Course Project Module-1 Submission Related Instructions

If the course project is a machine learning/deep learning-related one, then write a single Python/R program that performs all preprocessing tasks.

Non-image datasets and non-phishing detection-related

- Check if the entire column has the same value or not. If yes, print all such column numbers on the terminal and also store them in an output file, **RollNumber-Duplicate-Column.txt.**
- Check if any duplicate rows exist in the dataset or not. If yes, print all such row numbers on the terminal and also store them in an output file, **RollNumber-Duplicate-Row.txt.**
- If any column needs normalization, then apply the appropriate normalization technique.
- If any missing value is observed, apply the appropriate missing value substitution technique.
- Generate a co-relation map (heatmap) for the entire dataset and store the output with the filename **RollNumber-Heatmap.JPEG**
- Store the pre-processed dataset with a file name **RollNumber-Pre-processed_Dataset.File-Extension.**
- Upload the Python/R program, **original dataset (before pre-processing), pre-processed dataset, RollNumber-Heatmap.JPEG**, **RollNumber-Duplicate-Row.txt. and RollNumber-Duplicate-Column.txt** onto Moodle before the submission deadline.

Phishing detection-related course project

- Download the URL dataset (not the pre-processed dataset).
- Write a program to upload URLs from a dataset onto the Virustotal website and get the label (benign or phishing) from the Virustotal website. Accordingly, change the URL label.
- Create your own dataset by extracting features from the URL dataset and save the created dataset with the filename **RollNumber_URLfeaturedataset.csv.** Refer to the research paper uploaded to Moodle to see what features need to be extracted and how to extract those features from the URL dataset.
- Check if the entire column has the same value or not. If yes, print all such column numbers on the terminal and also store them in an output file, **RollNumber-Duplicate-Column.txt.**
- Check if any duplicate rows exist in the dataset or not. If yes, print all such row numbers on the terminal and also store them in an output file, **RollNumber-Duplicate-Row.txt.**
- If any column needs normalization, then apply the appropriate normalization technique.
- If any missing value is observed, apply the appropriate missing value substitution technique.
- Generate a co-relation map (heatmap) for the entire dataset and store the output with the filename **RollNumber-Heatmap.JPEG**

- Store the pre-processed dataset with a file name **RollNumber-Pre-processed_Dataset.File-Extension.**
- Upload the Python/R program, **original dataset (before pre-processing), pre-processed dataset, RollNumber-Heatmap.JPEG**, **RollNumber-Duplicate-Row.txt. and RollNumber-Duplicate-Column.txt** onto Moodle before the submission deadline.

Image datasets related to course project

After preprocessing the image dataset, store the results in the output file in the below-mentioned format with a file name of **RollNumber-PreInfFile.excel.**

Upload the Python/R program, original dataset (created dataset before pre-processing), RollNumber-PreInfFile.excel, onto a Moodle before the submission deadline.

Sl No	Image file Name	Size of the image in terms of pixel M×N	Type of image (RGB/ Gray Scale/ (Black-and-White)	Image type (TIFF/BMP/JPEG)