MINI PROJECT 4

MACHINE LEARNING FOR ANALYSIS AND PREDICTION OF ATTRITION

Objective

The objective of this mini project is to enable practice in data analysis and prediction by 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 and explore the data, clean it, and analyse it by statistics
- select the most relevant features of an employee for machine learning operations on prediction of the attrition

2. Supervised machine learning: classification

- train, test, and validate two machine learning models for classification and prediction of attrition (e.g. Decision Tree and Naïve Bayes)
- apply appropriate methods and measures for assessing the validity of the models and recommend the one with highest accuracy

3. Unsupervised machine learning: clustering

- apply at least one clustering algorithm (e.g. K-Means) for segmentation of the employees in groups of similarity
- evaluate the quality of the clustering by calculating a silhouette score and recommend the cluster configuration with higher score

4. Machine Learning application

- create and deploy on the localhost an interactive prototype of Streamlit application, visualizing stages and results of your work
- enable input of user data and making predictions on attrition by use of the classification model, created in p.2 above.
- test the application with various previously unknown input data and record the results.

Notes

Feel free to replace the original dataset with another one, in support of your exam project, if it is appropriate source of classification and clustering. In that case, you need to formulate alternative questions for answering.

Submit a link to the Github repository of your solution, where in the readme file provide answers of the following questions:

- Which machine learning methods did you choose to apply in the application?
- How accurate is your solution of prediction?
- Which are the most decisive factors for quitting a job?
- 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?

Have success and fun!

the instructor