Lab - 2
CSL2010: Introduction To Machine Learning
AY 2022-23

General Instructions

1. You need to upload a zip ***<Lab2_Your_Roll_No>.zip***, which contains one file for the task in ***<Lab2_Your_Roll_No>.py*** format and the report for the entire assignment in ***<Lab2_Your_Roll_No>.pdf*** format.
2. Provide your colab file link in the report. **Make sure that your file is accessible**.
3. Submit a single report, mentioning your observations for all the tasks.
4. Report/Cite any resources you have used while attempting the assignment.
5. Attempt Q1 and one question from Q2-Q5 during the lab.

A csv file has been provided to you at this link. The given dataset contains 12 columns and description for each column is available here. In the given dataset, “salary_in_usd” is the target variable. Answer the following based on this.

Q1) Assign a type to each of the following features manually as well as using a code if possible: (a) Job Title, (b) salary, (c) company size [3 Marks]

Q2) Write a function to handle missing values in the dataset (e.g., any NA, NaN values), and demonstrate/discuss its functioning in the report. [4 Marks]

Q3) Write a function to reduce noise (any error) in individual attributes and demonstrate/discuss its functioning in the report. [2 marks]

Q4) Write a function to encode all the categorical features in the dataset according to the type of variable jointly, and demonstrate/discuss its functioning in the report. [5 Marks]

Q5) Write a function to normalize / scale the features either individually or jointly, and demonstrate/discuss its functioning in the report. [2 Marks]

Q6) Write a function to create a random split of the data into three subsets (train, validation and test sets) in the ratio of 70:20:10 respectively.
(a) Using numpy operation [2 Marks]
(b) Using library function [2 Marks]