Agenda

- Problem Statement
- Objective
- Dataset & Data Description
- Solution Steps:
- Load data
- Understand your data: Data Analysis and Visualizations (EDA)
- Pre-process the data
- Prepare train and test datasets
- Choose a model
- Train your model
- Evaluate the model (F1-score calculation)
- Optimize: repeat steps 5 - 7
- Conclusion
- Prediction on New Test data
- Load the new test data
- Fill missing values if any
- Preprocessing and cleaning the data
- Predict the target values

Problem Statement

Objective

A hospital in the province of Greenland has been trying to improve its care conditions by looking at historic survival of the patients. They tried looking at their data but could not identify the main factors leading to high survivals.

You are the best data scientist in Greenland and they've hired you to solve this problem. Now you are responsible for developing a model that will predict the chances of survival of a patient after 1 year of treatment (Survived_1_year).

Dataset & Data Description

The dataset contains the patient records collected from a hospital in Greenland. The "Survived_1_year" column is a target variable which has binary entries (0 or 1).

- Survived_1_year == 0, implies that the patient did not survive after 1 year of treatment
- Survived_1_year == 1, implies that the patient survived after 1 year of treatment

Data Description:

- ID_Patient_Care_Situation: Care situation of a patient during treatment
- Diagnosed_Condition: The diagnosed condition of the patient
- ID_Patient: Patient identifier number
- Treatment_with_drugs: Class of drugs used during treatment
- Survived_1_year: If the patient survived after one year (0 means did not survive; 1 means survived)
- Patient_Age: Age of the patient
- Patient_Body_Mass_Index: A calculated value based on the patient's weight, height, etc.
- Patient_Smoker: If the patient was a smoker or not
- Patient_Rural_Urban: If the patient stayed in Rural or Urban part of the country
- Previous_Condition: Condition of the patient before the start of the treatment (This variable is splitted into 8 columns - A, B, C, D, E, F, Z and Number_of_prev_cond. A, B, C, D, E, F and Z are the previous conditions of the patient. Suppose for one patient, if the entry in column A is 1, it means that the previous condition of the patient was A. If the patient didn't have that condition, it is 0 and same for other conditions. If a patient has previous condition as A and C, columns A and C will have entries as 1 and 1 respectively while the other column B, D, E, F, Z will have entries 0, 0, 0, 0, 0 respectively. The column Number_of_prev_cond will have entry as 2 i.e. 1 + 0 + 1 + 0 + 0 + 0 + 0 + 0 = 2 in this case.)