The task ahead of us

It's clear that the problem that we're facing is a supervised problem, as we have in our power historical data that we can use to create the models. Thus, we the problem is either a classification or a regression problem. To know this we just have to notice that we're not looking into a specific numeric value, but instead we're looking for discrete values to determine if the client will file a claim on their travel or not. Therefore the predictive task has to be a classification problem, because we're only concerned of the discrete values.

Features that may be useful

Considering that we can know everything of the client that is asking for the insurance I believe that the following features could be very useful to be able to create a better model

- *Travel destination location*. This feature is important because there are some travel destinations that exist a higher chance of something going wrong, mainly because of crime but it can also be due to bad weather conditions in certain dates.
- *Start of travel month*. The month the client will be traveling to the destination, this could help to detect possible cases of filing claims due to damage caused by weather conditions.
- *Travel length* (How much time the client will be traveling). Longer travels normally leads to having a higher chance of something happening, thus increasing the chance of filing a claim
- *Insurance history* (There were any previous claims that client has filed?). If the person has previously filed claims may be because they're constantly putting themselves in dangerous scenarios.
- *Age*. Mainly based on the idea that younger people are more inclined to more dangerous activities specially if they're traveling for pleasure
- *Type of insurance* (or how much is the insurance). If the insurance company provides different type of insurances it may be possible to know which is the type is more common to file a claim
- *Business travel?* A business travel may be safer and less probable to file a claim.
- *Rating of the hotel expected to be hosted*. A lower rating may be due to having a more hazardous environment inside and/or outside the hotel

Learning procedures

To be able to create the model there are some possible learning procedures we can choose, these are the recommended:

- Decision Trees
- Naive Bayes Classifier
- SVMs
- Classification Neural Networks

Evaluate the performance of the system

To evaluate the performance of the system we'll first divide our dataset 70-30 (70% for training and 30% for testing). Using this splits it will be used the testing dataset to evaluate using a metric to determine the performance of the model trained with the training dataset. This metric can be either of this:

- Accuracy (Check how many times the classifier predicted correctly), using this metric we should then look for having the **bigger** accuracy possible
- Categorical Cross-Entropy (A loss function specific for classification problems), using this metric we should then look for having the **smaller** loss possible

In this case I recommend to use the accuracy, as it's easier for everyone to understand what the metric is and why having a bigger accuracy means a higher performance