Solution for Quiz-2

CSCE 590-1: From Data to Decisions with Open Data: A Practical Introduction to AI

Prof. Biplav Srivastava, Spring 2021

Student: Avineet Kumar Singh

GitHub

Question 1: Classification

German credit dataset is a popular dataset in ML. It can be found at in multiple formats at (.csv, .arff):

https://www.openml.org/t/31

https://archive.ics.uci.edu/ml/datasets/statlog+(german+credit+data)

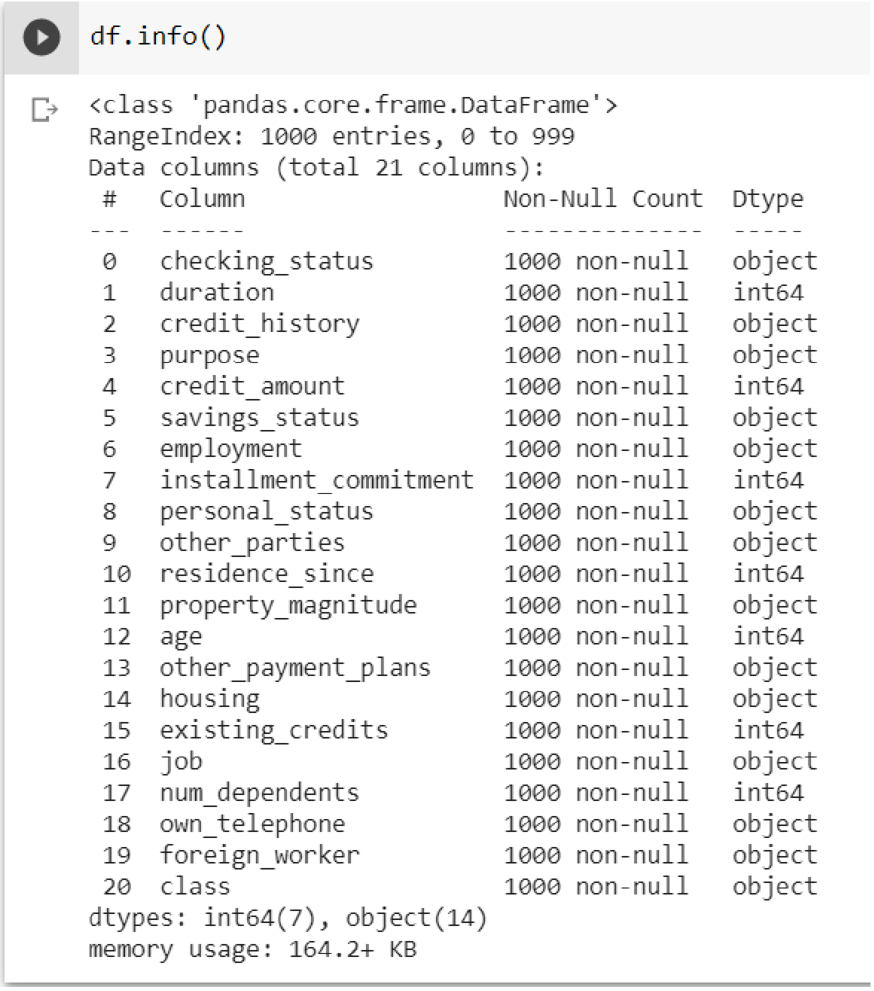
1(a): Download the data and pre-process in any way necessary. How many data items and features does it have? What are their types? [10 points]

Solution:

Number of data items: 1000,

Features: 21

Type of data:



```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1000 entries, 0 to 999
Data columns (total 21 columns):
 #   Column                  Non-Null Count  Dtype 
---  ------                  --------------  ----- 
 0   checking_status         1000 non-null   object
 1   duration                1000 non-null   int64 
 2   credit_history          1000 non-null   object
 3   purpose                 1000 non-null   object
 4   credit_amount           1000 non-null   int64 
 5   savings_status          1000 non-null   object
 6   employment              1000 non-null   object
 7   installment_commitment  1000 non-null   int64 
 8   personal_status         1000 non-null   object
 9   other_parties           1000 non-null   object
 10  residence_since         1000 non-null   int64 
 11  property_magnitude      1000 non-null   object
 12  age                     1000 non-null   int64 
 13  other_payment_plans     1000 non-null   object
 14  housing                 1000 non-null   object
 15  existing_credits        1000 non-null   int64 
 16  job                     1000 non-null   object
 17  num_dependents          1000 non-null   int64 
 18  own_telephone           1000 non-null   object
 19  foreign_worker          1000 non-null   object
 20  class                   1000 non-null   object
dtypes: int64(7), object(14)
memory usage: 164.2+ KB
```

Numeric features: 7

Nominal features: 14

Data Preprocessing is performed using two methods:

1) Using Sklearn (giving numbers to categories/label encoding the data)

2) Using Pandas(one hot encoding)

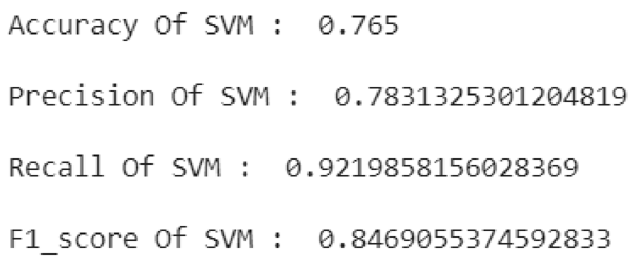
1(b) Perform classification on the class label with at least two methods.

Present model accuracy, recall and F1 statistics. If possible, print model structure.

Solution:

Code: https://github.com/AVINEET-Singh/csce-590-submissions/blob/main/D2D_Quiz2_Q1.ipynb

Classifier 1: **SVM**



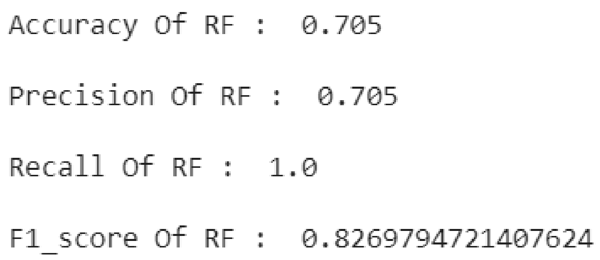
```
Accuracy Of SVM :  0.765

Precision Of SVM :  0.7831325301204819

Recall Of SVM :  0.9219858156028369

F1_score Of SVM :  0.8469055374592833
```

Classifier 2: **Random Forest**



```
Accuracy Of RF :  0.705

Precision Of RF :  0.705

Recall Of RF :  1.0

F1_score Of RF :  0.8269794721407624
```

<u>Question 2:</u> Clustering

Cluster the data with any method without giving the number of classes. Now compare the clusters with the classes. Find the homogeneity, completeness, and v-score metrics.

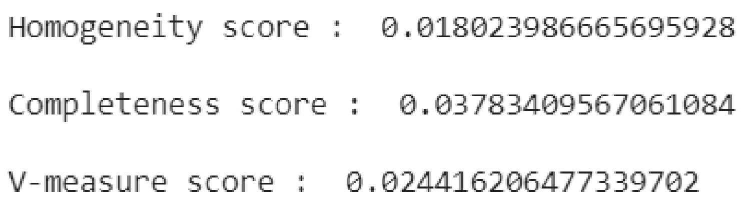
<u>Solution:</u>

<u>Code:</u> https://github.com/AVINEET-Singh/csce-590-submissions/blob/main/D2D_Quiz2_Q2.ipynb

<u>Clustering Method:</u> DBSCAN

Identified clusters: [-1 0 1 2 3 4 5]

Clustering Performance Evaluations:



```
Homogeneity score :  0.018023986665695928

Completeness score :  0.03783409567061084

V-measure score :  0.024416206477339702
```

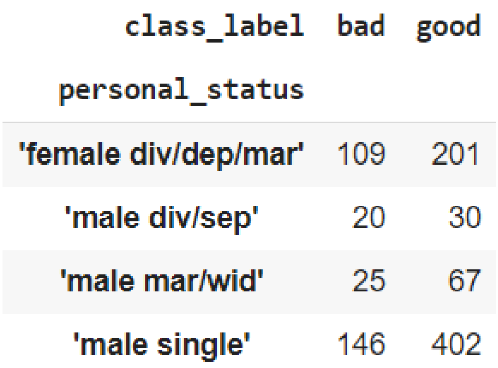
<u>Question 3:</u> Bonus:

The dataset has attributes for age and personal_status. What is the distribution of class with respect to these attributes? Is there a age or personal_status group that can perceive bias? Feel free to pre-process data to gain insights – e.g., binning for age.

<u>Solution:</u>

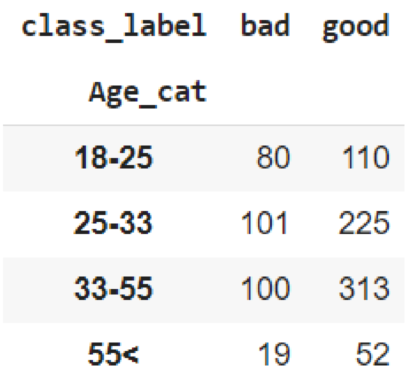
<u>Code</u>: https://github.com/AVINEET-Singh/csce-590-submissions/blob/main/D2D_Quiz2_Q3.ipynb

distribution of 'personal_status' based on 'class' label:



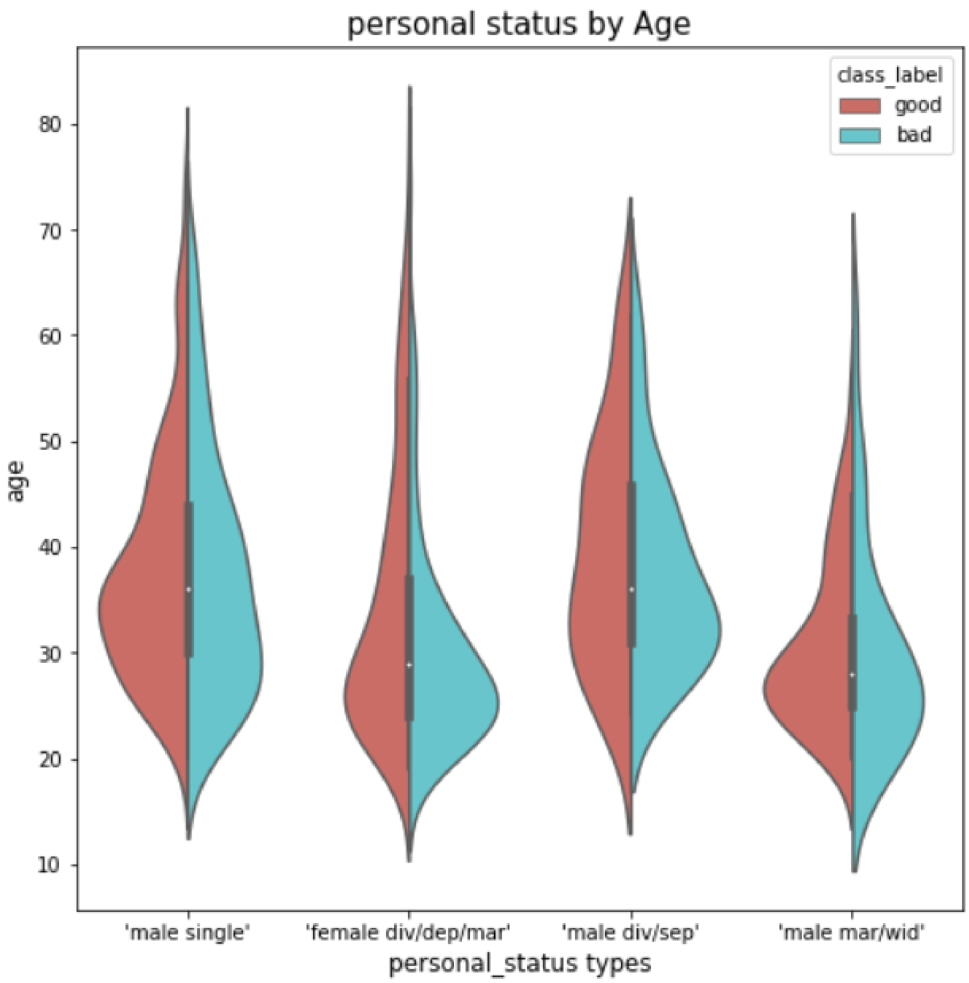
class_label	bad	good
personal_status		
'female div/dep/mar'	109	201
'male div/sep'	20	30
'male mar/wid'	25	67
'male single'	146	402

distribution of 'age' based on 'class' label(After Categorizing age):



class_label	bad	good
Age_cat		
18-25	80	110
25-33	101	225
33-55	100	313
55<	19	52

Plotting 'age' and 'personal_status' groups based on 'class' label



<u>Study of Biasness in dataset (as per age and personal_status group)</u>

<u>Based on the above tables and figures following biasness could be identified:</u>

1) Categories for female customers are low as male customers are divided into more categories as per personal status. Even if the dataset is small, the categories for personal status should be balanced among both the genders. This could be considered as '**Association bias'**.

2) As per personal status, the ratio of good credit risks to bad credit risks customers is higher for 'male-single' customers. It means there is a slight biasness in considering marital status for determining good and bad credit risks.

3) As per age category, the ratio of good credit risks to bad credit risks customers is higher for age category of '33-55'. People in this category are considered more stable which may not be true.

4) As per the plot, for 'female div/dep/mar' the number of bad credit risks increase as compared to good credit risks after the age of 35(approx.), which is not prevalent in other categories at that age.