

Authors: Patrick O'Connell (oconne16) & Jahnvi Patel (jpate201)

1. What is your data and task (including data preprocessing steps)?

Dataset for this project is acquired from *Young People Survey* from *kaggle.com*. The main task of the project is to predict a person's "Empathy" as either "very empathetic" or "not very empathetic". The dataset provided in the .csv files is either numerical or categorical. The numerical values were sorted based on 0 being the least and 5 being the most. While, the categorical data was sorted more delicately based on specific keywords (i.e. "never smoked" = 0, or "doctorate degree" = 5). The empty fields were filled with the total average of the column. After loading and cleaning up the data, we normalized the data so all the categories can have the same scale for a fair comparison between them.

2. What ML solution did you choose and, most importantly, why was this an appropriate choice?
3. How did you choose to evaluate success, including baselines, experimental setup (e.g., % train/dev/test), metrics?
Experimental Setup: Train, validation and test dataset were chosen based on rule-of-thumb discussed in class. Given 1110 responses, 810 were chosen as training data, 100 were chosen as validation data and the remaining 200 responses were set aside as test data.
4. What software did you use and why did you choose it?
5. What are the results?
6. Show some examples from the development data that your approach got correct and some it got wrong: if you were to try to fix the ones it got wrong, what would you do?