Advanced Data Analysis in Python: Homework 3

Shayan Rahimi Shahmirzadi – Student ID: 0080986

This homework aims to review and practice fundamental machine learning concepts. The idea is to build a predictive model of whether a respondent likely voted in their last presidential election. For this purpose, the "cses4_cut.csv" file is used containing a subset of the CSES Wave Four data set.

As requested, different models and approaches have been tested. Here is a shortlist of what has been implemented:

	Model	Accuracy
5	Random Forest	86.99%
3	Linear Discriminant Analysis	84.07%
0	Logistic Regression	83.08%
2	Support Vector Machine	82.75%
6	K-Nearest Neighbors	81.14%
1	Decision Tree	78.43%
4	Quadratic Discriminant Analysis	69.94%
7	Bayes	69.88%

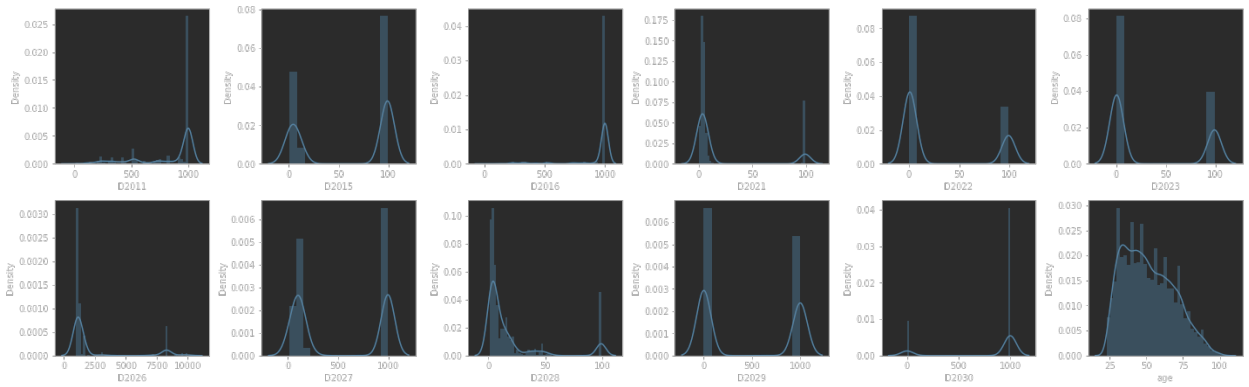
Dimensionality-reduction with feature selection

We can reduce overfitting, improve accuracy, and reduce training time with feature selection, for this purpose, “sklearn.feature_selection.SelectKBest” is used, and 12 features were with the highest score.

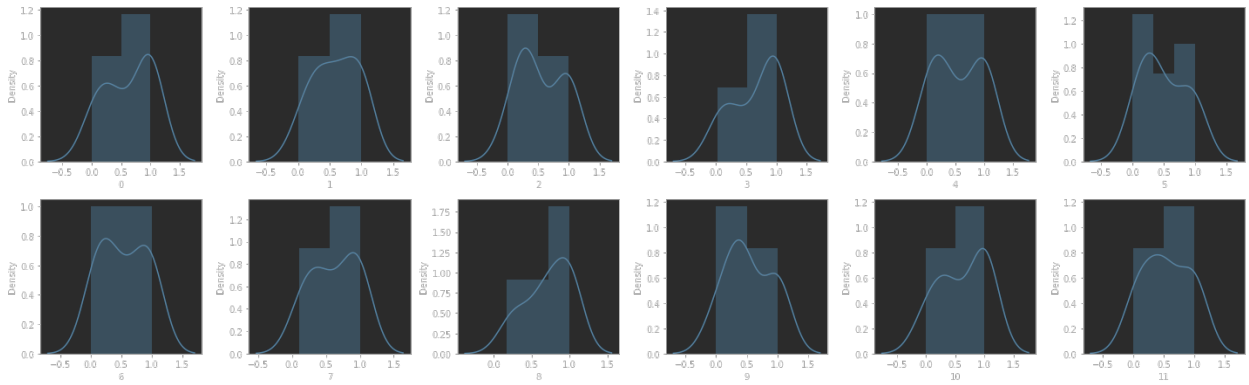
Pre-processing:

There are some unwanted data in the data set like:

1. REFUSED
2. VOLUNTEERED: DON'T KNOW
3. MISSING



These data disrupt the distribution of data. I used the quantile transformer method "sklearn.preprocessing.QuantileTransformer" to solve this problem. This method transforms the features to follow a uniform or a normal distribution. Therefore, this transformation tends to spread out the most frequent values for a given feature. It also reduces the impact of outliers.



Classifiers with dimensionality-reduction and preprocessing

After preprocessing and feature selection, I re-trained the models. Results are as follows:

	Model	Accuracy
5	Random Forest	86.85%
3	Linear Discriminant Analysis	84.07%
0	Logistic Regression	83.08%
2	Support Vector Machine	82.75%
6	K-Nearest Neighbors	81.14%
1	Decision Tree	78.76%
4	Quadratic Discriminant Analysis	69.94%
7	Bayes	69.88%

Optimizing the model and its hyperparameters

I took the top 5 classifiers and regressors and looped them until I found the best hyperparameters. Results are as follows:

	Model	Accuracy
3	Random Forest	86.25%
1	Support Vector Machine	85.95%
4	K-Nearest Neighbors	84.44%
0	Logistic Regression	83.89%
2	Linear Discriminant Analysis	83.81%