Project Report

DSCI.315.UHSF1 Data Mining and Warehousing

-Manikanta Mamidi (0616233)

Breast Cancer (Binary Classification Prediction for type of Breast Cancer)

Breast cancer is the most prevalent cancer among women globally, representing approximately 25% of all cancer cases. In 2015 alone, it affected more than 2.1 million individuals. This disease typically begins when cells in the breast grow uncontrollably, often forming tumors that can be detected through X-rays or felt as lumps in the breast tissue.

One of the primary challenges in breast cancer detection is accurately classifying tumors as either malignant (cancerous) or benign (non-cancerous). This classification is critical, as it directly influences the course of treatment and patient outcomes. Machine learning, particularly Support Vector Machines (SVMs), can play a significant role in enhancing the accuracy of these classifications.

In this task, you are asked to perform a detailed analysis of breast cancer tumors using the Breast Cancer Wisconsin (Diagnostic) Dataset. The objective is to develop a machine learning model using SVMs to differentiate between malignant and benign tumors effectively.

Objective:

- Data Preparation, Cleaning and Analysis
- Build classification models to predict whether the cancer type is Malignant or Benign.
- Also fine-tune the hyperparameters & compare the evaluation metrics of various classification algorithms.

Breast Cancer Classification Analysis:

1. Overview

- The analysis focused on a dataset containing information about breast cancer tumors, categorized as either malignant (cancerous) or benign (non-cancerous). The goal was to develop machine learning models to accurately classify these tumors based on the features provided.

2. Data Preparation

- Dropping Unnecessary Columns: The dataset included an id column that wasn't useful for the classification task, so it was removed. Another column, Unnamed: 32, contained only NaN values and was also dropped.
- Target Variable and Features: The diagnosis column was identified as the target variable (the outcome we want to predict), while the other columns were used as input features for the models.
- No Missing Values: The target variable didn't have any missing values, ensuring clean data for training.

3. Data Splitting

The dataset was divided into a training set and a testing set using a 70-30 split. This ensures the model has enough data to learn from, while keeping a separate set to test its performance on unseen data.

4. Data Scaling

Before feeding the data into the models, the features were standardized using StandardScaler. This step is important for algorithms like Logistic Regression and SVM, which are sensitive to the scale of the data.

5. Model Training and Evaluation

Various machine learning models were trained and evaluated on the dataset. Below are the models used and their respective performance:

i. Logistic Regression
   - Training Accuracy: 96.4%
   - Test Accuracy: 96.4%
   - Overview: Logistic Regression performed exceptionally well, with high accuracy on both the training and test sets.

ii. Decision Tree Classifier
   - Training Accuracy: High, with potential overfitting
   - Test Accuracy: 92.8%
   - Overview: The Decision Tree model showed decent performance but might have overfitted the training data.

iii. Random Forest Classifier
   - Training Accuracy: High
   - Test Accuracy: 94.7%
   - Overview: As an ensemble of decision trees, Random Forest provided strong performance and was less prone to overfitting.

iv. Bagging Classifier
   - Test Accuracy: 92.8%
   - Overview: Bagging improved the stability of the Decision Tree model, though the accuracy was slightly lower than that of Random Forest.

v. AdaBoost Classifier
   - Test Accuracy: 88.8%
   - Overview: AdaBoost, a boosting method, showed good performance but was outperformed by other models.

vi. Gradient Boosting Classifier
   - Test Accuracy: 88.8%
   - Overview: Similar to AdaBoost, Gradient Boosting was effective but not the top performer.

vii. LightGBM Classifier
   - Test Accuracy: High
   - Overview: LightGBM, known for its speed and efficiency, performed well and is often used in large-scale applications.

viii. Stacked Generalization
   - Test Accuracy: Varies depending on the base models used
   - Overview: Stacked Generalization combines multiple models to improve overall prediction accuracy, leveraging the strengths of each model.

ix. Voting Classifier
   - Test Accuracy: Varies depending on the base models used
   - Overview: The Voting Classifier aggregates predictions from different models to produce a more robust final prediction.

x. K-Nearest Neighbors (KNN)
   - Training Accuracy: 98.2%
   - Test Accuracy: 94.7%
   - Overview: KNN performed well but showed signs of overfitting, as the accuracy dropped slightly from training to test data.

xi. Support Vector Classifier (SVC)
- Training Accuracy: 96.4%
- Test Accuracy: 96.4%
- Overview: After tuning its hyperparameters, SVC matched the performance of Logistic Regression, making it one of the top models.

6. Model Comparison

All the models were compared based on their accuracy scores. Logistic Regression and SVM emerged as the best performers, both achieving an accuracy of 96.4%. Ensemble methods like Stacked Generalization and Voting Classifier also performed well, highlighting the benefits of combining multiple models.

7. Conclusion

In this analysis, Logistic Regression and SVM were identified as the best models for classifying breast cancer tumors, each achieving an accuracy of 96.4%. While other models like Random Forest and ensemble methods also showed strong performance, Logistic Regression and SVM were the most consistent and reliable for this task.