CSE 574: Introduction to Machine Learning
Assignment 1 (to be completed individually)
Python tasks using dataset to analyze income inequality
Points: 100; Date of submission: 4 October 2024

Description: For this project, you will be introduced to the adult dataset. This dataset consists of census data popularly known as census income dataset. Your task will be to write a python code to identify the individuals who individually earn more than USD 50,000/year. Your will need to preprocess the dataset to obtain this result, and, also apply a classifier, preferably a random classifier. This part of the project will be challenging as we did not cover those topics, however, by continuous self-learning you should be able to understand the concepts of simple pre-processing and classifier. For accuracy, you should aim for atleast 85% and above accuracy. You should conduct the following tasks.

1. Load the dataset (Points: 5)
2. Output the structure of the dataset (Points: 10)
3. Clean the dataset, handle the missing values and encode the categorical values (Points: 15)
4. Explore the data to understand better, for example, draw a bar plot to identify the distribution of the population in the dataset by age, followed by distribution of income by gender. (Points: 20)
5. Apply predictive modeling to the data to predict whether an individual earns more than $5π0K a year. (Points: 30)
6. In conclusion, prepare a project report with the code to show the diagrams and output generated and submit a poster using the template provided. Use creativity to represent the results in the poster format. In your poster, you need to cite the dataset as "Becker, B. & Kohavi, R. (1996). Adult [Dataset]. UCI Machine Learning Repository. https://doi.org/10.24432/C5XW20."(Points: 20)

Submission: Place all your files in a folder and name it LASTNAME_FIRSTNAME_UBPERSONNUMBER_CSE574_PROJ1.zip

Note: Any kind of plagiarism, like copying code from the internet, copying report data and figures will result in 0 points for this project.