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 caregories 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?
   We used variety of algorithms to see which algorithm would run best with the given dataset. Ultimately, Random Forest gave us the best results with a testing accuracy of 79%, approximately 15% better than the baseline.

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 1010 responses, 810 were chosen as training data, 100 were chosen as validation data and the remaining 100 responses were set aside as test data. Success of the algorithm is based on its performance compared to the baseline collected. If the algorithm performs better than the baseline accuracy then it is marked as a success.

4. What software did you use and why did you choose it?

5. What are the results?

Algorithm	**Test Accuracy**
Baseline	63.63636363636364
Sqaured Loss	68.68686868686869
Logistic Loss	68.68686868686869
Hinge Loss	70.70707070707071
Decision Tree	70.70707070707071
KNN	72.72727272727273
Random Forest	78.78787878787878

Figure 1: Testing Accuracy of "empathetic" vs "not very empathetic"

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?