Big Homework

Syed Taqi Ali

This report presents the results of lazy classification on three datasets: Breast Cancer, Heart, and Employee. The analysis includes the tuning of decision function parameters through a cross-validation procedure. The code and details are available in the GitHub repository https://github.com/Taqiali5/OSDA_BigHome_Assingment.git

1. Employee

The dataset has 8 entities, one for each employee.

Education: The educational qualifications of employees, including degree.

Joining Year: The year each employee joined the company, indicating their length of service.

City: The location or city where each employee is based or works.

Payment Tier: Categorization of employees into different salary tiers.

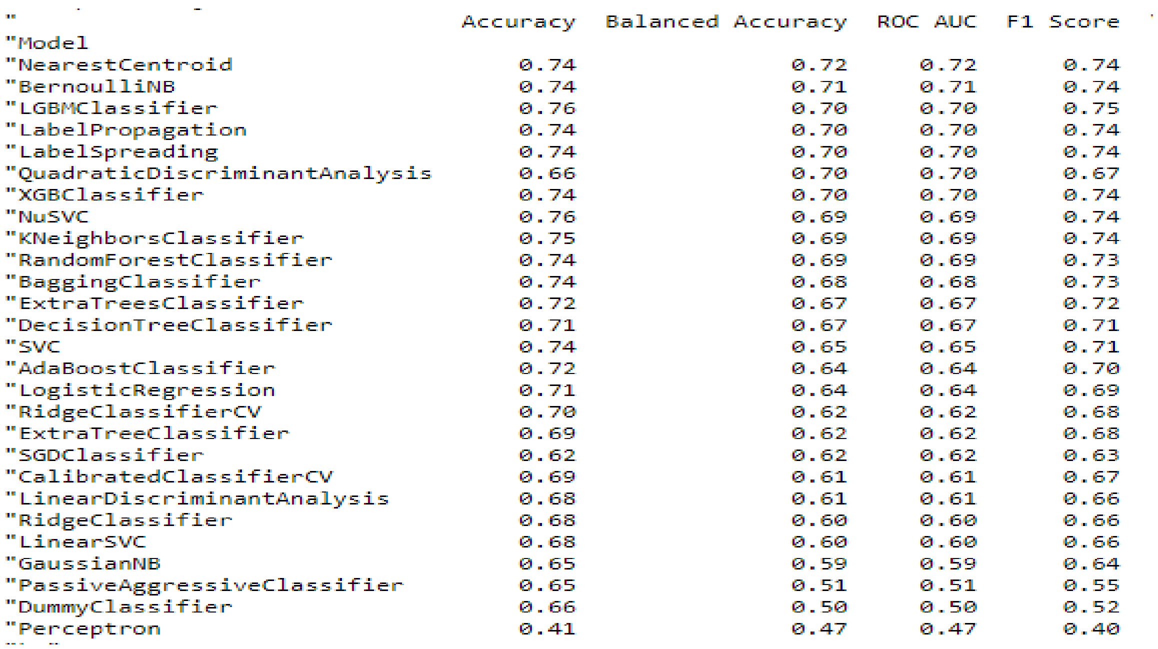
Age: The age of each employee, providing demographic insights.

Gender: Gender identity of employees, promoting diversity analysis.

Ever Benched: Indicates if an employee has ever been temporarily without assigned work.

Experience in Current Domain: The number of years of experience employees have in their current field.

Leave or Not: a target column

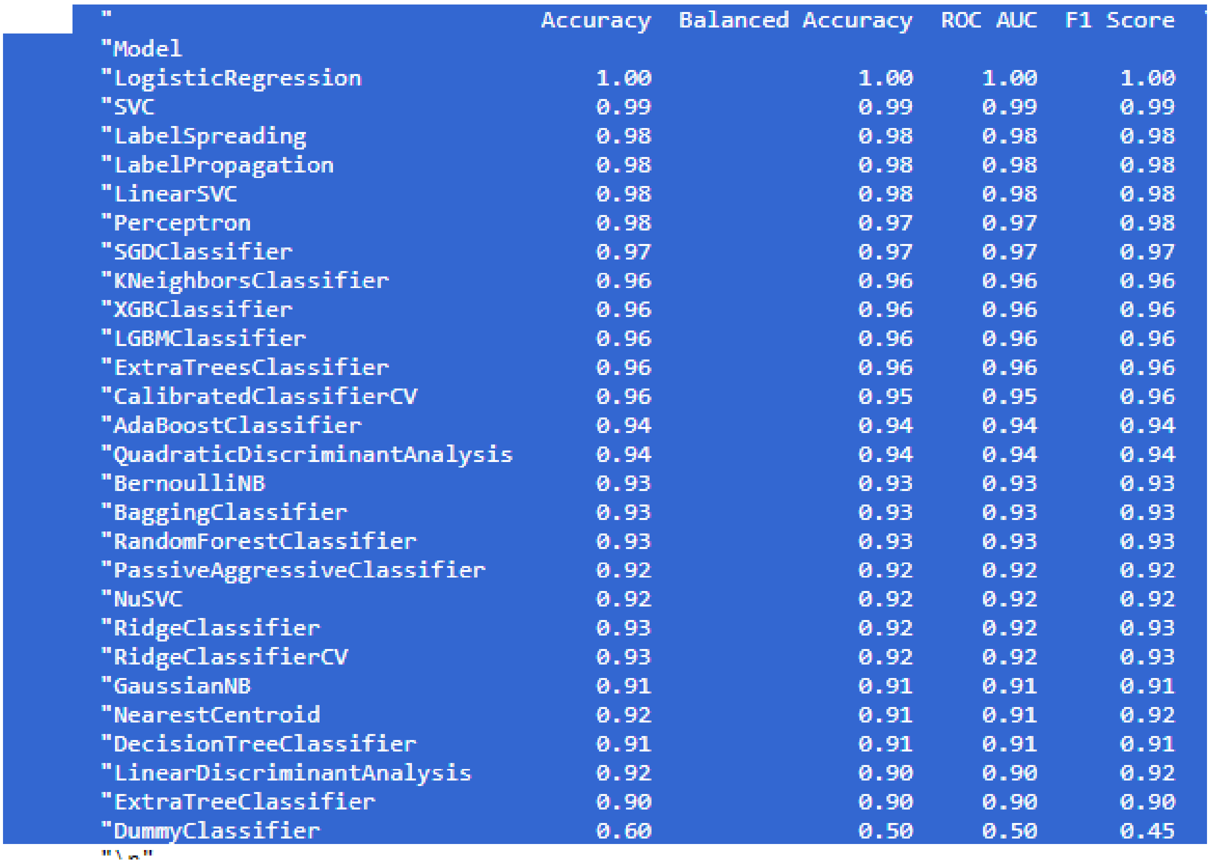


```
"                                  Accuracy  Balanced Accuracy  ROC AUC  F1 Score
"Model
"NearestCentroid                       0.74               0.72     0.72      0.74
"BernoulliNB                           0.74               0.71     0.71      0.74
"LGBMClassifier                        0.76               0.70     0.70      0.75
"LabelPropagation                      0.74               0.70     0.70      0.74
"LabelSpreading                        0.74               0.70     0.70      0.74
"QuadraticDiscriminantAnalysis         0.66               0.70     0.70      0.67
"XGBClassifier                         0.74               0.70     0.70      0.74
"NuSVC                                 0.76               0.69     0.69      0.74
"KNeighborsClassifier                  0.75               0.69     0.69      0.74
"RandomForestClassifier                0.74               0.69     0.69      0.73
"BaggingClassifier                     0.74               0.68     0.68      0.73
"ExtraTreesClassifier                  0.72               0.67     0.67      0.72
"DecisionTreeClassifier                0.71               0.67     0.67      0.71
"SVC                                   0.74               0.65     0.65      0.71
"AdaBoostClassifier                    0.72               0.64     0.64      0.70
"LogisticRegression                    0.71               0.64     0.64      0.69
"RidgeClassifierCV                     0.70               0.62     0.62      0.68
"ExtraTreeClassifier                   0.69               0.62     0.62      0.68
"SGDClassifier                         0.62               0.62     0.62      0.63
"CalibratedClassifierCV                0.69               0.61     0.61      0.67
"LinearDiscriminantAnalysis            0.68               0.61     0.61      0.66
"RidgeClassifier                       0.68               0.60     0.60      0.66
"LinearSVC                             0.68               0.60     0.60      0.66
"GaussianNB                            0.65               0.59     0.59      0.64
"PassiveAggressiveClassifier           0.65               0.51     0.51      0.55
"DummyClassifier                       0.66               0.50     0.50      0.52
"Perceptron                            0.41               0.47     0.47      0.40
```

2. Breast Cancer Dataset

Objects: 500 instances of Malignant and Benign samples.

- Mean Compactness
- Mean Concavity
- Mean Concave Points
- Mean Symmetry
- Mean Fractal Dimension
- Mean Radius
- Mean Texture
- Mean Perimeter
- Mean Area
- Mean Smoothness

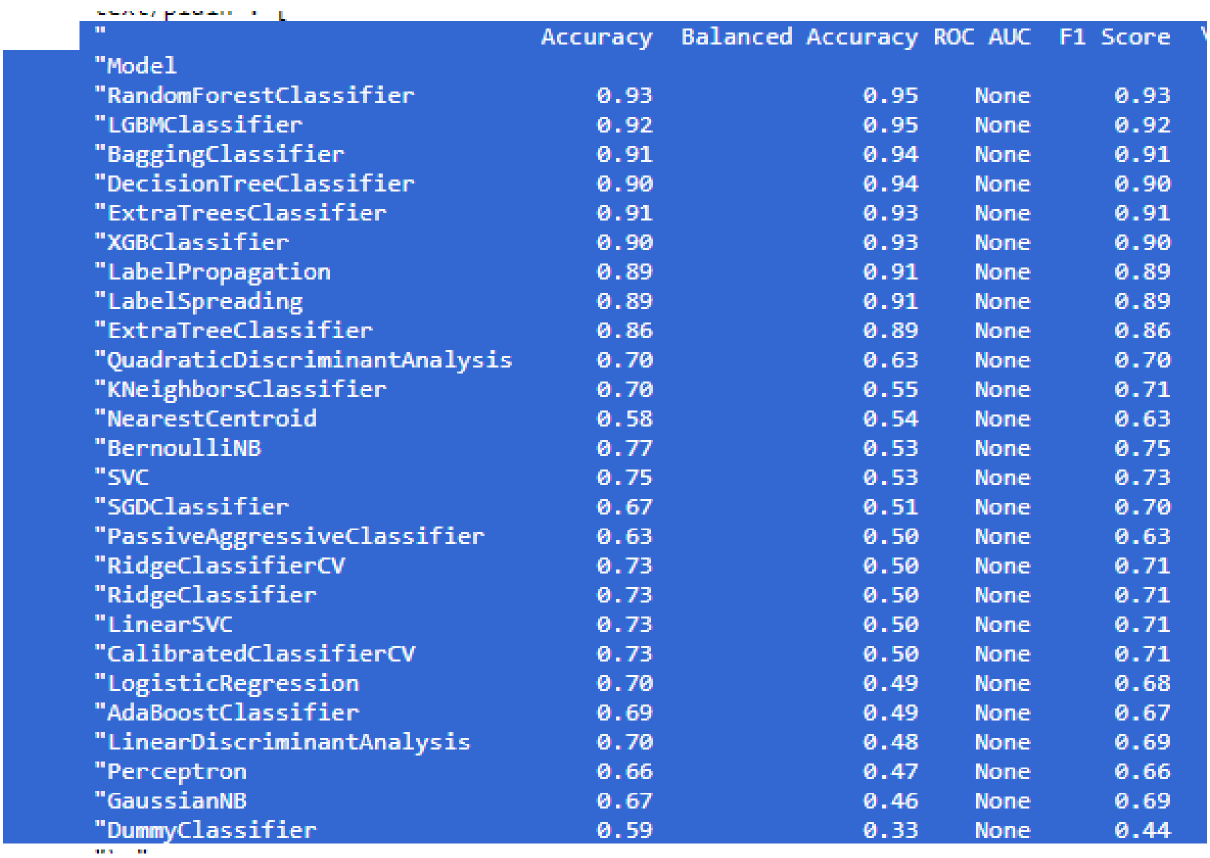


```
"                               Accuracy  Balanced Accuracy  ROC AUC  F1 Score
"Model
"LogisticRegression                 1.00               1.00     1.00      1.00
"SVC                                0.99               0.99     0.99      0.99
"LabelSpreading                     0.98               0.98     0.98      0.98
"LabelPropagation                   0.98               0.98     0.98      0.98
"LinearSVC                          0.98               0.98     0.98      0.98
"Perceptron                         0.98               0.97     0.97      0.98
"SGDClassifier                      0.97               0.97     0.97      0.97
"KNeighborsClassifier               0.96               0.96     0.96      0.96
"XGBClassifier                      0.96               0.96     0.96      0.96
"LGBMClassifier                     0.96               0.96     0.96      0.96
"ExtraTreesClassifier               0.96               0.96     0.96      0.96
"CalibratedClassifierCV             0.96               0.95     0.95      0.96
"AdaBoostClassifier                 0.94               0.94     0.94      0.94
"QuadraticDiscriminantAnalysis      0.94               0.94     0.94      0.94
"BernoulliNB                        0.93               0.93     0.93      0.93
"BaggingClassifier                  0.93               0.93     0.93      0.93
"RandomForestClassifier             0.93               0.93     0.93      0.93
"PassiveAggressiveClassifier        0.92               0.92     0.92      0.92
"NuSVC                              0.92               0.92     0.92      0.92
"RidgeClassifier                    0.93               0.92     0.92      0.93
"RidgeClassifierCV                  0.93               0.92     0.92      0.93
"GaussianNB                         0.91               0.91     0.91      0.91
"NearestCentroid                    0.92               0.91     0.91      0.92
"DecisionTreeClassifier             0.91               0.91     0.91      0.91
"LinearDiscriminantAnalysis         0.92               0.90     0.90      0.92
"ExtraTreeClassifier                0.90               0.90     0.90      0.90
"DummyClassifier                    0.60               0.50     0.50      0.45
"\n"
```

3. **Heart**

The features are

Age, Sex, cp, trestbps, chol, fbs, restecg, thalach, exang, oldpeak, slope, ca, thal and target.



```
text/plain" : [
"                                  Accuracy  Balanced Accuracy ROC AUC  F1 Score
"Model
"RandomForestClassifier                0.93               0.95    None      0.93
"LGBMClassifier                        0.92               0.95    None      0.92
"BaggingClassifier                     0.91               0.94    None      0.91
"DecisionTreeClassifier                0.90               0.94    None      0.90
"ExtraTreesClassifier                  0.91               0.93    None      0.91
"XGBClassifier                         0.90               0.93    None      0.90
"LabelPropagation                      0.89               0.91    None      0.89
"LabelSpreading                        0.89               0.91    None      0.89
"ExtraTreeClassifier                   0.86               0.89    None      0.86
"QuadraticDiscriminantAnalysis         0.70               0.63    None      0.70
"KNeighborsClassifier                  0.70               0.55    None      0.71
"NearestCentroid                       0.58               0.54    None      0.63
"BernoulliNB                           0.77               0.53    None      0.75
"SVC                                   0.75               0.53    None      0.73
"SGDClassifier                         0.67               0.51    None      0.70
"PassiveAggressiveClassifier           0.63               0.50    None      0.63
"RidgeClassifierCV                     0.73               0.50    None      0.71
"RidgeClassifier                       0.73               0.50    None      0.71
"LinearSVC                             0.73               0.50    None      0.71
"CalibratedClassifierCV                0.73               0.50    None      0.71
"LogisticRegression                    0.70               0.49    None      0.68
"AdaBoostClassifier                    0.69               0.49    None      0.67
"LinearDiscriminantAnalysis            0.70               0.48    None      0.69
"Perceptron                            0.66               0.47    None      0.66
"GaussianNB                            0.67               0.46    None      0.69
"DummyClassifier                       0.59               0.33    None      0.44
```

Explanation:

The report tells machine learning pipeline for classifying data from three distinct datasets: breast cancer, employee, and heart. The breast cancer dataset is loaded and pre-processed, with the 'diagnosis' column mapped to binary values for benign and malignant diagnoses. The Lazy Classifier from the lazy predict library is then applied to quickly assess the performance of multiple machine learning models on this dataset. The simplicity and automation of Lazy Classifier make it a convenient tool for gaining initial insights into the baseline performances of various algorithms without the need for extensive manual tuning. This code structure is not limited to the breast cancer dataset; it serves as a versatile template that can be easily adapted to other datasets.

For a more comprehensive analysis, the code can be extended to include the employee and heart datasets. By loading and preprocessing each dataset similarly, one can conduct lazy classification to compare and contrast the performance of machine learning models across the diverse datasets. The Lazy Classifier's ability to automatically evaluate and select suitable algorithms streamlines the analysis, providing a quick overview of model performances across different domains. This approach facilitates a holistic understanding of how various classifiers perform on datasets with varying characteristics, allowing for informed model selection and parameter tuning in subsequent stages of the machine learning pipeline.

Conclusion:

In conclusion, this report outlines the outcomes of a lazy classification approach applied to three distinct datasets: Breast Cancer, Employee, and Heart. The analysis involved the tuning of decision function parameters through a cross-validation procedure, with the code and detailed results accessible in the provided GitHub repository. The Breast Cancer dataset comprised 500 instances of Malignant and Benign samples, with specific features such as Mean Compactness, Mean Concavity, and others contributing to the classification process.

For the Employee dataset, which contained information on eight employees, diverse features like Education, Joining Year, City, Payment Tier, Age, Gender, Ever Benched, and Experience in the Current Domain were considered. The target column "Leave or Not" indicated whether an employee took leave or not. Lastly, the Heart dataset included features like Age, Sex, and various medical parameters. The LazyClassifier from the lazypredict library was employed to quickly evaluate the performance of multiple machine learning models on each dataset.

The report underscores the adaptability of the code structure, which serves as a versatile template applicable to various datasets. LazyClassifier's automation and simplicity make it a valuable tool for obtaining initial insights into the baseline performances of different algorithms, enabling a quick overview of model performances across diverse datasets. This approach fosters a comprehensive understanding of classifier behaviours in different domains, facilitating informed model selection and parameter tuning in subsequent stages of the machine learning pipeline.