

<https://www.overleaf.com/project/631ee3913cec3013af9ba12e>

Summer Project On

Title

By

Chetan Dhandge(2021510010)

Utkarsh Doras (2021510013)

Under the guidance of

Internal Supervisor

Prof. Pooja Raundale



Department of Master Of Computer Application
Sardar Patel Institute of Technology
Autonomous Institute Affiliated to Mumbai University
2022-23

CERTIFICATE OF APPROVAL

This is to certify that the following students

Chetan Dhandge(2021510010)
Utkarsh Doras (2021510013)

Have satisfactorily carried out work on the project
entitled

**“Breast Cancer Prediction Using
Machine Learning”**

Towards the fulfilment of project, as laid down
by
Sardar Patel Institute of Technology
during year
2022-23.

Project Guide:
Pooja Raundale

PROJECT APPROVAL CERTIFICATE

This is to certify that the following students

Chetan Dhandge(2021510010)
Utkarsh Doras (2021510013)

Have successfully completed the Project report on

“Breast Cancer Prediction Using Machine Learning”,

which is found to be satisfactory and is approved

at

**SARDAR PATEL INSTITUTE OF TECHNOLOGY,
ANDHERI (W), MUMBAI**

INTERNAL EXAMINER

EXTERNAL EXAMINER

HEAD OF DEPARTMENT

PRINCIPAL

Contents

Abstract	i
Objectives	i
List Of Figures	ii
List Of Tables	ii
1 Introduction	1
1.1 Problem Definition	1
1.2 Objectives and Scope	1
1.2.1 Objectives	1
1.3 1. INTRODUCTION	2
1.4 RELATED WORK	2
1.5 DATASET AND FEATURES	2
1.6 Data Preprocessing	2
1.7 METHODS	3
1.8 Feature selection	3
1.9 Results:-	3
1.10 System Requirements	5
2 Software Requirement Specification (SRS) and Design	6
2.1 Purpose	6
2.2 Definition	6
2.3 Overall Description	6
2.3.1 Product Functions	6
3 Project Analysis and Design	7
3.1 Methodologies Adapted	7
3.2 Modules	8
3.2.1 Activity diagram	8
3.2.2 Work Breakdown Structure	9
3.2.3 PERT Chart	10
3.2.4 Gantt Chart	10
4 Project Implementation and Testing	14
4.1 Input Fields	14
4.2 Code 1	15
4.3 Code 2	15
4.4 Code 3	16
5 Conclusion	17
6 Bibliography	18
6.1 Web References	18

Breast Cancer Prediction Using Machine Learning

Abstract

Breast cancer is one of the most common cancer and is causing a huge number of deaths in women. The high incidence and mortality of breast cancer is due to its considerably low accuracy of diagnosis.

In this paper, we explore machine learning models that can be applied to help increasing the accuracy of the diagnosis of breast cancer. The main problem of the project is to detect breast cancer based on a set of features calculated from a digitized image of the Fine Needle Aspiration (FNA) of a breast mass from a patient. We present a diagnosis model using both traditional and deep learning machine learning models. Classic machine learning models including Logistic Regression, Decision Tree, Random Forest, etc. are tested on the Breast Cancer Wisconsin dataset. Additionally, we checked which algorithm gives the most optimal prediction. This paper demonstrates that machine learning models can be used for an automatic diagnosis for breast cancer.

Objectives

- The objective of this report is to train machine learning models to predict whether a breast cancer cell is Benign or Malignant. Data will be transformed and its dimension reduced to reveal patterns in the dataset and create a more robust analysis.
- As previously said, the optimal model will be selected following the resulting accuracy, sensitivity, and f1 score, amongst other factors. We will later define these metrics.
- We can use machine learning method to extract the features of cancer cell nuclei image and classify them. It would be helpful to determine whether a given sample appears to be Benign ("B") or Malignant ("M").
- The machine learning models that we will applicate in this report try to create a classifier that provides a high accuracy level combined with a low rate of false-negatives (high sensitivity).

List of Figures

3.1.1Diagrammatic Representation of Waterfall Model	7
3.2.1Activity Diagram	8
3.2.2Work Breakdown Structure	9
3.2.3PERT Chart	11
3.2.4Gantt Chart	12
4.1.1 Login and Register	14

List of Tables

1.5.1 Hardware Requirements on Server Side	5
1.5.2 Hardware Requirements on Client Side	5
1.5.3 Software Requirements on Server Side	5
1.5.3 Software Requirements on Client Side	5
4.2.1 Use Case Table - Register	13

1 Introduction

1.1 Problem Definition

To eliminate redundancy in collecting data in current physical training and placement process and to squash the time gap and miscommunication in the current process.

1.2 Objectives and Scope

1.2.1 Objectives

- The objective of this report is to train machine learning models to predict whether a breast cancer cell is Benign or Malignant. Data will be transformed and its dimension reduced to reveal patterns in the dataset and create a more robust analysis.
- As previously said, the optimal model will be selected following the resulting accuracy, sensitivity, and f1 score, amongst other factors. We will later define these metrics.
- We can use machine learning method to extract the features of cancer cell nuclei image and classify them. It would be helpful to determine whether a given sample appears to be Benign ("B") or Malignant ("M").
- The machine learning models that we will applicate in this report try to create a classifier that provides a high accuracy level combined with a low rate of false-negatives (high sensitivity).

Breast Cancer Prediction Using Machine Learning

1.3 1. INTRODUCTION

Breast cancer is one of the most common cancer in women and the second leading cause of women's cancer death. Despite the lack of effective treatment, the low accuracy of diagnosis is also a major cause of the high incidence and mortality of breast cancer. Mammography is a traditional method used for diagnosing breast cancer. According to UCHealth's report, only 78

1.4 RELATED WORK

There have been many studies applying different machine learning techniques on medical analysis. In terms of traditional machine learning methods, Chaurasia et al. used Simple Logistic to reduce the dimension of feature space and applied RepTree and RBF Network to evaluate the performance. Dubey et al. used K-means algorithm to evaluate the impact of clustering using centroid initialization and achieved 92

1.5 DATASET AND FEATURES

The Wisconsin Breast Cancer (Diagnostic) dataset has been extracted from the UCI Machine Learning Repository. 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. Class distribution: 357 benign, 212 malignant

Number of instances: 569; Number of attributes: 32

Attributes: Ten real-valued features are computed for each cell nucleus a) radius (mean of distances from center to points on the perimeter) b) texture (standard deviation of gray-scale values) c) perimeter d) area e) smoothness (local variation in radius lengths) f) compactness ($\text{perimeter}^2 / \text{area} - 1.0$) g) concavity (severity of concave portions of the contour) h) concave points (number of concave portions of the contour) i) symmetry j) fractal dimension ("coastline approximation" $- 1$)

1.6 Data Preprocessing

We first cleaned the dataset by removing samples with empty values. There are 683 samples after removing invalid samples. We then realized that the dataset with 683 samples is rather small. To enhance the dataset, we generated a new dataset by copying original dataset and add Gaussian noise to it. Afterward, we appended the generated dataset to the original dataset. This process doubles the dataset to 1398 samples. Then we rescaled data to $[0,1]$ using MinMaxScaler. By plotting scatter matrix of the features, we realized the data is significantly right skewed, which may make the model biased towards the majority of the features. Thus, we took the square root of the feature data to mitigate the data skew problem.

Breast Cancer Prediction Using Machine Learning

1.7 METHODS

We used 3 traditional models for the classification of breast cancer cases. Feature selection is applied to increase the rate of accurate prediction. Finally, we compare the performance of all the models applied and choose the one with the highest performance.

(1) Logistic Regression (LR):

Logistic regression predicts the probability of the default class (e.g. Class 2 in this case) and transforms the probability into a binary value (0 or 1) for classification using "sigmoid" function as shown in Equation 1.

$$f(x) = 1/(1 + \exp(-x)) \dots \dots \dots (1).$$

(2) Decision Tree (DT):

The decision tree are presented with a tree structure. The test objects are classified by their feature values. A node in a decision tree represents an instance, outcomes of the test represented by branch, and the leaf node epitomized the class label.

(3) Random Forest (RF):

Random forest is a set of individual decision trees. Each decision tree spits out a class prediction. It decides the class of the test object by aggregating the votes from different decision trees.

1.8 Feature selection

In many machine learning algorithms, there is a decrease of accuracy when the number of features is redundant [16]. In order to improve the accuracy of the models and avoid overfitting, we performed feature selection on the data. For the traditional models, we used two techniques to select features from the dataset. For the Decision Tree and Random Forest model, we generate the feature importance of the last training result and choose features accordingly. Given the importance of the j th feature I_j , we drop two features that has minimum feature importance: $\hat{j}_{drop} = \arg\min_j (I_j)$. For the remaining five models, we used the correlation matrix with heatmap visualization. Correlation represents how the features in the dataset are related to each other. By using the heatmap visualization, it is easier to identify which features are highly correlated. Using the seaborn library, we could plot the heatmap for better view. For each group of highly correlated features, we choose only one feature to represent all that are in the group. This way most of the information in the features is reserved, and the redundant information is dropped to avoid overfitting.

1.9 Results:-

Model 0 (Logistic Regression)

precision recall f1-score support

0 0.96 0.99 0.97 67 1 0.98 0.94 0.96 47

accuracy 0.96 114 macro avg 0.97 0.96 0.96 114 weighted avg 0.97 0.96 0.96

114

Breast Cancer Prediction Using Machine Learning

Accuracy : 0.9649122807017544 Model 1 (Decision Tree) 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 Model 2 (Random Forest) 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

Accuracy : 0.9736842105263158

Breast Cancer Prediction Using Machine Learning

1.10 System Requirements

- Hardware Requirements on Server Side

Table 1.5.1: Hardware Requirements on Server Side

Processor	Dual Core Processor or Above
RAM	Minimum 4 GB RAM
Storage	Minimum 10 GB Hard Disk Space for smooth run

- Hardware Requirements on Client Side

Table 1.5.2: Hardware Requirements on Client Side

Device	Android Device with Touch Screen minimum 5" inch Display
Processor	Dual Core Processor or Above
RAM	Minimum 2 GB RAM
Storage	Minimum 250 MB Storage Space

- Software Requirements on Server Side

Table 1.5.3: Software Requirements on Server Side

Operating System	OS Independent
Browser	Google Chrome 5.0.0 or equivalent

- Software Requirements on Client Side

Table 1.5.3: Software Requirements on Client Side

Operating System	Android/IOS Smartphone
Server	Not Required

2 Software Requirement Specification (SRS) and Design

2.1 Purpose

The purpose of our project is to develop an UI application that can help user (Doctors) to predict cancer

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

2.2 Definition

To build a Breast Cancer Prediction web-app so the users (doctors) can predict cancer easily.

2.3 Overall Description

2.3.1 Product Functions

The product function includes:

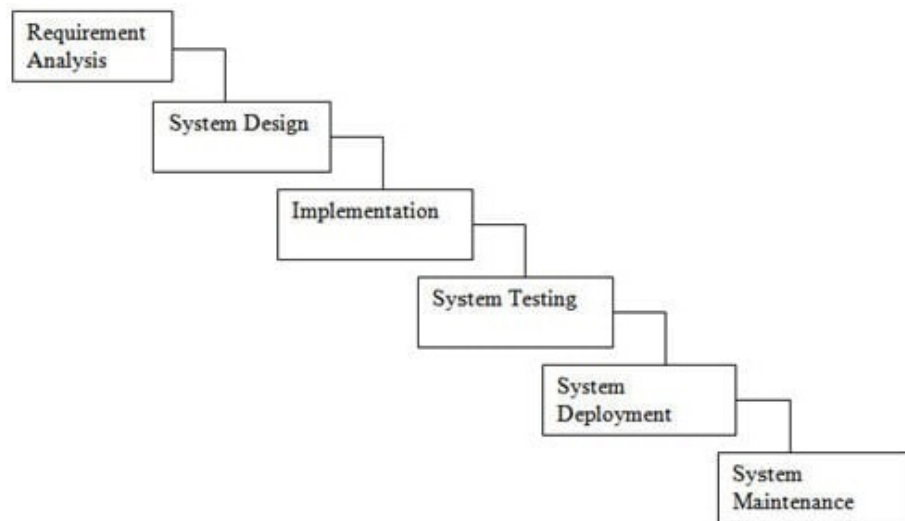
1. Input field: Users are required to fill the input field with correct information
2. Predict: This will predict if patient has cancer or not.

3 Project Analysis and Design

3.1 Methodologies Adapted

In Waterfall model, very less customer interaction is involved during the development of the product. Once the product is ready then only it can be demonstrated to the end users.

Once the product is developed and if any failure occurs then the cost of such issues is very high, because we need to update everything from document till the logic.

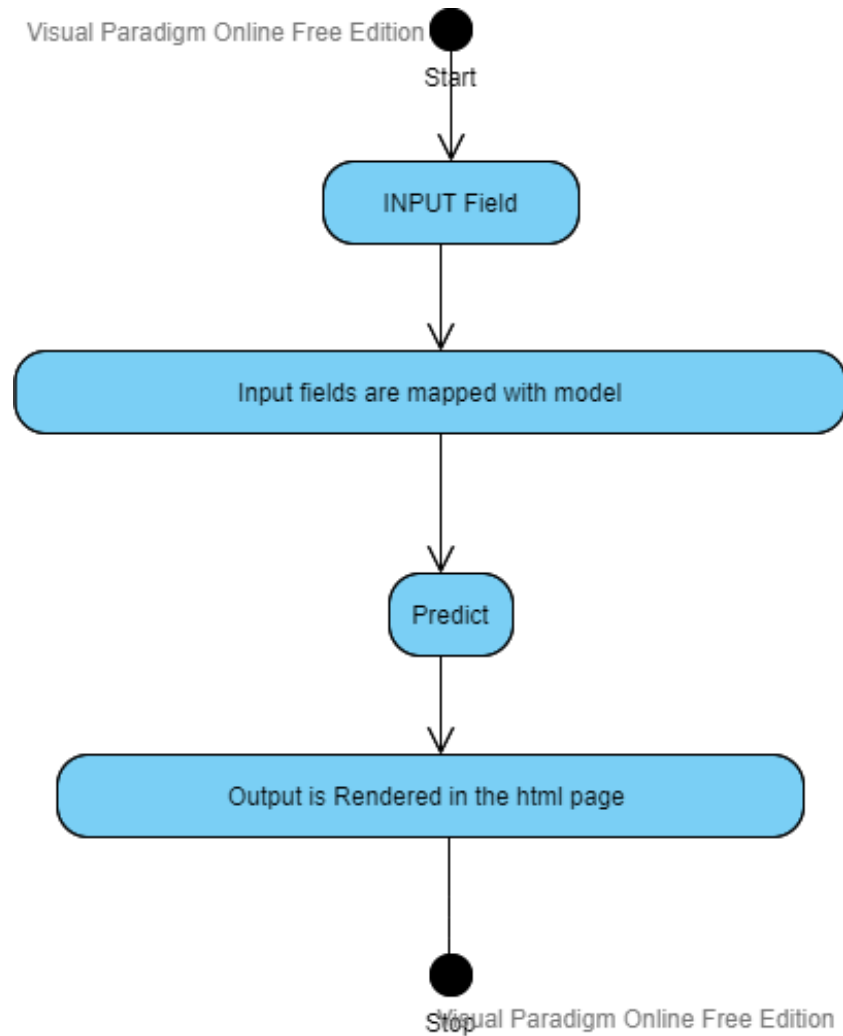


3.1.1: Diagrammatic Representation of Waterfall Model

Breast Cancer Prediction Using Machine Learning

3.2 Modules

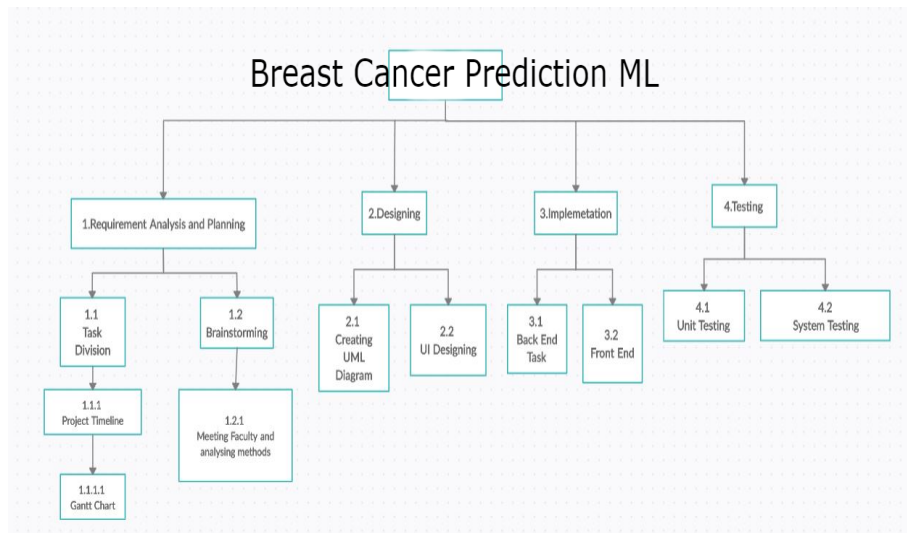
3.2.1 Activity diagram



3.2.1: Activity Diagram

Breast Cancer Prediction Using Machine Learning

3.2.2 Work Breakdown Structure



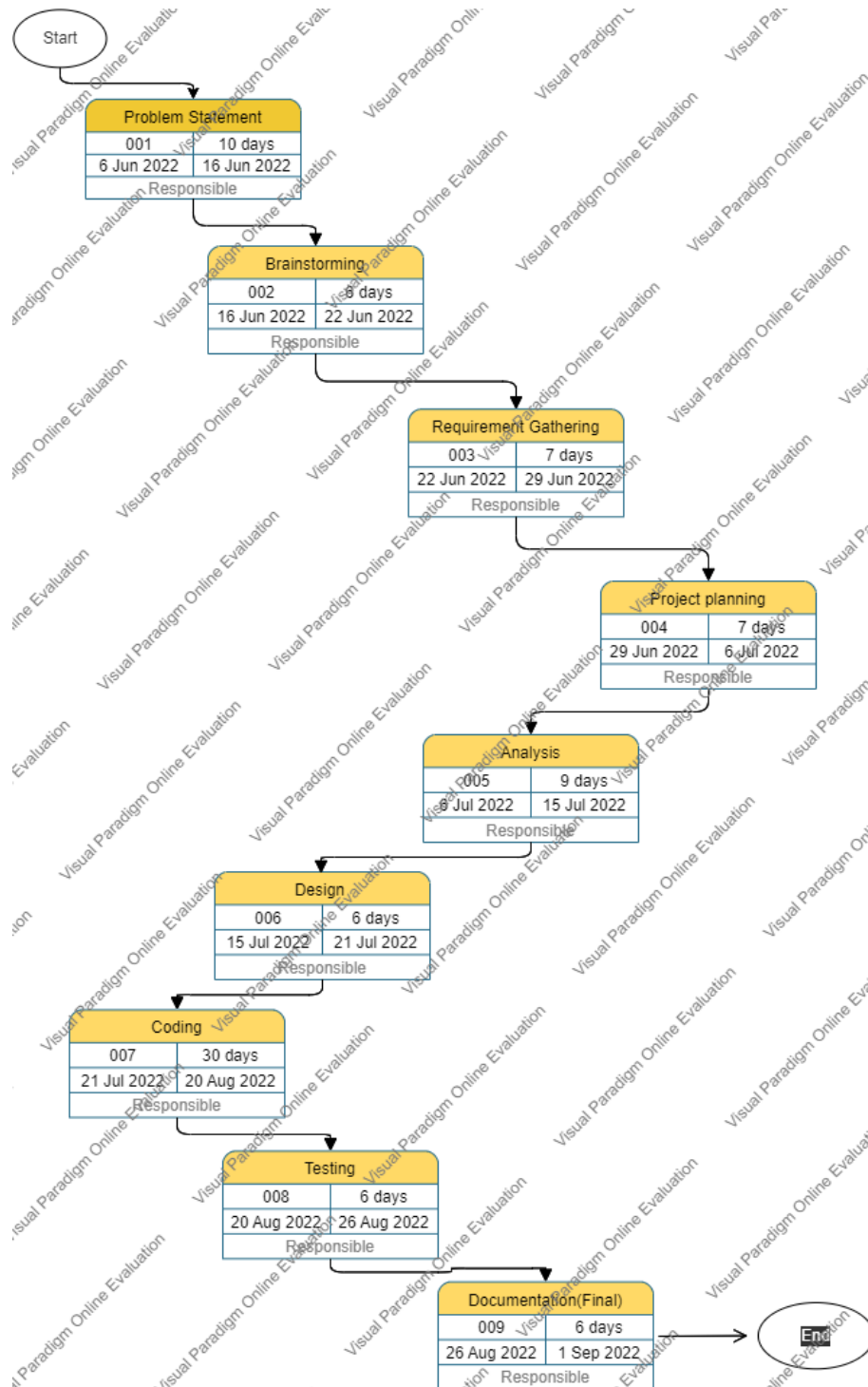
3.2.2: Work Breakdown Structure

Breast Cancer Prediction Using Machine Learning

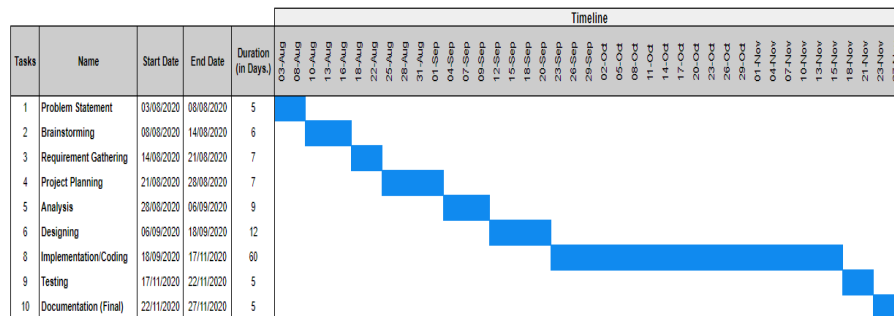
3.2.3 PERT Chart

3.2.4 Gantt Chart

Breast Cancer Prediction Using Machine Learning



3.2.3: PERT Chart



Breast Cancer Prediction Using Machine Learning

Use Cases:

1. Predict

Table 4.2.1: Use Case Table - Register

Use Case ID	1
Use Case Name	Predict
Actor	Doctors
Pre-Condition	-
Post-Condition	Users can see Prediction
Flow of events	Fill input fields

Breast Cancer Prediction Using Machine Learning

4 Project Implementation and Testing

4.1 Input Fields

Breast Cancer Prediction

radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness_mean	concavity_mean
concave_points_mean	symmetry_mean	fractal_dimension_mean	radius_se	texture_se	perimeter_se	area_se
smoothness_se	compactness_se	concavity_se	concave_points_se	symmetry_se	fractal_dimension_se	radius_worst
texture_worst	perimeter_worst	area_worst	smoothness_worst	compactness_worst	concavity_worst	concave_points_worst
symmetry_worst						

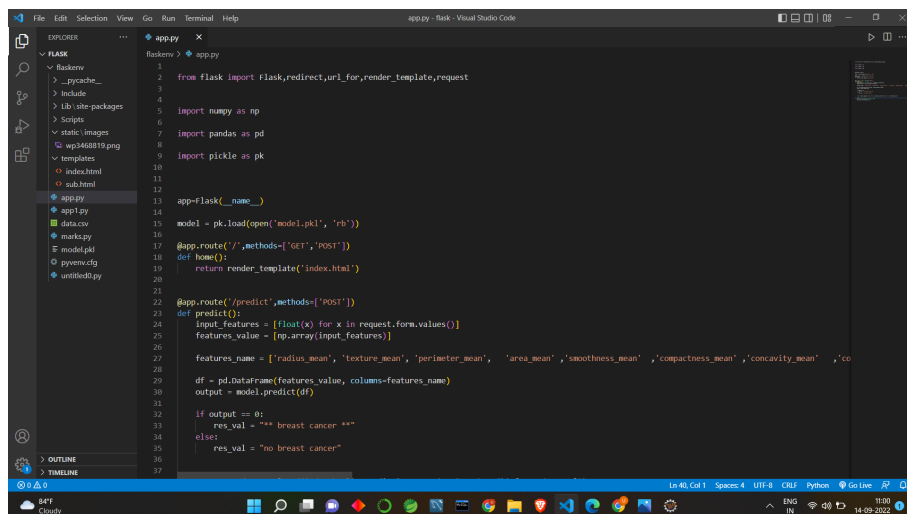
Predict Cancer

Support Breast Cancer

4.1.1: Login and Register

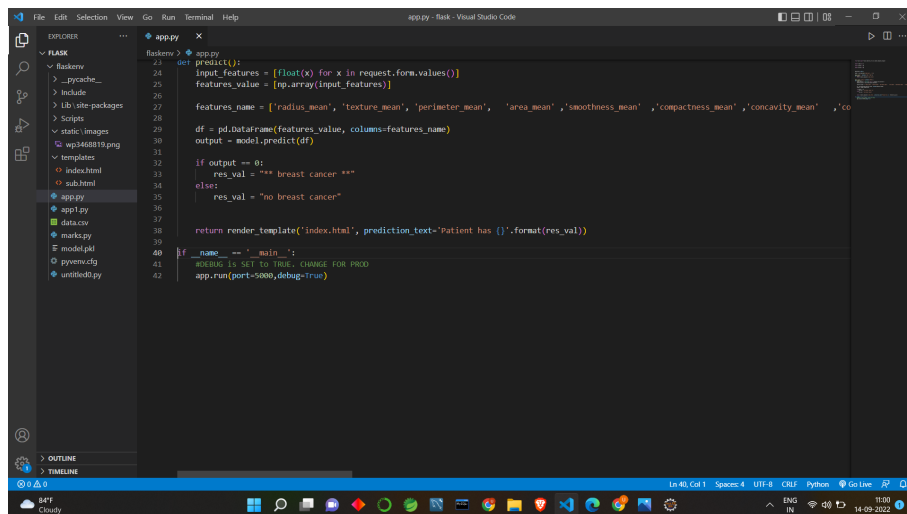
Breast Cancer Prediction Using Machine Learning

4.2 Code 1



```
1 from flask import Flask, redirect, url_for, render_template, request
2
3
4 import numpy as np
5
6 import pandas as pd
7
8 import pickle as pk
9
10
11
12
13 app = Flask(__name__)
14
15 model = pk.load(open('model.pkl', 'rb'))
16
17 @app.route('/', methods=['GET', 'POST'])
18 def home():
19     return render_template('index.html')
20
21
22 @app.route('/predict', methods=['POST'])
23 def predict():
24     input_features = [float(x) for x in request.form.values()]
25     features_value = np.array(input_features)
26
27     features_name = ['radius_mean', 'texture_mean', 'perimeter_mean', 'area_mean', 'smoothness_mean', 'compactness_mean', 'concavity_mean', 'co
28
29     df = pd.DataFrame(features_value, columns=features_name)
30     output = model.predict(df)
31
32     if output == 0:
33         res_val = "breast cancer"
34     else:
35         res_val = "no breast cancer"
```

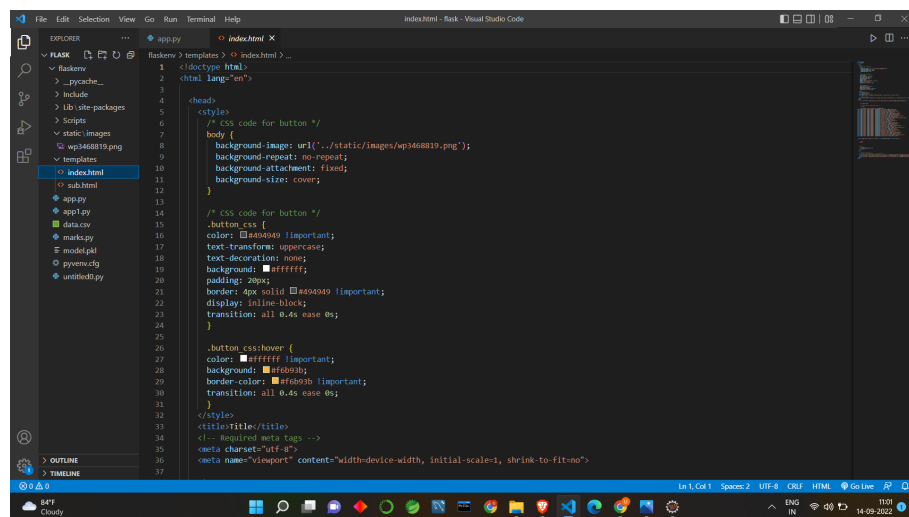
4.3 Code 2



```
24 def predict():
25     input_features = [float(x) for x in request.form.values()]
26     features_value = np.array(input_features)
27
28     features_name = ['radius_mean', 'texture_mean', 'perimeter_mean', 'area_mean', 'smoothness_mean', 'compactness_mean', 'concavity_mean', 'co
29
30     df = pd.DataFrame(features_value, columns=features_name)
31     output = model.predict(df)
32
33     if output == 0:
34         res_val = "breast cancer"
35     else:
36         res_val = "no breast cancer"
37
38     return render_template('index.html', prediction_text='Patient has {}'.format(res_val))
39
40 if __name__ == '__main__':
41     MODELS IS SET TO TRUE. CHANGE FOR PROD
42     app.run(port=5000, debug=True)
```

Breast Cancer Prediction Using Machine Learning

4.4 Code 3



```
1 <!doctype html>
2 <html lang="en">
3
4 <head>
5   <style>
6     /* CSS code for button */
7     body {
8       background-image: url('../static/images/wp3468819.png');
9       background-repeat: no-repeat;
10      background-attachment: fixed;
11      background-size: cover;
12    }
13
14    /* CSS code for button */
15    .button_css {
16      color: #808080 !important;
17      text-transform: uppercase;
18      text-decoration: none;
19      background: #ffffff;
20      padding: 20px;
21      border: 4px solid #808080 !important;
22      display: inline-block;
23      transition: all 0.4s ease 0s;
24    }
25
26    .button_css:hover {
27      color: #ffffff !important;
28      background: #808080;
29      border-color: #808080 !important;
30      transition: all 0.4s ease 0s;
31    }
32  </style>
33  <title>title</title>
34  <!-- required meta tags -->
35  <meta charset="utf-8">
36  <meta name="viewport" content="width=device-width, initial-scale=1, shrink-to-fit=no">
37
```

5 Conclusion

To analyze the medical data, many methods of data mining and machine learning are available. One of the most important challenges in machine learning and data mining areas is to build the accuracy and computationally efficient classifiers for medical applications. Random forest is one of the advanced ensemble learning algorithms and is a very flexible classifier. Forest random algorithm (RF) is an algorithm that forms a family of classification methods that depend on a combination of several decision trees. The random forest also runs efficiently in large databases. However, there is a weakness in the random forest; it is good at classification but not as good as for regression. The main step of the random forest is: 1. From the original dataset, specify n tree bootstrap samples. 2. Grow one tree on each bootstrap data set. Randomly select the m tree variable for separation on each tree node, then grow all three so that each terminal node has no less than a node size case. 3. Collect information from n trees to predict new data such as majority voting for classification. 4. Finally, calculate the level of error out-of-bag (OOB) using data outside the bootstrap sample. We have classified breast cancer using random forest method. The result in this paper is more than 97

6 Bibliography

6.1 Web References

- [1.] <https://www.kaggle.com/code/junkal/breast-cancer-prediction-using-machine-learning>
- [2.] <https://towardsdatascience.com/building-a-simple-machine-learning-model-on-breast-cancer>
- [3.] https://www.researchgate.net/publication/331233978_Predicting_Breast_Cancer_using_Logistic_Regression_and_Multi-Class_Classifiers
- [4.] <https://stackoverflow.com/>
- [5.] <https://www.draw.io/>
- [6.] <https://github.com/gmineo/Breast-Cancer-Prediction-Project/blob/master/Report.Rmd>
- [7.] <https://www.geeksforgeeks.org/>