HW 3

Advanced Data Analysis in Python

The purpose of this assignment is to review and practice fundamental machine learning concepts.

First of all, when we examine the data set given to us, we see that the 'voted' feature is given as a categorical variable.(True or False) So we replace this feature to 1 and 0. (one-hot encoding) Then, we divide our data, which we converted into dataframe with the help of pandas, into features and labels.(X and y) We determine y as 'voted' column because we aim to build a predictive model of whether or not a respondent likely voted in their last presidential election.

```
X = mydata.drop(columns=['voted'])

y = mydata['voted']
```

Then we divide our data into training and test sets.

```
Xtrain, Xtest, ytrain, ytest = train_test_split(X,  y, test_size=0.4, random_state=0)
```

I will use the Random Forest Classifier algorithm to train the datasets we obtained for all features. This algorith has several hyperparameters. In order to find the best approach on this algorithm, GridSearchCV is used and a couple of hyperparameters were determined before.

```
grid = {
   'n_estimators': [200, 500],
   'max_features': ['auto', 'sqrt', 'log2'],
   'max_depth' : [4,5,6,7,8],
   'criterion' :['gini', 'entropy']
}
```

After we run

```
CV_RF = GridSearchCV(estimator=model, param_grid=grid, cv= 5)

CV_RF.fit(Xtrain, ytrain)
```

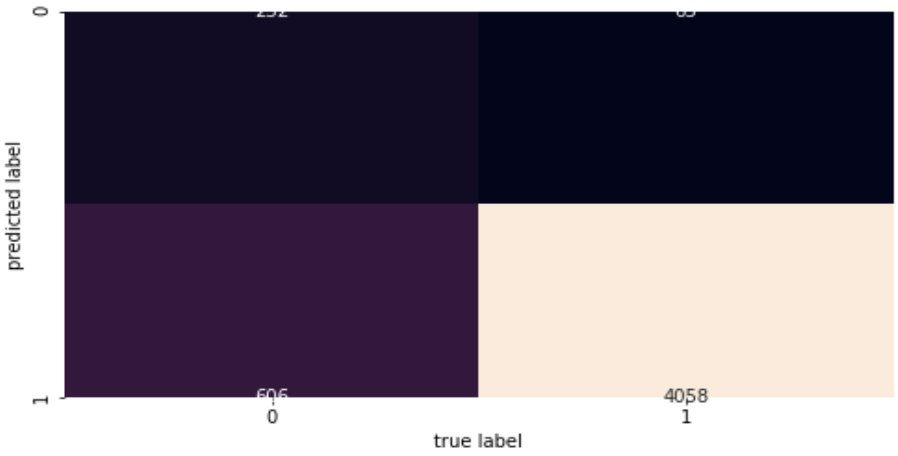
These modules, we get best score and best parameters as:

```
Best score: 0.8583668005354752
Best parameters: {'criterion': 'entropy', 'max_depth': 8, 'max_features
': 'auto', 'n_estimators': 500}
```

And we observe some metrics about classification reports below:

	precision	recall	f1-score	support
0	0.29	0.79	0.43	317
1	0.98	0.87	0.92	4664
accuracy			0.87	4981
macro avg	0.64	0.83	0.68	4981
weigted avg	0.94	0.87	0.89	4981

And confusion matrix:



Result scores of cross validation is : array([0.8361709 , 0.62955823])

Above, we worked on for all feauters. Now, we used the Dimensionality- Reduction Technique.(PCA test) In this approach we determine n_components as 10 and 20 in order to compare the results. When we compare the result, we see that, when we choose n_component s=10 then we get better accuracy score. (0.86) So, if we use dimensionality reduction techniq ue for selecting feauters we should choose n_components = 10.