# **Decision Tree**

## **Classify based on “type” of wine**

**Without PCA**

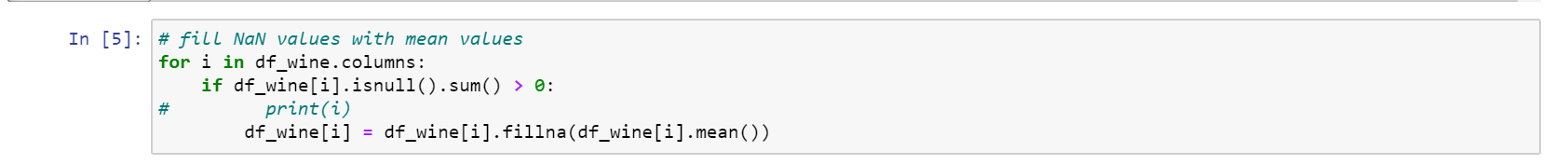
Data before data engineering

First, import the required libraries(pandas, sklearn, …) and read the csv file of dataset.

A screenshot of a cell phone

Description automatically generated

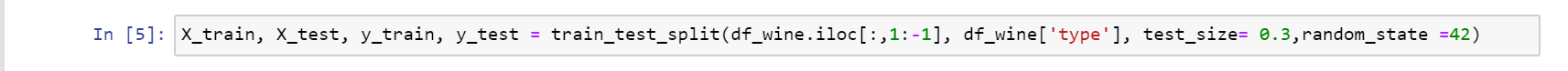
Data after data engineering



Replace the null values in dataset with mean value.

**Model train/test split**

Take ‘fixed acidity’, ’volatile acidity’, ’citric acid’, ’residual sugar’, ’chlorides’, ’free sulfur dioxide’, ’total sulfur dioxide’, ’density’, ’pH’, ’sulphates’, ’alcohol’ (col 1:11) as the independent variables to predict the wine type which is the dependent variable. Then, split the dataset for training and testing as per (70%-30%) ratio.



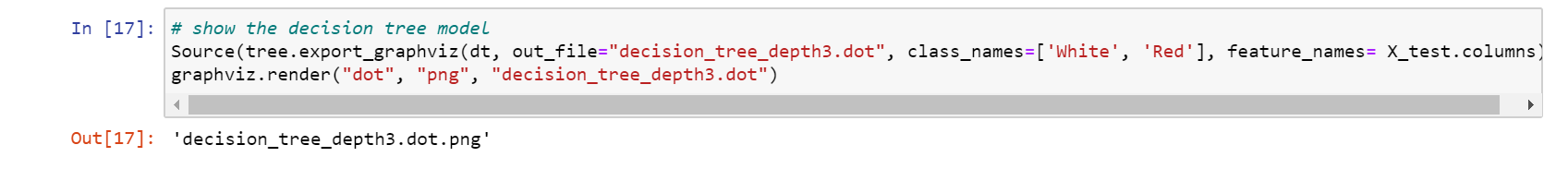
Instantiate a decision tree model

We built and train the model with decision tree classifier and got the accuracy score of 94.92% with depth = 3.

A screenshot of a social media post

Description automatically generated

**Visualize the Decision tree**



A picture containing receipt

Description automatically generated

**Predict data**

**A screenshot of a cell phone

Description automatically generated**

Duration of training



The confusion matrix for the model is as below.

A screenshot of a cell phone

Description automatically generated

**With PCA**

Feature extraction with PCA

A screenshot of a social media post

Description automatically generated

In this training dataset, we performed some feature extraction with n\_components = 5 on PCA, and then do the training.

A screenshot of a social media post

Description automatically generated

With the modified features by PCA, build a decision tree model and train the model. The accuracy score resulted is 0.982, which is better than training the model without PCA.

However, the duration of training this model with PCA(as per below) is not much different from without PCA.

A picture containing screenshot

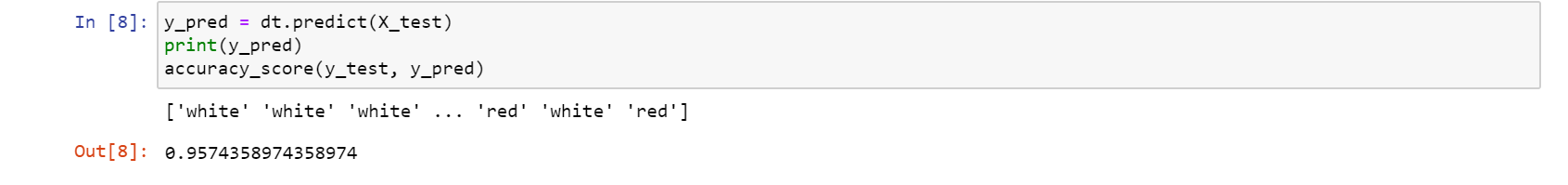
Description automatically generated

**Drop columns/features**

As per correlation matrix, ‘free sulfur dioxide’ and ‘total sulfur dioxide’ correlation is 0.72, which can be considered as high. So, we try drop one of them and drop ‘free sulfur dioxide’ after reading csv file. The result of removing this column is shown below.

A screenshot of a cell phone

Description automatically generated

****

The accuracy after dropping ‘free sulfur dioxide’ is slightly higher than including it in dataframe( 94.9 % to 95.7%).

Another correlation score around 0.5 are ‘residual sugar’ with ‘total sulfur dioxide’ and ‘density’. So, try dropping ‘free sulfur dioxide’ as well as ‘residual sugar’ and the result is shown in pictures.

A screenshot of a cell phone

Description automatically generated

A picture containing screenshot

Description automatically generated

However, the accuracy score of dropping these two columns is no difference with including both of them (94.9%).

**Impact of data/feature engineering**

As in this dataset, there are many features of which wine type is dependent, we tried with all features or with few features using PCA. From our observation, the accuracy score of the model with PCA application is slightly higher than the one without PCA, which resulted (98.2% and 94.9%) in each training.

As for another feature engineering, which is dropping certain columns with high correlation with each other, it is found out that by removing one of the two columns which have high correlation score(0.72) makes the accuracy score of training model higher than without removing it. However, when tried dropping the column which has correlation with another column( around 0.5) doesn’t make difference in related to accuracy score, remaining at 0.949.

**Other observations**

The accuracy score is getting higher as the value of max\_features in Decision tree classifier is higher. In this dataset, the wine type is depended on the tested independent variables and the more features made the accuracy score high.

Conclusion of decision tree for wine type

|  |  |  |
| --- | --- | --- |
| Model training | Accuracy | Duration |
| With PCA | 98.2% | 0.0203 |
| Without PCA | 94.9% | 0.0094 |