CS 412 - Mini Project

Repository link: https://github.com/kvkadakia/Machine-Learning-Project

Problem solved: You are working for a non-profit that is recruiting student volunteers to help with Alzheimer’s patients. You have been tasked with predicting how suitable a person is for this task by predicting how empathetic he or she is.

We make use of the Young People Survey dataset, please find the link to the dataset below: https://www.kaggle.com/miroslavsabo/young-people-survey/

Predicting a person’s “empathy” on a scale from 1 to 5.

Dataset preprocessing steps:

1. Cleaning the data: Here, I clean up the data firstly by replacing NAN values. I simply find the NAN values and replace them by the column median.

2. One Hot Encoding: Categorical variables are converted into a form that could be provided to the ML algorithm to do a better job in prediction.

Solution:

I make use of Recursive feature elimination in order to select the features that are best able to predict Empathy. Now I perform cross validation in order to evaluate various models and compare with baseline models. LogisticRegression, LinearDiscriminantAnalysis, KneighborsClassifier, DecisionTreeClassifier, GaussianNB, SVC , Ensemble models are evaluated and SVC gives best performance.

Experimental setup:

The data is split into train, test and validation in the following manner:

```
X_train, X_test , Y_train , Y_test = train_test_split(X,Y,test_size=0.25,random_state=26)
X_train, X_val, Y_train, Y_val = train_test_split(X_train, Y_train, test_size=0.2, random_state=1)
```

Evaluation:

10 fold cross validation is performed in order to compare various models and find out the best model which works from the given set of models. Importantly configured with the same random seed to ensure that the same splits to the training data are performed and that each algorithms is evaluated in precisely the same way. Now I use SVC to evaluate the model.

Result:

After tuning hyper parameters on the validation set we test using SVC which is the best model found by cross validation. The selected model gives an accuracy of 45.84% on test data.