PREDICTIVE ANALYTICS

Homework 3

Feature Selection and Data Classification Assignment

Fall 2024

Github link: https://github.com/saurav16997/Predicting-Breast-Cancer-using-Data-Mining

Results

This project aims to predict breast cancer survivability using data from the SEER database. The approach involves data preprocessing, feature selection, model training, and hyperparameter tuning.

Step 1: Data Cleaning and Preprocessing

Handling Missing Values: Applied Multiple Imputation by Chained Equations (MICE) for numeric variables and k-Nearest Neighbors (kNN) imputation for categorical variables.

Outlier Detection and Handling: Used Z-score method to detect outliers. For 'Tumor Size' and 'Reginol Node Positive', outliers were clipped at the 99th percentile.

Feature-Target Separation: Separated features (independent variables) from the target variable (survivability status).

Step 2: Feature Engineering

Standardization: Applied standardization to numerical features, scaling them to have a mean of 0 and standard deviation of 1.

Encoding Categorical Variables: Implemented one-hot encoding for categorical variables.

Step 3: Dimensionality Reduction

Applied Principal Component Analysis (PCA) to reduce the number of features from 40 to 14 principal components, retaining 95% of the variance.

Step 4: Model Implementation and Evaluation

Implemented and evaluated six different algorithms:

K-Nearest Neighbors (KNN)

Naïve Bayes

Decision Tree (C4.5)

Random Forest

Gradient Boosting

Neural Network

Model Performance Summary:

Rank	Model	Accuracy
1	**Gradient Boosting**	**0.898137**
2	**KNN**	**0.896894**
3	**Random Forest**	**0.895652**
4	**Naïve Bayes**	**0.888199**
5	**Neural Network**	**0.881988**
6	**Decision Tree**	**0.850932**

Step 5: Hyperparameter Tuning

Performed hyperparameter tuning on two models:

Random Forest:

Accuracy: **0.903**

Gradient Boosting:

Accuracy: **0.90062**

Conclusion

This project demonstrates a comprehensive approach to predicting breast cancer survivability. Through careful data preprocessing, feature engineering, and model evaluation, we achieved promising results. The Random Forest and Gradient Boosting models showed the highest accuracy, with further improvements seen after hyperparameter tuning.