CSCE 590-1: From Data to Decisions with Open Data: A Practical Introduction to AI
Prof. Biplav Srivastava, Spring 2021

Quiz 2 / Feb 18, 2021/ Instructions

- Return answer to quiz as .pdf and GitHub link by 5:00 pm on Monday, Feb 22, 2021 by posting to your shared folder (e.g., Google folder mentioned in spreadsheet) and email to biplav.s@sc.edu.
- Ask question by email. Or, office hour of Monday, Feb 22, 2021 can be used to clarify questions. Timing: 11:30am-12:30pm.
- Total points = 100 + 20 (bonus) = 120, Obtained =

Student Name: Vedant Khandelwal

GitHub link with model: https://github.com/khvedant02/Vedant-CSCE590-submission/

Question 1: Classification
[10 + 60 = 70 points]

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]

It has a total of 1000 data items
It has a total of 20 features and one prediction class columns
Columns with categorical data are: checking_status, credit_history, purpose, saving_status, employment, personal_status, other_parties, property_magnitude, other_payment_plans, housing, job, own_telephone, foreign_worker, class, installment_commitment, residence_since, existing_credits, num_dependents.
The continuous columns are: duration, credit_amount , age

Detailed explanation in the python notebook in the Github repository

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.
[2 x 30 = 60 points]

Model - Accuracy - Precision - Recall - F1-Score - AUC
SupportVectorClassifier - 81.00 - 0.759333 - 0.773333 - 0.760409 - 0.675265
RandomForestClassifier - 72.00 - 0.518400 - 0.720000 - 0.602791 - 0.500000
NaiveBayes - 70.00 - 0.693702 - 0.436667 - 0.422830 - 0.697696
LogisticRegression - 81.67 - 0.759053 - 0.766667 - 0.761922 - 0.772707

more details are present in the python notebook in the Github Repo

Question 2: Clustering
[30 points]
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.

The details are present in the python notebook in Github Repo

Question 3: Bonus:
[20 points]

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.

The details are present in the python notebook in Github Repo