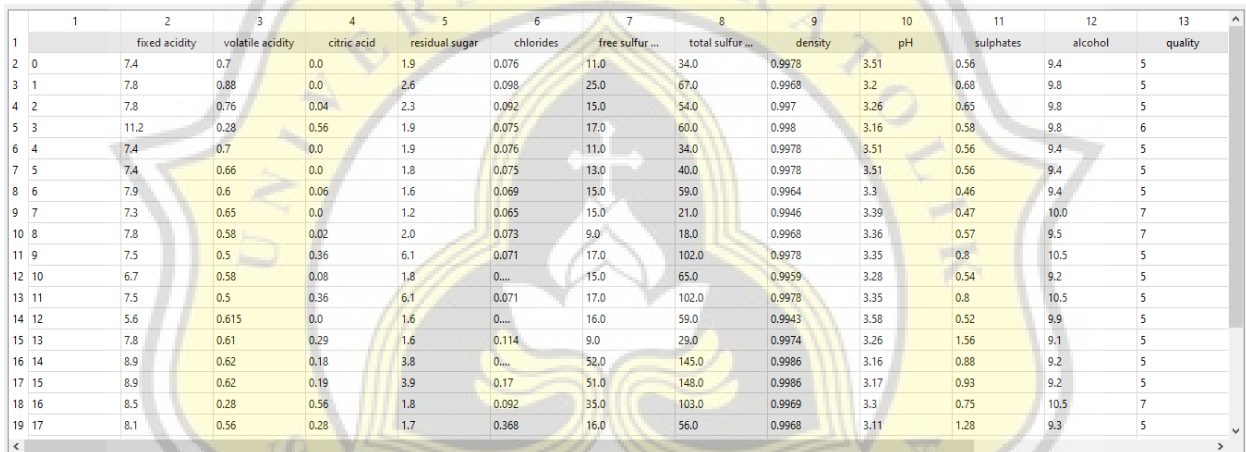


# CHAPTER 4

## ANALYSIS AND DESIGN

### 4.1. Collecting data

The dataset I used was taken from the site <https://www.kaggle.com/datasets/rajyellow46/wine-quality> This data is taken from a set of data from the Portuguese "Vinho Verde". With a total of 6000 data and 12 attributes in the dataset fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulfates, alcohol, and quality.



1	2	3	4	5	6	7	8	9	10	11	12	13
	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur ...	total sulfur ...	density	pH	sulphates	alcohol	quality
2 0	7.4	0.7	0.0	1.9	0.076	11.0	34.0	0.9978	3.51	0.56	9.4	5
3 1	7.8	0.88	0.0	2.6	0.098	25.0	67.0	0.9968	3.2	0.68	9.8	5
4 2	7.8	0.76	0.04	2.3	0.092	15.0	54.0	0.997	3.26	0.65	9.8	5
5 3	11.2	0.28	0.56	1.9	0.075	17.0	60.0	0.998	3.16	0.58	9.8	6
6 4	7.4	0.7	0.0	1.9	0.076	11.0	34.0	0.9978	3.51	0.56	9.4	5
7 5	7.4	0.66	0.0	1.8	0.075	13.0	40.0	0.9978	3.51	0.56	9.4	5
8 6	7.9	0.6	0.06	1.6	0.069	15.0	59.0	0.9964	3.3	0.46	9.4	5
9 7	7.3	0.65	0.0	1.2	0.065	15.0	21.0	0.9946	3.39	0.47	10.0	7
10 8	7.8	0.58	0.02	2.0	0.073	9.0	18.0	0.9968	3.36	0.57	9.5	7
11 9	7.5	0.5	0.36	6.1	0.071	17.0	102.0	0.9978	3.35	0.8	10.5	5
12 10	6.7	0.58	0.08	1.8	0.0...	15.0	65.0	0.9959	3.28	0.54	9.2	5
13 11	7.5	0.5	0.36	6.1	0.071	17.0	102.0	0.9978	3.35	0.8	10.5	5
14 12	5.6	0.615	0.0	1.6	0.0...	16.0	59.0	0.9943	3.58	0.52	9.9	5
15 13	7.8	0.61	0.29	1.6	0.114	9.0	29.0	0.9974	3.26	1.56	9.1	5
16 14	8.9	0.62	0.18	3.8	0.0...	52.0	145.0	0.9986	3.16	0.88	9.2	5
17 15	8.9	0.62	0.19	3.9	0.17	51.0	148.0	0.9986	3.17	0.93	9.2	5
18 16	8.5	0.28	0.56	1.8	0.092	35.0	103.0	0.9969	3.3	0.75	10.5	7
19 17	8.1	0.56	0.28	1.7	0.368	16.0	56.0	0.9968	3.11	1.28	9.3	5

Figure 4.1 Data sample

### 4.2. AdaBoost

AdaBoost is an ensemble algorithm that utilizes bagging and boosting to develop improved predictor accuracy. AdaBoost builds a stumps forest. A stump is a tree made up of only one branch and 2 leaves. Stumps that have the largest error have little influence when making decisions.

- Add Dataset

In this first step, open the orange data mining application, then import the CSV data that has been downloaded from the kaggle.com site. Then import data with the CSV file format

- Data Splitting

This step has a relationship with the data we imported earlier, the data we imported earlier was 6000 data and all the data will be divided using the data sampler widget using the orange data mining application. If we use fixed proportion data of 70%, the sample data has 4200 data and 1800 data becomes the remaining data or testing data.

- Inputting Model/Algorithm

This step uses tools in the orange data mining application, this time using the AdaBoost model to get predictions with parameters, base estimator tree, number of estimators 50, and learning rate 1.00000. The best output on this algorithm is using the default settings on this widget.

- Prediction

Then from the use of orange data mining is to make predictions and the prediction output is MSE, RMSE, MAE, R2.

$$MSE = \frac{1}{n} \sum_{i=0}^n (target - prediction)^2 \quad (1)$$

$$RMSE = \sqrt{MSE} \quad (2)$$

$$MAE = \frac{1}{n} \sum_{i=0}^n |target - prediction| \quad (3)$$

Indeks	explanation
MSE	Mean Square Error
N	Total Data
RMSE	Root mean Square error
MAE	Mean absolute error
$\Sigma$	Summation

### 4.3. Random Forest

The random forest method is one of the decision tree methods. A decision tree is a tree-like flowchart that has a root node used to collect data. A decision tree classifies a data sample that has no known class. A random forest is a combination of each good tree then combined into one model. The random forest also relies on a random vector value with the same distribution in all trees each decision tree has a maximum depth.

- Add Dataset

The first step is the same as AdaBoost by importing a CSV file using the widget in orange data mining. And data obtained from kaggle.com.

- Data Splitting

After adding a dataset in orange data mining that can divide data into training data and testing data. For example, 6000 data are using fixed 70% making the training data into 4200 training data and 1800 testing data.

- Inputting model/Algoritma

In this step use tools in the orange data mining application and the name of the tool used is a random forest by configuring the number of trees, training replications, and also balancing the class distribution. This step aims to ensure good accuracy and can produce good results.

- Prediction

Then from the use of orange data mining is to make predictions and the prediction output is MSE, RMSE, MAE, R2.

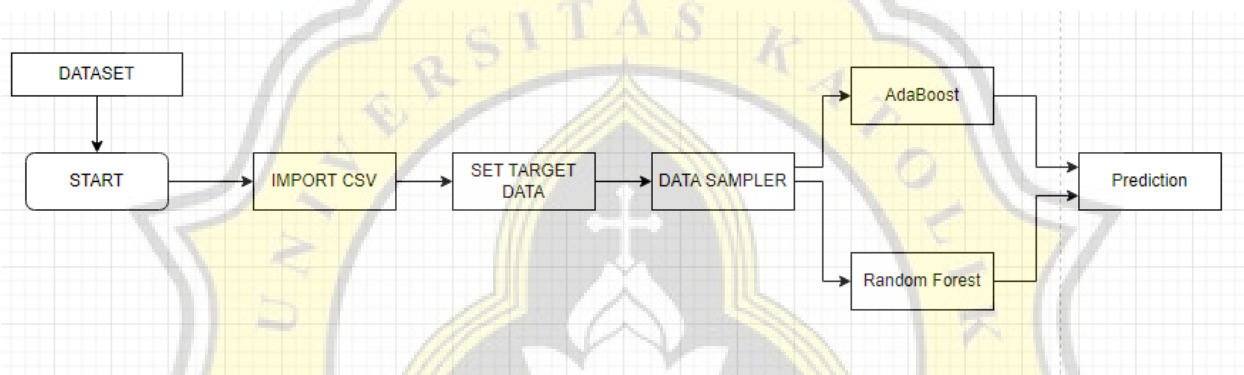
$$MSE = \frac{1}{n} \sum_{i=0}^n (target - prediction)^2 \quad (1)$$

$$RMSE = \sqrt{MSE} \quad (2)$$

$$MAE = \frac{1}{n} \sum_{i=0}^n |target - prediction| \quad (3)$$

Indeks	explanation
MSE	Mean Square Error
N	Total Data
RMSE	Root mean Square error
MAE	Mean absolute error
$\Sigma$	Summation

#### 4.4. Desain



**Figure 4.2** Flowchart

The flowchart above explains the work steps in the Orange data mining application. The first thing to do is to enter the dataset on the CSV import tab, then determine the target data that will be used in this project. Then enter the data sampler tab section then the data is processed using the AdaBoost and Random Forest methods. After the data is processed, the last step is to find out the results of the data using the prediction tab.

#### 4.5. Result

After conducting experiments in the Orange data mining application using two algorithms and getting data results, then getting prediction output results, namely MSE, MAE, RMSE, and R2. Both algorithms have their advantages and disadvantages in terms of determining the error value and accuracy level.