**MUFFAKHAM JAH COLLEGE OF ENGINEERING TECHNOLOGY**

(Affiliated to Osmania University)

Mount Pleasant, 8-2-249, Road No. 3, Banjara Hills, Hyderabad-34.



**DEPARTMENT OF INFORMATION TECHNOLOGY**

***CERTIFICATE***

This is to certify that the Mini Project work titled “**Cancer Analysis**” is a bonafide work prescribed by the Osmania University for B.E III/IV IT V Semester during the academic year 2019-2020 carried out by **Nafeesa (1604-17-737-018)**

Faculty Incharge

A

Mini Project Report

On

**CANCER ANALYSIS**

By,

**Nafeesa (1604-17-737-018)**

Of

III/IV B.E. Sem-V (IT-A)

Under the Guidance of

Mr. Mohammad Riyazuddin Sir



DEPARTMENT OF INFORMATION TECHNOLOGY

**MUFFAKHAM JAH COLLEGE OF ENGINEERING AND TECHNOLOGY** (Affiliated to Osmania University)

Mount Pleasant, 8-2-249, Road No. 3, Banjara Hills, Hyderabad-34.

**ACKNOWLEDGEMENT**

The austerity and satisfaction that one gets on completing a project cannot be fulfilled without mentioning the people who made it possible with gratitude.

We are grateful to the almighty God who helped us all the way throughout the project and also has molded us into what we are today. We express our sincere thanks to our parents who encouraged us always to achieve our goals.

We offer our sincere thanks to Muffakham Jah College of Engineering and Technology for allowing us to do our mini project in their esteemed institution.

We show gratitude to **Dr. N. Seetharamaiah sir, M.Tech, PhD** and **Principal** for having provided all the facilities and support. We would like to thank **Dr. Mousmi Ajay Chaurasia,** (**Head of the Department, Information Technology)** for her expert guidance and encouragement at various level of project.

We are thankful to our guide **Mr. Mohammad Riyazuddin Sir Assistant Professor of the Department** and for his sustained inspiring Guidance and cooperation throughout the process of this project report.

We express our deep sense of gratitude and thanks to all the **Teaching** and **Non-Teaching Staff** of our college who stood with us and helped us to make it a successful venture.

**ABSTRACT**

**Breast Cancer Analysis: (using Data Science (ML))**

**Data science** is a multidisciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from structured and unstructured data.Data science is broadly divided into Artificial Intelligence, Machine Learning and Deep Learning. We are going to analyse the dataset of cancer patients using Machine Learning. Machine learning is an application of artificial intelligence (AI) that provides systems the ability to automatically learn and improve from experience without being explicitly programmed. Machine learning focuses on the development of computer programsthat can access data and use it to learn for themselves.

Breast cancer is one of the most common cancers among women worldwide, representing the majority of new cancer cases and cancer related deaths according to global statistics, making it a significant public health problem in today’s society.

The aim of our project is to build an effective ML algorithm which could help in detecting the tumor in its early stage by the analysis of patient details. Here after scanning the reports, we can use the input parameters and predict whether the patient is suffering from cancer or not. Here we are just helping the patients to detect the cancer and reducing the time of doctors by predicting the stage of cancer using the latest technology.We have used supervised learning model in Machine learning and the dataset is collected from the Google AI dataset. The method used is classification and we have worked with all the 5 models which comes under classification.Based on the accuracy we can decide to go for which model.

**CONTENTS:**

**Title**  **Page No.**

1. **Introduction 1**
2. **Literature survey 2-9**
3. **System Analysis 10-11**
4. **Software Requirement specifications 12-17**
5. **System Design 18-19**
6. **Implementation 20-24**
7. **Testing 25-26**
8. **Output Screenshots 27-30**
9. **Conclusion 31-32**
10. **Future Enhancement 33-34**
11. **Bibliography 34-35**

**CHAPTER 1**

**INTRODUCTION**

**INTRODUCTION**

Breast Cancer is one of the leading cancers among women leading to death. Early detection and diagnosis might help to treat the disease. It is easy for an experienced and expertise doctor to detect cancer easily through specific patterns through observation. However, when it comes to giving probabilities they might fail sometimes.

Exact detection of cancer is not possible even by the expert. This leads to the idea of bringing computational analysis to get better results. A research shows that Machine Learning may improve the accuracy of diagnosis.

In Brause’s work, the result shows that the most experienced physician can diagnose upto 79.97% accuracy while, 91.1% correct diagnosis is achieved using Machine Learning.

Tumors are classified into two types - Malignant and Benign

With an unfortunate increasing rate of Breast Cancer also comes huge amount of data. This data can be made useful in furthering medical research, and much more to the application of data science and machine learning.

In this Project we analyze the Breast Cancer Dataset and compare the efficiency of various Machine Learning Algorithms so as to know and build the most efficient model.

**Objective:**

The goal is to classify whether the breast cancer is benign or malignant by building an efficient Machine Learning Model.To achieve this we have used machine learning classification methods to fit a function that can predict the discrete class of new input.

Also, we aim to investigate the best Machine Learning Algorithm from those used for this project.

**Scope:**

* In medical and research fields

1

**CHAPTER 2**

**LITERATURE SURVEY**

**LITERATURE SURVEY**

**Machine learning:**

Machine learning is a branch of artificial intelligence that aims at solving real life problems. It provides the opportunity to learn without being explicitly programmed and it is based on the concept of learning from data. The advantage of machine learning (ML) methods is that it improves overtime with experience as it is kept exposed to more and more data.

So in general, machine learning is about learning to do better in the future based on what was experienced in the past. The emphasis of machine learning is on automatic methods. In other words, the goal is to devise learning algorithms that do the learning automatically without human assistance. However, training that data is the most important part of Machine Learning.

It updates easily by adding a new patient‘s record.

The application of machine learning models on human disease diagnosis aids medical experts based on the symptoms at an early stage, even though some diseases exhibit similar symptoms. One of the important problems in multivariate techniques is to select relevant features from the available set of attributes.

**Python :**  We used Python Programming Language for the implementation of the code.

Python is a better programming language for Machine Learning and Data Science as it has the following benefits:

* It offers a rich Library Ecosystem
* Easily understood Language
* It is platform independent
* It offers a variety of Data Visualization Tools to pictorially view the data
* Python is an open source and that is the reason it is fully updated by the community in the form of library updation, these libraries or modules can be used by anyone around the world.

**Phases/ Methodology :**

**Phase 0 — Data Preparation**

We will use the UCI Machine Learning Repository for breast cancer dataset.

The dataset used in this story is publicly available and was created by Dr. William H. Wolberg, physician at the University Of Wisconsin Hospital at Madison, Wisconsin, USA. To create the dataset Dr. Wolberg used fluid samples, taken from patients with solid breast masses and an easy-to-use graphical computer program called Xcyt, which is capable of performing the analysis of cytological features based on a digital scan.

3

Attribute Information:

1. ID number 2) Diagnosis (M = malignant, B = benign)

Ten real-valued features are computed for each cell nucleus:

1. radius (mean of distances from center to points on the perimeter)
2. texture (standard deviation of gray-scale values)
3. perimeter
4. area
5. smoothness (local variation in radius lengths)
6. compactness (perimeter² / area — 1.0)
7. concavity (severity of concave portions of the contour)
8. concave points (number of concave portions of the contour)
9. symmetry
10. fractal dimension

# Phase 1 — Data Exploration

**Data Exploration** is about describing the data by means of statistical and visualization techniques. We explore data in order to bring important aspects of that data into focus for further analysis.

We will be using ***Jupyter Notebook application*** to work on this dataset.

# Phase 2 — Categorical Data

Categorical data are variables that contain label values rather than numeric values. The number of possible values is often limited to a fixed set.

For example,in our case Benign and Malignant are two labels

4

We will use Label Encoder to label the categorical data. Label Encoder is the part of SciKit Learn library in Python and used to convert categorical data, or text data, into numbers, which our predictive models can better understand.

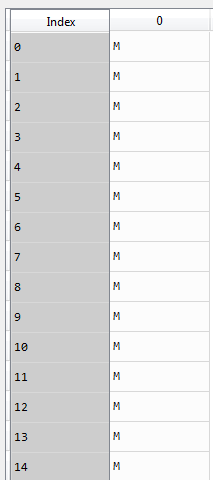
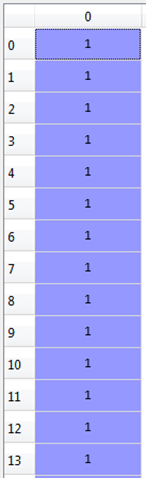
 

Fig2.1: Diagnosis Data without Encoding Fig2.2: Diagnosis Data Encoded

* **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.

5

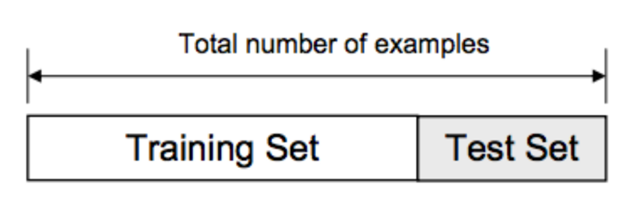


Fig2.3: Training and testing data set

**Phase 4 — Model Selection**

This is the most exciting phase in Applying Machine Learning to any Dataset. It is also known as Algorithm selection for Predicting the best results. Usually Data Scientists use different kinds of Machine Learning algorithms to large data sets.

**KNN Algorithm:**

Our analysis uses KNN algorithm. The **K Nearest Neighbours** algorithm is a well-known pattern recognition method, which is one of the best text classification algorithms. It is a simple algorithm that stores all available cases and classifies new cases based on a similarity measure (e.g., distance functions). KNN has been used in statistical estimation and pattern recognition already in the beginning of 1970’s as a non-parametric technique.

**Algorithm:**

A case is classified by a majority vote of its neighbours, with the case being assigned to the class most common amongst its K nearest neighbours measured by a distance function. If K = 1, then the case is simply assigned to the class of its nearest neighbour.

KNN algorithm fairs across all parameters of considerations. It is commonly used for its easy of interpretation and low calculation time.

6

# Let’s take a simple case to understand this algorithm. Following is a spread of red circles (RC) and green squares (GS) :

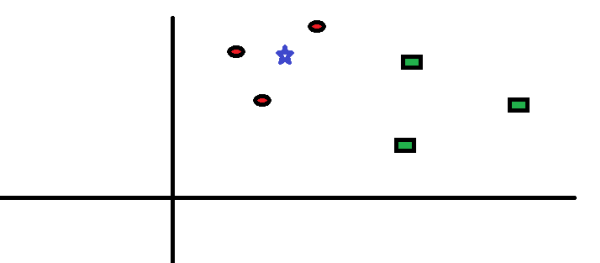


Fig 2.4: KNN Pictorially

You intend to find out the class of the blue star (BS) . BS can either be RC or GS and nothing else. The “K” is KNN algorithm is the nearest neighbors we wish to take vote from. Let’s say K = 3. Hence, we will now make a circle with BS as center just as big as to enclose only three datapoints on the plane. Refer to following diagram for more details:

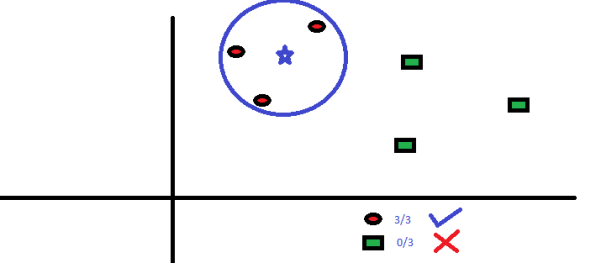


Fig2.5: KNN Algorithm found solution

7

The three closest points to BS is all RC. Hence, with good confidence level we can say that the BS should belong to the class RC. Here, the choice became very obvious as all three votes from the closest neighbor went to RC.

**Support Vector Machine Algorithm (SVM):**

It is a linear model for classification and regression problems. It can solve linear and non-linear problems and work well for many practical problems. The idea of SVM is simple: The algorithm creates a line or a hyperplane which separates the data into classes.

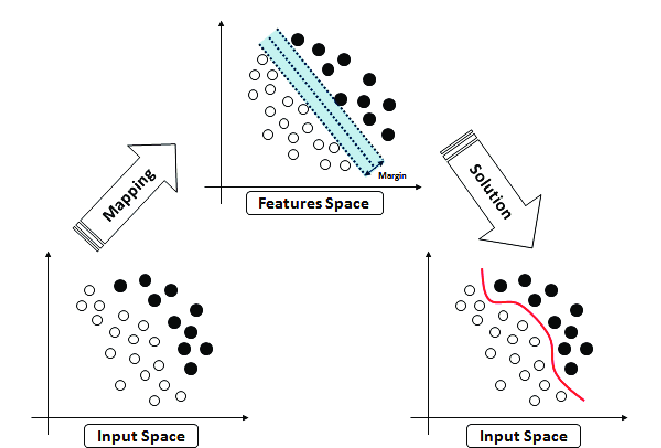
****

Fig 2.6: Support Vector Machine algorithm working

**Decision tree classifier:**

1. Place the best attribute of the dataset at the root of the tree.
2. Split the training set intosubsets. Subsets should be made in such a way that each subset contains data with the same value for an attribute.
3. Repeat step 1 and step 2 on each subset until you find leaf nodes in all the branches of the tree.

8

**Random Forest :**

Random forest classifier creates a set of decision trees from **randomly** selected subset of training set. It then aggregates the votes from different decision trees to decide the final class of the test object.

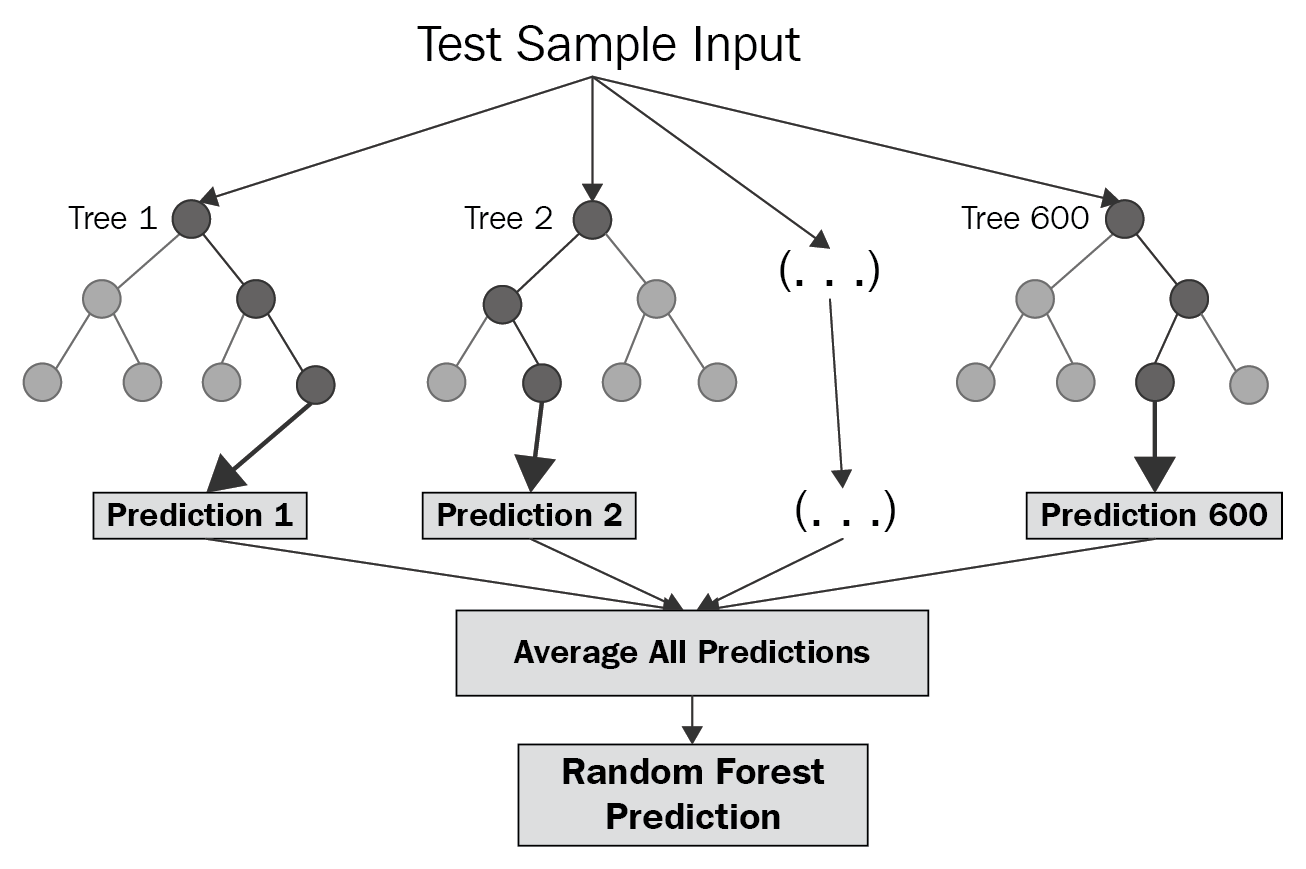


Fig 2.7: Random Forest working

9

**CHAPTER 3**

**SYSTEM ANALYSIS**

10

**SYSTEM ANALYSIS**

**EXISTING SYSTEM**

Doctors often use additional tests to find or diagnose breast cancer.

**Breast ultrasound**, a machine that uses sound waves to make detailed pictures, called sonograms, of areas inside the breast.

**Diagnostic mammogram**: If you have a problem in your breast, doctors may have you get a diagnostic mammogram.

**Magnetic resonance imaging (MRI)**, a kind of body scan that uses a magnet linked to a computer. **Biopsy**: This is a test that removes tissue or fluid from the breast to be looked at under a microscope and do more testing.

**Staging breast cancer**: Tests and procedures used to stage breast cancer may include, blood tests, such as a complete blood count, Mammogram of the other breast to look for signs of cancer, Breast MRI, Bone scan, Computerized tomography (CT) scan, Positron emission tomography (PET) scan.

**DRAWBACK**

Complete information about your cancer's stage may not be available until after you undergo breast cancer surgery.

Even after these procedures there is no clarity about the stage of the cancer, which is a huge issue when it comes to giving the right treatment according to the stages. Determining the stage of the breast cancer is a difficult task and gives no accurate result.

**PROPOSED SYSTEM**

In our project we look for the best machine learning algorithm that can predict the most accurate results and no human assistance is required in future. Once the new patient’s record is taken it provides the results based on it’s past experience.

**ADVANTAGE**

Only by taking the patient details the software is able to predict even the early stage of the patient without wasting their time and money, so that they could immediately proceed for the accurate procedures, which in turn may save their lives and also not produce any false information regarding their health.

11

**CHAPTER 4**

**SOFTWARE REQUIREMENTS SPECIFICATIONS**

12

**Software Requirements Specification**

**1. Introduction**

**1.1 Purpose**

Purpose is to train the train the dataset or in other words to build a model and test it’s efficiency of Cancer Detection using Machine Learning Algorithm and also to study Machine Learning Approach for breast Cancer diagnosis through implementation.

**1.2 Scope**

Breast cancer is the most common type of cancer among women world wide. Early detection is key in the treatment of breast cancer.

A research shows that Machine detects early stage cancer better than the expert himself. Therefore, steps can be taken There are steps you can take to detect breast cancer at early stage resulting in curable treatment.

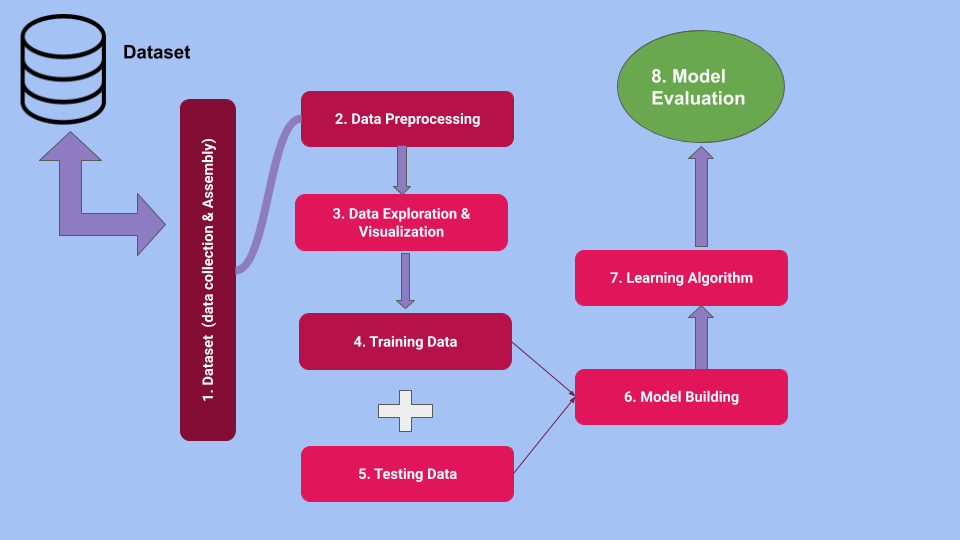
## Recommended Screening

**Mammography.** The most important screening test for breast cancer is the mammogram. A mammogram is an X-ray of the breast. It can detect breast cancer up to two years before the tumor can be felt by you or your doctor.

In our project, the software that is used in this project is Jupyter notebook(to provide python environment). We are able to diagnose using a machine learning (KNN Algorithm - K Nearest Neighbours). We are training the machine with all the possible inputs and outputs and finally testing our model for its efficiency.

13

**2. Overall Description of the model building:**



**2.2 Functions**

It performs the following actions:

* Get trained with all possible inputs and outputs.
* Is able to predict the stage of a cancer by taking the details of the patient.
* Has a good accuracy.

**2.3 User Classes and Characteristics**

Breast cancer is characterized by dysregulation of multiple cellular pathways  and different sensitivities to treatment, which may contribute to the wide range of survival durations for patients with metastatic disease. Some types of breast cancers are more aggressive than others. “Aggressive breast cancer” is not a standard term commonly used in the breast cancer literature. However, the ability to identify factors associated with aggressive breast cancer and to predict prognosis and treatment response has a considerable impact on patient management.

14

**2.4 Operating Environment**

The software that is used in this project is Jupyter Notebook which comes along with Anaconda Navigator.

The open-source Anaconda Distribution is the easiest way to perform Python/R data science and machine learning on Linux, Windows, and Mac OS X. With over 15 million users worldwide, it is the industry standard for developing, testing, and training on a single machine, enabling individual data scientists to:

* Quickly downloads 1,500+ Python/R data science packages
* Manage libraries, dependencies, and environments with [Conda](https://conda.io/docs/)
* Analyze data with scalability and performance with NumPy, pandas, Matplotlib etc,.

**3. Project Requirements:**

**3.1 Hardware Requirements:**

* Any system (Windows 10)

**3.2 Software Requirements:**

* Anaconda Navigator
* Jupyter Notebook with python 3.6.0

**4. System Features**

**4.1 Anaconda Navigator**

**4.1.1 Description and Priority**

It is a desktop graphical user interface (GUI) included in Anaconda distribution that allows you to launch applications and easily manage conda packages, environments and channels without using command-line commands. Navigator can search for packages on Anaconda Cloud or in a local Anaconda Repository. It is available for Windows, macOS, and Linux.

**4.1.2 Stimulus/Response Sequences**

Stimulus: User clicks on run the program

Response: Model efficiency is displayed

15

**4.1.3 Functional Requirements**

REQ-1: Final Model should give >=90% accuracy in predicting the correct output of cancer

**4.2 Jupyter Notebook**

**4.2.1 Description and Priority**

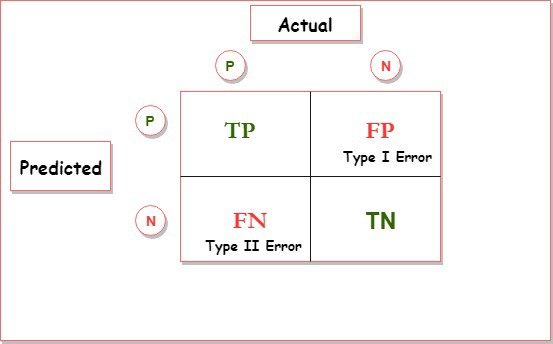
#### The Jupyter Notebook is an open-source web application that allows you to create and share documents that contain live code, equations, visualizations and narrative text. Uses include: data cleaning and transformation, numerical simulation, statistical modeling, data visualization, machine learning, and much more.

16

**5. Other Nonfunctional Requirements**

**5.1 Performance Requirements**

It is very important for us to have a good efficiency or accuracy rate of the model that we are building with respect to the confusion Matrix as shown below :



It needs to give true only to the patients who suffer from the Cancer & false if cancer is not detected in the reports.

Giving False assumptions/results may lead patients to dangerous situations

Therefore, performance or Accuracy must be >=90% at least for KNN Algorithm .

**5.2 Software Quality Attributes**

Correctness: The correctness of the output is very important in Cancer Analysis

Potability: The code is portable on all types of systems

Reusability : This code can be further used to build other machine learning models by changing approaches and in finding the best Machine Learning Algorithm to give maximum efficiency, based on which we could use the best model for Cancer Analysis after comparing all the algorithms.

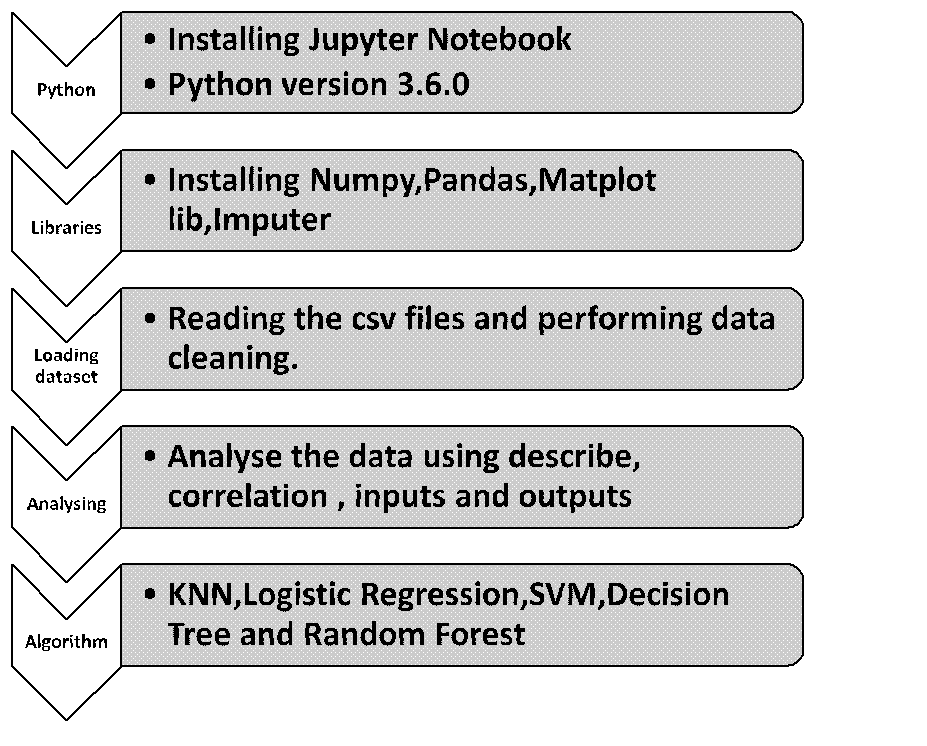
17

**CHAPTER 4**

**SYSTEM DESIGN**

18

**SYSTEM DESIGN:**



19

**CHAPTER 5**

**IMPLEMENTATION**

20

**CODE:**

import os

os.getcwd()

'C:\\Users\\DELL\\Desktop\\Data-Sets'

os.chdir("C:/Users/DELL/Desktop/Data-Sets")

os.getcwd()

'C:\\Users\\DELL\\Desktop\\Data-Sets'

import numpy as np

import pandas as pd

from sklearn import preprocessing

data1=pd.read\_csv("cancer.csv")

data1

import matplotlib.pyplot as plt

**#dividing the inputs into x and y to outputs its called feature engineering**

x=data1.iloc[:,1:31]# here 1 to 31 are inputs

y=data1.iloc[:,0] # 0 is the output that is diagnosis

**#here we are replacing 0 to benin and 1 to malign which is called recoding**

data1['diagnosis'].replace(['Malignant','Benign'],[1,0],inplace=True)

**#normalization is done when there are outliers**

minmax=preprocessing.MinMaxScaler(feature\_range=(0,1))

minmax.fit(x).transform(x)

from sklearn.model\_selection import train\_test\_split

x\_train,x\_test,y\_train,y\_test=train\_test\_split(x,y,test\_size=0.2)

x['area\_mean'].describe()

Out[543]:

count 569.000000

mean 654.889104

std 351.914129

min 143.500000

25% 420.300000

50% 551.100000

75% 782.700000

max 2501.000000

Name: area\_mean, dtype: float64

21

**CROSS-VALIDATION**

from sklearn import cross\_validation,neighbors

from sklearn.cross\_validation import train\_test\_split

x\_train,x\_test,y\_train,y\_test=train\_test\_split(x,y,test\_size=0.5)

#before these step that 1 and 0 are just assignment now we are saying that 0 and 1 are integers

y\_train=y\_train.astype('int')

y\_train

y\_test=y\_test.astype('int')

y\_test

from sklearn.metrics import accuracy\_score,recall\_score,roc\_auc\_score,confusion\_matrix

print("\n Accuracy score:%f" %(accuracy\_score(y\_test,y\_pred)\*100))

print("\n Recall score:%f" %(recall\_score(y\_test,y\_pred)\*100))

print("\n ROC score:%f" %(roc\_auc\_score(y\_test,y\_pred)\*100))

print(confusion\_matrix(y\_test,y\_pred))

Accuracy score:92.982456

Recall score:87.155963

ROC score:91.873436

[[170 6]

[ 14 95]]

In [557]:

**#dixal prob image**

plt.figure(dpi=150)

plt.hist(probas,bins=20)

plt.title('Classification Probabilities')

plt.xlabel('Probability')

plt.ylabel('# of Instances')

plt.xlim([0.5,1.0])

plt.legend(y\_test)

plt.show()

x\_train\_std=minmax.fit\_transform(x\_train)

x\_test\_std=minmax.transform(x\_test)

22

**KNN Classifier**

#fitting knn classifier

clf=neighbors.KNeighborsClassifier

clf.fit(x\_train,y\_train)

from sklearn.model\_selection import cross\_val\_score,cross\_val\_predict

#knn model name is clf, acc is accuracy, cv=3 is training,validation,test and that -1 is to get the best value

clf\_acc=cross\_val\_score(clf,x\_train\_std,y\_train,cv=3,scoring="accuracy",n\_jobs=-1)

array([0.98958333, 0.93617021, 0.95744681])

**logistic regression**

from sklearn.linear\_model import LogisticRegression

lr=LogisticRegression()

lr.fit(x\_train,y\_train)

lr.acc=cross\_val\_score(lr,x\_train\_std,y\_train,cv=3,scoring='accuracy',n\_jobs=-1)

array([0.98958333, 0.91489362, 0.95744681])

**Support Vector Machine**

from sklearn.svm import SVC

svc=SVC(kernel='rbf',probability=True)

svc\_classifier=svc.fit(x\_train,y\_train)

svc\_acc=cross\_val\_score(svc\_classifier,x\_train\_std,y\_train,cv=3,scoring="accuracy",n\_jobs=-1)

svc\_proba=cross\_val\_predict(svc\_classifier,x\_train\_std,y\_train,cv=3,method='predict\_proba')

svc\_acc

array([0.96875 , 0.91489362, 0.94680851])

**Decision tree classifier**

from sklearn.tree import DecisionTreeClassifier

dtc\_clf=DecisionTreeClassifier()

dtc\_clf.fit(x\_train,y\_train)

23

dtc\_clf\_acc=cross\_val\_score(dtc\_clf,x\_train\_std,y\_train,cv=3,scoring="accuracy",n\_jobs=-1)

dtc\_proba=cross\_val\_predict(dtc\_clf,x\_train\_std,y\_train,cv=3,method='predict\_proba')

dtc\_clf\_scores=dtc\_proba[:,1]

dtc\_clf\_acc

array([0.9375 , 0.89361702, 0.89361702])

**Building Random Forest Classifier**

from sklearn.ensemble import RandomForestClassifier

rmf=RandomForestClassifier(max\_depth=3,random\_state=0)

rmf\_clf=rmf.fit(x\_train,y\_train)

rmf\_clf\_acc=cross\_val\_score(rmf\_clf,x\_train\_std,y\_train,cv=3,scoring="accuracy",n\_jobs=-1)

rmf\_proba=cross\_val\_predict(rmf\_clf,x\_train\_std,y\_train,cv=3,method='predict\_proba')

rmf\_clf\_scores=rmf\_proba[:,1]

array([0.95833333, 0.90425532, 0.92553191])

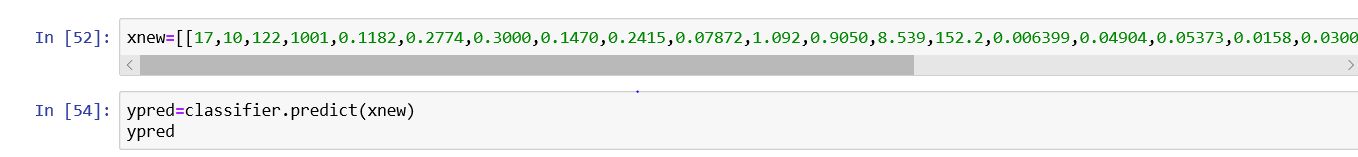
24

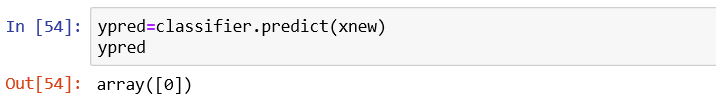
**CHAPTER 6**

**TESTING**

25

**TESTING WITH OUR OWN INPUTS :**

****

****

where, 0 is Malignant from code and 1 indicates Benign. Therefore, the patient has Malignant

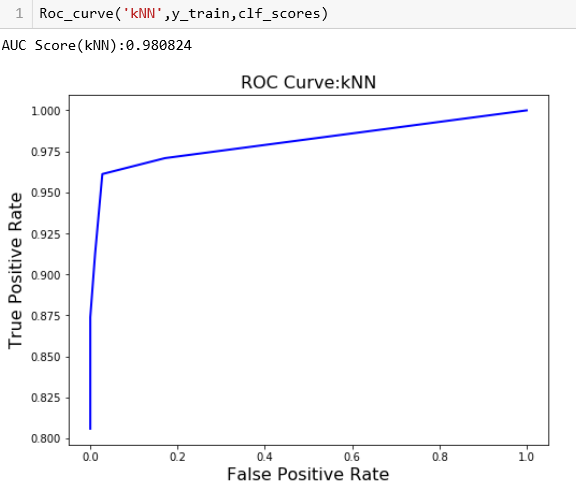
26

**CHAPTER 7**

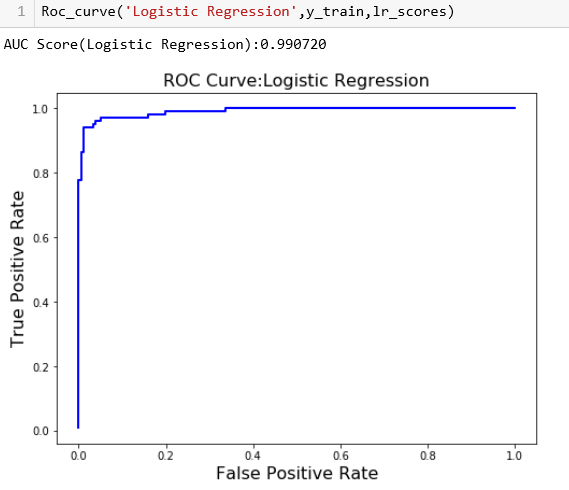
**OUTPUT SCREENSHOTS**

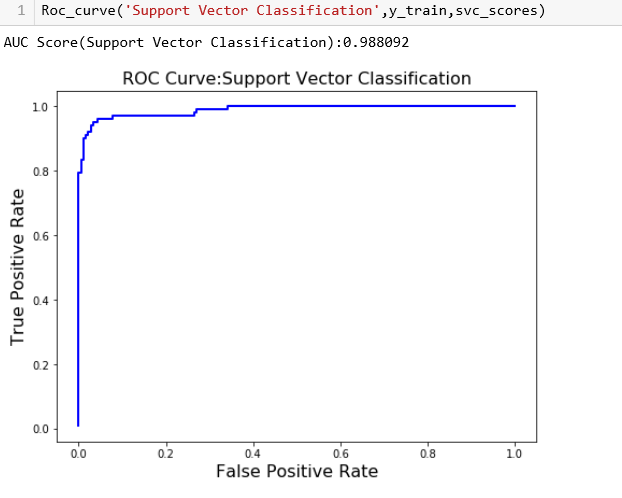
27

**Output Screenshots:**

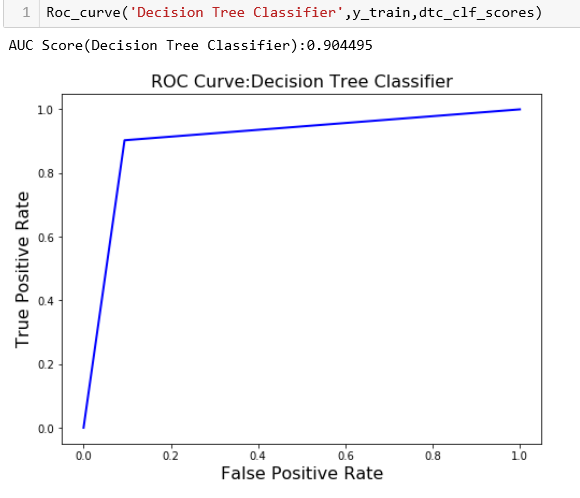
****

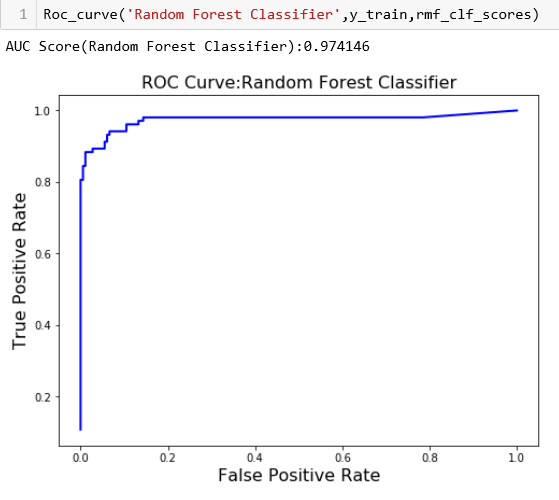
28

****

****

29

****

****

30

**CHAPTER 9**

**CONCLUSION**

31

**CONCLUSION:**

Machine Learning is undoubtedly an effective alternative in Breast Cancer prediction whose decision making capabilities is found to be much better than human and it easily keeps getting better with experience. The user needs to just add a new record and the algorithm does it’s task without Human Assistance yet, gives better results as confirmed in our project.

32

**CHAPTER 10**

**FUTURE ENHANCEMENTS**

33

**FUTURE ENHANCEMENT**

In this model, we have worked on the backend of the project and we would like to work on the front end of the model so that when we provide the inputs or scan the documents, we can the result on the screen detecting which stage of cancer. Not only for cancer we would like to combine all the healthcare departments and build a single platform where by using the data science we can predict from which disease the patient is suffering from. Not only the prediction we would also like to implement how to cure that disease by providing what kind of tablet to be consumed. As an engineer it is also our duty to save people's lives by using the technologies which we earn and implement them in the real world.

34

**CHAPTER 11**

**BIBLIOGRAPHY**

34

**BIBLIOGRAPHY**

1. Cancer Dataset - [www.googlescholar.com](http://www.googlescholar.com)
2. code references - kaggle   
   <https://www.kaggle.com/junkal/breast-cancer-prediction-using-machine-learning>   
   <https://medium.com/datadriveninvestor/breast-cancer-detection-using-machine-learning-475d3b63e18e>
3. Introduction to Machine Learning Algorithms - <https://www.analyticsvidhya.com/blog/2017/09/common-machine-learning-algorithms/>
4. For analysing Model’s superiority over human expert prediction - Research Paper from IEEE Xplore - by[Naresh Khuriwal](https://ieeexplore.ieee.org/author/37086397984) ; [Nidhi Mishr](https://ieeexplore.ieee.org/author/37708285900)a <https://ieeexplore.ieee.org/document/8748777>
5. Software to get jupyter Notebook - [www.anaconda.com](http://www.anaconda.com)

35