Abstract

Cancer is a serious public health issue worldwide and it is one of the leading causes of female deaths worldwide. It has caused more deaths than any other diseases such as tuberculosis or malaria. The World Health Organization (WHO) agencies for a cancer research report that 17.1 million new cancer cases are recorded in 2018 worldwide. Breast cancer is among the 4 leading cancers in women worldwide (ie, lung, breast and bowel, etc). Studies have shown that early detection and appropriate treatment of breast cancer significantly increase the chances of survival. Further, accurate classification of benign tumors can prevent patients from undergoing unnecessary treatments. Thus, the correct diagnosis of BC and classification of patients into malignant or benign groups is the subject of much research.

This machine learning project uses a dataset that can help determine the likelihood that a breast tumor is malignant or benign. Various factors are taken into consideration, including the lump's thickness, number of bare nuclei, and mitosis. Machine learning (ML) has become a vital part of medical imaging research and with the recent advancements, it has made a significant impact on improving the diagnostics capabilities of CAD systems. The goal is to classify whether the breast cancer is benign or malignant with higher accuracy.

Table of Contents

Title **Page No.**

List of Tables

List of Figures

Abbreviations

Abbreviation	**Description**
IARC	International Agency for Research on Cancer
BC	Breast Cancer
WDBC	Wisconsin Diagnostic Breast Cancer
ML	Machine Learning
NB	Naive Bayes Algorithm
SVM	Support Vector Machine Algorithm
GRNN	General Regression Neural Network
KNN	K-Nearest Neighbour Algorithm
EM	Expectation-Maximization
CSV	Comma-seperated Values

CHAPTER 1

Introduction

1.1 Motivation

Abstract Breast cancer is a common cancer in women, and one of the major causes of death among women around the world. Invasive ductal carcinoma (IDC) is the most widespread type of breast cancer with about 80% of all diagnosed cases. Early accurate diagnosis plays an important role in choosing the right treatment plan and improving survival rate among the patients. In recent years, efforts have been made to predict and detect all types of cancers by employing artificial intelligence.

1.2 Problem Definition

Cancer is a serious public health issue worldwide and the second leading cause of death in the United States. According to the International Agency for Research on Cancer (IARC), about 18.1 million new cases and 9.6 million deaths caused by cancer were reported in 2018. As shown in Fig. 1, breast cancer is a common cancer and one of the major causes of death worldwide with 627,000 deaths among 2.1 million diagnosed cases in 2018.[1]

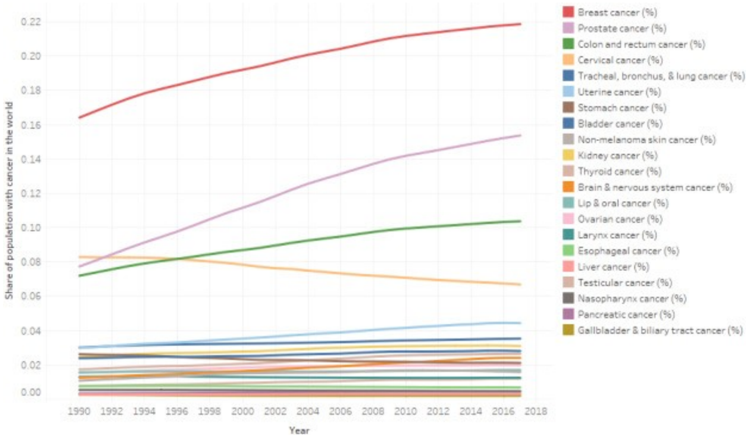


Figure 1.1: Share of Population with Cancer

The early diagnosis of BC can improve the prognosis and chance of survival significantly, as it can promote timely clinical treatment to patients. Further accurate classification of benign tumors can prevent patients undergoing unnecessary treatments. Thus, the correct diagnosis of BC and classification of patients into malignant or benign groups is the subject of much research. Because of its unique advantages in critical features detection from complex BC datasets, machine learning (ML) is widely recognized as the methodology of choice in BC pattern classification and forecast modelling. Classification and data mining methods are an effective way to classify data. Especially in medical field, where those methods are widely used in diagnosis and analysis to make decisions.

1.3 Objective of Project

• 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 malignant. To achieve this we have used machine learning classification methods to fit a function that can predict the discrete class of new input.

1.4 Limitations of Project

The limitations of this project are:

• All the algorithms we use may not give 100% accuracy.

•The dataset we use now doesn't give all the information efficiently.

CHAPTER 2

Literature Survey

2.1 Introduction

Early detection is the best way to increase the chance of treatment and survivability. Recently, multiple classifiers algorithms are applied on medical datasets to perform predictive analysis about patients and their medical diagnosis . For example, using machine learning techniques to assess tumor behavior for breast cancer patients. One problem is that there is a class imbalance in the training data, since the probability of not having this disease is higher than the one of having it. In recent years, several studies have applied data mining algorithms on different medical datasets to classify Breast Cancer. These algorithms show good classification results and encourage many researchers to apply these kinds of algorithms to solve challenging tasks. A list of some literature studies related to this method is presented below[2].

2.2 Existing System

Paper title	Datasets	Algorithms	Results
Integration of data mining 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%
A study on prediction of breast cancer recurrence using data mining techniques, 2017	WPBC	Classification: KNN, SVM, NB and C5.0, Clustering: K-means, EM, PAM and Fuzzy c-means	Classification accuracy is better than clustering, SVM & C5.0: 81%
Predicting breast cancer recurrence using effective classification and feature selection technique, 2016	WPBM	NB, C4.5, SVM	NB: 67.17%, C4.5: 73.73%, SVM: 75.75%

Table 2.1: Existing systems

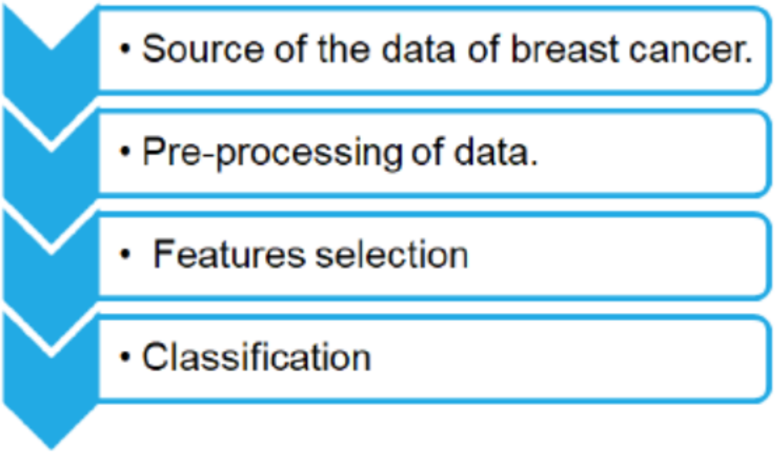
2.3 Disadvantages of Existing system

The existing models have been doing excellent job but they fail to provide the maximum accuracy with consistency. Existing models used Datasets like WPBC, WPBM, Breast Cancer which are not very efficient as compared to the WDBC Dataset which we have used in our proposed project.

2.4 Proposed System

- This proposed system presents a comparison of major machine learning (ML) algorithms. The dataset used is obtained from the Wisconsin Diagnostic Breast Cancer (WDBC) which is very efficient in classification of breast tumor. For the implementation of the ML algorithms, the dataset was preprocessed and then partitioned into the training set and testing set.

- A comparison between different algorithms will be made. The algorithm that gives the best results will be supplied as a model to the user. We are using supervised learning algorithms because the performance of supervised learning algorithms is better than the performance of unsupervised learning algorithms[3]



- Source of the data of breast cancer.
- Pre-processing of data.
- Features selection
- Classification

Figure 2.1: Proposed Diagnosis Model

CHAPTER 3

ANALYSIS

3.1 Software Requirement Specification

3.1.1 User requirement

- Initialize the Application

3.1.2 Software requirements

- Python
- Jupyter Notebook
- Numpy
- Pandas
- Matplotlib
- Seaborn
- sklearn

3.1.3 Hardware Requirements

- Laptop/ Personal Computer (PC)
- Random Access Memory (RAM): 1 GB or above
- Central Processing Unit (CPU): 1.7 GHz Processor and above
- Operating System (OS): Windows 8 and above

3.2 Content diagram or Architecture of Project

In our proposed we use WDBC dataset, the data is splitted into Training Set and Testing Set. They both are divided 70 percent and 30 percent respectively. After Splitting the dataset we have to preprocess the dataset and apply various machine learning algorithms. The model with the highest accuracy is used to predict the output.

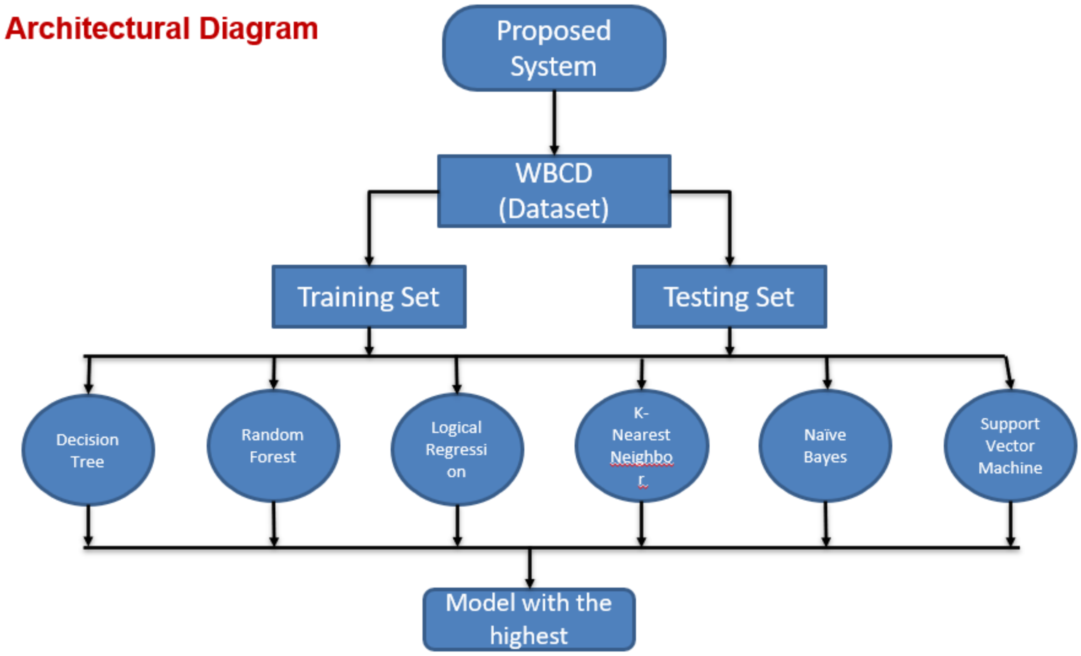


Figure 3.1: Basic Architecture Diagram of Proposed System

3.3 Algorithms

3.3.1 Logical Regression

Logistic regression is a supervised learning classification algorithm used to predict the probability of a target variable. It is used for predicting the categorical dependent variable using a given set of independent variables. Logistic regression model predicts P(Y=1) as a function of X.

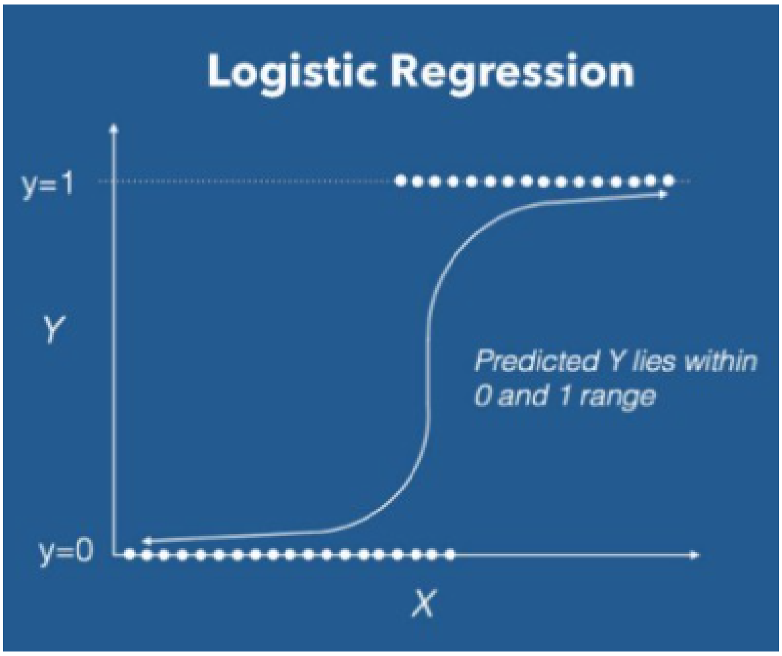


Figure 3.2: Logical Regression

3.3.2 K-Nearest Neighbors Classifier

K-NN algorithm assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories. Ex: In this data point is moved to category A as it have more no of neighbours.

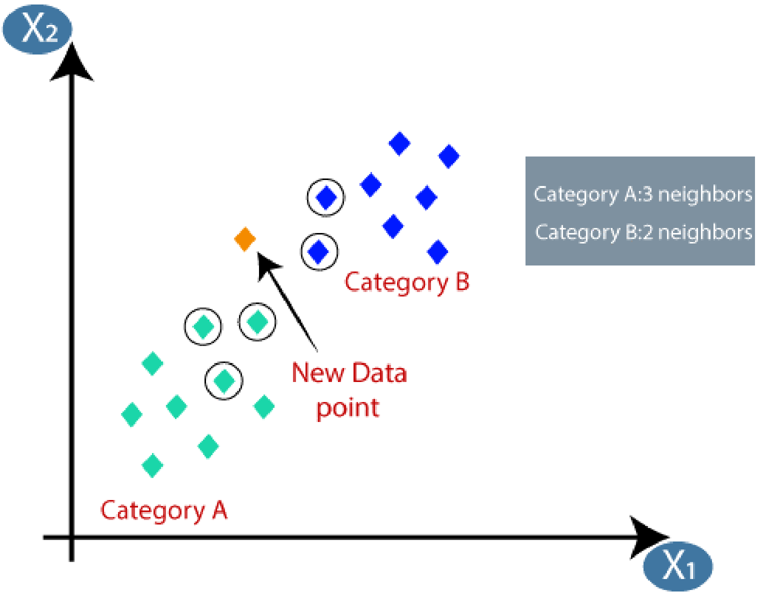


Figure 3.3: KNN Algorithm

3.3.3 Support Vector Machine Algorithm

SVM is used for Classification as well as Regression problems. The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future. This best decision boundary is called a hyperplane.

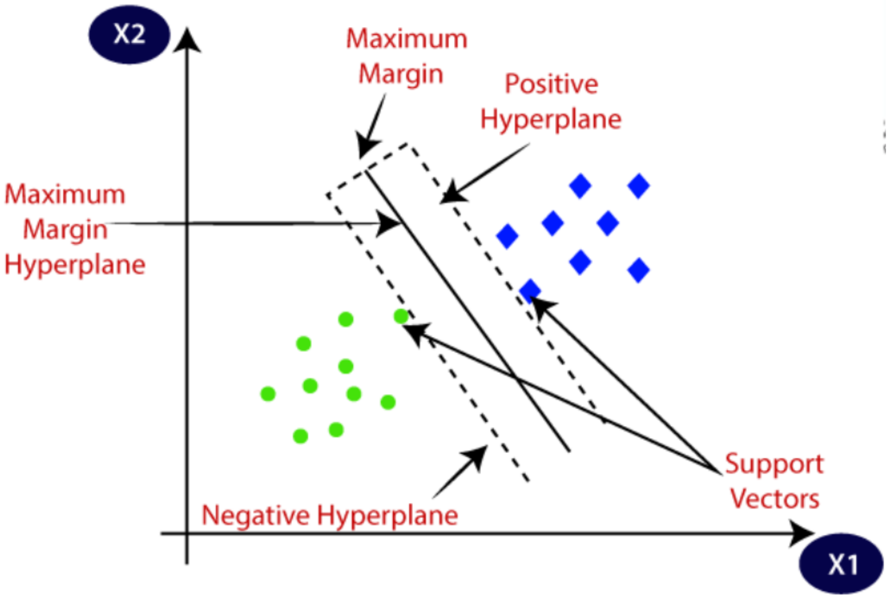


Figure 3.4: SVM Algorithm

3.3.4 Gaussian Naïve Bayes

Naïve Bayes is a probabilistic classifier, which means it predicts on the basis of the probability of an object. The Gaussian model assumes that features follow a normal distribution. This means if predictors take continuous values instead of discrete, then the model assumes that these values are sampled from the Gaussian distribution

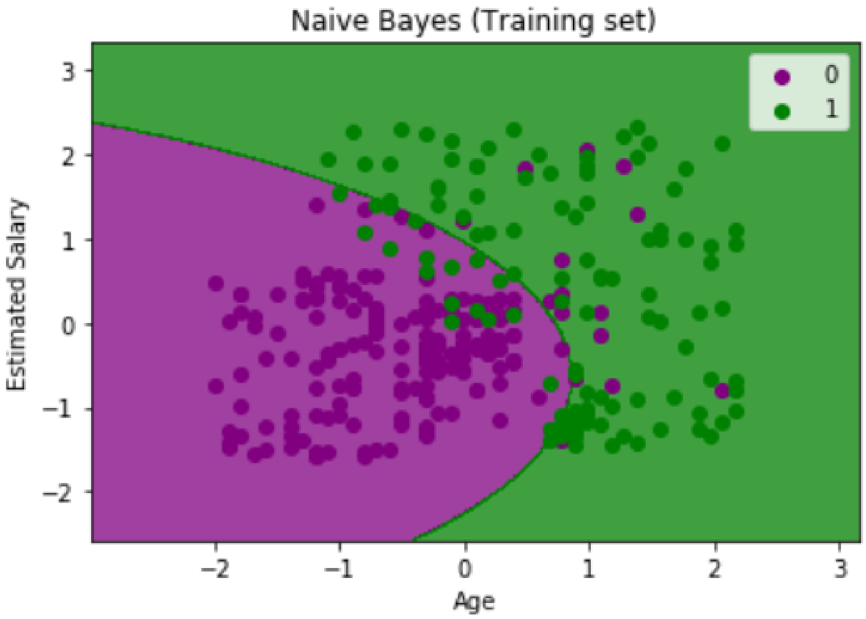


Figure 3.5: Naïve Bayes Algorithm

3.3.5 Decision Tree Regression

It decomposes a dataset into a series of smaller subsets while simultaneously developing an associated decision tree incrementally. The final result is a tree with decision nodes and leaf nodes which grows level wise.

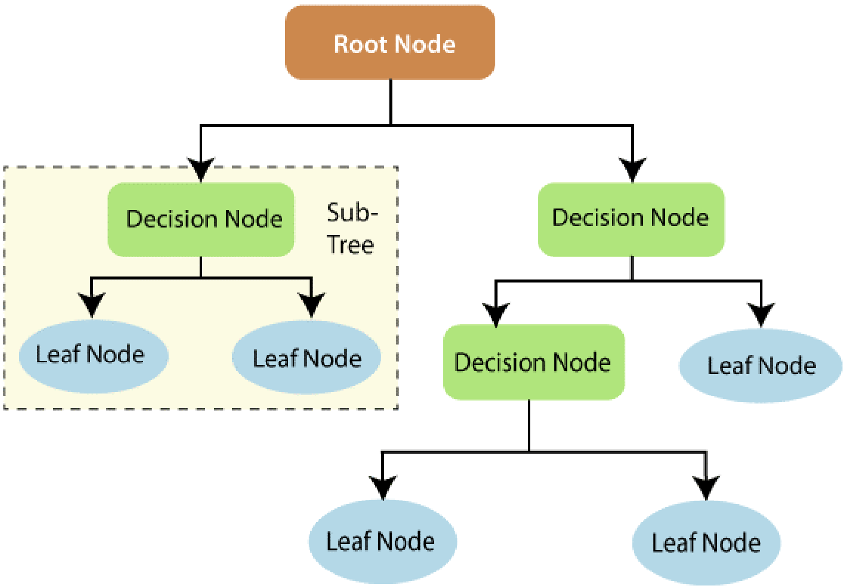


Figure 3.6: Decision Tree Algorithm

3.3.6 Random Forest Algorithm

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. In this technique we combines predictions from multiple machine learning algorithms to make a more accurate prediction than a single model.

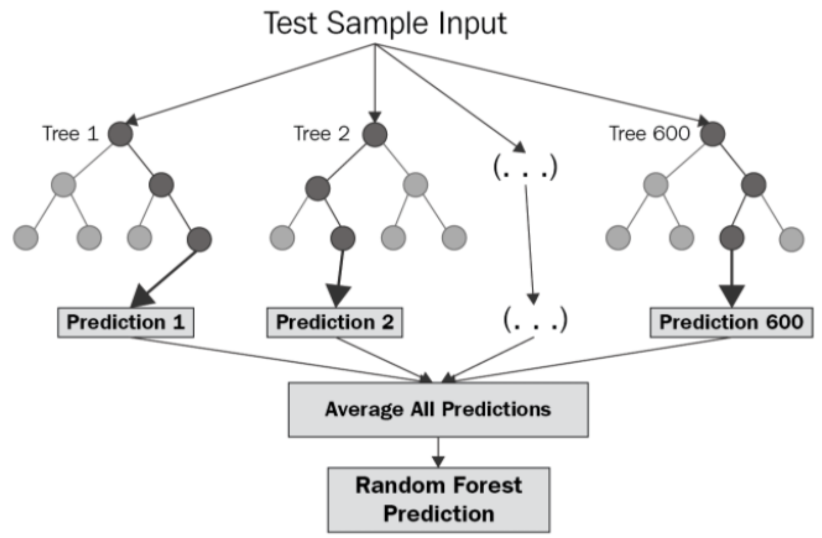


Figure 3.7: Random Forest Algorithm

CHAPTER 4

DESIGN

4.1 Introduction

Here in this proposed system, we have developed our application using python and some basic python modules to read and write files as we are finding accuracy using different algorithms in machine learning.

4.2 UML diagram

Below is the Use-Case diagram of our system as it helps to capture the functional requirements and easily traceable. Also helps to development guidelines to programmers, to a test case and finally into user documentation.

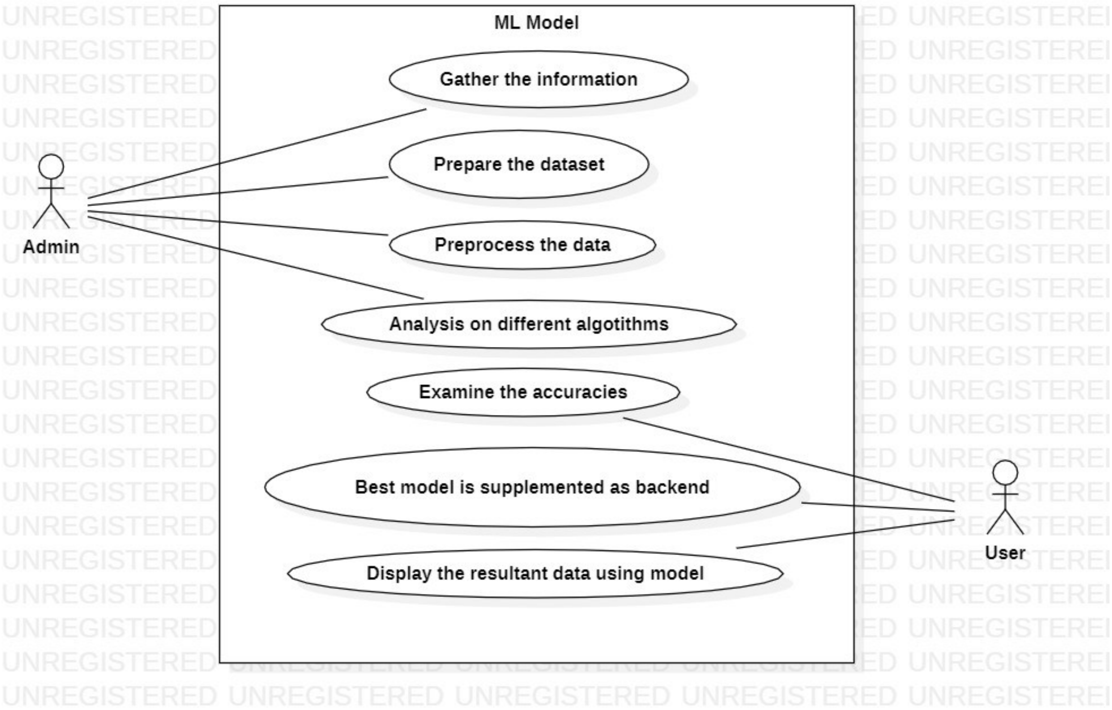
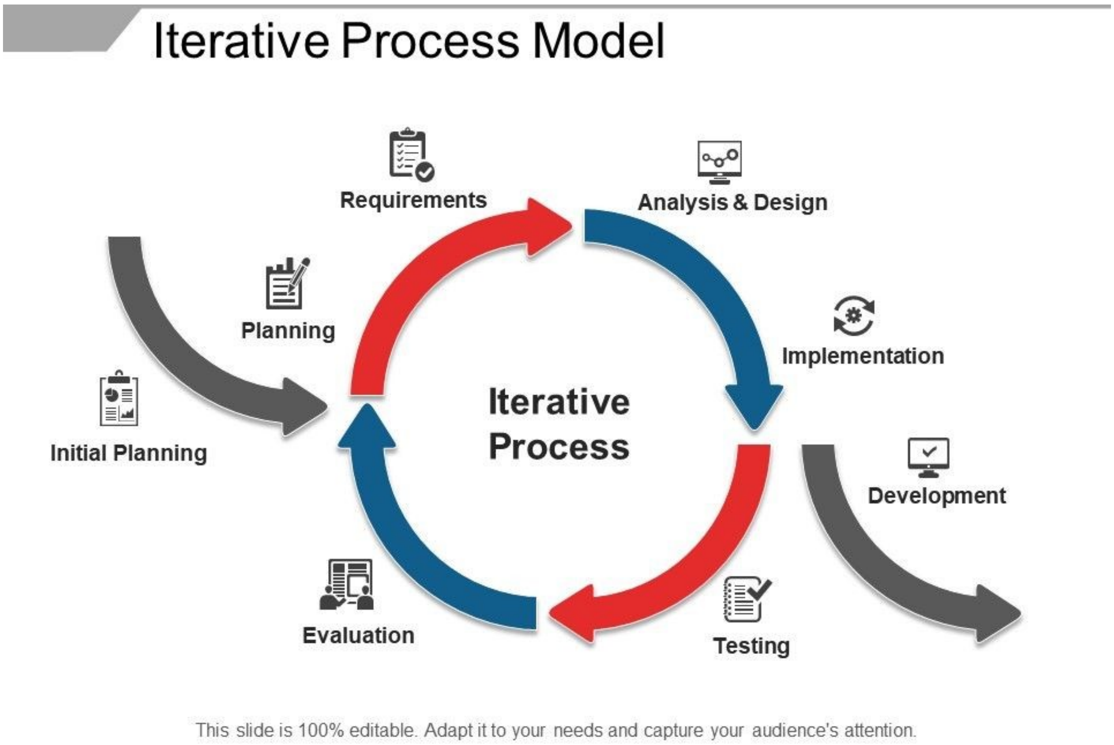

ML Model
Gather the information
Prepare the dataset
Preprocess the data
Analysis on different algotithms
Examine the accuracies
Best model is supplemented as backend
Display the resultant data using model
Admin
User

Figure 4.1: Use-Case diagram of proposed system

4.3 Process Model

In our proposed system we used iterative process model. Firstly, we will gather the required dataset and load into the program. After that, we will move on to the design phase where we will pre-process the data and split it into training and testing data and implement an algorithm. We will repeat this process by various algorithms and take out the best algorithm which provides highest accuracy.



Iterative Process Model

This slide is 100% editable. Adapt it to your needs and capture your audience's attention.

Figure 4.2: Process Model

CHAPTER 5

IMPLEMENTATION & RESULTS

5.1 Introduction

We have developed application, compatible for Windows OS, MacOS and Linux as this is a python program implemented in Jupiter notebook.

This dataset is taken from Breast Cancer Wisconsin (Diagnostic) Data Set from Kaggle. Features are computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. They describe characteristics of the cell nuclei present in the image. The Dataset contains of 357 benign tumor and 212 malignant tumor patients. Our dataset contains nearly 33 attributes which can be able to determine whether the tumor is benign or malignant.

5.2 Method of Implementation

5.2.1 Data Preparation Exploratory Analysis

In Our Dataset [4], the outputs of the diagnosis are determined in the variable 'M' and 'B'. The count of the number of patients with Malignant (M) cancerous and Benign (B) non-cancerous cells are represented visually in below graph.

The dataset used in this work are vulnerable to missing and imbalanced data therefore, before performing the experiments, a large fraction of this work will be for preprocessing the data in order to enhance the classifier's performance. Preprocessing will focus on managing the missing values and the imbalanced data. To manage the missing attributes, all the instances with missing values are removed.

Then we have to encode the categorical data values into 0's and 1's (M means 1, B means 0)

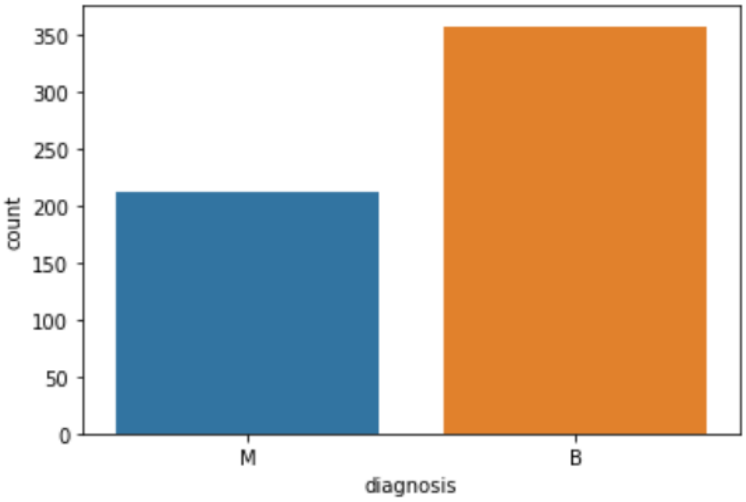


Figure 5.1: Chart displaying Malignant (cancerous) & Benign(non-cancerous) diagnosis

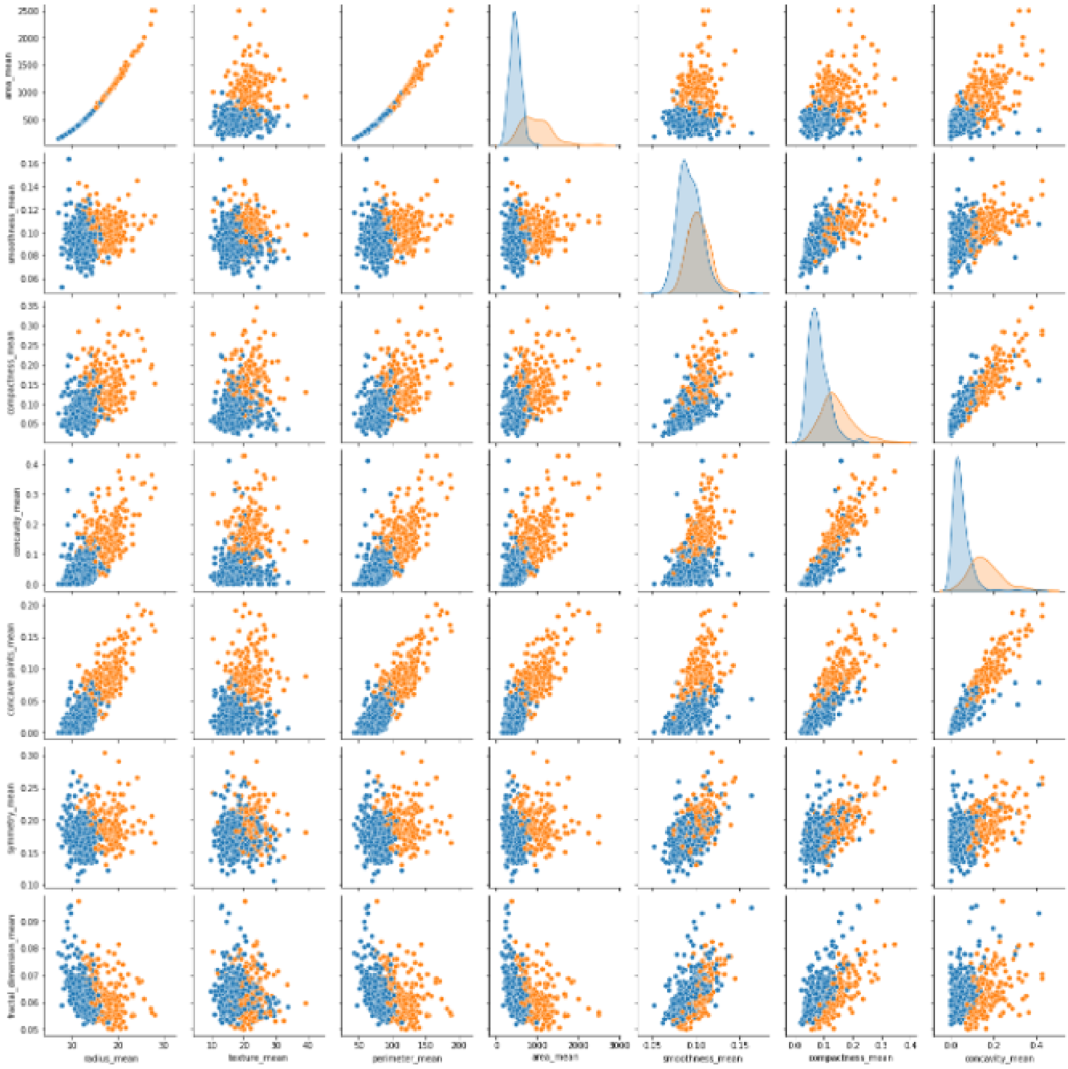
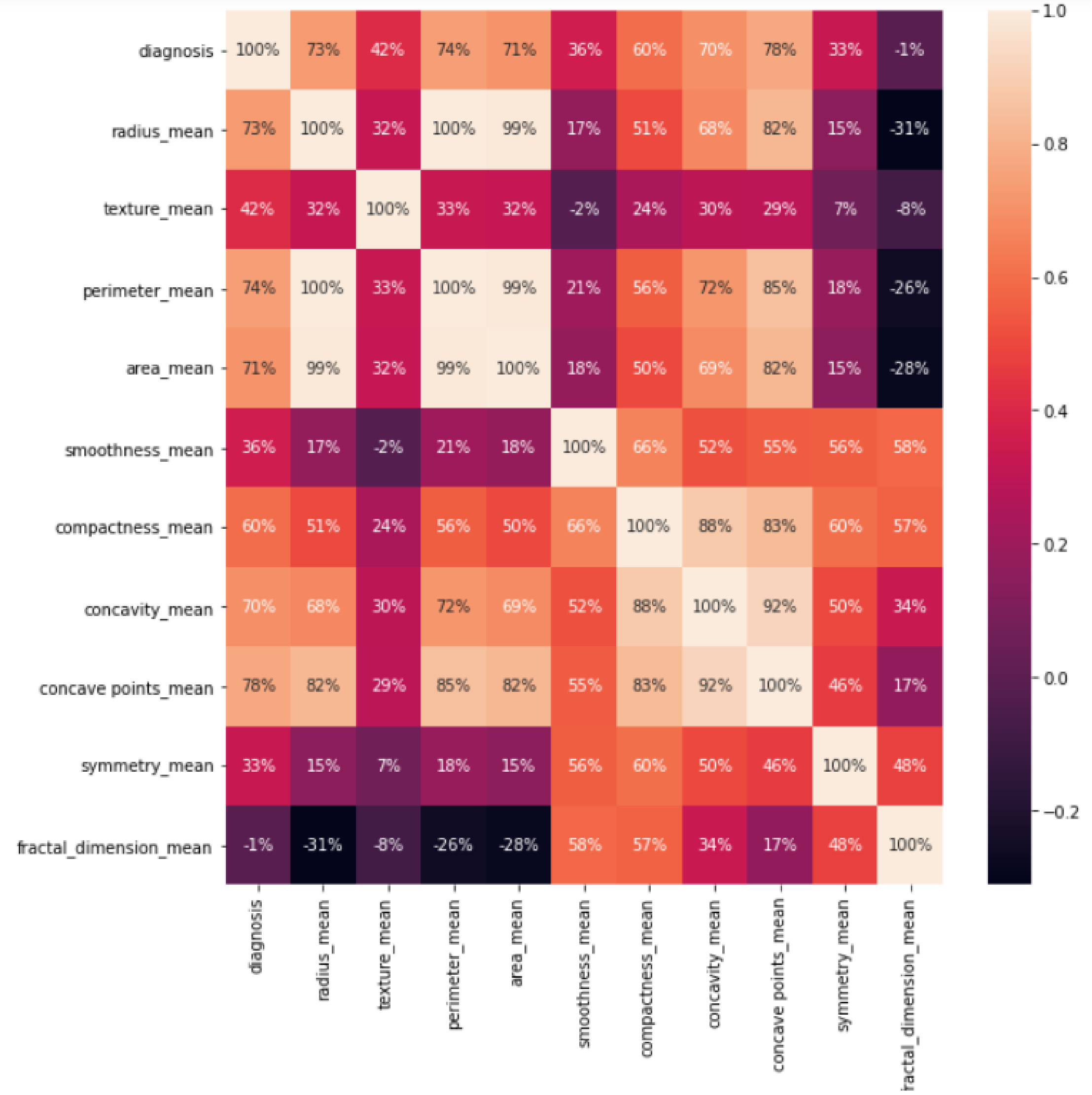


Figure 5.2: Pair plot of all of the columns highlighting the diagnosis points in Orange (1) & Blue (0)

The next visualization was to get an idea of how different paramenter are effecting the output. This was displayed by creating a distribution plot using the seaborn package.

A correlation matrix was then created to find the relationship between the parameters.We can see that all the means values are effecting the output of the diagnosis And the remaining parameters have little to no effect. After this we have designed a model using supervised algorithms (We are using supervised because we have labeled data).



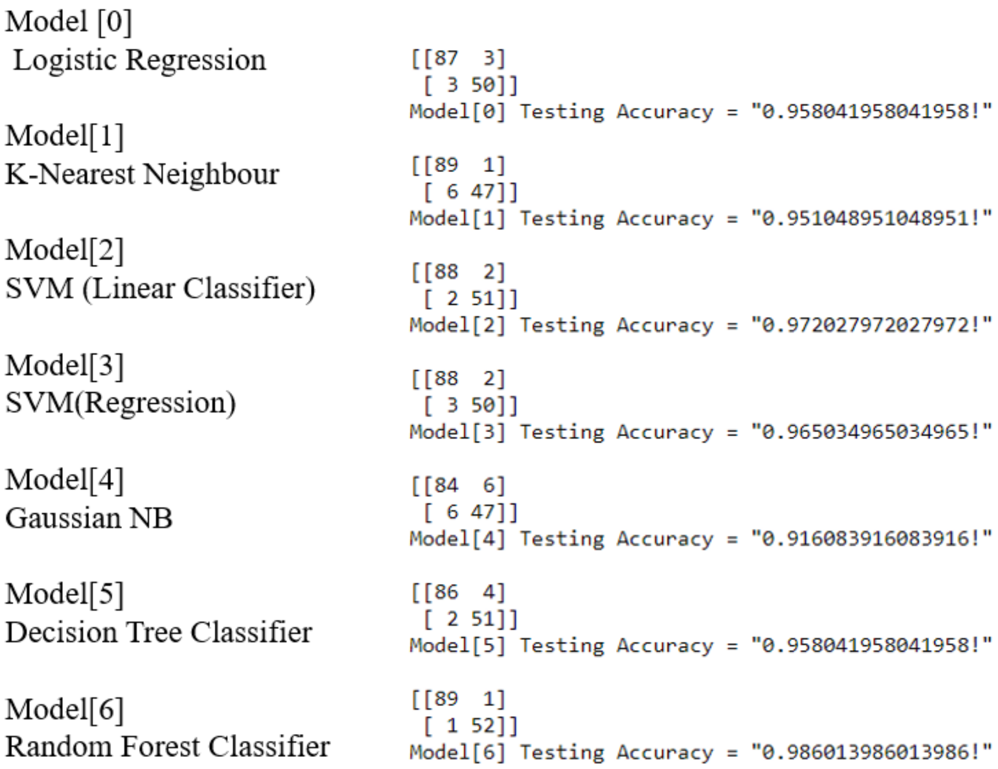
	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness_mean	concavity_mean	concave points_mean	symmetry_mean	ractal_dimension_mean
diagnosis	100%	73%	42%	74%	71%	36%	60%	70%	78%	33%	-1%
radius_mean	73%	100%	32%	100%	99%	17%	51%	68%	82%	15%	-31%
texture_mean	42%	32%	100%	33%	32%	-2%	24%	30%	29%	7%	-8%
perimeter_mean	74%	100%	33%	100%	99%	21%	56%	72%	85%	18%	-26%
area_mean	71%	99%	32%	99%	100%	18%	50%	69%	82%	15%	-28%
smoothness_mean	36%	17%	-2%	21%	18%	100%	66%	52%	55%	56%	58%
compactness_mean	60%	51%	24%	56%	50%	66%	100%	88%	83%	60%	57%
concavity_mean	70%	68%	30%	72%	69%	52%	88%	100%	92%	50%	34%
concave points_mean	78%	82%	29%	85%	82%	55%	83%	92%	100%	46%	17%
symmetry_mean	33%	15%	7%	18%	15%	56%	60%	50%	46%	100%	48%
fractal_dimension_mean	-1%	-31%	-8%	-26%	-28%	58%	57%	34%	17%	48%	100%

Figure 5.3: Heat Map of Correlation

5.3 Output Screen

5.3.1 Models Comparision

The key to a fair comparison of machine learning algorithms is ensuring that each algorithm is evaluated in the same way on the same data. You can achieve this by forcing each algorithm to be evaluated on a consistent test here we compared with 6 different algorithms



Model [0]
Logistic Regression

```
[[87  3]
 [ 3 50]]
Model[0] Testing Accuracy = "0.958041958041958!"
```

Model[1]
K-Nearest Neighbour

```
[[89  1]
 [ 6 47]]
Model[1] Testing Accuracy = "0.951048951048951!"
```

Model[2]
SVM (Linear Classifier)

```
[[88  2]
 [ 2 51]]
Model[2] Testing Accuracy = "0.972027972027972!"
```

Model[3]
SVM(Regression)

```
[[88  2]
 [ 3 50]]
Model[3] Testing Accuracy = "0.965034965034965!"
```

Model[4]
Gaussian NB

```
[[84  6]
 [ 6 47]]
Model[4] Testing Accuracy = "0.916083916083916!"
```

Model[5]
Decision Tree Classifier

```
[[86  4]
 [ 2 51]]
Model[5] Testing Accuracy = "0.958041958041958!"
```

Model[6]
Random Forest Classifier

```
[[89  1]
 [ 1 52]]
Model[6] Testing Accuracy = "0.986013986013986!"
```

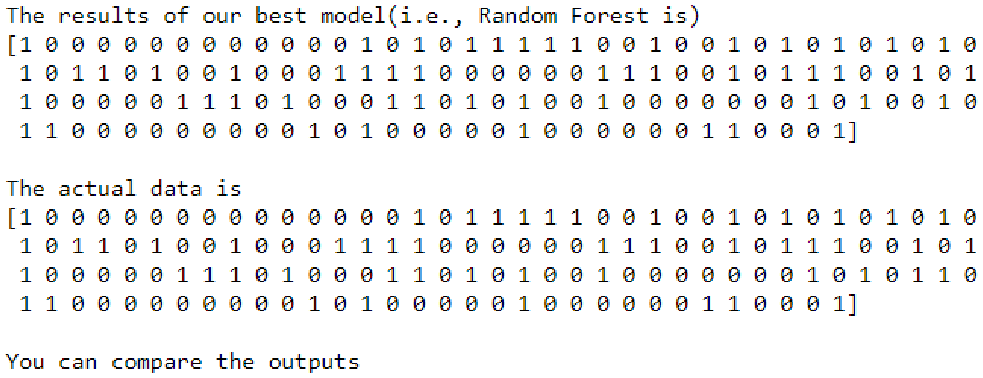
Figure 5.4: Testing accuraries of various algorithms

Output:

After testing our model with various algorithms we have taken Random Forest Classifier to predict the end output.

We have displayed the output which we got from our algorithm and the actual data from the dataset to test how many cases are correct manually.

Here 1 means cancerous(Malignant) and 0 means non-cancerous(Benign).



```
The results of our best model(i.e., Random Forest is)
[1 0 0 0 0 0 0 0 0 0 0 0 0 1 0 1 0 1 1 1 1 1 0 0 1 0 0 1 0 1 0 1 0 1 0 1 0
 1 0 1 1 0 1 0 0 1 0 0 0 1 1 1 1 0 0 0 0 0 0 1 1 1 0 0 1 0 1 1 1 0 0 1 0 1
 1 0 0 0 0 0 1 1 1 0 1 0 0 0 1 1 0 1 0 1 0 0 1 0 0 0 0 0 0 0 1 0 1 0 0 1 0
 1 1 0 0 0 0 0 0 0 0 0 1 0 1 0 0 0 0 0 1 0 0 0 0 0 0 1 1 0 0 0 1]

The actual data is
[1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 1 1 1 1 1 0 0 1 0 0 1 0 1 0 1 0 1 0 1 0
 1 0 1 1 0 1 0 0 1 0 0 0 1 1 1 1 0 0 0 0 0 0 1 1 1 0 0 1 0 1 1 1 0 0 1 0 1
 1 0 0 0 0 0 1 1 1 0 1 0 0 0 1 1 0 1 0 1 0 0 1 0 0 0 0 0 0 0 1 0 1 0 1 1 0
 1 1 0 0 0 0 0 0 0 0 0 1 0 1 0 0 0 0 0 1 0 0 0 0 0 0 1 1 0 0 0 1]

You can compare the outputs
```

Figure 5.5: Output of the Model

5.4 Result Analysis

Model Accuracy Analysis Diagram

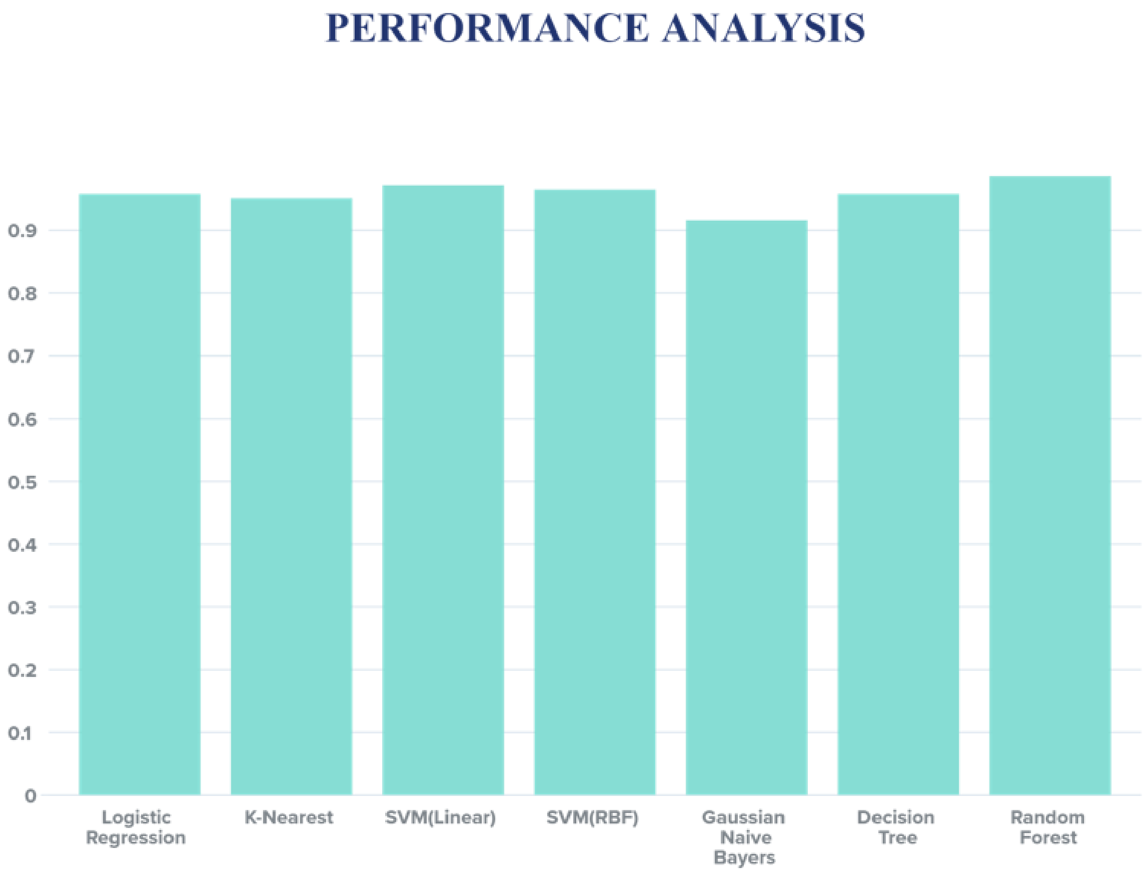


Figure 5.6: Model Accuracy analysis diagram

EXISTING SYSTEM	**PROPOSED SYSTEM**
The output data is based on datasets like WBC, WPMB, etc. which is not very efficient	The output data is based on WDBC dataset which is very effective in classifying the tumor.
The existing models have been doing excellent job but they fail to provide the maximum accuracy with consistency.	Our project mainly focuses by training and testing the datasets with the upcoming algorithms to achieve their accuracy.
Existing model gives GRNN and J48 as the best models with around 91% accuracy.	The proposed system gives Random Forest Classifier as the best model with 98% accuracy

Table 5.1: Existing vs Proposed System

CHAPTER 6

TESTING & VALIDATION

6.1 Introduction

In this testing phase, we need our application to run only when dataset is properly imported.

6.2 Design of test cases and scenarios

6.2.1 Scenario-1

User tries to run application with invalid data.

6.2.2 Scenario-2

User runs the application with invalid file path.

6.3 Validation

Case	Test Case	Expected Result	Actual Result
Null Values	1. Invalid data 2. Missing values in the file	1. Validation Error 2. Validation Error	1. Validation Error 2. Validation Error
Input File Path	1. Invalid Path (which doesn't exist) 2. Valid Path (but we uploaded a mp3 file) 3. Valid Path (valid CSV file)	1. Path Not Found 2. Expect an unknown file error. 3. Uploads Successfully	1. Path Not Found 2. Error: Parser Error 3. No Error

Table 6.1: Validations table

CHAPTER 7

Conclusion

Breast cancer is considered to be one of the significant causes of death in human. Early detection of breast cancer plays an essential role to save human's life. AI is set to change the medical industry in the coming decades — it wouldn't make sense for pathology to not be disrupted too. Currently, ML models are still in the testing and experimentation phase for cancer prognoses.[5] As datasets are getting larger and of higher quality, researchers are building increasingly accurate models.

We have proposed a project which uses multiple machine learning algorithms and select a model which will be able to classify and predict the cancer into benign or malignant with the best accuracy. These projects help the real-world patients and doctors to gather as much information as they can. Machine learning algorithms can be used for medical oriented research, it advances the system, reduces human errors and lowers manual mistakes. In the future, the same experiment will be applied to different classifiers and different datasets.

REFERENCES

[1] *U.S. Breast Cancer Statistics — Breastcancer.org.* URL: https://www.breastcancer.org/symptoms/understand_bc/statistics.

[2] Siham A. Mohammed, Sadeq Darrab, Salah A. Noaman, and Gunter Saake. "Analysis of Breast Cancer Detection Using Different Machine Learning Techniques". In: *Data Mining and Big Data.* Ed. by Ying Tan, Yuhui Shi, and Milan Tuba. Singapore: Springer Singapore, 2020, pp. 108–117. ISBN: 978-981-15-7205-0.

[3] Abderrahim Ghadi, Hajar Saoud, and M. Ghailani. "Proposed approach for breast cancer diagnosis using machine learning". In: Oct. 2019. ISBN: 978-1-4503-6289-4. DOI: 10.1145/3368756.3369089.

[4] *Breast Cancer Wisconsin (Diagnostic) Data Set — Kaggle.* URL: https://www.kaggle.com/uciml/breast-cancer-wisconsin-data/tasks?taskId=299.

[5] Meriem Amrane, Saliha Oukid, Ikram Gagaoua, and Tolga Ensarİ. "Breast cancer classification using machine learning". In: *2018 Electric Electronics, Computer Science, Biomedical Engineerings' Meeting (EBBT).* 2018, pp. 1–4. DOI: 10.1109/EBBT.2018.8391453.