Glass Classification

Name: Jerrin C. Redmon
Link:https://www.kaggle.com/datasets/danushkumarv/glass-identification-data-set

ABSTRACT

The glass identification dataset is a data set of the properties of different kinds of glasses from 214 different instances. The goal of this dataset is to use the different distribution of elemental properties of various types of glass to be able to identify where a glass shard originated from.

I. **BACKGROUND**

-Description provided by the source: "Vina conducted a comparison test of her rule-based system, BEAGLE, the nearest-neighbor algorithm, and discriminant analysis. BEAGLE is a product available through VRS Consulting, Inc.; 4676 Admiralty Way, Suite 206; Marina Del Ray, CA 90292 (213) 827-7890 and FAX: -3189. The study of the classification of types of glass was motivated by criminological investigation. At the scene of th, the glass left can be used as evidence...if it is correctly identified!"

II. **EXPLORATORY ANALYSIS**

- This data set contains 214 samples with 11 columns of various data types

Table 1: Data Types

Variable	Data Type
Id	Integer
RI (refractive index)	Float
Na (Sodium)	Float
Mg (Magnesium)	Float
Al (Aluminum)	Float
Si (Silicon)	Float
K (Potassium)	Float
Ca (Calcium)	Float
Ba (Barium)	Float
Fe (Iron)	Float
Type of glass	Integer

Table X: Experiment Parameters

Experiment Number	Parameters
1	Logistic Regression with 80/20 split and stratify it
2	Random Forest Classification
3	Support Vector Classification
4	Decision Tree Classifier
5	Performed parameter tuning using Grid search CV

7	Artificial neural network with 3 dense layers with ReLU activation function and softmax classification

Note:The the dataset contained no categorical variables, but on an important note, the first 4 models had the dataset reshaped due to an imbalance. For the ANN the data was scaled before use, which results in higher accuracy.

A.   *Tools Used*

-The following tools were used for this analysis: Python running in Google colab on a linux machine was used for all analysis and implementation. In addition to base Python, the following libraries were also used: Pandas 0.18.1, Numpy 1.11.3, Matplotlib 1.5.3, Seaborn 0.7.1, SKLearn Libraries, Keras, Tensorflow

III.   RESULTS

```
              precision    recall  f1-score   support

           1       0.33      1.00      0.49        14
           2       0.00      0.00      0.00        15
           3       0.00      0.00      0.00         3
           5       0.00      0.00      0.00         3
           6       0.00      0.00      0.00         2
           7       0.00      0.00      0.00         6

    accuracy                           0.33        43
   macro avg       0.05      0.17      0.08        43
weighted avg       0.11      0.33      0.16        43
```

	model	best_score	best_params
0	SVC	0.840563	{'C': 100, 'kernel': 'rbf'}
1	Logistic_Regession	0.813242	{'penalty': 'none', 'solver': 'newton-cg'}
2	Random_Forest	0.867998	{'criterion': 'entropy', 'max_depth': 5}
3	Decision_Tree	0.843341	{'criterion': 'gini', 'max_depth': 4}

IV.   CONCLUSION

From what I gathered the Glass classification data set performed quite well with the main 4 models with random forest performing the bast. However despite several attempts and trails with ANN, the best accuracy was only 42%.