Medi-Chal Project

Thomas Gerspacher

March 30th

1 Challenge

The objective is to build an **automatic reader** for the challenge's results which will have **error rate** as a metric. This task is due on next week.
The dataset of the challenge is about **bio-degradability** of chemical compenents.
48 students will participate to the challenge and the language will be R.

2 Presentation of SDV paper by Andrew

Even if the results are kind of blurry, some ideas can be retained and tested. Notably, how they handled missing values, how they encoded categorical variables or their idea of using Gaussian Copula.

3 Medi-Chal

We did a recap of some key aspects of the project.

In order to assess the performance and improve robustness of our generative model, we chose multiple datasets.

- MIMIC III
- Forest fires
- Adult (a1a libsvm)
- Artificial data

Note: Those datasets might change.

One of the task we need to look at is the preprocessing of the input : how to handle missing values, normalisation, how to encode variables? About encoding, some ideas were raise and solutions were proposed in the *Synthetic Data Vault* paper such as frequency-based encoding and others were proposed by Isabelle

like target-based encoding. Researches and tests still need to be performed to choose the best approaches.

Another task is the metric we choose to implement. A small list was already proposed by Adrien and Thomas in their proposal. Maximum minimum discrepancy must be added to this list.