# Homework 5 Mini Project

**Name: Shobhit Lamba**

**UIN: 655612480**

For my project, I have chosen to work on Task 1: Predicting how empathetic a person is based on the Young People Survey dataset .

The ML solution I employed works as follows- First I split the data into train/dev/test datasets in a 60%-20%-20% ratio. Then I trained 9 separate classifiers on the training data. The classifiers used are: Gaussian and Multinomial Naïve Bayes, Multi-Layer Perceptron, SVC, Decision Tree Classifier, Logistic Regression Classifier, AdaBoost, Random Forest and KNN. When evaluating them, I discovered that the accuracy peaked at **~40%** for MLP with most classifiers returning an accuracy below that. To get a better result, I turned to ensemble methods and applied Voting Ensemble. This gave me a peak of **<=45%** accuracy. While this is still a fair increase considering the baseline performance and that a random function would give us a **20%** average accuracy, I believed that it could still get a little better. I tried stacking ensemble methods. I used Bagging Classifier on all the baseline classifiers except Random Forest and AdaBoost, and then used voting classifier on all those base classifiers. Finally, this gave me an average accuracy of **~50%** on dev data and **~47%** on test data. A point to be noted here is that I used a for loop to find the best count of features needed to be trained on and the process was entirely done by the program and not manual.

I chose to evaluate on basis of accuracy. As mentioned above, a random classifier will give an average accuracy of 20% on a dataset with 5 classes. It is fair to say my model is performing much better.

I did all my work in python using the Anaconda environment, working primarily with sklearn, numpy and pandas libraries. I did so because these libraries have a wide array of functionalities available for us to work with and are easy to use.

To a great extent my approach got the right results, but I had to leave some part of the data out so that I can easily work with and move forward with my ML model. Some of the features that were categorical were left out because my classifier only dealt with digits. But that must have resulted in missing some features relevant to the classification, such as gender, lying etc. For example, at some places during validation, the classifier predicted 3, neutral empathy while it was 5, high. If I had included gender, I could have used that to make the correct prediction (As one can notice from the data, female gender showed more empathy in general).

Also, because we were asked not to select the features on our own, human understanding of the data was missing in the classification. There were many aspects of the data which were directly related to the target class, but were missing from the features selected by the algorithm.

All, the results and the reasoning behind why everything was done the way it was, is included in the **jupyter notebook**.