BIRLA INSTITUTE OF TECHNOLOGY & SCIENCE, PILANI
K. K. BIRLA GOA CAMPUS

Second Semester 2022-23
Data Mining (WILP)
Assignment-1

Given the following dataset, build a model(s) that can label test tuples into one of the 29 predefined classes. You may use any of the classification algorithm(s).

Competition Link: https://www.kaggle.com/t/44dbab6c5cbb47c9a202be762a632ea7
Start Time: 25 April 2023, 6 AM
End Time: 05 May 2023, 11.30 PM

A. General Instructions:

1. Create a Kaggle account (if you don't have one) and join the competition.
2. **Change the Team name to your id: 20XXXXXXXXX**. (ANY OTHER USERNAME WILL BE REMOVED FROM THE COMPETITION AND WILL BE CONSIDERED MALPRACTICE (Component Maximum Marks * -1)).
3. Download the train file ("X_train.csv") ("Y_train.csv") and test file ("X_test.csv") . You are given a train dataset (labels ranging between 1 and 29) and a test dataset with 8 attributes.
4. You need to build a classification model(s) that can assign given test tuples into one of the 29 classes.
5. You need to upload a CSV file in the format of the sample submission file (given on Kaggle) to submit your solution.
6. Your submission will be evaluated on 50% of the dataset to give your result on the public leaderboard. At the end of the competition, your **selected submission** (You can select maximum **one** submission) will be evaluated on the remaining 50% of the data and will be shown on the private leaderboard. However, the final score will be based on your performance on 100% data. The evaluation metric is **categorical accuracy**.
7. You are given 10 days with 5 submissions on each day. You are requested to start working on your model immediately to avoid pressure at the last moment. No requests will be entertained to increase the submission deadline or to increase the number of submissions allowed per day.

B. Final Assignment Submission Instructions:

Same "final code" should be submitted in both KAGGLE and TAXILA. In case of a mismatch, zero marks will be awarded for this assignment.

1. **Assignment Submission Format: (KAGGLE)**
   a. You kaggle submission file name SHOULD always be your **FULL ID NUMBER** "20XXXXXXXXX"
   b. **Do at least one submission before ~~27/04/2023~~ 29/04/2023**
      i. if you face any issue on kaggle, you can raise a request to hemantr@goa.bits-pilani.ac.in and f20190115@goa.bits-pilani.ac.in

ii. After ~~27/04/2023~~ 29/04/2023, we will not respond to your admission issues on kaggle.

c. Select **one** public submission to be considered for your final evaluation before the deadline. (By default the latest submission is considered)

d. The result obtained on running the notebook submitted on TAXILA must match your score on the private leaderboard. In case of a mismatch, zero marks will be awarded for this assignment.

2. Assignment Submission Format: (TAXILA)

A zip file (**20XXXXXXXXX**.zip) consisting of the following: (each file should be named as ***"ID_*****")***

a. (ID_psc) Portable source code:
   i. Must contain all required packages/libraries.
   ii. Path for any required file(s) should not be local to your machine
   iii. Instructor should be able to run your code after direct download.

b. (ID_psc) Portable source code (pdf format) different methods you tried with outputs.

c. (ID_fsc) Final Jupiter notebook (pdf format) that gave you your best result with outputs.

d. (ID_report) Report in PDF format (max 2 pages. 11pt. Times New Roman.)
   i. Insights, inferences, results, and conclusions are drawn from the assignment.
   ii. Proper references to the source code and figures.

e. (ID_fig1…) Figures (depends on the type of the assignment)
   i. Self-explanatory caption to the figures. ~~1.jpg, q1.jpg, abc.jpg~~

f. (ID_readme) README.txt
   i. Step by step instructions to run your code.
   ~~ii. Download package 1, download xyz.jar, install MySQL~~

C. Assignment Submission Policy:

- Submission accepted through **(Taxila + Kaggle) only.**
- No assignment will be accepted by **email or after the deadline**.

D. Malpractice and Plagiarism:

Plagiarism will be checked for every submission.

- The rule is very simple
- If **(Plagiarism % from Turnitin Report)** > 30
  - o Will be awarded **"Component Maximum Marks * -1"**

Any act by the student that violates the above mentioned instructions (such as having team name different from 20XXXXXXXXX or creating multiple accounts, etc) will be considered as malpractice in lab and will result in strict disciplinary action against the student in addition to being awarded **Component Maximum Marks * -1.**