

**MINI PROJECT REPORT**

**On**

**“BREAST CANCER DETECTION”**

**BACHELOR OF TECHNOLOGY IN COMPUTER**

**SCIENCE AND INFORMATION TECHNOLOGY**

**Submitted to-**

**(MR. ROHIT AGARWAL SIR)**

**(Department of Computer Science & Information Technology)**

**Submitted by-**

**Student name TANISHA AGARWAL Roll No. 1900680110050**

**5TH SEMESTER**

**DEPARTMENT OF COMPUTER**

**SCIENCE AND INFORMATION TECHNOLOGY**

**MEERUT INSTITUTE OF ENGINEERING AND TECHNOLOGY, MEERUT**

# DECLARATION

We hereby declare that the project entitled - “Breast Cancer Detection”, which is being submitted as Mini Project in department of Computer Science and Information Technology to Meerut Institute of Engineering and Technology, Meerut (U.P.) is an authentic record of our genuine work done under the guidance of Prof. “Rohit Agarwal Sir” of Computer Science and Information Technology, Meerut Institute of Engineering and Technology, Meerut.

Date: 18/12/21 name of student: TANISHA AGARWAL

Place: MIET (Roll no.): 1900680110050

# CERTIFICATE

This is to certify that mini project report entitled – “BREAST CANCER DETECTION” submitted by “TANISHA AGARWAL” has been carried out under the guidance of Prof. “ROHIT AGARWAL SIR” of Computer Science and Information Technology, Meerut Institute of Engineering and Technology, Meerut. This project report is approved for Mini Project (KCS 554) in 5th semester in “COMPUTER SCIENCE AND INFORMATION TECHNOLOGY DEPARTMENT” from Meerut Institute of Engineering and Technology, Meerut.

**Internal Examiner**

**Date:**

.

# ACKNOWLEDGEMENT

I express my sincere indebtedness towards our guide Prof., “ROHIT AGARWAL SIR” of Computer Science and Information Technology, Meerut Institute of Engineering and Technology, Meerut for his valuable suggestion, guidance and supervision throughout the work. Without his king patronage and guidance the project would not have taken shape. I would also like to express my gratitude and sincere regards for his kind approval of the project. Time to time counseling and advises.

I would also like to thank to our HOD DR. (Prof.) “Swati Sharma Mam”, Department of Computer Science and Information Technology, Meerut Institute of Engineering and Technology, Meerut for her expert advice and counselling from time to time.

I owe sincere thanks to all the faculty members in the department of Computer Science and Information Technology for their kind guidance and encouragement time to time.

Date:18/12/21 Student name: TANISHA AGARWAL

**Table of contents**

**Description page no.**

[Declaration i](#_Toc4453)

[Certificate ii](#_Toc4454)

[Acknowledgement iii](#_Toc4455)

**Chapter 1 Introduction**

**Chapter 2 System Design**

(Work flow/flow chart/DFD/ working principle/ constructional details of individual components)

**Chapter 3 Technology Bucket**

3.1 Description of Data science

**Chapter 4** Outputscreens

**Appendices** Implementation code

**References**

**INTRODUCTION**

Cancer is a disease that occurs when there are changes or mutations that take place in genes that help in cell growth. These mutations allow the cells to divide and multiply in a very uncontrolled and chaotic manner. These cells keep increasing and start making replicas which end up becoming more and more abnormal. These abnormal cells later on form a tumor. Tumors, unlike other cells, don’t die even though the body doesn’t need them.

The cancer that develops in the breast cells is called breast cancer. This type of cancer can be seen in the breast ducts or the lobules. Cancer can also occur in the fatty tissue or the fibrous connective tissue within the breast. These cancer cells become uncontrollable and end up invading other healthy breast tissues and can travel to the lymph nodes under the arms.

There are two types of cancers. Malignant and Benign. Malignant cancers are cancerous. These cells keep dividing uncontrollably and start affecting other cells and tissues in the body. They spread to all other parts of the body and it is hard to cure this type of cancer. Chemotherapy, radiation therapy and immunotherapy are types of treatments that can be given for these types of tumors. Benign cancer is non-cancerous. Unlike malignant, this tumor does not spread to other parts of the body and hence is much less risky that malignant. In many cases, such tumors don’t really require any treatment.

Breast cancer is most commonly diagnosed in women of ages above 40. But this disease can affect men and woman of any age. It can also occur when there’s a family history of breast cancer. Breast Cancer has always had a high mortality rate and according statistics, it alone accounts for about 25% of all new cancer diagnoses and 15% of all cancer deaths among women worldwide. Scientists know about the dangers of it from very early on, and hence there’s been a lot of research put into finding the right treatment for it.

Let's have a quick overview of what we are going to cover.

So the objective and requirements or the goal here is to we have extracted features of breast cancer patient cells and normal person cells. As a Machine learning engineer / Data Scientist has to create an ML model to classify malignant and benign tumor. To complete this ML project we are using the supervised machine learning classifier algorithm.

The development tools we need are Python and python libraries.

The required tool will need a google collab.

**SYSTEM DESIGN**

**DESIGN GOALS**

Under our model, the goal of our project is to create a design to achieve the following:

**ACCURACY**

Only accurate outcomes can help make this model a good one. It can be reliable only when all the outcomes are correct and can be trusted. As this data is required for healthcare purposes, it is important that no errors occur.

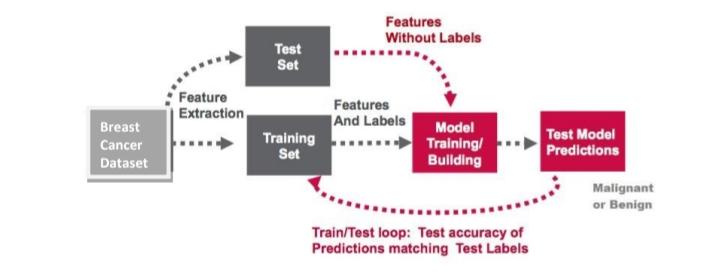
**EFFICIENCY**

The model should be efficient as there is no requirement of manual data entry work or any work by doctors. It takes less time to predict outcomes after all the ML algorithms have been used on the data.

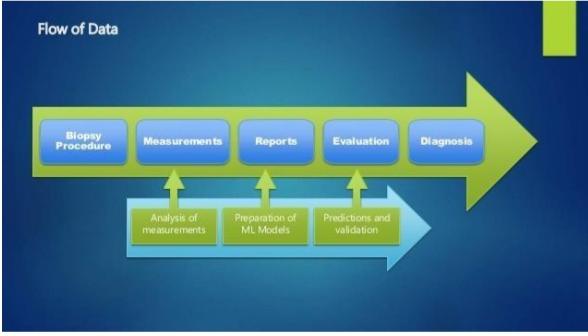
**SYSTEM ARCHITECTURE**

As this project does not have any UI, the architecture is basically the dataset and the features of the dataset. It is trying to understand the dataset and try making the system as simple and easy as possible.

The dataset is first split into training and testing set. The training set if first exposed to the machine learning algorithms so that the system understands what data gives what type of outcome.

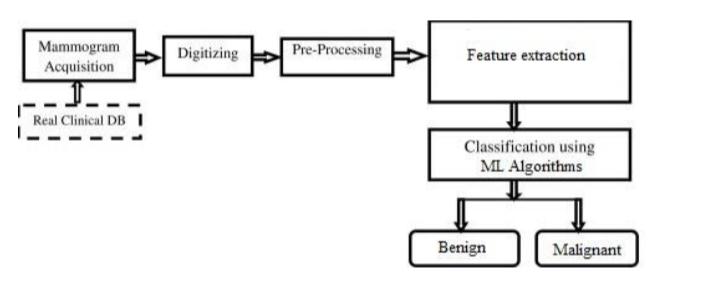
After the system is trained, the testing data is used to test whether the system can correctly predict the class of the data. It checks the percentage accuracy of the model. dada

**DATA FLOW DIAGRAM**



**EXPECTED OUTCOME**

The outcome of this model is to correctly check and predict whether a patient has breast cancer or not. If yes, the model should also be able to tell if the patient has malignant or benign type of cancer.



**TECHNOLOGY BUCKET**

**DATA SCIENCE**

Data science continues to evolve as one of the most promising and in-demand career paths for skilled professionals. Today, successful data professionals understand that they must advance past the traditional skills of analyzing large amounts of data, data mining, and programming skills. In order to uncover useful intelligence for their organizations, data scientists must master the full spectrum of the data science life cycle and possess a level of flexibility and understanding to maximize returns at each phase of the process.

**HISTORY OF DATA SCIENCE**

Data Science has revolutionized several different aspects of our world. Let's take a look then at when and where data science comes from.

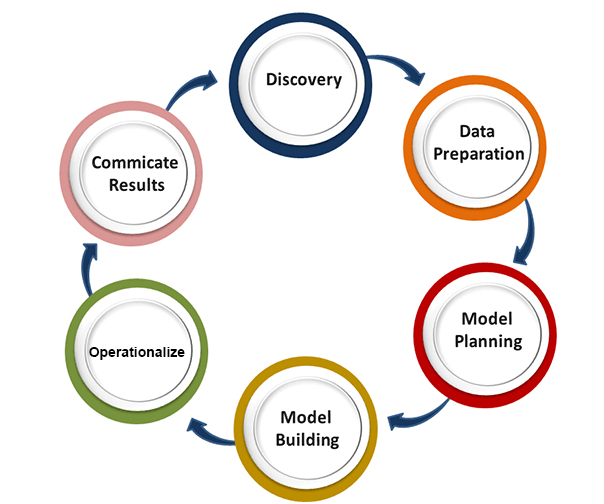
* In 1962, John W. Tukey wrote in “The Future of Data Analysis” - The first milestone in the history of data science is globally recognized for the bright American mathematician John Tukey. The influence of John Tukey in statistical terms is enormous, but the most famous coinage attributed to him is related to computer science. In fact, it should be mentioned that he was the first to introduce the term "bit" as a contraction of "binary digit."
* In 1974, Peter Naur published the Concise Survey of Computer Methods, which surveyed data processing methods across a wide variety of applications. The term “data science” becomes clearer, as he puts his own definition on it: “The science of dealing with data, once they have been established, while the relation of the data to what they represent is delegated to other fields and sciences.”
* In 1977, the International Association for Statistical Computing (IASC) was founded.
* In 1989, Gregory Piatetsky-Shapiro organized and chaired the first Knowledge Discovery in Databases (KDD) workshop.
* In 1994, BusinessWeek published a cover story on “Database Marketing.”
* In 1996, on the occasion of the conference of the International Federation of Classification Societies (IFCS), for the first time, the term “data science” is included in the title of the conference (“Data science, classification, and related methods”). In the same year, Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth publish “From Data Mining to Knowledge Discovery in Databases.”
* In 1997, during his inaugural lecture as the H. C. Carver Chair in Statistics at the University of Michigan, Jeff Wu called for statistics to be renamed “data science” and statisticians to be renamed “data scientists.”

**BENEFITS OF DATA SCIENCE**

* Increases business predictability. ...
* Ensures real-time intelligence. ...
* Favors the marketing and sales area. ...
* Improves data security. ...
* Helps interpret complex data. ...
* Facilitates the decision-making process.

## **Data Science Lifecycle**

The life-cycle of data science is explained as below diagram.



The main phases of data science life cycle are given below:

**1. Discovery:** The first phase is discovery, which involves asking the right questions. When you start any data science project, you need to determine what are the basic requirements, priorities, and project budget. In this phase, we need to determine all the requirements of the project such as the number of people, technology, time, data, an end goal, and then we can frame the business problem on first hypothesis level.

**2. Data preparation:** Data preparation is also known as Data Munging. In this phase, we need to perform the following tasks:

* Data cleaning
* Data Reduction
* Data integration
* Data transformation,

After performing all the above tasks, we can easily use this data for our further processes.

**3. Model Planning:** In this phase, we need to determine the various methods and techniques to establish the relation between input variables. We will apply Exploratory data analytics(EDA) by using various statistical formula and visualization tools to understand the relations between variable and to see what data can inform us. Common tools used for model planning are:

* SQL Analysis Services
* R
* SAS
* Python

**4. Model-building:** In this phase, the process of model building starts. We will create datasets for training and testing purpose. We will apply different techniques such as association, classification, and clustering, to build the model.

Following are some common Model building tools:

* SAS Enterprise Miner
* WEKA
* SPCS Modeler
* MATLAB

**5. Operationalize:** In this phase, we will deliver the final reports of the project, along with briefings, code, and technical documents. This phase provides you a clear overview of complete project performance and other components on a small scale before the full deployment.

**6. Communicate results:** In this phase, we will check if we reach the goal, which we have set on the initial phase. We will communicate the findings and final result with the business team.

**Application**  
**1) Image recognition and speech recognition**

Data science is currently using for Image and speech recognition. When you upload an image on Facebook and start getting the suggestion to tag to your friends. This automatic tagging suggestion uses image recognition algorithm, which is part of data science.  
When you say something using, "Ok Google, Siri, Cortana", etc., and these devices respond as per voice control, so this is possible with speech recognition algorithm.

**2)Gamingworld**  
In the gaming world, the use of Machine learning algorithms is increasing day by day. EA Sports, Sony, Nintendo, are widely using data science for enhancing user experience.

**3)Internet search:**  
When we want to search for something on the internet, then we use different types of search engines such as Google, Yahoo, Bing, Ask, etc. All these search engines use the data science technology to make the search experience better, and you can get a search result with a fraction of seconds.

**4)Transport:**  
Transport industries also using data science technology to create self-driving cars. With self-driving cars, it will be easy to reduce the number of road accidents.

**5)Healthcare:**  
In the healthcare sector, data science is providing lots of benefits. Data science is being used for tumor detection, drug discovery, medical image analysis, virtual medical bots, etc.

**6)Recommendation systems:**  
Most of the companies, such as Amazon, Netflix, Google Play, etc., are using data science technology for making a better user experience with personalized recommendations. Such as, when you search for something on Amazon, and you started getting suggestions for similar products, so this is because of data science technology.

**7)Riskdetection:**  
Finance industries always had an issue of fraud and risk of losses, but with the help of data science, this canbe rescued.  
Most of the finance companies are looking for the data scientist to avoid risk and any type of losses with an increase in customer satisfaction.

**IMPLEMENTATION AND OUTPUT**

**PREPARING THE DATA**

**Step 1:** The first step in the machine learning process is to prepare the data. This includes importing all the packages that will help us organize and visualize the data. The packages used are as follows:

import pandas as pd

import numpy as np

from sklearn.datasets import load\_breast\_cancer

from sklearn.preprocessing import StandardScaler

from keras.models import Sequential

from keras.layers import Dense

**Step 2:** After importing all the necessary packages, we need to load the dataset. We use the

help of Pandas to load the data set.

data = load\_breast\_cancer()

**STEP 3**: Now train ,validate and test the data

#500 Training

X\_train = feature[:500]

y\_train = label[:500]

#35 Validation

X\_val = feature[500:535]

y\_val = label[500:535]

#34 Testing

X\_test = feature[535:]

y\_test = label[535:]

**Step 4**: to check how many data points are Malignant and benign.

for i in range(30):

  sample = X\_test[i]

  sample = np.reshape(sample, (1,30))

  if (model.predict(sample)[0][0] > 0.5):

    print("-Benign")

  else:

    print("-Malignant")

  if (y\_test[i] == 1):

    print("\*Banign")

  else:

    print("\*Melignant")

  print("-----------")

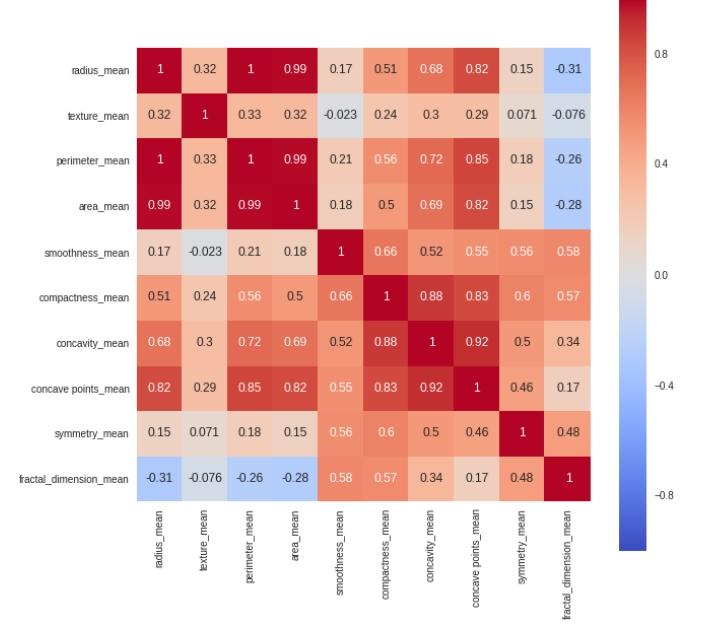
**VISUALIZING THE DATA**

We need to build visualizations of the data in order to decide how to proceed with the machine

learning tools. The Seaborn and the Matplotlib packages will be used for this purpose. We use

the mean values of the features. So first we will have to separate those features in the list to

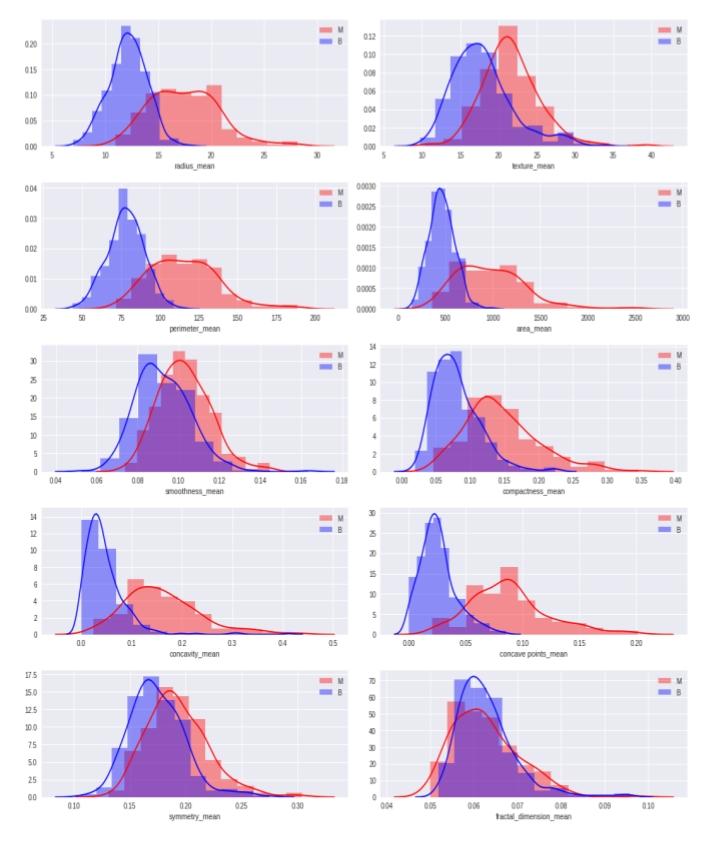
make some work easier and the code more readable



We can also see how the malignant or benign tumors cells can have (or not) different values

for the features plotting the distribution of each type of diagnosis for each of the mean

features.



**Steps to use in Python:**

**Loading the dataset:**

Scikit-learn comes with a few small standard datasets that do not require downloading any

file from any external website. The dataset includes several data about the breast cancer

tumors along with the classification labels, viz., malignant or benign. It can be loaded using

the following function:

The data set has 569 instances or data of 569 tumors and includes data on 30 attributes or

features like the radius, texture, perimeter, area, etc. of a tumor. We will be using these

features to train our model.

**Installing the necessary modules:**

For this machine learning project, we will be needing the ‘Scikit-learn’ Python module by

running the following command in the command prompt:

pip install scikit-learn

**Step #1:** Importing the necessary module and dataset.

We will be needing the ‘Scikit-learn’ module and the Breast cancer wisconsin (diagnostic)

dataset.

**Step #2:** Loading the dataset to a variable.

The important attributes that we must consider from that dataset are ‘target-names'(the

meaning of the labels), ‘target'(the classification labels), ‘feature\_names'(the meaning of

the features) and ‘data'(the data to learn).

**Step #3:** Organizing the data and looking at it

To get a better understanding of what the dataset contains and how we can use the data

to train our model, let us first organize the data and then see what it contains by using

the print() function.

**Step #4:** Organizing the data into Sets.

For testing the accuracy of our classifier, we must test the model on unseen data. So, before

building the model, we will split our data into two sets, viz., training set and test set. We

will be using the training set to train and evaluate the model and then use the trained model

to make predictions on the unseen test set.

The sklearn modlue has a built-in function called the train\_test\_split(), which automatically

divides the data into these sets. We will be using this function two split the data.

The train\_test\_split() function randomly splits the data using the parameter test\_size.

What we have done here is that, we have split 33% of the original data into test data (test).

The remaining data (train) is the training data. Also, we have respective labels for both the

train variables and test variables, i.e. train\_labels and test\_labels.

**Step #5**: Building the Model.

There are many machine learning models to choose from. All of them have their own

advantages and disadvantages. For this model, we will be using the Naive Bayes algorithm

that usually performs well in binary classification tasks. Firstly, import

the GaussianNB module and initialize it using the GaussianNB() function. Then train the

model by fitting it to the data in the dataset using the fit() method.

**Step #6:** Evaluating the trained model’s accuracy.

As we have predicted values now, we can evaluate our model’s accuracy by comparing it

with the actual labels of the test set, i.e., comparing predictions with test\_labels. For this

purpose, we will be using the built-in accuracy\_score() function in the sklearn module

Epoch 1/10

100/100 [==============================] - 1s 5ms/step - loss: 0.2430 - accuracy: 0.9180 - val\_loss: 0.0521 - val\_accuracy: 1.0000

Epoch 2/10

100/100 [==============================] - 0s 2ms/step - loss: 0.0728 - accuracy: 0.9780 - val\_loss: 0.0045 - val\_accuracy: 1.0000

Epoch 3/10

100/100 [==============================] - 0s 2ms/step - loss: 0.0560 - accuracy: 0.9800 - val\_loss: 0.0063 - val\_accuracy: 1.0000

Epoch 4/10

100/100 [==============================] - 0s 2ms/step - loss: 0.0306 - accuracy: 0.9880 - val\_loss: 0.0842 - val\_accuracy: 0.9714

Epoch 5/10

100/100 [==============================] - 0s 2ms/step - loss: 0.0352 - accuracy: 0.9860 - val\_loss: 0.0200 - val\_accuracy: 1.0000

Epoch 6/10

100/100 [==============================] - 0s 2ms/step - loss: 0.0218 - accuracy: 0.9900 - val\_loss: 0.0277 - val\_accuracy: 0.9714

Epoch 7/10

100/100 [==============================] - 0s 2ms/step - loss: 0.0241 - accuracy: 0.9900 - val\_loss: 0.0384 - val\_accuracy: 0.9714

Epoch 8/10

100/100 [==============================] - 0s 2ms/step - loss: 0.0127 - accuracy: 0.9980 - val\_loss: 0.0478 - val\_accuracy: 0.9714

Epoch 9/10

100/100 [==============================] - 0s 2ms/step - loss: 0.0055 - accuracy: 0.9980 - val\_loss: 0.0029 - val\_accuracy: 1.0000

Epoch 10/10

100/100 [==============================] - 0s 2ms/step - loss: 0.0032 - accuracy: 1.0000 - val\_loss: 0.1426 - val\_accuracy: 0.9714

<keras.callbacks.History at 0x7f73f335b350>