QMBU450 - Homework 3 - Report

By Selin Öztürk, 60160

To start working on this project, I investigated the dataset and changed the column codes with their human-friendly explanations. Meanwhile, I realized that I had to preprocess the dataset as there were some categorical attributes which were specified by numbers. Their numeric representation would falsely affect the accuracy of the model, so I decided to identify them to apply one-hot encoding. There were around 30 columns, many of which were categorical, and applying one-hot encoding to each of them would result in a big, sparse matrix that might have increase the complexity of the model. Moreover, I saw that some columns are mostly consisting of "7, 8, 9, 97, 98, 99, 997, 998, 999" which correspond to `NaN`, and removing the rows containing these `NaN` values would cause a lot of information. So, I decided to go over each column to decide if I should keep working with them or not. If a column has at most 1200 `NaN` entries, I removed those entry rows. Else, if a column has more than 1200 `NaN` entries, I dropped that column entirely. In addition, I removed some columns as their definitions in the website were missing or inconsistent. In the end, in exchange for keeping high number of entries, I had to disregard around 20 columns and I was left with 8. Some of the columns (such as `age`) were already numerical, and some of them (such as `EDUCATION`) were categorical but representing the levels, thus I decided to treat them as numeric values. Then, I applied one-hot encoding to the remaining 4 columns and numerically encoded the Boolean values in the target column `voted`, and I was done with preprocessing and I split the dataset into training and test groups.

Moving on with data analysis, I looked for the best methods for binary classification and I identified three supervised machine learning methods: `GaussianNB`, `SGDClassifier` and `LinearSVC`. `GaussianNB` yielded a mean accuracy of `0.7725` with 5-fold cross validation. `SGDClassifier` had an accuracy of `0.8399` as `LinearSVC` had `0.8469`.

These accuracies were satisfactory, but the computation of `LinearSVC` model, which gave the best result, took a very long time. Moreover, these methods are black box models and it gave no space for interpreting the classification results. Thus, I looked for white box models and I decided to implement `DecisionTreeClassifier`, which allowed a ruleset export to identify the factors that affected people's voting tendencies. I applied hyperparameter tuning by setting `max_depth` attribute of `DecisionTreeClassifier` to [3, 10], and `max_depth=4` gave the optimum accuracy. After a 5-fold cross validation, I received a mean accuracy of `0.8605`, which turned out to be the best result I have received. In the following page, you can see the decision tree ruleset and my findings:

```
|--- age <= 29.50
|   |--- age <= 27.50
|   |   |--- EDUCATION <= 4.50
|   |   |   |--- age <= 26.50
|   |   |   |   |--- class: 0
|   |   |   |--- age >  26.50
|   |   |   |   |--- class: 0
|   |   |--- EDUCATION >  4.50
|   |   |   |--- RELIGIOUS SERVICES ATTENDANCE <= 3.50
|   |   |   |   |--- class: 0
|   |   |   |--- RELIGIOUS SERVICES ATTENDANCE >  3.50
|   |   |   |   |--- class: 0
|   |--- age >  27.50
|   |   |--- EDUCATION <= 4.50
|   |   |   |--- CURRENT EMPLOYMENT STATUS_1 <= 0.50
|   |   |   |   |--- class: 0
|   |   |   |--- CURRENT EMPLOYMENT STATUS_1 >  0.50
|   |   |   |   |--- class: 1
|   |   |--- EDUCATION >  4.50
|   |   |   |--- age <= 28.50
|   |   |   |   |--- class: 1
|   |   |   |--- age >  28.50
|   |   |   |   |--- class: 1
|--- age >  29.50
|   |--- RELIGIOUS SERVICES ATTENDANCE <= 1.50
|   |   |--- age <= 42.50
|   |   |   |--- age <= 30.50
|   |   |   |   |--- class: 0
|   |   |   |--- age >  30.50
|   |   |   |   |--- class: 1
|   |   |--- age >  42.50
|   |   |   |--- MARITAL STATUS_1 <= 0.50
|   |   |   |   |--- class: 1
|   |   |   |--- MARITAL STATUS_1 >  0.50
|   |   |   |   |--- class: 1
|   |--- RELIGIOUS SERVICES ATTENDANCE >  1.50
|   |   |--- age <= 37.50
|   |   |   |--- EDUCATION <= 6.50
|   |   |   |   |--- class: 1
|   |   |   |--- EDUCATION >  6.50
|   |   |   |   |--- class: 1
|   |   |--- age >  37.50
|   |   |   |--- MARITAL STATUS_1 <= 0.50
|   |   |   |   |--- class: 1
|   |   |   |--- MARITAL STATUS_1 >  0.50
|   |   |   |   |--- class: 1
```

People who are less than 27.5 years old don't vote.

People who are older than 27.5 with lower education levels vote if they have a full time job. Else, they don't vote.

People who are older than 27.5 with higher education levels vote.

30 y/o people who don't really attend religious services don't vote.

People older than 30.5 and younger than 42.5 who don't really attend religious services vote.

People older than 42.5 who don't really attend religious services vote.

People older than 29.5 who attend religious services vote.