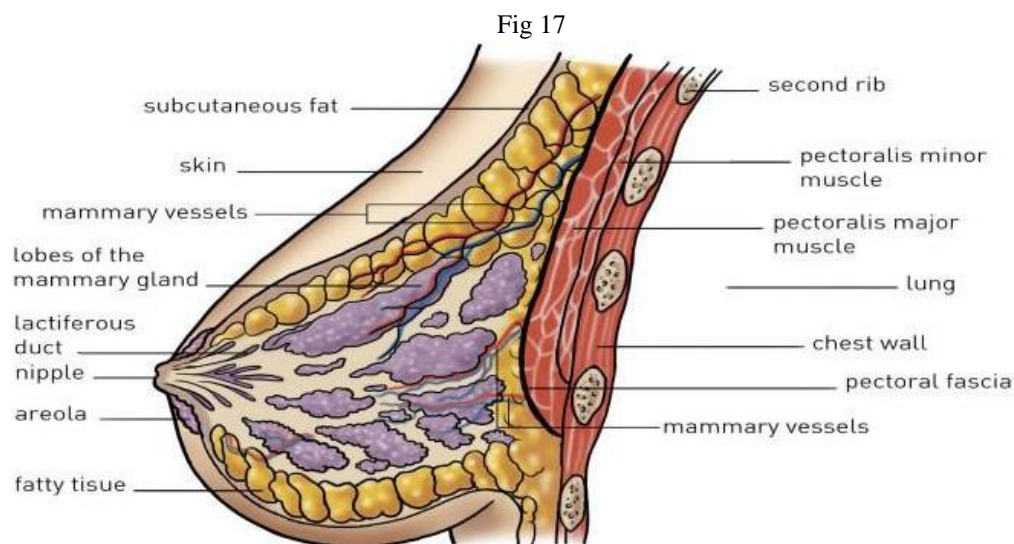
### MOTIVATION

The causes of he cancer rates in India resemble those found seen in other parts of the world by the constraint [23]: carcinogens, or substances that cause uncontrollably and disorderly cell proliferation, are caused by biological, chemical, and other environmental factors. Under unusual circumstances, carcinogens interact with the DNA of healthy cells, starting a series of intricate, multi-step processes that lead to unchecked cell division or tumor development. Cancer may be caused by internal or external factors. While internal variables include hormones and immune systems, external variables are environmental factors like food, smoking, radiation, or other infectious agents. A significant divergence has been noted and documented due to lifestyle decisions, fashion preferences, and dietary practices.

Finding the cancer is essential to developing the best possible treatment. An unusual bulge surrounding breast tissue or an unusual discharge from the breasts are examples of subjective indicators or superficial symptoms that are initially used to diagnose breast cancer. In order to confirm the presence of cancer cells, additional tests are performed following the manual screening. While there are alternative methods available, mammography and clinical breast examinations performed by a qualified medical professional remain the gold standards for confirming the presence of cancer. However, it still doesn't offer a success percentage of 100%. These kinds of incorrect diagnoses are known as "false positives," and they result in additional testing and diagnostic procedures that could be upsetting for the patients. . Therefore, in order to effectively detect cancer cells, we need new approaches that produce better and more precise results.

## ANATOMY OF BREAST

Corrupt cellular division is the root cause of breast cancer. As a result, it is essentially a cancer that arises from ineffective cell division. Understanding the anatomy and physiology of the breasts is essential to understanding breast cancer. Each breast is made up of 15 to 20 parts termed lobes that surround the nippular area radially or like spokes on a wheel. The smallest portions of these lobes are referred to as lobules. There are tiny "bulbs" to create milk at the end of these lobules. Milk is delivered to the nipples by tiny tubes called ducts that link these bulbs. There is fat in the voids created by the lobes and channels.



1. Mammary Glands
  - a. The modified sweat glands known as mammary glands are made up of 15 to 20 secretory lobule and duct series.
2. Lactiferous duct

- a. Numerous alveoli are drained by a single, tube-like structure called a lactiferous duct.
3. Connective Tissue Stroma
  - a. We refer to the connective tissue stroma that surround the mammary glands as supporting structures. Its constituent parts are mostly fibrous and fatty. Later, the fibrous stroma creates suspensory ligaments that divide the breasts' secretory lobules and attach the breasts to the pectoral fascia.
4. Pectoral Fascia
  - a. Pectoral fascia is the sheet of fibrous tissue that runs from the breast to the pectoralis major muscles. It serves as the suspensory ligament's attachment site.

## **Breast Conditions**

The following are several breast conditions that may impact the breasts.

- **Breast Cancer:**

Breast cancer develops when defective or malignant cells proliferate uncontrollably and Cancer cells can spread to various parts of the body. While breast cancer affects individuals of both genders, it tends to occur more frequently in women and common in the former. Breast cancer can be identified on the surface by looking for lumps, changes in skin tone, or red nipple discharge.

.

- **Breast Calcification:**

Microcalcification is the term for tiny calcium deposits that are difficult to see in mammograms because of their small size and weak contrast. The size of these calcium deposits ranges from 0.33 to 0.7 mm. Because isolated microcalcifications are smaller than cluster microcalcifications, they are more difficult to detect. Three or more microcalcifications with an area of one centimeter make up a cluster of microcalcification. Finding microcalcifications is essential for detecting cancer in its early stages. Being able to distinguish between benign and malignant microcalcification is crucial because the former can lead to cancer.

- **Simple Breast cyst:**

This is a benign or tumor that resembles a sac and is typically Containing liquid . Typically, women in their 30s or 40s exhibit it. Breast cysts have the potential to drain fluid and cause tenderness. 15

- **Phyllodes tumor:**

This is a rare tumor that grows quickly and usually affects women in their 40s. It can be benign or malignant. The tumor is often substantial in size and appears on ultrasonography to be a fibroadenoma.

- **Breast fibroadenoma:**

It's a common solid tumor that isn't malignant and causes a unproblematic but moveable mass within the breast[23].

- **Fibrocystic breast disease:**

A mass within the breast that is not malignant but is uncomfortable and fluctuates in size during the menstrual cycle by the constraint[23].

## **Types of cancer**

### **A. Benign**

These tumors do not pose a risk to the patient's life. They do not cause cancer. A benign tumor has a low likelihood of regrowing and is easily removed. This is as a result of the benign tumors' cells not spreading.

### **B. Malignant**

Malignant tumors are cancerous, in contrast to benign ones. Malignant tumor cells are aberrant and divide quickly. It is possible for malignant tumor cells to spread to other body parts and develop new cancers. These cells behave aggressively, attacking the surrounding tissues.

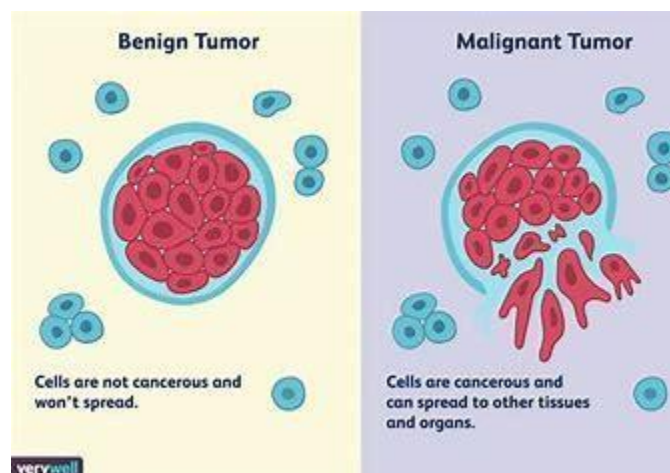


Fig 18

## Stages of cancer

Stages are used to categorize cancerous tumors. Tumor cell characteristics under a microscope are used to define a tumor's stage. A stage aids in identifying the sort of cancer the body is experiencing and in planning the patient's course of therapy. It establishes the extent to which cancer has developed and dispersed throughout the body. There are numerous staging systems available. TNM is the most commonly utilized system. The letter "T" stands for tumor size, the letter "N" for number of affected lymph nodes, and the letter "M" for metastases, or the degree to which cancer has metastasized to other areas of the body in human organs through the circulatory systems.

A lower stage is preferable since it suggests least advanced cancer that is more likely to have stayed within the body which improve treatment outcomes.

Cancer stages are categorized as follows:

Stage 4 involves cancer spreading to an organ different from its origin.

Stage 3 signifies a sizable cancer presence along with involvement in the lymph nodes.

Stage 2 denotes a larger cancer, with uncertainty about its spread to the lymph nodes.

Stage 1 indicates the presence of a small cancer confined to its originating organ.

Stage 0 represents precancerous conditions.

Based on the microscopic appearance of the cells, breast cancer is classified. Cancerous breast cancers predominate. One kind of cancer is carcinoma. The cells that border organs and tissues, such as the breast, are affected by this kind of cancer. Cancer has no known cure, and there is no reliable means to prevent it from occurring. Statistics on breast cancer are horrific when compared to other recognized cancers. An increasing number of patients in India are being diagnosed with breast cancer in individuals across all age groups, genders, and religions. For Indian women, it is the most common type of cancer. It has been noted that one in every 22 Indian women will acquire breast cancer. One of the two people who get it passes away. India has a distinct breast cancer situation. from the west

. In this case, it is detrimental to younger women, and over half of them give in at an advanced stage. It is predicted that by 2030, breast cancer would kill more Indian women than all other cancers combined. The most effective way to deal with this epidemic is to recognize it early. Mammography is now the most effective early detection procedure and it has been shown to work. Due to earlier detection and treatment, a decline has been noted in both fatalities from critical cases of breast cancer and routine mammography patients.

Programs for routine examinations, however, give the radiologist conducts numerous examinations increasing the likelihood of an inaccurate diagnosis.[23]

Mammograms can sometimes be quite challenging to read and understand in their original form. Mammograms are breast exams using low-energy X-rays that aid in the early diagnosis of cancer. Electronic images of the breasts are provided via digital mammography. The device can identify even the tiniest tumors and aids in the identification of malignant cells.

## CHAPTER 3

### APPLYING BREAST CANCER DATA SET ON DIFFERENT MODELS OF MACHINE LEARNING

On the basis of data we have first we'll differentiate it between the types of cancer using feature distribution

```
diagnosis_colors = ['#68b377', '#d66970']

[ ] plt.figure(figsize=(8, 4))

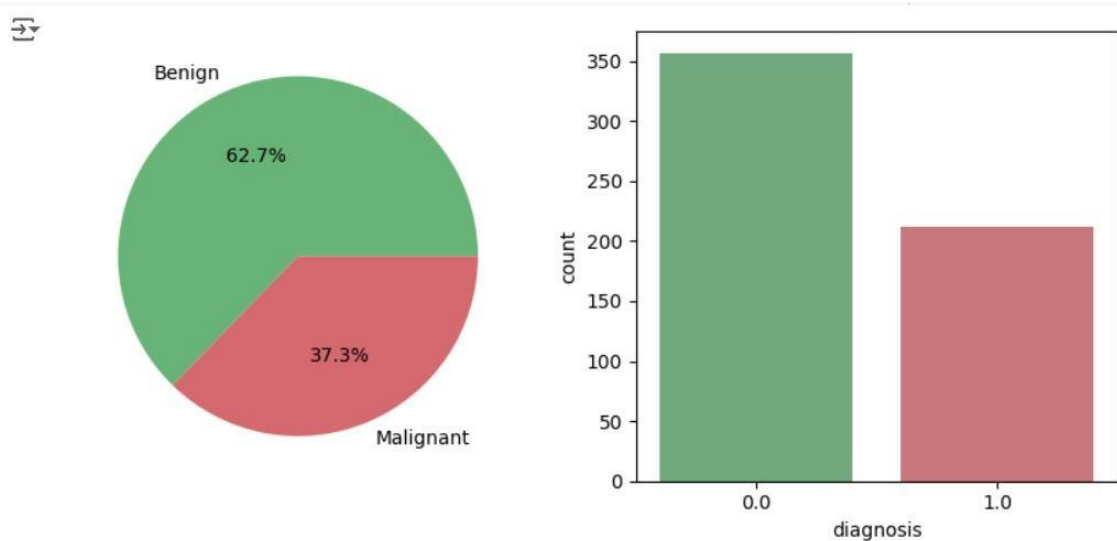
plt.subplot(1, 2, 1)
plt.pie(data['diagnosis'].value_counts(), autopct='%1.1f%%', labels=['Benign', 'Malignant'], colors=diagnosis_colors)

plt.subplot(1, 2, 2)
sns.countplot(data=data, x='diagnosis', palette=diagnosis_colors)

plt.tight_layout()
plt.show()
```

The code will provide us with the graph which makes it easy to read and analysis about the data

Fig 19



## HISTOGRAM

Using histogram graph we'll find the relation of Benign and Malignant with respect to each of its parameter but before finding the relation between parameters we have to find the all the parameter uniquely

Following code will help in to do so

```
data.info()
```

Data columns (total 31 columns):

#	Column	Non-Null Count	Dtype
0	diagnosis	569 non-null	object
1	radius_mean	569 non-null	float64
2	texture_mean	569 non-null	float64
3	perimeter_mean	569 non-null	float64
4	area_mean	569 non-null	float64
5	smoothness_mean	569 non-null	float64
6	compactness_mean	569 non-null	float64
7	concavity_mean	569 non-null	float64
8	concave points_mean	569 non-null	float64
9	symmetry_mean	569 non-null	float64
10	fractal_dimension_mean	569 non-null	float64
11	radius_se	569 non-null	float64
12	texture_se	569 non-null	float64
13	perimeter_se	569 non-null	float64
14	area_se	569 non-null	float64
15	smoothness_se	569 non-null	float64
16	compactness_se	569 non-null	float64
17	concavity_se	569 non-null	float64
18	concave points_se	569 non-null	float64
19	symmetry_se	569 non-null	float64
20	fractal_dimension_se	569 non-null	float64
21	radius_worst	569 non-null	float64
22	texture_worst	569 non-null	float64
23	perimeter_worst	569 non-null	float64
24	area_worst	569 non-null	float64
25	smoothness_worst	569 non-null	float64
26	compactness_worst	569 non-null	float64
27	concavity_worst	569 non-null	float64
28	concave points_worst	569 non-null	float64
29	symmetry_worst	569 non-null	float64
30	fractal_dimension_worst	569 non-null	float64

dtypes: float64(30), object(1)

Here we can see data has total 31 columns So now it will easy to plot histograms with respect to all the parameters

```
▶ y_column = data.columns[0]
  x_columns = data.drop(['diagnosis'], axis=1).columns

▶ plt.figure(figsize=(15, 12))

  for i, col in enumerate(data.columns[1:]):

    plt.subplot(5, 6, i+1).set_title(col)

    x_col = data[x_columns[i]].values
    y_col = data[y_column].values

    plt.hist(x_col[y_col == 0.0], label='B', color='green', alpha=0.5, bins=25)
    plt.hist(x_col[y_col == 1.0], label='M', color='red', alpha=0.5, bins=25)

    plt.ylabel('Counts')
    plt.legend(loc='best')

  plt.tight_layout()
  plt.show()
```

Fig 20



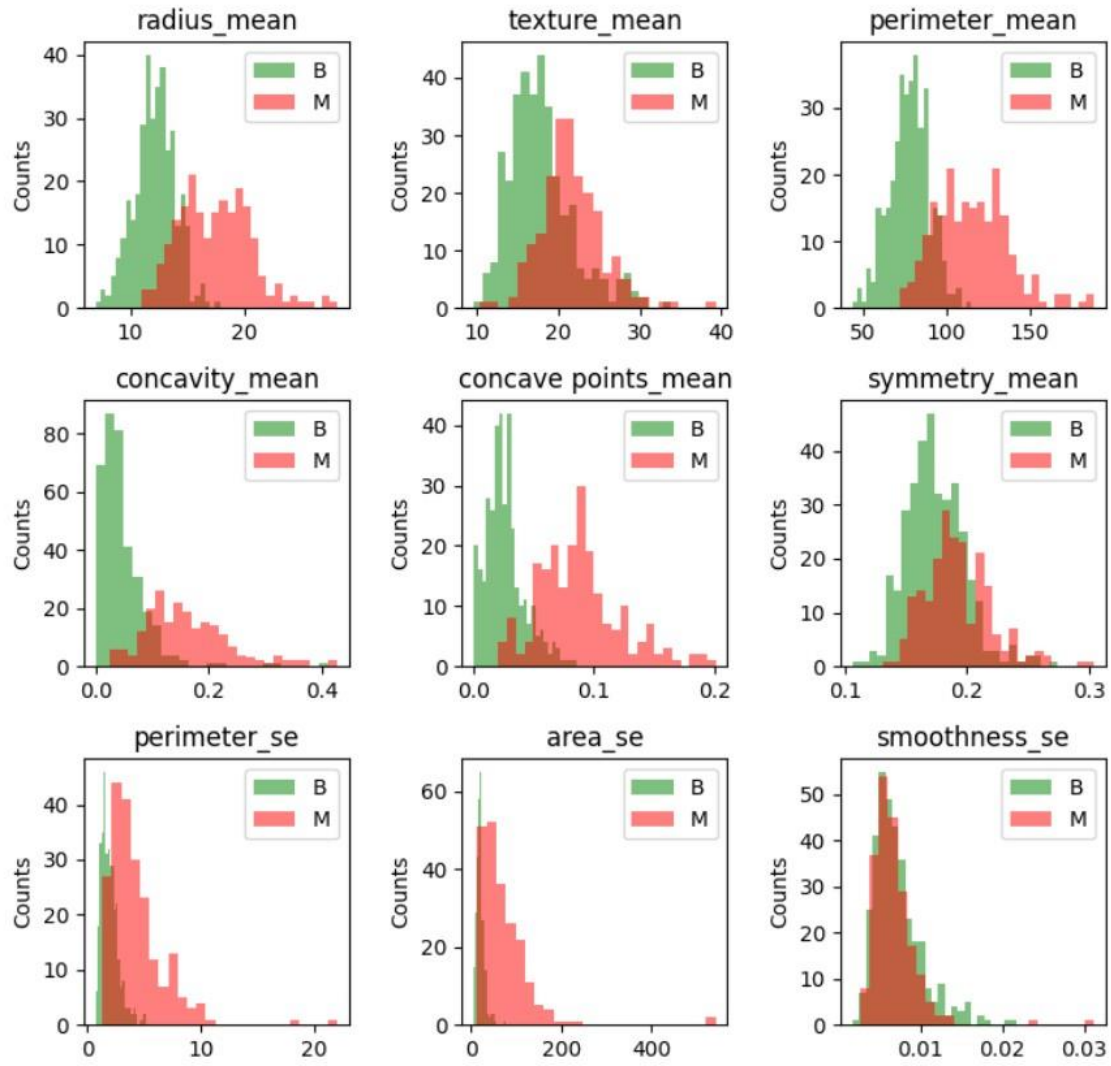


Fig 21

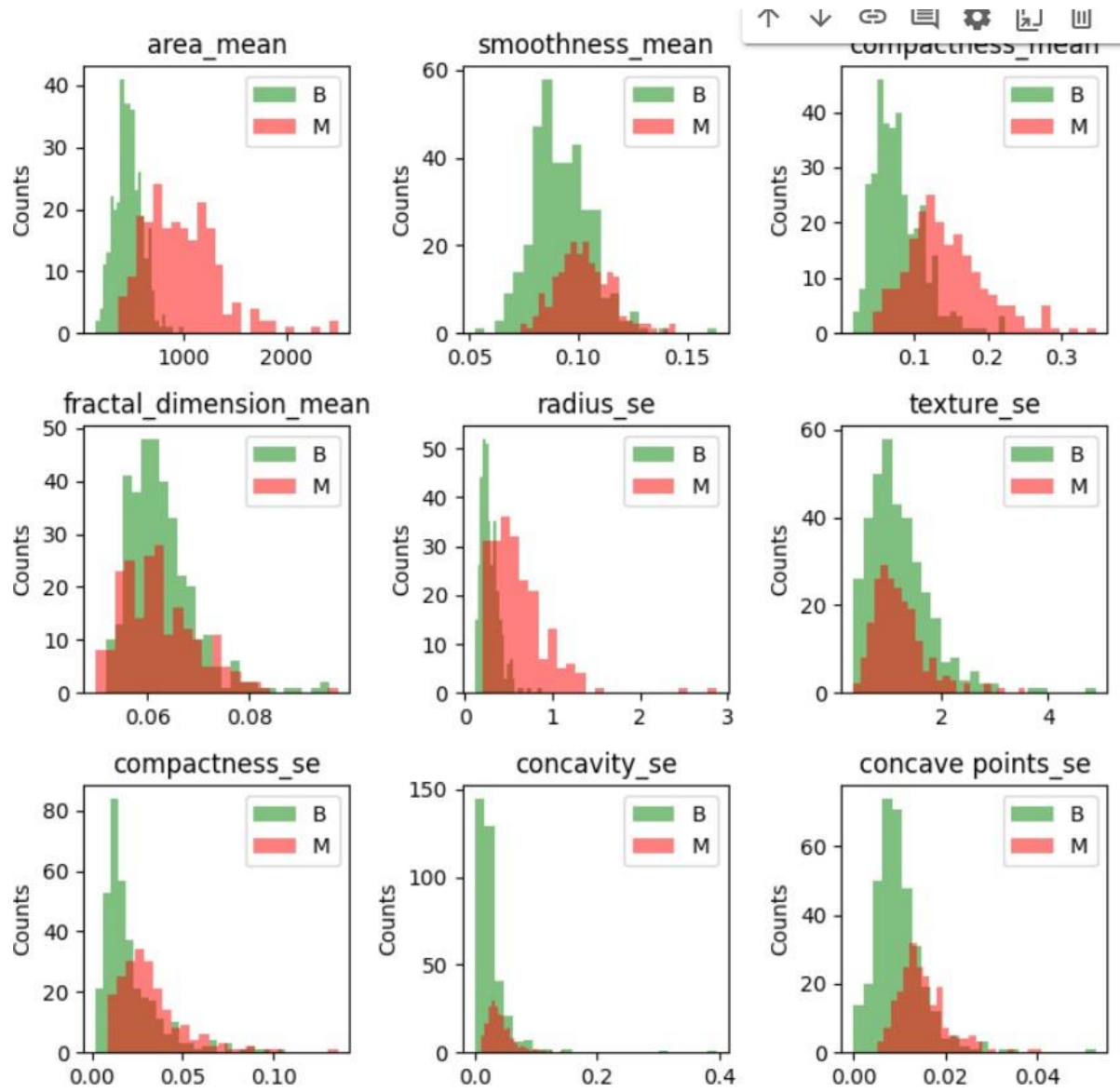


Fig 22

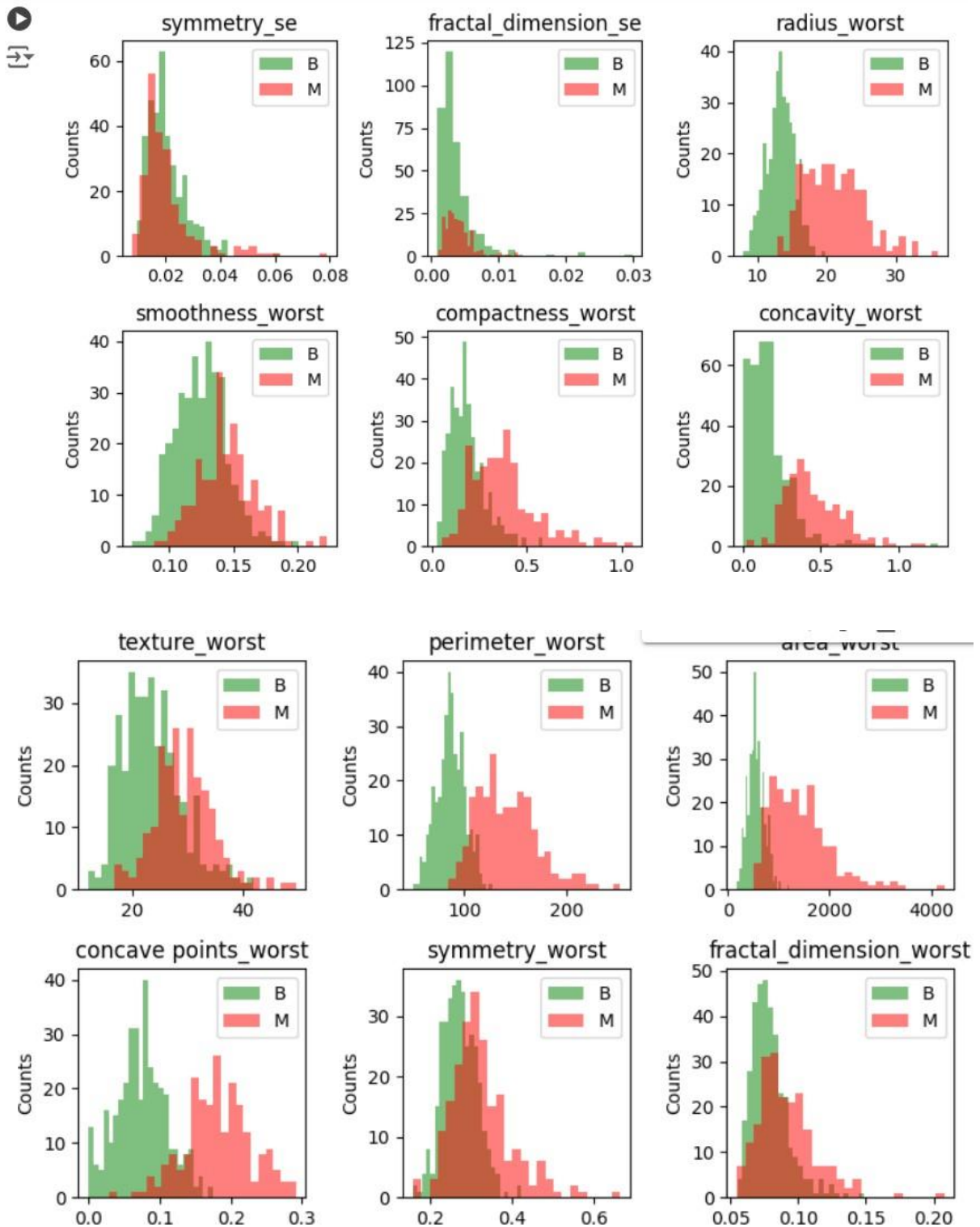


Fig 23

To understand the relationships between different variables within a dataset we can use Correlation matrix

### Why Use Correlation Matrix?:

A tabulated representation illustrating the coefficients" of correlation between a correlation matrix refers to a compilation of different variables .It provides an overview of the connections (correlations) between every potential set of values in a dataset. The correlation

coefficient, which shows how closely two variables are related, is present in each cell of the matrix.

- **Increase Model Accuracy:** Redundancy can be avoided by identifying associated features. It may not be very beneficial to include both features in your model if they are significantly connected.
- **Reduced Computational Cost:** Training and inference require less processing when there are fewer features.

## CODE FOR CORRELATION MATRIX

```
plt.figure(figsize=[12, 10])
sns.heatmap(data.corr(), annot=True, fmt = '.1f', annot_kws={"fontsize": 8}, linewidths=0.25, center= 0.3, cmap= 'coolwarm', square=True)

plt.tight_layout()
plt.show()
```

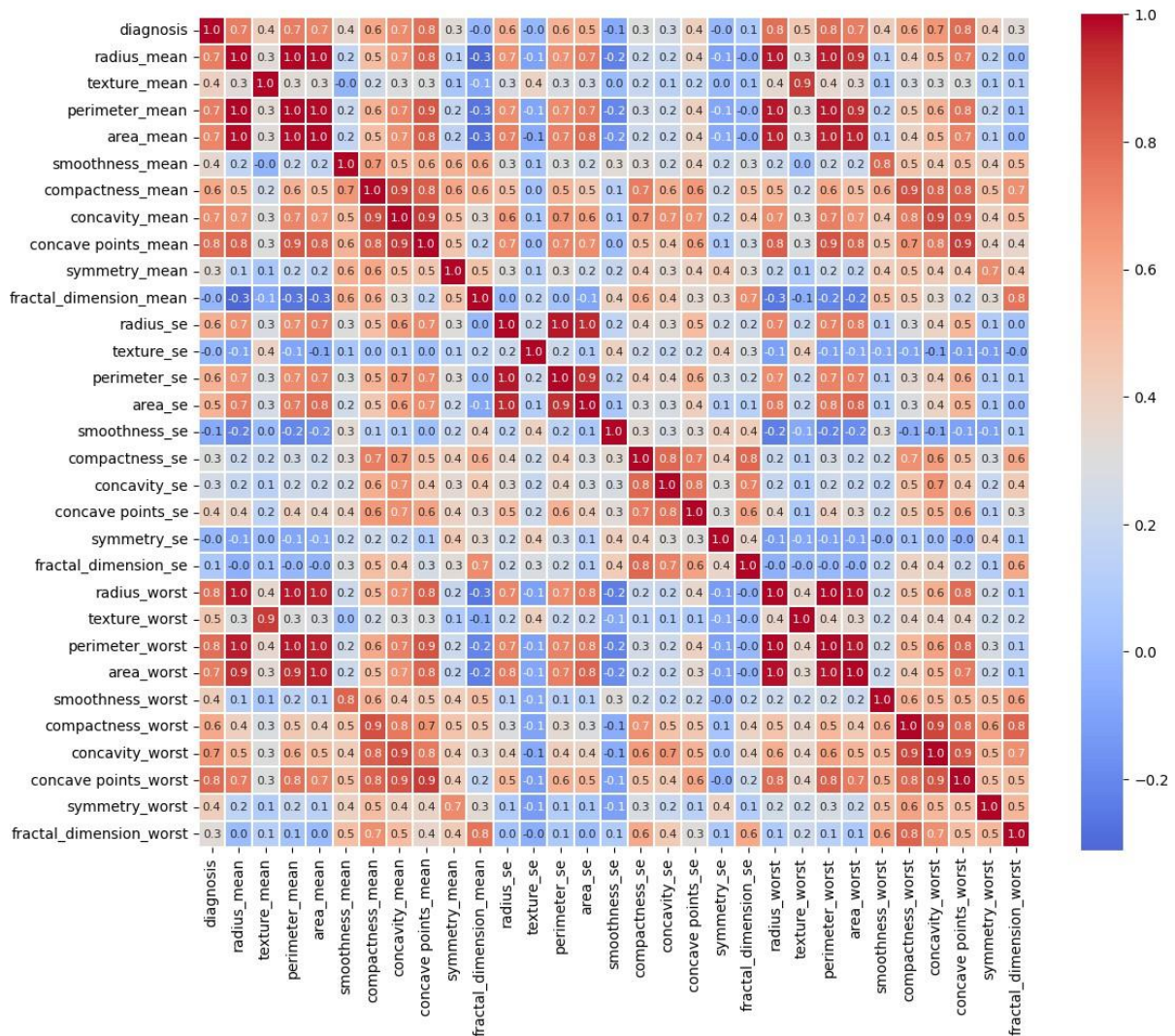


Fig 24

## SCATTER PLOT

A scatter plot is a two-dimensional coordinate system's graphical depiction of data points.

1. Visualizing Relationships: o We can see the relationship between two variables by using scatter plots.

o We can see patterns, trends, and possible correlations by charting data points as individual dots.

2. We can evaluate the correlation among two variables with the use of scatter plots.

3. A high association is shown if the dots cluster around a straight line that has a positive or negative slope.

4. A positive slope denotes a positive correlation, meaning regarding one variable, rises, the other also tends to rise.

5. An inverse link (negative correlation) is shown by a negative slope.

6. The absence of a distinct pattern indicates little to no association.

The code for scatter plot will be little longer as our data has 31 columns

The graph will look like as



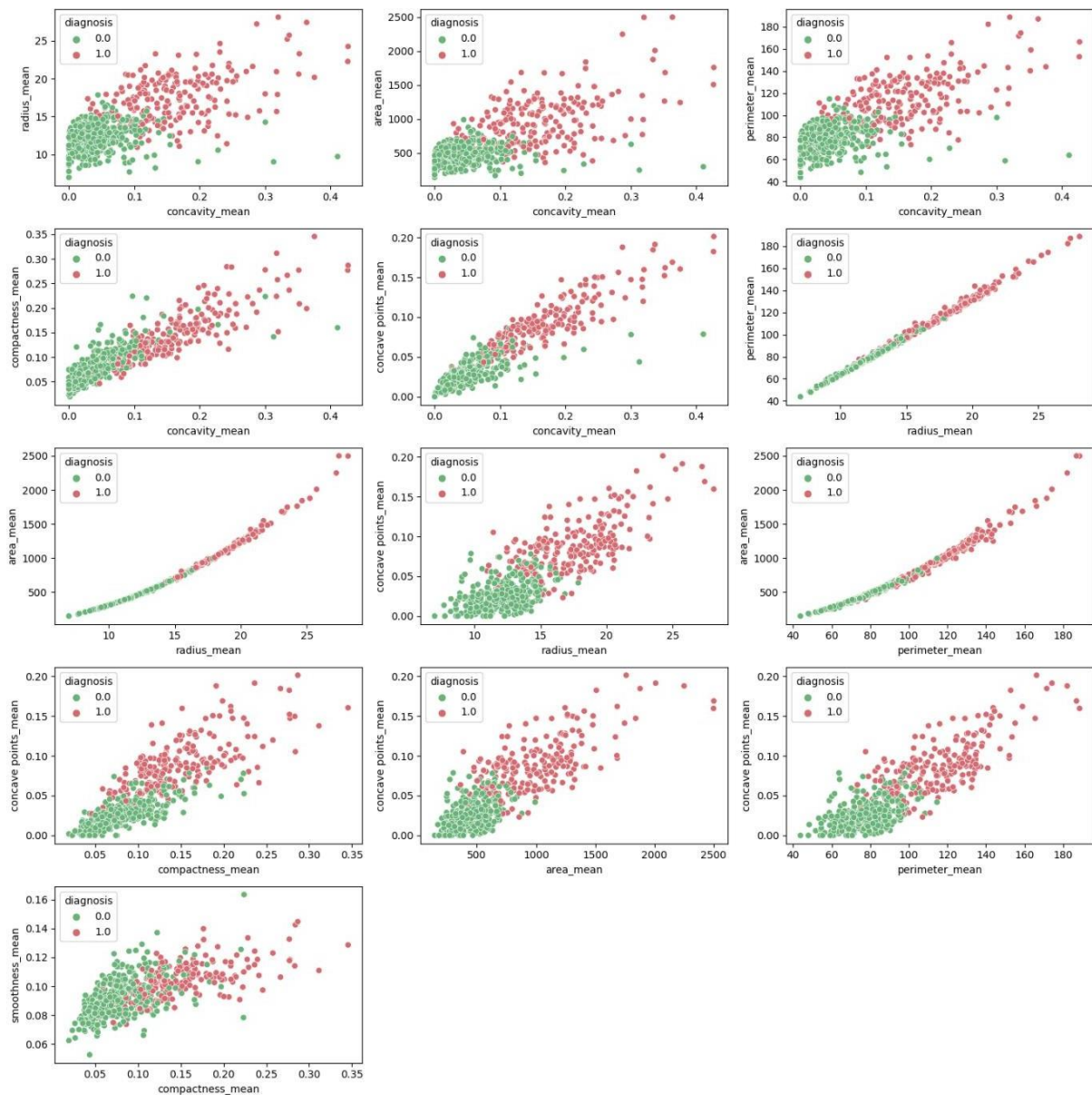


Fig 25

## BOX PLOT

In machine learning, a box plot—also called a whisker plot or box-and-whisker diagram—is an effective tool for analysing and understanding data distributions. Let's examine the benefits of box plots:

Visualization of Summary Statistics:

- o A box plot offers a succinct overview of the most important statistics in a dataset.

The following elements are visible:

Minimum: The dataset's smallest value (apart from outliers).

Identify the threshold at which the lowest 25% of the dataset falls found is Identified as the initial quarter, typically denoted as Q1.

Median (Q2): The dataset's median value, below which 50% of the data falls.

The Third Quartile (Q3) Identify the point at which half of the dataset lies below.  
Maximum: The highest value inside the dataset (apart from anomalies).

1. Recognizing Outliers: Data points that substantially vary from the general trend are known as outliers.
  - o Potential outliers outside of the whiskers (upper and lower boundaries) are graphically indicated by box plots.

## 2 Comparison Between Datasets:

- a. Box plots allow us to comparison dispersals across different datasets.
- b. By plotting multiple box plots side by side, we can analyse variations in medians, quartiles, and ranges.

## 3 Skewness and Symmetry:

- c. Skewed distributions (asymmetric) have longer tails on one side.
- d. Box plots reveal skewness:
  - i. If the median is not centred within the box, the delivery is skewed.
  - ii. Right-skewed: Median closer to Q1.
  - iii. Left-skewed: Median closer to Q3.

## 4 Feature Selection and Engineering:

- e. Box plots guide feature selection:
  - i. Features with little difference (small IQR) may not donate much to the model.
  - ii. Features with large difference (large IQR) are educational.

In summary, box plots provide a compact and informative representation of data distributions, facilitate outlier detection, and assist in making informed decisions during the machine learning process.

## CODE FOR BOXPLOT

```
plt.figure(figsize=[15,20])

i = 0

for col in data.columns[1:]:
    plt.subplot(6, 5, i+1)
    sns.boxplot(data=data, x=col, hue='diagnosis', palette=diagnosis_colors,)
    plt.legend(loc='upper right', title='diagnosis')
    i = i + 1

plt.tight_layout()
plt.show()
```

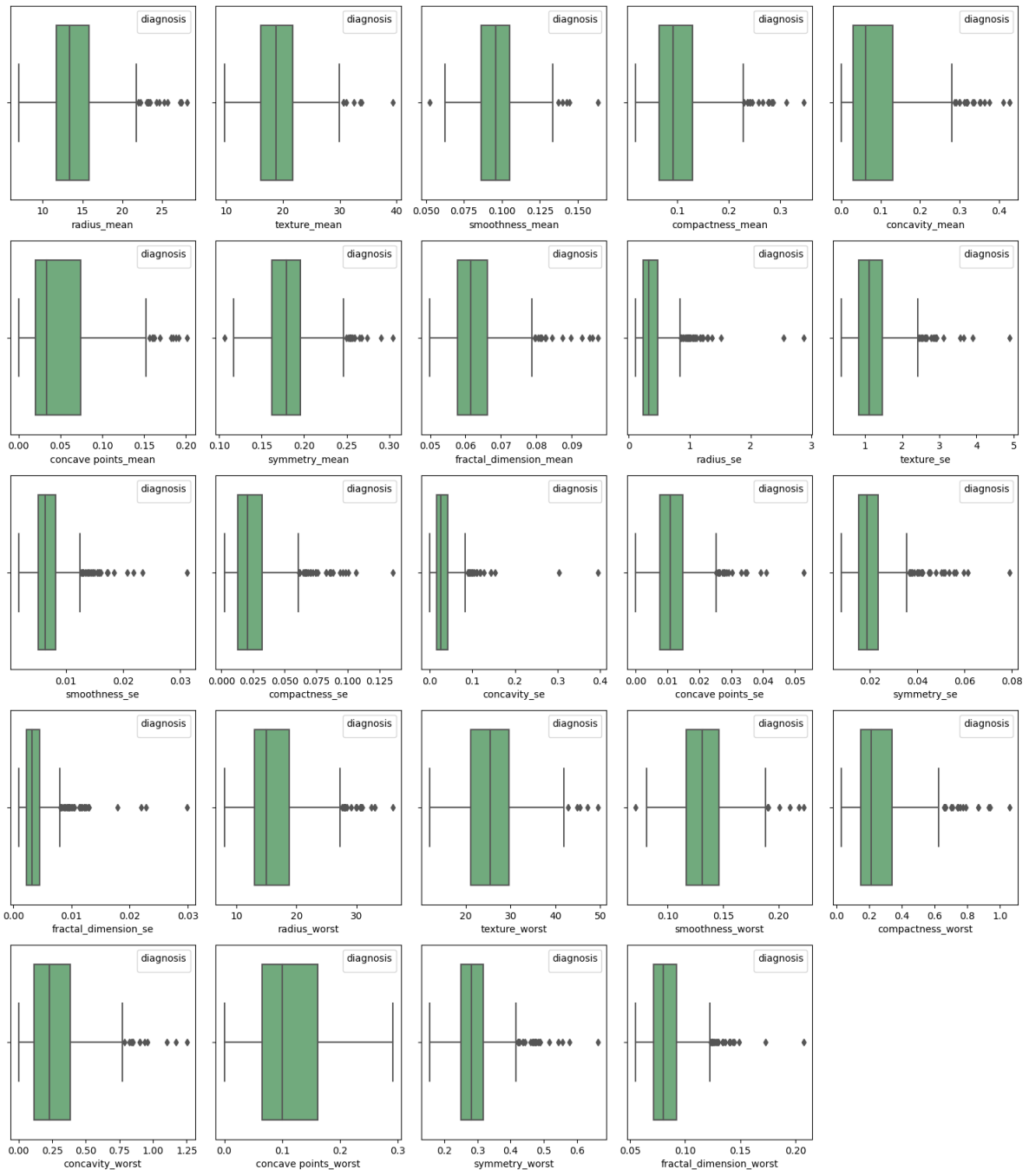


Fig 26



## MODELS

### LOGISTIC REGRESSION

```
1 = data.copy()


[ ] l.drop(['compactness_worst', 'concave points_worst', 'radius_worst', 'concavity_mean'], axis=1, inplace=True)

[ ] x = l.drop(['diagnosis'], axis=1)
    y = l['diagnosis']
    X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.30, random_state=42)
    scaler = StandardScaler()
    X_train = scaler.fit_transform(X_train)
    X_test = scaler.transform(X_test)

    X_train.shape, X_test.shape, y_train.shape, y_test.shape

((398, 20), (171, 20), (398,), (171,))
```


Fig 27



```
log = LogisticRegression(random_state=42)
log.fit(X_train, y_train)

y_train_log = log.predict(X_train)
y_test_log = log.predict(X_test)

print("Train quality:")
report(y_train_log, y_train)
print("\nTest quality:")
report(y_test_log, y_test)
```



```
Train quality:
Accuracy: 0.9874
Precision: 0.9732
Recall: 0.9932
f1_score: 0.9831

Test quality:
Accuracy: 0.9766
Precision: 0.9841
Recall: 0.9538
f1_score: 0.9688
```

Fig 27

▶ auc\_curve(log)

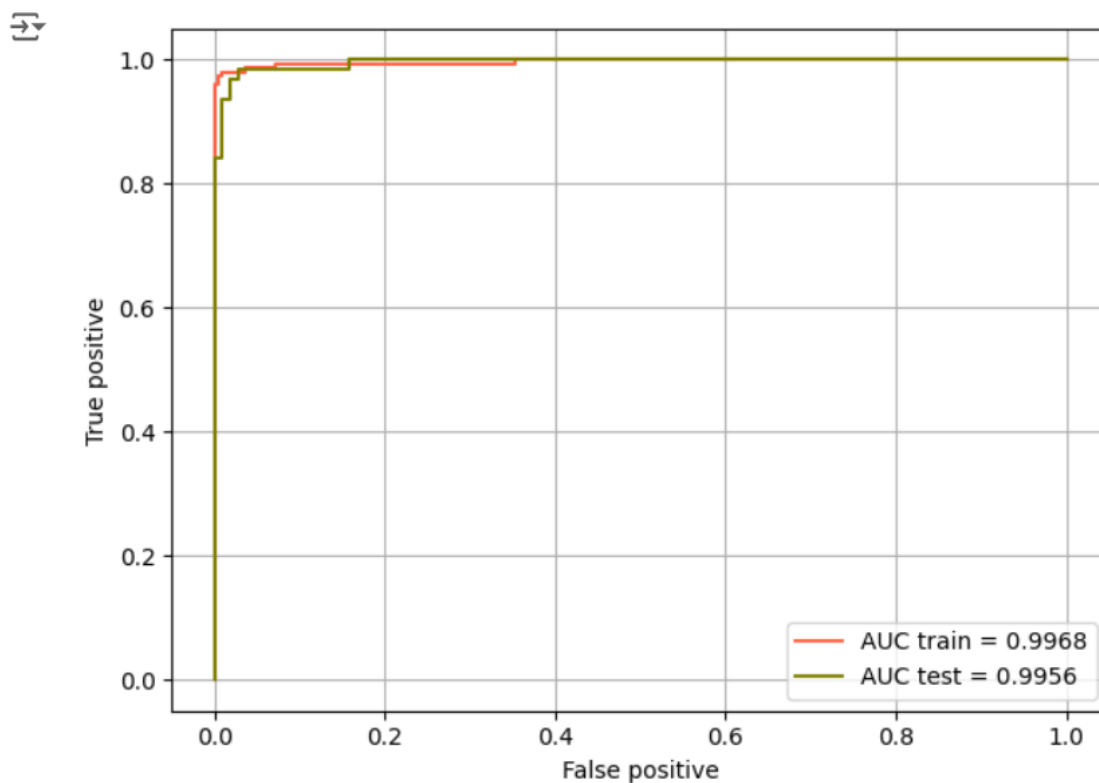


Fig 28

### Feature importance in logistic regression

```
▶ feature = l.drop(['diagnosis'], axis=1).columns
importance = log.coef_[0]
f_i = pd.DataFrame(index=feature,
                    data={'Importance': importance})
f_i.sort_values(['Importance'], inplace=True)

f_i.plot.barh(color=['teal'])

plt.tight_layout()
plt.show()
```

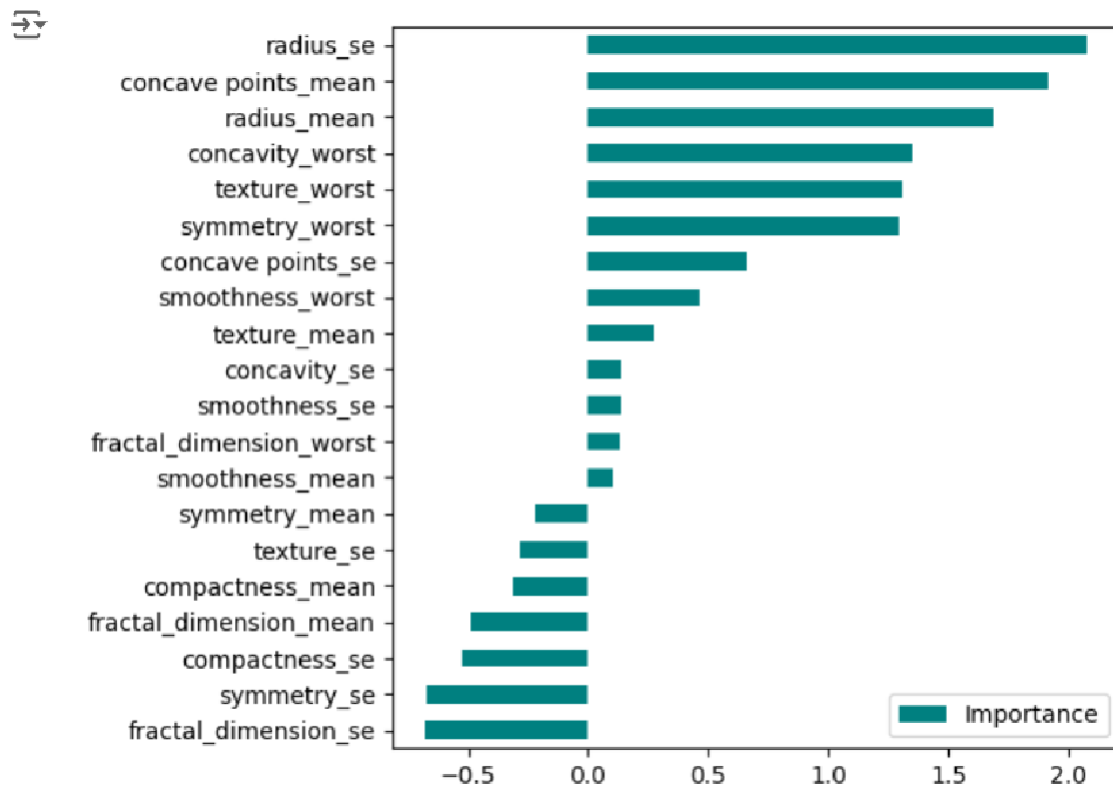


Fig 29

## DECISION TREE

```
[ ] t = data.copy()
```

```
[ ] X = t.drop(['diagnosis'], axis=1)
    y = t['diagnosis']
    X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.30, random_state=42)
    scaler = StandardScaler()
    X_train = scaler.fit_transform(X_train)
    X_test = scaler.transform(X_test)
```

```
X_train.shape, X_test.shape, y_train.shape, y_test.shape
```

```
((398, 24), (171, 24), (398,), (171,))
```

```
[ ] tree = DecisionTreeClassifier(random_state=42)
    tree.fit(X_train, y_train)

    y_train_tree = tree.predict(X_train)
    y_test_tree = tree.predict(X_test)

    print(classification_report(y_true=y_test, y_pred=y_test_tree))
```

Fig 30

	precision	recall	f1-score	support
0.0	0.96	0.92	0.94	108
1.0	0.87	0.94	0.90	63
accuracy			0.92	171
macro avg	0.91	0.93	0.92	171
weighted avg	0.93	0.92	0.92	171

```

y_train_tree = gs_tree.predict(X_train)
y_test_tree = gs_tree.predict(X_test)

print(classification_report(y_true=y_test, y_pred=y_test_tree))

```

	precision	recall	f1-score	support
0.0	0.97	0.95	0.96	108
1.0	0.92	0.95	0.94	63
accuracy			0.95	171
macro avg	0.95	0.95	0.95	171
weighted avg	0.95	0.95	0.95	171

## FEATURE IMPORTANCE

```

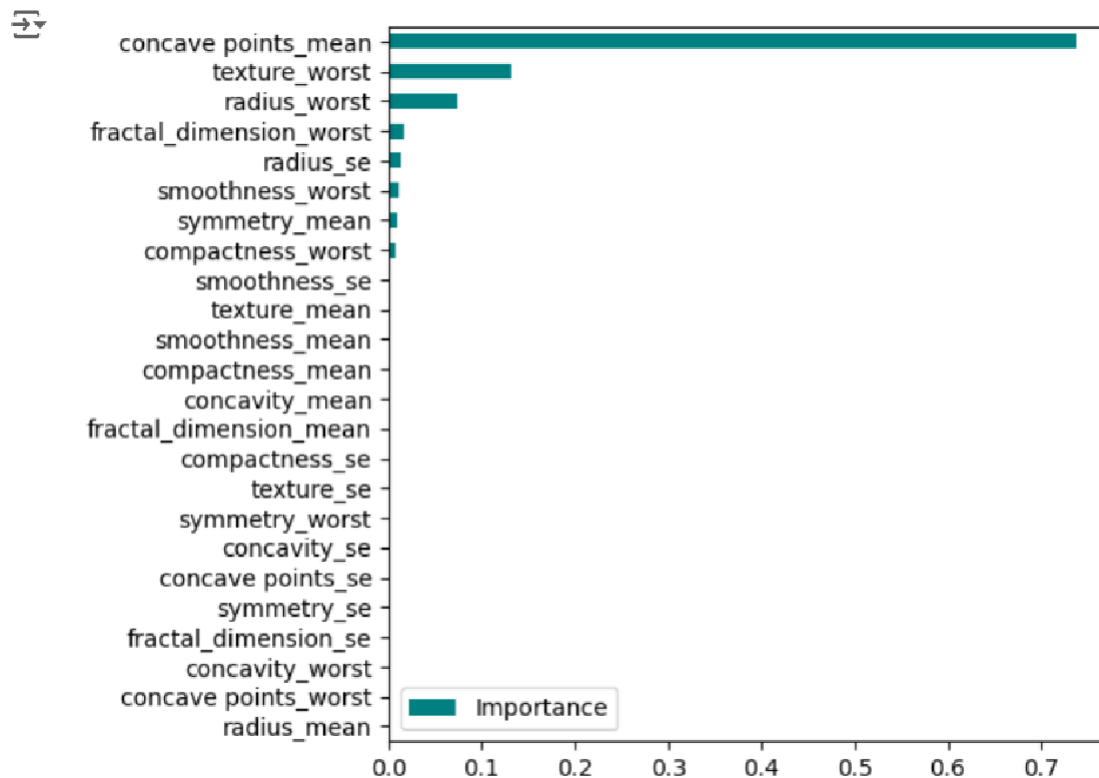
▶ feature = t.drop(['diagnosis'], axis=1).columns
importance = tree.feature_importances_
f_i = pd.DataFrame(index=feature,
                    data={'Importance': importance})
f_i.sort_values(['Importance'], inplace=True)

f_i.plot.barh(color=['teal'])

plt.tight_layout()
plt.show()

```

Fig 30



## KNN MODEL

```
[ ] X = k.drop(['diagnosis'], axis=1)
    y = k['diagnosis']
    X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.30, random_state=42)
    scaler = StandardScaler()
    X_train = scaler.fit_transform(X_train)
    X_test = scaler.transform(X_test)

    X_train.shape, X_test.shape, y_train.shape, y_test.shape
```

⇒ ((398, 24), (171, 24), (398,), (171,))

---

```
[ ] knn = KNeighborsClassifier()
    knn.fit(X_train, y_train)

    y_train_knn = knn.predict(X_train)
    y_test_knn = knn.predict(X_test)

    print("Train quality:")
    report(y_train_knn, y_train)
    print("\nTest quality:")
    report(y_test_knn, y_test)
```

⇒ Train quality:  
Accuracy: 0.9724  
Precision: 0.9262  
Recall: 1.0  
f1\_score: 0.9617

Test quality:  
Accuracy: 0.9649  
Precision: 0.9524  
Recall: 0.9524  
f1\_score: 0.9524

---

```

def auc_curve(model):
    y_pred_proba_train = model.predict_proba(X_train)[: , 1]
    y_pred_proba_test = model.predict_proba(X_test)[: , 1]

    fpr_train, tpr_train, _ = roc_curve(y_train, y_pred_proba_train)
    fpr_test, tpr_test, _ = roc_curve(y_test, y_pred_proba_test)

    auc_train = round(roc_auc_score(y_train, y_pred_proba_train), 4)
    auc_test = round(roc_auc_score(y_test, y_pred_proba_test), 4)

    plt.plot (fpr_train, tpr_train, label=(f'AUC train = {auc_train}'), color='tomato')
    plt.plot (fpr_test, tpr_test, label=(f'AUC test = {auc_test}'), color='olive')
    plt.ylabel('True positive')
    plt.xlabel('False positive')

    plt.legend()
    plt.grid()
    plt.tight_layout()
    plt.show()

```

```
[ ] auc_curve(knn)
```

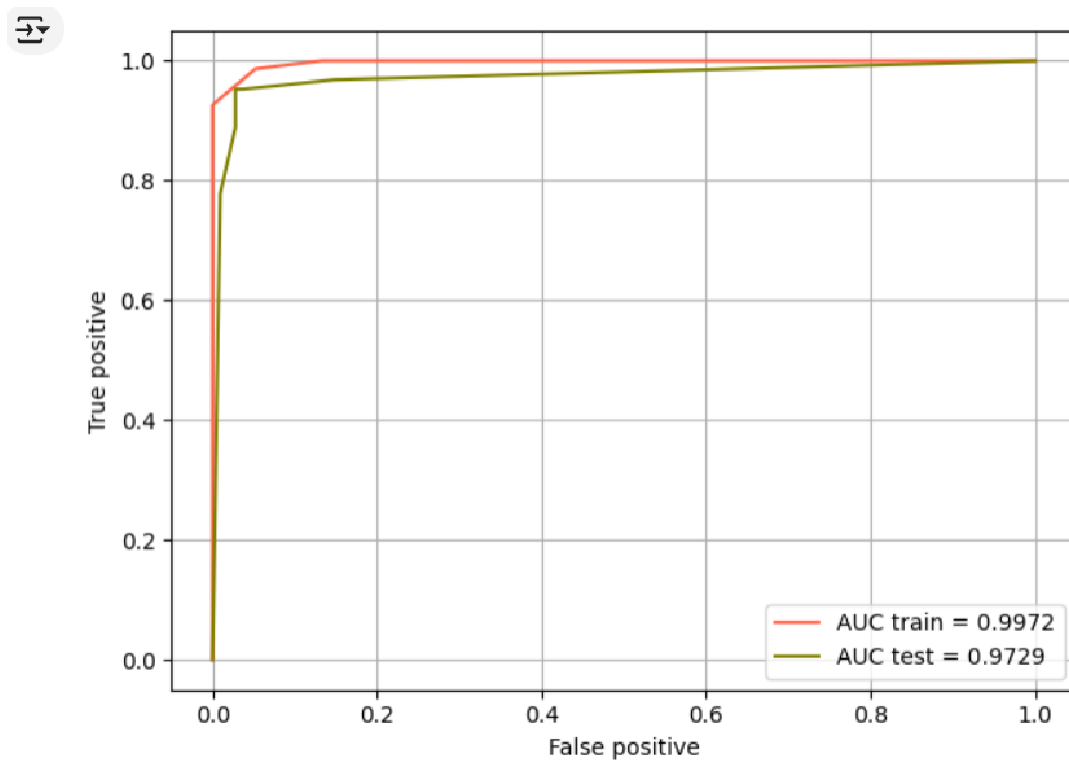


Fig 31



## XG BOOST

```
▶ b = data.copy()
```

```
[ ] X = b.drop(['diagnosis'], axis=1)
    y = b['diagnosis']
    X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.30, random_state=42)
    scaler = StandardScaler()
    X_train = scaler.fit_transform(X_train)
    X_test = scaler.transform(X_test)

    X_train.shape, X_test.shape, y_train.shape, y_test.shape
```

```
↔ ((398, 24), (171, 24), (398,), (171,))
```

```
▶ xgb = xgb.XGBClassifier(random_state=42)
  xgb.fit(X_train, y_train)

  y_train_xgb = xgb.predict(X_train)
  y_test_xgb = xgb.predict(X_test)

  print(classification_report(y_true=y_test, y_pred=y_test_xgb))
```

```
↔
```

	precision	recall	f1-score	support
0.0	0.97	0.97	0.97	108
1.0	0.95	0.95	0.95	63
accuracy			0.96	171
macro avg	0.96	0.96	0.96	171
weighted avg	0.96	0.96	0.96	171

Fig 32

## CONCLUSION

When developing a machine learning model for breast cancer detection or diagnosis, several factors are critical in defining the best fit model. The importance of each factor can vary depending on the specific goals and constraints of the project, but generally, the following aspects are key:

**Accuracy:** The model's ability to correctly identify both positive (cancerous) and negative (non-cancerous) cases. This includes sensitivity (recall) and specificity.

**Sensitivity (Recall):** The ability of the model to correctly identify patients with breast cancer. High sensitivity reduces the risk of false negatives, which is crucial in medical diagnostics.

**Specificity:** The ability of the model to correctly identify patients without breast cancer. High specificity reduces the risk of false positives, which can reduce unnecessary anxiety and medical procedures for patients.

**Precision:** The proportion of true positive results in the predicted positive cases. High precision means fewer false positives.

### **Comparative Analysis:**

**Accuracy:** Logistic Model has higher train (0.9874) and test (0.9766) accuracy compared to KNN.

**Precision:** Logistic Model also has higher train (0.9732) and test (0.9841) precision.

**Recall:** KNN has perfect recall on the training set (1.0), but on the test set, Model 2 performs slightly better (0.9538 vs. 0.9524).

**F1 Score:** Logistic Model has higher train (0.9831) and test (0.9688) F1 scores.

### **Conclusion:**

Logistic Model is performing better overall, with higher accuracy, precision, and F1 scores on both the training and test sets. This indicates that Logistic Model is more reliable and balanced in its predictions for breast cancer detection.

## **REFERENCES**

1. <https://www.geeksforgeeks.org/machine-learning/>
2. <https://www.javatpoint.com/machine-learning>
3. <https://link.springer.com/article/10.1007/s42979-021-00592-x>
4. [https://www.researchgate.net/publication/347059772\\_Machine\\_LearningA\\_Review](https://www.researchgate.net/publication/347059772_Machine_LearningA_Review)
5. <https://www.sciencegate.app/keyword/314>
6. <https://www.sciencedirect.com/science/article/pii/S2666764921000485>
7. <https://www.geeksforgeeks.org/types-of-machine-learning/>
8. <https://www.javatpoint.com/types-of-machine-learning>
9. <https://www.datacamp.com/blog/what-is-machine-learning>
10. [What is Machine Learning? Definition, Types, Tools & More | DataCamp](#)
11. [https://drive.google.com/file/d/1MYl-TD-rmSz9ArFekRO0SD\\_dnqkWRXEa/view?usp=sharing](https://drive.google.com/file/d/1MYl-TD-rmSz9ArFekRO0SD_dnqkWRXEa/view?usp=sharing)
12. <https://drive.google.com/file/d/1mlidbT7QCyX8oYd38-8SwrTJjsIKmjC6/view?usp=sharing>
13. <https://drive.google.com/file/d/1cG-2ANQhBwqDJdNS13xL0dVbtDVD4xfY/view?usp=sharing>
14. <https://paperswithcode.com/>
15. [https://www.hindawi.com/journals/nrp/2014/858403/?msclid=ce7f52e16a281dff40d00df523c028bb&msclid=09cf9361ca24143aa19ef536a31f7cf0&utm\\_source=bing&utm\\_medium=cpc&utm\\_campaign=HDW\\_MRKT\\_GBL\\_SUB\\_BNGA\\_PAID\\_DYNAJOUR\\_X\\_Partners\\_Others&utm\\_term=Mental%20Illness&utm\\_content=JOUR\\_X\\_Partner\\_MIJ](https://www.hindawi.com/journals/nrp/2014/858403/?msclid=ce7f52e16a281dff40d00df523c028bb&msclid=09cf9361ca24143aa19ef536a31f7cf0&utm_source=bing&utm_medium=cpc&utm_campaign=HDW_MRKT_GBL_SUB_BNGA_PAID_DYNAJOUR_X_Partners_Others&utm_term=Mental%20Illness&utm_content=JOUR_X_Partner_MIJ)
16. <https://www.hindawi.com/journals/nrp/2014/858403/#literature-review>
17. <https://www.mayoclinic.org/diseases-conditions/breast-cancer/symptoms-causes/syc-20352470>
18. <https://www.medicalnewstoday.com/articles/327488>
19. <https://www.webmd.com/breast-cancer/understanding-breast-cancer-symptoms>
20. <https://www.coursera.org/articles/what-is-machine-learning>
21. <https://www.coursera.org/articles/types-of-machine-learning>
22. <https://builtin.com/machine-learning/types-of-machine-learning>
23. <http://dspace.dtu.ac.in:8080/jspui/bitstream/repository/20177/1/Mansi%20and%20Rashi%20M.Sc.pdf>
24. <https://www.ljmu.ac.uk/-/media/files/ljmu/students/academic-calendar2425.pdf>
25. <https://www.whitecliffe.ac.nz/>
26. <https://nexusfordlearning.com/>



## **DELHI TECHNOLOGICAL UNIVERSITY**

(Formerly Delhi College of Engineering)  
Shahbad Daulatpur, Main Bawana Road, Delhi-42

### **PLAGIARISM VERIFICATION**

Title of the thesis – Prediction of Breast Cancer Machine Learning Models

Total Pages 50

Name of Scholars - Nidhi Bhati (2K22/MSCMAT/26)

Supervisor- Prof. Anjana Gupta

Department -Applied Mathematics

This is to report that the above thesis was scanned for similarity detection. Process and outcome is given below:

Software used: Turnitin

Similarity Index: 3%

Total Word Count: 6725

Date: 3 June, 2024

PAPER NAME

INTRODUCTION 2.pdf

AUTHOR

. .

WORD COUNT

6725 Words

CHARACTER COUNT

36319 Characters

PAGE COUNT

50 Pages

FILE SIZE

2.2MB

SUBMISSION DATE

Jun 3, 2024 10:11 PM GMT+5:30

REPORT DATE

Jun 3, 2024 10:12 PM GMT+5:30

● 3% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.

- 2% Internet database
- 0% Publications database
- Crossref database
- Crossref Posted Content database
- 2% Submitted Works database

● Excluded from Similarity Report

- Bibliographic material
- Quoted material
- Cited material
- Small Matches (Less than 10 words)



**DELHI TECHNOLOGICAL UNIVERSITY**  
(Formerly Delhi College of Engineering)  
Shahbad Daulatpur, Main Bawana Road, Delhi-42

**CERTIFICATE OF FINAL THESIS SUBMISSION**

1. Name: Nidhi Bhati
2. Roll No.: 2K22/MSCMAT/26
3. Thesis title: “Prediction of Breast Cancer using Machine Learning Models”.
4. Degree for which the thesis is submitted: M.Sc. Mathematics
5. Faculty of the University to which the thesis is submitted: Anjana Gupta
6. Thesis Preparation Guide was referred to for preparing the thesis.  
YES ☐ NO ☐
7. Specifications regarding thesis format have been closely followed.  
YES ☐ NO ☐
8. The contents of the thesis have been organized based on the guidelines.  
YES ☐ NO ☐
9. The thesis has been prepared without resorting to plagiarism. YES ☐ NO ☐
10. All sources used have been cited appropriately. YES ☐ NO ☐
11. The thesis has not been submitted elsewhere for a degree. YES ☐ NO ☐
12. All the correction has been incorporated. YES ☐ NO ☐
13. Submitted 2 hard bound copies plus one CD. YES ☐ NO ☐

(Signature of Candidate)

Name(s): Nidhi Bhati

Roll No.: 2K22/MSCMAT/26