MODEL FOR CANCER PREDICTION

On

MINOR PROJECT REPORT

SUBMITTED IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE AWARD OF THE DEGREE OF

BACHELOR OF TECHNOLOGY

(COMPUTER SCIENCE AND ENGINEERING)

Submitted By:

Sumit Kumar (1905828)

Digvijay Kumar (1905782)

Vishwanathan Anand (1905833)

Submitted To:

Prof. Harkomalpreet Kaur

Department of Computer Science and Engineering

Guru Nanak Dev Engineering College

Ludhiana-141006

Abstract

Women are seriously threatened by breast cancer with high morbidity and mortality. The lack of robust prognosis models results in difficulty for doctors to prepare a treatment plan that may prolong patient survival time. Hence, the requirement of time is to develop the technique which gives minimum error to increase accuracy. Four algorithm SVM, Logistic Regression, Random Forest and KNN which predict the breast cancer outcome have been compared in the paper using different datasets. All experiments are executed within a simulation environment and conducted in JUPYTER platform. Aim of research categorises in three domains. First domain is prediction of cancer before diagnosis, second domain is prediction of diagnosis and treatment and third domain focuses on outcome during treatment. The proposed work can be used to predict the outcome of different technique and suitable technique can be used depending upon requirement. This research is carried out to predict the accuracy. The future research can be carried out to predict the other different parameters and breast cancer research can be categorises on basis of other parameters.

ACKNOWLEDGEMENT

We are highly grateful to **Dr Sehijpal Singh, Principal, Guru Nanak Dev Engineering College (GNDEC)**, Ludhiana, for providing this opportunity to carry out the minor project work at making 'REAL TIME HUMAN AND OBJECT DETECTION' .

The constant guidance and encouragement received from **Dr. Parminder Singh H.O.D. CSE department, GNDEC** Ludhiana has been of great help in carrying out the project work and is acknowledged with reverential thanks.

We would like to express a deep sense of gratitude and thanks profusely to **our teachers**, without them wise counsel and able guidance, it would have been impossible to complete the project in this manner.

We express gratitude to other faculty members of the computer science and engineering department of GNDEC for their intellectual support throughout the course of this work.

Finally, We are indebted to all whosoever have contributed in this report work.

SUMIT KUMAR

DIGVIJAY KUMAR

VISHWANATHAN ANAND

LIST OF FIGURES

LIST OF TABLES

TABLE OF CONTENTS

Chapter 1

INTRODUCTION

1.1 Introduction to project:

Breast cancer (BC) 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.

In 2020, there were 2.3 million women diagnosed with breast cancer and 685 000 deaths globally. As of the end of 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.

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.

1.2 Project Category

Our project comes under category of application of Machine Learning

1.3 OBJECTIVE

1 Model to predict breast cancer from dataset.

2 To check accuracy of different ML models using 80% data as training and 20% data as testing.

Chapter 2

Feasibilitty Analysis

1.Economical - From the economical point of view, the project is quite feasible because it involves basic and open-source software. Thus, the cost of development is almost negligible as no high-end hardware is required.

2.Operational - Operationally this project requires coding to in python programming to detect breast cancer prediction in human.

3.Technical – The proposed project can be implementable at zero cost and run with basic software and hardware requirements. It can run on a machine with an operating system at least Single Core 1.0 GHz, 64MB Graphics Card, 2Gb of RAM, and a disk space of 2GB.

Chapter 3

SYSTEM REQUIRMENT

3.1Hardware Requirments:

Hardware	Minimum Requirements
Any Computing Device	2 GHz minimum, multi-core processor
Disk Space	Atleast 2 GB
Memory(RAM)	Atleast 4 GB, preferably higher

3.2 Software Requirments:

Type	Software Required
Operating System	Windows 10
Coding Language	Python
Libraries	Numpy, pandas, matplotlib,Seaborn,sklearn

3.3 SOFTWARE REQUIRMENT ANALYSIS

VS Code:

Visual studio code , also known as VS code is a source code editor made by microsoft for windows,linux,macOS.Features include support of debugging ,syntax highlighting,intelligent code completion ,snippets, code refactoring and embedded giT.

Google Colab:

Colab notebooks are Jupyter notebooks that run in the cloud and are highly integrated with Google drive, making them easy to set up, access,

1) PANDAS

It is mainly used for data analysis. Pandas allows importing data from various file formats such as comma-separated values, JSON, SQL, Microsoft Excel. Pandas allows various data manipulation operations such as merging, reshaping, selecting, as well as data cleaning, and data wrangling features.

2) NUMPY

Numpy is a library for the Python programming language, adding support for large, multidimensional arrays and matrices, along with a large collection of high-level mathematical functions to operate on these arrays.

3) SEABORN

Seaborn is a data visualization library built on top of matplotlib and closely integrated with pandas data structures in Python. Visualization is the central part of Seaborn which helps in exploration and understanding of data.

4) MATPLOTLIB

Matplotlib is a library in Python and it is numerical – mathematical extension for NumPy library.pyplot is a state-based interface to a Matplotlib module which provides a MATLAB-like interface. There are various plots which can be used in Pyplot are Line Plot, Contour, Histogram, Scatter, 3D Plot, etc.

5) SKLEARN

Scikit-Learn, also known as sklearn is a python library to implement machine learning models and statistical modelling. Through scikit-learn, we can implement various machine learning models for regression, classification, clustering, and statistical tools for analyzing these models.

6) HEROKU APP

Heroku is a container-based cloud Platform as a Service (PaaS). Developers use Heroku to deploy, manage, and scale modern apps. Our platform is elegant, flexible, and easy to use, offering developers the simplest path to getting their apps to market.

3.4 TECHNOLOGIES USED

1)Data Analysis:

Data Analysis involves extraction, cleaning, transformation, modeling and visualization of data with an objective to extract important and helpful information.

2)Data Mining:

Data mining could be called as a subset of Data Analysis. It is the exploration and analysis of huge knowledge to find important patterns and rules Data mining could also be a systematic and successive method of identifying and discovering hidden patterns and data throughout a big dataset. Moreover, it is used to build machine learning models that are further used in artificial intelligence.

3)Machine Learning:

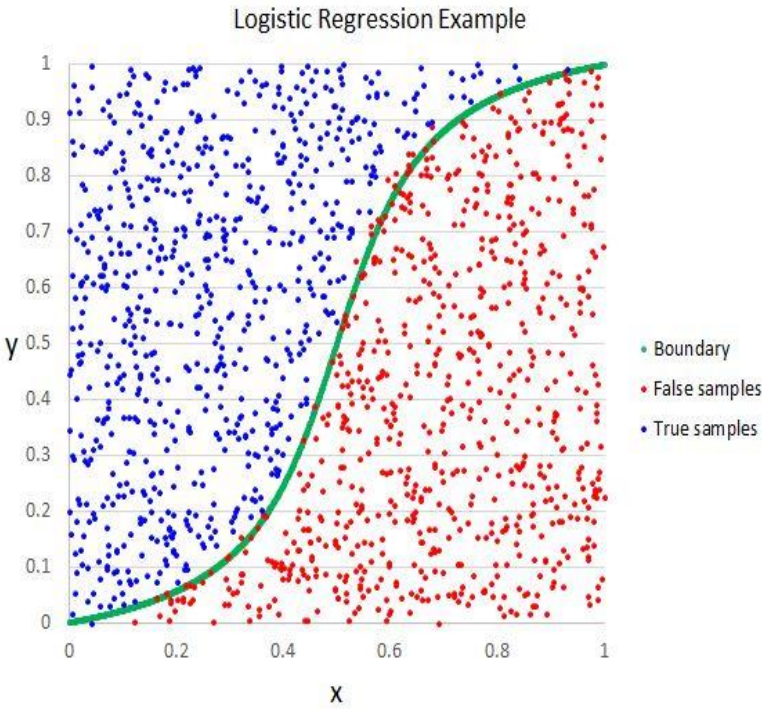
Machine learning is an emerging subdivision of artificial intelligence. Its primary focus is to design systems, allow them to learn and make predictions based on the experience. It trains machine learning algorithms using a training dataset to create a model. The model uses the new input data to predict fake news. Using machine learning, it detects hidden patterns in the input dataset to build models. It makes accurate predictions for new datasets. The dataset is cleaned and missing values are filled. The model uses the new input data to predict fake news and then tested for accuracy.

MACHINE LEARNING MODELS:-

(a).Logistic Regression:

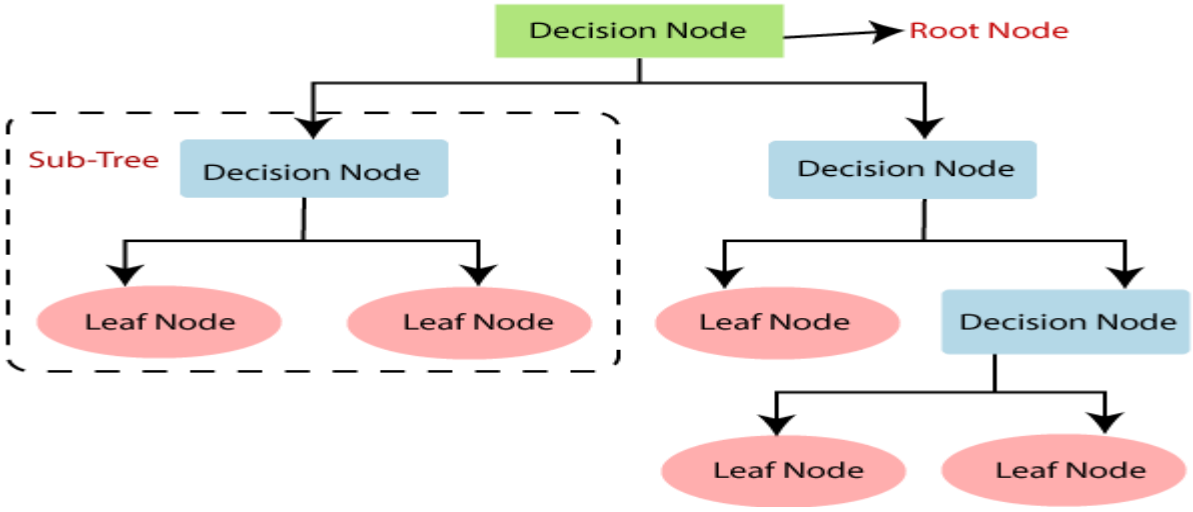
Logistic regression is a statistical method that is used for building machine learning models where the dependent variable is dichotomous: i.e. binary. Logistic regression is used to describe data and the relationship between one dependent variable and one or more independent variables. The independent variables can be nominal, ordinal, or of interval type. The name “logistic regression” is derived from the concept of the logistic function that it uses. The logistic function is also known as the sigmoid function. The

The name “logistic regression” is derived from the concept of the logistic function that it uses. The logistic function is also known as the sigmoid function. The value of this logistic function lies between zero.



(b)Decision tree:-

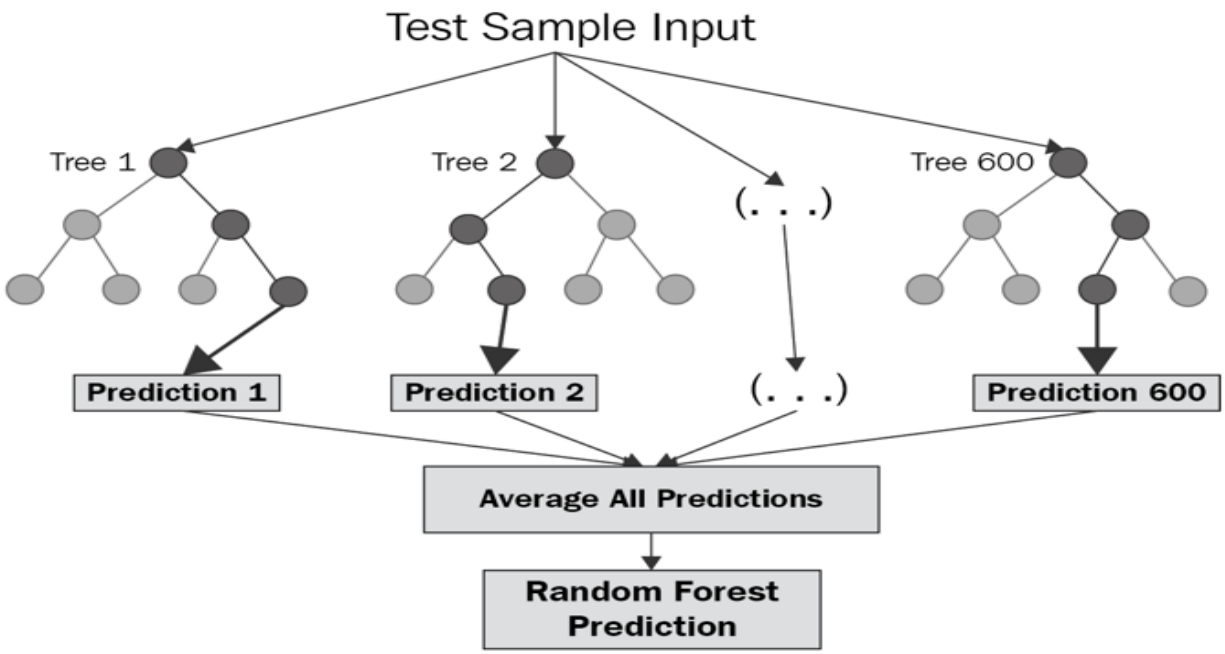
decision Trees (DTs) are a non-parametric supervised learning method used for classification and regression. The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features. A tree can be seen as a piecewise constant approximation.



(c)Random Forest:-

"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."

It is based on the concept of ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model.



4)Python

Python is an interpreted, object-oriented, high-level programming language with dynamic semantics. Its high-level built-in data structures, combined with dynamic typing and dynamic binding, make it very attractive for Rapid Application Development. Python's simple, easy to learn syntax emphasizes readability and therefore reduces the cost of program maintenance. Python supports modules and packages, which encourages program modularity and code reuse.

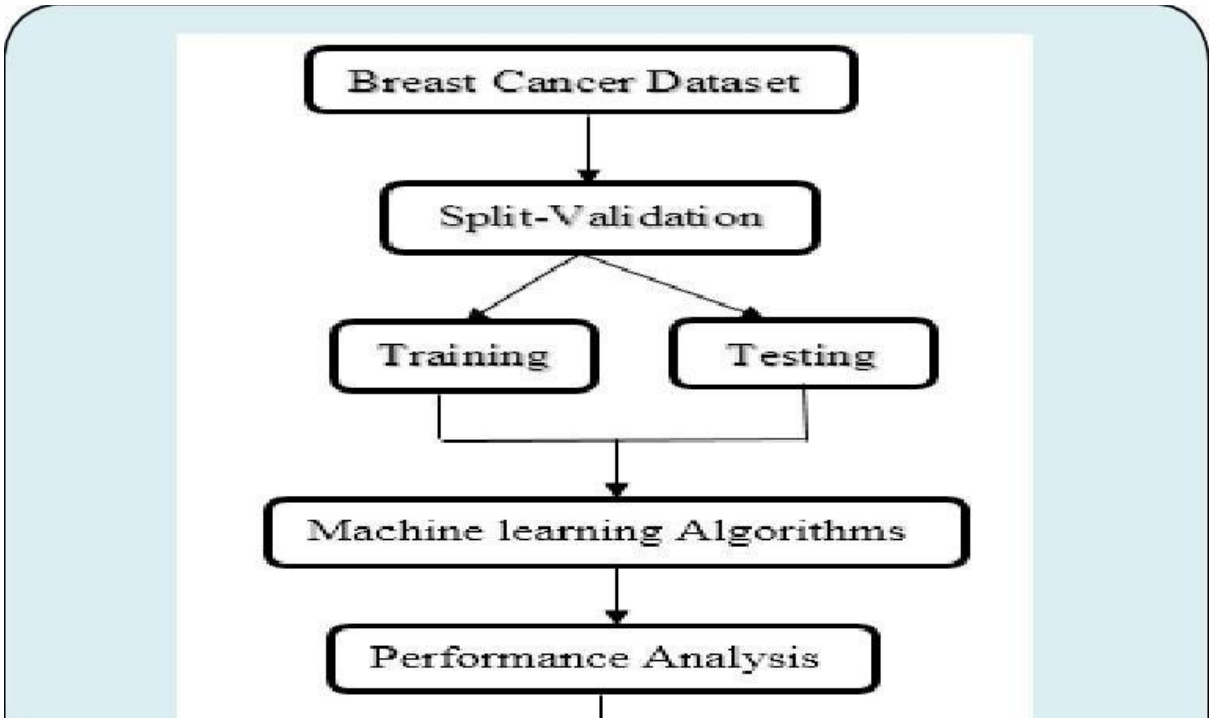
The Python interpreter and the extensive standard library are available in source or binary form without charge for all major platforms, and can be freely distributed.

5)Flask

Flask isa Python-based microframework used for developing small scale websites. Flask is very easy to make Restful API's using python. As of now, we have develop a model i.e model. pkl which can predict a class of the data based on a various attribute of the data.

SYSTEM DESIGN

4.1 FLOWCHART



4.2 PROCESS OF IMPLIMENTATION

1.Data collection

Our first task is to collect data.We will use UCI MACHINE LEARNING RESPOSITORY breast cancer dataset.(University Of Wisconsin Hospital at Madison, Wisconsin, USA).it has 569 rows and 32 coloumns.Then we categories data.

2.Data prepration

a)Data exploration:

we use jupyter notebook for this.then we import different librairies like pandas,numpy to prepare our data.

b)Categorical data:

for diagnosis column where we have two value(M=malignant,B=benign)we changes them to (M=1,B=0),using label encoder from sklearn.

c)Spiliting of data

we split our dataset using SciKit-Learn library in Python using the train_test_split method.

80% Training data and 20%test data.

3.MODEL SELECTION

We use different ML models like-Logistic Regression,Nearest Neighbour,Support Vector Machine,Naives Bayes,Decision Tree,Random Forest Classification.Then we check accuracy of each of them and then we select our model for final prediction which has best acuuracy.

4.TRAINING

AT the heart of the machine learning process is the training of the model.Bulk of the “learning” is done at this stage. If we view our model in mathematical terms, the inputs i.e., our 2 features would have coefficients. These coefficients are called the weights of features. There would also be a constant or y-intercept involved. This is referred to as the bias of the model. The process of determining their values is of trial and error. Initially, we pick random values for them and provide inputs. The achieved output is compared with actual output and the difference is minimized by trying different values of weights and biases. The iterations are repeated using different entries from our training data set until the model reaches the desired level of accuracy.

5.EVALUATION

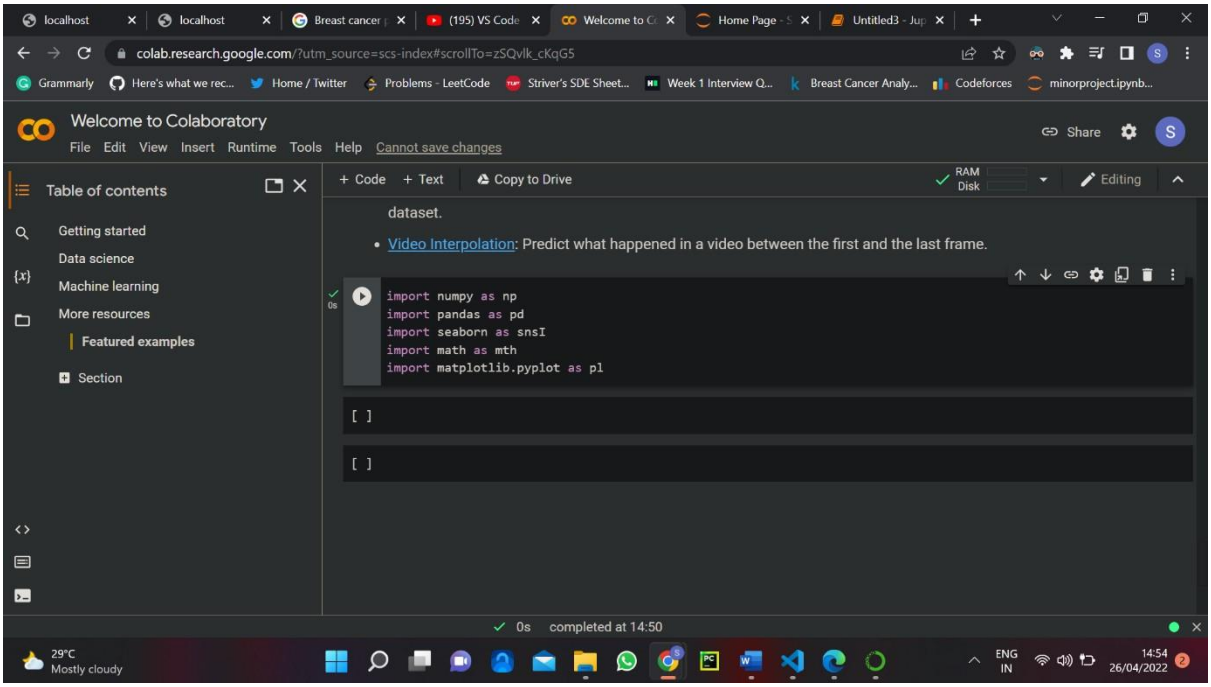
With the model trained, it needs to be tested to see if it would operate well in real world situations. That is why the part of the data set created for evaluation is used to check the model’s proficiency.In this ,we use five performance measures to evaluate all the classifiers: true positive, false positive, ROC curve, standard deviation (Std) and accuracy (AC).

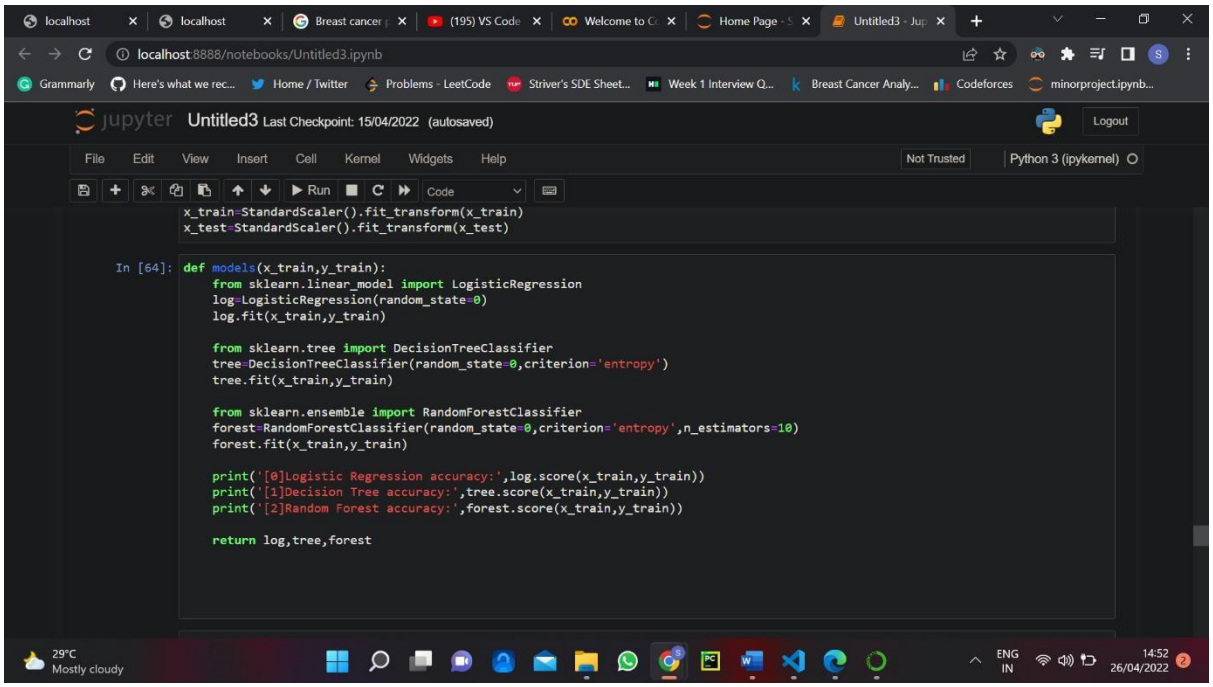
AC=(TP+TN)/(TP+TN+FP+FN).AC=(TP+TN)/(TP+TN+FP+FN).

Where TP, TN, FP and FN denote true positive, true negative, false positive and false negative, respectively.

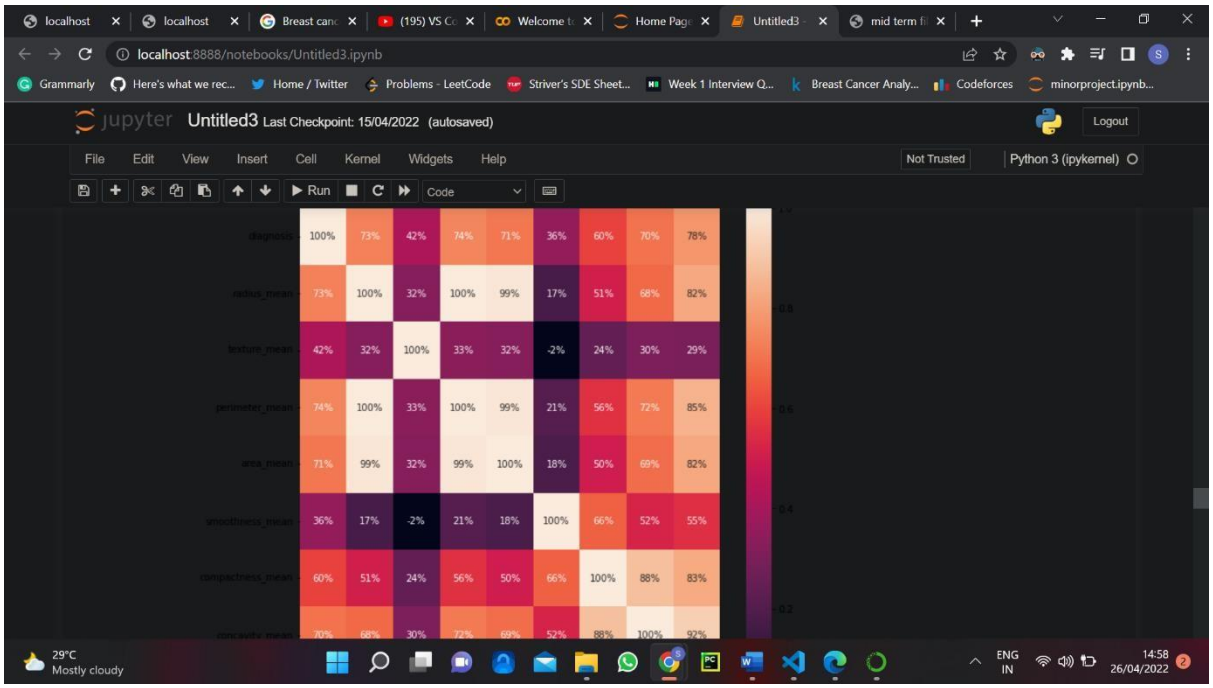
Chapter 5

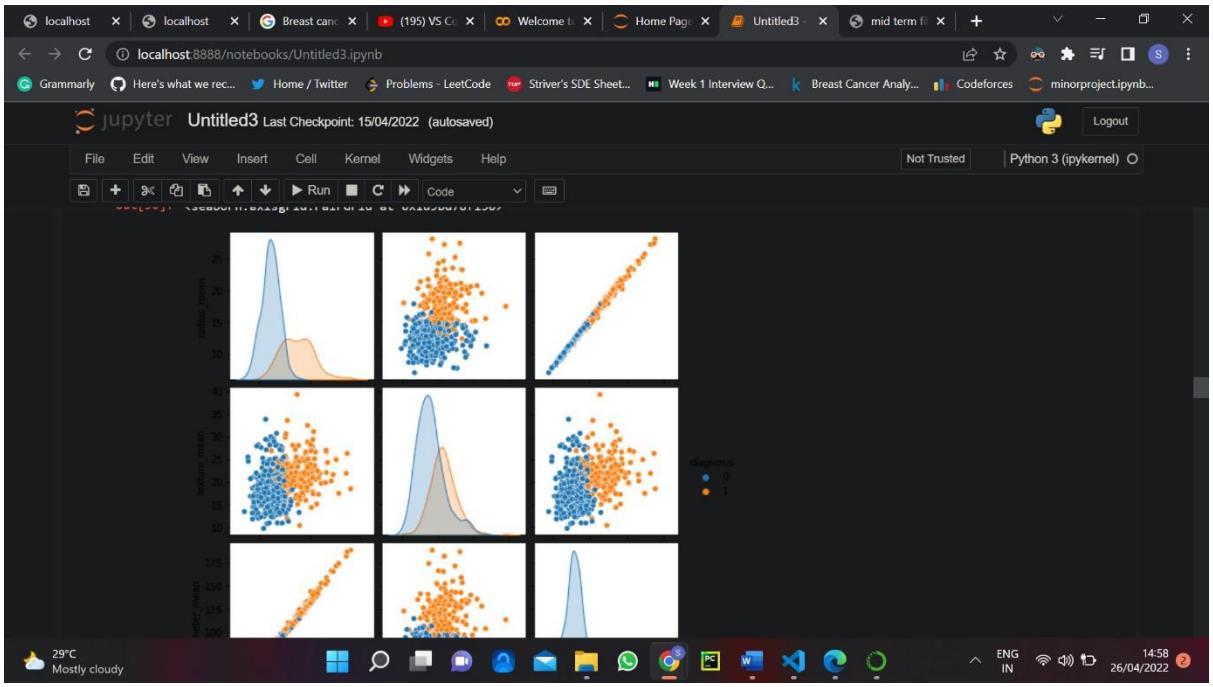
5.1 SCREENSHOTS OF IMPLEMENTATION

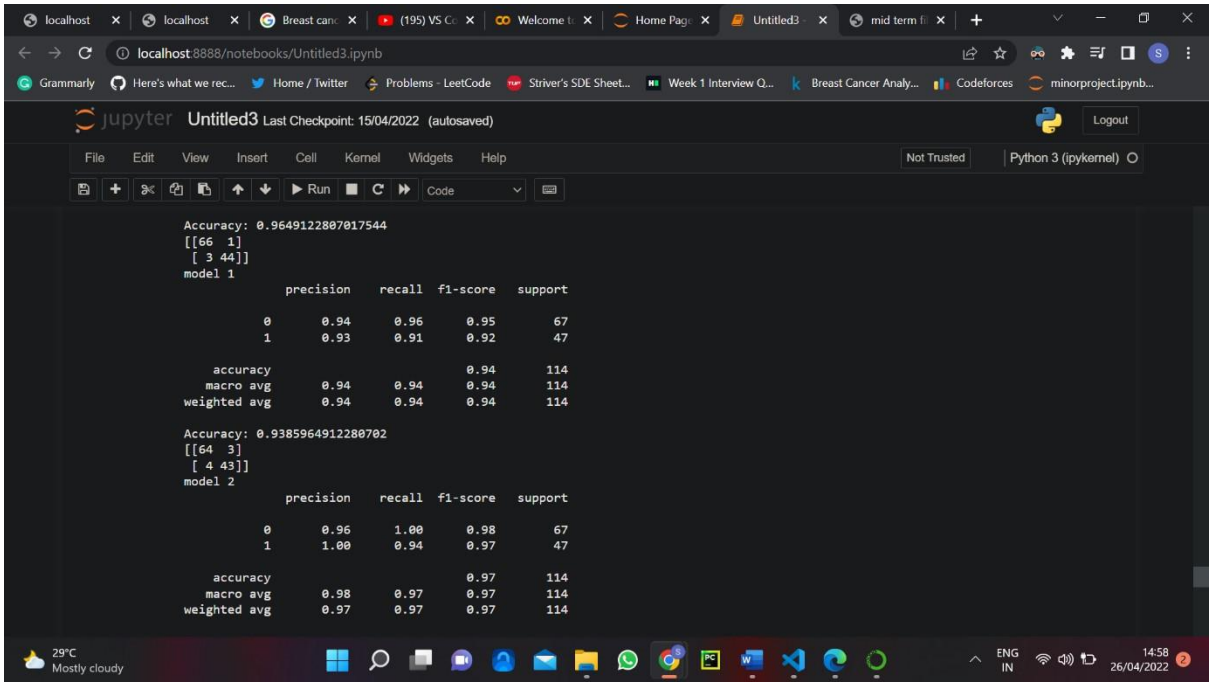


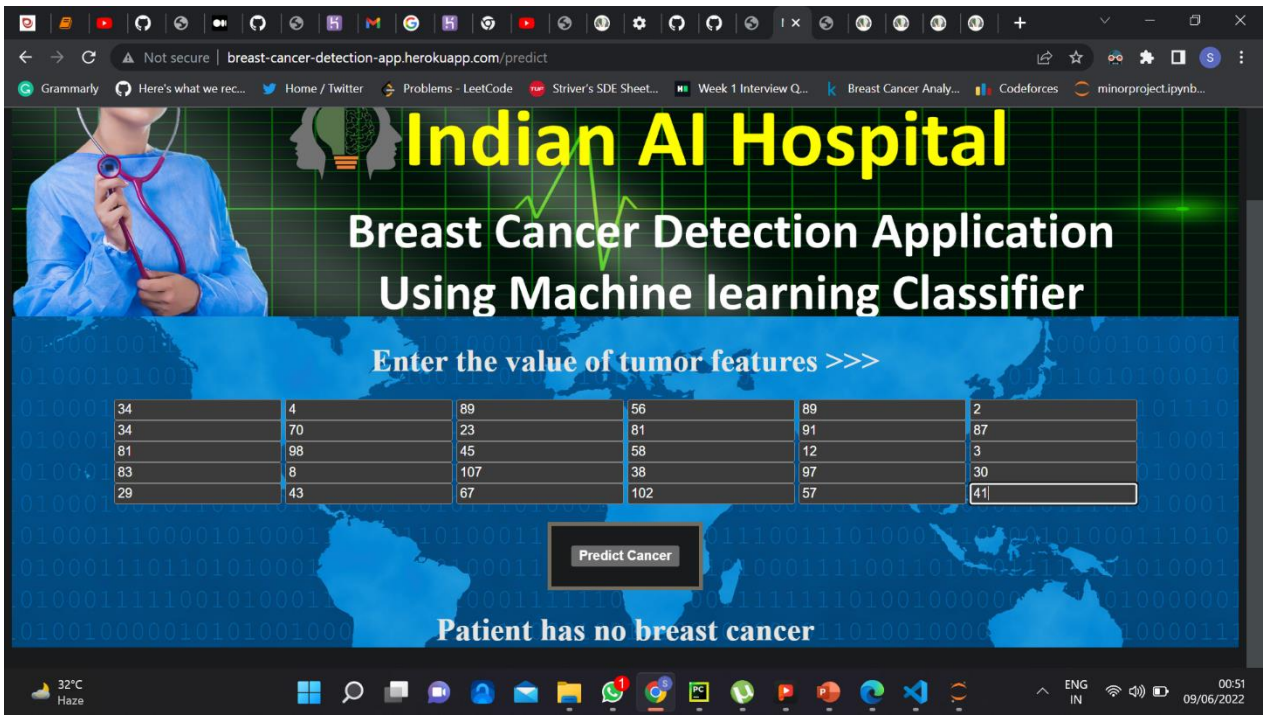


5.2 SCREENSHOTS OF RESULT






localhost:8888/notebooks/Untitled3.ipynb
jupyter Untitled3 Last Checkpoint: 15/04/2022 (autosaved)
Logout
File Edit View Insert Cell Kernel Widgets Help
Not Trusted
Python 3 (ipykernel)
Run
Code
Accuracy: 0.9649122807017544
[[66 1]
[3 44]]
model 1
precision recall f1-score support
0 0.94 0.96 0.95 67
1 0.93 0.91 0.92 47
accuracy 0.94 114
macro avg 0.94 0.94 0.94 114
weighted avg 0.94 0.94 0.94 114
Accuracy: 0.9385964912280702
[[64 3]
[4 43]]
model 2
precision recall f1-score support
0 0.96 1.00 0.98 67
1 1.00 0.94 0.97 47
accuracy 0.97 114
macro avg 0.98 0.97 0.97 114
weighted avg 0.97 0.97 0.97 114
29°C
Mostly cloudy
ENG
IN
14:58
26/04/2022


Not secure | breast-cancer-detection-app.herokuapp.com/predict
Indian AI Hospital
Breast Cancer Detection Application
Using Machine learning Classifier
Enter the value of tumor features >>>
34 4 89 56 89 2
34 70 23 81 91 87
81 98 45 58 12 3
83 8 107 38 97 30
29 43 67 102 57 41
Predict Cancer
Patient has no breast cancer
32°C
Haze
ENG
IN
00:51
09/06/2022

Chapter 6

6.1 Conclusion

In this review, we discussed the concepts of ML while we outlined their application in cancer prediction/prognosis. Most of the studies that have been proposed the last years and focus on the development of predictive models using supervised ML methods and classification algorithms aiming to predict valid disease outcomes. Based on the analysis of their results, it is evident that the integration of multidimensional heterogeneous data, combined with the application of different techniques for feature selection and classification can provide promising tools for inference in the cancer domain.

6.2 Future Scope

Machine Learning models are getting better than pathologists at accurately predicting the development of cancer.

Every year, Pathologists diagnose 14 million new patients with cancer around the world. That's millions of people who'll face years of uncertainty.

Pathologists have been performing cancer diagnoses and prognoses for decades. Most pathologists have a 96–98% success rate for diagnosing cancer. They're pretty good at that part.

The problem comes in the next part. According to the Oslo University Hospital, the accuracy of prognoses is only 60% for pathologists. A prognosis is the part of a biopsy that comes after cancer has been diagnosed, it is predicting the development of the disease.

REFRENCE:-

• Source code from github.

• https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8442074/ (RESEARCH PAPER)

•https://www.kaggle.com/code/nitin7060/breast-cancer-detection

• https://youtu.be/HXnDyrraRb0.

.