CSSM 502 : ADVANCED DATA ANALYSIS WITH PYTHON

HW3 Report

Ezgi Aydın

The purpose of this homework report is to summarize outcomes of fundamental machine learning practice over "cses4 cut.csv" data. This cses4 data includes several different demographic values with voted or not values. My aim is to build a predictive model of whether a respondent likely voted in their last presidential election or not.

To do so, I firstly important all of the essential libraries including NumPy, pandas, sklearn, seaborn and matplotlib. After that, I observed the data, performed some data manipulation specially to see null values. My initial strategy was to drop columns including null values for more than half of it, however my model broke by doing so. Therefore, I only observed the null values and dropped those columns manually. After I created cleanDf I performed following modeling operations on this dataframe.

As next I have defined my train and test sets, x train, x test, y train and y test parameter using train test split. After that I have calculated classifiers accuracies for 3 different stages. First without any reduction or feature selection, second with manipulation, dimensionality-reduction and pre-processing, and as third with optimized model and its hyperparameters. Classifiers and their accuracies for each step is as follows:

1. Classifiers before reduction:

```
Classifiers before reduction:
```

	Model	Accuracy
6	Random Forest	86.25%
1	K-Nearest Neighbors	85.19%
2	Linear Discriminant Analysis	83.38%
4	Support Vector Machine	82.07%
3	Decision Tree	77.85%
5	Quadratic Discriminant Analysis	76.15%
7	Bayes	73.89%
0	Logistic Regression	NaN

2. Classifiers with dimensionality-reduction and pre-processing:

```
Classifiers with dimensionality-reduction and pre-processing:
```

	Model	Accuracy
6	Random Forest	85.01%
4	Support Vector Machine	84.22%
2	Linear Discriminant Analysis	83.58%
0	Logistic Regression	83.38%
1	K-Nearest Neighbors	83.04%
5	Quadratic Discriminant Analysis	79.88%
3	Decision Tree	78.14%
7	Bayes	77.27%

3. Classifiers with optimized model and its hyperparameters:

```
Classifiers with optimized model and its hyperparameters:
```

	Model	Accuracy
3	Random Forest	85.25%
1	Support Vector Machine	84.82%
4	K-Nearest Neighbors	83.89%
2	Linear Discriminant Analysis	83.58%
0	Logistic Regression	83.38%

4. Hyperparameters:

```
Best score is: 0.8524625267665952 with estimator: 500 criterion: entropy
Best score is: 0.8482334047109207 with c: 5 kernel: precomputed2
Best score is: 0.8358137044967879 with solver: svd
Best score is: 0.8337794432548179 with penalty l2
Best score is: 0.8388650963597432 with number of neighbors: 9
```

For feature selection, I have used SelectKBest(chi2, k='all').fit_transform(X, y) function from sklearn feature selection. Calculated k scores, ordered columns referring to their k scores and selected top 10 of them for better accuracy. Named this new df highscorers and visualized them. In this visualization I saw that they are not in Gaussion form, so I transformed them to normal distribution using quantile transformer again from sklearn library. With this step I also muted the outliers and focused on the most frequent values of the data automatically. Finally, before the last classifier accuracy test, I tried to optimize hyperparameters by performing top 5 scoring classifiers in loops. At this stage I struggled with the parameters, and to solve this problem I researched online from ml resources for approximately 2 days. This experience also showed me that I need to practice sklearn library and hypermeters more.