Republic of the Philippines
Laguna State Polytechnic University
Province of Laguna
College of Computer Studies

Breast Cancer Diagnosis Analysis Report

Introduction

This report presents the analysis of the Breast Cancer Diagnosis Dataset using K-Nearest Neighbors (KNN) and Logistic Regression. The dataset contains various tumor characteristics, and the goal is to classify tumors as benign or malignant.

Data Exploration Process

- Data Loading

I started by loading the breast cancer dataset from a CSV file, which contains various tumor characteristics.

- Data Cleaning

I checked for missing values and outliers. Fortunately, there were no significant outliers, and all data points were complete.

- Feature Selection

I focused on relevant features for classification, such as radius, perimeter, area, and texture metrics, which are crucial in differentiating tumor types.

- Normalization

To ensure that all features contributed equally to the classification process, I standardized the features using normalization techniques.

K-Nearest Neighbors (KNN) Classifier

- Implementation

I trained a KNN classifier with n_neighbors=5 and evaluated its performance on the test data.

- Accuracy

The accuracy of the KNN model was calculated, and the confusion matrix was presented to show the performance of the classifier.

- Experiment with Different Values of n_neighbors

I experimented with different values of n_neighbors (3, 5, 7, 9) and plotted the accuracy against the number of neighbors. The optimal value of n_neighbors was determined based on the highest accuracy.

Logistic Regression Classifier

- Implementation

I trained a Logistic Regression model and evaluated its performance on the test data.

- Accuracy

The accuracy of the Logistic Regression model was calculated, and the confusion matrix and classification report were presented to show the performance of the classifier.

Comparison of KNN and Logistic Regression

- Accuracy Comparison

I compared the accuracies of the KNN and Logistic Regression models. The classification reports for both models were also compared.

Grid Search for Hyperparameter Tuning

- KNN Hyperparameter Tuning

I performed Grid Search Cross-Validation (GridSearchCV) to tune the hyperparameters of the KNN model. The best combination of parameters and the corresponding accuracy were reported.

Cross-Validation for Logistic Regression

- K-Fold Cross-Validation

I performed k-fold cross-validation on the Logistic Regression model (k=5) and reported the cross-validated accuracy.

Visualizing the Decision Boundary

- Principal Component Analysis (PCA)

I reduced the dimensionality of the dataset to 2D using PCA and visualized the decision boundaries of the KNN and Logistic Regression models.