Name: Sarfo K. Frimpong

Course: Data Mining

Assignment: Final Project Proposal

Project Proposal: Breast Cancer Classification

Abstract

Breast cancer remains one of the most prevalent cancers worldwide, making early and accurate diagnosis essential. This project focuses on developing a machine learning–based breast cancer classification system using the widely studied Wisconsin Diagnostic Breast Cancer (WDBC) dataset. The primary objective is to build, compare, and evaluate classification models capable of distinguishing between benign and malignant tumors. The workflow includes dataset preprocessing, feature normalization, implementation of 10-fold cross-validation, and performance benchmarking using multiple classifiers such as Support Vector Machines (SVM), Logistic Regression, Decision Trees, and K-Nearest Neighbors (KNN). Preprocessing steps include handling missing values, scaling all numeric features using standardization, and encoding the cancer diagnosis labels. The 10-fold cross-validation structure ensures model robustness by allowing every data point to serve as both training and testing data. Model performance will be assessed using metrics such as accuracy, precision, recall, F1-score, and ROC-AUC. The expected outcome is that SVM

with an RBF kernel will achieve the highest accuracy due to its strong performance on high-dimensional biomedical data. Ultimately, this project aims to demonstrate how supervised machine learning can support medical professionals by increasing diagnostic accuracy and reducing false negatives, contributing to improved patient outcomes.

Keywords

Breast Cancer; Machine Learning; Classification; Support Vector Machine; Cross-Validation.

Dataset Description

The project uses the Wisconsin Diagnostic Breast Cancer (WDBC) dataset from the UCI Machine Learning Repository. It contains 569 samples with 30 numeric features extracted from digitized FNA biopsy images. The dataset consists of two target classes: Malignant (212 samples) and Benign (357 samples). The features include measurements such as radius, texture, perimeter, area, smoothness, compactness, concavity, and fractal dimensions.

Method Description

Preprocessing includes encoding diagnosis labels (M=1, B=0), standardizing features using Z-score normalization, and checking for duplicates or missing values. A 10-fold cross-validation method will be used, with each fold containing 90% training data and 10% testing data. Multiple classifiers including SVM (RBF), Logistic Regression, KNN, and Decision Trees will be trained and

optimized using grid search for hyperparameter tuning. Each fold's prediction performance will be recorded and averaged across all folds.

Expected Results

Performance will be evaluated using accuracy, precision, recall, F1-score, and ROC-AUC. Based on prior findings, the SVM classifier with C=1 and gamma=0.01 is expected to provide the highest accuracy of approximately 98%.

References

1. UCI Machine Learning Repository: Wisconsin Diagnostic Breast Cancer Dataset.

2. Cristianini, N., & Shawe-Taylor, J. (2000). An Introduction to Support Vector Machines.

3. Pedregosa et al. (2011). Scikit-Learn: Machine Learning in Python.

4. Wolberg W.H., Mangasarian O.L. (1995). Breast Cancer Wisconsin Dataset Documentation.