Cancer Tumor Detection Using Machine Learning

A PROJECT REPORT
SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE COMPLETION OF CS4200-MAJOR PROJECT

BACHELOR OF TECHNOLOGY
IN
COMPUTER SCIENCE AND ENGINEERING

SUBMITTED BY

Mandava vasu
(Enrollment No. 21CS002392)

FACULTY OF COMPUTING AND INFORMATIC
SIR PADAMPAT SINGHANIA UNIVERSITY
UDAIPUR 313601, INDIA

JAN, 2025

Cancer Tumor Detection Using Machine Learning

a Project Report
Submitted in partial fulfillment of the requirements
for CS4200-Major Project

BACHELOR OF TECHNOLOGY
in
Computer Science & Engineering

submitted by

Mandava Vasu
(Enrollment No. 21CS002392)

Under the guidance of
Prof. Alok Kumar
(Project Coordinator)
and
Dr. Harish Tiwari
(Supervisor)

FACULTY OF COMPUTING AND INFORMATICS
SIR PADAMPAT SINGHANIA UNIVERSITY
UDAIPUR 313601, India

JAN, 2025

Faculty of Computing and Informatics
Sir Padampat Singhania University
Udaipur, 313601, India

CERTIFICATE

I, **Mandava Vasu**, hereby declare that the work presented in this project report entitled **"Cancer Tumor Detection Using Machine Learning "** for the completion of CS4200-Major Project and submitted in the **Faculty of Computing and Informatics** of the **Sir Padampat Singhania University, Udaipur** is an authentic record of my own work carried out under the supervision of **Prof. Alok Kumar, Professor**, and **Prof. Harish Tiwari**. The work presented in this report has not been submitted by me anywhere else.

Mandava Vasu
(21CS002392)

This is to certify that the above statement made by the candidate is true to the best of my knowledge and belief.

Prof. Alok Kumar
Professor
Project Coordinator

Dr. Harish Tiwari
Professor
Supervisor

Place: Udaipur
Date:

Acknowledgements

Inscribing these words of gratitude feels akin to painting a masterpiece on the canvas of appreciation. This incredible path of learning and exploration would not have been possible without the unflinching support and encouragement of the great individuals who have paved the road for my accomplishment.

I reserve a special place in my heart for my beloved parents, whose unwavering love, unwavering support, and unwavering belief in my abilities have been the bedrock upon which my dreams have flourished. Their persistent support, sacrifices, and unshakable trust in my abilities have been the driving factors behind my quest for knowledge and academic pursuits.

First and foremost, I owe a tremendous debt of gratitude to my esteemed supervisor, **Prof. Alok Kumar**, whose guidance and advice have been the compass guiding me through the many twists and turns of this thesis. His stimulating conversations, insightful feedback, kind advice, and boundless forbearance have challenged me to push the boundaries of my capabilities and inspired me to strive for academic excellence. I am very thankful for the trust you put in me and the chances you gave me to grow both professionally and personally. I am grateful beyond words for the opportunity to have worked under your guidance, and I hope my thesis serves as a fitting tribute to your hard work, knowledge, and encouragement.

I like to thank **Dr. Amit Kumar Goel**, Dean, Faculty of Computing and Informatics Department, and **Dr. Chandrashekhar Goswami**, Deputy Dean, Faculty of Computing and Informatics Department, for their extended support.

I would like to extend a heartfelt thank you to **ruthvik**, my incredible classmates and friends, who have been a constant source of support, camaraderie, and inspiration. Their presence has made the often-trying process of writing a thesis into one that is filled with joy and fun. Finally, I want to thank everyone who helped me grow as a scholar and made this trip unforgettable.

Mandava Vasu

Abstract

Breast cancer has been identified as the second leading cause of death among women worldwide after lung cancer and hence, it becomes extremely crucial to identify it at an early stage, which can considerably increase the chances of survival. Around 1.1 million cases were recorded in 2004 .The most important part in cancer detection is to be able to differentiate between benign and malignant tumors. implementing Supervised Machine Learning classifiers such as Logistic Regression Classifier, Gaussion Naïve Bayes, Decision tree classifier algorithms by splitting a data into train and test sets. Our aim is to provide a comprehensive view on prediction of breast cancer through Machine Learning through both image and data analyses, which can play a pivotal role in prevention of misdiagnosis in future.

Contents

Certificate ii

Acknowledgements iii

Abstract iv

Contents v

List of Figures vi

1 Introduction 1
- 1.1 Project Introduction . . . 1
- 1.2 Project Overview . . . 1
  - 1.2.1 Objectives . . . 2
- 1.3 Data set . . . 2

2 LITERATURE SURVEY 3
- 2.1 Existing System . . . 3
- 2.2 Proposed System . . . 3
- 2.3 Methodology . . . 3
  - 2.3.1 Dataset Description . . . 4
  - 2.3.2 Dataset Analysis . . . 4
- 2.4 Training and Testing . . . 4

3 SYSTEM ANALYSIS and SYSTEM DESIGN 6
- 3.1 Functional Requirements . . . 6
- 3.2 Performance Requirements . . . 6
- 3.3 Software Requirements . . . 6
- 3.4 Hardware Requirements . . . 7
- 3.5 System Architecture . . . 7
  - 3.5.1 Data Flow Diagram . . . 7

4 IMPLEMENTATION AND RESULTS 9
- 4.1 Languages/Technology Used . . . 9
- 4.2 Used Methods/Algorithm . . . 9
- 4.3 Results(Accuracy)/OutputScreens . . . 9

5 Conclusions and Future Scope 15
- 5.1 Conclusions . . . 15
- 5.2 Future Scope . . . 16

List of Publications 17

List of Figures

3.1 System Architecture 7
3.2 Data Flow Diagram 8
3.3 Class Diagram 8

4.1 code executions 10
4.2 code executions 11
4.3 code executions 12
4.4 code executions 13
4.5 code executions 14

Chapter 1

Introduction

1.1 Project Introduction

Breast cancer is the most common cancer worldwide and leading cancer compared to other types of cancer.Cancer is a syndrome associated with an imbalance of replication cells and cellresponse in the body, causing abnormal cell growth or known as a tumor The tumor is classified as non-cancerous (benign) or cancerous (malignant). Benign tumors do not invade nearby tissues or spread to other areas of the body.

A malignant tumor consists of cancer cells that can invade and kill surrounding tissues and can attack different parts of the body. Cancer cells can spread to other organs causing systemic complications.

Early detection can be done by performing a regular self-examination more accurately, through early screening at nearby public and private health facilities before a person experiences more severe cancer symptoms

1.2 Project Overview

This project is related to Predict Whether the Cancer is Benign or Malignant by using Machine Learning Techniques. The present report starts with a general idea of the project and by representing its objectifs. Then the given dataset will be prepared and setup. An exploratory data analysis is carried out in order to develop a machine learning algorithm that could predict whether a breast cancer cell is benign or malignant until a nal model. Results will be explained. Finally, the report will end with some concluding Best Machine Learning algorthim to get more accuracy.

1.2.1 Objectives

The objective of this project is to report on breast cancer where we took advantage of those available technological advancements to develop prediction models for breast cancer survivability. We used five algorithms(Logistic Regression,K-Means Nearest Neighbor,Support Vector Machine,Decision Tree Algorithm,Random Forest Classifier) to develop the prediction models using a large dataset.The Results (based on average accuracy Breast Cancer dataset) indicated that the Random Forest Classifier is the best predictor with 98.6

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 i have used machine learning classification methods to fit a function that can predict the discrete class of new input.

1.3 Data set

1.5.1 ID number 2) Diagnosis (M = malignant, B = benign) 3–32) 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 (perimeterZ / 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 ("coastline approximation" — 1) The mean, standard error and "worst" or largest (mean of the three largest values) of these features were computed for each image, resulting in 30 features. For instance, field 3 is Mean Radius, field 13 is Radius SE, field 23 is Worst Radius. 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 i have used machine learning classification methods to fit a function that can predict the discrete class of new input.

Chapter 2

LITERATURE SURVEY

2.1 Existing System

The existing system uses Support Vector Machine (SVM) one of the main issues with this is that it need the data to be linearly separable. The system also does not provide enough preprocessing and visualization or Exploratory Data Analysis(EDA).

2.2 Proposed System

We will be implementing Supervised Machine Learning classifiers such as Logistic Regression Classifier, Gaussion Naïve Bayes, Decision tree classifier algorithms by splitting a data into train and test sets. It is done by both image and data analyses.

2.3 Methodology

We have configured a series of steps to come up with the most reliable results in order to determine whether stage of the tumor is malignant (cancerous) or benign (non-cancerous). Our overall methodology can be presented in following subsections

- **A.** Dataset Description
- **B.** Dataset Analysis
- **C** Training and Testing.

2.3.1 Dataset Description

Dr. William H. Wolberg of the University of Wisconsin Hospital in Madison, Wisconsin, USA has developed the WBCD dataset used for this paper which is publicly accessible. This dataset includes 357 and 212 cases of benign and malignant breast cancer respectively as shown in Fig. The dataset comprises 32 columns, with the ID number being the first column and the diagnosis outcome (0-benign and 1-malignant) being the second column. The rest of the columns (3-32) contain three measurements (mean, standard deviation, and mean of worst) of ten features. These features represent the shape and size of the target cancer cell nucleus. The sample of cells is collected from a breast through Fine Needle Aspiration (FNA) procedure in biopsy test. For each cell nucleus, these features are determined by analyzing under a microscope in a pathology laboratory. All values of the features are stored up to four significant digits. There were no null entries in the dataset.

2.3.2 Dataset Analysis

For the dataset analysis, the whole dataset has been considered. In Fig. 2, the mean radius feature of the dataset is counter-plotted. From the figure, it can be observed that suspected patients not bearing cancer have a mean radius of around 1, whereas suspected patients bearing cancer have a mean radius of more than 1,the correlation among the features of the WBCD dataset is shown in a heatmap. Correlation heatmap shows a 2D correlation matrix between two discrete dimensions where the first dimension value is considered as a row and the second dimension value as a column of the heatmap. In this heatmap, the colored cells in a monochromatic scale are used to show the resultant correlation between the features of the dataset. Increasing intensity of color represents increasing correlation. The value of the color of the cells is proportional to the number of measurements that match the dimensional values. The dimensional value (-1 to +1) is calculated from the linearity between the pair of features. If both variables vary and move in the same direction, positive correlation is acquired. In case of negative correlation, increase in one variable is associated with a decrease in the other and vice versa.The 'fractal dimension mean', 'texture se', and 'symmetry se' features are associated very less negatively and other remaining features are highly positively correlated. Fig. 3. Correlation between the different features. The less correlated features can be removed as they have too less impact on the target. We have omitted these features for enhancing the accuracy of the implemented classifiers.

2.4 Training and Testing

Initially, the dataset is read from the CSfile. The data entries from the dataset are analyzed on the basis otheirfeaturesbefore they were used for further step. Then, we split the dataset into two portions randomly:training set and testing set . Not every feature within the dataset is useful and capable of giving same contribution to the result.

According to the data analysis, we have done feature selection to eliminate less correlated features which will increase the accuracy. Then, the dataset is ready for the application of the ML algorithms to examine their performance. Afterthis step, we have accomplished the performance analysis by the comparative study of the resultant testing and training accuracy. Fig. 5 depicts the overall workflow of the study.

Machine learning is an automated approach to learn where algorithms are programmed to gain experience from past datasets for predicting future. In this project, we have used the following ML algorithms: • Logistic Regression. • Support Vector Machine (SVM) • Random Forest. • Naive Bayes. • Decision Tree. • K-Nearest Neighbors (KNN)

Chapter 3

SYSTEM ANALYSIS and SYSTEM DESIGN

3.1 Functional Requirements

1. Logistic Regression
2. Decision Tree Classifier
3. Random Forest Classifier
4. KNN Neighbour
5. Support Vector Machine

3.2 Performance Requirements

1. The system must be of recent version.
2. Data should be collected with utmost care.
3. Robust and Scalability

3.3 Software Requirements

1. **Operating System :**Windows 7 , Windows 8, (or higher versions)

2. **Language :**Python 3.5 and other libraries likes numpy , pandas, matplotlib seaborn and scikit learn .
3. Mozilla Firefox(or any browser)

3.4 Hardware Requirements

1. **Processor:**Pentium 3,Pentium 4 and higher
2. **RAM :** 2GB/4GB RAM and higher
3. **Hard disk :** 40GB and higher

3.5 System Architecture

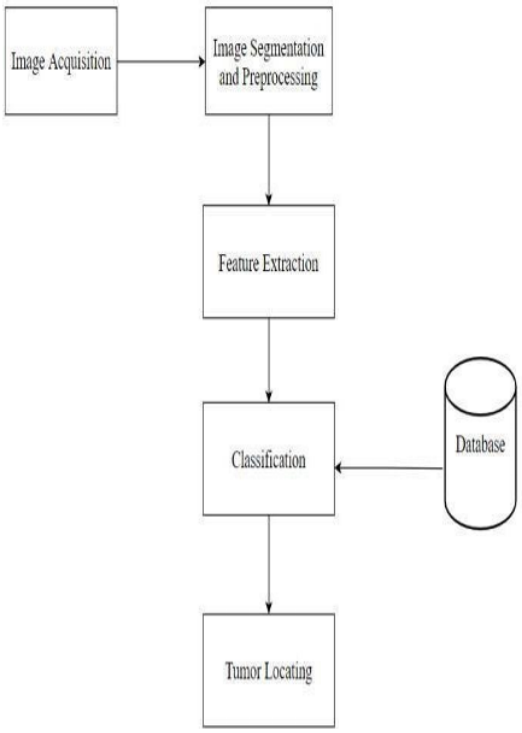
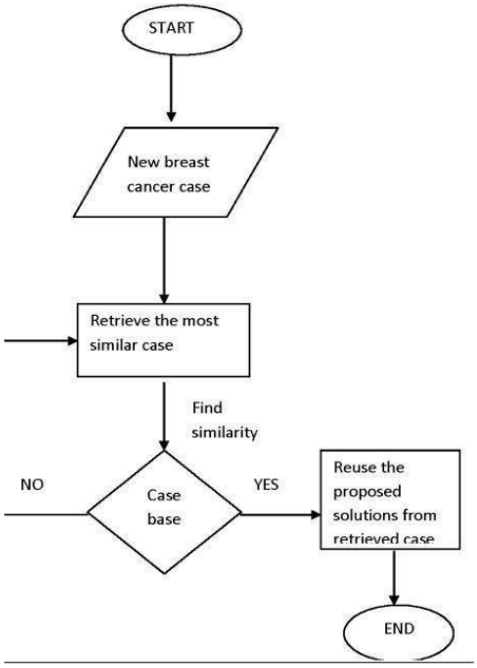


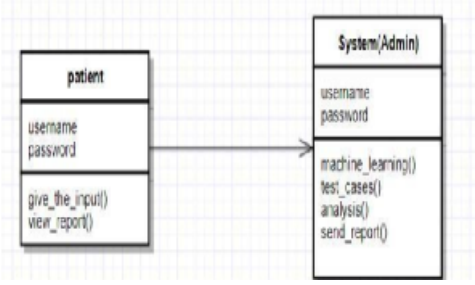
Figure 3.1: System Architecture

3.5.1 Data Flow Diagram



H

Figure 3.2: Data Flow Diagram



H

Figure 3.3: Class Diagram

Chapter 4

IMPLEMENTATION AND RESULTS

4.1 Languages/Technology Used

Python

4.2 Used Methods/Algorithm

1. Support Vector Machine
2. Random Forest Classifier
3. Decision Tree Classifier
4. K-neighbours Classifier
5. Logistic Regression

4.3 Results(Accuracy)/OutputScreens

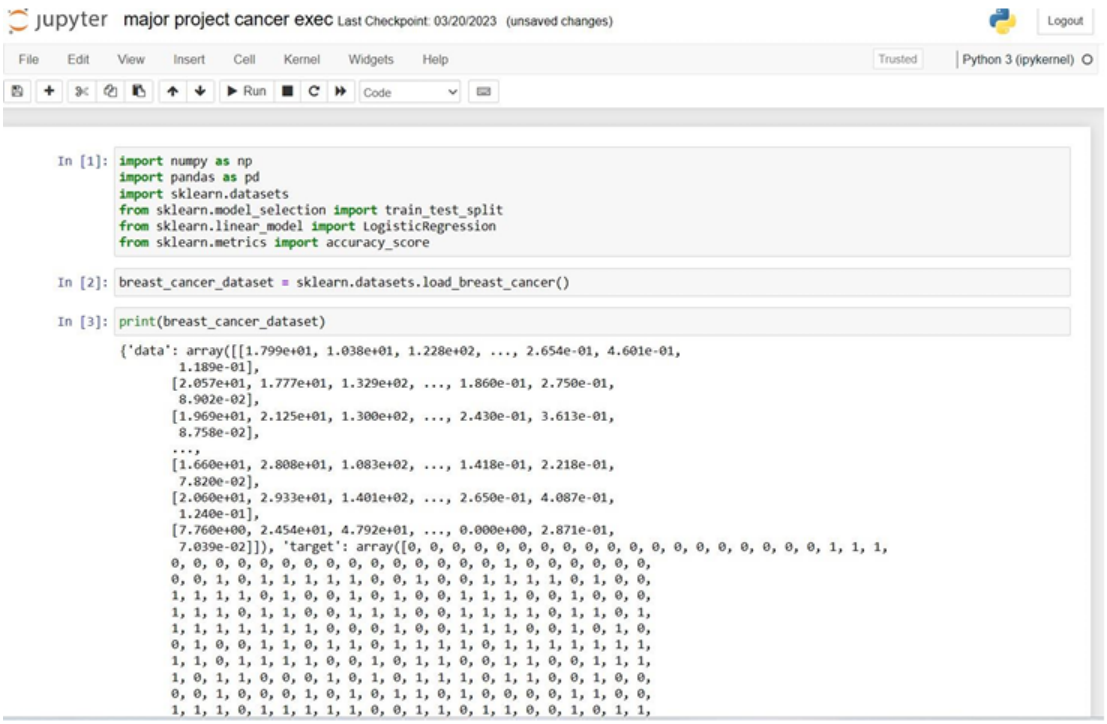
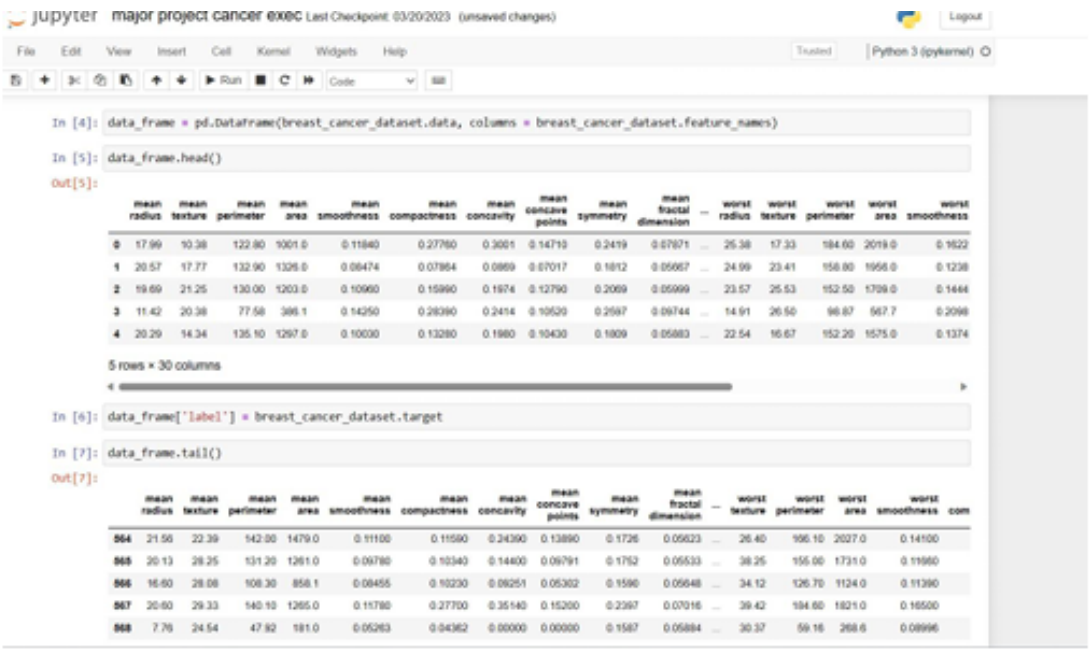


Figure 4.1: code executions



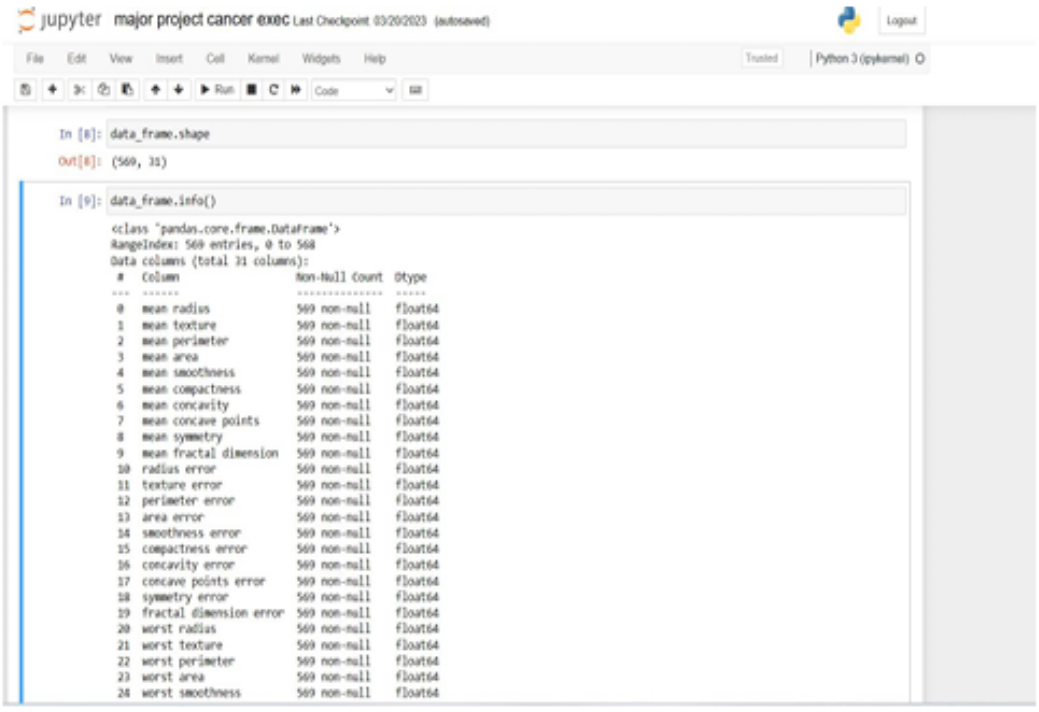


Figure 4.2: code executions

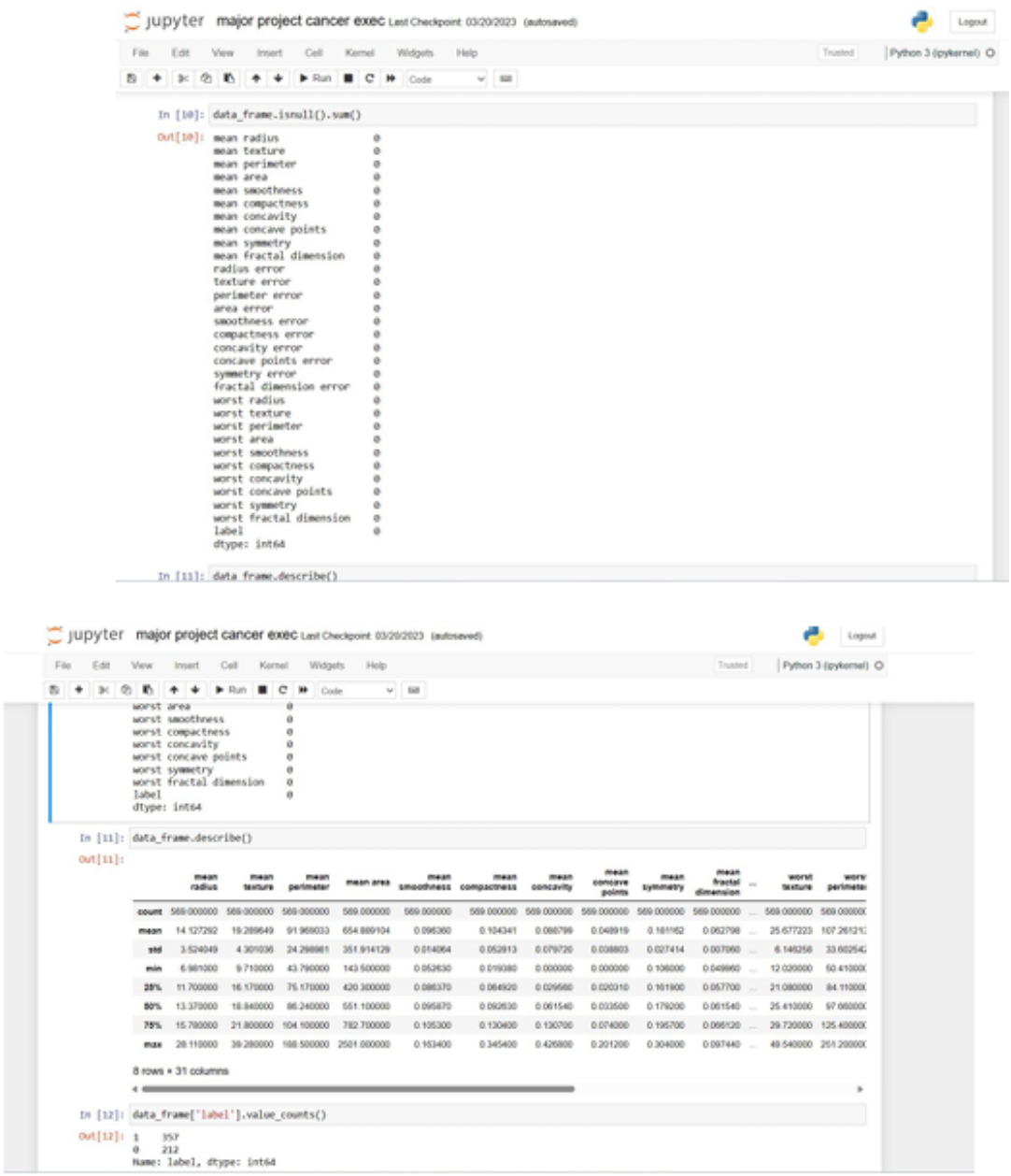
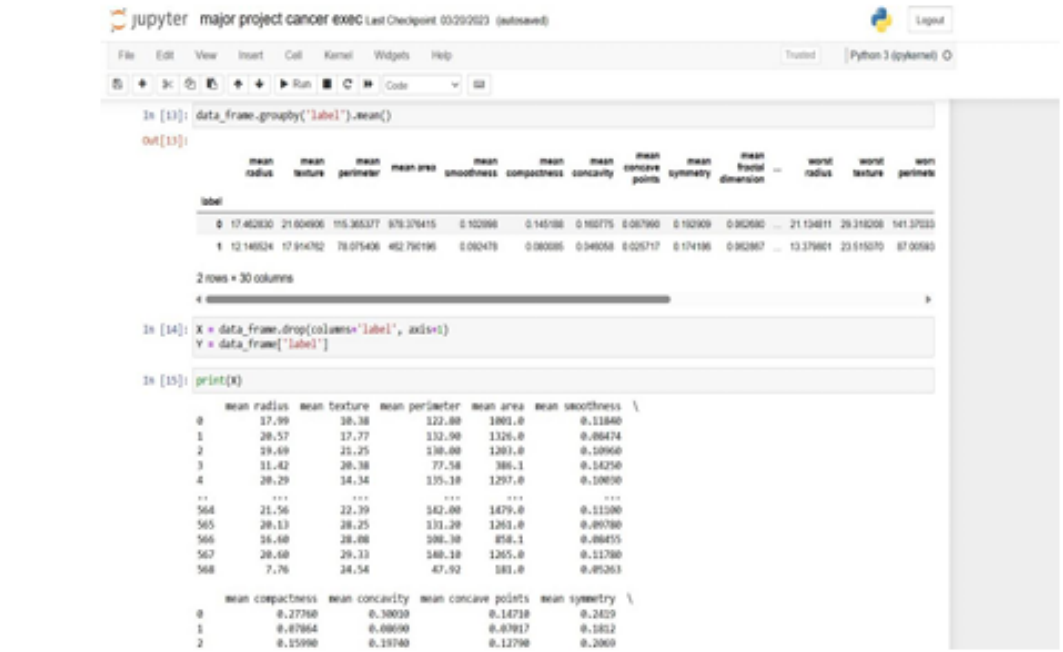


Figure 4.3: code executions



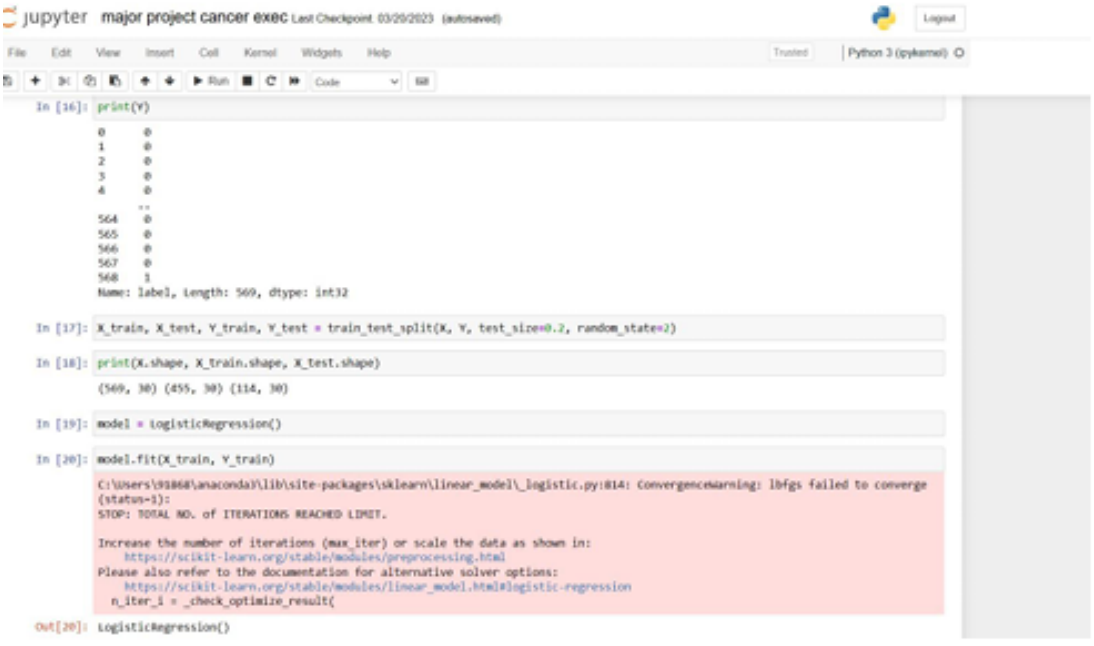


Figure 4.4: code executions

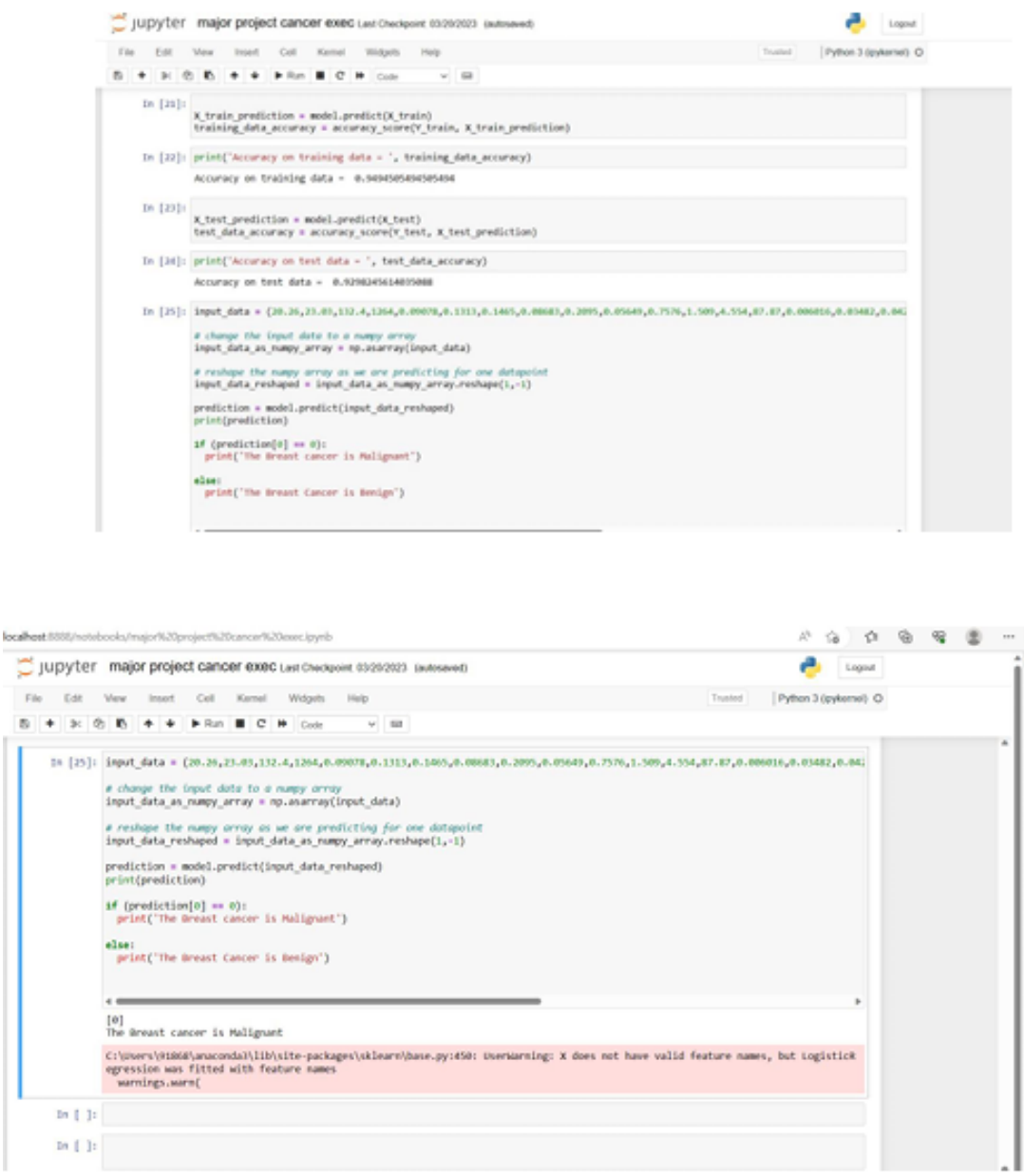


Figure 4.5: code executions

Chapter 5

Conclusions and Future Scope

5.1 Conclusions

In recent years, various breast cancer detection methods using ML algorithms and biosensors have been developed. In addition, breast screening and big data analysis techniques are increasing over time. Most databases require preprocessing depending on the ML model used. However, there are some differences in the coefficient of determination, and different models use different metrics depending on the dataset. Biosensors, on the other hand, require analytical materials that are used to generate electrical signals to identify biomarkers, depending on their application. Like the ML input, the analyte must be processed before testing. The specifications of each biosensor depend on the analyte used. This project described various ML approaches and their applications in the diagnosis and prognosis of breast cancer used to analyze data from the benchmark database WBCD. The application of data mining technology in the medical field is crucial to ensure that it supports the decision-making process. However, such algorithms require high precision and high performance, and the selection of the appropriate method depending on the working context and the data being processed. This study uses five learning algorithms (SVM, Random Forest, Naive Bayes, K-NN, applied to breast cancer datasets) and compares them by many criteria of accuracy, duration, sensitivity, and specificity. The Random Forest Classifier has the lowest error rate of and the shortest turnaround time, demonstrating performance at on some levels than others. Build a classification model and find that the random forest classification algorithm gives the best results for your dataset. Well, it's not always applicable to all datasets. To choose a model, you should always analyze the dataset before applying the machine learning model.

- The ability to predict user ratings more accurately with lower RMSE and MAE values.
- Improved recommendation relevance, as evidenced by higher Precision@10 scores.
- Enhanced robustness in handling challenges such as data sparsity and cold-start problems.

Overall, the proposed method offers a promising approach to social recommendation by leveraging graph-based models to incorporate social connections and user opinions.

5.2 Future Scope

Breast cancer affects about 12 percent of women worldwide, and the trend is increasing. In rare cases, men can also get breast cancer. Every 14 seconds, somewhere in the world, women are diagnosed with breast cancer. Early-stage breast cancer continues to grow and can eventually spread to other parts of the body. Early detection is essential to overcome cancer. The most important risk factors for breast cancer are gender and age. Cancer is a serious health problem with irregular cell rejuvenation that can spread to different parts of the body. Breast cancer cell proliferation is uncontrolled, and when the cancer grows rapidly, the cells become amorphous. About one in eight women develop breast cancer in her lifetime. 2.1 million people are diagnosed each year, most of them between the ages of 40 and 70. Early detection dramatically increases survival. Doctors do not identify all breast cancer patients. That's why machine learning engineers / data scientists work because they have math skills and computing skills. I've trained all supervised classification algorithms to be more accurate, but you can always try some of the popular ones.

These future directions could further enhance the capabilities of the GraphRec+ framework, making it more scalable, accurate, and adaptive to dynamic user preferences and evolving recommendation environments.

References

1. Wang, D. Zhang and Y. H. Huang "Breast Cancer Prediction Using Machine Learning" (2018), Vol. 66, NO. 7.

2. B. Akbugday, "Classification of Breast Cancer Data Using Machine Learning Algorithms," 2019 Medical Technologies Congress (TIPTEKNO), Izmir, Turkey, 2019, pp. 1-4.

3. Keles, M. Kaya, "Breast Cancer Prediction and Detection Using Data Mining Classification Algorithms: A Comparative Study." Tehnicki Vjesnik - Technical Gazette, vol. 26, no. 1, 2019, p. 149+.

4. Delen, D.; Walker, G.; Kadam, A. Predicting breast cancer survivability: A comparison of three data mining methods. Artif. Intell. Med. 2005, 34, 113–127.

5. R. K. Kavitha1, D. D. Rangasamy, "Breast Cancer Survivability Using Adaptive Voting Ensemble Machine Learning Algorithm Adaboost and CART Algorithm" Volume 3, Special Issue 1, February 2014

6. Vikas Chaurasia and S.Pal, "Using Machine Learning Algorithms for Breast Cancer Risk Prediction and Diagnosis" (FAMS 2016) 83 (2016) 1064 – 1069

7. N. Khuriwal, N. Mishra. "A Review on Breast Cancer Diagnosis in Mammography Images Using Deep Learning Techniques", (2018), Vol. 1, No. 1.

8. Y. Khourdifi and M. Bahaj, "Feature Selection with Fast Correlation-Based Filter for Breast Cancer Prediction and Classification Using Machine Learning Algorithms," 2018 International Symposium on Advanced Electrical and Communication Technologies (ISAECT), Rabat, Morocco, 2018, pp. 1-6.

9. R. M. Mohana, R. Delshi Howsalya Devi, Anita Bai, "Lung Cancer Detection using Nearest Neighbour Classifier", International Journal of Recent Technology and Engineering (IJRTE), Volume-8, Issue-2S11, September 2019

10. Ch. Shravya, K. Pravalika, Shaik Subhani, "Prediction of Breast Cancer Using Supervised Machine Learning Techniques", International Journal of Innovative Technology and Exploring Engineering (IJITEE), Volume-8 Issue-6, April 2019.

11. Haifeng Wang and Sang Won Yoon, "Breast Cancer Prediction Using Data Mining Method", Proceedings of the 2015 Industrial and Systems Engineering Research Conference, [14] Abdelghani Bellaachia, Erhan Guven, "Predicting Breast Cancer Survivability Using Data Mining Techniques"

12. P. Sinthia, R. Devi, S. Gayathri and R. Sivasankari, "Breast Cancer detection using PCPCET and ADEWNN", CIEEE' 17, p.63-65