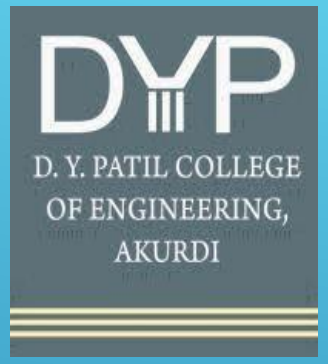


Dr. D. Y. Patil Pratishthan's

D. Y. Patil College of Engineering, Akurdi, Pune-44

Department of Computer Engineering

A. Y. 2021-22
Sem - 2

PROJECT BASED LEARNING-II

Class- S.E. Computer
Division - A

GROUP NUMBER: A1_1

BREAST CANCER DETECTION USING MACHINE LEARNING ALGORITHMS

GROUP MEMBERS :

Aditya Sadakal – SECO2122A001

Aman Shukla – SECO2122A003

Shrishti Sethi – SECO2122A013

Swapnil Bonde – SECO2122A018

GUIDED BY :

Mrs. Dhanashree Phalke

CONTENTS

PROBLEM STATEMENT :-

To detect if the patient has Breast Cancer or not(Malignant or Benign), by making use of the Machine Learning(ML) algorithms.

MOTIVATION

In order to identify the real life problem from societal need point of view and to apply knowledge in ways leading to deep, meaningful learning, Breast Cancer Detection using Machine Learning algorithms has been chosen.

The doctors do not identify each and every breast cancer patient. That's the reason Machine Learning Engineer / Data Scientist comes into the picture because they have knowledge of math and computational power.

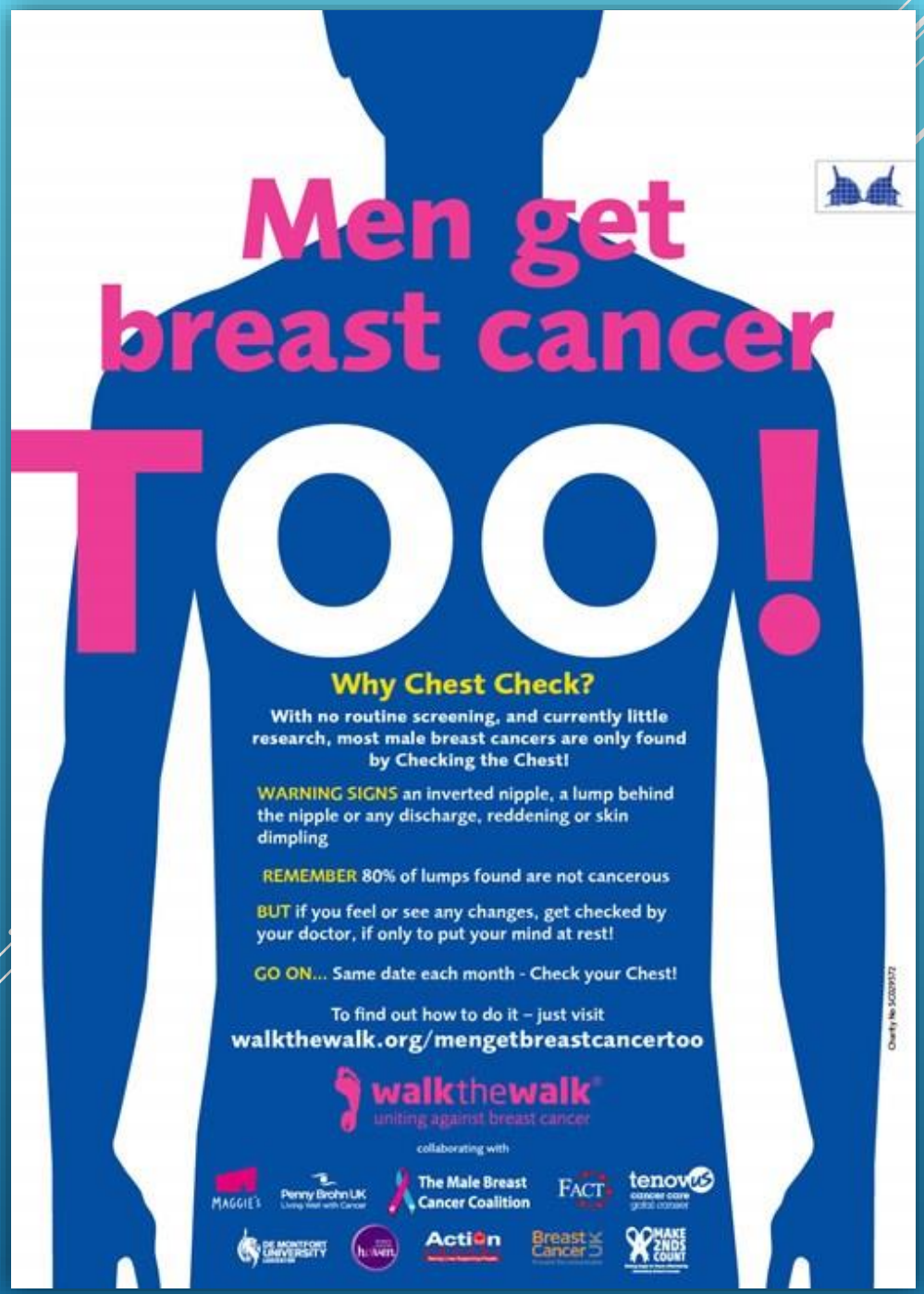
INTRODUCTION

- Breast cancer is a malignant cell growth in the breast. If left untreated, the cancer spreads to other areas of the body
- Breast cancer is a prevailing cause of death, and it is the only type of cancer that is widespread among women worldwide.
- In 2020, there were 2.3 million women diagnosed with breast cancer and 6,85,000 deaths globally.
- Breast cancer occurs in every country of the world in women at any age after puberty but with increasing rates in later life.

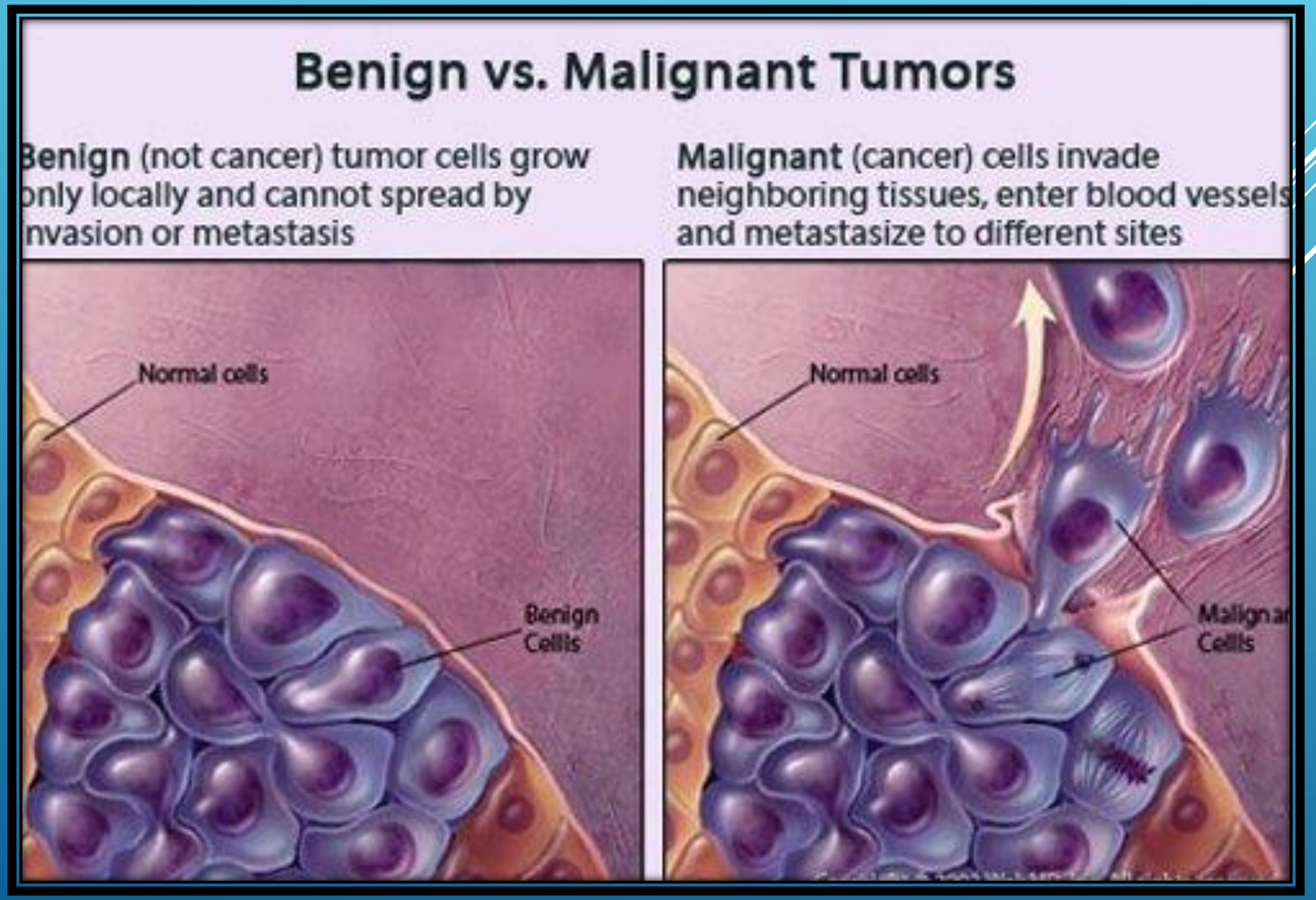
BREAST CANCER IN MEN

Men do not get Breast Cancer, it affects women only – MYTH.

It is estimated that approximately 2,190 men are diagnosed with breast cancer every year and 410 die. (Although this ratio is very less as compared to women).



Cancerous v/s Non-Cancerous Tumours



Definitions

Benign Tumors	**Malignant** Tumors

• Small
• Slow-growing
• Non-invasive
• Well-differentiated
• Stay localized
  • Stay where they are.
  • Can't invade or metastasize.
	• Large
	• Fast-growing
	• Invasive
	• Poorly-differentiated
	• Metastasize
	  • Infiltrate, invade, destroy surrounding tissue.
	  • Then metastasize to other parts of body.

OBJECTIVES of the project:

- To create an ML model to identify whether the breast cancer is benign or malignant.
- To observe which features are most helpful in predicting malignant or benign cancer and to see general trends that may aid us in model selection and hyper parameter selection.
- To predict and diagnose breast cancer, using machine-learning algorithms, and find out the most effective one with respect to confusion matrix, accuracy and precision.
- To study the existing cancer detection models in depth and present the highly accurate and efficient results.

With the potential rise in the number of BC cases in India, the distress reaching is alarming. We hope our efforts will save the life of breast cancer patients.

SCOPE OF PROJECT

- The analysis of the results signify that the dataset when put together with different machine learning techniques provide best suited models for inference in this domain.

- For Future research, we plan to evaluate different algorithms that can help us assist in accurate classification of breast cancer as either benign or malignant. In future more data could be added to the proposed dataset which would help in better training of the machine learning models and would work more accurately. Also, to improve the performance of proposed algorithm , applying other datasets for breast cancer classification can prove useful.

- The findings of the present study can be a good starting point for breast cancer classification. Better results can be achieved by looking into other datasets and various machine learning algorithms. The aim is to reduce the error rates and attain maximum accuracy in future.

Literature Review:

- In recent years, several studies have applied Machine Learning algorithms on different medical datasets to classify Breast Cancer. These algorithms show good classification results, and encourage many researchers to apply these kind of algorithms to solve challenging tasks.
- Moreover, Machine Learning algorithms are used widely in medical fields to predict and classify abnormal events to create a better understanding of any incurable diseases such as cancer. The outcomes of using Machine Learning in classification are promising for breast cancer detection.
- Therefore, Machine Learning approach is used in this work. A list of some literature studies related to this method is presented in the following table:

Paper title	Datasets	Algorithms	Results
ML classification techniques and ensemble learning for predicting the type of breast cancer recurrence, 2019	Breast Cancer	NB, SVM, GRNN and J48	GRNN & J48 accuracy: 91% NB & SVM: 89%
Comparative study on different classification techniques for breast cancer dataset, 2014	Breast Cancer	J48, MLP, rough set	J48: 79.97%, MLP: 75.35%, rough set: 71.36%
Analysis of feature selection with classification: breast cancer datasets, 2011	WBC WDBC Breast Cancer	Decision Tree with and without feature selection	Feature selection enhances the results WBC: 96.99% WDBC: 94.77% Breast Cancer: 71.32%

SYSTEM ARCHITECTURE

In this project the architecture is basically the dataset and the features of the dataset . The algorithms take the features of the dataset as input and give labels as malignant or benign tumor to each of the record in the dataset.

The dataset is first split into training and testing set.

The training set is first given as input to the machine learning algorithms so that the system understands what data gives what type of outcome.

After the system is trained, the testing data is used to test whether the system can correctly predict the class of the data. It checks the percentage accuracy of the model.

SYSTEM ARCHITECTURE FLOW CHART

BREAST CANCER DATASET

For the models to operate efficiently we will use the Breast Cancer Wisconsin(Diagnostic) Data Set. It is a dataset of Breast Cancer patients with Malignant and Benign tumor.

Data Set information :
Features are computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. They describe characteristics of the cell nuclei.
This data set has 569 rows (212 malignant cases and 357 benign cases) with 30 numeric features. The outcomes are either 1 - malignant, or 0 - benign.

The breast cancer dataset was obtained from the University of Wisconsin Hospitals, Madison from Dr. William H. Wolhberg.

It is given by Kaggle from UCI Machine Learning Repository :
https://www.kaggle.com/uciml/breast-cancer-wisconsin-data.

ATTRIBUTES IN THE DATASET:

- Diagnosis: The diagnosis of the breast tissues(1 = Malignant, 0 = Benign)
- Mean_radius: Mean of distances from center to points on the perimeter.
- Mean_texture: Standard Deviation of gray-scale values
- Mean_perimeter: Mean size of the core tumor
- Mean_area: Area of the tumor cell
- Mean_smoothness: Mean of local variation in radius lengths

DATA COLLECTION AND PREPROCESSING

- The **Data-Collection** step involves the collection of the data, reading and analyzing the data and data visualization through suitable Python Matplotlib library.

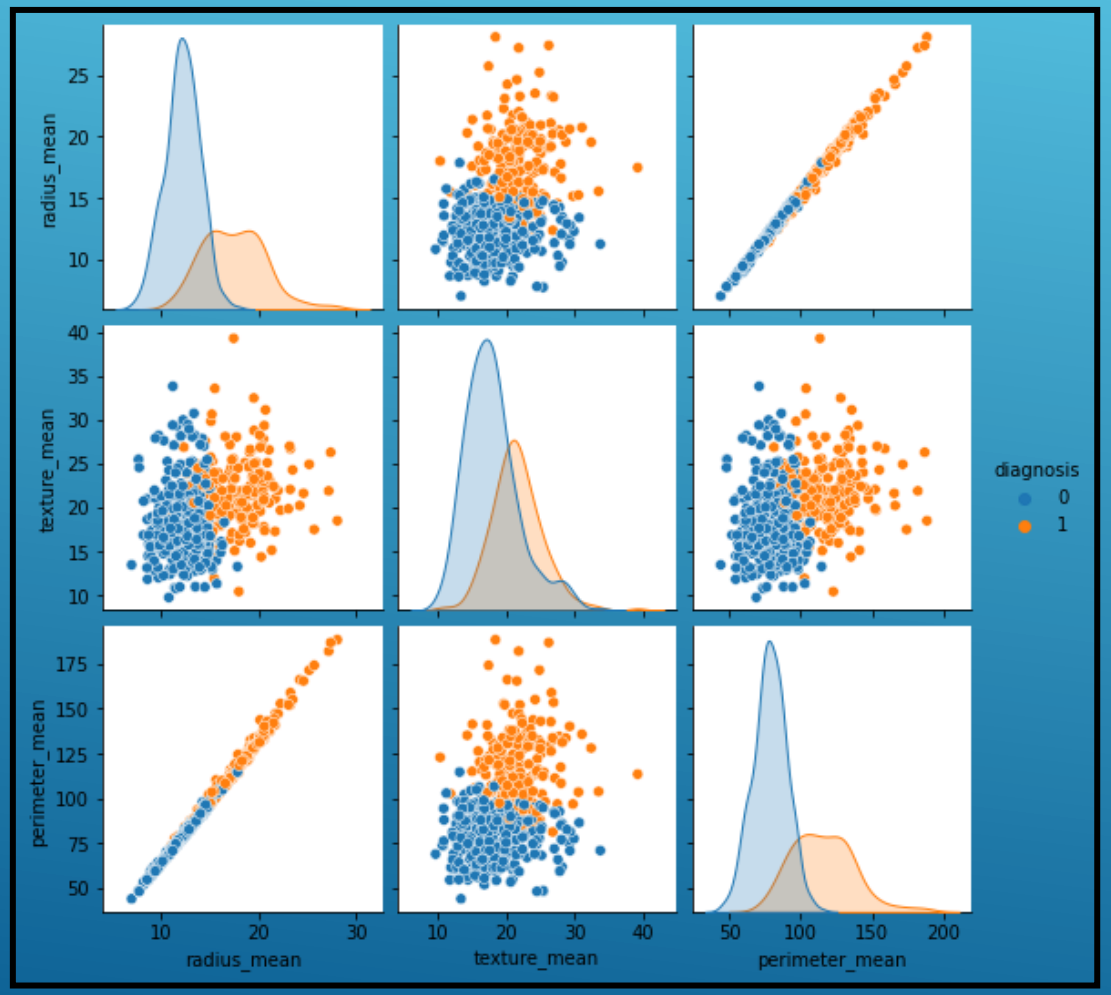
- The proposed dataset is Breast Cancer Wisconsin (diagnostic) dataset made publicly available on Kaggle from the UCI machine learning repository.

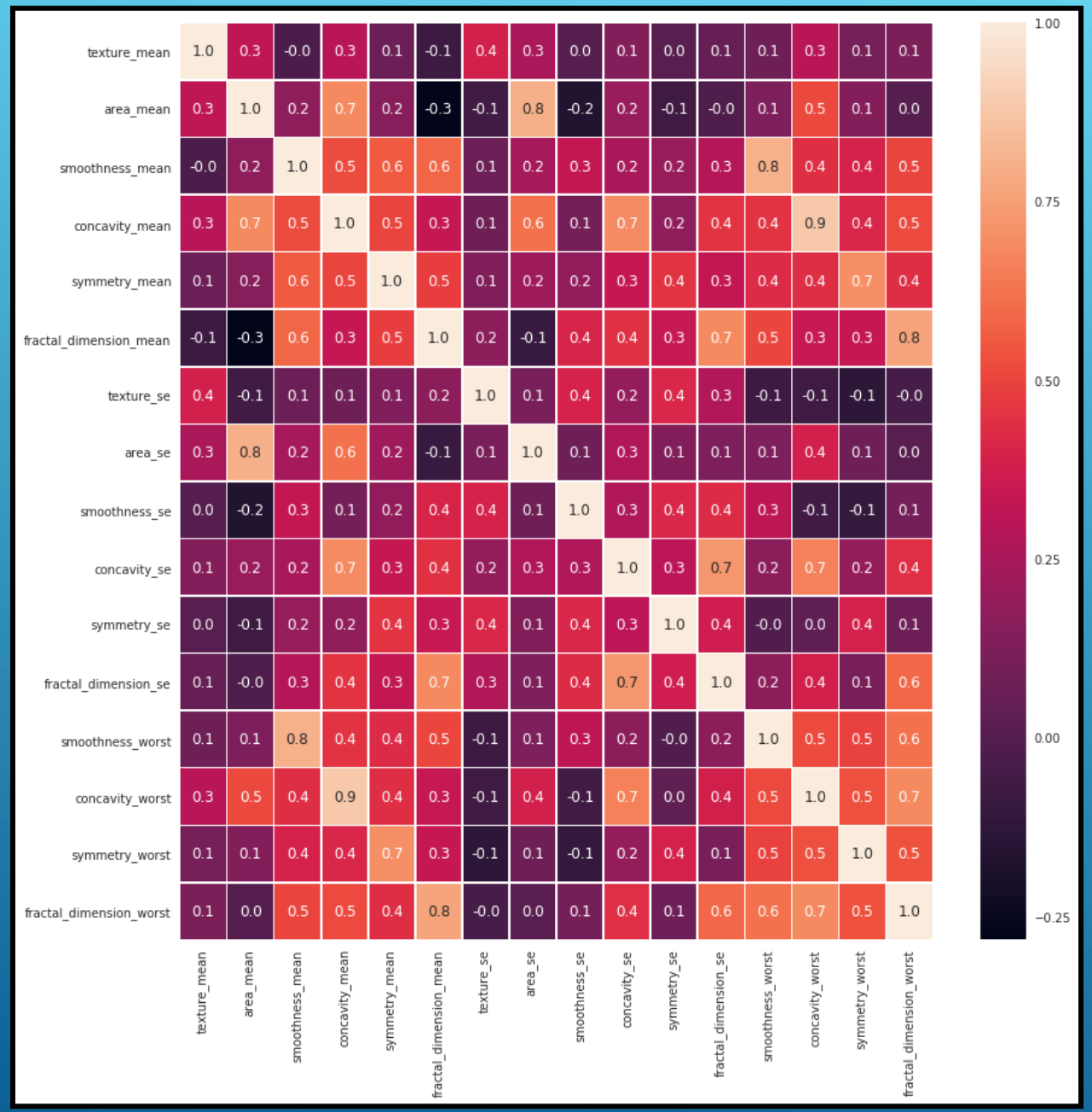
- The presented dataset has **569 rows (212 malignant, or cancerous cases and 357 benign, or non-cancerous cases)** with 30 numeric features. The outcomes are either 1 - malignant, or 0 - benign. Features are computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. They describe characteristics of the cell nuclei.

- **Data preprocessing** involves transforming raw data into understandable form. This step is very important because the quantity and quality of the collected data will directly determine how good the model will be.

Correlation Pair Plot

Pair plot is used to understand the best set of features to explain a relationship between two variables or to form the most separated clusters. It also helps to form some simple classification models by drawing some simple lines or make linear separation in our data set.





❑ Correlation Heatmap

- ❑ Heatmaps visualize the data in a 2 dimensional format in the form of coloured maps. The colour maps use hue, saturation, or luminance to achieve colour variation to display various details. This colour variation gives visual cues to the readers about the magnitude of numeric values. Heatmaps represent data in an easy to understand manner.
- ❑ Heatmaps can describe the density or intensity of variables, visualize patterns, variance, and even anomalies. Heatmaps show relationships between variables. These variables are plotted on both axes. Patterns can be recognised in the cell by noticing the colour change. It only accepts numeric data and plots it on the grid, displaying different data values by varying colour intensity.

DATA PREPARATION

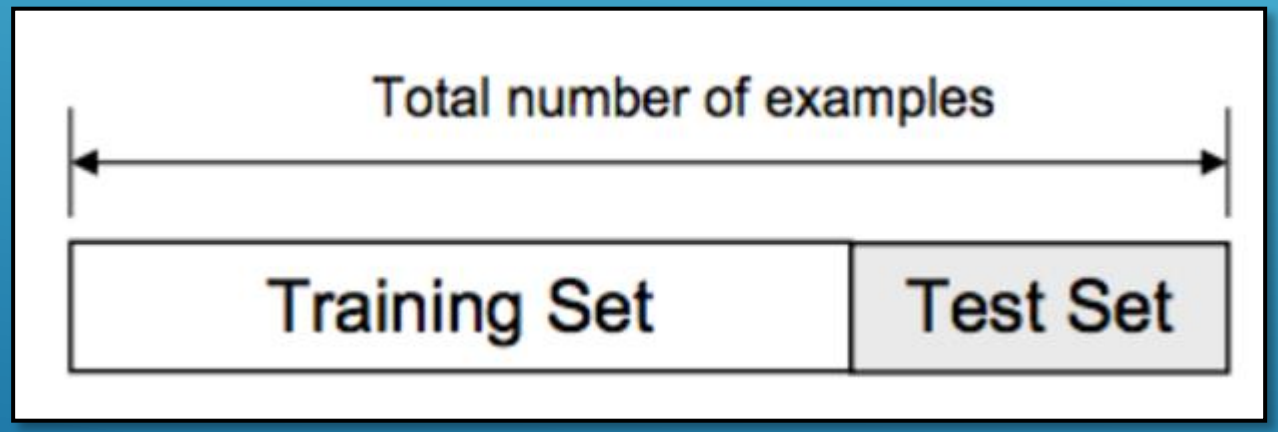
- For achieving better results from the applied model in Machine Learning the format of the data has to be in a proper manner, this is where term Data Preparation is used. Data preparation is a supreme phase of machine learning in which we load our data and prepare it for use in the machine learning training.
  - Example: Random Forest algorithm does not support null values, therefore to execute random forest algorithm null values has to be managed from the original raw data set.

- Another aspect of Data Preparation and analysis is that the data set should be formatted in such a way that more than one Machine Learning algorithms are executed in one data set, and the best out of them is chosen.

- Steps involved in Data Preparation are :
  - Data Cleaning
  - Data Integration
  - Data Transformation
  - Data Reduction

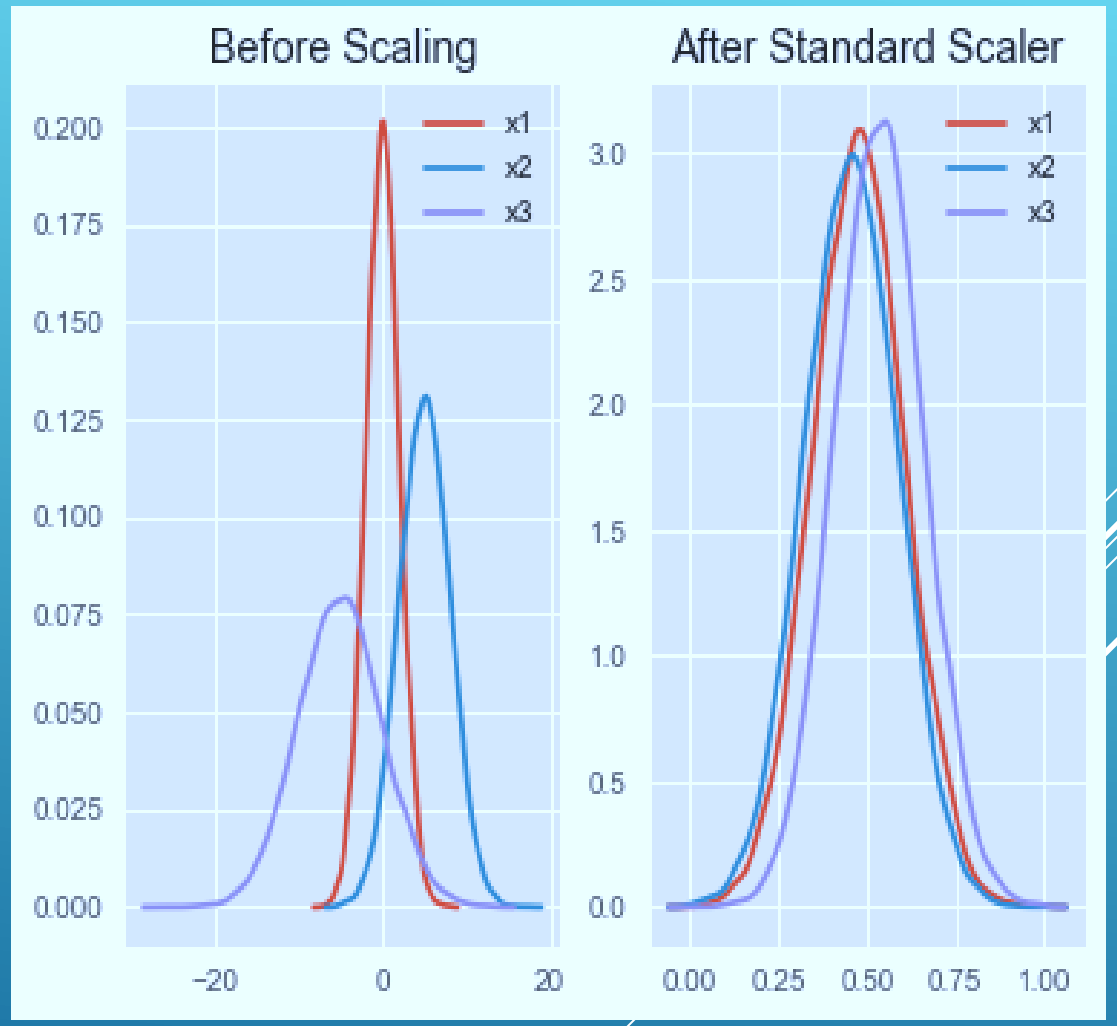
Splitting the dataset

The data we use is usually split into training data and test data. The training set contains a known output and the model learns on this data in order to be generalized to other data later on. We have the test dataset (or subset) in order to test our model's prediction on this subset. We will do this using SciKit-Learn library in Python



FEATURE SCALING

- The dataset contains features varying highly in magnitude, units and range, it is needed to bring all features to the same level of magnitudes. This is done by Features Scaling.
- It is a technique of bringing down the values of all the independent features of our dataset on the same scale. This phase helps to do calculations in algorithms very quickly.
- If we didn't do feature scaling then the machine learning model gives higher weightage to higher values and lower weightage to lower values. Also, takes a lot of time for training the machine learning model.



MODEL SELECTION

- Supervised learning is the method in which the machine is trained on the data which the input and output are well labelled. The model can learn on the training data and can process the future data to predict outcome. They are grouped to Regression and Classification techniques.

- Unsupervised learning is giving away information to the machine that is neither classified nor labelled and allowing the algorithm to analyze the given information without providing any directions. In unsupervised learning algorithm the machine is trained from the data which is not labelled or classified making the algorithm to work without proper instructions.

- In the proposed dataset, the outcome variable or Dependent variable which is Y has only two set of values, either M (Malignant) or B (Benign). Thus, Classification Algorithms of supervised learning are applied. Five different classification algorithms have been chosen for evaluation on the proposed dataset.

RESULTS

- Machine Learning uses data to test and answer questions. This is the phase where data is tested and after analyzing the predictions, conclusions are drawn.

- In this study five different classification algorithms: Random Forest Classifier Algorithm, Decision Tree Classifier Algorithm, Support Vector Machine Classifier Algorithm, K-Nearest Neighbor Classifier Algorithm and Logistic Regression Algorithm are applied on the Breast Cancer Wisconsin (Diagnostic) dataset. The accuracy rate, performance of all the five classification algorithms were then assessed to find out that Support Vector Machine Classifier Algorithm beats all other algorithms providing the highest accuracy rate of 98.24%

ALGORITHMS

Supervised machine learning classifier algorithm will be used in the ML project. The different algorithms that would be used are:

1. Logistic Regression
2. Decision tree Classifier
3. Random forest Classifier
4. k-Nearest Neighbor Classifier
5. Support Vector Machine Classifier

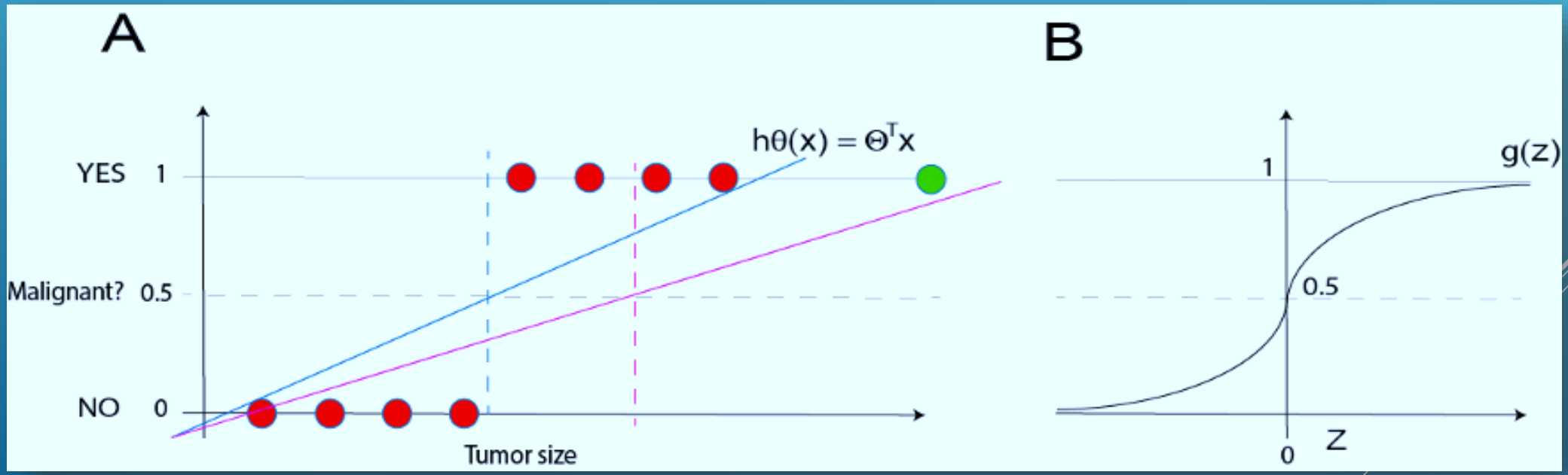
The dataset will be used to train the ML algorithms and then test the same models and the one with the highest accuracy will be selected and thus build the best model.

LOGISTIC REGRESSION CLASSIFIER

- Logistic regression is a supervised learning classification algorithm which is used to predict the probability of a target variable. It is used when the data is linearly separable and the outcome is binary or dichotomous in nature.
- The dependent variable is binary in nature having data coded as either 1 (stands for success/yes) or 0 (stands for failure/no).
- Logistic regression is usually used for Binary classification problems i.e predicting the output variable that is discrete in two classes. A few examples of Binary classification are Yes/No, Pass/Fail, Win/Lose, Cancerous/Non-cancerous, etc.
- Logistic regression is easier to implement, interpret, and very efficient to train. And performs well when the dataset is linearly separable.

WORKING OF LOGISTIC REGRESSION CLASSIFIER

- Linear Hypothesis seems not to add further information to our predictions. That happens because classification is not a linear function(refer fig A).
- Logistic Hypothesis uses a function which is a *sigmoid function that* is non-linear. It calculates the probability that the Diagnosis output can be 0 or 1(refer fig B).



DECISION TREE CLASSIFIER

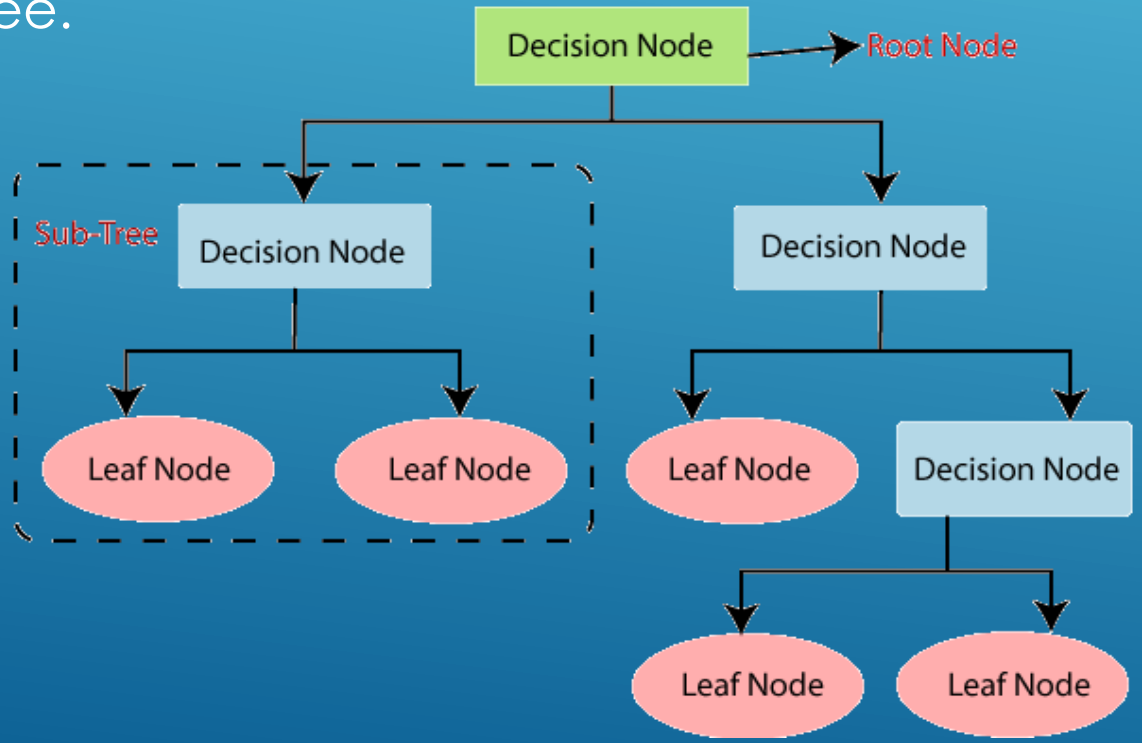
- It is a tree structured classifier, where internal nodes represent the features of a dataset, branches represent the decision rules and each leaf node represents the outcome.
- There are two nodes : Decision Node which is used to make any decision and have multiple branches and leaf nodes are the output of those decisions ,containing no further branches. The decisions or the test are performed on the basis of features of the given dataset.
- It is a graphical representation for getting all the possible solutions to a problem/decision based on given conditions.
- There is less requirement of data cleaning compared to other algorithms.

WORKING OF DECISION TREE CLASSIFIER

Step 1 : Place the best attribute of the dataset at the **root** of the tree.

Step 2 : Split the training set into **subsets**. Subsets should be made in such a way that each subset contains data with the same value for an attribute.

Step 3 : Repeat step 1 and step 2 on each subset until you find **leaf nodes** in all the branches of the tree.

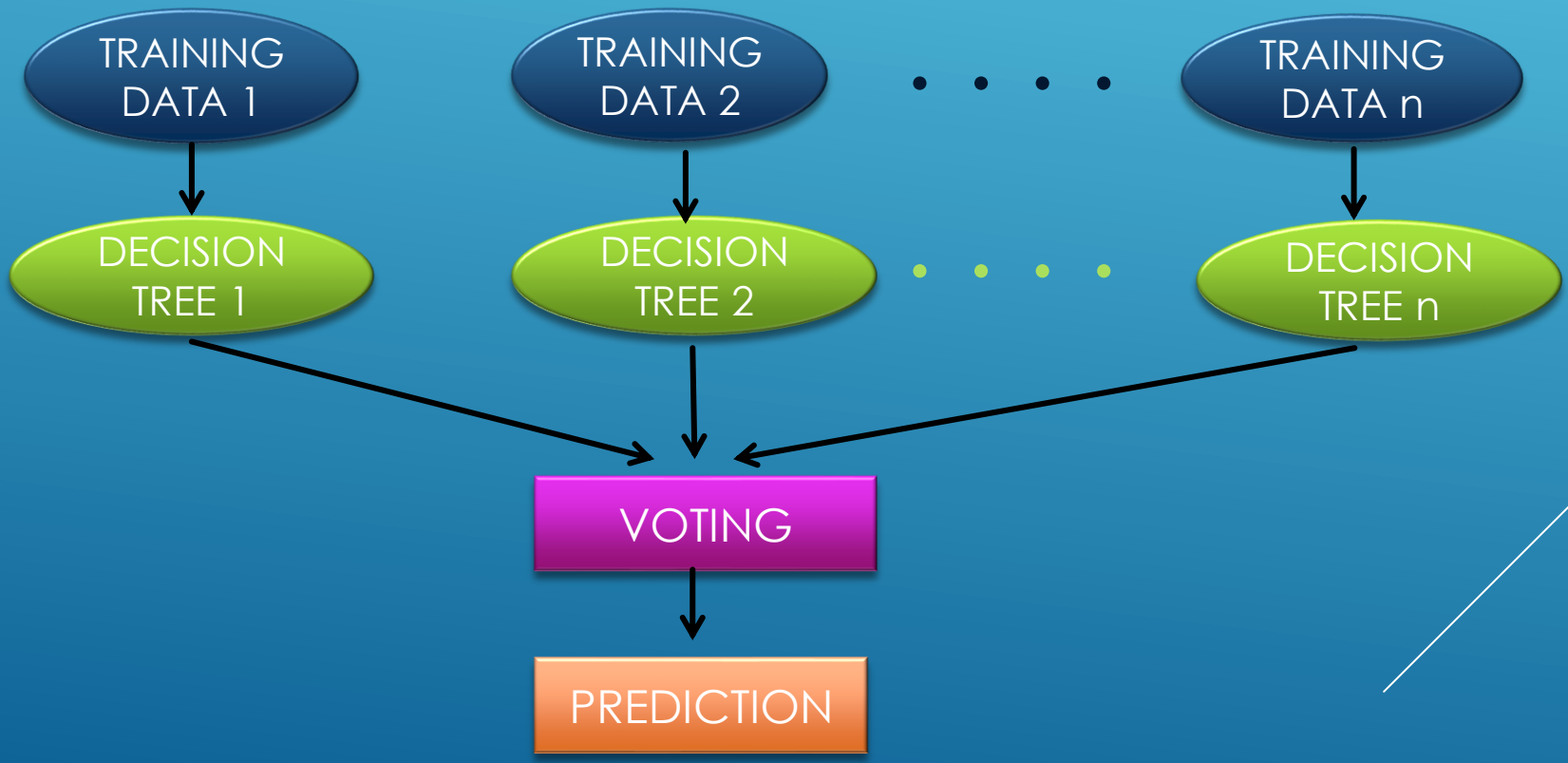


RANDOM FOREST CLASSIFIER

- Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset.
- The greater number of trees in the forest leads to higher accuracy and prevents the problem
  of overfitting.
- It takes less training time as compared to other algorithms.
- It predicts output with high accuracy, even for the large dataset it runs efficiently.
- It can also maintain accuracy when a large proportion of data is missing.

WORKING OF RANDOM FOREST CLASSIFIER

Step-1: Select random K data points from the training set.
Step-2: Build the decision trees associated with the selected data points (Subsets).
Step-3: Choose the number N for decision trees that you want to build.
Step-4: Repeat Step 1 & 2.

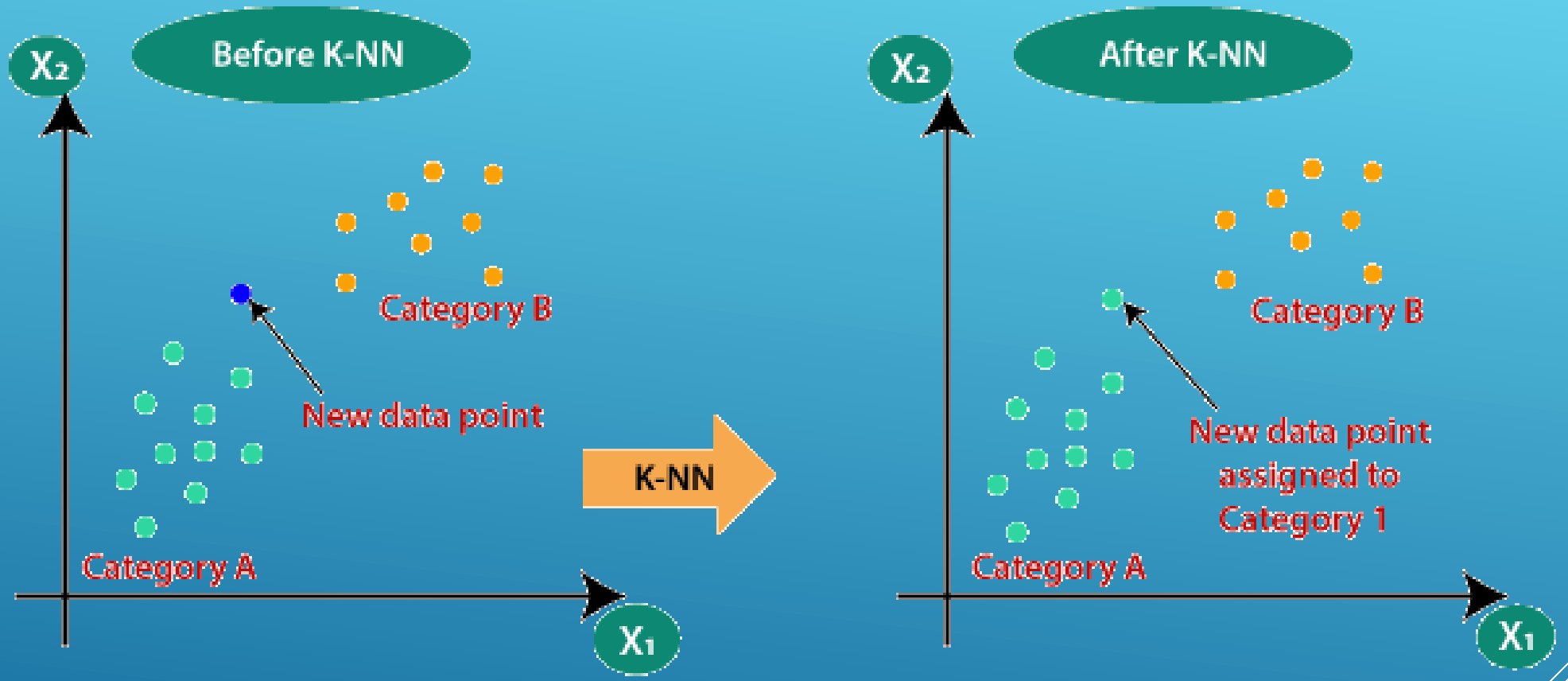


K-Nearest Neighbour Classifier

- It is a supervised machine learning algorithm. The algorithm can be used to solve both classification and regression problem statements.
- **KNN for classification:** KNN can be used for classification in a supervised setting where we are given a dataset with *target labels*. For classification, KNN finds the k nearest data points in the training set and the target label is computed as *the mode* of the target label of these k nearest neighbours.
- **KNN for Regression:**KNN can be used for regression in a supervised setting where we are given a dataset with *continuous target values*. For regression, KNN finds the k nearest data points in the training set and the target value is computed as *the mean* of the target value of these k nearest neighbours.

Working of KNN

1.Load the data
2.Initialize K to your chosen number of neighbors
3. For each example in the data
3.1 Calculate the distance between the query example and the current example from the data.
3.2 Add the distance and the index of the example to an ordered collection
4. Sort the ordered collection of distances and indices from smallest to largest (in ascending order) by the distances
5. Pick the first K entries from the sorted collection
6. Get the labels of the selected K entries
7. If regression, return the mean of the K labels
8. If classification, return the mode of the K labels


Before K-NN
X2
Category B
New data point
Category A
X1
K-NN
After K-NN
X2
Category B
New data point
assigned to
Category 1
Category A
X1

SUPPORT VECTOR MACHINE

What is a Support Vector Machine (SVM)?

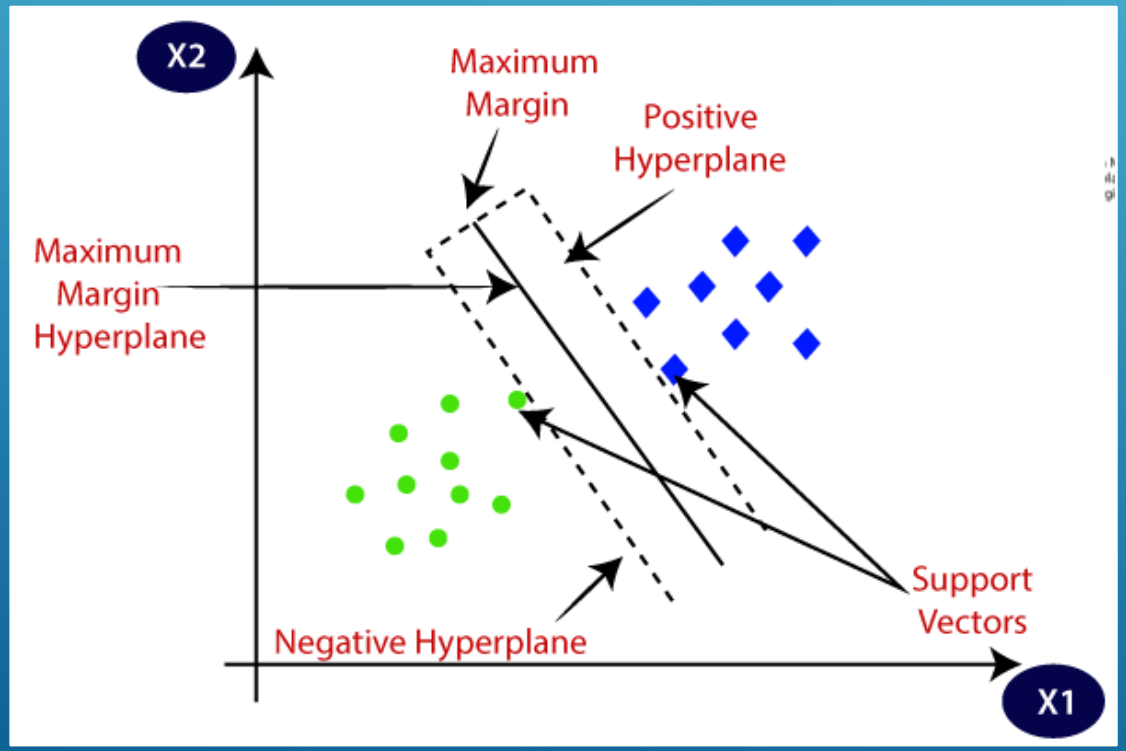
- A Support Vector Machine (SVM) is a binary linear classification whose decision boundary is explicitly constructed to minimize generalization error. It is a very powerful and versatile Machine Learning model, capable of performing linear or nonlinear classification, regression and even outlier detection.

- SVM is well suited for classification of complex but small or medium sized datasets.

How does SVM classify?

It's important to start with the intuition for SVM with the special linearly separable classification case.

If classification of observations is "linearly separable", SVM fits the "decision boundary" that is defined by the largest margin between the closest points for each class. This is commonly called the "maximum margin hyperplane (MMH)".

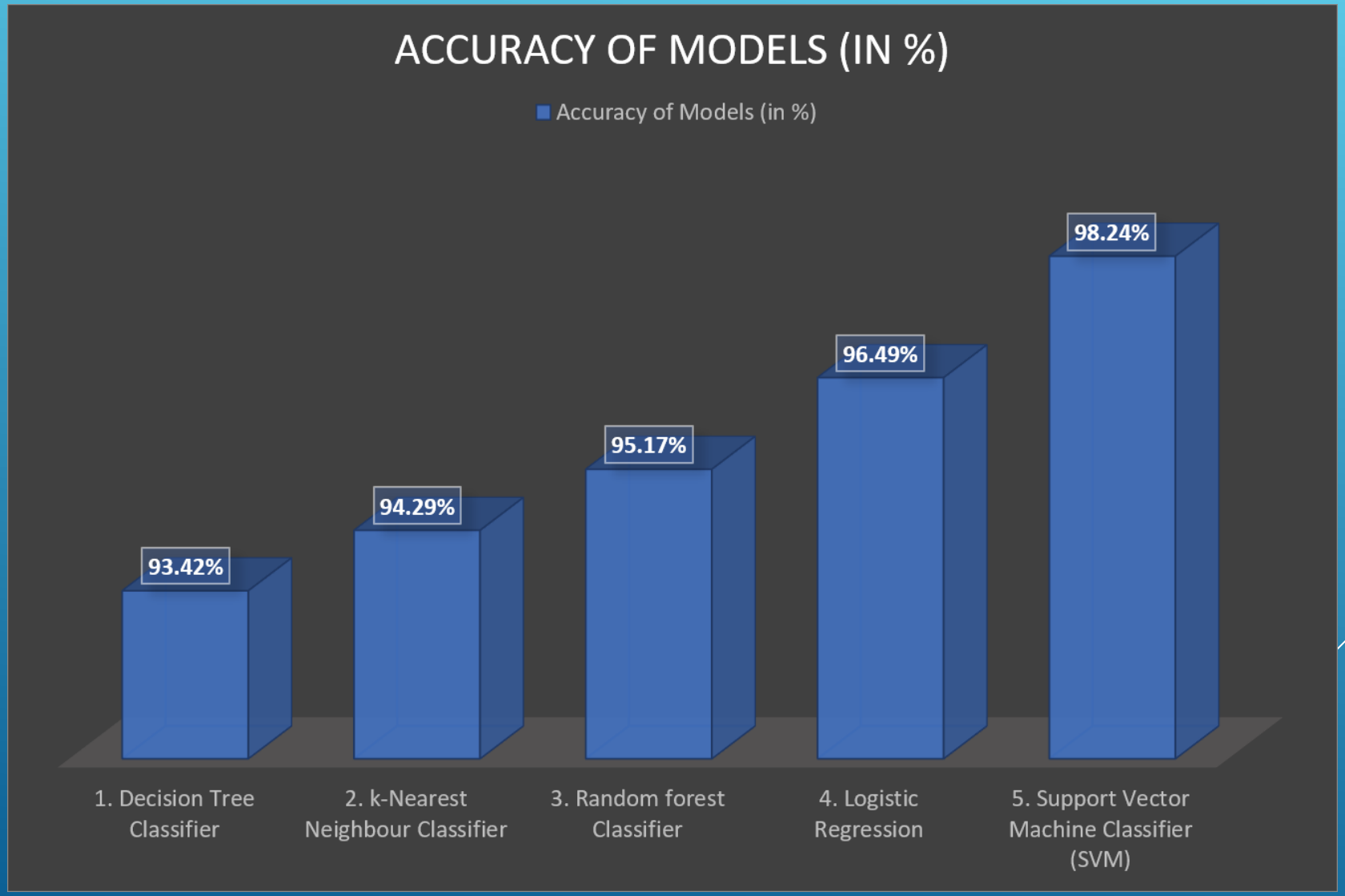


Advantages of SVM:

- Effective in high dimensional spaces.
- Still effective in cases where number of dimensions is greater than the number of samples.
- Uses a subset of training points in the decision function (called support vectors), so it is also memory efficient.
- Versatile: different Kernel functions can be specified for the decision function. Common kernels are provided, but it is also possible to specify custom kernels.

RESULTS:

After applying the different classification models, we have got below accuracies with different models:



Comparative Analysis between the different algorithms used:

ML Algorithm Used	Train - Test dataset Ratio = 70:30 **Accuracy (in %)**	Train - Test dataset Ratio = 60:40 **Accuracy (in %)**	Train - Test dataset Ratio = 80:20 **Accuracy (in %)**
1. Decision Tree Classifier	94.15%	93.42%	93.42%
2. k-NN Classifier	93.34%	92.45%	94.29%
3. Random Forest Classifier	95.90%	95.17%	95.17%
4. Logistic Regression	96.49%	96.49%	96.49%
5. **SVM Classifier**	96.50%	97.36%	**98.24%**

RESULTS AND CONCLUSION

- The aim is to classify the types of tumour and learn the various machine learning techniques such that the most suitable model capable enough not only to classify the tumour but also provide a high accuracy is found.

- The different machine learning techniques like Logistic Regression, Decision Tree, Random Forest and SVM are applied to the Wisconsin Breast Cancer Diagnosis(WBCD) dataset taken from the UCI machine Learning repository. A 80%-20% training-testing split was used for evaluation and then the results were evaluated based on which the accuracy found is as following :-
  - Decision Tree Classifier : 93.42%
  - K-Nearest Neighbour Classifier : 94.29%
  - Random Forest Classifier : 95.17%
  - Logistic Regression Classifier : 96.49%
  - Support Vector Machine Classifier : 98.24%

- Support Vector Classifier was the model that performed the best on the test data with an accuracy score of 98.2% and therefore was chosen to detect the cancer cells in patients.

References:

1. Dataset Reference:
Breast Cancer Wisconsin (Diagnostic) data set [CrossRef]
2. List of datasets for machine-learning research (from Wikipedia) [CrossRef]
3. WHO (World Health Organization) Fact Sheet –Top 10 causes of death [CrossRef]
4. S. Sharma, A. Aggarwal and T. Choudhury, "Breast Cancer Detection Using Machine Learning Algorithms," 2018 International Conference on Computational Techniques, Electronics and Mechanical Systems (CTEMS), 2018, pp. 114-118, doi: 10.1109/CTEMS.2018.8769187.[CrossRef]
5. Mohammed, S.A., Darrab, S., Noaman, S.A., Saake, G. (2020). Analysis of Breast Cancer Detection Using Different Machine Learning Techniques. In: Tan, Y., Shi, Y., Tuba, M. (eds) Data Mining and Big Data. DMBD 2020. Communications in Computer and Information Science, vol 1234. Springer, Singapore. [CrossRef]
6. Mohammed Amine Naji, Sanaa El Filali, Kawtar Aarika, EL Habib Benlahmar, Rachida Ait Abdelouhahid, Olivier Debauche, Machine Learning Algorithms For Breast Cancer Prediction And Diagnosis, Procedia Computer Science, Volume 191, 2021, Pages 487-492, ISSN 1877-0509 [CrossRef]
7. E. A. Bayrak, P. Kırcı and T. Ensari, "Comparison of Machine Learning Methods for Breast Cancer Diagnosis," 2019 Scientific Meeting on Electrical-Electronics & Biomedical Engineering and Computer Science (EBBT), 2019, pp. 1-3, doi: 10.1109/EBBT.2019.8741990. [CrossRef]