# Data Science – Classification

In the role of a data scientist, you provide relevant and analytical insights to existing data sets. This data comes most of a time in a raw manner, and a lot of time data transformation and accurate presentation is the most time consuming job. Also most of this data is currently stored in Data Lakes, Data Warehouse and other analytical environment. In the end of a ML or DL exercise, a data scientist develops also to make a complete application that is deployable. The following exercise is meant to pursue these multiple skills.

The goal of this exercise is to check your analytical skills and application developer knowledge to fulfil the work of a data scientist. You are allowed to choose whatever environment you like, but the exercise should be sharable and deployable. For traceability, one can use as a starter a Notebook, but please in the end provide a proper application framework.

Given the file attached, the task is to create an analysis on all relevant features for a binary classification problem, being able to model it and create a proper ruleset in order to determine the combination of features providing the classification results. If you can execute this problem in any Cloud environment (Azure preferred), that would be the differential technical solution. Extra credit goes for the ones using Spark framework.

## Task

Build a small **application***, also locally but preffered in any cloud you want*, based on the insurance\_claims.csv in order to:

1. The classification target attribute is the fraud\_reported column: Y (Yes) / N (No)
2. **Analyse** the original columns from data sets, identify and remove all kind of outliers and data issues that would endanger the accurate modelling. **Show** why those transformations were needed and prepare a wrapper for future similar data
3. **Divide** the data properly into training, testing and prediction data sets
4. Create the **classification model**, prove its efficiency on the testing model and apply it for prediction data set showing also the basis ruleset for the model
5. Create an **application** including the previous sets and imagine the application should work on any future data with the same structure (columns and types) as the one in the csv file.

## Requirements

* You create a small notebook including all steps above
* You create a deployable application according to point 5
* You are able to demonstrate and explain us the setup
* You are able to demonstrate and explain us the setup.

## Nice to have

* You copy the file in any Cloud provider (Azure preferred) and run the exercise using Cloud infrastructure
* Usage of any CI/CD tool to test, build and deploy your project (Gitlab, Circle CI, Travis or any of your preference).