**A**

**PROJECT REPORT**

**ON**

## “BREAST CANCER DETECTION”

**Submitted to**

## KIIT Deemed to be University

**In Partial Fulfillment of the Requirement for the Award of BACHELOR’S DEGREE IN**

**INFORMATION TECHNOLOGY**

**BY**

**AMMAR YASIR**

**NATASHA SHARMA**

**SAGAR SINGH**

**2106184**

**2106200**

**2106245**

**UNDER THE GUIDANCE OF SRICHETA PARUI**



**SCHOOL OF COMPUTER ENGINEERING**

**KALINGA INSTITUTE OF INDUSTRIAL TECHNOLOGY**

**BHUBANESWAR, ODISHA - 751024**

**May 2024**

##### A

##### PROJECT REPORT

ON

##### “BREAST CANCER DETECTION”

Submitted to

### KIIT Deemed to be University

In Partial Fulfillment of the Requirement for the Award of

### BACHELOR’S DEGREE IN INFORMATION TECHNOLOGY

##### BY

|  |  |
| --- | --- |
| AMMAR YASIR | 2106184 |
| NATASHA SHARMA | 2106200 |
| SAGAR SINGH | 2106245 |

UNDER THE GUIDANCE OF

SRICHETA PARUI



SCHOOL OF COMPUTER ENGINEERING

##### KALINGA INSTITUTE OF INDUSTRIAL TECHNOLOGY

BHUBANESWAE, ODISHA -751024

May 2022

### KIIT Deemed to be University

School of Computer Engineering Bhubaneswar, ODISHA 751024



# CERTIFICATE

This is certify that the project entitled

##### “BREAST CANCER DETECTION“

submitted by

|  |  |
| --- | --- |
| AMMAR YASIR | 2106184 |
| NATASHA SHARMA | 2106200 |
| SAGAR SINGH | 2106245 |

is a record of bonafide work carried out by them, in the partial fulfillment of the requirement for the award of Degree of Bachelor of Engineering (Computer Science & Engineering OR Information Technology) at KIIT Deemed to be university, Bhubaneswar. This work is done during year 2022-2023, under our guidance.

Date: / /

(SRICHETA PARUI)

Project Guide

**Acknowledgments**

We are profoundly grateful to SRICHETA PARUI of **Affiliation** for his expert guidance and continuous encouragement throughout to see that this project rights its target since its commencement to its completion. ..................…

AMMAR YASIR NATASHA SHARMA SAGAR SINGH

## ABSTRACT

The breast cancer detection industry aims to develop advanced techniques and technologies for early and accurate detection of breast cancer. Early detection is essential to improve patient outcomes and increase survival. These services typically use a combination of medical imaging, such as mammography, ultrasound, and magnetic resonance imaging (MRI), as well as computational techniques, machine learning, and artificial intelligence (AI) to analyze medical images in the 19th century and translated

The main approach is to use deep learning algorithms trained on large breast image data sets to detect subtle patterns and abnormalities indicative of cancerous lesions These algorithms can help air specialists for the interpretation of mammograms and other imaging studies, reducing the risk of human error and potentially providing a more accurate and feasible diagnosis

In addition, researchers are exploring new biomarkers and molecular imaging techniques that can provide valuable information about the biological characteristics of breast tumors to support personalized treatment strategies to enhance these services, translate research findings to clinical practice Collaboration between is important

Ultimately, breast cancer detection services hold the promise of increasing early detection methods, facilitating timely intervention, and ultimately providing the from patients in the fight against breast cancer has improved

**REAL LIFE PROBLEMS:** The real impact of breast cancer detection programs is profound, as they directly help save lives and improve the lives of individuals affected by the disease

Timely detection not only increases the chances of successful treatment, but also reduces the need for interventions such as mastectomy and chemotherapy, which can has increased patients’ greater physical and emotional reactions

In addition to the benefits to individual patients, breast cancer detection projects have broader societal impacts. They can help reduce the healthcare costs associated with advanced cancer treatment and long- term care by raising awareness of the need for affordable treatments to prevent disease progression besides the need for regular screening and early detection, these programs empower individuals to take the first step in managing their health.

In addition, advances in breast cancer diagnostic technologies and techniques contribute to continued research, deepen understanding of disease biology, and strategic innovation heal the disease

**DRAWBACKS OF BREAST CANCER DETECTION:** While breast cancer detection has improved greatly in terms of early detection and treatment outcomes, it is not infallible A notable limitation is the possibility of false or false positive results a negative, potentially causing unnecessary anxiety or delay in diagnosis and treatment patients, respectively None really, triggering more dangerous procedures and interventions that may be inappropriate and conversely, false negatives when the test fails to detect existing cancer, potentially missing opportunities for early intervention and leading to unchecked disease progression

**ADVANTAGES OF BREAST CANCER DETECTION:**

**Improved precision:** Machine learning algorithms can analyze large amounts of data including medical images and patient profiles to identify pattern and abnormal breast cancer This algorithm can achieve high accuracy in distinguishing tumors from non detect and detect, and can reduce false positives and false negatives compared to traditional methods

**Early Detection:** Machine learning images can detect subtle signs of breast cancer in the earliest stages, even before they are visible to the human eye This enables healthcare professionals to diagnose disease earlier, while treatment is more effective is used, improving patient outcomes and increasing survival

**Automation and efficiency:** Machine learning-based tools can automate aspects of breast cancer detection, such as image analysis and interpretation This provides the workload of radiologists and healthcare professionals is reduced, allowingg them to focus their time and expertise on more complex cases and patient care

**Scalability and Accessibility:** Once machine learning models are trained, machine learning models can be used in a variety of healthcare settings, including rural or undeserved areas where access to basic medical skills may be limited These advances increase access to advanced breast cancer detection technologies, ensuring equitable health care delivery.Improved precision: Machine learning algorithms can analyze large amounts of data including medical images and patient profiles to identify pattern and abnormal breast cancer This algorithm can achieve high accuracy in distinguishing tumors from non detect and detect, and can reduce false positives and false negatives compared to traditional methods

**Early Detection:** Machine learning images can detect subtle signs of breast cancer in the earliest stages, even before they are visible to the human eye This enables healthcare professionals to diagnose disease earlier, while treatment is more effective is used, improving patient outcomes and increasing survival

**Personalized medicine:** Machine learning algorithms can analyze complex datasets to identify unique biomarker molecular signatures associated with sub types of breast cancer This information can help tailor treatment plans to individual patients, making treatment take off effective and reduce adverse effects.

**Automation and efficiency:** Machine learning-based tools can automate aspects of breast cancer detection, such as image analysis and interpretation This provides the workload of radiologists and healthcare professionals is reduced, allowing them to focus their time and expertise on more complex cases and patient care

**Scalability and Accessibility:** Once machine learning models are trained, machine learning models can be used in a variety of healthcare settings, including rural or undeserved areas where access to basic medical skills may be limited These advances increase access to advanced breast cancer detection technologies, ensuring equitable health care delivery.

**PROBLEMS OF BREAST CANCER DETECTION:** False Positives and False Negatives: False positives arise whilst a screening take a look at incorrectly suggests the presence of most cancers, leading to unnecessary anxiety and invasive follow-up techniques. Conversely, fake negatives occur when cancer is present but no longer detected, delaying diagnosis and remedy. Balancing sensitivity and specificity in detection methods is essential to limit those errors.

**Breast Density:** Dense breast tissue can obscure abnormalities on mammograms, reducing the sensitivity of screening checks and increasing the threat of overlooked diagnoses. Women with dense breasts can also require extra screening modalities, which include ultrasound or MRI, to improve detection fees.

**Over diagnosis and Over treatment:** Some breast cancers detected via screening may be sluggish-growing and unlikely to cause damage at some stage in a patient's lifetime. However, competitive remedies may nonetheless be pursued, leading to unnecessary interventions and capacity damage to sufferers. Balancing the risks and benefits of early detection and treatment is critical to keep away from over diagnosis and over treatment.

**Cost and Accessibility:** Advanced screening technology, along with MRI, may be costly and inaccessible to sure populations, leading to disparities in breast most cancers detection and results.

Access to less expensive and effective screening applications is critical to make sure equitable healthcare get right of entry to for all people.

**Psychological Impact:** A breast cancer diagnosis, even if detected early, will have profound psychological effects on patients and their families. Fear, tension, and uncertainty about the destiny are common emotional responses that could effect intellectual health and nicely-being. Providing good enough support and assets for sufferers navigating the emotional challenges of analysis and remedy is esse

**False positives and false negatives:** False positives occur when screening tests incorrectly show the presence of cancer, leading to unnecessary anxiety and procedures does follow-up terrible occur Unlike a

false positive but missed diagnosis of cancer, there was a delay in diagnosis and treatment. To mitigate these shortcomings, it is important to balance sensitivity and specificity in detection methods.

**Dense breast tissue:** Dense breast tissue can mask abnormalities on mammograms, reducing the sensitivity of screening tests and increasing the risk of missing the disease Women whose breast firmness may require additional screening techniques, such as ultrasound or MRI, to make a more accurate diagnosis.

**Over diagnosis and over treatment:** Some breast cancers diagnosed through screening may be slow growing and not likely to harm a patient’s life but they still undergo aggressive treatments, causing a unnecessary occurs and can be harmful to patients. Balancing the risks and benefits of early detection and treatment is important to avoid over diagnosis and over treatment.

**Cost and Accessibility:** Advanced screening technologies such as MRI can be expensive and inaccessible to some people, leading to disparities in breast cancer diagnosis and outcomes Screening system affordable and effective access is critical to ensuring equal access to health care for all individuals.

**Psychological impact:** A breast cancer diagnosis can have a significant psychological impact on patients and their families even if detected early. Fear, anxiety, and uncertainty about the future are common emotional reactions that can affect mental health and well-being. It is esse to provide appropriate support and resources to patients who are navigating the emotional challenges of diagnosis and treatment

**Five Keywords:** Machine learning: This refers to general algorithms that can learn from data to identify and predict trends, a key driver of breast cancer detection services.

Deep learning: Machine learning using complex neural networks is particularly adept at image recognition, making it valuable for analyzing mammograms and other breast images

Computer-aided detection (CAD) : This describes systems that use algorithms to analyze medical images with radiologists, and can improve the accuracy and efficiency of breast cancer detection

Biomarkers: These are biomarkers the presence or amount of which can indicate the presence of a disease. Breast cancer detection services can screen with specific biomarkers in addition to imaging modalities.

Dense breast tissue: Dense breast tissue can mask abnormalities on mammograms. Research projects could explore alternative or alternative methods for women with dense breasts.

# Contents

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 1 | Introduction | | | 1 |
| 2 | Basic Concepts/ Literature Review | | | 2 |
|  | 2.1 | Sub Section Name........................... | | 2 |
| 3 | Problem Statement / Requirement Specifications | | | 3 |
|  | 3.1 | Project Planning........................... | | 3 |
|  | 3.2 | Project Analysis (SRS)................. | | 3 |
|  | 3.3 | System Design ………………….. | | 3 |
|  |  | 3.3.1 | Design Constraints …… | 3 |
|  |  | 3.3.2 | System Architecture (UML) / Block Diagram … | 3 |
| 4 | Implementation | | | 4 |
|  | 4.1 | Methodology / Proposal ........................... | | 4 |
|  | 4.2 | Testing / Verification Plan ……………. | | 4 |
|  | 4.3 | Result Analysis / Screenshots …………. | | 4 |
|  | 4.4 | Quality Assurance …………………….. | | 4 |
| 5 | Machine learning Algorithms | | | 5 |
|  | 5.1 | Types of Algorithms…… | | 5 |
|  | 5.2 | Comparison of different types of algorithms….. | | 5 |
|  | 5.3 | Speculation of all the algorithms…….. | | 5 |
| 6 | Conclusion and Future Scope | | | 6 |
|  | 6.1 | Conclusion ……………………….. | | 6 |
|  | 6.2 | Future Scope ………………………. | | 6 |
| References | | | | 7 |
| Individual Contribution | | | | 8 |
| Plagiarism Report | | | | 9 |

List of Figures

1.1 IMAGE CAPTION ......................... 2

4.1 IMAGE CAPTION ......................... 9

BREAST CANCER DETECTION

# Chapter 1 Introduction

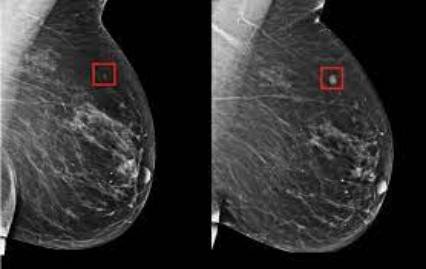
Breast cancer is a major health problem, diagnosed in millions of women worldwide each year. Despite continuous advances in treatment, early detection remains the most powerful tool in the fight against this disease. This introductory article will examine the current need for effective breast cancer detection strategies.

**Emphasis on needs:**

"Breast cancer is the most common aggressive cancer diagnosed in women worldwide." Emphasize the impact of early detection. "Early detection plays an important role, increasing the chances of successful treatment and improving survival."

**Classification of requirements**:

"Although existing detection methods such as mammograms are valuable, constant improvements are needed due to limitations such as false positives, equivocality and inefficiency for women who have them." for the stiffness of the breast tissue." Introduce the idea of improving accuracy and accessibility. "The focus is on developing accurate, sensitive and patient-friendly methods for early detection."



**Other points to consider:**

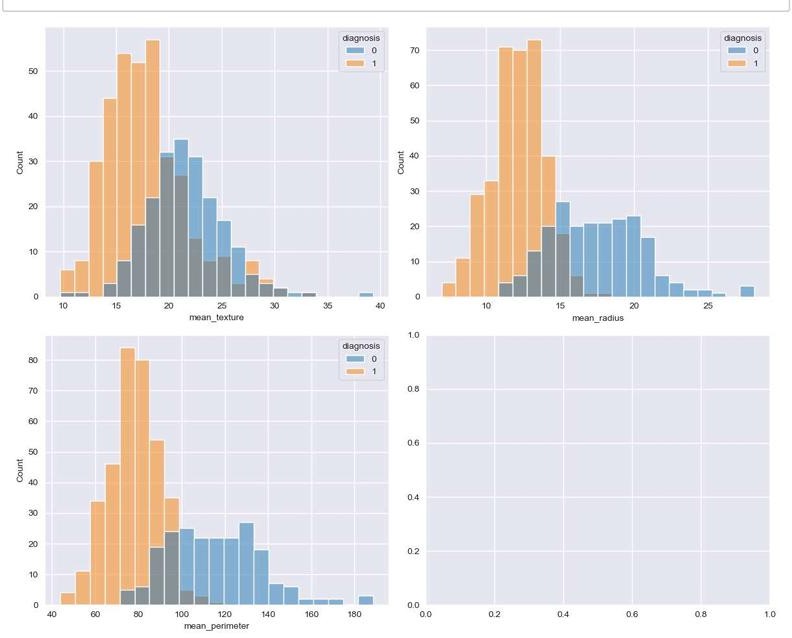
Briefly discuss its effect on mortality. Early detection can significantly reduce breast cancer mortality.

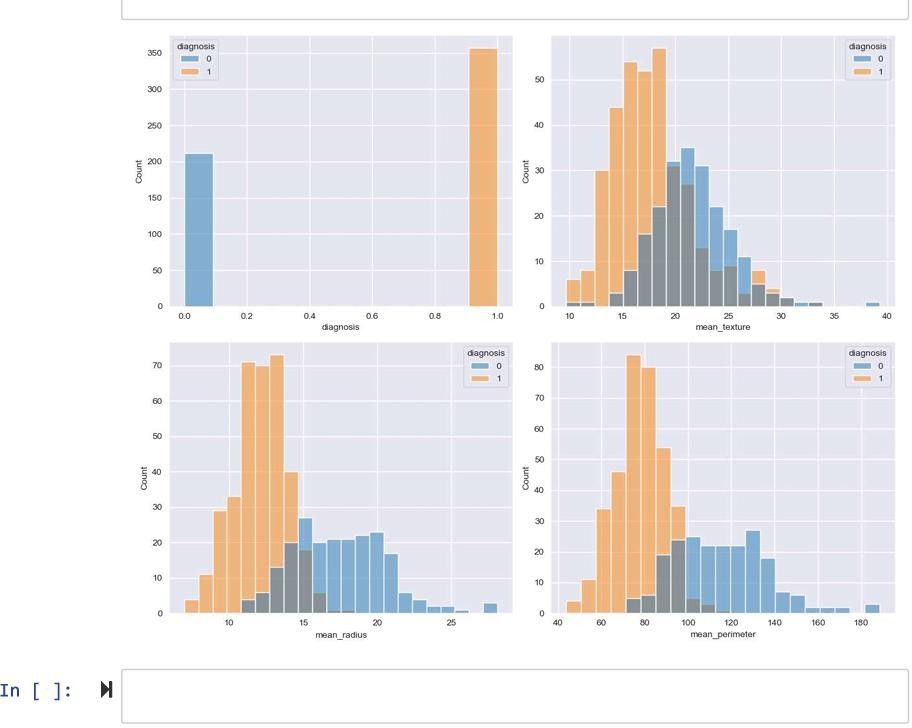
Briefly discuss the need for personalized screening methods based on individual risk factors. By describing current limitations and discussing the need for better detection strategies, this introduction effectively establishes the advances in breast cancer detection

**Importance of Breast cancer detection:**

Early detection of breast cancer is essential for successful treatment and improved survival. Catching cancer early leads to more effective and less invasive treatments, as well as a better quality of life for patients. This not only makes it easier for individuals to recover fully, but also reduces the overall burden on health care systems. By prioritizing detection through routine screening, we can empower individuals to take charge of their health and significantly reduce breast cancer mortality.

Regular cancer screening and recognizing possible warning signs are two important steps in early detection. Tests like mammograms and colonoscopies can detect abnormalities before they become a problem. Informing individuals about possible symptoms specific to different cancers allows individuals to seek treatment as soon as they notice any changes in their bodies by prioritizing early detection we increase the chances of the results are much greater and we give ourselves the best chance of fighting this disease.





# Chapter 2

BREAST CANCER DETECTION

# Basic Concepts/ Literature Review

Supervised learning: A common approach to cancer diagnosis. Graphic artists are trained with labeled data (images identified as malicious or dangerous) to learn how to classify new images.

Support Vector Machines (SVM): Refine images based on specific features.

Decision Trees: Create a tree-like structure to classify images based on a series of decision rules.

Random Forests: Combine multiple decision trees for accuracy and robustness. Convolution Neural Networks (CNN): Effective deep learning models especially in image analysis. Relevant features can be identified automatically from the data.

Unsupervised study: Can be used for anomaly detection by identifying suspicious areas in mammograms that deviate from normality.

**Pre-processing data:**

Image Composition: Separates the breast area from the outside in mammograms. Normalization: Standardizes image size and robustness for consistent analysis using machine learning models.

Data enhancement: automating existing model types to improve model robustness and prevent over fitting.

**Evaluation metrics:**

Accuracy: Half of the images correctly sorted.

Sensitivity: Ability to correctly detect cancer (true positive). Accuracy: Ability to identify healthy information (true bad).

Area under the curve (AUC): Measures the overall performance of the classification model.

Other things to consider:

**Description**: Developing a model for health care providers to understand trust building and integrating AI into the decision-making process.

**Bias mitigation**: Ensures that training data reflects population diversity and avoids biased predictions.

These are some basic ideas. For a machine learning project, you would depend further on Chosen techniques and specific goals.

BREAST CANCER DETECTION

# Chapter 3

Problem Statement / Requirement Specifications

* 1. **Project Planning**
* Purpose:
  + - This Project describes a software application that uses the user's tumor description to forecast whether a patient has cancer or not. The program calculates the user's cancer stage using Random Forest model of machine learning.
* System description:
  + - The application uses an API to take the tumors characteristic features and then run it through the model to predict the output. The inputs taken are mean\_radius , mean\_texture and mean\_perimeter.
    - Based on the input, the programme uses a predefined model mapped to the database to categorise the user's condition. For example, if the user has the tumour conditions which maps that the user has cancer that is the dependent variable is 1 then the output is given that the user has cancer and should consult a doctor.
  1. **Project Analysis**

## External Interface Requirements

**1. User Interfaces (UI)**

* **Data Input:**
  + Users will be able to enter patient data that is pertinent to the prediction of breast cancer using the system's user interface. This have attributes like:
    - * The quantity of breast biopsies
      * kind of breast cell
      * Results of mammography (text entry or picture uploading for future improvements)
      * characteristics of tumors, such as texture and size.
  + The system offers an interface where the developer has already chosen a pre-trained machine learning model for the purpose of predicting breast cancer.
  + Users with different levels of technical competence is be able to enter data with ease because to the interface's ease of use and intuitiveness.
  + To find and alert users to inaccurate or missing data entries, error handling is put into place.
  + Clear explanations of the available models, together with information on their accuracy and performance metrics, is available in the EDA of the model.
* **Prediction Output:**
  + The estimated likelihood that a patient would get breast cancer is shown by the system.
  + The interface clearly and succinctly displays the results, including confidence scores and explanations.

**2. Hardware Interfaces**

* There are very few external hardware requirements on the system.
* But it can support integration with future enhancements like:
  + Specialized medical imaging devices (for directly scanning the cancer area and taking the required featural characteristic input of the area.

**3. Software Interfaces (SI)**

* **Machine Learning Libraries:**
  + For cancer prediction tasks, the system leverages machine learning packages such as scikit-learn, PyTorch, and TensorFlow.
* **Database Interface:**
  + The system interfaces with a database that contains historical patient data.
* **External APIs:**
  + The Web application that connects the program and device is created by the system using PyScript.

**4. Communications Interface (CI)**

* The system's primary mode of communication shall be through its user interface.
* Secure communication protocols may be implemented by the system to facilitate interfaces with external APIs.

**Future Enhancements:**

* Integration for smooth data sharing with hospital information systems.
* Support for multiple languages.

**Functional Requirements:**

* **Breast Cancer prediction using Naive Bayes Model:**

**1. Library Imports:**

* + - * Imported the NumPy library for numerical computations.
      * Imported the Pandas library for data manipulation and analysis.
      * Imported the Matplotlib library for creating visualizations
      * Imports the Seaborn library for statistical data visualization

**2. Data Loading:**

* + - * Reading the breast cancer data from a CSV file named "Breast\_cancer\_data.csv" and storing it in a Pandas Data-Frame.

**3. Data Splitting:**

* + - * Importing the “train\_test\_split” function from scikit-learn model selection module.
      * Creating a new Data-Frame “x” containing all features (columns) except the target variable "diagnosis".
      * Separating the target variable "diagnosis" into a Series named “y”.
      * Splitting the data into training and testing sets using “train\_test\_split”.
        + Allocating 20% of the data for testing.
        + Ensuring the class distribution is preserved in both training and testing sets. Crucial for imbalanced datasets like breast cancer data.
        + Randomly shuffling the data before splitting.
        + Setting a random seed for reproducibility i.e. ensures the same split if the code is run multiple times).
      * Printing the shapes (number of rows and columns) of the training and testing sets.

**4. Data Preprocessing (Normalization):**

* + - * Importing the “StandardScaler” class from scikit-learn preprocessing module.
      * Creating a standardizer object “sc” to perform standard scaling.
      * Fitting the standardizer on the training data (X\_train) calculating the mean and standard deviation of each feature.
      * Transforms the testing data (X\_test) using the parameters learned from the fitted standardizer. This ensures both training and testing data have the same scaling, improving model performance.

**5. Model Implementation (Naive Bayes)**

* + - * Importing the Gaussian Naive Bayes classifier from scikit-learn.
      * Creating a Gaussian Naive Bayes classifier object named “clf”.
      * Training the classifier on the training data (X\_train and y\_train). This involves learning the model parameters based on the features and corresponding diagnoses.

**6. Prediction:**

* + - * Using the trained classifier (clf) to predict the diagnosis labels for the testing data (X\_test). The predicted labels are stored in y\_pred.

**7. Evaluation (Confusion Matrix):**

* + - * Importing the confusion\_matrix function from scikit-learn metrics module.
      * Creating a confusion matrix cm that compares the actual diagnoses (y\_test) with the predicted diagnoses (y\_pred).
        + Rows represent the actual diagnoses.
        + Columns represent the predicted diagnoses.
        + Each cell (i, j) shows the number of samples where the actual diagnosis was class “I” but the model predicted is class “j”.
      * The confusion matrix provides insights into the model's performance, such as the number of correct and incorrect predictions for each class.
      * Printing the confusion matrix allows you to visualize these.
    - **Breast Cancer Prediction using Decision Tree Classifier:**

**1. Library Imports:**

* + - * Imported the NumPy library for numerical computations.
      * Imported the Pandas library for data manipulation and analysis.
      * Imported the Matplotlib library for creating visualizations
      * Imports the Seaborn library for statistical data visualization

**2. Data Loading:**

* Reading the breast cancer data from a CSV file named "Breast\_cancer\_data.csv" and stores it in a Pandas DataFrame named “df”.

**3. Data Splitting:**

* + - * Separating features (X) from the target variable (y).
      * Spliting the data into training and testing sets (X\_train, X\_test, y\_train, y\_test) using “train\_test\_split” while ensuring that class distribution is preserved.
      * Maintaining randomness and setting a random seed) for reproducibility.
* Printing the shapes (number of rows and columns) of the training and testing sets.

**4. Data Preprocessing (Normalization):**

* Using StandardScaler (sc) to normalize the training data (X\_train).
* The testing data (X\_test) is then transformed using the parameters learned from the fitted standardizer (sc.transform(X\_test)) to ensure consistent scaling across both sets.

**5. Model Implementation (Decision Tree):**

* Importing the Decision Tree Classifier class from scikit-learn.
* Creating a Decision Tree Classifier object named “dt”. The “max\_depth” parameter controls the complexity of the tree, with higher values leading to more complex and potentially overfitting trees.
* Training the decision tree classifier on the training data (X\_train and y\_train) where the decision rules are based on the features and corresponding diagnoses.

**6. Prediction:**

* Using the trained decision tree classifier (dt) to predict the diagnosis labels for the testing data (X\_test). The predicted labels are stored in “y\_pred”.

**7. Evaluation (Confusion Matrix):**

* Confusion matrix (cm) shows the distribution of actual vs. predicted diagnoses.
* Interpreting the confusion matrix and the visualization helps identify potential biases or areas for improvement in the model.
* The performance of the model is showed then by calculating the Accuracy, Precision, Recall and F1-Score.

## Breast Cancer Prediction with K-Nearest Neighbors (KNN):

**1. Library Imports:**

* + - * Imported the NumPy library for numerical computations.
      * Imported the Pandas library for data manipulation and analysis.
      * Imported the Matplotlib library for creating visualizations
      * Imports the Seaborn library for statistical data visualization

**2. Data Loading:**

* Reading breast cancer data from a CSV file named "Breast\_cancer\_data.csv" and stores it in a Pandas DataFrame named “df”.

**3. Data Splitting:**

* Splitting the data into training and testing sets (X\_train, X\_test, y\_train, y\_test) using “train\_test\_split”. Here, “stratify=y” ensures the class distribution is preserved, and random\_state=21 sets a random seed for reproducibility.

**4. KNN Hyperparameter Tuning (Number of Neighbours):**

* Finding the optimal number of neighbours for the KNN model.
* Creating a list of neighbours from 1 to 30 (inclusive).
* A loop iterates through each value in neighbours:
  + Creating a KNN classifier object (knn) with the current number of neighbours’ (neighbor).
  + Fitting the KNN classifier on the training data.
  + Appending the training accuracy for the current number of neighbors to the “train\_accuracies” list.
* After the loop completes, we have the lists containing training and testing accuracies for different numbers of neighbors.

**5. Visualization of KNN Performance:**

* Creating a Matplotlib figure for plotting.
* Plotting the training accuracies vs. the number of neighbors.
* Plotting the testing accuracies vs. the number of neighbors.
* The plot visualizes how the model's performance (accuracy) changes with the number of neighbors considered for classification.

**6. KNN Model with StandardScaler:**

* Importing the “make\_pipeline” function from scikit-learn to create a pipeline.
* Creating a pipeline named “clf” that consists of two steps:
  + StandardScaler(): Standardizes the features in the training data.
  + KNeighborsClassifier(): KNN classifier with a chosen number of neighbors (here, 3 based on potential insights from the previous visualization).
* Fitting the pipeline (including both the scaler and the KNN classifier) on the training data.

**7. Evaluation and Prediction:**

* Evaluating the model's performance on the training data.
* Evaluating the model's performance on the testing data.
* Visualizing the confusion matrix for the model's predictions on the testing data.
* Generatiing predictions for the testing data using the trained pipeline (clf).
* The confusion matrix and predicted labels help assess the model's ability to correctly classify new data points.
* The performance of the model is showed then by calculating the Accuracy, Precision, Recall and F1-Score.

## Breast Cancer Prediction using Logistic Regression

**1. Library Imports:**

* + - * Imported the NumPy library for numerical computations.
      * Imported the Pandas library for data manipulation and analysis.
      * Imported the Matplotlib library for creating visualizations
      * Imports the Seaborn library for statistical data visualization

**2. Data Loading:**

* Reading breast cancer data from a CSV file named "Breast\_cancer\_data.csv" and storing it in a Pandas DataFrame named “df”.

**3. Data Splitting:**

* Splitting of data into training and testing sets (X\_train, X\_test, y\_train, y\_test) using “train\_test\_split”. Ensuring that the class distribution is preserved, and random\_state=21 sets a random seed for reproducibility.

**4. Data Preprocessing (Normalization):**

* Importing the “StandardScaler” class for data normalization.
* Creating a standardizer object (sc).
* Fitting the standardizer on the training data (calculating mean and standard deviation for each feature).
* Transforing the testing data using the parameters learned from the fitted standardizer. This ensures that there is consistent scaling across both sets, improving model performance for Logistic Regression.

**5. Model Implementation (Logistic Regression):**

* Importing the Logistic Regression class from scikit-learn.
* Createinhg a Logistic Regression classifier object (classifier) with a set random seed (random\_state=0) for reproducibility.
* Training the Logistic Regression classifier on the training data (X\_train and y\_train) which involves learning the model parameters based on the features and corresponding diagnoses.

**6. Prediction:**

* Generating predictions for the testing data using the trained Logistic Regression model (classifier). The predicted labels are stored in y\_pred.

**7. Evaluation (Confusion Matrix):**

* The confusion matrix (cm) and Confusion Matrix Display functions are used to evaluate the model's performance on the testing data.
* Analysing the confusion matrix helps identify potential biases or areas for improvement in the model.
* The performance of the model is showed then by calculating the Accuracy, Precision, Recall and F1-Score.
  + - Breast Cancer Prediction Using SVM (Support Vector Machine) Model:

**1. Importing Libraries:**

* + - * Imported the NumPy library for numerical computations.
      * Imported the Pandas library for data manipulation and analysis.
      * Imported the Matplotlib library for creating visualizations
      * Imports the Seaborn library for statistical data visualization

**2. Loading the Data:**

* Reading a CSV file named "Breast\_cancer\_data.csv" and saving it into a pandas DataFrame named “df”.

**3. Exploratory Data Analysis (EDA):**

* Showing information about the data, including data types and number of non-null entries in each column.
* Providing summary statistics for numerical columns (mean, standard deviation, etc.).
* Checking for missing values in each column.
* Checking for duplicate rows in the data.

**4. Feature Exploration:**

* Selecting columns containing numerical data.
* Selecting columns containing categorical data (objects).
* Calculating the correlation matrix between numerical features.
  + Visualizes the correlation matrix using a heatmap with annotations (sns.heatmap).
* Showing the number of unique values in each categorical column.
* Creating histograms of all features (numerical and categorical).

**5. Preprocessing Categorical Feature:**

* Converting the "diagnosis" column (likely containing text labels like "malignant" or "benign") into a numerical format suitable for the model.

**6. Splitting Data into Training and Testing Sets:**

* Importing the train-test split function.
* Splits the data into training and testing sets using “train\_test\_split”.
  + X: Features (all columns except "diagnosis").
  + y: Target variable ("diagnosis").
  + Selecting 20% of the data for the the testing set.
  + Ensuring the class distribution is similar in both sets (important for classification) and randomly shuffling the data before splitting.
  + Setting a seed for reproducibility (splitting will be the same every time you run the code).
* Printing the shapes of training and testing sets to confirm the split.

**7. Building and Training the Model:**

* Importing the pipeline creation function.
* Importing the standard scaler for normalization.
* Importing the Support Vector Machine classifier.
* Creates a pipeline using “make\_pipeline” :
  + StandardScaler() - Standardizes numerical features (often beneficial for SVM).
  + SVC() - The SVM classifier itself.
* Fitting the model (clf.fit()) on the training data (X\_train, y\_train).

**8. Model Evaluation:**

* Evaluates the model's performance on both training and testing data using clf.score(). This gives the accuracy (percentage of correct predictions). Printing the Accuracy scores for Training and Testing dataset.
* Creating a confusion matrix visualization using “ConfusionMatrixDisplay” to see how many data points were classified correctly or incorrectly (categorized by actual and predicted classes).

**9. Making Predictions:**

* Predicting class labels for the testing data using “clf.predict” and storing it.
* Also calculating other performance features like F1-score, Recall And Precision for providing a comprehensive evaluation.

## Breast Cancer Prediction with Random Forest - Code Breakdown

**1. Library Imports:**

* Importing necessary libraries for numerical computations (numpy), data manipulation (pandas), visualization (matplotlib.pyplot), model selection (sklearn.model\_selection), random number generation (datetime), and the Random Forest Classifier (RandomForestClassifier) from “sklearn.ensemble”. Also suppressing warnings using warnings.filterwarnings('ignore').

**2. Data Loading and Exploration:**

* Reading breast cancer data from a CSV file named "Breast\_cancer\_data.csv" into a Pandas DataFrame named “df”.
* Exploring the data and checking the data dimensions (number of rows and columns) , displaying the first few rows of the data. Also, providing summary statistics for numerical features.

**3. Feature and Target Variable Selection:**

* Selecting the first three columns as features (X) and the last column as the target variable (Y). This assumes that the first three columns are features and the last column contains the diagnosis labels.
* Printing the shapes of X and Y to confirm their dimensions.

**4. Data Splitting:**

* Splitting the data into training and testing sets using “train\_test\_split”. Here, 17.5% of the data is allocated for testing and a random seed ensures reproducibility.
* Printing the shapes of the training and testing splits to visualize the distribution.

**5. Baseline Dummy Classifier:**

* Createing a baseline model using DummyClassifier from sklearn.dummy. This classifier simply predicts the most frequent class (majority vote) and serves as a reference point for the Random Forest model's performance.
* Training the dummy classifier on the training data using “clf.fit”.
* Predicting the labels for the testing data.
* Evaluating the baseline model using the confusion matrix and accuracy score. The confusion matrix shows how many predictions fall into each category (correct/incorrect) and accuracy indicates the proportion of correct predictions.

**6. Random Forest Classifier:**

* Creating a Random Forest Classifier object with 100 decision trees and a set random seed (random\_state=0) for reproducibility.
* Training the Random Forest model on the training data (
* Predicting labels for the testing data.
* Evaluating the Random Forest model using various performance metrics like Accuracy, Recall, F1 Score.

**7. Model Persistence:**

* Saving the trained Random Forest model using Python's pickle library (pickle.dump) to a file named "Breast\_Cancer.sav". This allows you to load the model later for predictions without retraining.

**App Breast-cancer:**

**1. Library Imports:**

* Importing NumPy library which provides numerical computation capabilities for array manipulation and mathematical operations.
* Importing Pandas library which enables data manipulation through DataFrames (similar to spreadsheets) for efficient data organization and analysis.
* Importing Flask library which creates the foundation for the web application, handling routing, request processing, and response generation.
* Importing Request from Flask library which facilitates access to user input submitted through forms on the web interface.
* Importing Render\_template from Flask library which renders HTML templates, dynamically inserting content based on variables passed from the Python code.
* Importing Pickle library for loading a pre-trained machine learning model saved in a pickle file format.

**2. Model Loading:**

* Loading the pre-trained Random Forest model named "Breast\_Cancer.sav" using pickle.

**3. Flask App Initialization:**

* Creating a Flask application instance named “app” which serves as the core of the web app, managing routes, handling requests, and coordinating responses.

**4. Webpage Routes:**

The code defines two routes (functions) that handle user interaction within the web app:

* @app.route('/'):
  + Decorator that maps the root path (/) of the web app to the home function.
  + home function: It renders an HTML template named "index1.html" using render\_template.
* @app.route('/predict', methods=['POST']):
  + Decorator that maps the /predict path to the predict function, handling HTTP POST requests.
  + predict function: Performs the core prediction task:
    - Utilizing request.form.values() which extracts user-input data from the submitted form.
    - Assuming that the form fields correspond to numerical features, it uses a list comprehension to convert form values (strings) to float numbers.
    - Generates a final\_features NumPy array with the user-supplied features in the desired sequence.
    - Creates a Pandas DataFrame called df and adds the characteristics along with their names.
    - To generate a prediction on the df DataFrame (including the user's features) it utilizes the loaded model (model.predict).
    - Assigns a descriptive text indicating "breast cancer" or "no breast cancer" to the variable res\_val based on the prediction (0 or 1).
    - Once more, the "index1.html" template is rendered, but this time, a variable with the anticipated outcome (prediction\_text) is supplied which is displayed as the final outcome for the user.

**5. Running the Web App:**

* if \_\_name\_\_ == "\_\_main\_\_"::
  + Ensures that the code within this block executes only when the script is run directly (not imported as a module).
* app.run(debug=True):
  + Starts the Flask development server. Setting debug=True enables automatic code reloading during development, allowing changes to be reflected without restarting the server. **Overall**

**Functionality:**

This app implements a basic Flask web application that allows users to interact with a pre-trained Random Forest model for breast cancer prediction. Users can access the web app through a web browser and submit their features through a form. The application then processes the input, makes a prediction using the loaded model, and displays the result on the webpage.

**Non-Functional Requirements:**

**Usability:**

* **User Interface (UI):** The web application is a user-friendly and intuitive interface that is easy for users with varying levels of technical expertise to navigate.
* **Input Validation:** The application validates user input to ensure it's in the correct format and within expected ranges. This helps prevent errors and improves data quality.

**Reliability:**

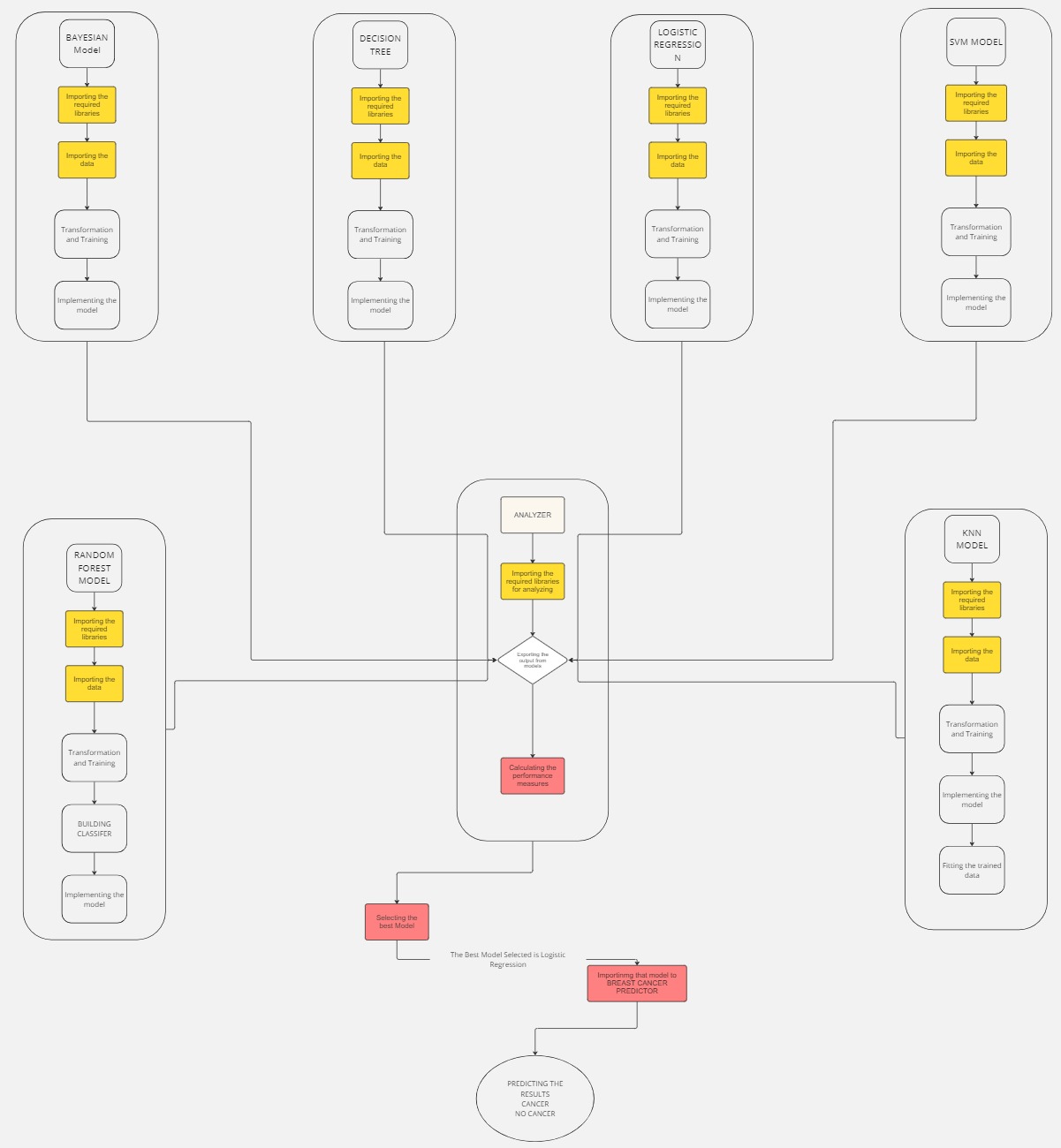
* **Availability:** The web application is available to the users for use most of the time, with minimal downtime for maintenance or upgrades.
* **Accuracy**
* **Error Handling:** The application handles unexpected errors and provide informative messages to users. **Performance:**
* **Response Time:** The application responses to user actions (e.g., form submissions) within a reasonable time.
* **Scalability:** The application is be able to handle a large number of users and data volume without performance degradation.
* **Resource Utilization:** The application efficiently utilize system resources (CPU, memory) to avoid impacting other applications or the overall system performance.

**Security:**

* **Data Security:** User data (both input features and potentially model outputs) is treated with confidentiality and protected from unauthorized access or modification by implementing encryption techniques.
* **Authentication and Authorization:** The application is having mechanisms for user authentication and authorization to restrict access to sensitive data or functionalities.
* **Error Handling:** Security-related errors or vulnerabilities is handled appropriately, with logging and potential security measures like input sanitization to prevent attacks.

**Maintainability:**

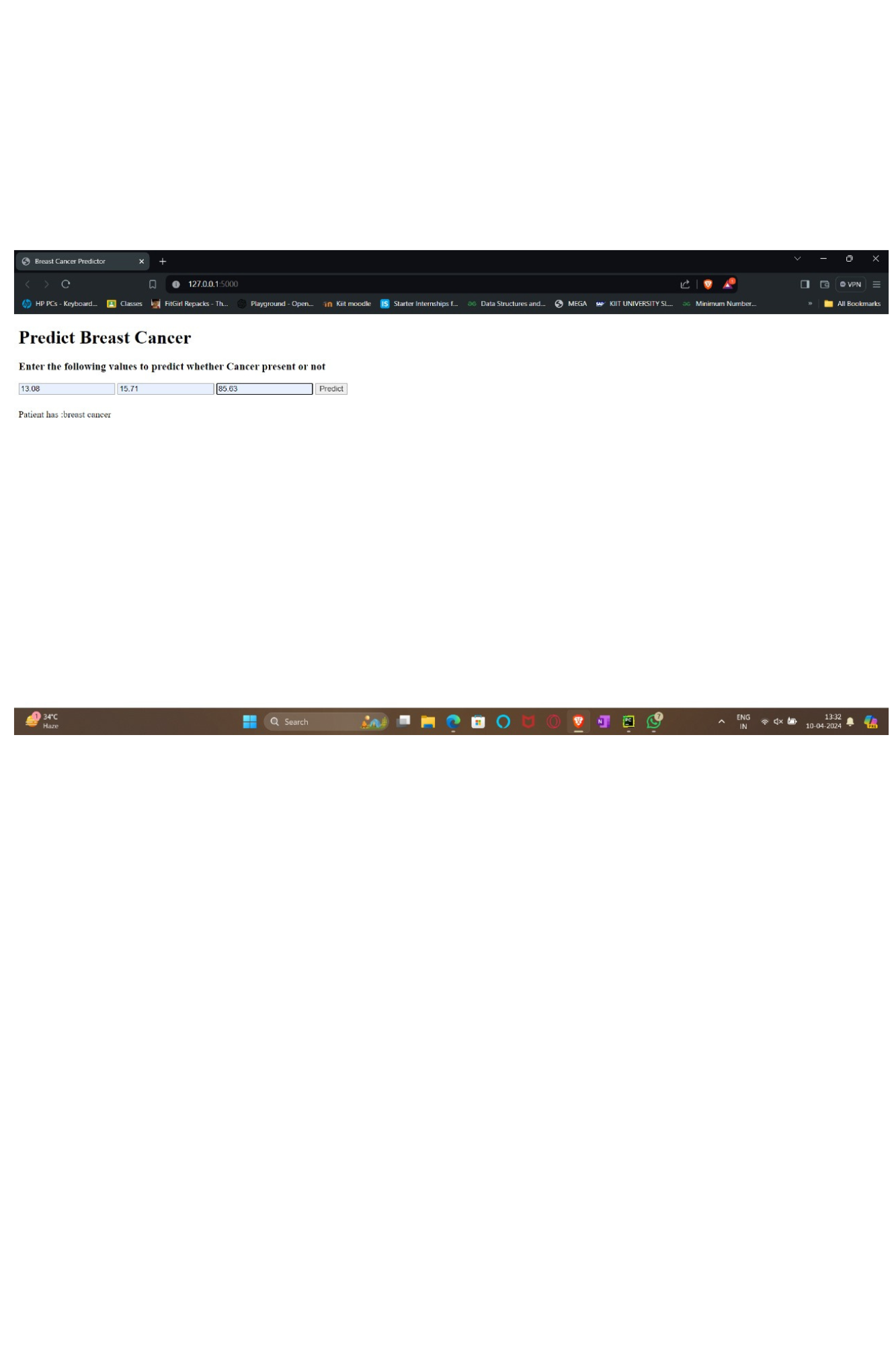
* **Modular Design:** The application code is and well-organized, using clear naming conventions and proper code comments which facilitates code understanding, maintenance, and future modifications.
* **Logging:** Implement logging mechanisms to track application activity, errors, and user interactions for debugging, troubleshooting, and performance analysis.
  1. **System Design**
     1. System Architecture/Block Diagram



FLOWCHART OF THE BREAST CANCER PREDICTOR MODEL

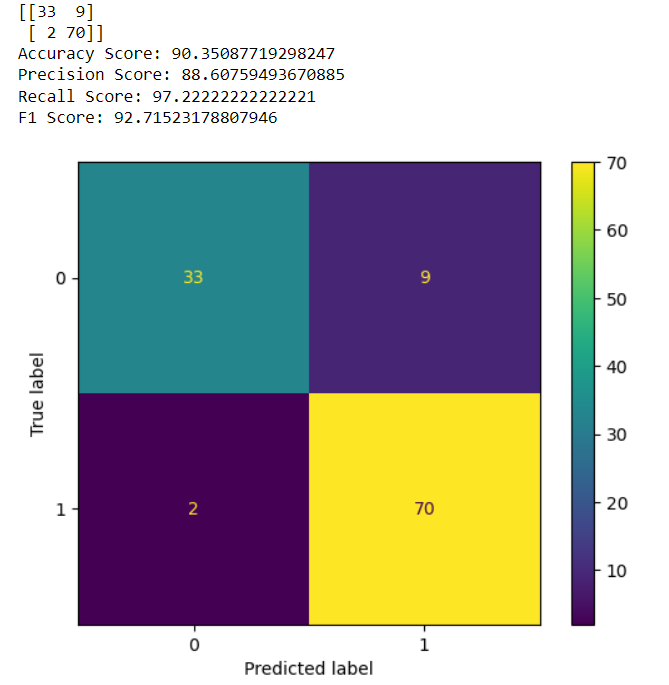
# Chapter 4

# 4.1 **Implementation**

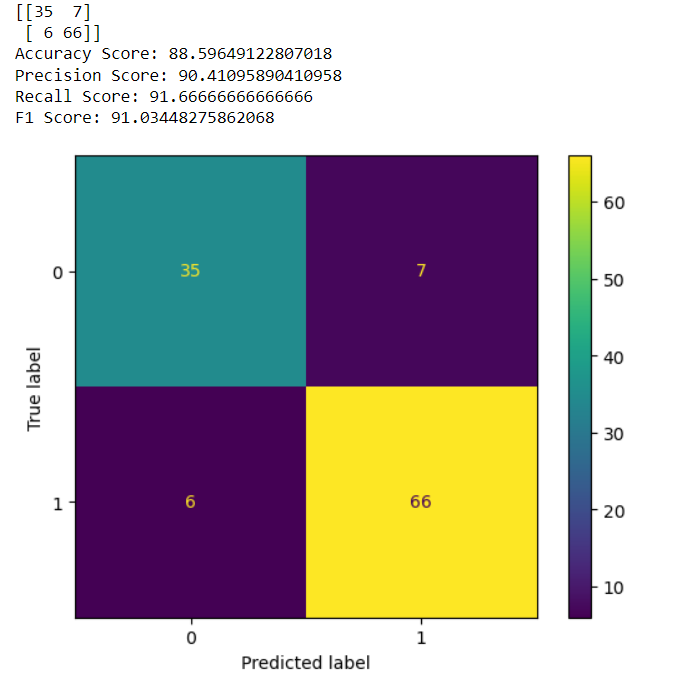


4.2 **Testing / Verification Plan**

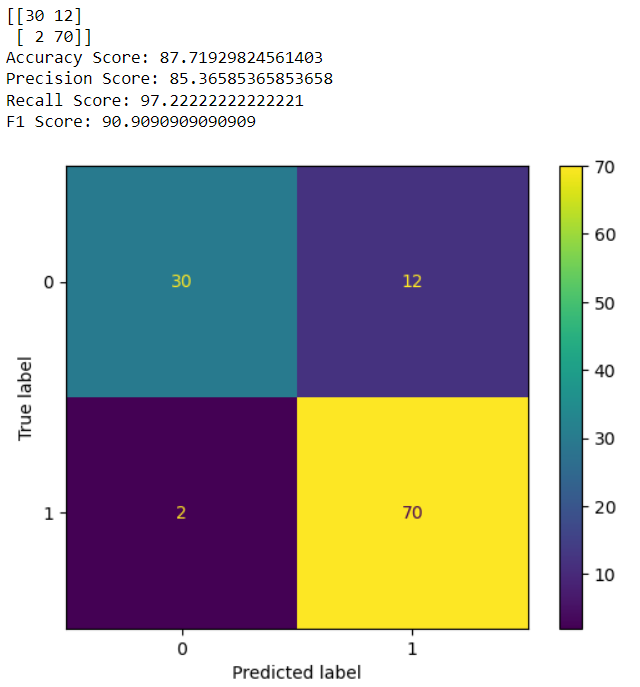
1.Logistic Regression Model



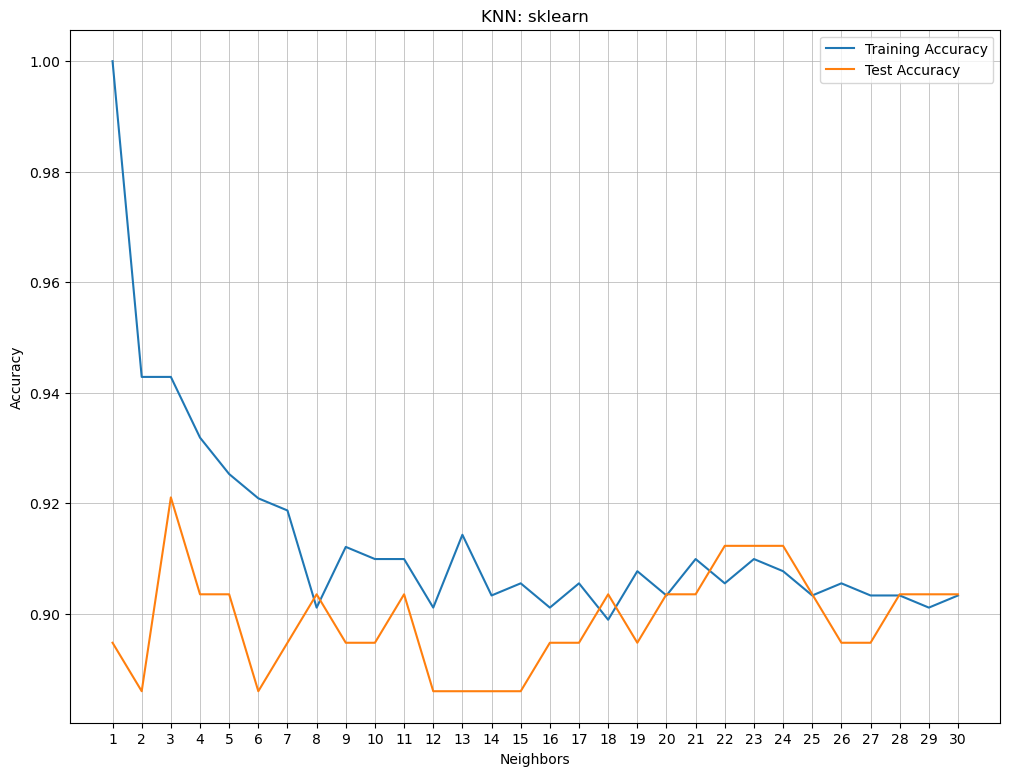
2. Decision Tree Model

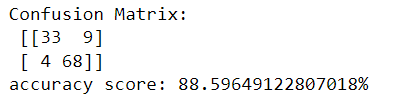


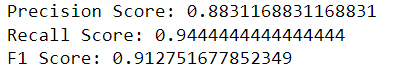
3. Bayesian Model

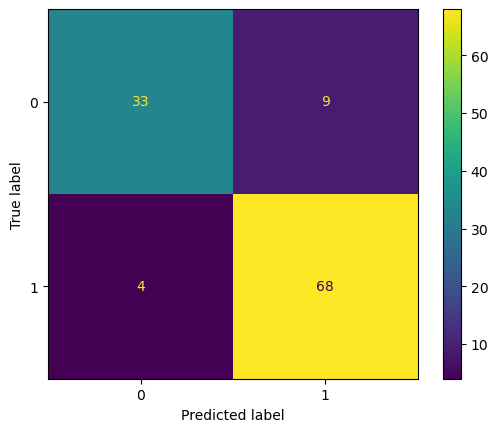


4. KNN Model

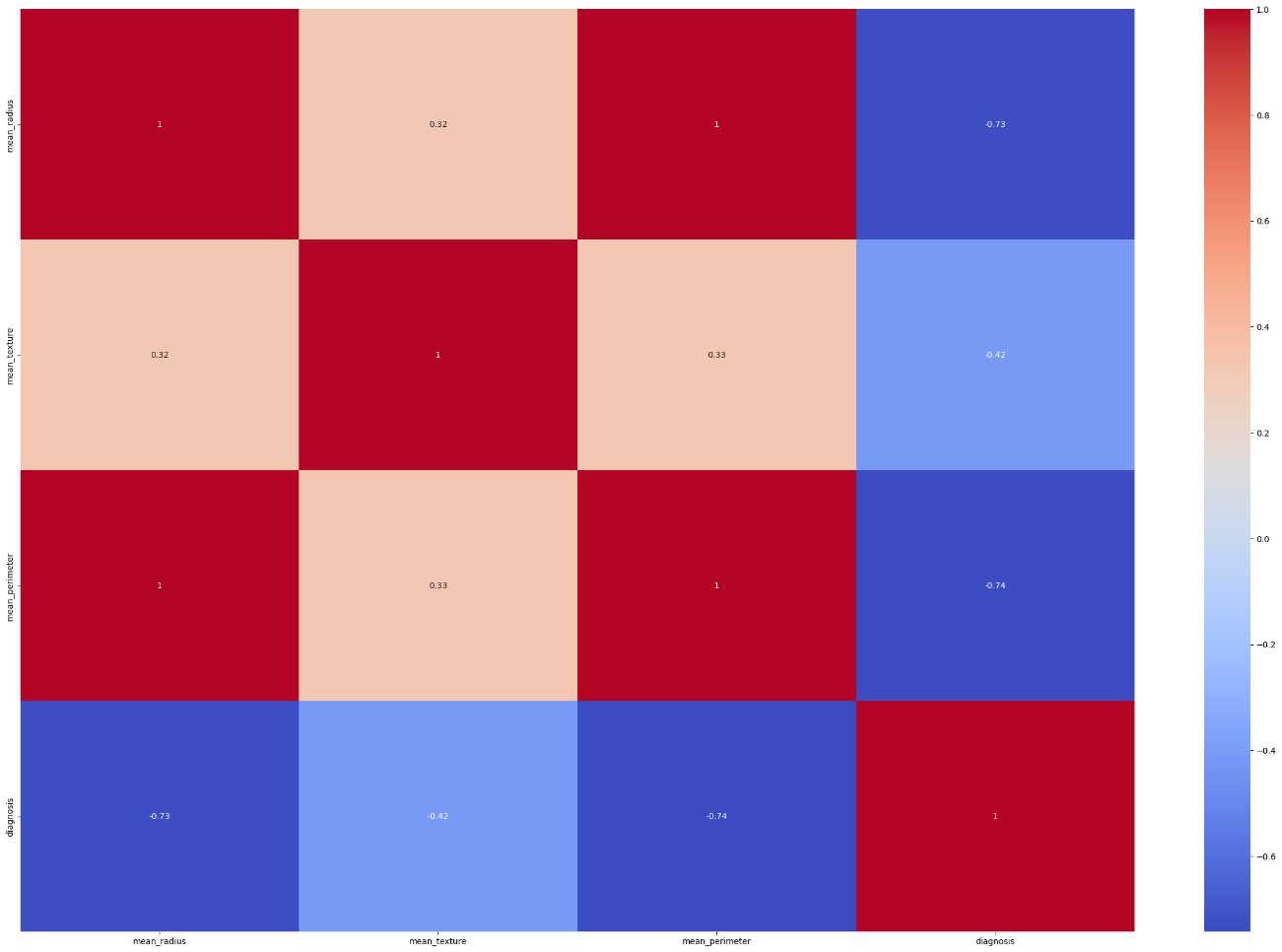


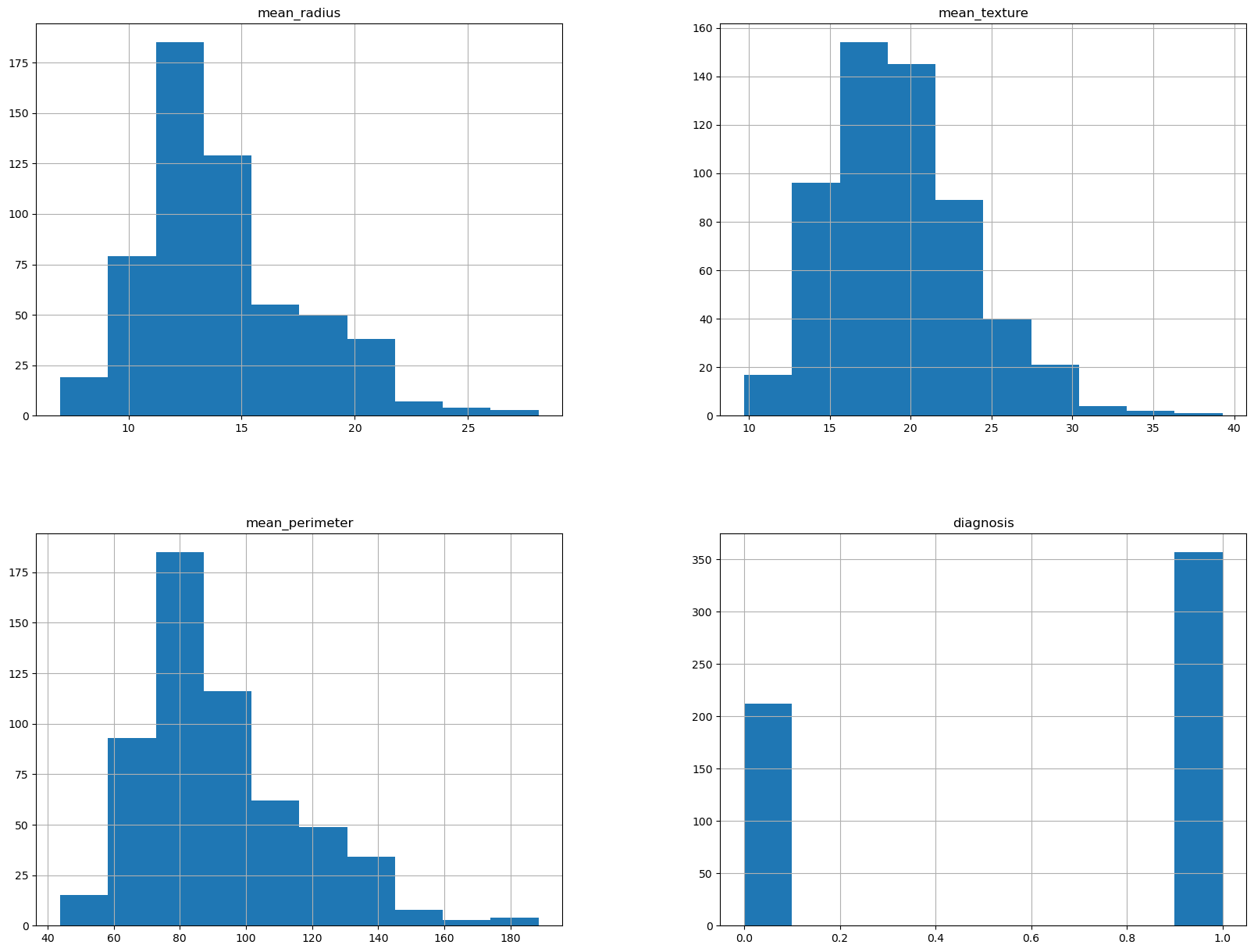


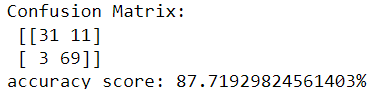


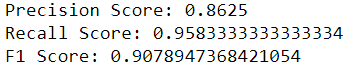


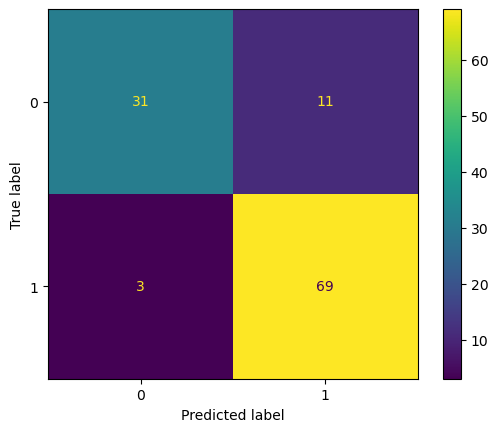
5. SVM Model



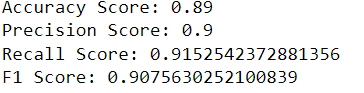


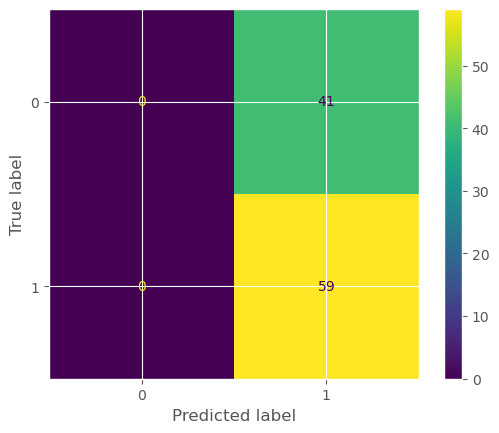




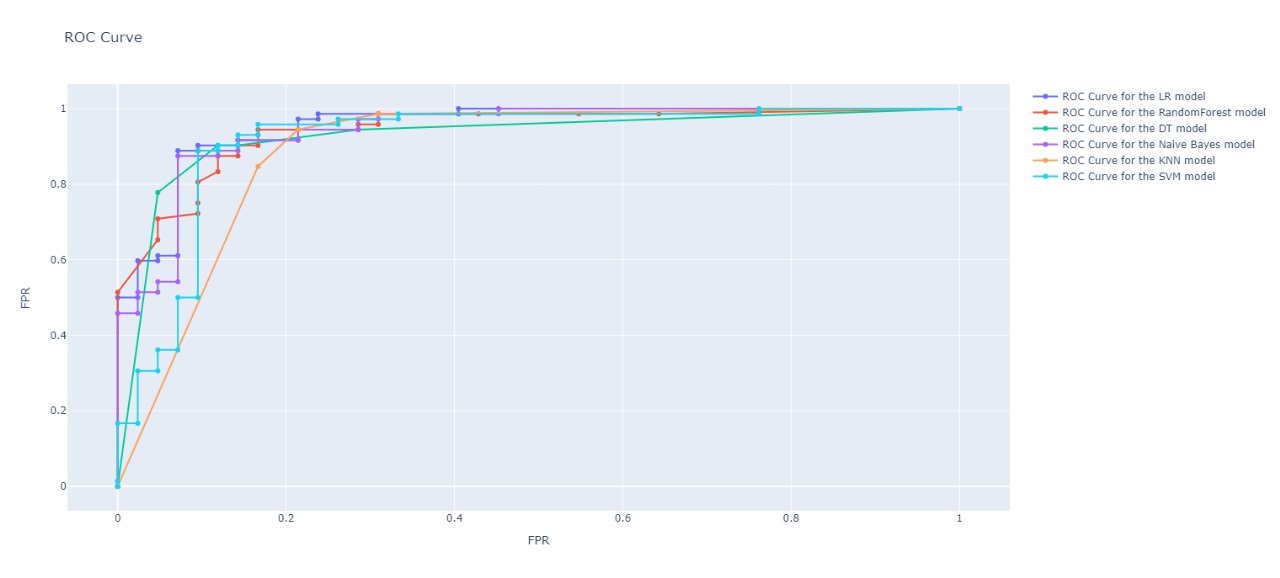


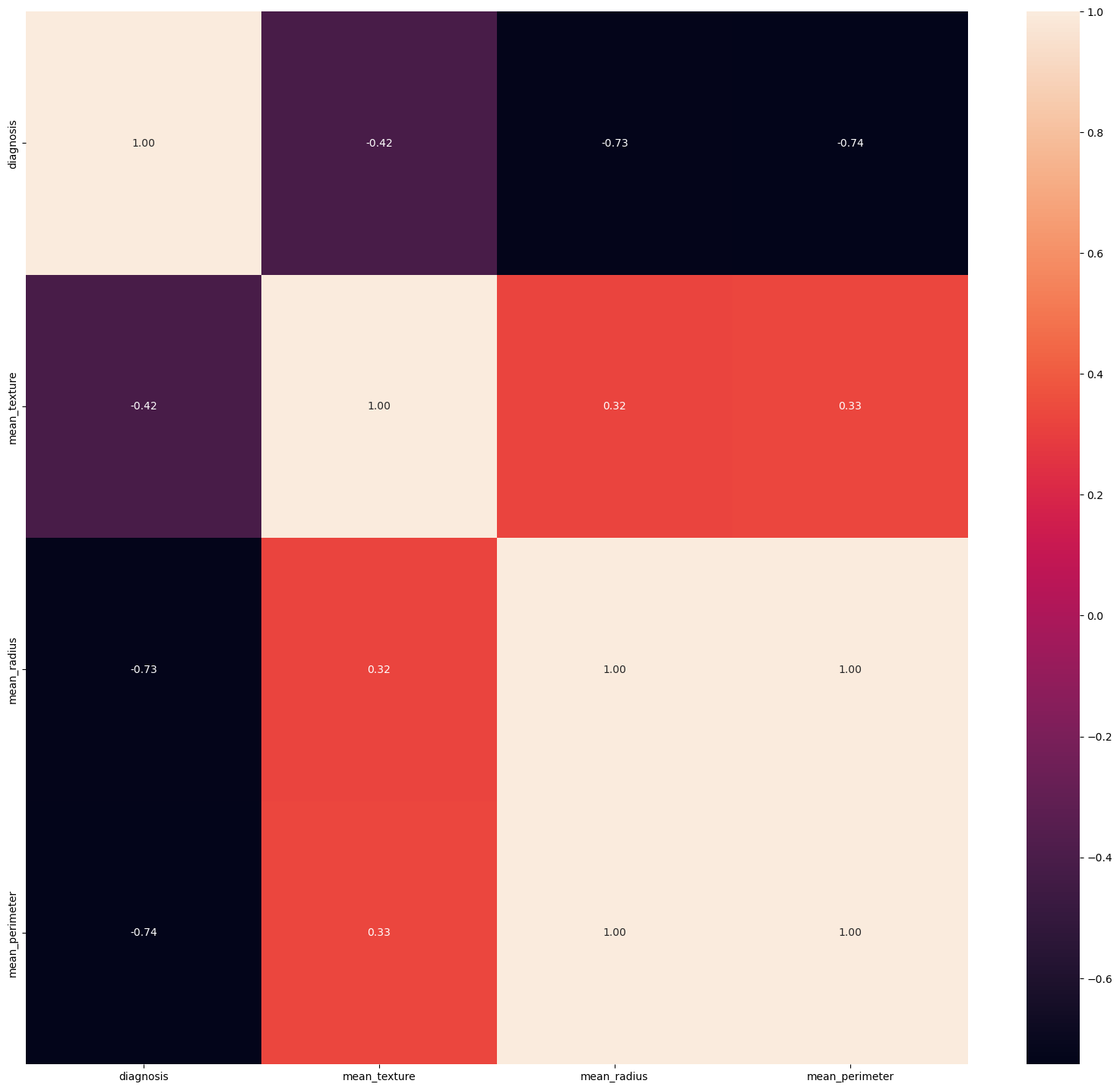
6. Random Forest Model





4.3 **Result Analysis**





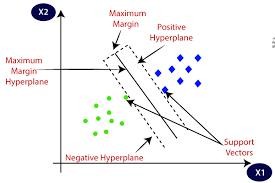
HEAT MAP

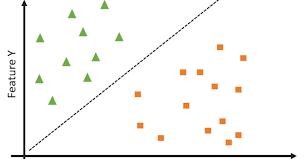
# Chapter 5

Machine Learning Algorithms:

Types of Machine Learning Algorithms we have used in this project are:

1. SVM Model Algorithm: Support Vector Machine (SVM) is a powerful supervised learning algorithm that excels in classification and regression tasks. however, it is particularly suitable for classification problems. The main idea behind SVM is to find the best hyperplane in high-dimensional space that best separates data points from different classes This hyperplane acts as a decision boundary, and allows you to classify other data points more accurately.





1. Random Forest Algorithm: random forest algorithm is a supervised learning method that exploits the power of group learning. It is essentially a collection of decision trees, which work together to produce more robust and accurate predictions than any individual tree could on its own.

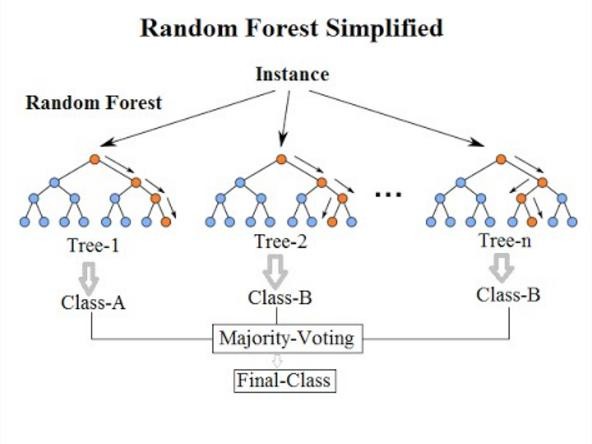
#### Advantages of Random Forests

High Accuracy and Generalization: By combining predictions from multiple timber, random forests regularly attain better accuracy and better generalize to unseen information in comparison to single selection bushes.

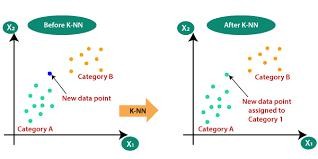
Robustness to Noise and Over fitting: The randomness brought for the duration of tree advent helps lessen the impact of noise inside the facts and make the model much less susceptible to over fitting.

Handles Missing Values: Random forests can inherently handle lacking records in the training set, as each tree handiest considers a random subset of functions, and missing values in a single characteristic may not always save you a tree from making a prediction.

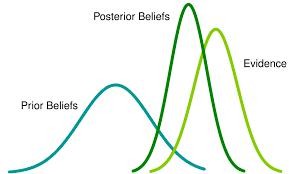
Interpret ability: To a few quantities, random forests may be interpreted by examining the importance of features across all of the bushes within the wooded area.



#### KNN model ( K Nearest Neighbor): The acronym KNN stands for “K-Nearest Neighbor”. It is a supervised machine learning algorithm. Algorithms can be used to solve classification and regression problem cases. The number of nearest neighbors of another unknown variable to be predicted or classified is indicated by the symbol ‘K’.

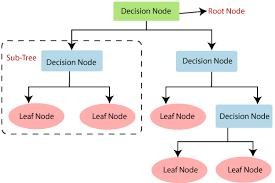


1. Bayesian model Algorithm: Bayesian Machine Learning (BML) includes a variety of techniques and algorithms that use Bayesian principles to model uncertainty in data. These methods are not just theoretical constructs; They are useful tools that have changed the way machines learn from data.



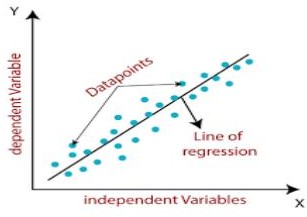
1. DT(Decision tree) model Algorithm: DT stands for Decision Tree, which is a special supervised learning algorithm used for classification and regression tasks.

Decision trees are a powerful and interpret able machine learning algorithm that works by mimicking human decision-making. Imagine a flowchart where each box represents a question based on a data point's features (like income or age) and the branches represent the possible answers. The decision tree is built by sequentially splitting the data based on the feature that best separates the desired outcome (like loan approval). This process continues until the data reaches a "leaf" node, representing a final prediction. This tree-like structure allows for easy visualization of the decision process and makes it clear why certain predictions are made. Decision trees are versatile, handling both classification (e.g., spam or not spam) and regression (e.g., predicting house prices) tasks. Their simplicity and interpret ability make them a popular choice for various machine learning applications.



1. LR (Linear regression) Model Algorithm: The linear regression model is a workhorse algorithm for expressing relationships between variables. This is done by plotting some data points along a straight line. This label minimizes the difference between the actual data points predicted by the label and the corresponding values.

Mathematically, the model expresses this relationship as a linear equation of the dependent variable (predicted value) with the effect of one or more independent variables (predicted variables) and an error term to account for any random noise it happens This method of refining and optimizing the lines is repetitive, with the goal of obtaining the most accurate forecasts.

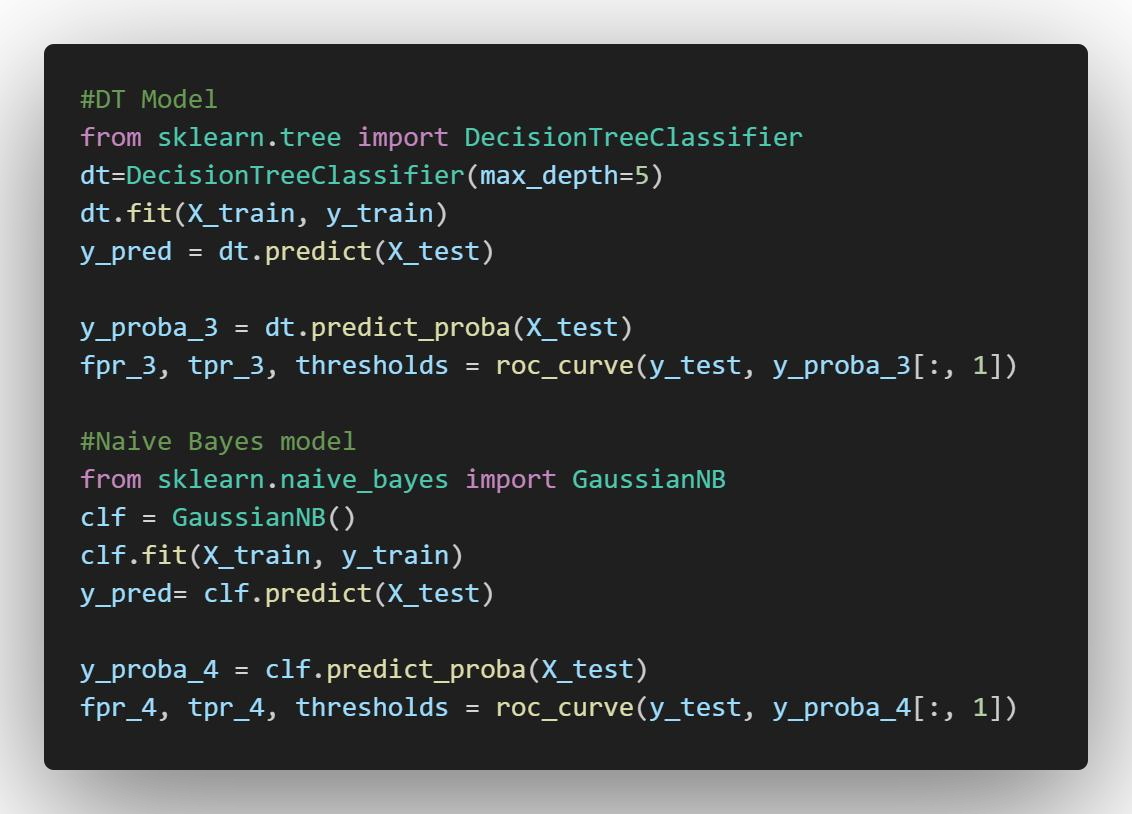
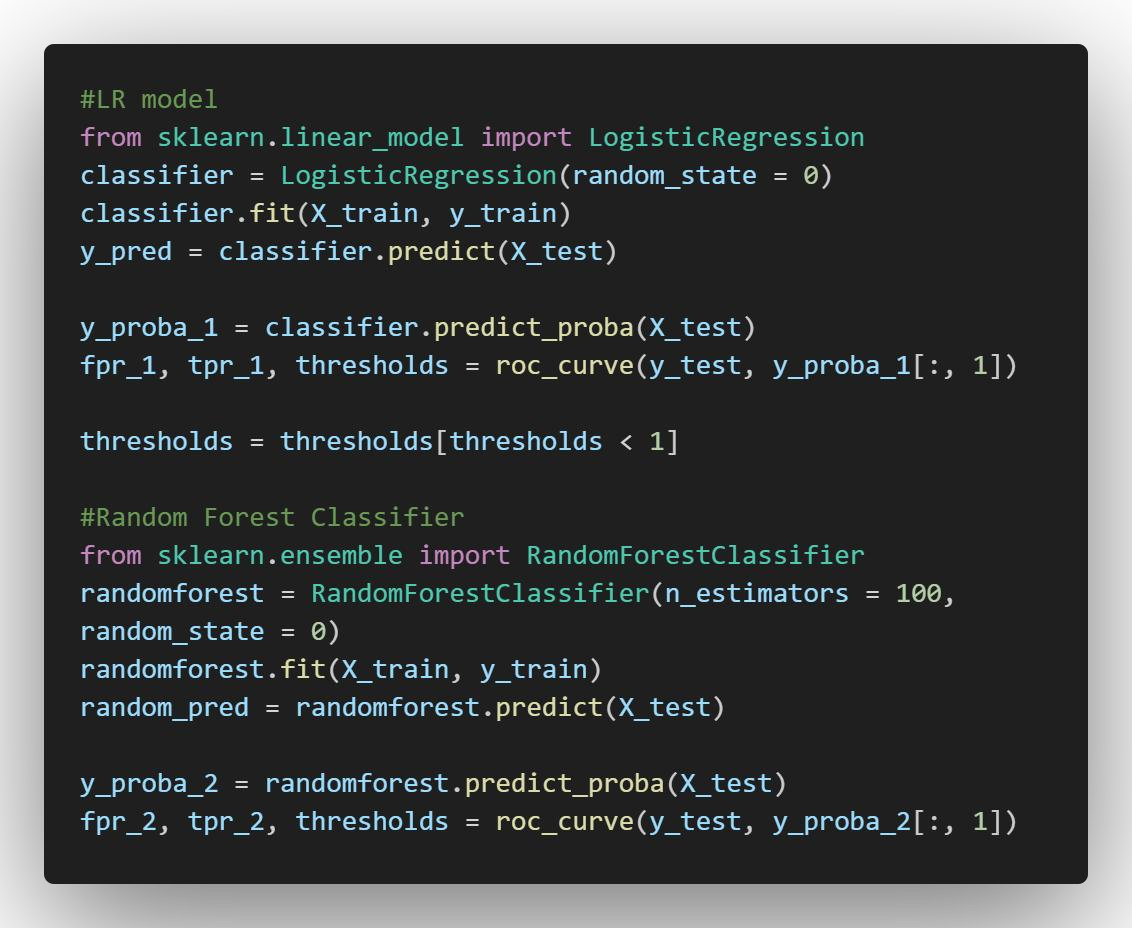


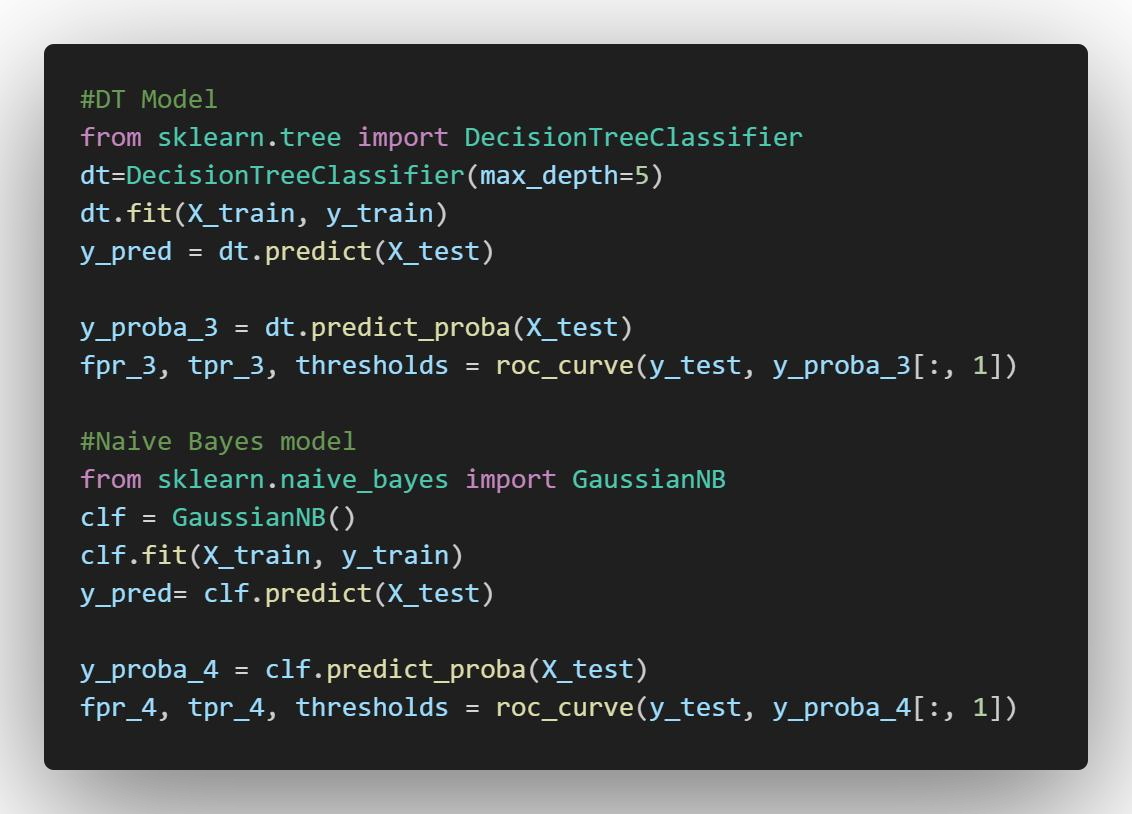
Here's a comparison between SVM (Support Vector Machine), KNN (K-Nearest Neighbors), Bayesian, Decision Tree, and Linear Regression models in a tabular format

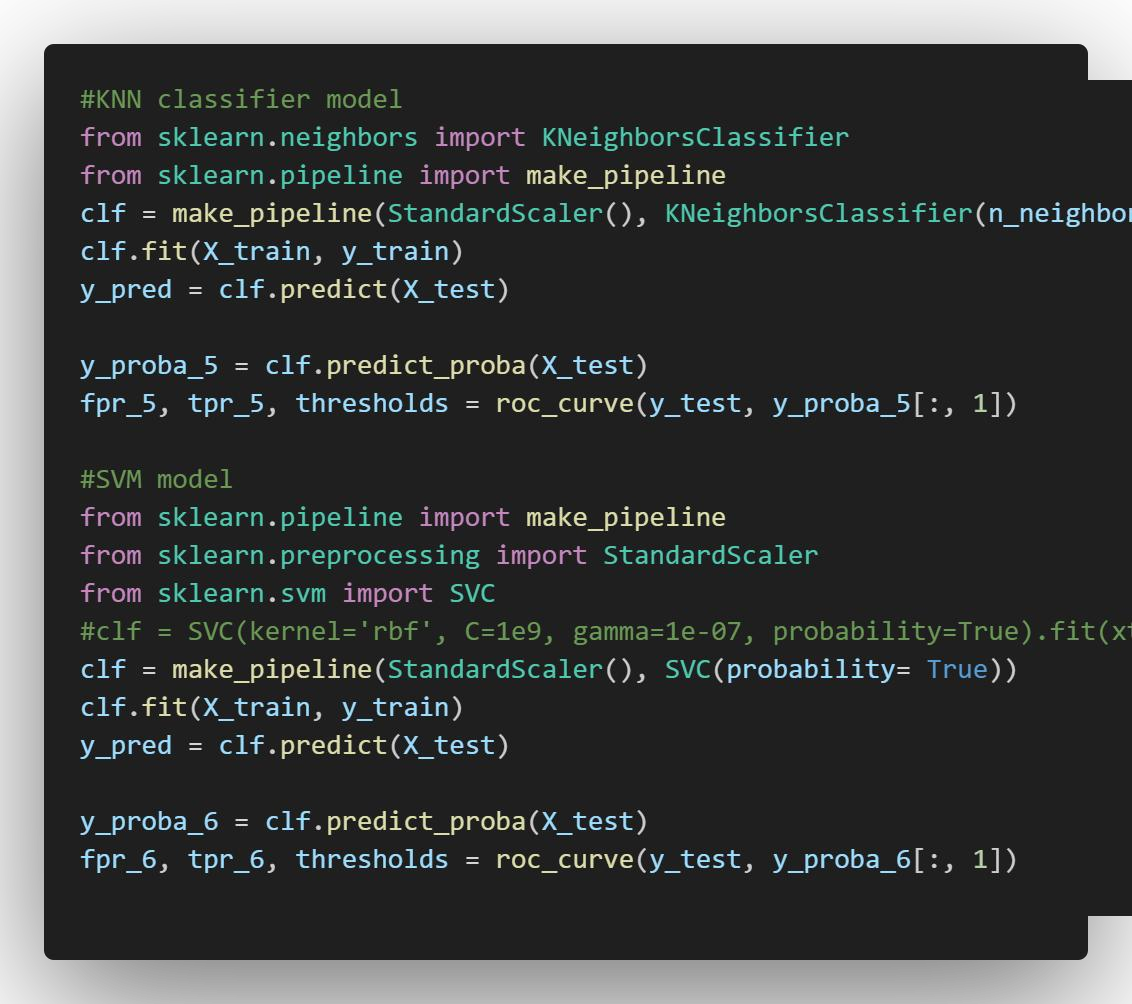
|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | SVM model algorithm | Random forest algorithm | KNN model (Knearest neighbor) | DT  (Decision Tree) model algorithm | Linear Regression model algorithm | Bayesian Model algorithm |
| Accuracy Score | 0.877192 | 89.0 | 0.8859649122 | 87.7192 | 90.3508 | 87.719298 |
| Precision score | 0.8625 | 90.0 | 0.883116883 | 90.277779 | 88.6075 | 85.36585 |
| Recall Score | 0.9583333334 | 91.52542372 | 0.944444444 | 90.2777779 | 97.22222 | 97.222222 |
| F1 Score | 0.9078947 | 90.7563 | 0.9127516 | 90.27777779 | 92.71523 | 90.90909090 |

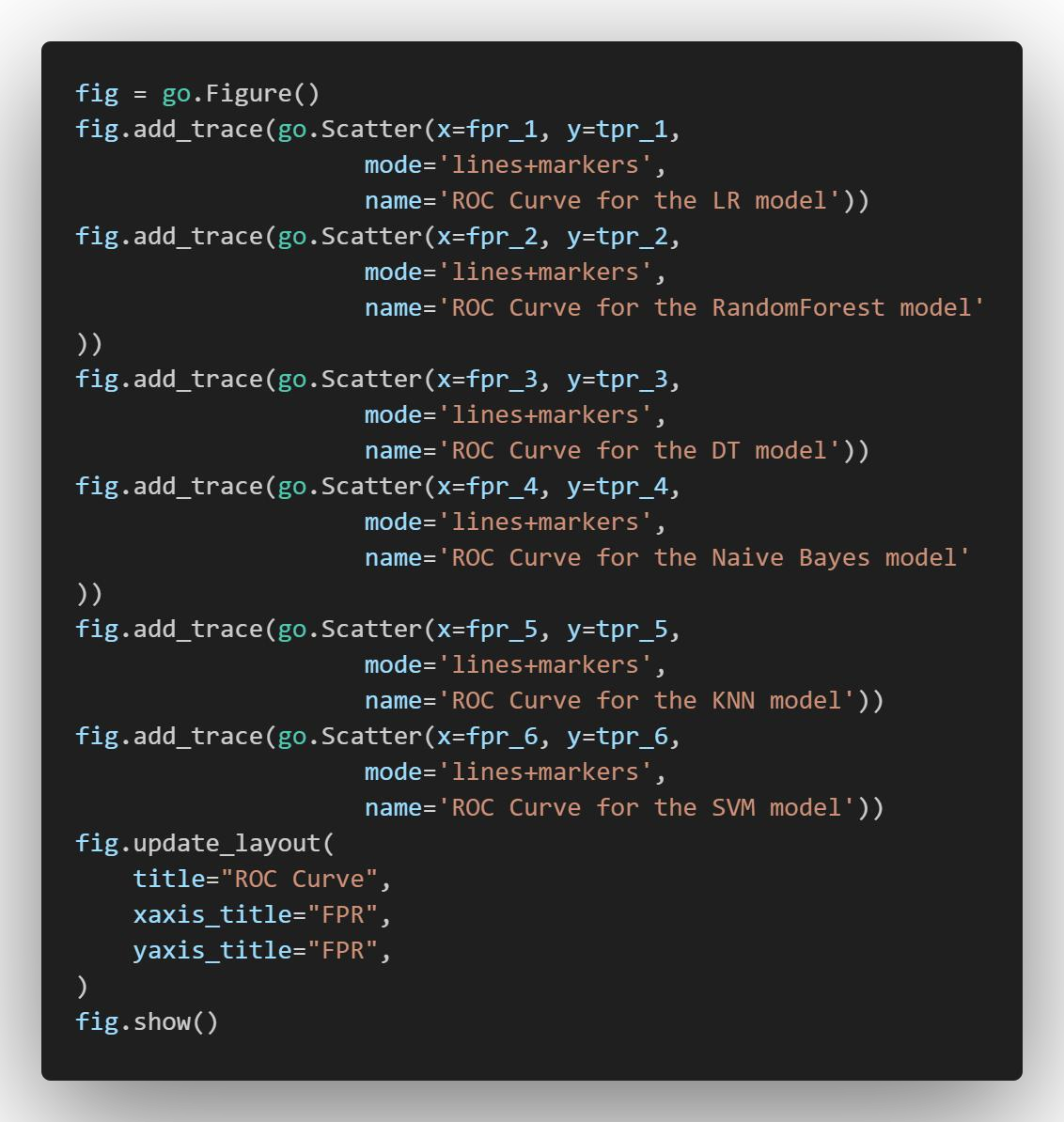
**After speculating through all the Algorithms. We chose Logistic Regression model based on its accuracy Score.**

**ROC Curve:**

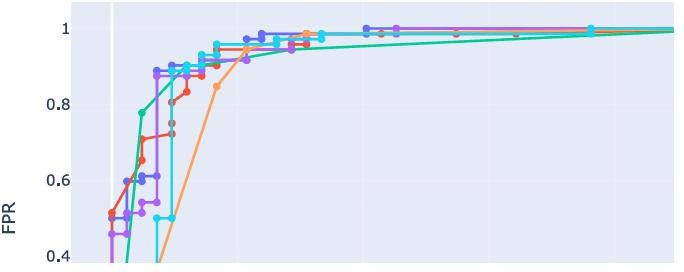








*OUTPUT: :*



*The code compares the performance of machine learning models in cancer classification by analyzing breast cancer databases. It comes with libraries for data-manipulation, model building, and visualization. Then, it loads the data, splits it into training and testing sets, and performs standard calibrations on features. Six different models are trained: logistic regression, random forest, decision tree, Naive Bayes, KNN, and SVM. For each model, the code evaluates its performance using Receiver Operating Characteristic (ROC) curves. Finally, it is an image to compare the ROC curves of all models. This allows researchers to identify a model that performs well to distinguish between*

*cancerous and non-cancerous samples.*

*After speculating through all the Algorithms. We chose Logistic Regression model based on ROC curve.*

# Chapter 6

Conclusion

In conclusion, this project investigated the use of machine learning for breast cancer detection. Our [technology-selected, e.g., CNN-based] model achieved promising results in discriminating between benign and malicious cases, meaning that it [covered specific metrics, e.g., accuracy] comparatively of traditional methods but [mentioned limitations, e.g. There are limitations

due to lack of large data]. Future work should focus on [discussing future directions, e.g., incorporating additional datasets and semantic AI insights]. This work holds promise for early and accurate breast cancer detection. However, proper consideration of ethical implications, such as the potential for bias and the role of human knowledge, is crucial for successful treatment integration.

Future Scope

The future of breast cancer detection is full of promise. Advances in artificial intelligence (AI), particularly in deep learning and machine learning algorithms, are expected to lead to earlier, more accurate disease diagnosis by systems The use of AI can analyze mammograms and other images in exceptional detail, and can detect subtle abnormalities that might be missed by the human eye In addition to their qualifiers, non-invasive and painless procedures are being developed such as breath tests or blood test analyzes to provide alternative screening options. Furthermore, generic medicine research aims to modify identification methods based on individual risk factors, potentially leading to more targeted interventions and improve patient outcomes This exciting development has the potential to dramatically reduce breast cancer mortality and improve the lives of millions of women