Breast Cancer Classification using Logistic Regression
Internship Task 4

1. Abstract

This project implements a robust binary classification system using Logistic Regression to predict whether a tumor is malignant or benign based on diagnostic features. The objective is to develop a reliable machine learning model, evaluate its performance using advanced metrics, and ensure high recall for medical diagnosis.

2. Problem Statement

To build an efficient and reliable classification model using Logistic Regression for early detection of breast cancer, ensuring high diagnostic accuracy and minimal false negatives.

3. Dataset Description

The Breast Cancer Wisconsin dataset consists of 569 samples with 30 diagnostic features extracted from digitized images of breast tissue. The target variable indicates whether the tumor is malignant or benign.

4. Methodology

• Data loading and inspection
• Feature scaling using StandardScaler
• Train-test split
• Logistic regression model training
• Model evaluation using confusion matrix, precision, recall, ROC-AUC
• Threshold tuning
• Feature importance analysis

5. Model Evaluation

The trained logistic regression model achieved high classification accuracy and excellent recall. Confusion matrix analysis revealed strong true positive rates. ROC-AUC evaluation confirmed the model's discriminative capability. Threshold tuning further enhanced recall, making the system more suitable for medical diagnosis.

6. Feature Importance Analysis

Feature coefficient analysis highlighted the most influential diagnostic parameters contributing to malignancy prediction. These features provide valuable insights into clinical decision-making and early detection of breast cancer.

7. Results and Discussion

The logistic regression classifier demonstrated strong predictive performance with high ROC-AUC scores. Advanced evaluation metrics and threshold optimization significantly improved the detection of malignant cases, reducing false negatives and enhancing clinical reliability.

8. Conclusion

This project successfully demonstrates a complete machine learning workflow for breast cancer classification using Logistic Regression. The integration of advanced evaluation metrics, threshold tuning, and feature interpretation ensures high model reliability, making it suitable for real-world healthcare applications.