Breast Cancer Prediction

Project Report

Industrial Training (ECS-599)

BACHELOR OF TECHNOLOGY (AI – ML - DL)

December, 2024

FACULTY OF ENGINEERING & COMPUTING SCIENCES

TEERTHANKER MAHAVEER UNIVERSITY, MORADABAD

PROJECT GUIDE:

Dr. Preeti Rani

SUBMITTED BY:

Abhishek Kumar Singh (TCA2364001)

DECLARATION

We hereby declare that this Project Report titled **Breast Cancer Prediction** submitted by us and approved by our project guide, Faculty of Engineering & Computing Sciences. Teerthanker Mahaveer University, Moradabad, is a bonafide work undertaken by us and it is not submitted to any other University or Institution for the award of any degree diploma / certificate or published any time before.

Project ID : BC_ML_Pred_2024

Student Name: Abhishek kumar singh Signature

Project Guide : Dr.Preeti Rani Signature

Table of Contents

Appendix

A: Data Flow Diagram (DFD)

B: Entity Relationship Diagram (ERD)

C: Use Case Diagram (UCD)

D: Data Dictionary (DD)

E: Screen Shots

1 Project Title

Breast Cancer Prediction Using Machine Learning Algorithms.

2 Problem Statement

Breast cancer is a leading cause of death, and early detection is crucial for improving survival rates. Traditional diagnostic methods are often time-consuming and prone to human error. This project aims to develop a machine learning-based system that analyzes clinical, genetic, and imaging data to predict the likelihood of breast cancer, enabling early diagnosis and more effective treatment plans. The goal is to enhance accuracy, reduce false negatives, and support healthcare providers in making informed decisions.

3 Project Description

This project aims to develop a machine learning-based system to predict the likelihood of breast cancer using clinical, genetic, and diagnostic data. By analysing datasets containing patient demographics, biopsy results, and imaging features, the model will identify patterns and predict cancer presence with high accuracy. Different machine learning algorithms, such as decision trees, random forests, and support vector machines, will be utilized and optimized to enhance prediction accuracy. The goal is to assist healthcare professionals in early detection, reducing false negatives, and improving treatment outcomes. Ultimately, this tool will aid in making informed, data-driven decisions for better patient care.

3.1 Scope of the Work

This project aims to develop a machine learning model for predicting breast cancer risk using clinical, genetic, and diagnostic data. Key tasks include:

1. **Data Collection & Pre-processing**: Gather relevant datasets (e.g., Breast Cancer Wisconsin dataset) and perform data cleaning, normalization, and feature selection.

2. **Model Development**: Implement and train various machine learning algorithms (e.g., decision trees, random forests, SVM) to predict breast cancer outcomes.
3. **Model Evaluation**: Evaluate model performance using accuracy, precision, recall, and ROC-AUC metrics to ensure reliable predictions.
4. **Deployment**: Build an interface for healthcare professionals to use the model for early cancer detection and decision-making.

The goal is to create an accurate, data-driven tool for early breast cancer diagnosis, improving detection and treatment outcomes.

3.2 Project Modules

1. **Data Collection and Pre-processing**:
   - Gather datasets (e.g., Breast Cancer Wisconsin dataset).
   - Handle missing data, normalize features, and perform feature selection.
2. **Exploratory Data Analysis (EDA)**:
   - Analyse data patterns, correlations, and distributions.
   - Visualize key features that influence breast cancer risk.
3. **Model Development**:
   - Train machine learning models (e.g., Logistic Regression, Decision Trees, Random Forests, SVM) to predict cancer classification (benign or malignant).
4. **Model Evaluation**:
   - Assess model performance using accuracy, precision, recall, F1-score, and ROC-AUC metrics.
5. **Model Optimization**:
   - Tune hyper parameters and refine models for improved prediction accuracy.
6. **Deployment**:
   - Develop a user-friendly interface or API for real-time predictions to assist healthcare professionals in early detection.

These modules collectively aim to create an effective and accurate breast cancer prediction system.

3.3 Context Diagram (High Level)

The context diagram represents the high-level flow of data and interactions between the key components in the breast cancer prediction system.

1. **External Entities:**
    - **Healthcare Providers**: Input patient data, such as medical history, demographics, and test results (e.g., biopsy, imaging).
    - **Patient**: Provides medical data (e.g., test reports, symptoms) to the system.
2. **System:**
    - **Breast Cancer Prediction Model**: Analyses input data using machine learning algorithms to predict whether a patient has benign or malignant breast cancer.
3. **Processes:**
    - **Data Pre-processing**: Clean and prepare data for analysis (e.g., normalization, missing values).
    - **Model Training**: Train various machine learning models (e.g., SVM, Random Forest) on historical data.
    - **Prediction**: Generate predictions (benign or malignant) based on input data using the trained model.
4. **Outputs:**
    - **Prediction Results**: The system outputs cancer predictions to the healthcare provider for review and further action (e.g., diagnosis, treatment planning).
5. **Feedback**:
    - Healthcare providers can provide feedback on prediction accuracy, which is used to update and improve the model.

This context diagram simplifies the interaction between the system, users, and external entities, focusing on the flow of data and outputs for cancer prediction.

4 Implementation Methodology

1. **Data Collection & Pre-processing**:
    - **Source**: Use datasets like the **Breast Cancer Wisconsin dataset**.
    - **Pre-processing**: Handle missing values, normalize features, and select relevant variables for analysis.
2. **Exploratory Data Analysis (EDA)**:
    - Analyse the data for patterns, correlations, and outliers.
    - Visualize key features to understand their relationship with cancer classification.
3. **Model Development**:
    - Implement machine learning algorithms (e.g., Logistic Regression, Decision Trees, Random Forests, SVM).
    - Split data into training and testing sets for model evaluation.
4. **Model Evaluation**:
    - Assess model performance using accuracy, precision, recall, F1-score, and ROC-AUC metrics.
    - Compare different models to select the most accurate one.
5. **Model Optimization**:
    - Tune hyper parameters and refine the model for better performance using cross-validation techniques.
6. **Deployment**:
    - Develop a user interface or API for real-time predictions to assist healthcare professionals in early cancer detection.

5 Technologies to be used

1. **Programming Languages**:

- **Python**: For data analysis, machine learning, and model development.

2. **Libraries and Frameworks**:
   - **Pandas**: For data manipulation and analysis.
   - **NumPy**: For numerical operations and data handling.
   - **Matplotlib/Seaborn**: For data visualization and exploratory analysis.
   - **Scikit-learn**: For machine learning algorithms and model evaluation.
3. **Data Storage**:
   - **CSV/Excel Files**: For storing and handling datasets.

5.1 Software Platform

a) Back-end

Python, Pandas, NumPy, Sklearn, Algorithms

5.2 Hardware Platform

RAM, Hard Disk, OS, VS Code, Browser etc.

5.3 Tools, if any

Python, Pandas, NumPy, Seaborn, Sklearn, Matplotlib etc.

6 Advantages of this Project

1. **Early Detection**:
   - Helps in identifying breast cancer at an early stage, improving treatment outcomes and survival rates.
2. **Accuracy and Efficiency**:
   - Machine learning models provide accurate predictions, reducing human error and speeding up the diagnosis process.
3. **Cost Reduction**:
   - By automating the prediction process, it reduces the need for costly diagnostic procedures like biopsies for every patient.

4. **Support for Healthcare Professionals**:
    - Assists doctors by providing data-driven insights, enabling more informed decisions and reducing diagnostic workload.
5. **Personalized Healthcare**:
    - Provides individualized predictions based on patient data, aiding in personalized treatment planning.
6. **Scalability**:
    - The system can be scaled to handle large datasets, making it suitable for deployment in hospitals or clinics globally.

7 Future Scope and further enhancement of the Project

1. **Integration with Medical Imaging**:
    - Incorporate AI-based image analysis (e.g., mammograms, ultrasounds) to enhance prediction accuracy.
2. **Real-time Prediction**:
    - Implement real-time prediction features for quicker diagnostic decision-making in clinical settings.
3. **Genetic Data Integration**:
    - Incorporate genetic information (e.g., BRCA gene mutations) to improve the model's accuracy and identify hereditary risks.
4. **Mobile Application**:
    - Develop a mobile app for healthcare professionals to access the prediction system on the go.
5. **Advanced Machine Learning Techniques**:
    - Explore deep learning and ensemble methods to improve prediction performance and handle complex cases.

8 Definitions, Acronyms, and Abbreviations

Abbreviation	Description

9 Conclusion

The Breast Cancer Prediction project demonstrates the power of machine learning in aiding early detection and improving diagnostic accuracy for breast cancer. By leveraging data from medical tests and patient information, the system can predict whether a tumour is benign or malignant with high precision. The use of advanced algorithms ensures efficiency and scalability, providing healthcare professionals with an effective tool for decision-making.

The project's future scope includes integrating medical imaging, incorporating genetic data, and expanding into real-time prediction systems, all of which will enhance the accuracy and applicability of the system in clinical environments. Overall, this project has the potential to significantly improve breast cancer detection, reduce healthcare costs, and ultimately contribute to better patient outcomes.

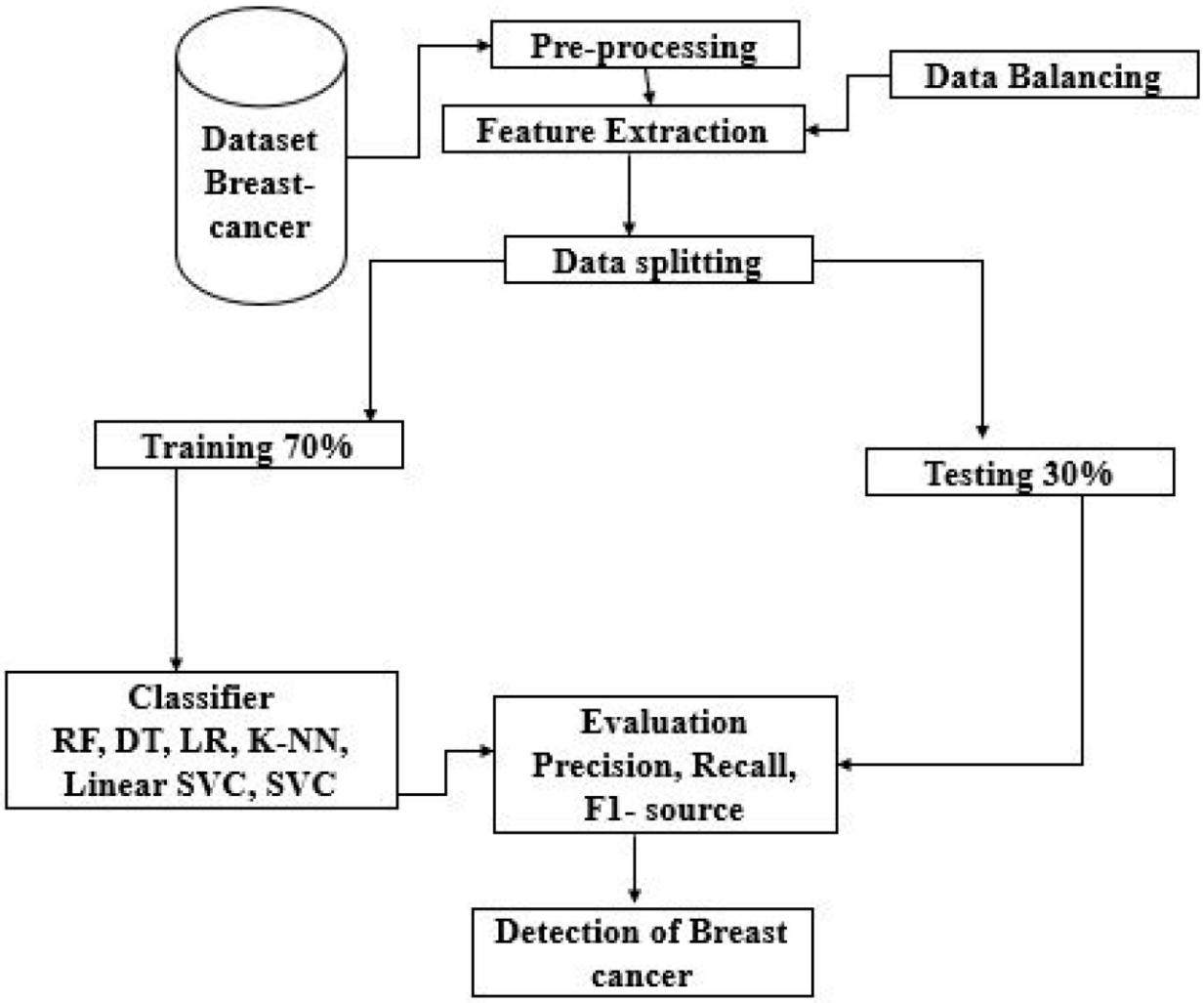
10 References

S#	Reference Details	Owner	Version	Date
1.	Project Synopsis	Abhishek Singh		29-11-24
2.	Project Requirements	Abhishek Singh		29-11-24

Annexure A

Data Flow Diagram (DFD)

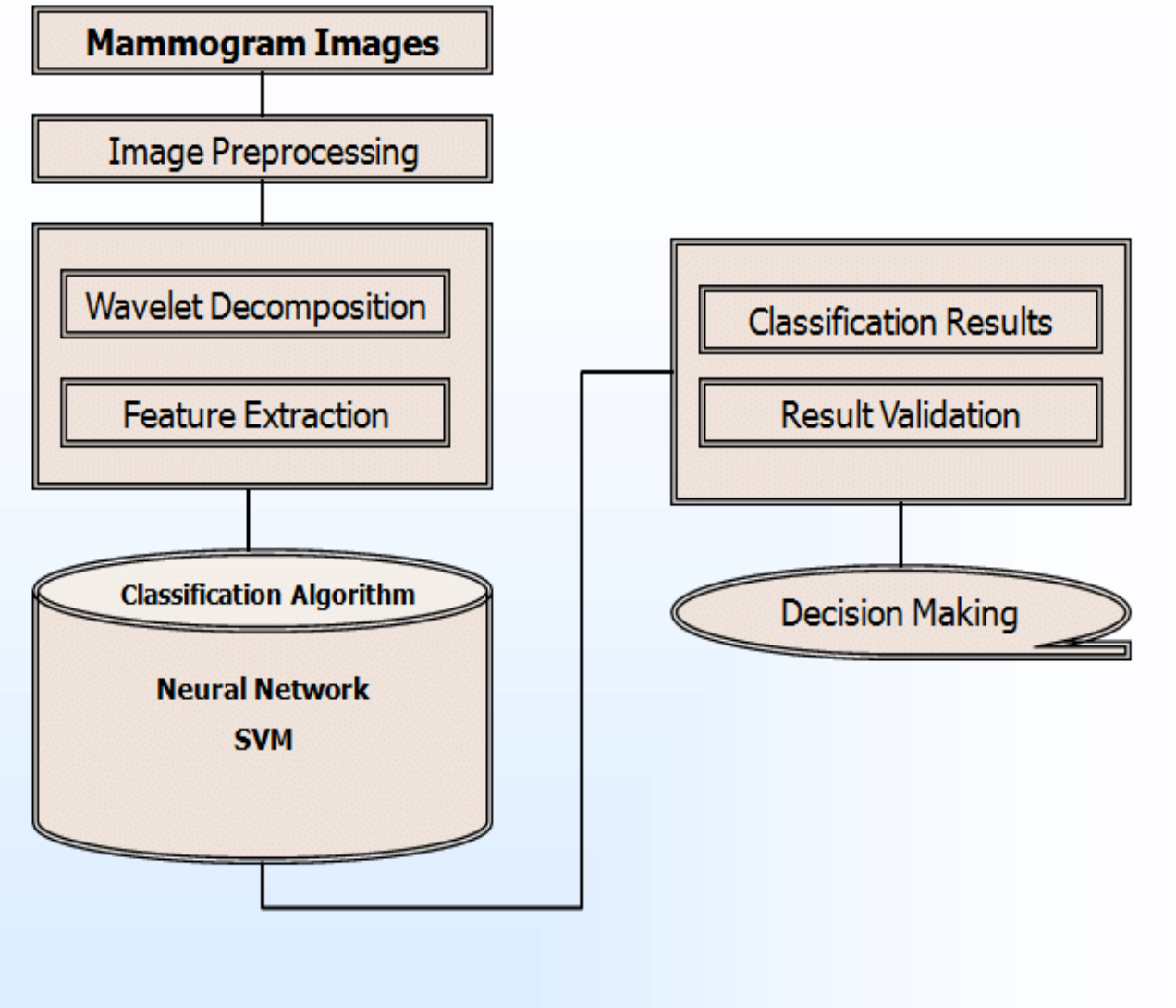
(Mandatory)



Annexure B

Entity-Relationship Diagram (ERD)

(Mandatory)



Annexure D

Data Dictionary (DD)

(Mandatory)

Example:

User Table (USR)

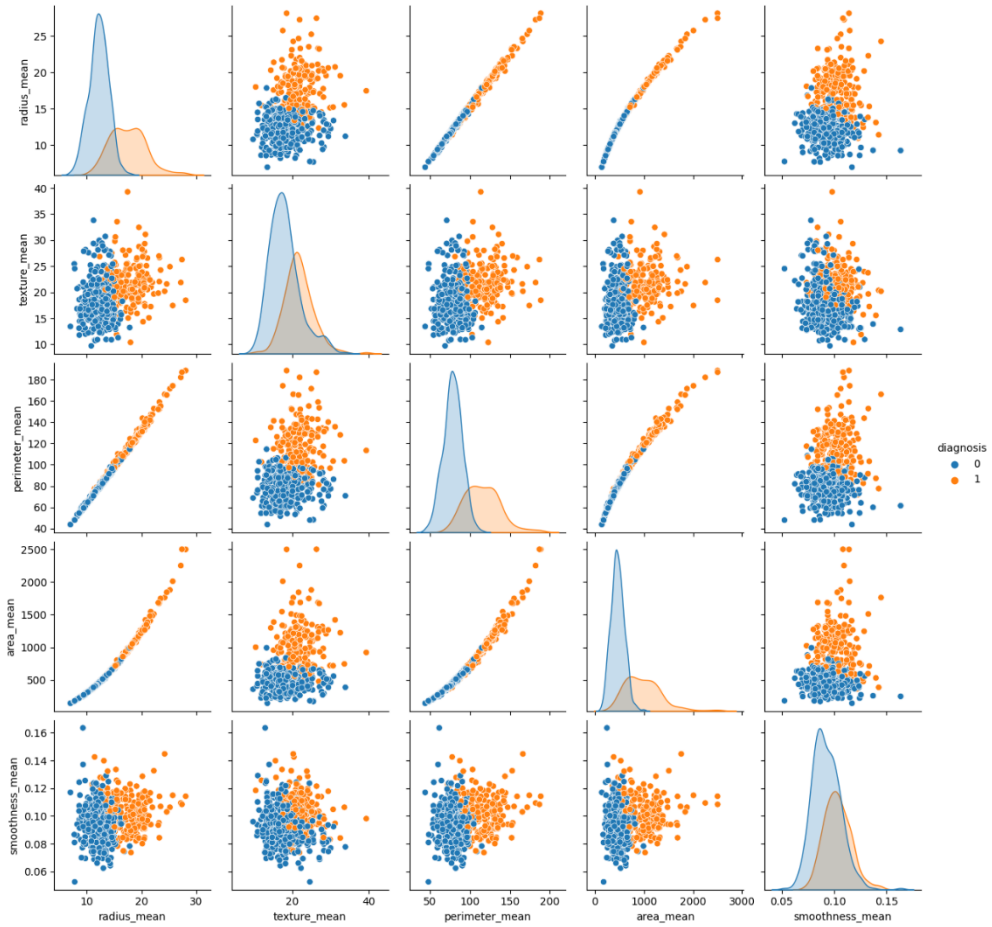
Fields	Data type	Description
USR-Name	Text	Admin name
USR-Password	Text	Admin password
USR-Contact-No	Number	Admin Contact
USR-Address	Text	City

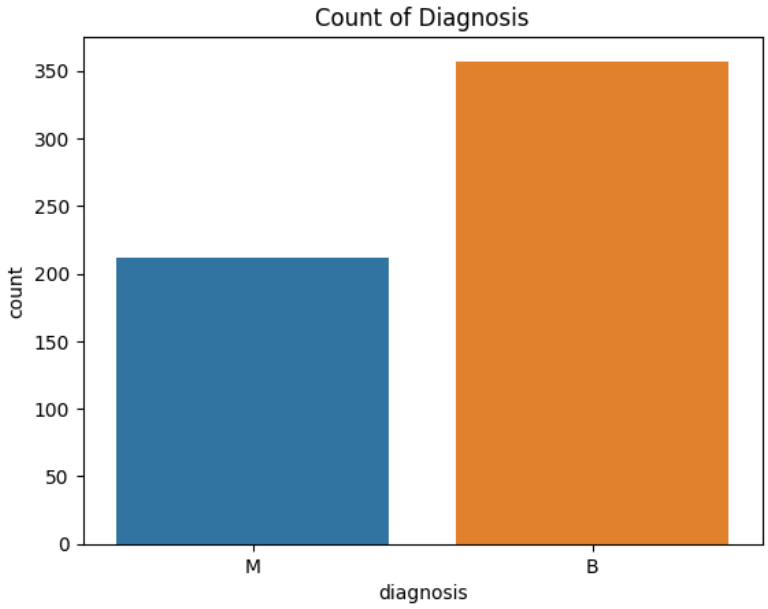
Supplier Table (SUPP)

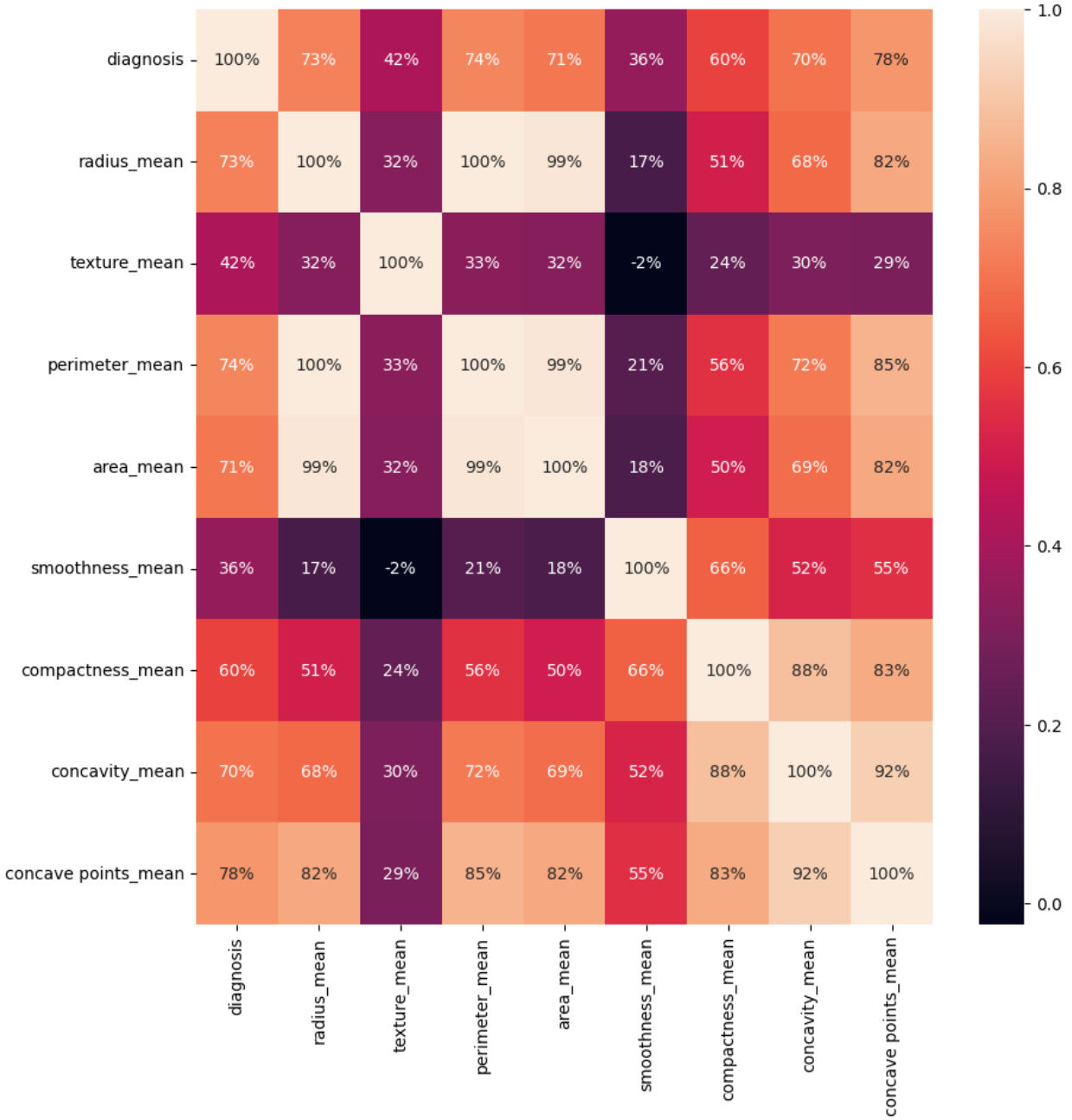
Fields	Data type	Description
SUPP-ID	Number	Supplier ID
SUPP-Name	Text	Supplier Name
SUPP-Address	Text	Supplier Address
SUPP-Contact	Number	Supplier Contact
SUPP-Credit-Limit	Number	Credit Limit

Annexure E

Screen Shots






diagnosis 100% 73% 42% 74% 71% 36% 60% 70% 78%
radius_mean 73% 100% 32% 100% 99% 17% 51% 68% 82%
texture_mean 42% 32% 100% 33% 32% -2% 24% 30% 29%
perimeter_mean 74% 100% 33% 100% 99% 21% 56% 72% 85%
area_mean 71% 99% 32% 99% 100% 18% 50% 69% 82%
smoothness_mean 36% 17% -2% 21% 18% 100% 66% 52% 55%
compactness_mean 60% 51% 24% 56% 50% 66% 100% 88% 83%
concavity_mean 70% 68% 30% 72% 69% 52% 88% 100% 92%
concave points_mean 78% 82% 29% 85% 82% 55% 83% 92% 100%
diagnosis
radius_mean
texture_mean
perimeter_mean
area_mean
smoothness_mean
compactness_mean
concavity_mean
concave points_mean
1.0
0.8
0.6
0.4
0.2
0.0