TECHNICAL REPORT UTS MACHINE LEARNING

Breast Cancer Dataset

I Made Bayu Satria Wardhana 1103204145

PROGRAM STUDI TEKNIK KOMPUTER

FAKULTAS TEKNIK ELEKTRO UNIVERSITAS

TELKOM

2023

1. Introduction

Breast cancer is one of the most commonly diagnosed cancers among women worldwide, and early detection plays a crucial role in improving patient outcomes. Machine learning algorithms can assist in classifying breast cancer as either malignant or benign based on a variety of features extracted from medical images. In this project, we explored the use of three different classification algorithms: Decision Tree, Random Forest, and Self-Training, to predict breast cancer outcomes based on features of the tumor.

2. Data

We used the Breast Cancer Wisconsin (Diagnostic) dataset from the Scikit-Learn library, which contains 569 instances and 30 features, including the mean, standard error, and maximum value of cell nucleus attributes such as radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry, and fractal dimension. The dataset also includes the target variable indicating whether the tumor is malignant or benign.

3. Data

We used the Seaborn library to create a pair plot, which shows pairwise relationships between different features of the dataset. We plotted the mean radius, mean texture, mean perimeter, and mean area features against each other and colored the points based on the target variable. The plot showed that the features have a clear separation between the two classes, making it suitable for classification.

4. Classification Algorithm

We trained three classification algorithms on the dataset: Decision Tree, Random Forest, and Self-Training.

- Decision Tree

Decision trees are a type of supervised learning algorithm used for classification problems. They construct a model in the form of a tree structure where each internal node represents a test on a feature, each branch represents the outcome of the test, and each leaf node represents a class label. We used the DecisionTreeClassifier class from Scikit-Learn library to train a decision tree model on our dataset. We achieved an accuracy of 0.91 on the test set.

- Random Forest

Random forests are a type of ensemble learning method for classification, regression, and other tasks that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes of the individual trees. We used the RandomForestClassifier class from Scikit-Learn library to train a random forest model on our dataset. We achieved an accuracy of 0.96 on the test set.

- Self-Training

  Self-training is a type of semi-supervised learning method where the algorithm starts with a small labeled dataset and iteratively adds to it by making predictions on the unlabeled data and adding the high-confidence predictions to the labeled dataset. We used the SelfTrainingClassifier class from Scikit-Learn library to train a self-training model on our dataset. We used the DecisionTreeClassifier as the base classifier and set the maximum number of iterations to 50. We achieved an accuracy of 0.92 on the test set.

5. Data Process

- Load the dataset: The first step is to load the breast cancer dataset into your Python environment using a library such as scikit-learn.
- Explore the data: Use exploratory data analysis (EDA) techniques to understand the structure of the data and identify any issues such as missing values, outliers, or imbalanced classes. This can be done using techniques such as scatterplots, histograms, and box plots.
- Preprocess the data: Once the issues with the data have been identified, preprocess the data to address them. This can include imputing missing values, scaling the features, and handling outliers. Additionally, feature selection or extraction techniques can be applied to reduce the dimensionality of the data.
- Split the data: Divide the data into training and testing sets using techniques such as cross-validation or holdout validation.
- Train models: Use machine learning algorithms such as decision trees, random forests, and self-training classifiers to train models on the training set.
- Evaluate models: Evaluate the performance of the models on the testing set using metrics such as accuracy, precision, recall, and F1-score.
- Tune hyperparameters: Fine-tune the hyperparameters of the models using techniques such as grid search or randomized search to optimize their performance.
- Select the best model: Select the best model based on its performance on the testing set.
- Deploy the model: Deploy the selected model in a real-world application and monitor its performance.

By following these steps, you can effectively process the breast cancer dataset and build machine learning models that can accurately classify breast cancer tumors.

6. Dataset Content

The dataset consists of 569 instances, each with 30 numeric features computed from digitized images of a fine needle aspirate (FNA) of a breast mass. These features describe characteristics of the cell nuclei present in the image. The target variable is a binary classification indicating whether the mass is malignant (M) or benign (B).

Here are the names and descriptions of the 30 features:

- mean radius: mean of distances from center to points on the perimeter
- mean texture: standard deviation of gray-scale values
- mean perimeter: mean size of the core tumor
- mean area: mean area of the core tumor
- mean smoothness: mean of local variation in radius lengths
- mean compactness: mean of perimeter^2 / area - 1.0
- mean concavity: mean of severity of concave portions of the contour
- mean concave points: mean for number of concave portions of the contour
- mean symmetry: mean symmetry
- mean fractal dimension: mean "coastline approximation" - 1
- radius error: standard error of distances from center to points on the perimeter
- texture error: standard error of gray-scale values
- perimeter error: error of size of the core tumor
- area error: error of area of the core tumor
- smoothness error: standard error of local variation in radius lengths
- compactness error: standard error of perimeter^2 / area - 1.0
- concavity error: standard error of severity of concave portions of the contour
- concave points error: standard error for number of concave portions of the contour
- symmetry error: standard error of symmetry
- fractal dimension error: standard error of "coastline approximation" - 1
- worst radius: "worst" or largest mean value for mean distances from center to points on the perimeter
- worst texture: "worst" or largest mean value for standard deviation of gray-scale values
- worst perimeter: "worst" or largest mean value for size of the core tumor
- worst area: "worst" or largest mean value for area of the core tumor

- worst smoothness: "worst" or largest mean value for local variation in radius lengths
- worst compactness: "worst" or largest mean value for perimeter^2 / area - 1.0
- worst concavity: "worst" or largest mean value for severity of concave portions of the contour
- worst concave points: "worst" or largest mean value for number of concave portions of the contour
- worst symmetry: "worst" or largest mean value for symmetry
- worst fractal dimension: "worst" or largest mean value for "coastline approximation" - 1

7. Conclusion

In this project, we explored the use of three classification algorithms: Decision Tree, Random Forest, and Self-Training, to predict breast cancer outcomes based on features of the tumor. The Random Forest algorithm performed the best, achieving an accuracy of 0.96 on the test set. The Decision Tree and Self-Training algorithms also performed well, achieving accuracies of 0.91 and 0.92, respectively. The Seaborn library was used to visualize the data trends, which provided useful insights into the separability of the dataset. Overall, the project demonstrates the effectiveness of machine learning algorithms in predicting breast cancer outcomes based on medical image features.