Dataset link: https://www.kaggle.com/code/hypnobear/absenteeism-at-work-dataset/notebook

**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.

we have removed unnecessary columns from the table.

Grouping the variables called classification.

In a nutshell, we have created a dummy variable name REASON which indicates the data about the reason of absence. goes like this.

**Reason\_0 -> when no reason given**

**Reason\_1 -> Various diseases**

**Reason\_2 -> Pregnancy**

**Reason\_3 -> Peculiar reason like poisoning**

**Reason\_4 -> light diseases.**

Cross checked the data during preprocessing if anything missed out or error occurring.

Created various checkpoint to save the work in various points.

If a person has been absent due to reason 0, this means they have been away from work for an unknown reason. Hence, this column acts like the baseline, and all the rest are represented in comparison to this.

Consequently, dropping this column would allow us to only conduct the analysis for the reasons we are aware of. And that’s exactly what we want to do - explore whether a specific known reason for absence induces an individual to be excessively absent from work. That’s why we don’t really need to keep in our data set information about someone who has been away due to an unknown reason.

**Machine Learning:**

This section will incorporate the work we 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 we 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.

**Connecting MySQL with jupyter Notebook:**

Here we used **pymysql** module in jupyter notebook which helps to connect with MySQL workbench directly from jupyter notebook. We created new Database and added new table. Then uploaded the values to that table.

here we are testing if our model is working properly or returning similar outputs we got in regression. we used sample data from our main data file and input it to our model from our linear regression.

**Analyzing the predicted outputs in Tableau:**

Finally, we will use Tableau to analyze 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.

<https://public.tableau.com/views/AbsenteeismAnalysi/AgevsProbability?:language=en-US&publish=yes&:display_count=n&:origin=viz_share_link>

**Age vs Probability**: Firstly, we will be visualizing Age vs probability. Here we can interpret that, for certain age there is a probability of absence (%)

**Reasons vs Probability:** we know, if the probability is over 50%; We can conclude most people were absent from their work because of REASON\_1.

**Transportation Expense vs Children:** The higher the Transportation Expense, the higher the chance of being absent from work. Also, the average cost is between 220 -240 for those who have 1 or 2 children’s