HW5: Problem 1-Empathy Analysis

Problem Statement:

have been tasked with predicting how suitable a person is for this task by predicting how empathetic he or she is to recruit volunteers to help with Alzheimer's patients. Using the Young People Survey dataset (https://www.kaggle.com/miroslavsabo/young-people-survey/), predict a person's "empathy" on a scale from 1 to 5.

ML Solution:

Data Cleaning and Pre-processing

1. **Missing Data**: We are replacing any data that's missing by the **mode** i.e. most frequent value of that column.
2. **Categorical Data**: most of the classifiers need numerical data to train on. Hence, we create *dummy* features for all the categorical attributes such as *Gender, Punctuality, Internet Usage* etc. These dummy features are based on each category and whether a person belongs to that category or not is marked by **1 or 0**, respectively.

After cleaning, we divide dataset into three parts: **60% of data is training set, 20% is validation set and rest is testing set**.

Tried Baseline Classifiers

Tried 'Assigning Most Frequent Class', 'Random Class' and untuned/default 'Decision Tree Classifier', 'SVM Classifier', 'Logistic Regression Classifier'. Using accuracy as our evaluation score, we get 26% as the highest accuracy out of these.

Feature Selection using Recursive Feature Elimination with Cross Validation(RFECV)

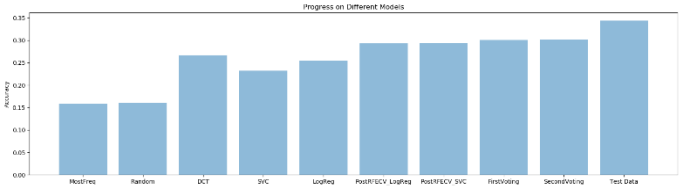
Ranked all the features using RFECV as their were a lot of features that might not be meaningful for our purpose of calculating Empathy and then selected those features which were ranked 3 or below. This gives us in general 10 to 15 features. I had tried Lasso coefficients too but RFECV ranking worked better. Improved accuracy by 3-6%.

Hyperparameter tuning using Grid Search with Cross Validation

Using Grid Search with 5 fold cross validation, we tune parameters for DCT, SVM and Random Forest Classifier models to get their best setting curated for our data.

Bagging and Voting Classifier

Applied bagging on the above three models mentioned which improves the stability of base classifier by model averaging. Then I applied Voting Classifier on the Bagged DCT, SVM, RFC which improved the accuracy by 1-3%. Then, I included few more classifiers at their default settings to the voting classifiers: K Nearest Neighbors (a clustering algorithm), Multi-layer Perceptron (a neural network) and Quadratic Discriminant Analysis (Bayes' Rule based decision boundary) which improves the accuracy by 0.1% to 0.4% on validation data set.



Received accuracy for testing data as 34.4% which is almost 4% more than the validation (30.6%).

Used Anaconda environment as it is the most convenient and consolidated place for all machine learning tools. Majorly used package was Pandas, Numpy for data handling and Sklearn for machine learning methods.

Judgment	Giving	Numbe	Smokir	Alcoh	Alcoh	Punct	Lying_	Inter	Interr	Genc	Gend	Left -	Educ	Educ	Educ	Educ	Only	Only	Villag	Villag	flats	Y_val	Pred
3	3	2	0	1	0	0	0	0	0	0	1	0	0	0	0	1	1	0	1	0	1	4	4
5	2	1	0	0	0	1	1	0	0	1	0	0	0	0	0	1	1	0	0	1	0	5	5
3	2	1	0	1	0	1	0	0	0	0	1	0	0	0	0	1	1	0	1	0	1	3	4
5	2	1	0	0	1	0	0	0	0	0	1	0	0	0	0	1	1	0	1	0	1	5	5
5	2	2	0	0	0	0	0	0	0	0	1	0	0	0	0	1	1	0	0	1	1	5	5
4	3	3	1	0	0	1	0	0	0	0	1	0	1	0	0	0	1	0	1	0	1	4	4
5	1	1	0	0	0	0	1	0	0	0	1	0	0	0	0	1	1	0	1	0	1	2	5
3	3	0	0	1	0	1	0	0	0	0	1	0	0	0	0	1	0	1	1	0	1	2	2
Comapring the actual 2 but predicted wrong, with correct 2																							
-2	2	-1	0	1	0	1	-1	0	0	0	0	0	0	0	0	0	-1	1	0	0	0		
Comparing the wrongly pred 2 as 5 with actual 5																							
0	-1	0	0	0	-1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0		

This image shows examples of correctly predicted(non-red rows). We can clearly see that the features for all predicted as 4 have very similar values. But, there is an entry which was 2 but was predicted as 5. When we compare the incorrectly predicted row (second red row) with a correct 2 and 5's features, we can see that it is different from 2 for many features where as almost similar to features of 5(same are green zeros, different are non-zeros). Hence, I believe the feature selection was weak as the features that would have uniquely identified is as 2 weren't present. I would probably try different set of rank thresholds and also different feature selection methods such as SelectKBest with various scoring algorithms as base. Also, we were asked to not manually go through the features but I believe a human insight could have also helped in the feature selection to be better.

Github link: https://github.com/SPanday/HW5_Problem1