**Final project: Odor classification**



SECTION 1

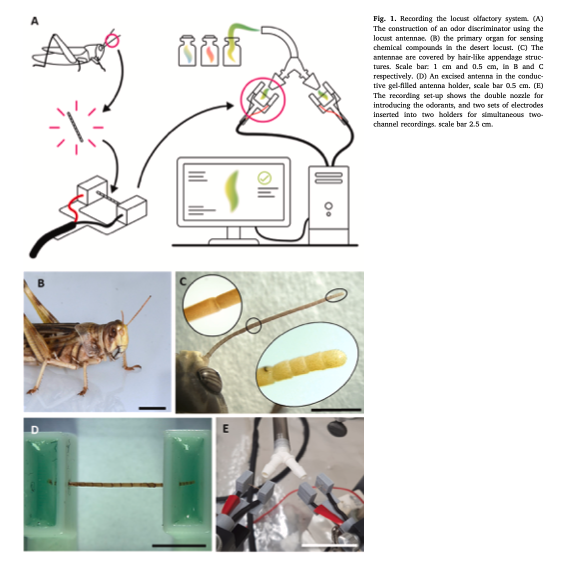
- Team members: Ben Sivan, Ethan, Samuel, Ella

Ella helped with EDA and writing the report

please add what you did

Descriptions of the project:

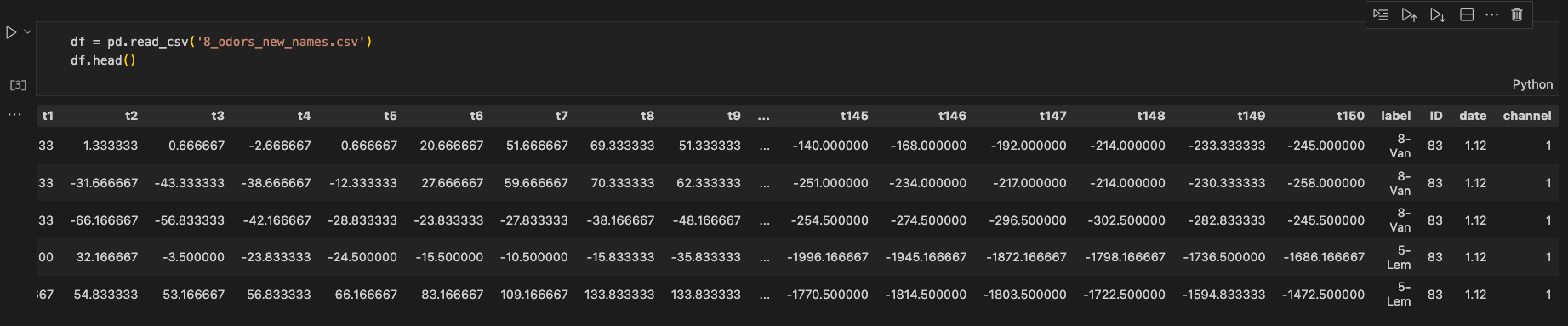
For our final project, we are collaborating with Neta Shvil a PhD student in Biology in Tel Aviv university. Her research focuses on neuroscience and how odour are perceived by animals. During her PhD, she studied more specifically how antennas on locus record and identify smells. Here we would like to continue her word and propose different and more elaborate algorithm for odour classification. Her experiment consists of 1 or 2 antennas cut off (on both sides) from an locus placed in a conductive gel and on which 2 electrodes are mounted on each extremity. This is linked to a recording computer. A smell is then send in the direction of the antenna producing a stimulus by the antenna and recorded at 100Hz. Hence, the data consist of a spectral recording of different smells. In total 8 smells were tested in this manner.



SECTION 2: EDA

The dataset is composed of about 972 rows and 155 columns. This data set has been slightly pre-processed by Neta and her students in order to (i) withdraw the physical influence of the blow of the smell and (ii) normalise the offset of each smell (making each record start at the t0) and (iii) subtract the blank.

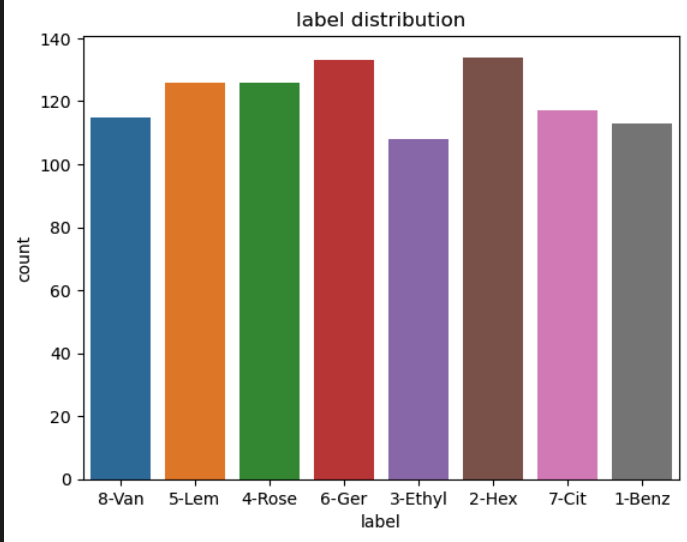
Here is an extract of the raw dataset :



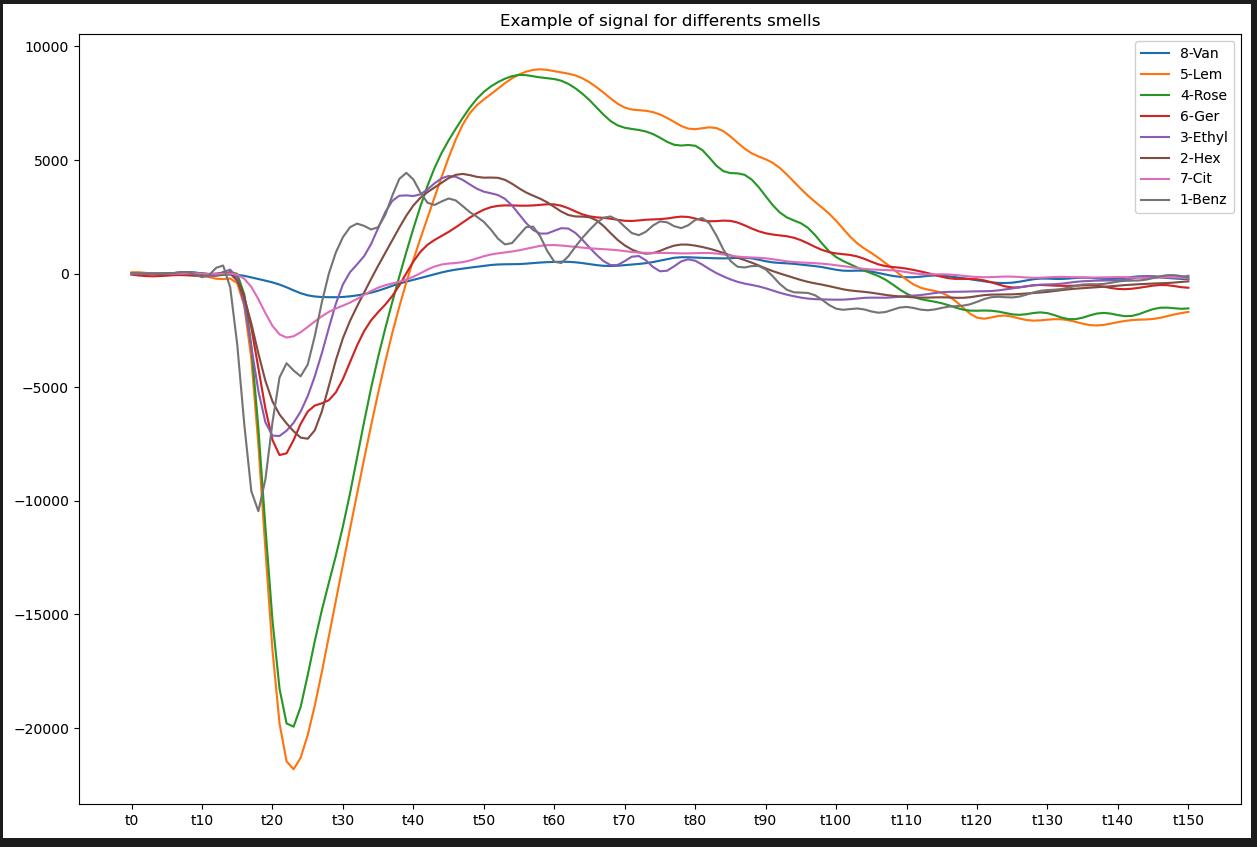
We checked that the dataset does not contains any nan or null values. The shape is (972, 155)

Below are chosen plots and comments to visualize the important characteristics of this dataset:

* We can see that we have a mostly well-balanced target feature with approx. 100 to 134 sample per label.



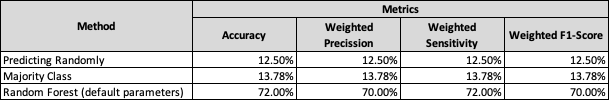
* Our target variable is categorical, we are on a classification problem.
* the ID feature represent the identification of the antenna. We will need to be careful during the splitting stage to make sure that the same antenna is not in train and test, to prevent data leakage
* our dataset is a signal, so the features are the time steps
* Here is an example of the differents smells signals :



we see that they are all significantly different so this leads us to be hopefull about the capacity to classify them.

SECTION 3: Baseline model

Since our dataset is overall quite balanced our classification metric will be accuracy. Below is a table of the metrics for 3 baseline models:



We can see on the first order that the random forest is significantly better than the two other baseline model. This suggest that there is room for improvement since the random forest model is simply used with no preprocessing and only the default parameters

SECTION 4: Next steps and questions

Moving forwards we need to concentrate

- data augmentation :our dataset is critically small, we need to figure out how to augment data in order to have more robust predictions/model. Need to look into SMOTE

- preprocessing: smoothing the curves ? since we are dealing with a time series/signal as an input the classical scaling and normalizing are not the best approach.

Models to test in the future

- Gaussian Mixture Model : soft clustering in order to classify the actual component of the mixed smell and not just with the label “mixed smell”