

Classification Algorithms : Economies Grew

Date: 10, August 2023

Name: Md Mohidul Haque

1. Structure

- The python script load and analysis the given data points, and then run three different binary classification algorithms to explain whether regions economies grew by more than 5%. Goodness of the models evaluated at the end.
- Treatments: Chosen three classification algorithms are *DecisionTree*, *RandomForest*, and *SupportVectorMachine*. To evaluate the goodness of trained model, we need to split the given data in train and test set. In not rich data situation 80% data used for train the model. After analysis, we can see that the given data is imbalanced: for this reason *Stratified Sampling* is used in *train_test_split* to split the data into train and test set.
- Response: The three mentioned binary classification algorithms will be trained on train set, and the test set will be used to evaluate those trained models.

2. How to run the code

The script can be executed directly (the input data sets has to be in the working directory). It will print the information about imbalanced data, and the values for models evaluation into the interpreter.

3. Code (Problem_01)

Please see the .py file.

4. Results and Interpretation (Problem_02)

Class 1: regions economies grew by more than 5%.

Class 0: regions economies didn't grow by more than 5%.

In the given data, 71% data points are for class 1, and 29% data points are for class 0. Therefore the given data is imbalanced.

To train and evaluate the algorithms *DecisionTree*, and *RandomForest* requires less than one seconds. On the other hand, *SupportVectorMachine* needs more than 3 minutes to train and evaluate the performance on unseen data points.

Evaluation scores:

Accuracy: tells us about the total amount of correct prediction made by our model.

Precision: low precision means the model has of low rate in predicting the class 1.

Recall: low recall means the model is assigning the wrong prediction to the class 1.

F1 score: harmonic mean of precision and recall. Low F1-score indicates poor balance between precision and recall in a binary classification model's performance.

The chosen three models performed well on the unseen data points: Accuracy, Precision, Recall, and F1 score are mostly more than 70%.