

BREAST_CANCER_ CLASSIFICATION

content:

1. introduction

2. Dataset Overview

3. Columns Description

4. Data Preprocessing

5. Exploratory Data Analysis (EDA) & Visualization

6. Feature Selection

7. Model Building – Logistic Regression

8. Model Building – Decision Tree

9. Final Model for the Application

10. Conclusion & Future Work

1. introduction:

Breast cancer is one of the most common and deadly diseases affecting women worldwide. Early detection is crucial for effective treatment and increased survival rates. This project aims to build a reliable machine learning model that can classify breast tumors as malignant or benign based on various cell features extracted from digitized images. The model will be integrated into a user-friendly web application to assist in real-time diagnosis support.

2. Dataset Overview:

>>> The dataset used in this project is the Breast Cancer Wisconsin Diagnostic Dataset, sourced from Kaggle. It contains 569 samples and 32 columns in total, including 30 numerical features representing characteristics of cell nuclei, one ID column, and one target variable called 'diagnosis' which indicates whether a tumor is malignant (cancerous) or benign (non-cancerous). The dataset has no missing values or duplicate records, making it suitable for training machine learning models.

3.Columns Description:

The dataset includes 30 numerical features that describe characteristics of the cell nuclei in breast tissue samples. Below are some of the key features used in this project:

- radius_mean: Average radius of the nucleus.
- texture_mean: Average variation in gray-scale values.
- perimeter_mean: Average perimeter of the nucleus.
- area_se: Standard error of the area.
- concavity_worst: Worst (largest) concavity of the nucleus, indicating how much the shape deviates inward.
- concave_points_worst: Worst number of concave points in the nucleus.
- radius_worst: Worst (largest) radius.
- concave_points_mean: Average number of concave points.
- perimeter_worst: Worst (largest) perimeter.

These features capture shape, size, and texture information of the cells, which help differentiate between benign and malignant tumors. Features related to concavity and concave points were found to have the strongest correlation with the tumor diagnosis.

Target and ID Column Description:

- diagnosis: This is the target variable indicating the diagnosis of the tumor. It is categorical, with two classes:

M for malignant (cancerous) tumors, encoded as 1.

B for benign (non-cancerous) tumors, encoded as 0.

- id: This column contains a unique identifier for each sample. It does not hold any predictive value and was removed during preprocessing.

4. Data Preprocessing:

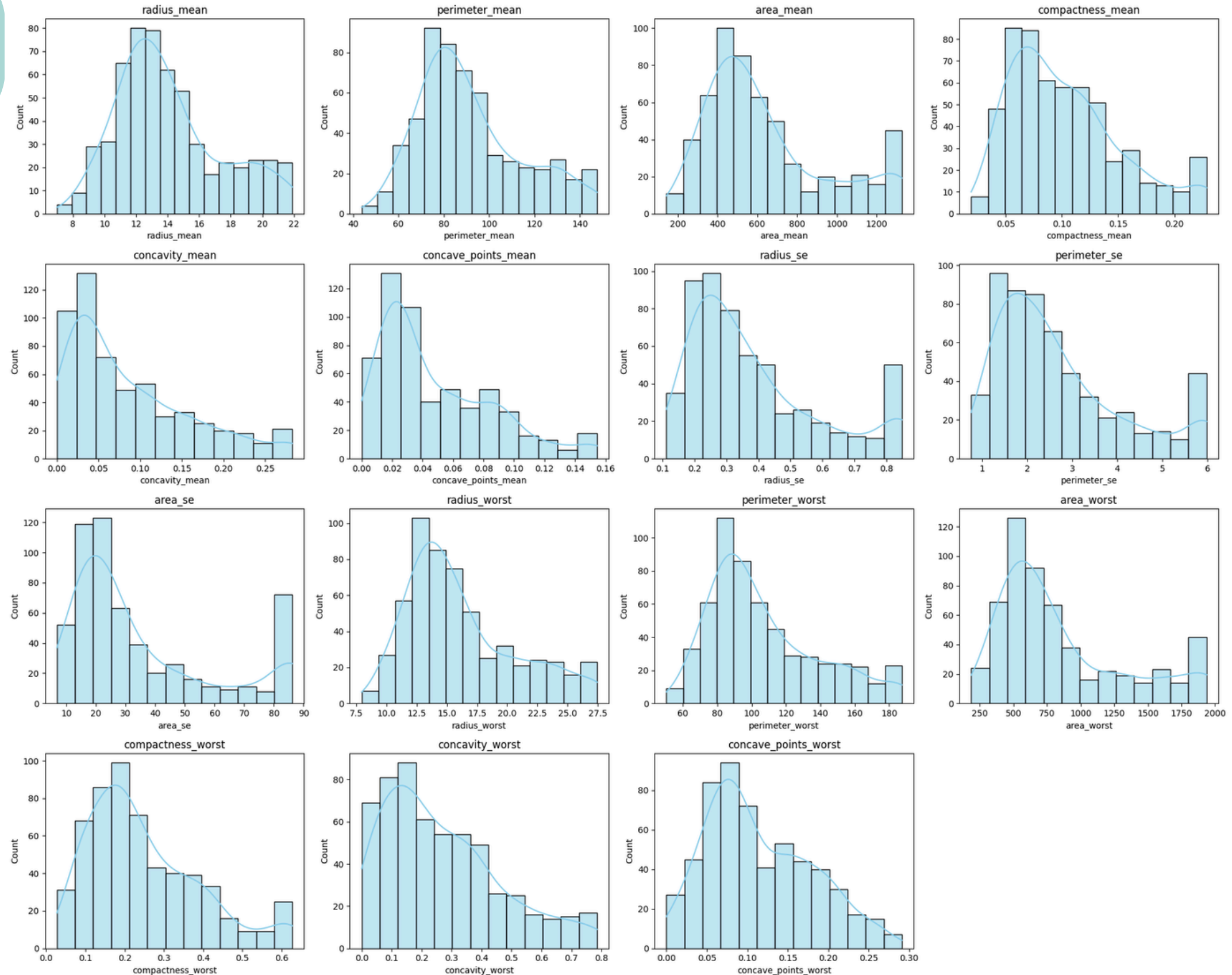
Data preprocessing is a critical step to prepare the dataset for effective modeling. The following tasks were performed:

- Converted the target variable from categorical labels ('M' for malignant and 'B' for benign) to numeric labels (1 and 0 respectively).
- Removed the 'id' column as it provides no predictive value.
- Checked for missing values and duplicates; none were found.
- Detected outliers in numerical features using boxplots and removed significant outliers to improve data quality.
- Applied MinMaxScaler to normalize feature values to the range [0,1], ensuring uniformity across all features for better model convergence.

5. Exploratory Data Analysis (EDA) & Visualization:

Exploratory data analysis was performed to better understand the dataset and reveal patterns:

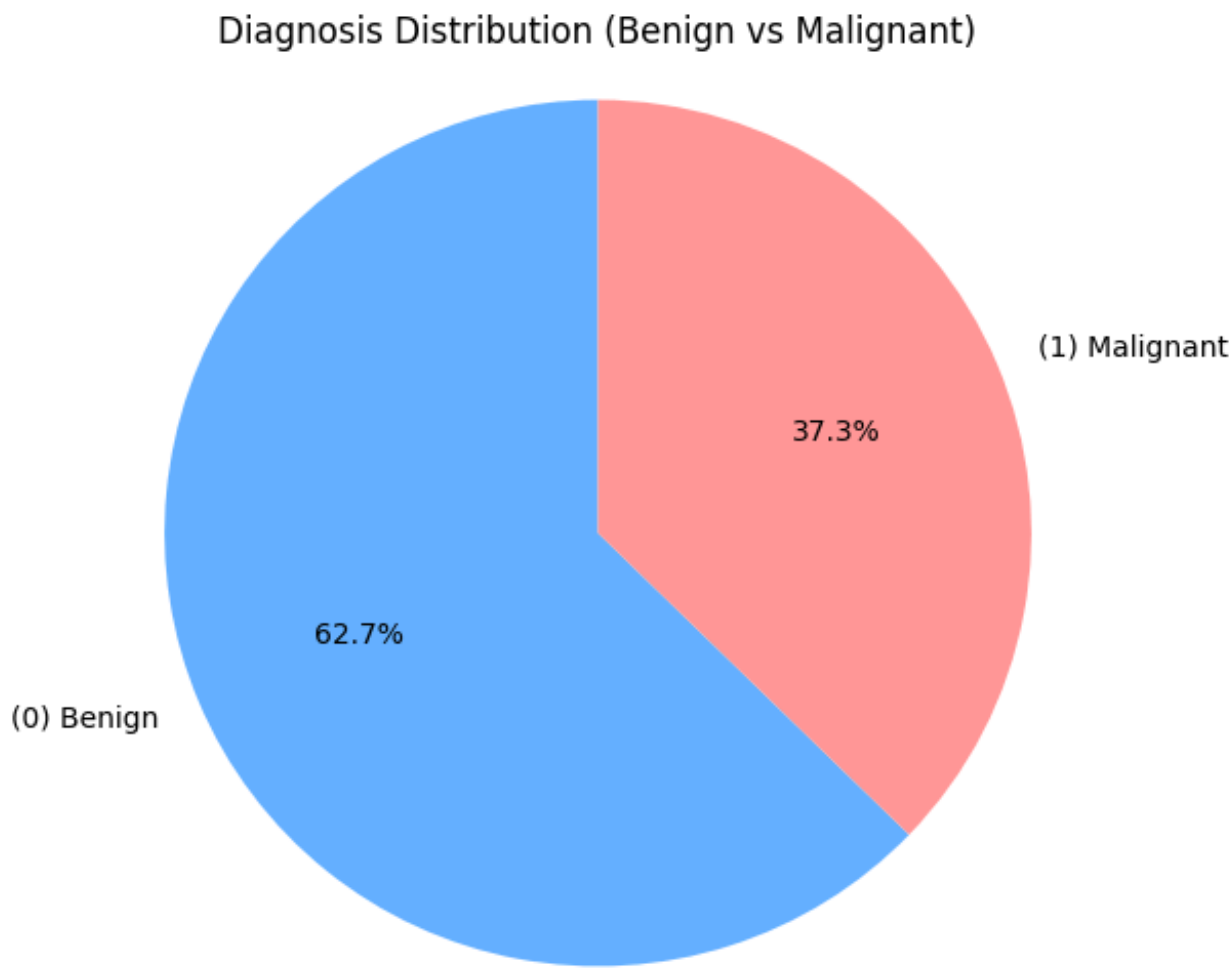
- Histograms to visualize the distribution of numerical features.



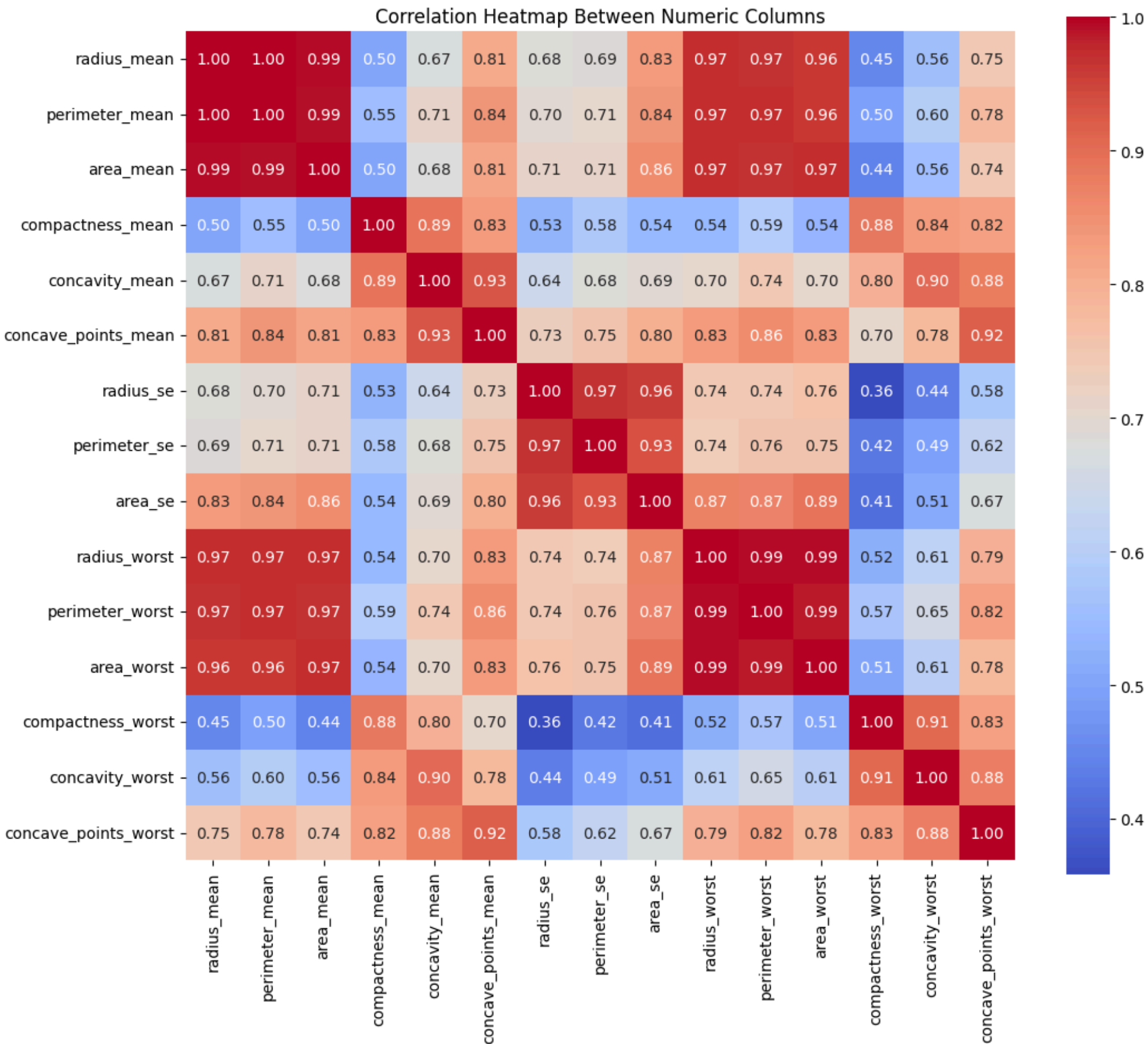
5. Exploratory Data Analysis (EDA) & Visualization:

Exploratory data analysis was performed to better understand the dataset and reveal patterns:

- Pie chart showing the target class distribution, with 63.7% benign and 37.3% malignant samples.



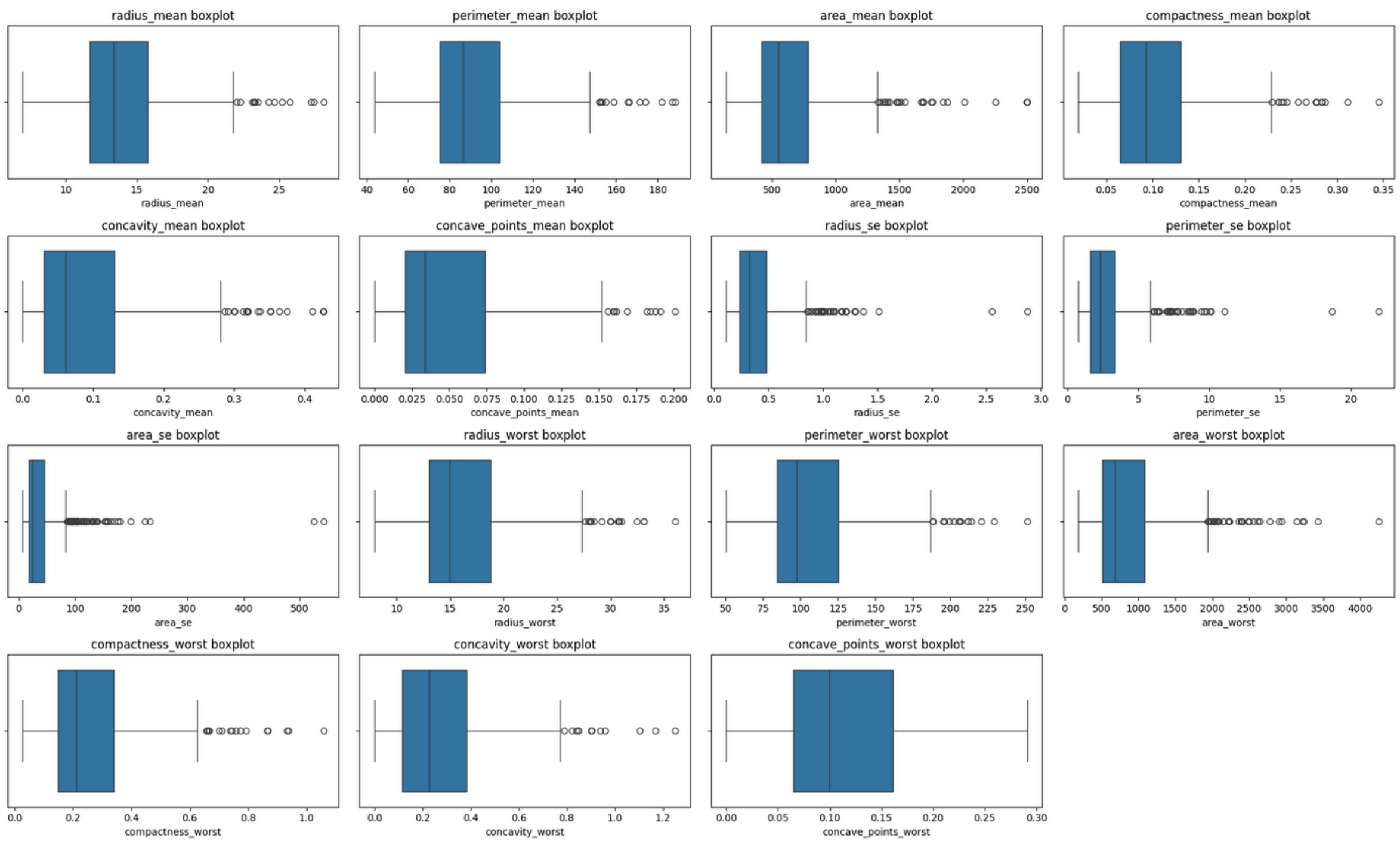
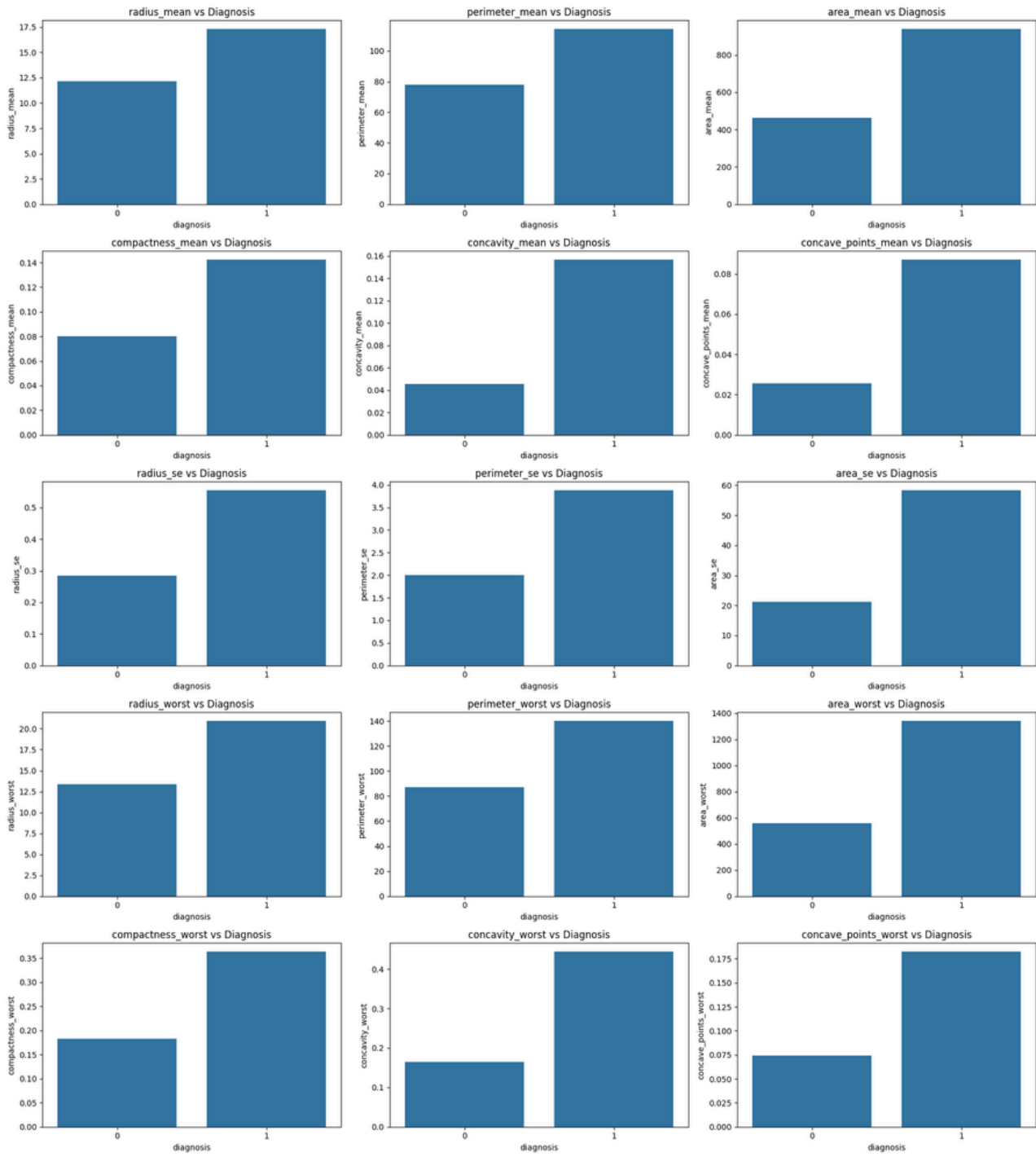
- Correlation heatmap illustrating the relationships between features and the target, helping identify highly correlated features.



5. Exploratory Data Analysis (EDA) & Visualization:

Exploratory data analysis was performed to better understand the dataset and reveal patterns:

- Bar plots comparing individual features against the target variable to observe their impact on classification.
- Boxplots to detect and visualize outliers in numerical features.



6. Feature Selection:

To improve model efficiency and interpretability, feature selection was carried out:

1. Features with correlation coefficients less than ± 0.5 with the target were dropped.
2. A Decision Tree Classifier was trained to calculate feature importance scores.

3. The top 7 most important features selected were:
concavity_worst, radius_mean,
concave_points_mean, area_se,
radius_worst, perimeter_worst,
concave_points_worst.
4. These features were normalized using MinMaxScaler and used for subsequent modeling.

7. Model Building – Logistic Regression:

1. Used the single most correlated feature `concave_points_worst` for a simple Logistic Regression model.

2. Split the data into training and testing sets.

3. Model evaluation on the test set showed:

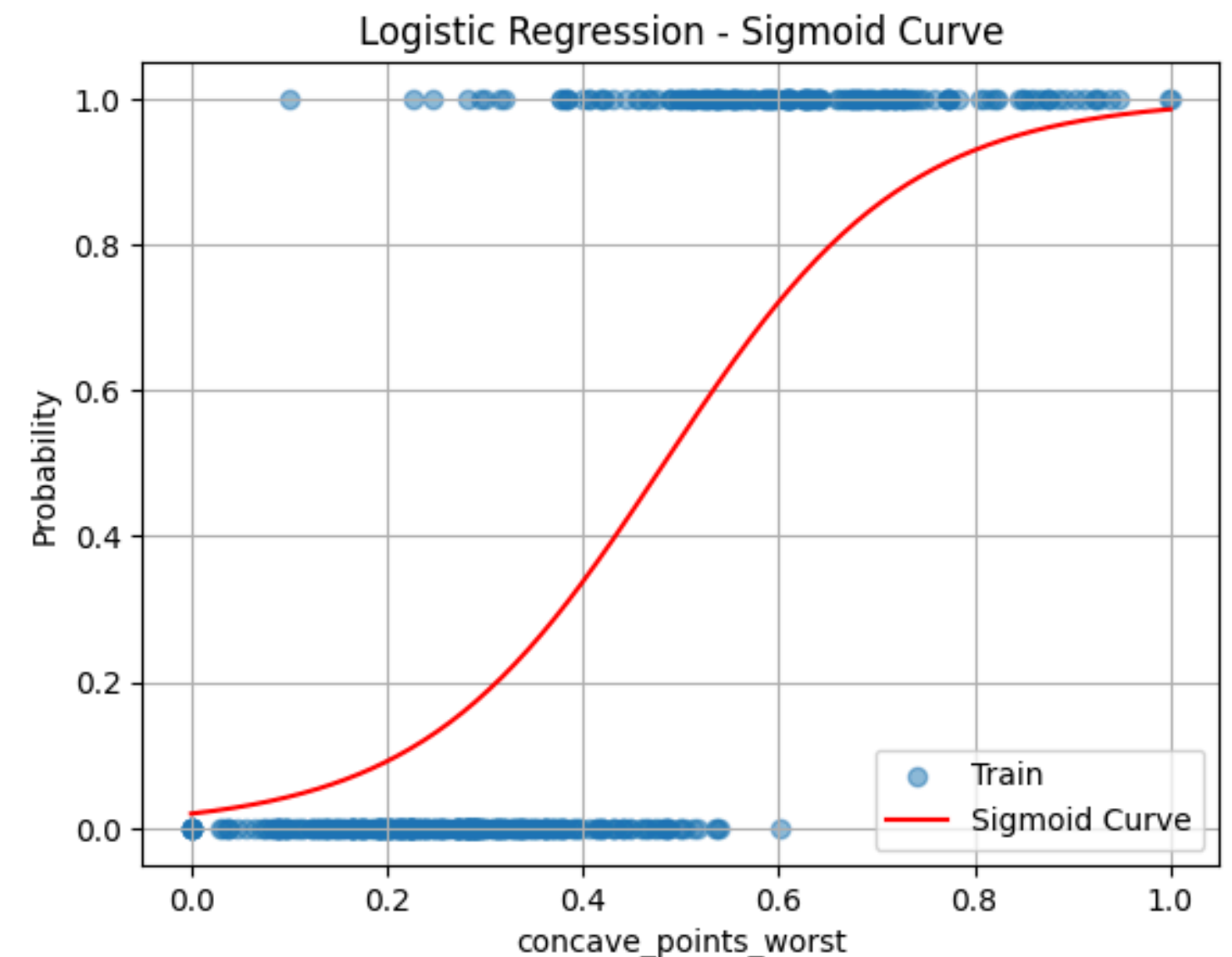
Accuracy: 91%

Precision: 92%

Recall: 84%

F1 Score: 88%

4. Visualized the sigmoid curve to represent the probability of malignancy relative to the feature.



8. Model Building – Decision Tree:

1. Trained a Decision Tree Classifier on the selected features with a maximum depth of 3 to reduce overfitting.

2. Performance metrics on the test set:

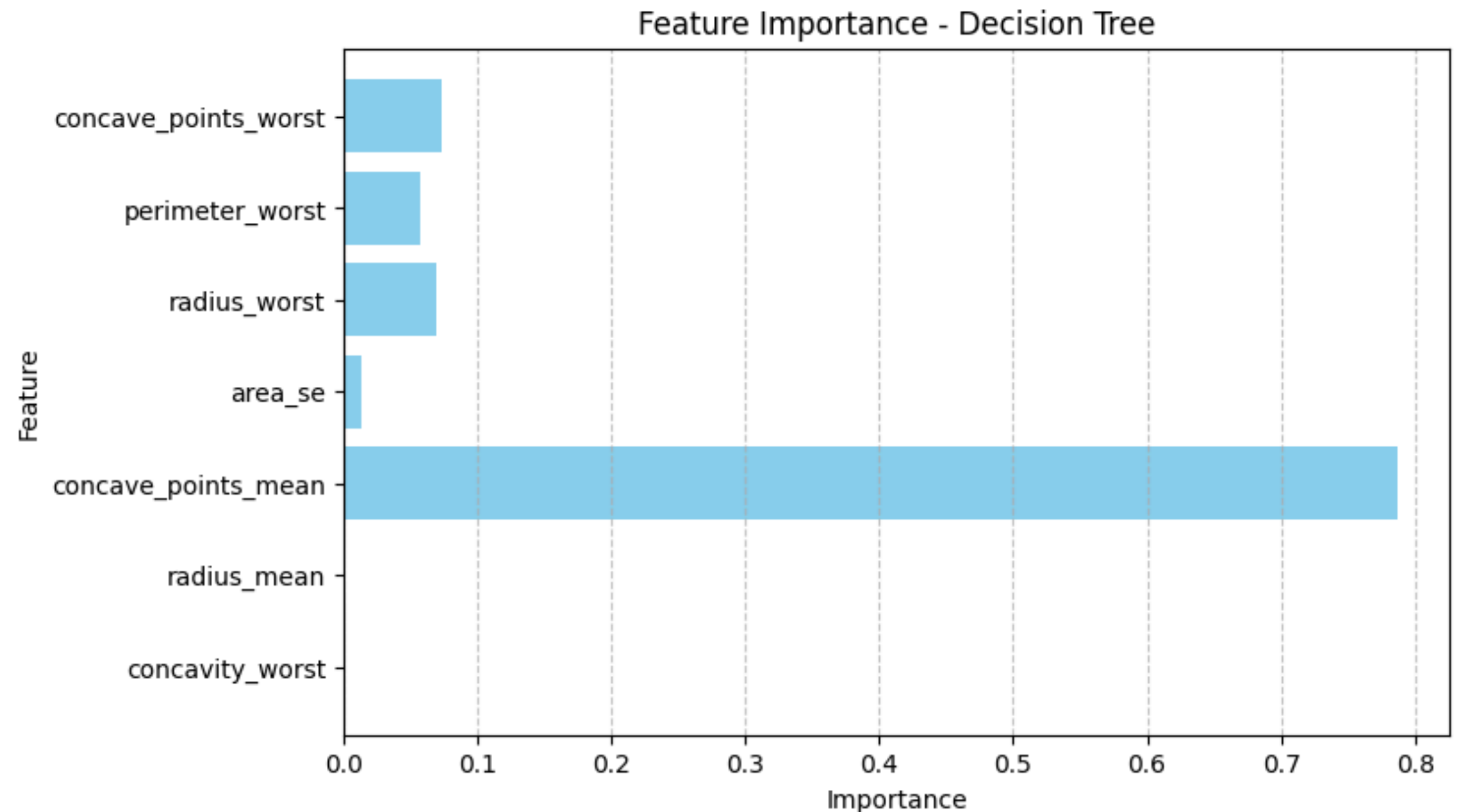
Accuracy: 94%

Precision: 95%

Recall: 88%

F1 Score: 92%

3. Visualized feature importance

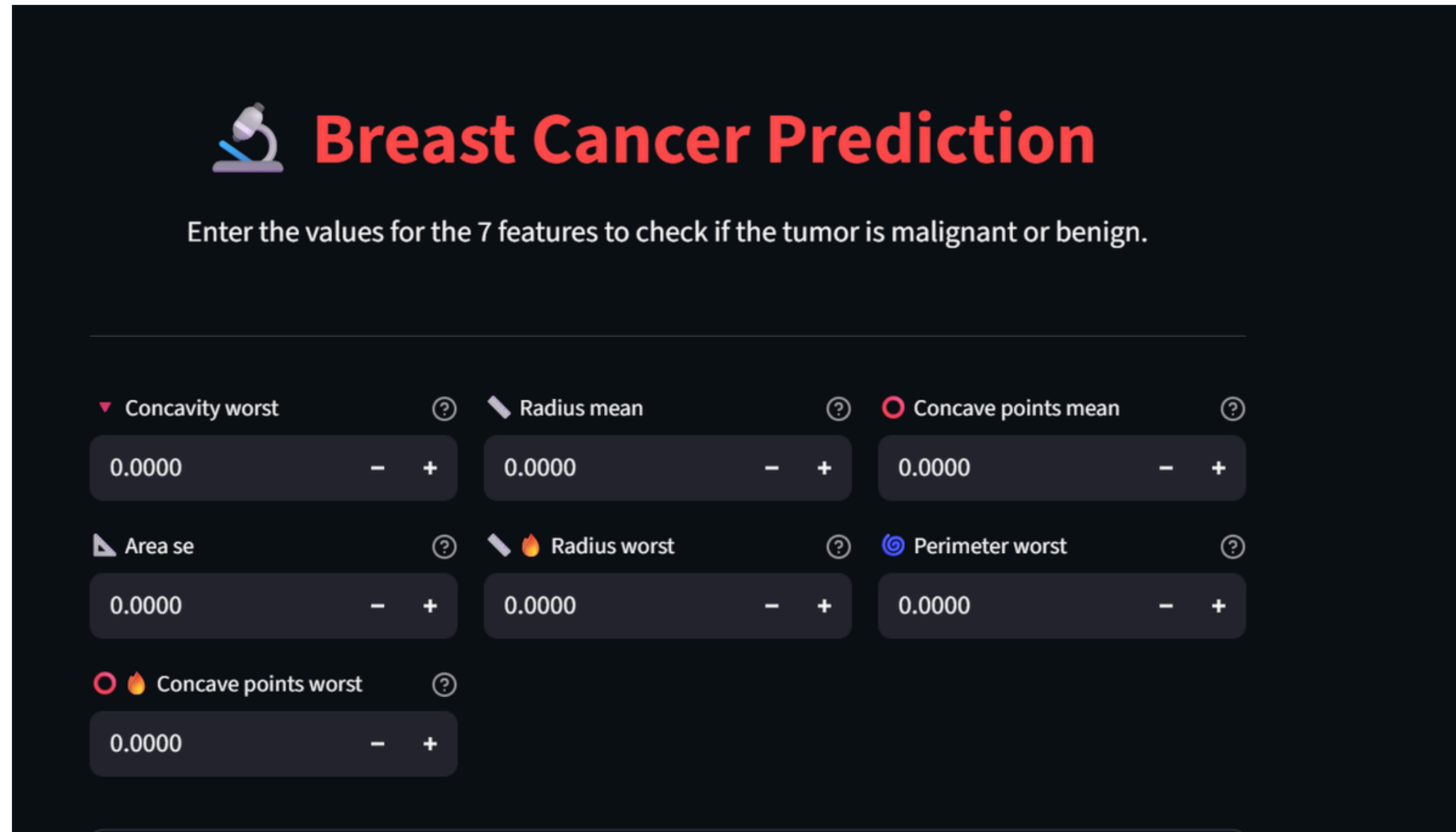


9. Final Model for the Application:

Saved the trained Decision Tree model and the MinMaxScaler using joblib.

These saved models are integrated into the web application, enabling real-time tumor classification based on user input

Limiting features to the selected seven reduces input complexity while maintaining high prediction accuracy.



The screenshot shows a web application titled "Breast Cancer Prediction" with a microscope icon. Below the title, it says "Enter the values for the 7 features to check if the tumor is malignant or benign." The interface features seven input fields, each with a feature name, a value of 0.0000, and minus/plus buttons. The features are: Concavity worst, Radius mean, Concave points mean, Area se, Radius worst, Perimeter worst, and Concave points worst. Each feature name is preceded by a small icon (triangle, circle, or flame) and a help icon (question mark).

Feature	Value
Concavity worst	0.0000
Radius mean	0.0000
Concave points mean	0.0000
Area se	0.0000
Radius worst	0.0000
Perimeter worst	0.0000
Concave points worst	0.0000

<https://breastcancerclassificationmodel-85.streamlit.app/>

10. Conclusion & Future Work:

- CONCLUSION:

THE PROJECT SUCCESSFULLY DEVELOPED AN ACCURATE AND INTERPRETABLE MODEL FOR BREAST CANCER CLASSIFICATION. THE DECISION TREE APPROACH BALANCES SIMPLICITY AND PERFORMANCE, MAKING IT WELL-SUITED FOR CLINICAL SUPPORT TOOLS.

- FUTURE WORK:

EXPLORE ENSEMBLE MODELS SUCH AS RANDOM FOREST OR XGBOOST TO POTENTIALLY IMPROVE ACCURACY.

PERFORM HYPERPARAMETER TUNING TO OPTIMIZE MODEL PARAMETERS.

EXPAND THE DATASET WITH MORE SAMPLES FOR BETTER GENERALIZATION.

Thank you