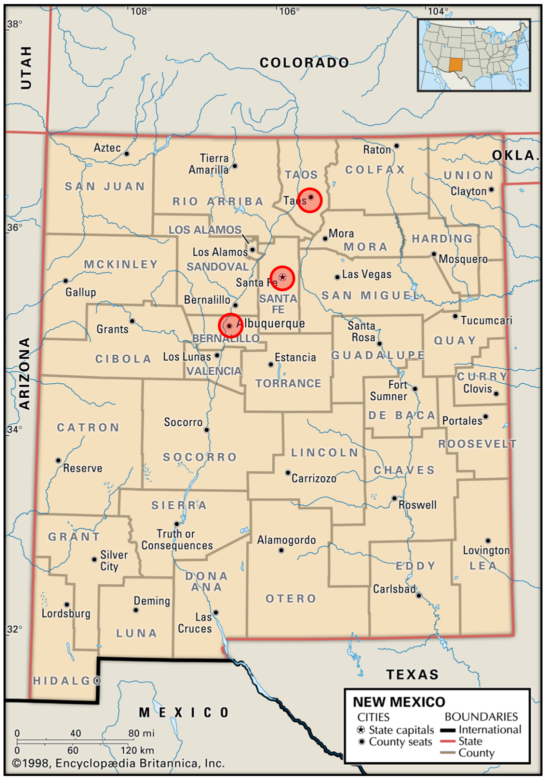
The Boolean pandemic

On January 1st, 2020, an epidemy was originated in Albuquerque, in New Mexico state, and spread on the following days to Santa Fe and Taos. It is estimated that the epidemy has already affected more than 1000 people at the end of February, with a mortality rate of more than 50%.



https://www.britannica.com/place/New-Mexico

While the conditions of the transmission of the virus is still unknown and there are no certainties of what leads a patient to survive or not to the virus, it seems there are some groups of people more prone to survive than others.

In this challenge, your goal is to build a predictive model that answers the question "**What are the people more likely to survive to the boolean pandemic?**" using the small quantity of data accessible of the patients – name, birthday date, severity of the disease, money of expenses associated to the treatment of each family, city and others.

As data scientists, your team is asked to analyze and transform as needed the data available and apply different models in order to answer in the more accurate way the defined question. Are you able to design a model that can predict if a patient will survive, or not, to the boolean pandemic?

The Dataset

Variable	Description
Patient_ID	The unique identifier of the patient
Family_Case_ID	The family identification
Severity	The severity of the disease, where 1 is the lowest and 3 the highest severity
Name	The name of the patient
Birthday_year	The year of birthday of the patient
Parents_Siblings_Infected	The number of parents and/or siblings infected
Partner_Children_Infected	The number of partner and/or children infected
Medical_Expenses_Family	The medical expenses associated to try to contain the epidemy on the family
Medical_Tent	The medical tent where the patient is / was lodged
City	The original city of the patient
Deceased	Flag if the patient deceased or not with the virus (0 = No, 1 = Yes)

The data has been split into two groups:

- Training set
- Test set

The training set should be used to build your machine learning models. In this set, you also have the ground truth associated to each patient, i.e., if the patient survived or not to the epidemy.

The test set should be used to see how well your model performs on unseen data. In this set you don't have access to the ground truth, and the goal of your team is to predict that value (0 or 1) by using the model you created using the training set.

The score of your predictions is the percentage of patients you correctly predict, using accuracy:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Deliverables

You should submit a jupyter notebook with all the steps needed to build and apply the model(s), and a csv file with the number of instances in the test set, containing the columns [Patient_ID, Deceased], and only those columns. The Deceased column should contain the prediction (1 for deceased, 0 for not deceased).

Evaluation

The project will be evaluated taking into account the following criteria:

- Model accuracy;
- The quality of the data exploration, pre-processing, modelling and assessment steps;
- Contributions based on self-study and creativity will be valued;
- The notebook structure and the conclusions / insights / review / justification of techniques of the developed processes in each of the stages of the process. For example, after the data exploration phase, you should write down at the end of this topic in markdown the main insights that you gather from the data during this phase.