INTL 472 ADVANCED DATA ANALYSIS IN PYTHON

DENIZ AYCAN
HOMEWORK 3 REPORT

For the third homework, I aimed to estimate how likely citizens are to vote by practicing machine learning models based on data of their status and identities. To estimate this, I used "cses4 cut.csv" file which contain a subset of the CSES Wave Four data set. I chose gender, union membership, socio-economic status and number of children in the household as variables. I extracted them and splitted into two groups, which are categorical variables (catx) and ordered variables(orderx). For both variables, I excluded the nan values and used SimpleImputer for transforming the data. For orderx, I used mean value strategy to substitute missing data. For catx, I used most frequent since the mean is not statistically meaningful. Later I scaled ordered ones using StandardScaler. As categorical variables cannot be ranked meaningfully, I used OneHotEncoder for catx for more meaningful results. Then I optimized the categorical variables used in the analysis with f_classif where k equals 6 by testing which k would work better. That way I finished preprocessing and variable selection steps.

After finishing data cleaning, I wanted to see which ML model would work with my dataset, therefore I did cross validation with ShuffleSplit, where I chose %30 of the data as test and n equals 5. I cross validated the following models: Decision Tree, Logistic Regression, Support Vector Machine and K-Nearest Neighbors and then performed train-test-split operations.

Cross Validation Scores (K=5) Follow as:

	Mean	STD	Results
Decision Tree	0.8238758029978586	0.0028428090037155647	[0.82949679 0.82307281 0.82253747 0.82173448 0.82253747]
Support Vector Machine	0.8247323340471091	0.002972900080709811	[0.83056745 0.82360814 0.82253747 0.82280514 0.82414347]
Logistic Regression	0.8247323340471091	0.002972900080709811	[0.83056745 0.82360814 0.82253747 0.82280514 0.82414347]
KNeighbors	0.7833511777301927	0.027859692821771302	[0.78479657 0.75856531 0.74678801 0.80513919 0.82146681]

Based on the results, I chose to continue the analysis by using Support Vector Machine and Logistic Regression, as they have higher and equal mean values. To improve the default scores of SVM and LR, I imported GridSearchCV. I trained my models, check for their accuracy scores and built confusion matrixes to visualize results. GridSearchCV results are as follows:

	Support Vector Machine	Logistic Regression
Accuracy score with default settings	0.8284261241970021	0.8284261241970021
Best optimized score	0.8182444061962133	0.8182444061962133
Best parameters	{'C': 1, 'gamma': 0.001}	{'C': 1}
Accuracy score with best settings	0.8284261241970021	0.8284261241970021

Unfortunately, accuracy scores, both for SVM and LR remained the same after improving the default settings. Therefore it can be said that using default setting would be sufficient in using this dataset. In addition, the similarities between SVM and LR are observed. Lastly, the code runs slowly due to unknown reasons.

Below are the figures: