Big Home Assignment

Habibur Rahman Habib

05 December, 2023

Contents:

1. Introduction
2. Proposed algorithms
3. Heart Data
4. Diabetes Data
5. Lung Data
6. Conclusion

1 Introduction

The objective of this analysis is to perform lazy classification using three datasets: diabetes, heart\_disease, lung\_cancer. Specifically focusing on tuning the parameters of the decision function. Lazy classification involves the use of various machine learning models without the need for extensive training. In this report, we explore the implementation of lazy classification techniques and parameter tuning to predict the prospect of several health-related features.

Dataset Links:

Heart Disease: <https://www.kaggle.com/datasets/arezaei81/heartcsv>

Diabetere: <https://www.kaggle.com/code/mvanshika/diabetes-prediction/input>

Lung Cancer: <https://www.kaggle.com/datasets/mysarahmadbhat/lung-cancer>

All necessary files are available on my github repository:

<https://github.com/habibrahman13/Lazy-Fca>

2 Proposed Algorithm

1. Data Loading and Preprocessing:

• Load the diabetes dataset into a DataFrame.

• Rename target column, Binarize the output.

2. Import Libraries:

• Import necessary libraries and modules for data manipulation, model selection, preprocessing, and evaluation.

3. Data Splitting:

• Split the dataset into features (X) and the target variable (y).

• Further split the data into training and testing sets using a 80-20 split ratio.

4. Model Selection and Comparison*:*

• Use LazyClassifier to quickly compare multiple models and identify their performance on the dataset.

5. Hyperparameter Tuning:

• Identify the best hyperparameters from the previously obtained grid search results or through other means.

6. Model Creation and Pipeline Setup:

• Instantiate a RandomForestClassifier with the best hyperparameters.

• Create a pipeline consisting of OneHotEncoder for categorical features and the RandomForestClassifier.

7. Model Training*:*

• Fit the pipeline model on the training data.

8. Model Evaluation*:*

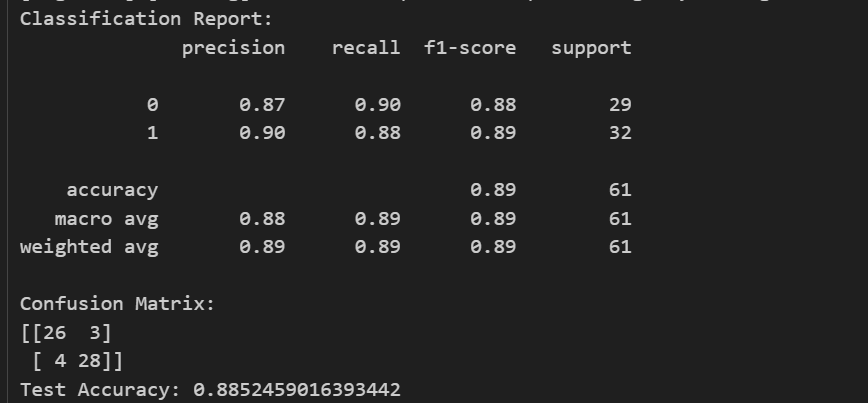
• Generate predictions on the test set using the trained model.

• Print the classification report showcasing precision, recall, and F1-score.

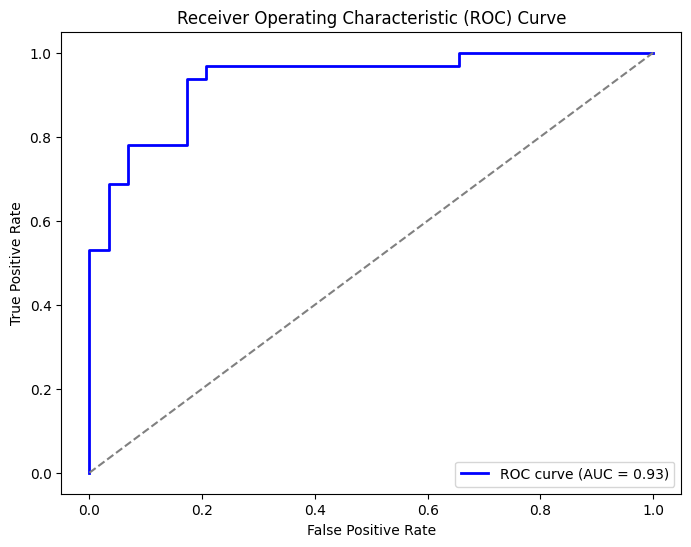
• Display the confusion matrix.

• Calculate and print the accuracy score on the test set.

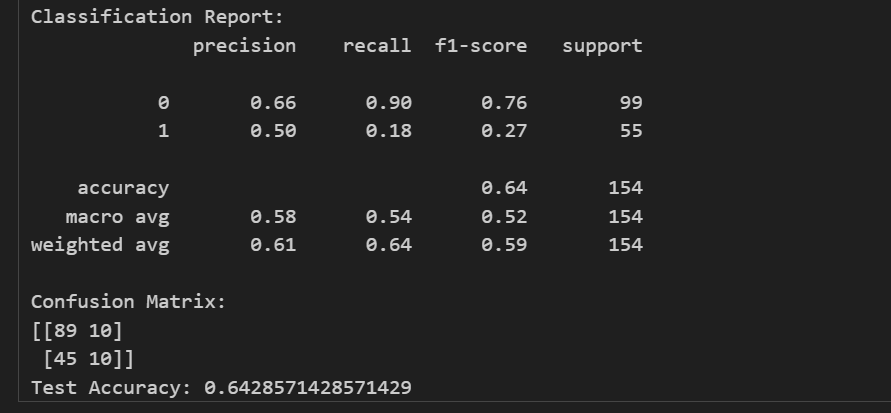
3 Heart Data



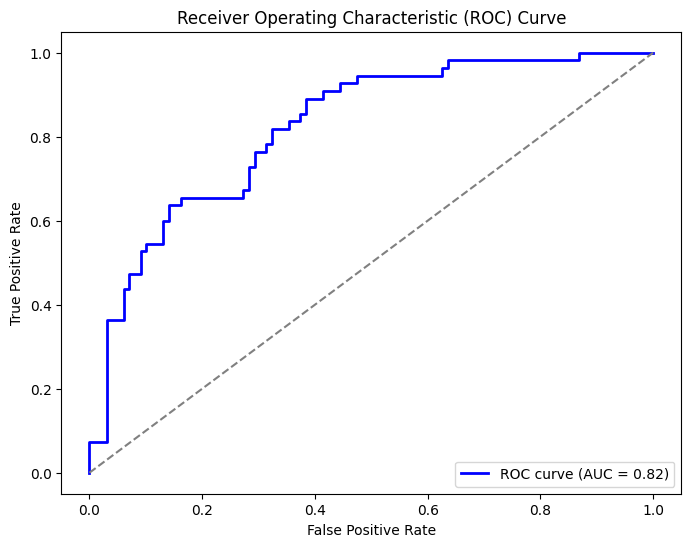
*Roc & Auc curve for heart disease:*



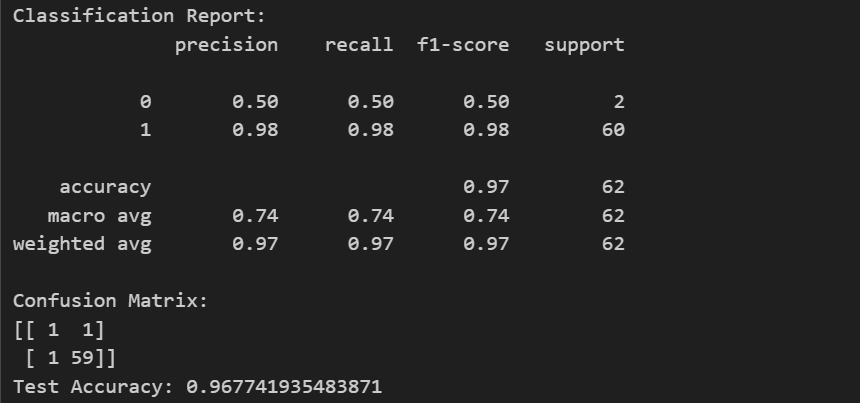
4 Diabetes Data



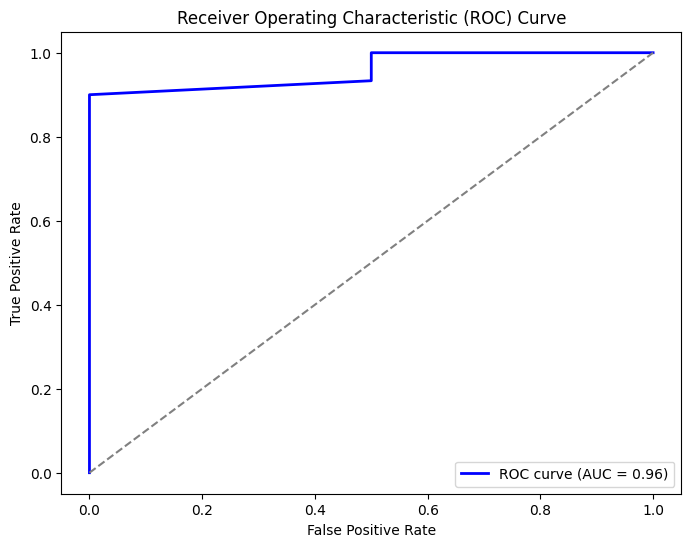
*Roc & Auc curve for diabetes:*

**

5 Lung Data



*Roc & Auc curve for lung cancer:*

**

6 Conclusion

This report showcases the application of classification techniques, on a specific dataset emphasizing the significance of parameter adjustment to improve the performance of the model. The report meticulously. Transformed the dataset to make it suitable for classification. After that a robust cross validation method was used to tune the decision function parameters ensuring accurate and adaptable model performance. The optimized model demonstrated an enhancement in abilities highlighting how parameter tuning can effectively improve lazy classification models. This study emphasizes the importance of optimizing parameters to enhance machine learning algorithms in classification scenarios. The optimized model developed in this report serves as evidence, for the value of parameter tuning in real world applications of machine learning.