Random Forest model on Red Wine Data

Pălăcian Ioana Teodora

Scope

- Build a model that will predict the quality of the wine based on its other features
- Use 1000 examples from the data set
- Achieve an accuracy of >60%

Data preprocessing

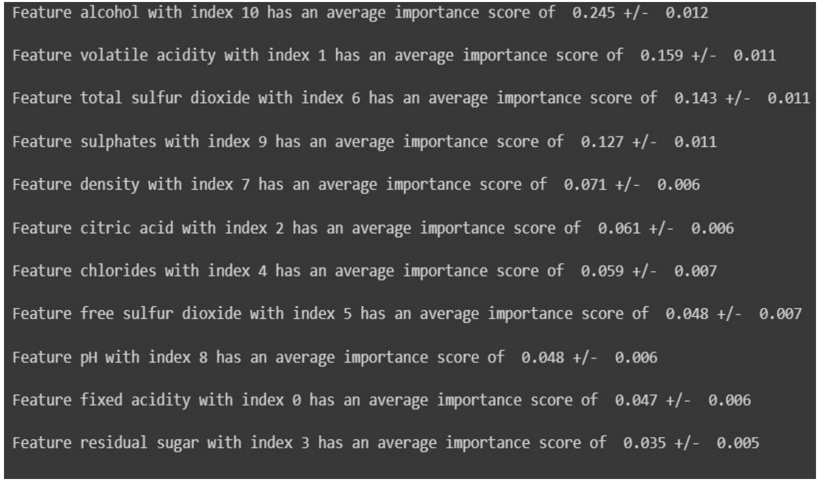
- Y = ‘quality’ feature of the data set
- X = the rest of the features used to predict the quality
- Data was split 30% for test and 70% for train
- Data was normalised using sklearn Standard Scaler which was fitted on the train data

Modeling

- Trained a Random forest model on train data
- Initial mean accuracy score: `0.6633333333333333`
    - Attempt to improve by computing feature importance and re-training the model using only the top features

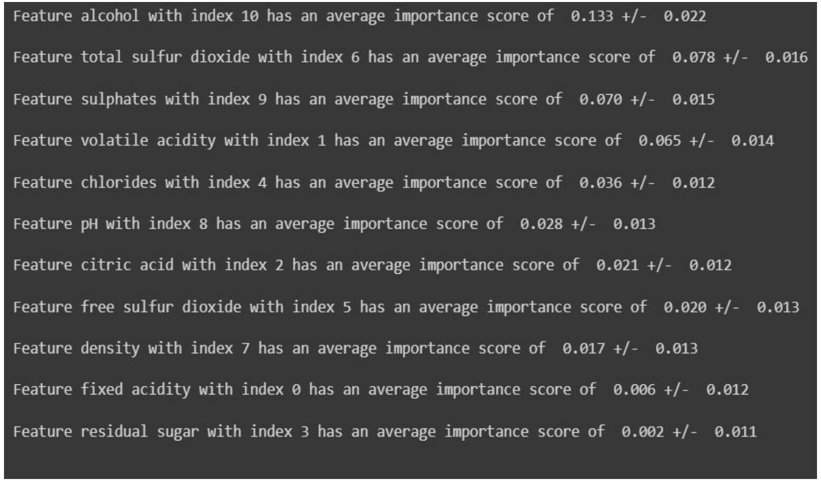
Modeling

Feature importance results on train data



```
Feature alcohol with index 10 has an average importance score of  0.245 +/-  0.012

Feature volatile acidity with index 1 has an average importance score of  0.159 +/-  0.011

Feature total sulfur dioxide with index 6 has an average importance score of  0.143 +/-  0.011

Feature sulphates with index 9 has an average importance score of  0.127 +/-  0.011

Feature density with index 7 has an average importance score of  0.071 +/-  0.006

Feature citric acid with index 2 has an average importance score of  0.061 +/-  0.006

Feature chlorides with index 4 has an average importance score of  0.059 +/-  0.007

Feature free sulfur dioxide with index 5 has an average importance score of  0.048 +/-  0.007

Feature pH with index 8 has an average importance score of  0.048 +/-  0.006

Feature fixed acidity with index 0 has an average importance score of  0.047 +/-  0.006

Feature residual sugar with index 3 has an average importance score of  0.035 +/-  0.005
```

Feature importance results on test data



```
Feature alcohol with index 10 has an average importance score of  0.133 +/-  0.022

Feature total sulfur dioxide with index 6 has an average importance score of  0.078 +/-  0.016

Feature sulphates with index 9 has an average importance score of  0.070 +/-  0.015

Feature volatile acidity with index 1 has an average importance score of  0.065 +/-  0.014

Feature chlorides with index 4 has an average importance score of  0.036 +/-  0.012

Feature pH with index 8 has an average importance score of  0.028 +/-  0.013

Feature citric acid with index 2 has an average importance score of  0.021 +/-  0.012

Feature free sulfur dioxide with index 5 has an average importance score of  0.020 +/-  0.013

Feature density with index 7 has an average importance score of  0.017 +/-  0.013

Feature fixed acidity with index 0 has an average importance score of  0.006 +/-  0.012

Feature residual sugar with index 3 has an average importance score of  0.002 +/-  0.011
```

Top 4 features are the same in both data sets, though they differ in order of importance

Modeling

Model re-trained with the first 4 most important features:

- Mean accuracy: `0.6533333333333333`
  - Less than the original model

Model re-trained with the first 3 most important features:

- Mean accuracy: `0.6633333333333333`
  - Equal to the original model - the highest so far

Modeling

Model re-trained with the first 2 most important features:

- Mean accuracy: `0.6`
  - Less than the original model

➔ Best models are the original random forest (with all of the features) and the third model (using only the top 3 features), with an accuracy of 66.3%