RESOLVE studio

Prueba Técnica de Conocimiento

INGENIERO DE DATOS

The purpose of this test is to find patterns of employee absenteeism in a courier company. Absenteeism is defined as the inability of employees to go to work, which for some companies may represent loss of productivity and competitiveness. So, finding patterns explaining and predicting this behaviour can be useful from a human resources planning viewpoint.

1. Download the dataset from:

https://archive.ics.uci.edu/ml/datasets/Absenteeism+at+work

The dataset contains 21 attributes and 741 instances. The attributes present in the dataset are: ID, Reason for absence, Month of absence, Day of the week, Seasons, Transportation expense, Distance from Residence to Work, Service time, Age, Work load Average/day, Hit target, Disciplinary failure, Education, Son, Social drinker, Social smoker, Pet, Weight, Height, Body mass index, Absenteeism time in hours

2. Perform Exploratory Data Analysis on this dataset and report your main findings.

3. Perform Clustering Analysis comparing at least two different techniques. What meaningful groups can you identified within the dataset? Notice that you may need to clean and/or preprocess the data before the analysis.

4. Perform Discriminatory Analysis, finding classification rules with at least three different techniques (one of them should be SVMs). Notice that further to the preprocessing of the previous task, you may need to define your prediction target (choose between binary or multi-class as you prefer). What attributes are more relevant to predict absenteeism?
NB. Analyses should be carried out using Python or R language, and results should be reported using Jupyter notebooks in additionally the analysis results should be returned through a python API Rest, this way other analysts can be use the information to have other insights.