IMPORTANT NOTICE

Please read the following notice carefully. It is a technical explanation of how we will solve our integration exercise.

Data Preprocessing:

First, we will preprocess the data. We devote a significant amount of time to this step as it is a crucial part of every analytical task.

We will start working on the 'Absenteeism_data.csv' file and take it to a usable state in a machine learning algorithm.

We suggest you go through all the videos provided in this section, so you know how much detail we go into when explaining the data preprocessing concepts. If you feel you are relatively new to this field or need refreshing, you'd benefit from completing the entire section.

Alternatively, if you are more experienced in data preprocessing, you could do the entire preprocessing as homework. Should you prefer to accept this challenge, you can download the 'data_preprocessing_homework.pdf' and follow the instructions provided in there.

Note: If you work correctly, regardless of whether you follow the lectures or do the homework on your own, you will transform the Absenteeism data into the data set attached here as 'df_preprocessed.csv'.

After the data preprocessing part has been completed, you will act as a member of the team of data scientists and machine learning engineers.

Machine Learning:

This section will incorporate the work you did in the preprocessing part into the code necessary for making the next step. Namely, to develop a model that will predict the probability of an individual being excessively absent from work.

This will be a logistic regression model. Numerous machine learning tools and techniques will help us at this stage. At the end, we will store our work as a Python module that we will call 'absenteeism_module' and will thus preserve it in a form suitable for further analysis.

Loading the 'absenteeism_module' and integrating Python and SQL:

In this section we will load the 'absenteeism_module' and use its methods to obtain predictions. Then, we will integrate Jupyter and MySQL Workbench to show you an example of how Python and SQL can be

connected. At the end, we will export the final version of the data set as a *.csv file that will be ready for further analysis and visualization.

Analyzing the predicted outputs in Tableau:

Finally, we will use Tableau to analyse three separate dependencies between the inputs of our model. The visualizations we will obtain with this software will help us a great deal while looking for insights.