Khandesh Education Society's

Pratap College, Amalner

(AUTONOMOUS)

Affiliated To North Maharashtra University, Jalgaon

Collage code - 120017

YEAR-2025-2026

A

PROJECT REPORT

ON

"BREAST CANCER PREDICTION SYSTEM WITH DATA INTELLIGENCE"

Department Of Computer Management

AMALNER

SUBMITTED BY

RUCHIKA KUMBHAR

Bachelor of Computer Application

PRN No: 2023096800010987

PROJECT GUIDE

Mr.BHAVESH SALUNKHE.

CERTIFICATE

This is to certify that the Project Report entitled "**Breast Cancer Prediction System with Data Intelligence**" for KES's Pratap College, Amalner. Submitted by **Miss. Ruchika Kumbhar** for the Partial fulfillment of the Bachelor of Computer Application Pratap College, Amalner (AUTONOMOUS) embodies the record of original work carried by her under my supervision.

H.O.D

Mrs. Dr. Varsha Pathak

Examiner-I Examiner-II

Date:

Place: Amalner

ACKNOWLEDGEMENT

Presentation of inspiration and motivation have always played a major key or role in the success.

I express my sincere thanks to all those who have provided us valuable Guidance towards the completion of this system as a part of syllabus of "Breast Cancer Prediction System with Data Intelligence". I express my sincere gratitude towards Dr. H. D. Jadhav of Pratap College Amalner, who have provided us valuable guidance in successful completion of Project.

I express my guidance Prof. Mr. Bhavesh Salunkhe Mam for giving me such a Guidance.

Lastly, I also thanks to all my friends who have their moral support for successful completion of this project.

I am extremely thankful to all whose valuable support on completion of my project in it 's presently.

Thanking You,

Miss. Ruchika Kumbhar

SUBMISSION

I Miss. **Ruchika Kumbhar**, Enrolment No.

2023096800010987

Student of TYBCA for academic year 2025-2026 humbly submit that I have completed from time to time the project work on skills and study as per the guidance of **Mr. Bhavesh Salunkhe.**

The following project work has not copied or it's any appreciable part from any other in convenient of academic ethics.

Signature of Student

Name:

Date:

INDEX TABLE

Introduction

Breast cancer is one of the most common and life-threatening diseases affecting women worldwide. Early detection and timely diagnosis play a crucial role in improving survival rates and treatment outcomes. However, manual diagnosis based solely on medical reports can sometimes lead to errors or delayed identification. To overcome this challenge, technology-driven solutions like machine learning and data visualization can significantly assist doctors and patients in making faster and more accurate decisions.

This project focuses on developing a **Breast Cancer Prediction Website** that allows users or healthcare professionals to input specific medical parameters such as cell size, texture, and other diagnostic features to predict the likelihood of breast cancer. The system utilizes a trained machine learning model that analyzes the entered data and provides a probability-based prediction, helping doctors in preliminary assessment.

To enhance user engagement and awareness, the website integrates a **Chatbot** that interacts with users to provide valuable information related to breast cancer symptoms, remedies, preventive measures, and general awareness. This feature not only improves user experience but also spreads medical knowledge in a simplified and accessible way. Additionally, the project includes an interactive Dashboard that visualizes historical and current breast cancer trends using graphical insights. This dashboard helps medical professionals, researchers, and policymakers to understand the rise, decline, or distribution of breast cancer cases over time.

Overall, this project aims to blend machine learning, web development, and data visualization to create a reliable, user-friendly, and informative platform that supports early diagnosis and awareness of breast cancer, ultimately contributing to better healthcare outcomes.

Need of the Project

1. Background Study

Breast cancer remains one of the most prevalent cancers worldwide, accounting for a significant percentage of cancer-related deaths among women. According to various medical studies, early detection greatly increases the chances of successful treatment. Traditionally, doctors rely on diagnostic tests and medical imaging to identify cancerous cells, but manual analysis can sometimes be time-consuming and prone to human error. With the advancement of technology, machine learning has proven to be a powerful tool in analyzing medical data and predicting disease outcomes with higher accuracy. By integrating these technologies into an accessible online platform, we can support healthcare professionals in making faster and more reliable predictions.

2. Problem Identification

Despite technological progress, there is still a gap between medical data analysis and real-time prediction tools available for doctors and patients. Many hospitals and diagnostic centers, especially in rural or underdeveloped areas, lack advanced systems to assist in early breast cancer detection. Manual interpretation of reports can lead to delays, misdiagnosis, or late-stage identification. Furthermore, general awareness about breast cancer symptoms, prevention, and lifestyle management is still limited among the population. There is also a lack of centralized systems to visualize and understand the trends or patterns of breast cancer cases over time.

3. How to Solve It

To address these issues, this project proposes a web-based Breast Cancer Prediction System integrated with Machine Learning, Chatbot assistance, and a Data Visualization Dashboard. The Machine Learning model analyzes diagnostic parameters entered by the user and predicts the chances of breast cancer, assisting doctors in quick and accurate decision making.

The Chatbot serves as an interactive guide, educating users about symptoms, preventive measures, and remedies, thus promoting awareness. The Dashboard displays historical and current breast cancer trends using visual insights such as graphs and charts, helping researchers and medical professionals track patterns effectively.

By combining these components, the system provides an efficient, user-friendly, and informative solution that bridges the gap between data analysis, awareness, and early detection contributing to better healthcare accessibility and outcomes.

Feasibility Study

A feasibility study is a crucial step in the software development process that helps determine whether a project idea can be successfully implemented within the available technical, financial, and operational constraints. It provides an in-depth analysis of the project's strengths, limitations, opportunities, and challenges. The main objective of a feasibility study is to evaluate whether the proposed system is practical, cost-effective, and beneficial for the intended users.

For the Breast Cancer Prediction Website, the feasibility study played an important role in identifying how machine learning and chatbot technologies can be effectively combined to assist medical professionals and users in predicting the chances of breast cancer. It ensured that the system could be built using available open-source tools, reliable datasets, and existing technical knowledge.

1. Survey / Interviews / Visit

- **Medical Professionals and Students:**

They emphasized the importance of early detection tools that could support diagnosis and provide an initial prediction to assist in clinical analysis. Many suggested including a feature that allows input of multiple tumor-related parameters like radius, texture, and smoothness for more accurate predictions.

They also recommended integrating a chatbot to answer general questions about symptoms, causes, and preventive care so that patients could stay informed.

- **General Users / Public:**

Most users expressed interest in an easy-to-use online platform where they could get a preliminary idea of breast cancer risk using basic medical values or test results. We found the idea of having an interactive chatbot useful, especially for gaining awareness about remedies and lifestyle habits.

No physical visits to hospitals or labs were conducted due to limited access

and ethical restrictions, but the responses collected from the online community and feedback sessions helped shape the functional and design requirements of the system.

These surveys and interactions confirmed that there is a real need for a machine learning-based predictive system that can not only assist doctors but also spread awareness among the general public.

2. Requirement Analysis

Requirement analysis is the process of gathering and understanding the needs of the system what functions it should perform and how users will interact with it. This stage ensures that the final product meets both technical and user expectations.

The requirements for the Breast Cancer Prediction Website were categorized as follows:

- **Software Requirements:** Python (for model training), Flask or Django (for backend integration), HTML/CSS/JavaScript (for frontend), and an IDE such as VS Code or PyCharm for development.

- **Data Requirements:** A breast cancer dataset (such as the Wisconsin Breast Cancer Dataset) for model training and validation.

- **Functional Requirements:**

These define what the system should do.

The system should allow users to enter tumor-related numerical parameters (such as mean radius, mean texture, and mean smoothness).
The input data should be processed and sent to the trained machine learning model for prediction. The model should predict the chances or probability of breast cancer (e.g., high or low risk, benign or malignant).

The chatbot should be able to answer questions related to breast cancer, such as symptoms, remedies, causes, and awareness tips. The system should display the output in a clear, easy-to-understand format for both medical and non-medical users.

- **Non-Functional Requirements:**

These specify how the system should perform.

i. Accuracy: The machine learning model should provide a reliable prediction based on the dataset used for training.

ii. Usability: The website should have a simple and user-friendly interface so that even non-technical users can operate it easily.

iii. Scalability: The system should be capable of handling additional input features or models in the future.

iv. Performance: The prediction process should be quick and responsive.

v. Security: User data should be handled securely and confidentially, especially if connected to sensitive health information.

vi. Compatibility: The system should be compatible with commonly used browsers and devices.

The requirement analysis confirmed that the system could be implemented efficiently using technologies like Python, Flask, HTML, CSS, and Machine Learning libraries such as Scikit-learn. The model would be trained using the Breast Cancer Wisconsin Dataset from Kaggle, which provides high-quality labeled data suitable for supervised learning algorithms.

3. Types of Feasibility Study

A comprehensive feasibility analysis involves examining several different types of feasibility to ensure that the project is practical and can be successfully executed.

- **Technical Feasibility**

This aspect determines whether the current technology and tools are sufficient to develop the proposed system. For this project, all the technical resources were readily available.

i. Frontend: HTML, CSS, and JavaScript were used for the user interface.

ii. Backend: Flask framework in Python handled server-side communication.

iii. Machine Learning Model: Trained using Scikit-learn and Kaggle dataset.

iv. Chatbot Integration: Implemented using rule-based logic and API responses.

All these technologies are open-source and compatible, which makes the system technically feasible for development.

- **Economic Feasibility**

This study examines the cost-effectiveness of the project. Since the entire development utilized free and open-source resources such as Python, Kaggle datasets, Flask, and ChatGPT for guidance, the cost of development was minimal. The system does not require any paid APIs or commercial software, making it economically viable for student-level or research projects.

- **Operational Feasibility**

Operational feasibility assesses how well the system will function in the real world and whether it will meet user needs. The breast cancer prediction system is designed to operate smoothly, requiring only simple input parameters to generate a result. It assists medical professionals in preliminary analysis and helps users gain awareness about their health. Thus, it is operationally efficient and valuable.

- **Time Feasibility**

Time feasibility evaluates whether the project can be completed within the given academic timeline. The system was designed with a structured development plan — starting from data preprocessing, model training, and website design to chatbot integration — all within the project duration. The availability of ready-to-use datasets and libraries helped save time and ensured timely completion.

- **Legal and Ethical Feasibility**

Since the project involves health-related data, ethical considerations were taken. The dataset used from Kaggle is open-source and does not contain personal or identifiable patient information. The system is designed strictly for educational and awareness purposes, ensuring compliance with ethical and legal standards.

Database Design

1. E-R Diagram

Entities & Attributes

1. User

user_id (PK) – Unique ID for each user
name – User's full name
email – User's email address
password – For login authentication
age – User's age
gender – User's gender
registration_date – Date when the user created the account

2. Patient_Record

record_id (PK) – Unique ID for each medical record
user_id (FK) – References the user who uploaded/entered the record
mean_radius
mean_texture
mean_smoothness
mean_compactness
mean_symmetry
mean_fractal_dimension
diagnosis_result – Final classification: Benign / Malignant
prediction_date – When the prediction was generated

3. Dataset

dataset_id (PK) – Unique identifier for the dataset used for training
dataset_name – Name of dataset (e.g., “Breast Cancer Wisconsin Dataset”)
source – Source of dataset (e.g., Kaggle)
num_records – Total number of records in the dataset
upload_date – When dataset was added for training

4. ML_Model

model_id (PK) – Unique identifier for the trained model
model_name – Algorithm used (e.g., Logistic Regression, Random Forest)
accuracy – Accuracy percentage achieved
training_date – Date the model was trained
dataset_id (FK) – References the dataset used for training

5. Prediction

prediction_id (PK) – Unique ID for each prediction
record_id (FK) – Refers to patient record predicted
model_id (FK) – Model used for prediction
prediction_result – Output label (Benign / Malignant)
confidence_score – Probability or confidence value of prediction
prediction_time – Timestamp when prediction occurred

.

Relationships

User → Patient_Record:

A user can have multiple patient records (1-to-Ma ny).
→ user_id in Patient_Record is a foreign key.

Dataset → ML_Model:

A dataset can train multiple ML models, but each model is trained on one dataset (1-to-Many).
→ dataset_id in ML_Model is a foreign key.

ML_Model → Prediction:

One model can generate multiple predictions (1-to-Many).
→ model_id in Prediction is a foreign key.

Patient_Record → Prediction:

One patient record can have one or more predictions (1-to-Many).
→ record_id in Prediction is a foreign key.

Text-Based ER Diagram (Structure View)

[USER]
user_id (PK)
name
email
password
age
gender
registration_date
|
| 1 --- M
|
[PATIENT_RECORD]
record_id (PK)
user_id (FK)
mean_radius
mean_texture
mean_smoothness
mean_compactness
|
[PREDICTION]
prediction_id (PK)
record_id (FK)
model_id (FK)
prediction_result
confidence_score
prediction_time
|
| M --- 1
|
[ML_MODEL]
model_id (PK)
model_name
accuracy

training_date
dataset_id (FK)
|
| M --- 1
|
[DATASET]
dataset_id (PK)
dataset_name
source
num_records
upload_date

2. Data Normalization

Data normalization is a data preprocessing technique used to adjust numerical values into a standard scale, without distorting differences in the range of values.

In the Breast Cancer Prediction System, normalization is crucial because the dataset contains various tumor measurement attributes — such as mean radius, mean texture, mean smoothness, etc. These attributes have different scales and units, which can bias the machine learning model during training.

By normalizing the data, each feature contributes equally to the model's performance and prevents features with larger values from dominating those with smaller ones.

1. Why Normalization Is Needed in This Project

The Breast Cancer Wisconsin Dataset (from Kaggle) contains several numerical columns with varying ranges.

For example:

mean radius ranges from 6.98 to 28.11
mean texture ranges from 9.71 to 39.28
mean smoothness ranges from 0.05 to 0.16

If the model is trained directly on these values:

The feature with the largest magnitude (e.g., mean radius) will have a greater influence on the model's outcome. Algorithms like Logistic Regression, KNN, and SVM are sensitive to data scale, meaning they perform better when all input features are on a similar scale.

Hence, normalization ensures:

- ☑ Faster convergence during model training
- ☑ Balanced feature contribution
- ☑ Improved accuracy and stability of predictions

2. Types of Normalization Techniques

There are several techniques to normalize data. For this project, the two most suitable methods are:

a. Min-Max Normalization

This technique scales the data to a fixed range, typically [0, 1].

Formula:

X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}}

Explanation:

: Original value

: Minimum value of the feature

: Maximum value of the feature

: Normalized value

Example: If mean_radius ranges from 6.98 to 28.11, and a particular record has a mean radius of 14.12:

X_{norm} = \frac {14.12 - 6.98}{28.11 - 6.98} = 0.33

So the normalized mean radius = 0.33, within [0, 1].

This method is ideal when you want to preserve relationships between values and when you use distance-based algorithms (like KNN or SVM).

b. Z-Score Normalization (Standardization)

This method transforms data to have a mean = 0 and standard deviation = 1.

Formula:

Z = \frac{X - \mu}{\sigma}

Explanation:

: Original value

: Mean of the feature

: Standard deviation of the feature

Example: If mean_radius has mean = 14.1 and standard deviation = 3.5, then for a record with value 20:

Z = \frac{20 - 14.1}{3.5} = 1.69

So, the standardized mean radius = 1.69.

Z-score normalization is used in models that assume normal distribution of features, like Logistic Regression, SVM, and Neural Networks.

3. Implementation in the Project

In your Python implementation (using libraries like Scikit-learn), normalization is typically done using the following code:

```
from sklearn.preprocessing import MinMaxScaler, StandardScaler
import pandas as pd

# Load dataset
data = pd.read_csv('data.csv')

# Separate features and target
X = data.drop(['diagnosis'], axis=1)
y = data['diagnosis']

# Option 1: Min-Max Normalization
scaler = MinMaxScaler()
X_normalized = scaler.fit_transform(X)

# Option 2: Z-Score Normalization
# scaler = StandardScaler()
# X_normalized = scaler.fit_transform(X)
```

This ensures that every feature (e.g., mean radius, mean texture, etc.) is rescaled before training the ML model.

4. Post-Normalization Data Check

After applying normalization, you can verify the results by checking the new range or mean/standard deviation:

```
print(X_normalized.min())
print(X_normalized.max())
```

or for standardization:

```
print(X_normalized.mean())
print(X_normalized.std())
```

- ☑ If using Min-Max, the values should lie between 0 and 1.
- ☑ If using Z-Score, the mean should be 0 and standard deviation 1.

3. Data Dictionary

The data dictionary provides a detailed description of each attribute (feature) in the dataset used for training and testing the Breast Cancer Prediction System. This dataset the Breast Cancer Wisconsin (Diagnostic) Data Set is sourced from Kaggle and originally from the UCI Machine Learning Repository.

It contains 569 records and 32 attributes, including one target variable (diagnosis) and 31 numerical predictor features that describe the characteristics of cell nuclei present in the image of a breast mass.The goal is to predict whether the tumor is Benign (B) or Malignant (M) based on these input features.

- **Null and Missing Values**

The dataset contains no null or missing values.
Each record is complete with all 32 attributes filled.
However, data cleaning is still applied to ensure type consistency and proper encoding of categorical variables.

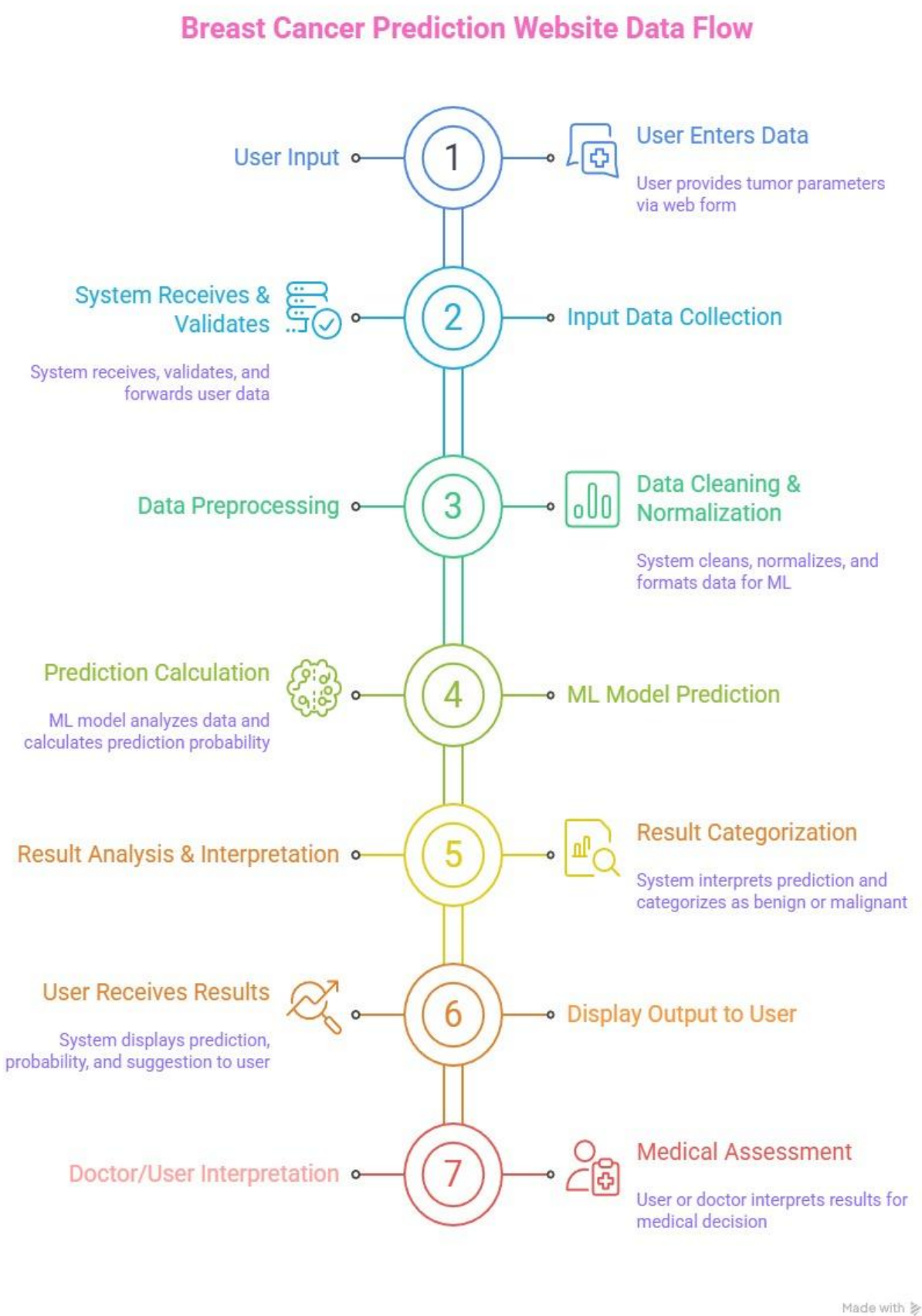
- **Data Encoding**

Since the target variable diagnosis is categorical (M/B), it is encoded numerically for model processing:

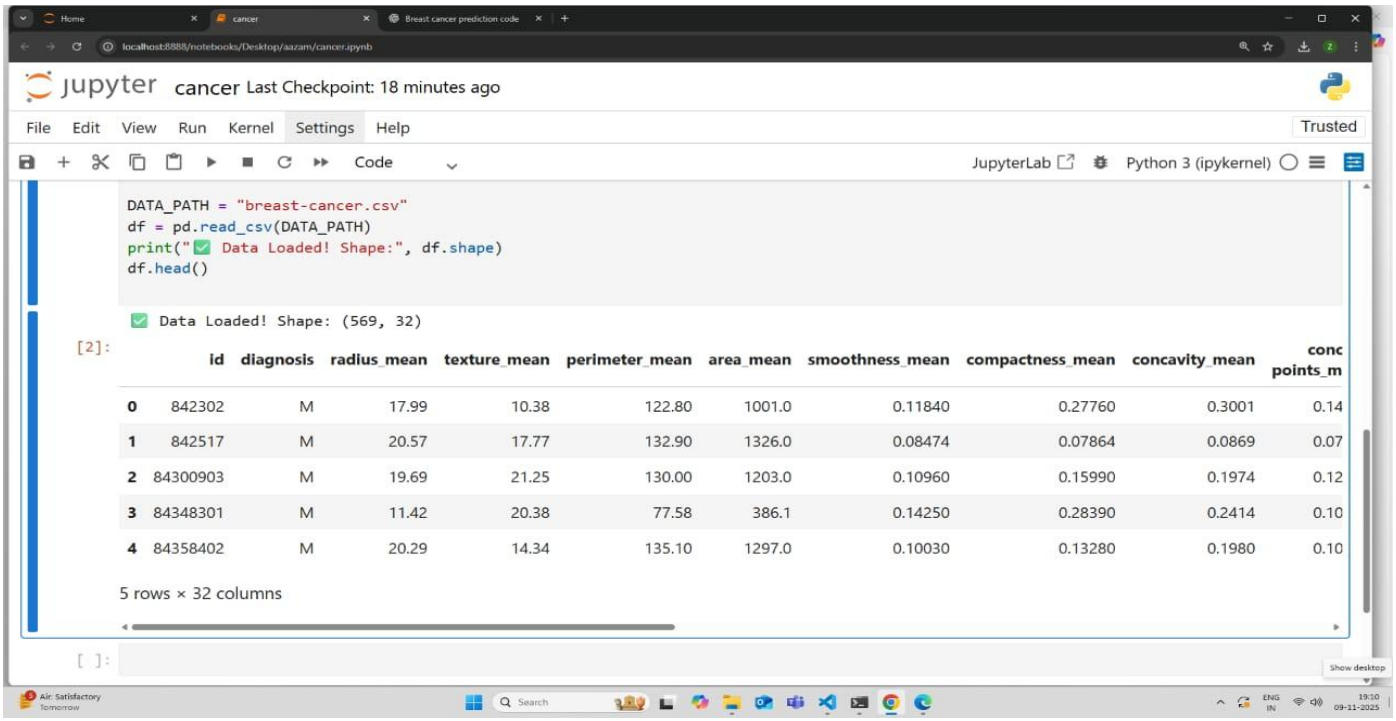
Original Value Encoded Value

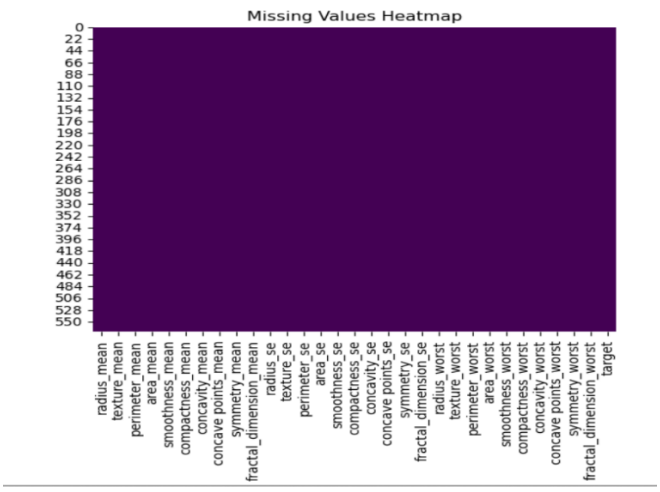
M (Malignant) 1
B (Benign) 0
9. Example Data Record (After Preprocessing.

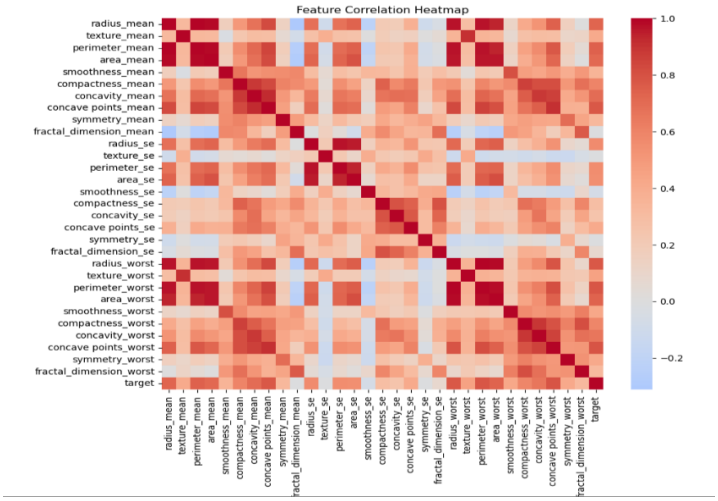
Data Flow

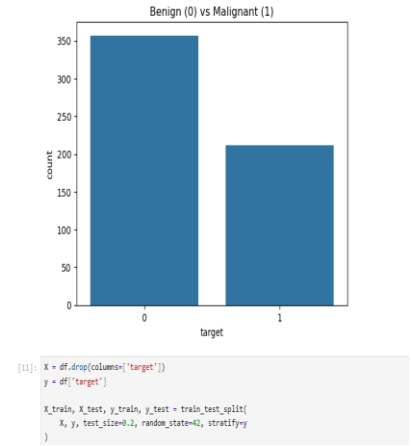

Breast Cancer Prediction Website Data Flow
User Input
1
User Enters Data
User provides tumor parameters via web form
System Receives & Validates
2
Input Data Collection
System receives, validates, and forwards user data
Data Preprocessing
3
Data Cleaning & Normalization
System cleans, normalizes, and formats data for ML
Prediction Calculation
ML model analyzes data and calculates prediction probability
4
ML Model Prediction
Result Analysis & Interpretation
5
Result Categorization
System interprets prediction and categorizes as benign or malignant
User Receives Results
System displays prediction, probability, and suggestion to user
6
Display Output to User
Doctor/User Interpretation
7
Medical Assessment
User or doctor interprets results for medical decision
Made with Napkin

Reports

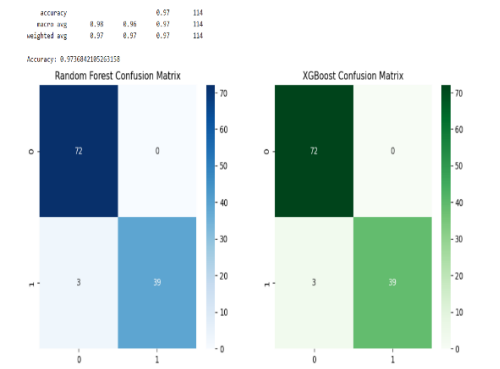

cancer Last Checkpoint: 18 minutes ago
File Edit View Run Kernel Settings Help
Trusted
Code
JupyterLab
Python 3 (ipykernel)
DATA_PATH = "breast-cancer.csv"
df = pd.read_csv(DATA_PATH)
print("✅ Data Loaded! Shape:", df.shape)
df.head()
✅ Data Loaded! Shape: (569, 32)
[2]:
id diagnosis radius_mean texture_mean perimeter_mean area_mean smoothness_mean compactness_mean concavity_mean conc points_m
0 842302 M 17.99 10.38 122.80 1001.0 0.11840 0.27760 0.3001 0.14
1 842517 M 20.57 17.77 132.90 1326.0 0.08474 0.07864 0.0869 0.07
2 84300903 M 19.69 21.25 130.00 1203.0 0.10960 0.15990 0.1974 0.12
3 84348301 M 11.42 20.38 77.58 386.1 0.14250 0.28390 0.2414 0.10
4 84358402 M 20.29 14.34 135.10 1297.0 0.10030 0.13280 0.1980 0.10
5 rows × 32 columns

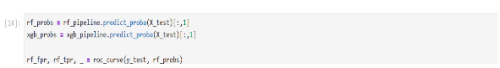

Missing Values Heatmap

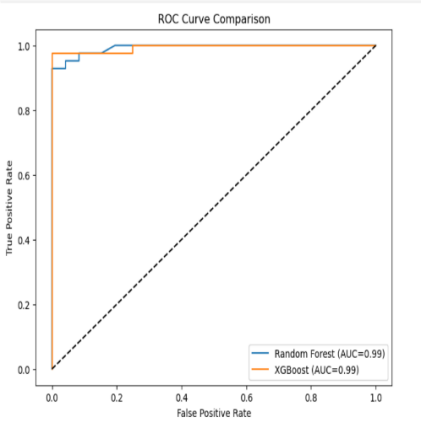

Feature Correlation Heatmap


Benign (0) vs Malignant (1)
count
target

[11]:
X = df.drop(columns=['target'])
y = df['target']
X_train, X_test, y_train, y_test = train_test_split(
X, y, test_size=0.2, random_state=42, stratify=y
)


accuracy 0.97 114
macro avg 0.98 0.96 0.97 114
weighted avg 0.97 0.97 0.97 114
Accuracy: 0.9736842105263158
Random Forest Confusion Matrix
XGBoost Confusion Matrix


[16]:
rf_probs = rf_pipeline.predict_proba(X_test)[:,1]
xgb_probs = xgb_pipeline.predict_proba(X_test)[:,1]
rf_fpr, rf_tpr, _ = roc_curve(y_test, rf_probs)


ROC Curve Comparison
True Positive Rate
False Positive Rate
Random Forest (AUC=0.99)
XGBoost (AUC=0.99)

Conclusion

a) What is Achieved?

The project successfully developed a web-based Breast Cancer Prediction System integrated with a Chatbot for awareness and assistance. The website allows users or healthcare professionals to input specific diagnostic parameters, which are then analyzed by a trained machine learning model to predict the likelihood of breast cancer. This helps in providing an early indication that can support doctors in making faster and more accurate assessments. Additionally, the integrated chatbot offers an interactive platform where users can ask questions related to symptoms, remedies, and preventive care, thus promoting health awareness and education. Overall, the system achieved its primary goal of combining machine learning and web technologies to contribute to early breast cancer detection and user awareness.

b) What is Required?

Although the system performs effectively, certain aspects can be enhanced. More real time and large-scale medical datasets are required to further improve the accuracy and reliability of the prediction model. In addition, collaboration with healthcare professionals is essential for validating the system's predictions with real patient data. Better data security mechanisms and privacy protocols are also needed to ensure that user information remains confidential. Lastly, hosting the system on a cloud-based platform would make it more accessible and scalable for public use.

c) Future Plan

In the future, the project can be expanded by integrating advanced deep learning algorithms such as Convolutional Neural Networks (CNNs) for more precise predictions based on medical images like mammograms. The chatbot can also be enhanced using Natural Language Processing (NLP) to make responses more human-like and context-aware. A mobile-friendly version of the website could be developed to reach a wider audience. With further development and medical validation, this system can be implemented in clinics and diagnostic centers to assist in early detection, reduce diagnostic workload, and create greater awareness about breast cancer prevention

Reference

The successful development of the Breast Cancer Prediction Website was made possible through the use of various credible platforms, datasets, and AI-powered tools that provided valuable guidance, data support, and technical assistance throughout the project.

1.World Health Organization (WHO):

The WHO website served as an authentic reference for understanding the global scenario of breast cancer. It provided real-time data and statistics regarding its prevalence, mortality rates, and early detection importance. This information helped in establishing the background and relevance of our project and demonstrated the real-world need for predictive healthcare solutions.

2. Kaggle – Breast Cancer Wisconsin (Diagnostic) Dataset:

The dataset obtained from Kaggle was the primary source for training and testing the machine learning model. It contained vital medical attributes such as tumor size, radius mean, texture, and compactness, which were used to analyze and predict the chances of breast cancer. Kaggle also offered a strong data science community with public notebooks and discussions that enhanced our data preprocessing, feature selection, and model evaluation process.

3.Wikipedia:

Wikipedia was referenced for obtaining general and medical knowledge about breast cancer, including its causes, symptoms, types, and treatment options. This information was used mainly to enrich the chatbot's responses, ensuring it could answer users' questions related to symptoms, remedies, and preventive measures accurately and understandably.

4. ChatGPT:

ChatGPT was a valuable AI-powered assistant throughout the project development. It helped in structuring documentation, generating project-related content, improving chatbot dialogues, and clarifying complex programming and design concepts. The tool also contributed to enhancing the language quality, technical clarity, and flow of the project report.

5. Napkin AI:

Napkin AI was used for creating professional and detailed process visualizations such as Data Flow Diagrams and System Flow Charts. It provided AI-generated diagrammatic representations that visually depicted how the system processes user input, executes the machine learning prediction, and returns the results. This tool made the documentation and presentation more interactive and comprehensible.