MINI PROJECT 3: MACHINE LEARNING FOR ANALYSIS AND PREDICTION

Objective

The objective of this mini project is to provide practice in data analysis and prediction by regression, classification and clustering algorithms.

Problem Statement

Attrition is the rate at which employees leave their job. When attrition reaches high levels, it becomes a concern for the company. Therefore, it is important to find out why employees leave, which factors contribute to such significant decision.

These and other related questions can be answered by exploration analysis and machine learning from the synthetic data provided by IBM to Kaggle (https://www.kaggle.com/datasets/pavansubhasht/ibm-hr-analytics-attrition-dataset/data).

Tasks

1. Data wrangling and exploration

- load, clean and explore the available data
- select and prepare the features of an employee that are most relevant for solving the following tasks

2. Supervised machine learning: linear regression

- train, test, and validate a machine learning model for prediction of the income of a new employee
- apply appropriate measures for assessing the quality of the model

3. Supervised machine learning: classification

- train, test, and validate a machine learning model for classification and prediction of employees' attrition
- apply appropriate methods for measuring of the accuracy of the model

4. Unsupervised machine learning: clustering

- apply a clustering algorithm for segmentation of the employees in groups of similarity
- evaluate the quality of the results by calculating a silhouette score and recommend a cluster configuration with the highest score

5. Implementation of the models in a Streamlit application

- store the models created above in files for future implementation
- create a web application with simple interface, where the users can select any of the three models and apply it for prediction of the behavior of their employees

Notes

1. In the readme file in your Github repository, answers of the following questions:
   - Which machine learning methods did you choose to apply in the application and why?
   - How accurate is your solution of prediction? Explain the meaning of the quality measures.
   - Which are the most decisive factors for quitting a job? Why do people quit their job?
   - What could be done for further improvement of the accuracy of the models?

- Which work positions and departments are in higher risk of losing employees?
- Are employees of different gender paid equally in all departments?
- Do the family status and the distance from work influence the work-life balance?
- Does education make people happy (satisfied from the work)?
- Which were the challenges in the project development?

2. Feel free to replace the attrition data set with your own data, related to the exam project. In that case you would use your data for solving the five tasks above, as well as for answering the re-formulated questions.

3. This is a group project. It provides 30 study points when completed.

Have success and fun,

the instructor