Assignment # 01

Data Science Tools & Techniques (MS DS)

Deadline: 13-Sep-2025 @ 11:59

Total Marks = 100

Instructions:

- Combine all your work in one folder. The folder must contain only deliverable files (no binaries, no exe files, etc.).
- Rename the folder as ROLL-NUM_SECTION (e.g., 22i-0001_A) and compress the folder as a zip file. (e.g. 23i-0001_A.zip). Do not submit a .rar file.
- All the submissions will be done on Google Classroom within the deadline. Submissions other than Google Classroom (e.g., email, etc.) will not be accepted.
- The student is solely responsible for checking the final zip files for issues like corrupt files, viruses in the file, and mistakenly executed sent. If we cannot download the file from Google Classroom for any reason, it will lead to zero marks in the assignment.
- Displayed output should be well-mannered and well-presented. Use appropriate comments and indentation in your source code.
- **Be prepared for a viva or anything else after the submission of the assignment.**
- If there is a syntax error in the code, zero marks will be awarded in that part of the assignment.
- Understanding the assignment is also part of the assignment.
- **Zero marks will be awarded to the students involved in plagiarism. (Copying from the internet is the easiest way to get caught.)**
- **Late Submission** will **not** be entertained, and **no retake** request will be accepted as per the course policy.

Tip: For timely completion of the assignment, start as early as possible.

Note: Follow the given instruction; failing to do so will result in a zero.

Q#1: (100 marks)

Perform the following tasks for this question.

Dataset:

For the tasks below, use the two datasets: **dataset # 1: Census-Income dataset** (https://archive.ics.uci.edu/dataset/117/census+income+kdd) and **dataset # 2: The Attached Dataset** (available in the attached .csv file).

Task 1: Data Pre-processing

- Identify the issues in these two datasets.
- Perform relevant encoding methods for the numerical data columns.
- Preprocess these two datasets and overcome the issues identified.

Task 2: Exploratory Data Analysis (EDA) Before and After Preprocessing

- Perform an initial EDA (summary statistics, missing values, distribution plots) before any preprocessing for both datasets.
- Perform EDA again after preprocessing and compare the results.

Deliverables:

1. Submit the Jupyter Notebook with code and output for all tasks.
2. Include a brief report (around 500 words) summarizing the key insights from each task and the importance of data filtering and preprocessing in the data science workflow.