[Project Code: RVNB]
Classification of Rice Varieties using Gaussian Naïve Bayes Learning Model

Project Duration: 22-Jan-2023 ~~ 11-Feb-2023
Submission Information: (via) CSE-Moodle

Objective:

The Rice dataset, collected by Commeo and Osmancik, contains information on various characteristics of rice samples, including information on grain length and width, chalkiness, and milled rice yield. The problem that researchers aim to solve using this dataset is to use the Gaussian Naive Bayes algorithm to classify the rice samples based on their quality. Specifically, the research objective is to train a Gaussian Naive Bayes model on the dataset and use this model to predict the quality of new rice samples based on their characteristics.

Attribute Information:

- Area: Returns the number of pixels within the boundaries of the rice grain.
- Perimeter: Calculates the circumference by calculating the distance between pixels around the boundaries of the rice grain.
- Major Axis Length: The longest line that can be drawn on the rice grain, i.e. the main axis distance, gives.
- Minor Axis Length: The shortest line that can be drawn on the rice grain, i.e. the small axis distance, gives.
- Eccentricity: It measures how round the ellipse, which has the same moments as the rice grain, is.
- Convex Area: Returns the pixel count of the smallest convex shell of the region formed by the rice grain.
- Extent: Returns the ratio of the region formed by the rice grain to the bounding box pixels.
- Class: Cammeo and Osmancik rices

The model will predict the class of the rice based on the seven specified attributes.

Note: *The program can be written in C / C++ / Java / Python programming language from scratch. No machine learning /data science /statistics package / library should be used for model creation.*

Data-Sets:
Rice_Cammeo_Osmancik.csv

Your Tasks:

1. Data Preprocessing: Split the dataset into parts: Test (30%) and Training (70%).
   The dataset is not divided into train and validation sets. The first task is to randomly partition the complete dataset into 5 parts: assign the first part as validation set and the rest for training the classifier. Repeat the process 5 times, assigning the validation sets in a round robin manner. *(5 fold cross-validation)*
2. Build a Gaussian Naive Bayes classifier from scratch for the dataset.
3. Classification Report:
   a. Create a classification report for both the trees in tabular form. (with and without pruning).
   b. You need to calculate precision, recall , f1-score and support of the model you made for each and every class (bad , good , ok , vgood).
   c. Also get the accuracy of the model.

```
Unpruned DT Classification report:
              precision    recall  f1-score   support

         bad       -          -        -         -
        good       -          -        -         -
          ok       -          -        -         -
       vgood       -          -        -         -

    accuracy                           0.92       467
    macroavg       0.94       0.85      0.89       467


Pruned DT Classification report:
              precision    recall  f1-score   support

         bad       -          -        -         -
        good       -          -        -         -
          ok       -          -        -         -
       vgood       -          -        -         -

    accuracy                           0.92       467
    macroavg       0.90       0.86      0.88       467
```

Fill the above blanks. The above results have also been provided for your reference and verification purpose. (Your answers may vary.)

Submission Details: (to be submitted under the specified entry in CSE-Moodle)

1. ZIPPED Code Distribution in CSE-Moodle
2. A brief (2-3 page) report/manual of your work
   (with your hyperparameter tuning results also presented in that report)

Submission Guidelines:

1. You may use one of the following languages: C/C++/Java/Python.
2. Your Programs should run on a Linux Environment.
3. You are **not** allowed to use any library apart from these (Also explore all these libraries if doing in Python, or equivalent of these):
   import numpy as np # linear algebra
   import pandas as pd # data processing, CSV file I/O (e.g. pd.read_csv)
   from sklearn.model_selection import train_test_split
   from sklearn.metrics import accuracy_score
   from sklearn.metrics import classification_report
   from sklearn.model_selection import KFold
   from sklearn import naive_bayes # sklearn Naive Bayes
   from sklearn.naive_bayes import GaussianNB # sklearn Gaussian Naive Bayes
   import operator
   from math import log
   from collections import Counter
   from statistics import mean

4. Your program should be standalone and should **not** use any *special purpose* library for Machine Learning. Numpy and Pandas may be used. And, you can use libraries for other purposes, such as generation and formatting of data.
5. You should submit the program file and README file and not the output/input file.

6. You should name your file as <GroupNo_ProjectCode.extension>.
   (e.g., *Group99_RVNB.zip* for code-distribution and *Group99_RVNB.pdf* for report)
7. The submitted program file *should* have the following header comments:
   # Group Number
   # Roll Numbers : Names of members (listed line wise)
   # Project Number
   # Project Title
8. Submit through CSE-MOODLE only.
   Link to our Course page: https://moodlecse.iitkgp.ac.in/moodle/course/view.php?id=508

You should not use any code available on the Web. Submissions found to be plagiarized or having used ML libraries (except for parts where specifically allowed) will be awarded zero marks.

For any questions about the assignment, contact the following TA:
Rijoy Mukherjee (Email: rijoy.mukherjee@gmail.com)