Young People Survey – Project Report

The dataset used is Young People Survey dataset (https://www.kaggle.com/miroslavsabo/young-people-survey/), the task at hand is to predict which student volunteer would be more suitable for helping Alzheimer's patients by predicting how empathetic he/she is on a scale of 1 to 5. The machine learning solution proposed involves the following steps:

1. **Elimination of null/missing values**: Eliminating the null/missing values in numerical features by replacing the null/missing values by the most frequently occurring number in that particular column. For categorical features, if a null/missing value is present the entire row is dropped. Similar process is followed for the column Empathy.

2. **Visualization of Relationships**: The relationships between various features is be visualized using various visualization techniques (Pie Charts, Correlation matrix, Barplot).

3. **Encoding of categorical features**: Categorical features are encoded using One Hot Encoding and Label Encoding (prevents feature explosion and dummy trap).

4. **Splitting the data**: The given data is split into training data (80%) and testing data (20%) to train and validate the performance of the model

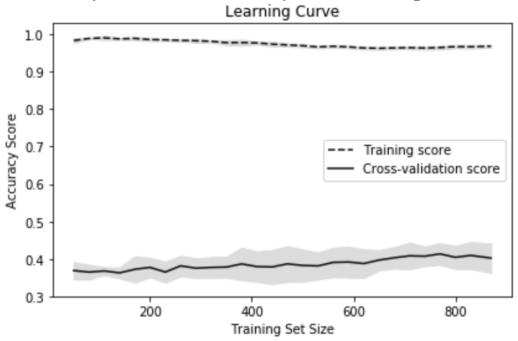
5. **Feature Selection**: There are 168 features available after encoding, selecting the important feature is important as it will improve the learning of the classifier. PCA is used for feature reduction and selection

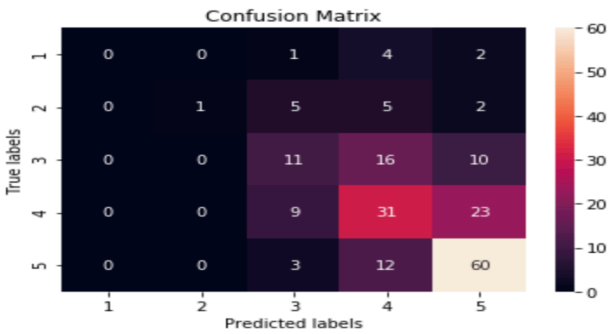
6. **Hyperparameter Tuning**: Used GridSearchCV to tune the hyperparamers (C and Gamma of SVC).

7. **Classification and evaluation**: In this step training and testing the performance of various classifiers is performed as it helps to determine which model works well with the data. The performance of the model is evaluated using accuracies, precision, recall and F-1 Scores.

This machine learning solution was an appropriate choice as the data had numeric and categorical features but no heavy text data in which case machine learning solution - Bag of Words, Unsupervised classification, Vectorization, etc. would be an ideal choice. The evaluation of success was done on the basis of test accuracy, confusion matrix, precision, recall and F-1 score values. The best accuracy is recorded for Kernelized Support Vector Classifier (RBF kernel) which is 52.82 % (Precision: 0.52, Recall: 0.53, F-1 Score:0.49) which is better than the baseline and other classifiers which gives 38.46% accuracy. The confusion matrix plots show which labels were falsely predicted, the model can be trained more on more such data points to decrease false classification.

Software used: Python, sklearn (various algorithms), seaborn, matplotlib, xgboost





Extra credit repo - https://sylvestersavioraj@bitbucket.org/sylvestersavioraj/iml_hw5_young_people_option-1.git