PG Certificate in Software Engineering for Data Science
Week 1 Assignment 2.1

In this assignment you will get familiar with data loading, preprocessing, training and testing phases involved in machine learning workflow. You can make use of the boiler plate code to begin with.

Programming Language: Python

Introduction

In this assignment, we will aim to solve the German Credit Risk Analysis problem. When a bank receives a loan application, based on the applicant's profile the bank has to make a decision regarding whether to go ahead with the loan approval or not.

To minimize loss from the bank's perspective, the bank needs a decision rule regarding who to give approval of the loan and who not to.

Dataset

The German Credit Data contains data on 20 variables and the classification whether an applicant is considered a Good or a Bad credit risk for 1000 loan applicants.

You can obtain the dataset from
https://archive.ics.uci.edu/ml/machine-learning-databases/statlog/german/german.data
and refer https://archive.ics.uci.edu/ml/machine-learning-databases/statlog/german/german.doc for description about the dataset.

Tasks

To get started, there is a credit_scoring_starter.ipynb notebook given with the assignment, which can be used to do all the tasks.

1. Load the data and display the first few rows of the data. The displayed data should be in human readable form.
2. Explore the dataset and plot the frequency of each label as a histogram. Also plot the histogram of values from any other attribute of choice.
3. Implement KNN classifier(using sklearn) and find the right number of neighbours that gives best performance for the credit scoring dataset.
4. Try implementing two other classifiers of your choice and report the performance.
5. Plot confusion matrix and ROC to analyze the results.

Bonus Tasks

6. Use k-fold validation while training inplace of normal random split in the starter code.

7. Implement KNN model (not using sklearn).
8. Determine the performance metric that best suits this problem statement.

Environment

We will be expecting that you run the notebook in python3.8 environment with the following packages:

- numpy
- scikit-learn
- pandas
- matplotlib
- seaborn

Submission

Make changes in the credit_scoring_starter.ipynb notebook and submit the same on the portal.