GRADED ASSIGNEMENT - 02.
(Mathematics and statistics for AI - Module 02)

1. Implementation details of each step -

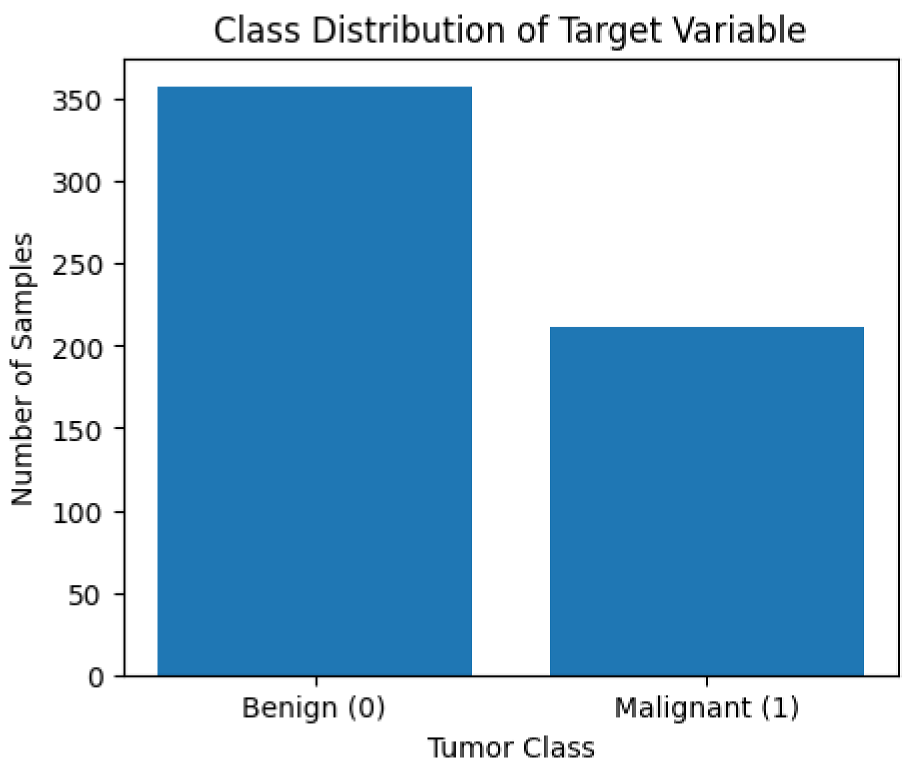
Part A: Data exploration and preprocessing.

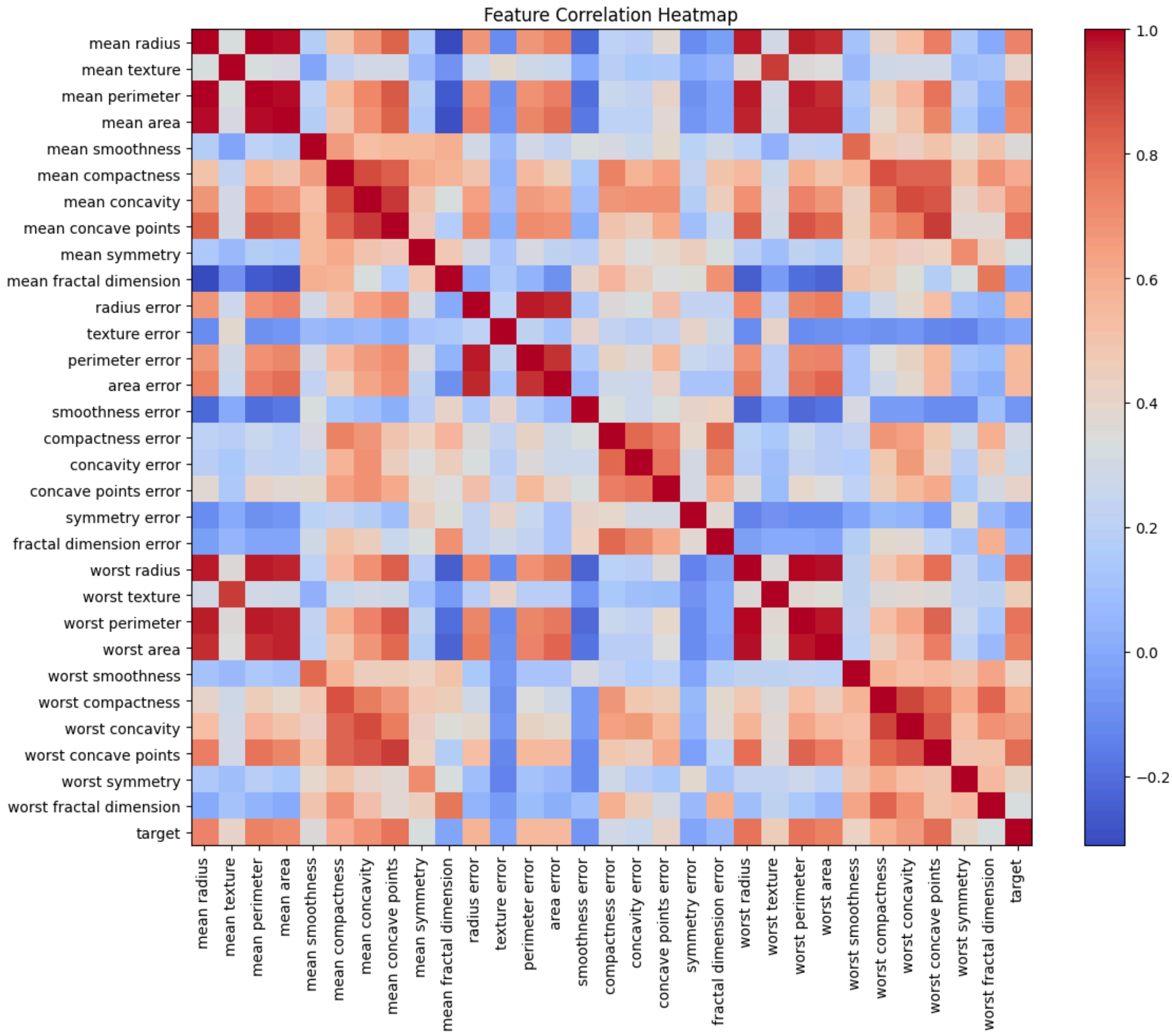
1. The dataset was first loaded and inspected to ensure data quality. Missing values were checked, and any identifier (ID) column present in the dataset was removed, as it does not contribute to the predictive performance of a logistic regression model.

2. To prepare the data for training, feature scaling was carried out using manual standardization. Before scaling, the dataset was split into training and testing sets to prevent data leakage and ensure a fair evaluation of the model.

3. The mean and standard deviation were calculated using only the training data and then applied consistently to both the training and test sets. Since the target variable was already provided in binary form, no additional label encoding was necessary.

4. EDA was performed to better understand the dataset. A class distribution plot was used to examine the balance between benign and malignant cases. Additionally, a correlation heatmap was generated to analyze relationships among features. The analysis revealed strong correlations between several size-related tumor features, which aligns with known medical characteristics of malignant tumors.





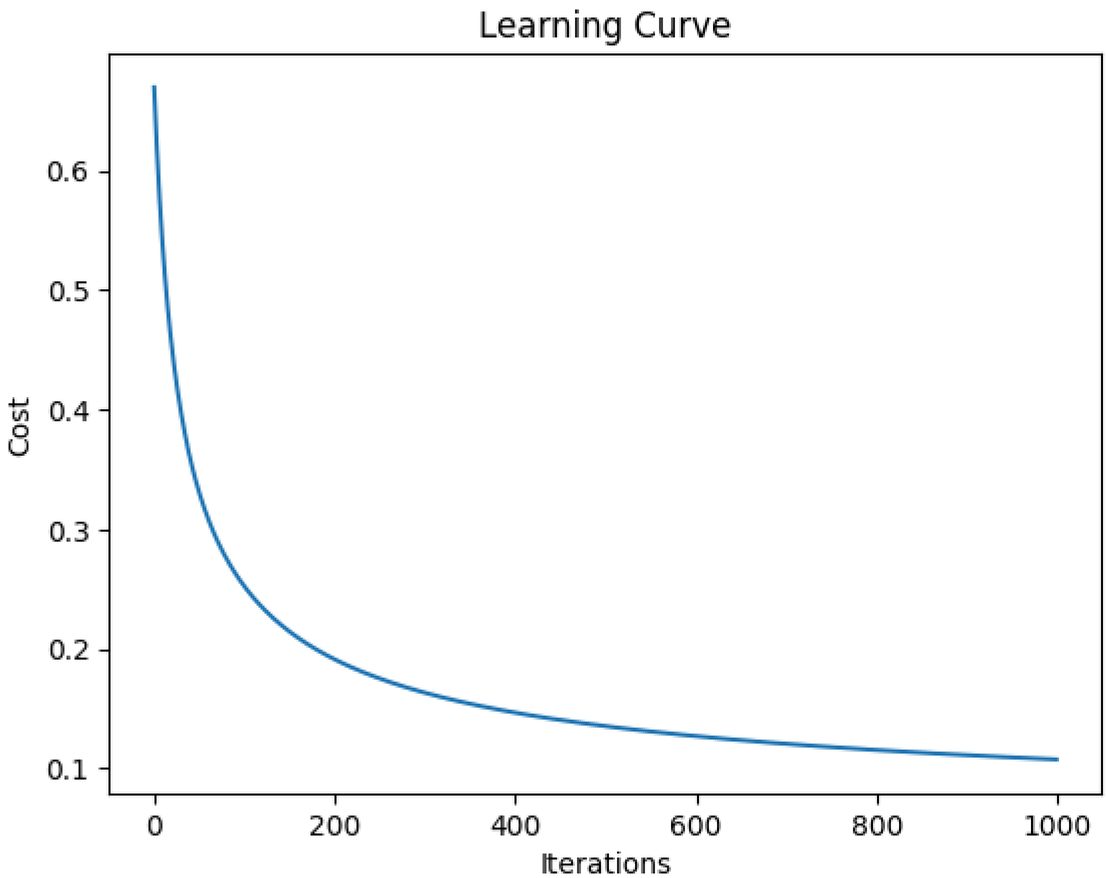
Part B: The mathematics of Logistic Regression.

1. The bias term was implemented separately from the feature matrix. The hypothesis function computes the weighted sum of input features and adds a bias term before applying the sigmoid function to obtain predicted probabilities.

2. The cost function measures how well the logistic regression model's predicted probabilities match the actual class labels. Binary cross-entropy loss was used, as it penalizes confident but incorrect predictions more heavily. The average loss over all training samples was computed using the predicted probabilities obtained from the hypothesis function.

3. The gradient computation measures how far the model's predictions are from the true labels and determines how each feature and the bias contribute to this error. Gradient descent then uses this information to iteratively update the model parameters in the direction that minimizes the loss. Over multiple iterations, the cost decreases, and the model learns to make more accurate predictions.

4. The learning curve shows a consistent decrease in the cost function over training iterations, followed by gradual stabilization. This indicates that gradient descent successfully converged to an optimal solution. The smooth and monotonic reduction in cost confirms the correctness of the gradient implementation, appropriate feature scaling, and a suitable learning rate.



Part C: Model Training and Evaluation.

1. The logistic regression model was trained using gradient descent on the training dataset. Model parameters, including weights and bias, were initialized with small random values and iteratively updated to minimize the binary cross-entropy loss. Training was performed for a fixed number of iterations, and convergence was verified using the learning curve.

2. After training, the learned weights and bias are used to compute predicted probabilities on the test data. These probabilities are then converted into class labels using a decision threshold of 0.5.

3. The trained model first computes probabilities for each test sample using the sigmoid function.
A threshold of 0.5 is applied:
Probability ≥ 0.5 → Malignant (1)
Probability < 0.5 → Benign (0)
This converts continuous probabilities into binary predictions.

4. Finally accuracy, precision and recall is calculated on the training set using standard formulae.

2. Summary of clustering results and insights -

Clustering was used as an exploratory analysis step to better understand the structure of the breast cancer dataset before and alongside model training. By grouping tumors based on their feature similarities, clustering helps reveal natural patterns in the data without using the actual diagnosis labels.

The clustering results showed that the data naturally forms **two main groups**, which closely correspond to **benign and malignant tumors**. One cluster mostly contained tumors with smaller, smoother, and more regularly shaped cell nuclei, which are typical characteristics of benign cases. The other cluster grouped tumors with larger sizes, irregular shapes, and higher concavity, which are commonly associated with malignant tumors.

Although clustering does not provide a diagnosis on its own, it clearly highlighted that **size and shape-based features play a major role in separating tumors**. This insight strongly supports the results obtained from the Logistic Regression model, where the same features were identified as the most important predictors.

Overall, the clustering analysis helped validate the dataset's structure and confirmed that malignant tumors are distinctly different from benign ones based on their physical characteristics. This reinforces confidence in using supervised models like Logistic Regression for accurate and reliable cancer prediction.