SHAH & ANCHOR

Mahavir Education Trust's

SHAH & ANCHOR KUTCHHI ENGINEERING COLLEGE

Chembur, Mumbai - 400 088

UG Program in Information Technology

- **Student Name: Purva Mahamunkar**

- **Roll Number: 12/28**

- **PRN : 22UF17661IT090**

- **Department: IT**

- **Batch/Division: BTECH-1**

- **Experiment Title: Perform Exploratory Data Analysis of Healthcare Data**

- **Date: 19/07/2025**

Experiment No. 1				
Date of Performance:	**19/07/2025**			
Date of Submission:	**26/07/2025**			
Program Execution/ formation / correction/ ethical practices	Timely Submission	Viva	Experiment Total	Sign with Date

Experiment No 1

2.1 Aim: Collect, Clean, Integrate and Transform Healthcare Data based on a Specific Disease

2.2 Course Outcome: Collect, Clean, Integrate, Transform Healthcare Data based on a Specific Disease

2.3 Learning Outcome: Collect, clean, preprocess healthcare data for use in AI/ML models.

2.4 Requirement: Healthcare Domain Knowledge, Data Collection Tools, Data Cleaning Techniques, Data Integration Methods.

2.5 Related Theory:

This experiment focuses on the collection, preprocessing, and transformation of healthcare data for the purpose of disease detection, using the Breast Cancer Dataset. The following steps were undertaken to ensure the dataset was clean, structured, and suitable for building an effective predictive model:

1. **Importing Libraries and Dataset**

- Essential Python libraries such as pandas, numpy, matplotlib, and seaborn were imported for data manipulation and visualization.
- The Breast Cancer dataset was loaded for analysis and model training.

2. **Exploratory Data Analysis (EDA)**

- Basic information about the dataset (such as number of rows, columns, and data types) was examined.

- Summary statistics were generated to understand the distribution and central tendencies of each feature.
- Initial visualizations (e.g., histograms, boxplots) were created to identify data patterns and anomalies.

3. **Handling Missing Values**

- The dataset was scanned for null or missing entries.
- Missing data were treated using appropriate imputation techniques, such as replacing with mean, median, or mode, depending on the feature's distribution.

4. **Encoding Categorical Data**

- Categorical variables were identified and transformed into numerical form using encoding techniques.
- Label encoding or one-hot encoding was applied to ensure compatibility with machine learning algorithms.

5. **Analyzing Correlation**

- A correlation matrix was generated to identify the strength and direction of relationships between features.
- A heatmap was visualized to detect multicollinearity and assess which features are strongly associated with the target variable.

6. **Splitting the Dataset**

- The dataset was divided into training and testing sets, typically in an 80:20 ratio.
- This step ensures that model performance can be evaluated on unseen data.

7. **Feature Scaling**

- Numerical features were standardized to ensure that all features contribute equally to the model.
- Techniques like StandardScaler (z-score normalization) were used to scale data within a similar range.

8. Applying Logistic Regression

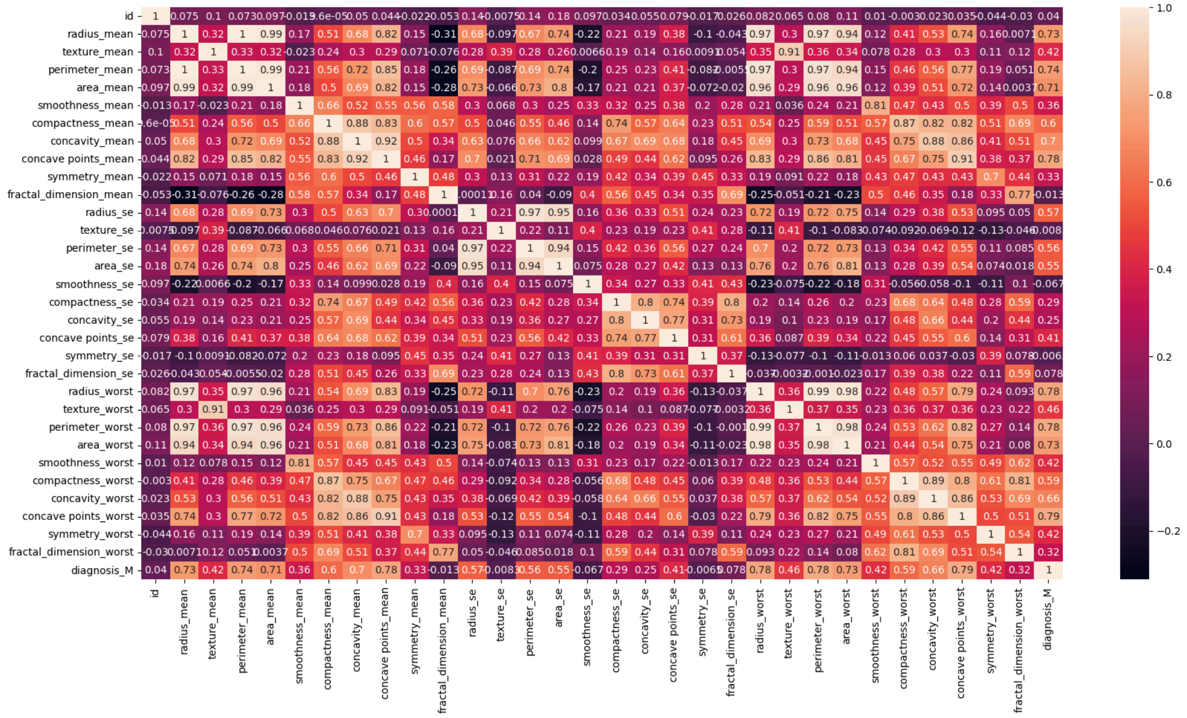
- A logistic regression classifier was implemented to model the probability of a binary outcome (malignant vs. benign).
- The model was trained on the preprocessed data and evaluated using metrics such as accuracy, precision, recall, and F1-score.

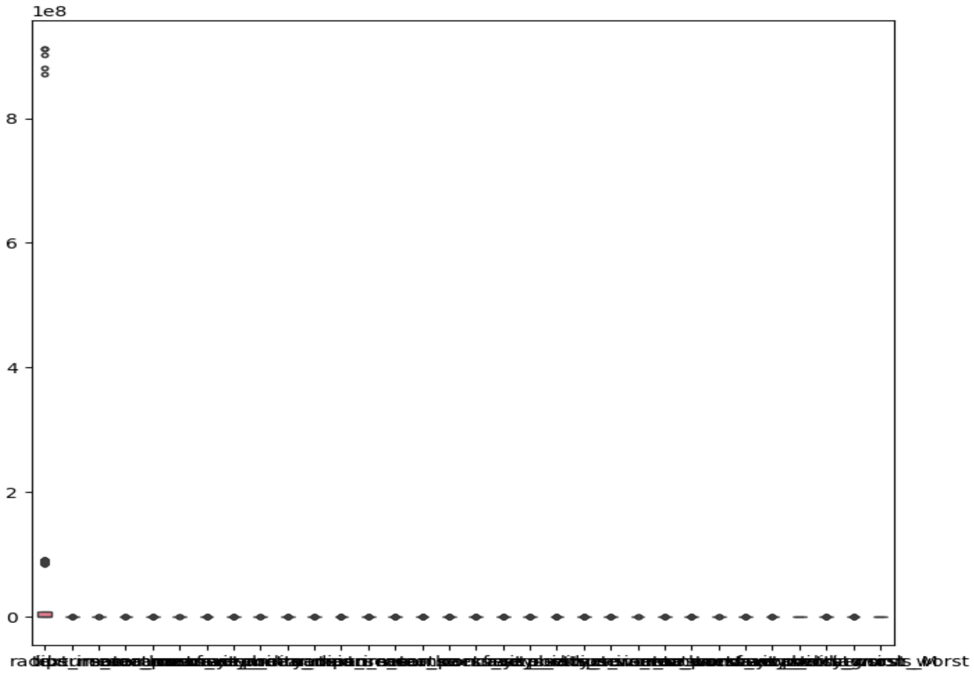
Each of the above steps contributes to building a robust and interpretable model for breast cancer detection. Proper preprocessing ensures data quality, while logistic regression, due to its simplicity and effectiveness, remains a widely used method in binary classification tasks in the medical domain.

2.6 Program and Output:

Completed Google Colab [AIML1 Link]

2.7 Result SnapShots:





2.8 Conclusion:

In this experiment, we successfully collected, cleaned, and transformed healthcare data to build a predictive model for breast cancer detection. By applying key preprocessing steps and implementing Logistic Regression, we demonstrated how clean, well-structured data significantly improves model performance. This experiment reinforced the importance of data preparation and introduced a reliable method for binary disease classification.

2.9 Questions:

1. **Why is feature scaling important in logistic regression?**

   Feature scaling is important in logistic regression because it ensures that all features contribute equally to the result. If features have different ranges, the model may give more importance to those with larger values. Scaling helps improve the convergence speed of the optimization algorithm and leads to better performance. It also makes the coefficients easier to interpret. In this experiment, we used StandardScaler to standardize the data.

2. **How did you handle missing values in the Breast Cancer dataset and why?**
   - First, I checked for missing values in the dataset to understand where and how much data was missing.
   - If only a few values were missing, I removed those rows to avoid affecting the overall analysis.
   - If missing data was present in important columns or in large amounts, I filled the missing values:
     - ➢ For numbers, I used the average (mean) or middle value (median).
     - ➢ For categories (like "benign" or "malignant"), I used the most common value (mode).
   - I chose this method to ensure that the dataset stays complete, balanced, and ready for machine learning.
   - This also helped improve the accuracy and reliability of the model by reducing data errors or bias.