Assignment 1

Due date: February 23, 2022 (12.00 PM)

The "assignment 1" objective is to assess how well you have grasped the machine learning concepts taught during the first three lectures. So far, we have identified the critical elements in the machine learning workflow and discussed the approach that can be taken to construct and train a machine learning model and how to use it during inference(i.e., using the trained model to make predictions on the test data). Similarly, we will follow the machine learning workflow to predict individuals identified as having "Hepatitis C".

Download the dataset(i.e., hcv_data_split.csv) and the ipython notebook(i.e., assignment_1_students.ipynb) and complete the missing sections according to the given instructions.

The dataset consists of the blood results from Hepatitis C patients and blood donors. Of 615 individuals, 75 patients were infected with hepatitis C. The rest of the individuals can be considered the control group, where the class distribution is approximately 1:7 favouring the control group. The column names: ALB, ALP, ALT, AST, BIL, CHE, CHOL, CREA, GGT, PROT represents laboratory values. The individuals' age and gender are also provided with the associated annotations indicating whether the person is infected with hepatitis C (i.e., 1) or not (i.e., 0). The dataset can be further divided into "train" and "test" based on the "split" column, where the positive and negative classes are proportionately distributed between the two datasets.

For more details on the dataset, please refer to the paper:

Hoffmann, G.F., Bietenbeck, A., Lichtinghagen, R., & Klawonn, F. (2018). Using machine learning techniques to generate laboratory diagnostic pathways—a case study. *Journal of Laboratory and Precision Medicine*.

To access the dataset:

https://archive.ics.uci.edu/ml/datasets/HCV+data

The objective of the assignment is to identify the best model out of the five models that produce the best "balanced accuracy".

You must also synthesize examples using the SMOTE(Synthetic Minority Oversampling TEchnique) technique so that the minority class can be oversampled, resulting in a more generalized model.

At the end of each model training session(i.e., without and with the synthesized data), you are required to report the best results(i.e., on both validation and test) produced by each algorithm with the associated hyperparameters.

Results submission:

- You can try as many combinations of hyperparameters and make sure to report them accordingly.

▼ Ensure that the corrector has only to provide the dataset path, and the rest of the code will get executed according to the provided instructions in the notebook.

- Please make sure that your program runs without runtime errors, and if not, you will receive zero marks. If you continuously face issues running the program, please let me know(i.e., before the deadline) so that I can look into your code. Given the task, it is important for you to learn to build a machine learning model from end to end.

💡 Submit the completed notebook with the group name, group participant names, email and student number at the top of the notebook in a markdown cell. Also, make sure to rename the notebook file to: "group_name.ipynb"

If you require further clarifications or are facing issues related to "assignment 1" please let me know.