HOMEWORK 5 REPORT

Harsh Shah

Link for github: https://github.com/harshshah6595/Machine-Learning-Mini-Project

Goal for project: The Data set is actually a survey which was taken from students at FSEV UK. The data set contains 150 columns and 1010 rows. The original questionnaire was in Slovak language and was later translated into English. The main goal of this project is to "PREDICT HOW LIKELY A PERSON WILL PAY MORE MOENY FOR GOOD, QUALITY OR HEALTHY FOOD".

Preprocessing steps: The first step was to remove all the rows where target column had missing values. Then the data categorical data was handled by using one hot encoding technique. This is because they can be useful for prediction. Outliers were removed to avoid anomalies. Then the missing values were filled. It was filled using median of the column. The reason is the data is mostly integer numbers and median will also give cross approximation of true values.

Techniques used: I used sklearn packages. It has most of the basic machine learning models. Tried various methods like scaling, PCA, Random forest, SVM, KNN, Decision tree, MLP, and my own model which is ensembling technique. It has various models in it which are extratrees, randomforest, adaboost, gradientboosting, naïve base, svm, mlp and bagging classifier. The reason of using voting classifier is because every classifier will put in their predictions and the class which max classifiers gave will be the actual prediction. This helps to increasing accuracy.

Experimental Setup: The whole dataset was divided into train, development and test set. The distribution is 64% training, 16% development and 20% test data set.

```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2,random_state=1)
X_train, X_val, y_train, y_val = train_test_split(X_train,y_train, test_size=0.20, random_state=1)
```

Evaluation: I evaluated different models based on their accuracy. I tried different baseline models. The accuracy of decision tree is a between 25 to 30%. The accuracy of knn is between 32-34%. Mlp performs similar and gives around 25-34%. My model which is ensembled using many classifiers gives accuracy from 40-43%.

Results: The ensembling model which was developeed gives better performance compared to other baseline models. It gives accuracy of 40 to 43%

References

http://www.cs.ccsu.edu/~markov/ccsu_courses/datamining-3.html

http://contrib.scikit-learn.org/imbalancedlearn/stable/generated/imblearn.over_sampling.SMOTE.html

http://haridas.in/outlier-removal-clustering.html
https://www.kaggle.com/miroslavsabo/young-people-survey/kernels

http://scikit-learn.org/