Project Based Learning-II

Work Book

Course Code: 210258

(2019 Course)

Second Year Engineering

Year 2021 - 2022, Sem II

Group ID:

A	1	-	1

Team Members:

Name of Student	Roll No.
1. Swapnil Bonde	SECO2122A018
2. Aditya Sadakal	SECO2122A001
3. Aman Shukla	SECO2122A003
4. Shrishti Sethi	SECO2122A013

Project Title: **Breast Cancer Detection Using Machine Learning Algorithms**

Name of Guide : **Mrs. Dhanashree Phalke**

CERTIFICATE

This is to certify that Mr./ Ms. **Swapnil Bonde, Aditya Sadakal, Aman Shukla, Shrishti Sethi** of Group No. **A1_1** Division **A** Branch Computer Engineering have successfully completed the work associated with **Project Based Learning II (210258**) titled as **Breast Cancer Detection Using Machine Learning Algorithms** and have submitted the work book associated under my supervision, in the partial fulfillment of Second Year Bachelor of Engineering (Choice Based Credit System) (2019 course) of Savitribai Phule Pune University.

Date:
Place:

Mrs. Dhanashree Phalke	Dr. Mrs. M. A. Potey
Guide	Head

Table of Contents

1. INTRODUCTION:

Breast Cancer is one of the leading causes of death among women worldwide. Breast cancer occurs in the breast cells, fatty tissues or the fibrous connective tissues within the breast. Cancer starts when the cells start growing abnormally. Breast cancer cells usually form a tumor that can be seen in an X-ray or can be felt as a lump. Breast cancer tumor falls in two categories that are Benign and Malignant. A benign tumor is non-cancerous that just grows abnormally and has a limited growth, however it can increase the chances of getting breast cancer. A malignant tumor on the other hand is more dangerous and life threatening. There is a self-test every woman can do monthly using her hand to check if she can feel a lump because of the abnormal growth of cells or another way is to get mammography test done. Breast cancer tumors tend to gradually worsen and grow faster which eventually cause death. It is often misunderstood that Breast cancer only occurs in women. Although it is common among women but men get diagnosed with Breast cancer too. Breast cancer is a hereditary disease. Age and family history can be the two factors responsible for increasing the risk of breast cancer.

Although breast cancer is life threatening, it can be cured and prevented if detected in the early stages. However, many women are diagnosed with cancer when it is too late. The major challenge after detection of lump in breast is to distinguish between benign and malignant. Timely prediction requires an accurate and authentic methodology and various machine learning techniques have become a popular tool in order to resolve this problem. Machine learning algorithms are applied to develop an intelligent system which can identify breast cancer in the early stage as possible in order to reduce the complications and increase the survival rates of the patients. Today’s advance machine learning algorithms have helped us design several advance models used which makes it easier in detection and diagnoses of breast cancer by training the model with the help of previously observed data of patients. This paper mainly gives a comparison between the performance of five classifier algorithms: Logistic Regression Classifier algorithm, Decision Tree Classifier algorithm, Random Forest Classifier algorithm, K- Nearest Neighbors Classifier algorithm and Support Vector Machine (SVM) Classifier algorithm. This research uses the Breast Cancer Wisconsin (Diagnostic) dataset that is publicly available from the UCI Machine learning Repository. The programming language used for the evaluation and the analysis of the different algorithms and models is Python. The performance of the machine learning algorithms needs to be of highest possible level so that the disease can be identified accurately and with higher precision level.

2. Motivation of Project

Breast Cancer is the most affected, widely spread lethal disease among women worldwide. The WHO [3] records said that in 2020, there were 2.3 million women diagnosed with breast cancer and 6,85,000 deaths were reported globally. By the end of year 2020, there were 7.8 million women alive who were diagnosed with breast cancer in the past 5 years, making it the world's most prevalent cancer. As of January 2022, there were more than 3.8 million women with a history of breast cancer which included the women currently being treated and women who have finished treatment. It is estimated that about 30% of the newly diagnosed cancers in women will be breast cancer. Breast cancer again become the most common cancer globally as of 2021, accounting for over 12% of all new annual cancer cases worldwide, according to World Health Organization (WHO). A study and research on breast cancer detection using the various machine learning algorithms is an attempt in order to identify the real-world problems from societal point of view and applying our knowledge in the right direction leading to deep, meaningful learning.

This paper attempts at proposing one such model with least error rate and best accuracy level in an attempt to save millions of lives by being able to identify breast cancer accurately and at an early stage.

3. Problem Definition and Scope

i. Problem Statement

To design a suitable model with highest accuracy rate possible which detects whether the patient has Breast Cancer or not (Benign or Malignant), by making use of the various Machine Learning Algorithms.

ii. Objectives

The objectives of this project are:

- To develop a machine learning model to classify the breast cancer tumor which gives great accuracy with minimum error.
- To predict and diagnose breast cancer using machine learning algorithms and find out the most effective one based on the performance of each classifier with respect to confusion matrix, accuracy and precision.
- 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 produce real-time data and predictions in a shorter duration of time using the machine learning techniques.
- To study the existing cancer detection models in depth and present the highly accurate and efficient results.

iii. Architecture

The system architecture of this project is an integrated form comprising of the six basic phases of machine learning which are as follows:

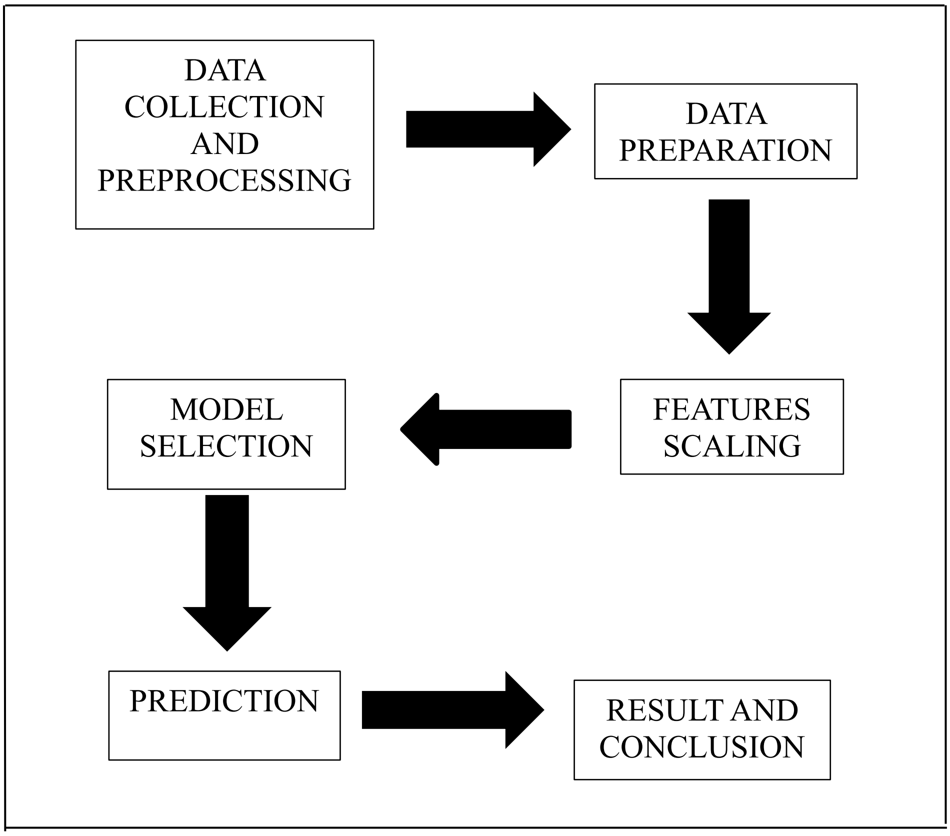


Fig 3.1: Phases of Machine Learning

The above six phases of Machine Learning can be explained as:

1) **DATA COLLECTION AND PREPROCESSING**:

This paper uses the **Breast Cancer Wisconsin (Diagnostic) data set**[1][2] which will be split into a training-testing ratio of 80%-20%.
The **Data-Collection** step involves the collection of the data, reading and analyzing the data and data visualization through suitable Python Matplotlib library, and exploratory analysis to get a thorough understanding of the data, which will be worked upon.

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.

Information about the Data Set:

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, or cancerous cases and 357 benign, or non-cancerous 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.

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.

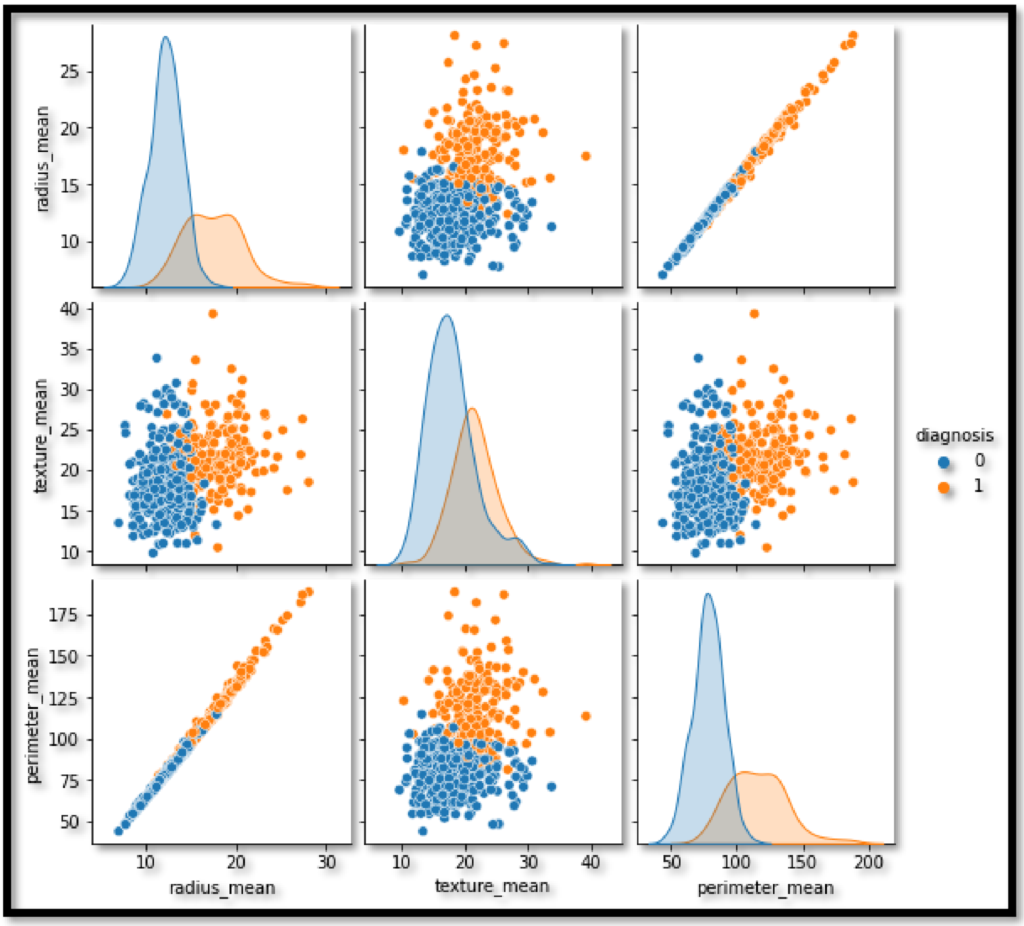
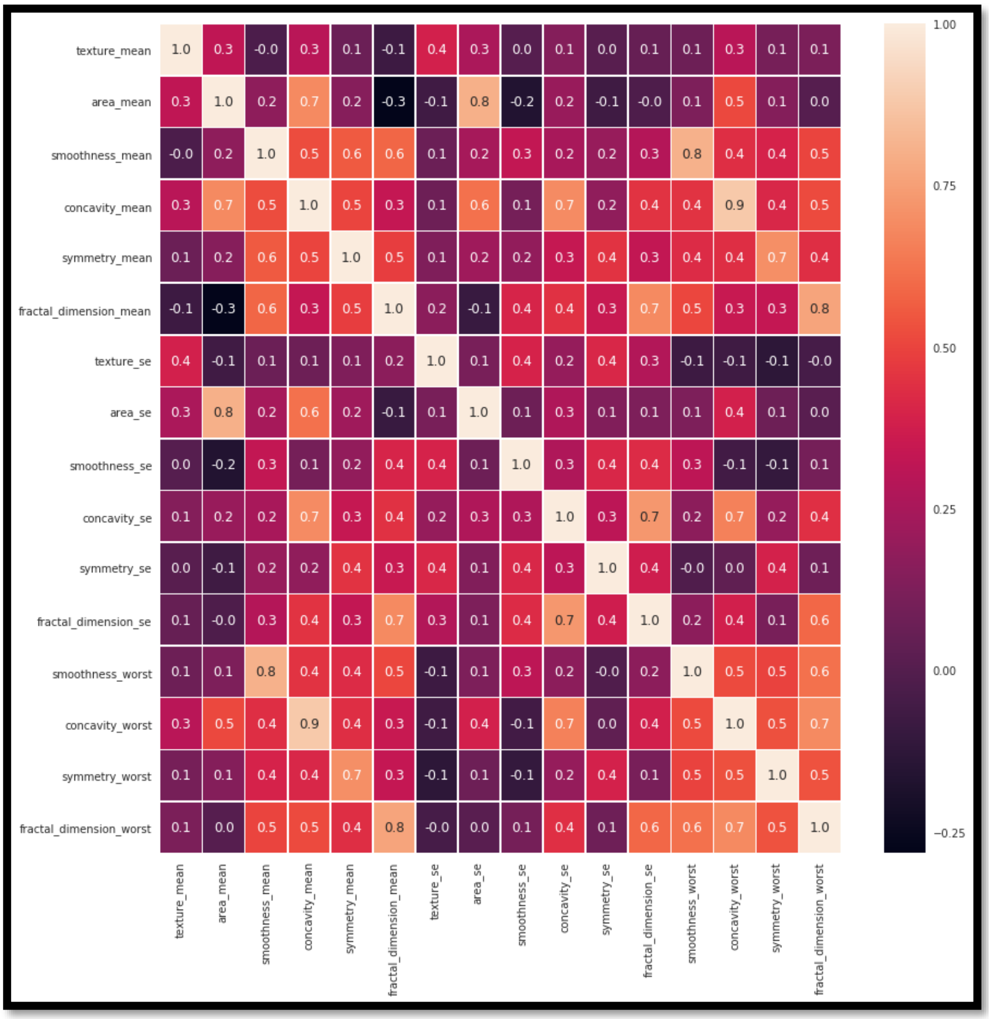


Fig 3.4: Correlation Pair plot

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.



	texture_mean	area_mean	smoothness_mean	concavity_mean	symmetry_mean	fractal_dimension_mean	texture_se	area_se	smoothness_se	concavity_se	symmetry_se	fractal_dimension_se	smoothness_worst	concavity_worst	symmetry_worst	fractal_dimension_worst
texture_mean	1.0	0.3	-0.0	0.3	0.1	-0.1	0.4	0.3	0.0	0.1	0.0	0.1	0.1	0.3	0.1	0.1
area_mean	0.3	1.0	0.2	0.7	0.2	-0.3	-0.1	0.8	-0.2	0.2	-0.1	-0.0	0.1	0.5	0.1	0.0
smoothness_mean	-0.0	0.2	1.0	0.5	0.6	0.6	0.1	0.2	0.3	0.2	0.2	0.3	0.8	0.4	0.4	0.5
concavity_mean	0.3	0.7	0.5	1.0	0.5	0.3	0.1	0.6	0.1	0.7	0.2	0.4	0.4	0.9	0.4	0.5
symmetry_mean	0.1	0.2	0.6	0.5	1.0	0.5	0.1	0.2	0.2	0.3	0.4	0.3	0.4	0.4	0.7	0.4
fractal_dimension_mean	-0.1	-0.3	0.6	0.3	0.5	1.0	0.2	-0.1	0.4	0.4	0.3	0.7	0.5	0.3	0.3	0.8
texture_se	0.4	-0.1	0.1	0.1	0.1	0.2	1.0	0.1	0.4	0.2	0.4	0.3	-0.1	-0.1	-0.1	-0.0
area_se	0.3	0.8	0.2	0.6	0.2	-0.1	0.1	1.0	0.1	0.3	0.1	0.1	0.1	0.4	0.1	0.0
smoothness_se	0.0	-0.2	0.3	0.1	0.2	0.4	0.4	0.1	1.0	0.3	0.4	0.4	0.3	-0.1	-0.1	0.1
concavity_se	0.1	0.2	0.2	0.7	0.3	0.4	0.2	0.3	0.3	1.0	0.3	0.7	0.2	0.7	0.2	0.4
symmetry_se	0.0	-0.1	0.2	0.2	0.4	0.3	0.4	0.1	0.4	0.3	1.0	0.4	-0.0	0.0	0.4	0.1
fractal_dimension_se	0.1	-0.0	0.3	0.4	0.3	0.7	0.3	0.1	0.4	0.7	0.4	1.0	0.2	0.4	0.1	0.6
smoothness_worst	0.1	0.1	0.8	0.4	0.4	0.5	-0.1	0.1	0.3	0.2	-0.0	0.2	1.0	0.5	0.5	0.6
concavity_worst	0.3	0.5	0.4	0.9	0.4	0.3	-0.1	0.4	-0.1	0.7	0.0	0.4	0.5	1.0	0.5	0.7
symmetry_worst	0.1	0.1	0.4	0.4	0.7	0.3	-0.1	0.1	-0.1	0.2	0.4	0.1	0.5	0.5	1.0	0.5
fractal_dimension_worst	0.1	0.0	0.5	0.5	0.4	0.8	-0.0	0.0	0.1	0.4	0.1	0.6	0.6	0.7	0.5	1.0

Fig 3.3: Heatmap for correlation of attributes

2) **DATA PREPARATION**:

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. It includes Data Cleaning, Data Integration, Data transformation and Data Reduction.

Data Cleaning is the first step implemented in Data Preparation. In this step, the primary focus is on handling missing data, noisy data, detection, and removal of duplicate data, and computed biases within the data.
Data Integration is used when data is gathered from various data sources and data are combined to form consistent data which is further put to use in the data preparation and analysis.
Data Transformation is used to convert the raw data into a specified format according to the need of the model. Data Transformation can be done by methods of Normalization, Aggregation or Generalization.
Data Reduction is referred to efficiently organizing the data after the transformation and scaling of duplicate data.

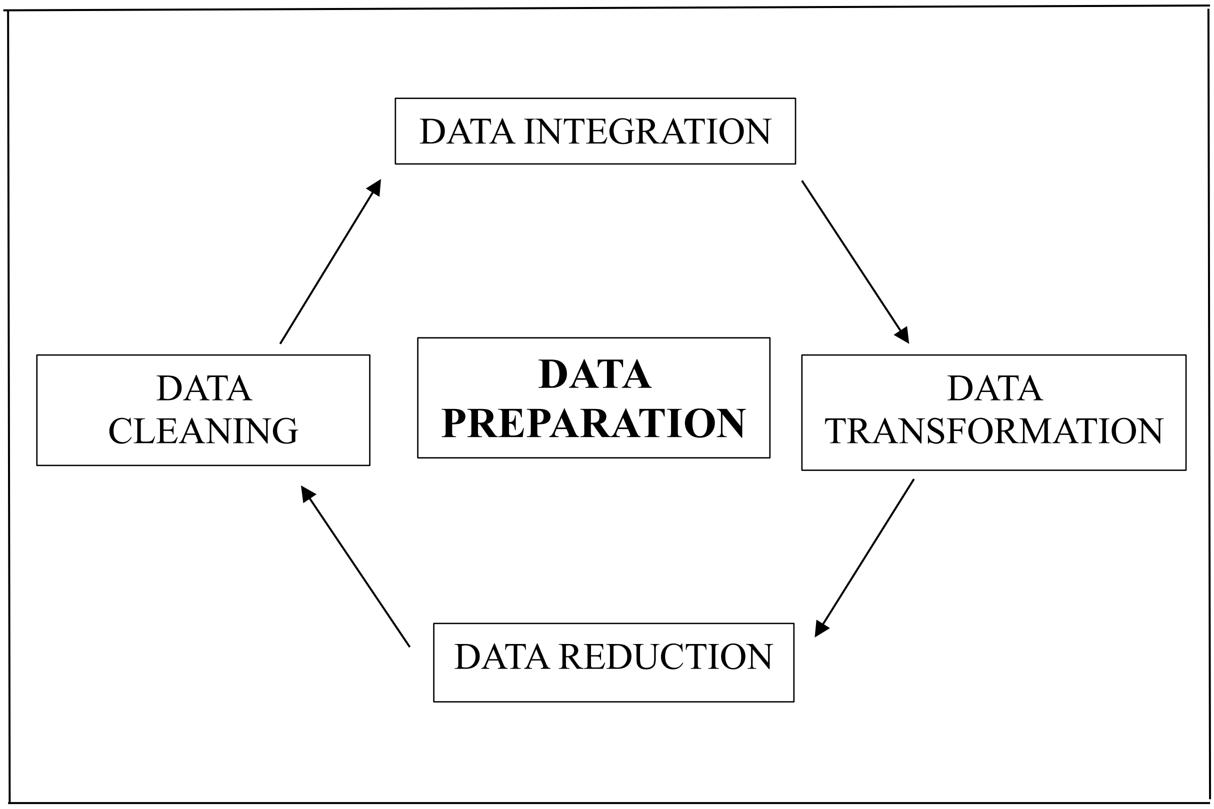


Fig 3.4: Process of Data Preparation.

3) **FEATURES 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.

Feature Scaling is again a very crucial phase of machine learning as 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. Therefore, ignoring this crucial step of scaling the features in machine learning is equal to compromising with the accuracy and efficiency of the designed model. Also, takes a lot of time for training the machine learning model.

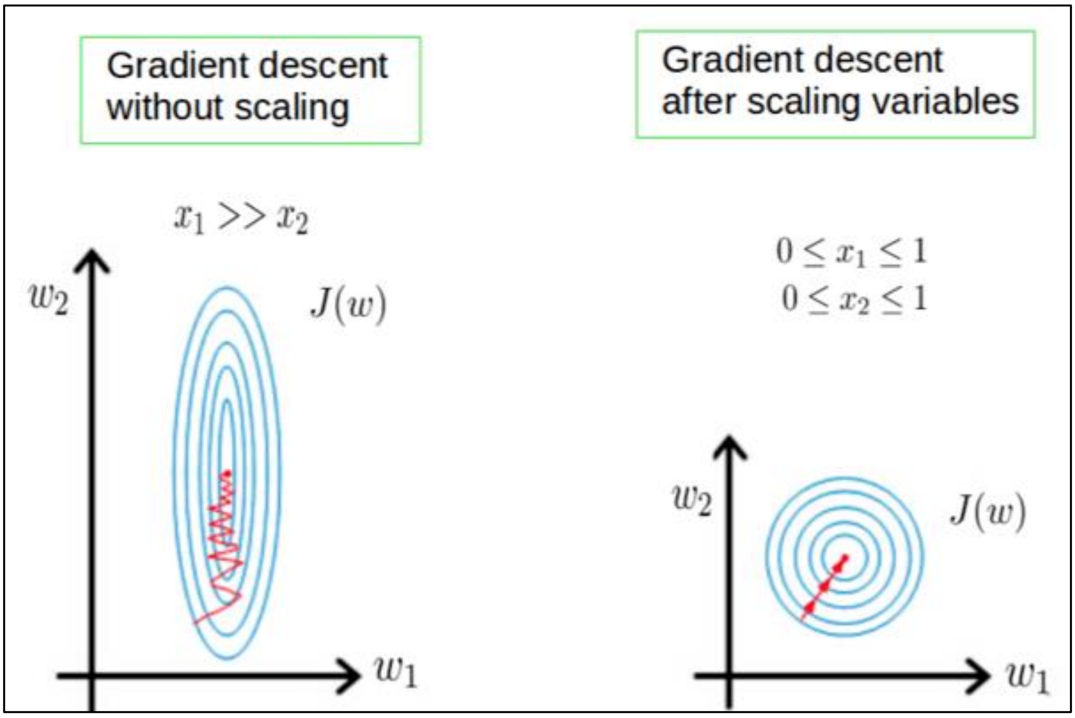


Fig 3.5: Feature Scaling

4) **MODEL SELECTION AND PREDICTION**:

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.

A regression problem is when the result is a real or continuous value, such as “cost” or “weight” where depending on the outcome of the previous data, value for testing data is calculated and outcome is produced.

A classification problem is when the result is a category like filtering emails “spam” or “not spam”, answering to questions with “yes” or “no”, dog breed detection etc. where depending on the trained data, tested data is only classified and not calculated.

Unsupervised Learning: 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.

5) **RESULT AND CONCLUSIONS**:

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 paper 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.

iv. Scope of the project

The research done on this project can prove to be a good starting point for breast cancer classification. The scanning of the outcomes exhibits that when the different machine learning techniques are evaluated on the dataset proposed in this paper then best suited models are provided for inference in this domain.

Apart from the algorithms studied in this research, other algorithms can also be evaluated which might help in assisting a further better prediction of classification of breast cancer. More data could be added to the used dataset based on the experience of many experiments and studies carried out by people which will again help in better training of the machine learning models and would work more accurately.

v. Algorithms [4][5]

The main objective of this study is to identify the effective and predictive algorithm for the detection of breast cancer, therefore we applied several machine learning classifier algorithms : Support Vector Machine (SVM), Random Forest, Decision Tree, Logistic Regression, K-Nearest Neighbor classifier on Breast Cancer Wisconsin (Diagnostic) Dataset and then evaluated the results obtained to define which model provides a higher accuracy.

The various machine leaning algorithms applied are :

I. Support Vector Machine (SVM) :-

It is a supervised learning method derived from statistical learning theory for the classification of both linear and non linear data. It classifies data into two classes over a hyperplane at the same time avoiding over-fitting the data by maximizing the margin of hyperplane separating. In p-dimensions, a hyperplane is described as follows –

$$\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p = 0 \qquad (1)$$

where β_0, β_1, β_2 … β_p are the hypothetical values and X_p are the data points in sample space of p dimension.

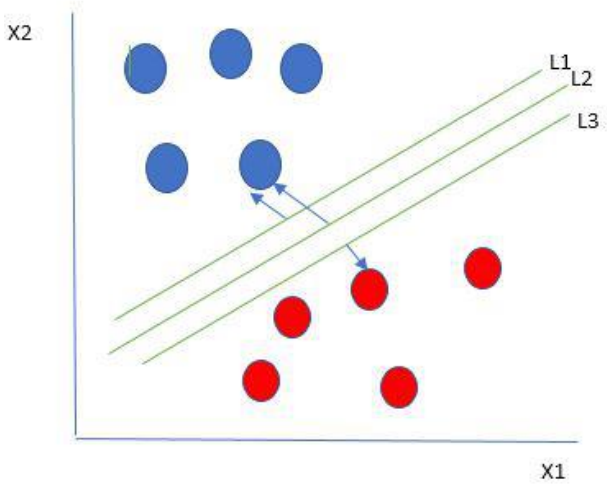


Fig 3.6: Linearly separable data

II. Decision Tree Classifier :-

The decision tree is one of the popular supervised machine learning algorithms used for the graphical representation of all the possible solutions. 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. The decisions are based on some conditions and are easy to interpret. It identifies and chooses the significant attributes that are helpful in classification. It selects only those attributes that return the highest information gain (IG). IG is defined as:

IG = E (Parent Node) – Average E (Child Nodes)

where Entropy (E) is defined as:

$E = \sum_{i} -\text{Prob}_{i} (\log_{2} \text{Prob}_{i})$

where Prob_i is the probability of class *i.*

III. Random Forest Classifier :-

Random forest consists of many individual decision trees that operate as an ensemble. Each individual tree in the random forest spits out a class prediction and the class with the most votes becomes our model's prediction. Random decision forest tries to build k different decision trees by picking a random subset S of training samples. It makes a final prediction based on the mean of each prediction. Random decision trees can interpret and handle irrelevant attributes in a simple manner. They are compact and can handle missing data.

IV. K-Nearest Neighbor Classifier :-

K-NN is a supervised classification algorithm. It takes a bunch of labeled points and uses them to learn how to label other points. To label a new point, it looks at the labeled points closest to that new point, which is its nearest neighbors, and has those neighbor's vote. To get better performance, KNN parameter tuning is done by choosing an appropriate value of k. The similarity between two points is calculated using, for example, the Euclidean distance.

Working of K-NN Classifier:

(1) Input the dataset and split it into a training and testing set.
(2) Pick an instance from the testing sets and calculate its distance with the training set.
(3) List distances in ascending order.
(4) The class of the instance is the most common class of the 3 first training instances(k=3)

V. Logistic Regression Classifier :-

Logistic regression is supervised classification algorithm. Logistic Regression is statistical model used for modelling binary classification problem using logistic function and many more complex extensions exist for logistic regression. Logistic regression is basically a model which uses the regression model to predict the probability that a given data object or entry belong to given category. Logistic Regression is based on the assumption that the linear function is followed by the data. Logistic Regression uses sigmoid function for modelling the data.

Equation of Linear Regression is as follows:

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots. + \beta_n X_n \qquad \ldots..(1)$$

Where y is the dependent variable and x1, x2…xn are explanatory variables.

Sigmoid Function:

$$p = 1/1 + e^{-y} \qquad \ldots..(2)$$

After applying sigmoid function in equation 2 on equation 1:

$$p = 1/1 + e^{-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots. + \beta_n X_n)}$$

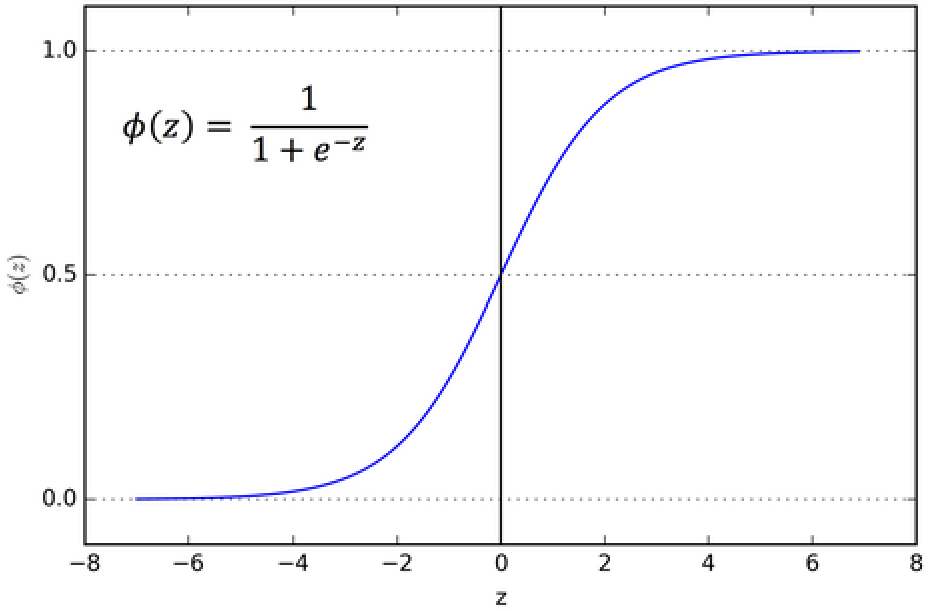


Fig 3.7: Sigmoid Function Graph

4. Hardware and Software Requirements

- Operating System:
    - Windows 7 or 10
    - Mac OS X 10.11 or higher, 64-bit
    - Linux: RHEL 6/7, 64-bit
- X86 64-bit CPU (Intel / AMD architecture)
- 4GB RAM
- 5GB free disk space
- IDE: Google Colab
- Python Libraries:
    - NumPy
    - Matplotlib
    - Pandas
    - Seaborn
    - Scikit Learn

5. Flowcharts

The flowchart of the different machine learning algorithms studied in this paper are as follows :

1) Support Vector Machine (SVM):

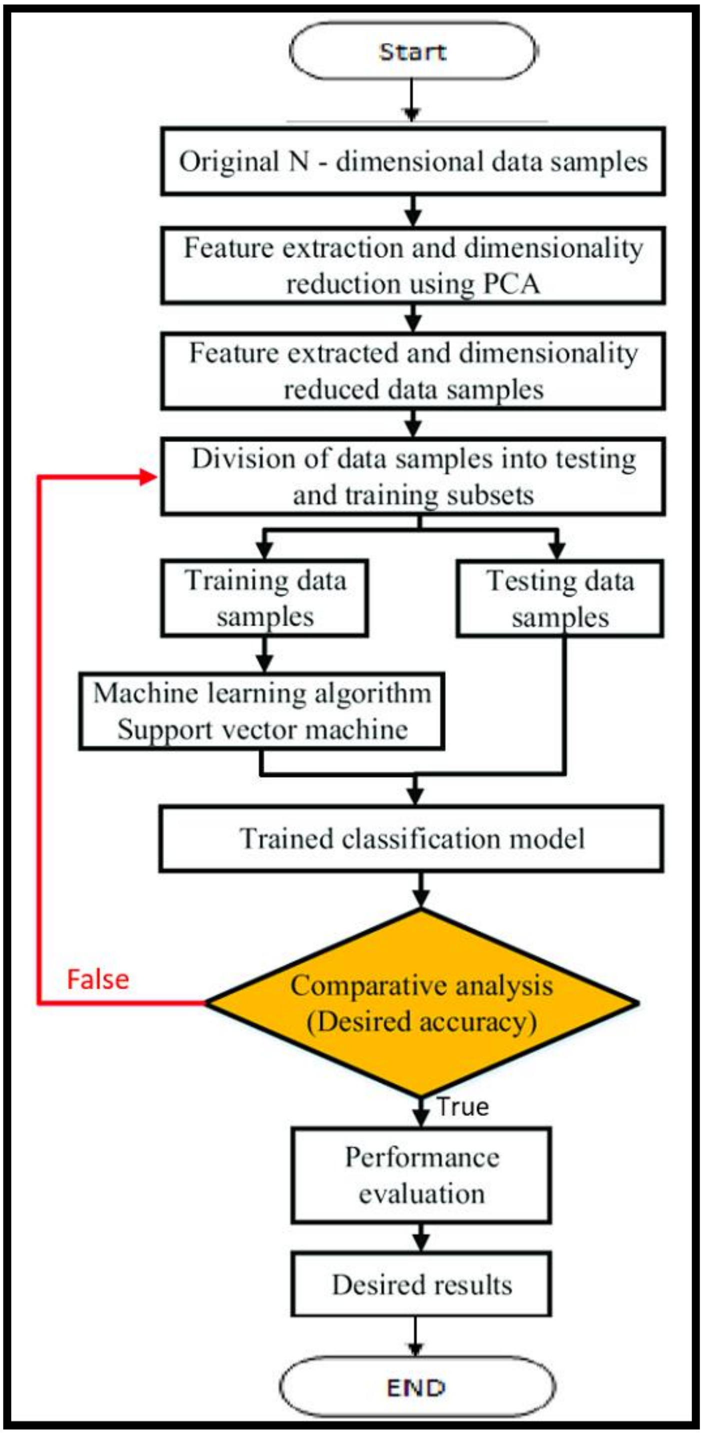

Start
Original N - dimensional data samples
Feature extraction and dimensionality reduction using PCA
Feature extracted and dimensionality reduced data samples
Division of data samples into testing and training subsets
Training data samples
Testing data samples
Machine learning algorithm Support vector machine
Trained classification model
Comparative analysis (Desired accuracy)
False
True
Performance evaluation
Desired results
END

Fig 5.1: SVM Flowchart

2) Decision Tree Classifier:

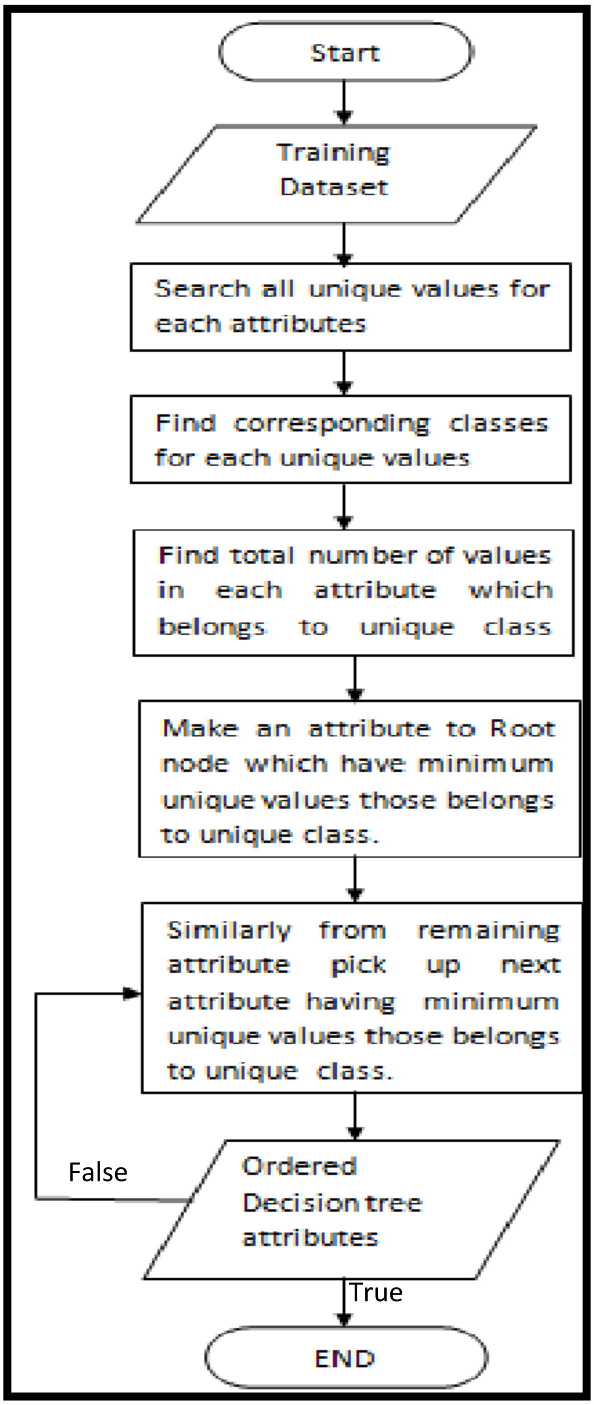

Start
Training Dataset
Search all unique values for each attributes
Find corresponding classes for each unique values
Find total number of values in each attribute which belongs to unique class
Make an attribute to Root node which have minimum unique values those belongs to unique class.
Similarly from remaining attribute pick up next attribute having minimum unique values those belongs to unique class.
False
Ordered Decision tree attributes
True
END

Fig 5.2: Decision Tree Classifier Flowchart

3) Random Forest Classifier:

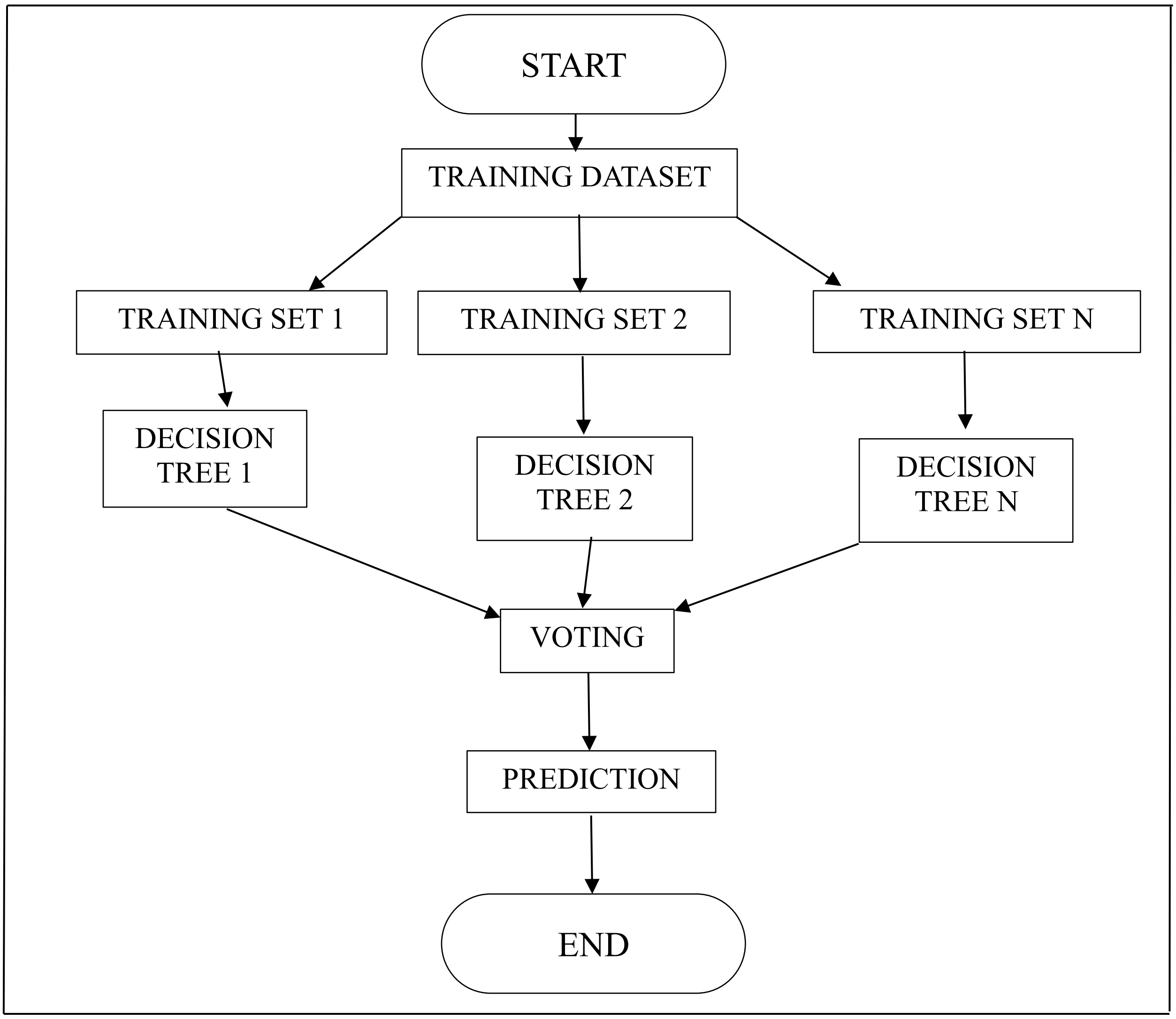


Fig 5.3: Random Forest Classifier Flowchart

4) K-Nearest Neighbor Classifier:

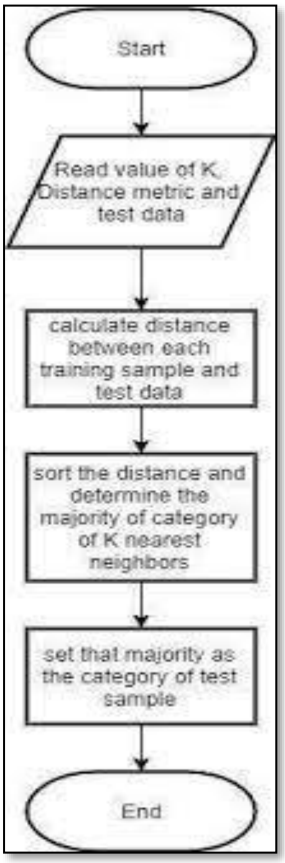

Start
Read value of K, Distance metric and test data
calculate distance between each training sample and test data
sort the distance and determine the majority of category of K nearest neighbors
set that majority as the category of test sample
End

Fig 5.4: K-Nearest Neighbor Classifier Flowchart

5) Logistic Regression:

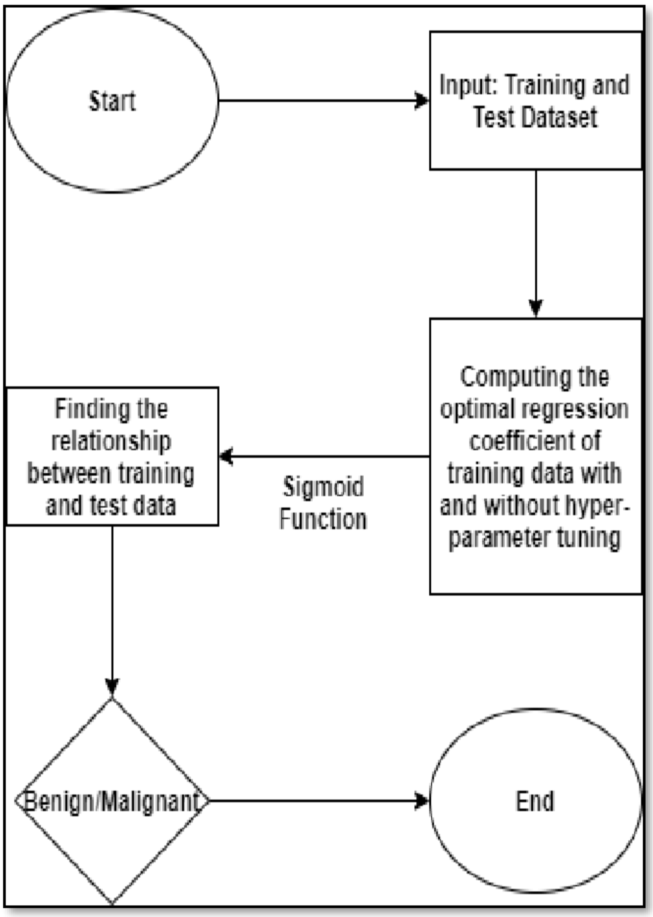

Start
Input: Training and Test Dataset
Computing the optimal regression coefficient of training data with and without hyper-parameter tuning
Sigmoid Function
Finding the relationship between training and test data
Benign/Malignant
End

Fig 5.5: Logistic Regression Classifier Flowchart

6. Benefits to the Society

- Early detection of breast cancer is crucial in order to cure the tumour, which is the basic aim of this paper.
- The mortality rate would drop down and a lot of valuable life can be saved once tumour is identified and cured at an early stage.
- This paper exhibits the study on various machine learning techniques and different machine learning algorithms which enhances one’s knowledge and creates a sense of awareness of the prevailing problems in the real-world.
- The research and methodology proposed in this study can be a starting point to accomplish further advancements and design even better models.
- The obtained model will be very helpful for the medical staff as well as general people.

7. Results and Outcomes [6][7]

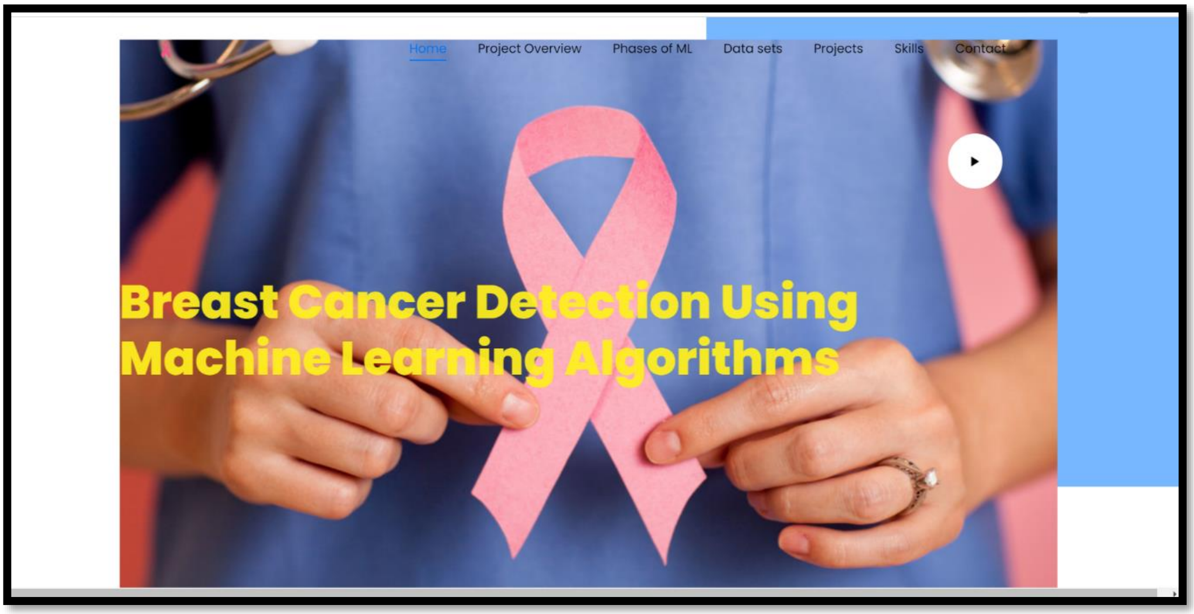
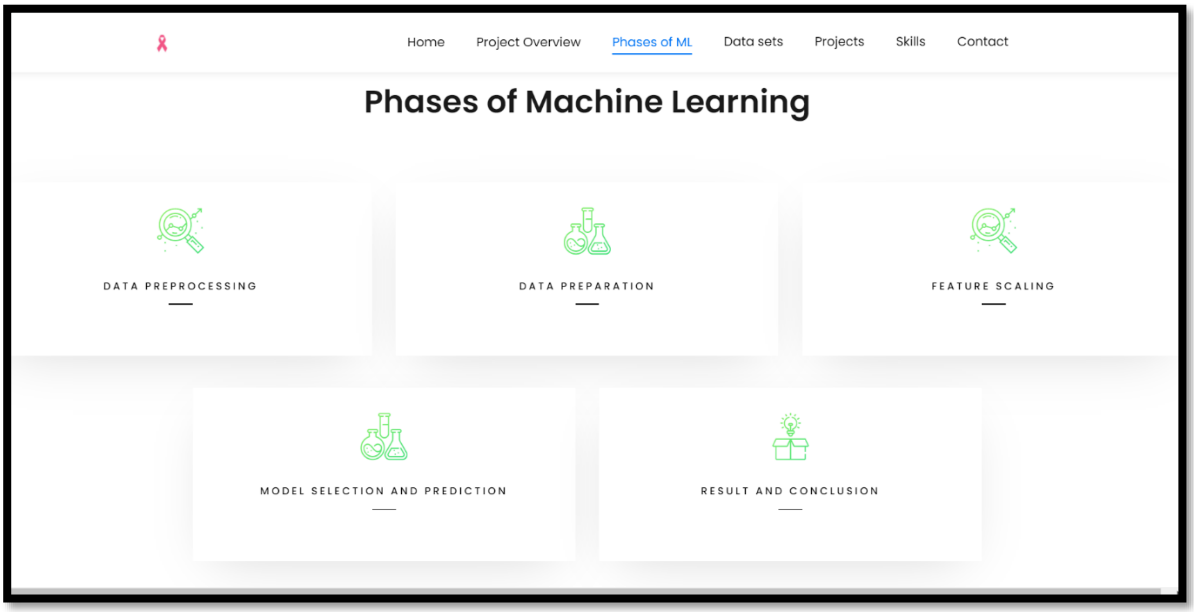


Fig 7.1: Home Page



Phases of Machine Learning

DATA PREPROCESSING

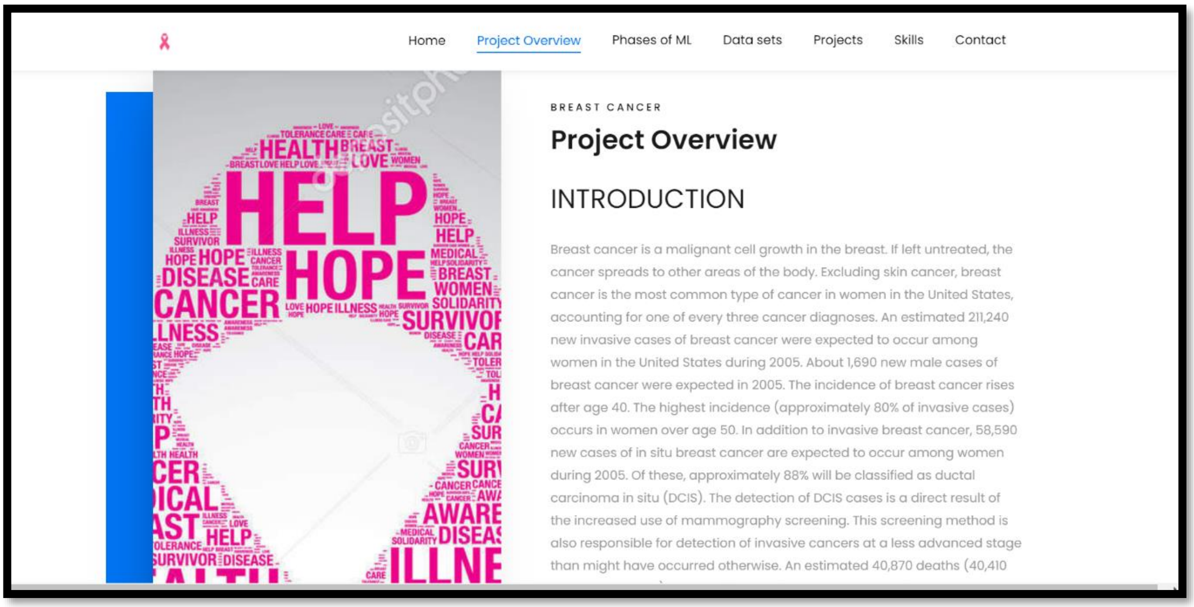
DATA PREPARATION

FEATURE SCALING

MODEL SELECTION AND PREDICTION

RESULT AND CONCLUSION

Fig 7.2: Phases of Machine Learning



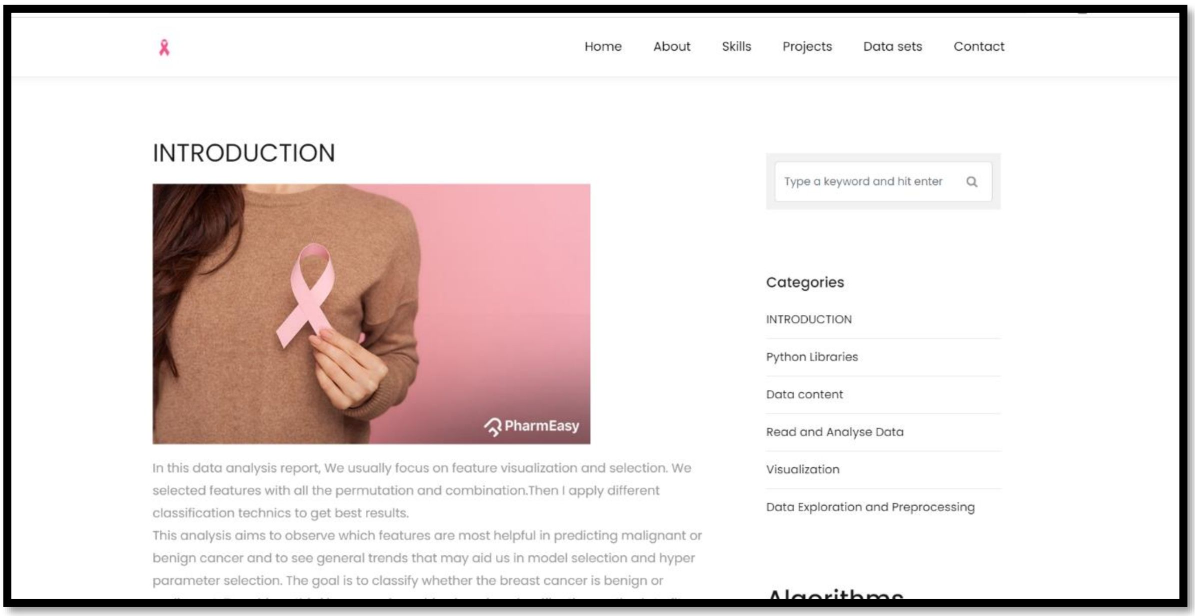
BREAST CANCER

Project Overview

INTRODUCTION

Breast cancer is a malignant cell growth in the breast. If left untreated, the cancer spreads to other areas of the body. Excluding skin cancer, breast cancer is the most common type of cancer in women in the United States, accounting for one of every three cancer diagnoses. An estimated 211,240 new invasive cases of breast cancer were expected to occur among women in the United States during 2005. About 1,690 new male cases of breast cancer were expected in 2005. The incidence of breast cancer rises after age 40. The highest incidence (approximately 80% of invasive cases) occurs in women over age 50. In addition to invasive breast cancer, 58,590 new cases of in situ breast cancer are expected to occur among women during 2005. Of these, approximately 88% will be classified as ductal carcinoma in situ (DCIS). The detection of DCIS cases is a direct result of the increased use of mammography screening. This screening method is also responsible for detection of invasive cancers at a less advanced stage than might have occurred otherwise. An estimated 40,870 deaths (40,410

Fig 7.3: Project Overview



INTRODUCTION

In this data analysis report, We usually focus on feature visualization and selection. We selected features with all the permutation and combination.Then I apply different classification technics to get best results.

This analysis aims 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. The goal is to classify whether the breast cancer is benign or

Categories

INTRODUCTION

Python Libraries

Data content

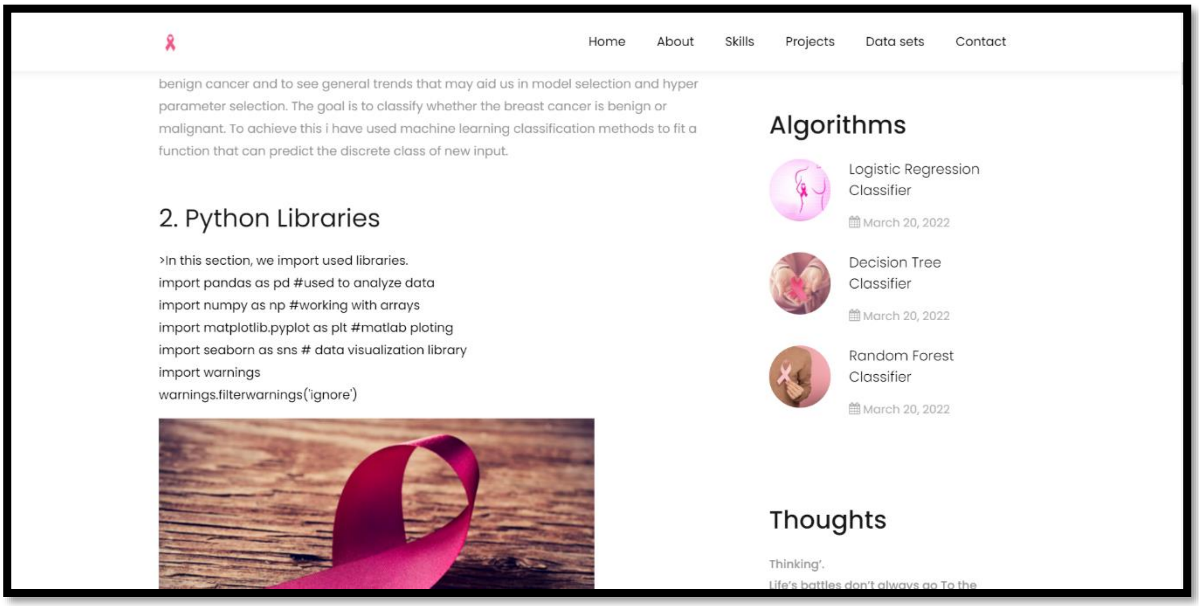
Read and Analyse Data

Visualization

Data Exploration and Preprocessing

Algorithms

Fig 7.4: Datasets and Algorithms Used



benign cancer and to see general trends that may aid us in model selection and hyper parameter selection. The goal is to classify whether the breast cancer is benign or malignant. To achieve this i have used machine learning classification methods to fit a function that can predict the discrete class of new input.

2. Python Libraries

```
>In this section, we import used libraries.
import pandas as pd #used to analyze data
import numpy as np #working with arrays
import matplotlib.pyplot as plt #matlab ploting
import seaborn as sns # data visualization library
import warnings
warnings.filterwarnings('ignore')
```

Algorithms

Logistic Regression Classifier
March 20, 2022

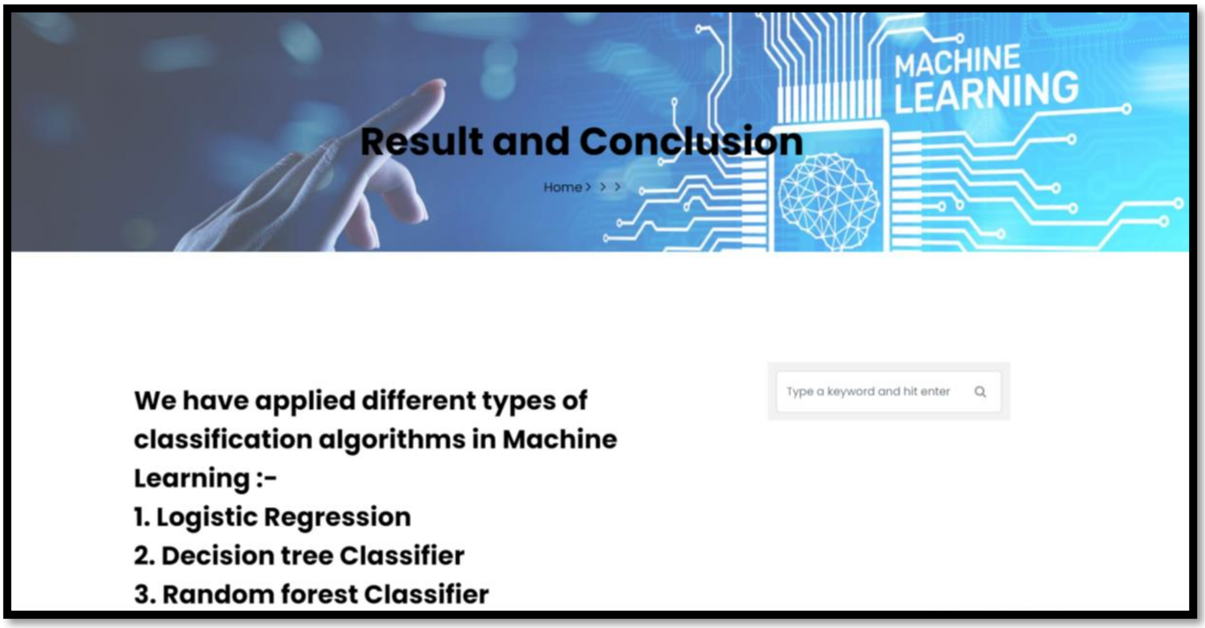
Decision Tree Classifier
March 20, 2022

Random Forest Classifier
March 20, 2022

Thoughts

Thinking'.
Life's battles don't always go To the

Fig 7.5: Algorithms Used in the project

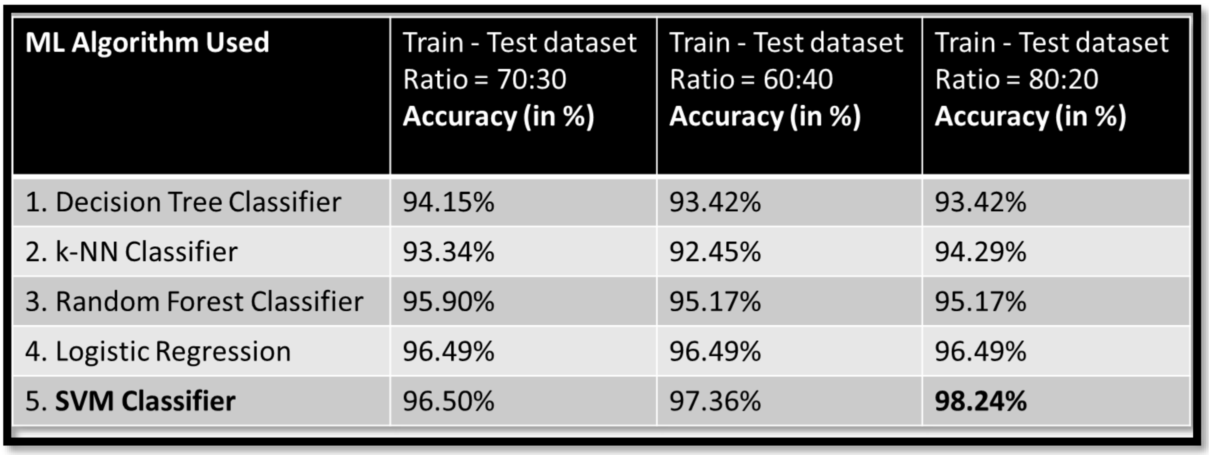


Result and Conclusion

We have applied different types of classification algorithms in Machine Learning :-
1. Logistic Regression
2. Decision tree Classifier
3. Random forest Classifier

Fig 7.6: Results and Conclusion Page

8. Conclusion



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%**

Table 8.1: Comparative study of the Accuracy of the ML Algorithms (in %)

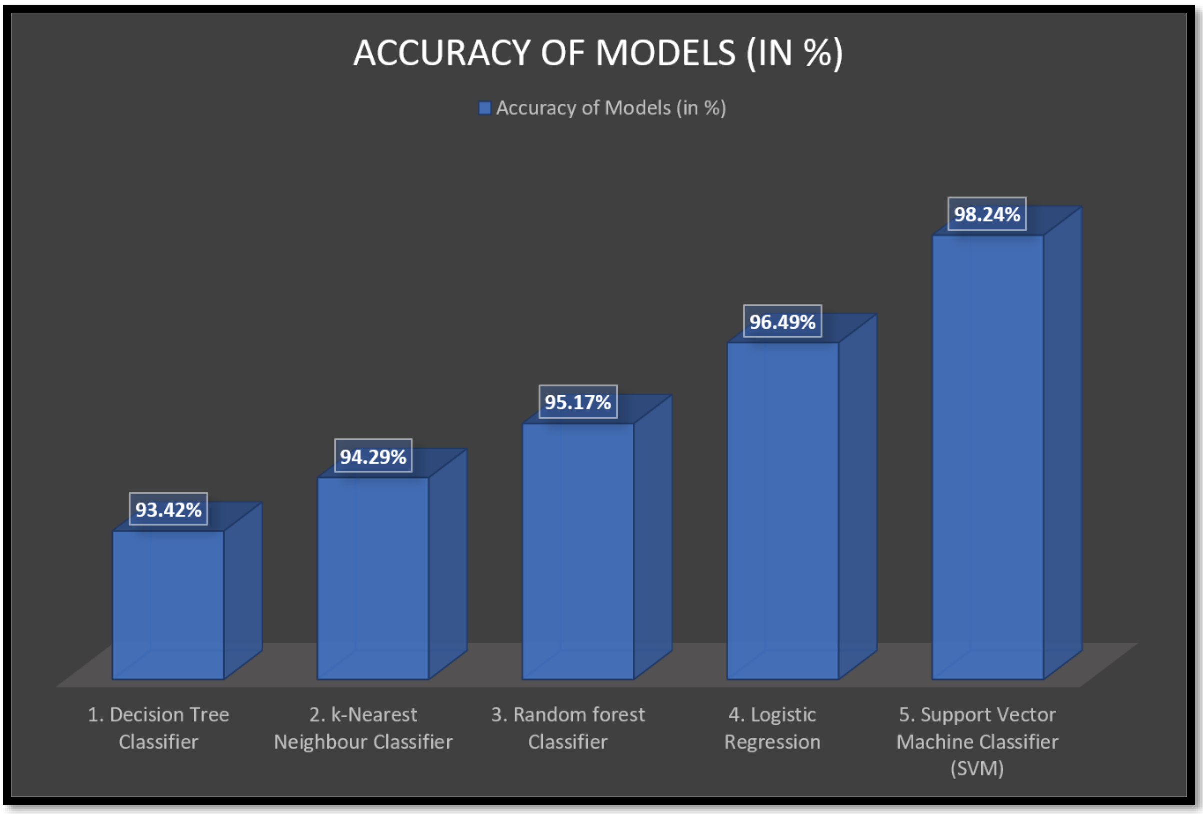


Fig 8.2: Comparative Analysis of the Accuracy of the applied ML models

This paper focuses on a very dangerous disease that causes death for many women worldwide. The aim of this study was to build an effective model successful at predicting the most severe cancer which is the breast cancer. The five different machine learning algorithms were applied on the Breast Cancer Wisconsin (Diagnostic) dataset after which different results obtained based on confusion matrix, accuracy, sensitivity, precision were evaluated. After an accurate comparison among the different classifier algorithms, we found that the **Support Vector Machine (SVM) Algorithm provided an accuracy of 98.24%** and outperforms all other algorithms. SVM has demonstrated its efficiency in breast cancer prediction and diagnosis and achieves the best performance in terms of accuracy and precision.

9. Future scope

This paper has evaluated the different machine learning algorithms on the proposed dataset which can be considered as a limitation of this work. Therefore, it is necessary for future works to apply these same algorithms and methods on other databases to confirm the results obtained via this database, as well as, in future works other machine leaning algorithms can be applied using new parameters on larger sets with more disease classes to obtain higher accuracy. In future more data would be added to the database which would increase help in better training of machine learning models and would work more accurately, which will also brief us about the relationship among various attributes. Evaluating different feature selection algorithms that can help us determine the smallest subset of features can assist in accurate classification in future.

10 . 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]