Predicting Crime Categories

In this task you will work with historical police incident reports from San Francisco, which are publically available.

Kaggle

For this assignment you will use the Kaggle platform, which hosts various machine learning competitions. At least one person per team will have to create an account to download the dataset and make submissions.

The competition you will join for this lab is: https://www.kaggle.com/competitions/sf-crime

The goal is to predict the category of an incident from temporal and geographical information.

The dataset contains the following columns:

- Dates — timestamp of the crime incident
- Category — category of the incident (only in the training data). **Variable to predict.**
- Descript — a description of the incident (only in the training data)
- DayOfWeek — the day of the week
- PdDistrict — name of the Police Department Disctrict
- Resolution — how the incident was resolved (only in the training data)
- Address — approximate street address of the incident
- X — geographical longitude
- Y — geographical latitude

Note: We heavily encourage you to augment this dataset with external data sources (except the raw data dump from the San Francisco Police)!

Formalities

- **DO NOT** reuse existing solutions from the kaggle competition and blog posts.
- You should work in groups of 2 - 3 students.
- At the end you should hand in a PDF report of 2-4 pages describing your solution.
- The report should include at least:
  - the names of all team members
  - the details of your model
  - if and what kind of external data you included
  - the score your best submission received on the kaggle leaderboard

- Every member of the team hands in the same report in Moodle (**everyone** makes a submission)
- This assignment is worth 30 Points.
- The deadline is Sunday 18.12.2022 at 23:59:59 (UTC+1).
- Some teams will be invited to present their solution during the last lecture of MLDM.

Some Hints to get you started

- Do you think that the time of day could be a good predictor?
- Is there a way to get an estimate of the police response time?
- What kind of information can you deduce from the incident location (maybe based on a map of San Francisco)?