Assignment 4

Topic: Classification

Deadline: 28th October, 2022

In this assignment, you are required to create a multiclass classifier to classify Penguin species, based on some given attributes which is a mix of both numerical and categorical attributes. Use the data provided in `train.csv` to train your classifier for the column 'Species'.

Tasks (100 points):

1. Build a multi-class classifier (all-vs-all) and train it on the given data.
   (correctness: 30 points, accuracy score on `test.csv`: 20 points)
   - Consider factors such as data cleaning, data skew, handling numerical vs categorical variables.
   - Running `` `teamName_classifier.py test.csv` `` should output a csv file with the predicted labels, which will then be checked with the actual labels to determine your model's accuracy score..
2. Experiment with various classifiers such as KNN, Decision Trees, Random Forests etc. Which classifier performs the best ? (10)
3. Explain the difference between using a one-vs-all and all-vs-all (also called many-vs-many) classifier on this dataset. (5)
4. The amount of training data is quite low. How do you deal with that? (5)
5. Consider both feature selection and feature engineering to improve your results. Write your findings in the report. (10)
6. Plot these three error metrics specific to multi-class classification: confusion matrix, F1 score, and ROC AUC score. Which is the best error metric for this dataset and why? (20)

<u>Submission instructions</u>:
Directory structure:
teamName_A3.zip
|
|--`teamName_classifier.py`
|--`teamName_extras.ipynb`
|--`teamName_Report.pdf`

`teamName_classifier.py` should contain the classifier only. You can conduct all of your analysis and testing in `teamName_extras.ipynb`. The final report must contain all of your plots and analysis.