ML Group Project – Salary Prediction

1. Introduction

This project aims to have students experience in a practical problem for predicting whether the salary of employees is less than 50K or greater than 50K. You will learn how to analyze and formulate problem and how to apply Machine Learning techniques in terms of preprocessing data, choosing learning algorithms, training and evaluation the models.

2. Dataset

The CSV file of the dataset will be given. The description of the dataset is provided as follow.

- **age**: Continuous.
- **workclass**: Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked.
- **fnlwgt**: Continuous. In other words, it is a continuous variable that indicates the number of people in the overall population that each entry in the dataset represents.
- **education**: Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool.
- **education-num**: continuous.
- **marital-status**: Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse.
- **occupation**: Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspct, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Forces.
- **relationship**: Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried.
- **race**: White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black.
- **sex**: Female, Male.
- **capital-gain**: continuous.
- **capital-loss**: continuous.
- **hours-per-week**: continuous.
- **native-country**: United-States, Cambodia, England, Puerto-Rico, Canada, Germany, Outlying-US(Guam-USVI-etc), India, Japan, Greece, South, China, Cuba, Iran, Honduras, Philippines, Italy, Poland, Jamaica, Vietnam, Mexico, Portugal, Ireland, France, Dominican-Republic, Laos, Ecuador, Taiwan, Haiti, Columbia, Hungary, Guatemala, Nicaragua, Scotland, Thailand, Yugoslavia, El-Salvador, Trinadad&Tobago, Peru, Hong, Holand-Netherlands.

3. Tasks

You are required to apply at least two learning algorithms and make the comparison between them.

The following describes what you have to submit:

- **Technical Report** which describes your whole work. The following are the main points to be included in your report:
    - Describe the project and its problem
    - Describe your team and tasks of each member
    - Draw and describe the pipeline of this machine learning project
    - Explain how you preprocess the data? why?
    - Describe input features and output
    - Describe what learning algorithms you use? why?
    - Outline how you train the learning algorithms
    - Describe how you evaluate the performance of the models
    - Make conclusion of your work

- **Source Code** in which the Jupyter Notebook and the dataset file must be included. You are suggested to use **ONLY ONE** Jupyter Notebook for implementing each learning algorithm.

- **Group Presentation Video**: each member has to make a group presentation video explaining your works and explain how to implement the project.

Note: Team leader is responsible for submitting the technical report, presentation video, and the source codes by zipping them with the format of “**[GroupNumber]-[ProjectName].zip**” and submit on Elearning website.

Deadline: The deadline for the submission is on ***20th July, 2024***.