

CHAPTER 4

ANALYSIS AND DESIGN

4.1. Analysis

In this explanation, the procedure in the Orange program is executed using a dataset in the form of.csv. This dataset offers a list of caffeine-containing beverages in general. For example, bottled beverages in stores specify the quantity of volume (ml), caffeine(mg), and calories levels. This sample dataset includes samples from many canned beverage brands sold in coffee shops and supermarkets around Europe. This dataset can be found on the website "kaggle.com". On the website there are several .csv files that can be used for processing in the classification process and can also be processed in the Orange program. In Figure 4.1.1 there is an image of the caffeine.csv dataset table which consists of several types of categories, as follows:

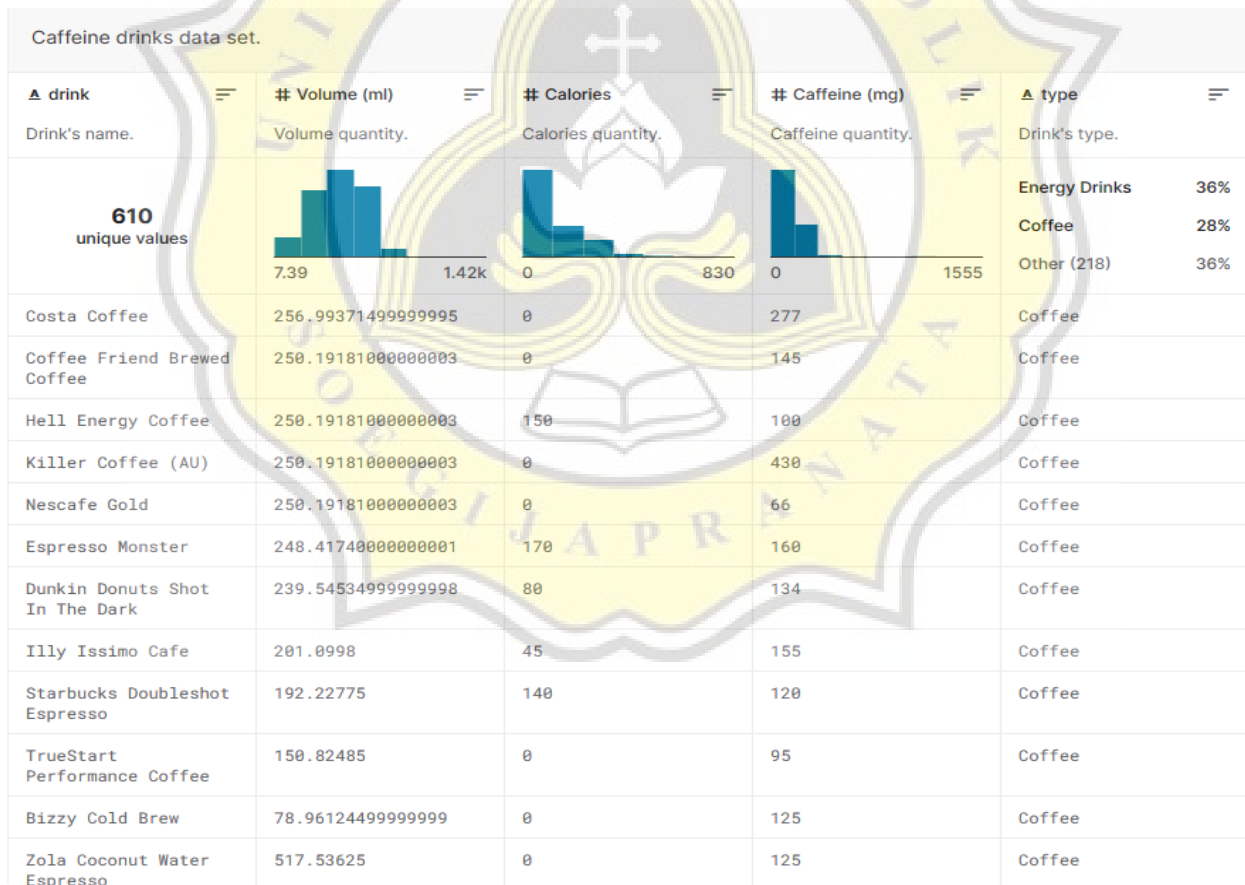


Figure 4.1. 1 Dataset Caffeine.csv

The table dataset is divided into various sections, including Drink, Volume (ml), Calories, Caffeine (mg), and Type. The Drink contains the drink's name. The components of the beverage product are then contained in Volume(ml). Calories is the calorie content of the beverage. Caffeine(mg) content is also present in some beverages. And Type is divided into six categories: coffee, energy drinks, energy shots, soft drinks, tea, and water.

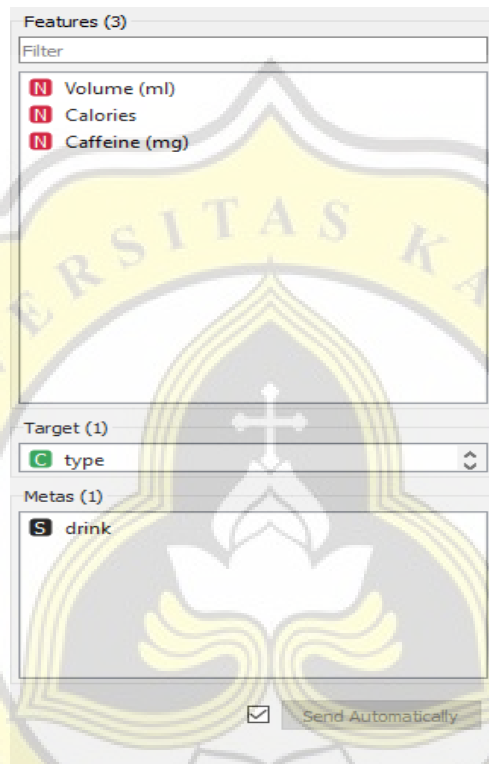


Figure 4.1. 2 Dataset selection of data attributes and composition of data domain

In figure 4.1.2 features explain about categorization to distinguish between regular characteristics, class attributes, and meta attributes in this character dataset. The approach targets the Type property using the Volume (ml), Calories, and Caffeine (mg) characteristics before classifying on the Drink attribute. To generate the target value for the type, Orange performs calculations on the Volume(ml), calories, caffeine(mg) attribute section and then calculates it using the algorithm used.

This dataset procedure will then be handled for additional examination. This project's analysis takes the form of computing each algorithm's accuracy, precision, recall, and f1-score, which will demonstrate how well it performed. The computations mentioned above are meant to evaluate how well each algorithm in the devised calculating system performs. Likewise, contrast how well the two algorithms performed. Then the analysis is processed using Orange Data Mining with the Confusion Matrix method to show the results of the data comparison. To find out the performance of the two algorithms, Figures 4.1.3 and 4.1.4 are described as follows:

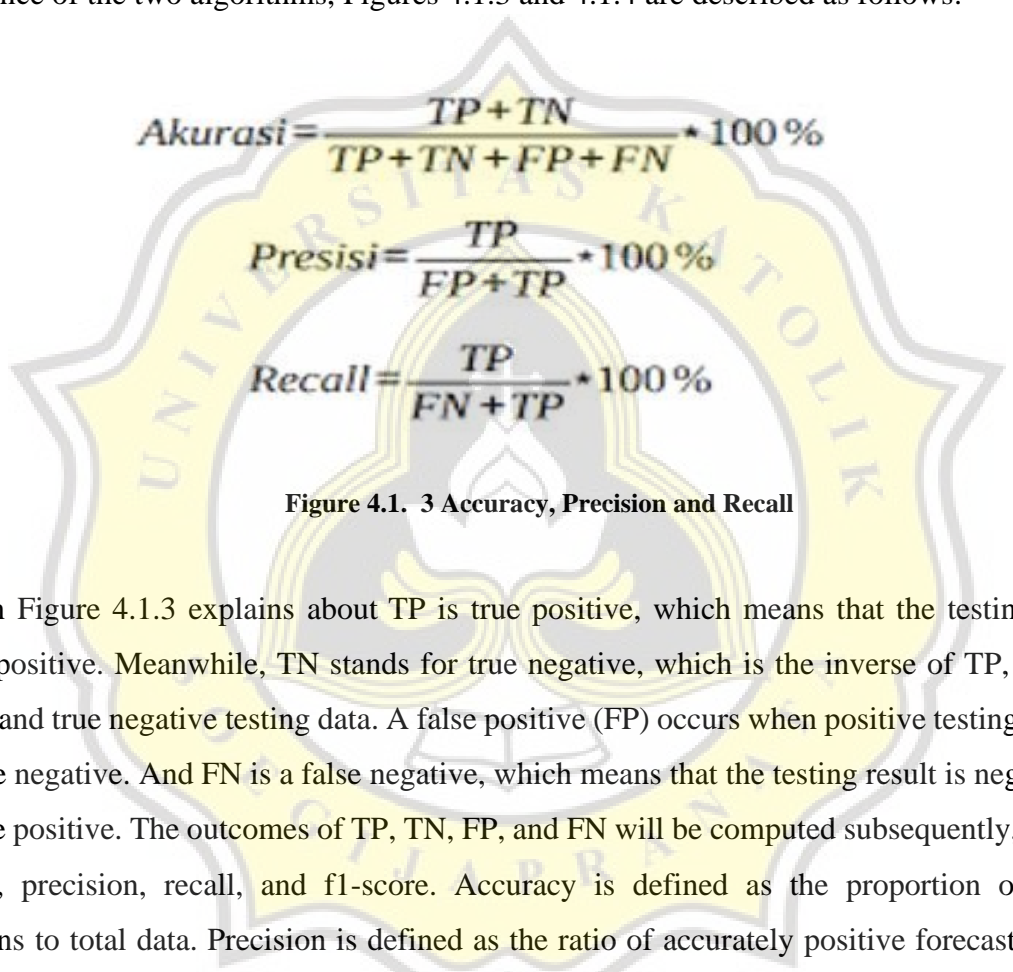

$$\text{Akurasi} = \frac{TP + TN}{TP + TN + FP + FN} * 100\%$$
$$\text{Presisi} = \frac{TP}{FP + TP} * 100\%$$
$$\text{Recall} = \frac{TP}{FN + TP} * 100\%$$

Figure 4.1. 3 Accuracy, Precision and Recall

In Figure 4.1.3 explains about TP is true positive, which means that the testing data is actually positive. Meanwhile, TN stands for true negative, which is the inverse of TP, meaning negative and true negative testing data. A false positive (FP) occurs when positive testing findings should be negative. And FN is a false negative, which means that the testing result is negative but should be positive. The outcomes of TP, TN, FP, and FN will be computed subsequently, yielding accuracy, precision, recall, and f1-score. Accuracy is defined as the proportion of correct predictions to total data. Precision is defined as the ratio of accurately positive forecasts to total positive predictions. Recall is the ratio of accurately positive forecasts to all correctly positive predictions.

$$F1 = 2. \frac{Precision \times Recall}{Precision + Recall}$$

Figure 4.1. 4 f1-score

Figure 4.1.4 explains that the F-1 score describes the comparison of the weighted average precision and recall. But if Accuracy is right, we can use it as a reference for algorithm performance if our dataset has a very close (symmetric) number of False Negatives and False Positives. But if the numbers are not close, then use the F1 Score as a reference.

4.2. K-Nearest Neighbor Analysis (K-NN)

The k-NN algorithm is a data mining method for classifying objects based on learning data with the closest distance to the object. The principle of k-NN is to find the shortest distance between what will be evaluated and the *k* closest neighbors in the training data. To determine the results of the predicted and actual values using the confusion matrix method. The following are some of the results obtained by K-NN in the confusion method matrix.

		Predicted						Σ
		Coffee	Energy Drinks	Energy Shots	Soft Drinks	Tea	Water	
Actual	Coffee	78	73	2	1	17	0	171
	Energy Drinks	56	149	0	5	6	7	223
	Energy Shots	13	0	23	0	0	0	36
	Soft Drinks	0	7	0	80	1	3	91
	Tea	13	16	0	6	28	0	63
	Water	1	8	0	11	3	3	26
Σ		161	253	25	103	55	13	610

Figure 4.2 1 Predicted K-NN Confusion Matrix

Figure 4.2.1 describes predictions with the KNN algorithm which are then displayed in the confusion matrix.. The reason for the many drink categories in this graphic is to learn the truth about what is processed using the confusion matrix technique. Then of the 223 types of Energy drink data which shows correct as much as 149 data then the rest of the data is considered wrong. Then from the 36 types of Energy Shots data that shows 23 data is correct then the rest of the data is considered wrong. Then from the 91 types of Soft Drinks data which shows correct as much as 80 data then the rest of the data is considered wrong. Then of the 63 types of Tea data which shows correct as much as 28 data then the rest are considered wrong. And of the 26 kinds of Water data that shows only 3 data are correct then the rest are wrong.

4.3. Support Vector Machine Analysis (SVM)

A system of algorithms called a support vector machine may be applied to both classification and regression. Based on structural risk reduction, SVM is able to divide the input space into two classes by processing data into a Hyperlane. The foundation of SVM theory is a collection of linear examples that may be classified into classes and separated by hyperlanes. Since the SVM idea turns into a two-class classification issue, positive and negative training sets are needed. To increase the margins of the two classes, SVM will attempt to divide the two as much as feasible. The linear approach and the kernel technique are the two methods used for classification.

This method uses the same as K-NN, namely the confusion matrix because the similarity in use must be with the dataset. The following is an example of data in the confusion matrix in SVM processing.

		Predicted						Σ
		Coffee	Energy Drinks	Energy Shots	Soft Drinks	Tea	Water	
Actual	Coffee	75	76	1	3	16	0	171
	Energy Drinks	28	179	0	5	11	0	223
	Energy Shots	9	0	27	0	0	0	36
	Soft Drinks	1	5	0	84	1	0	91
	Tea	9	19	0	10	25	0	63
	Water	2	9	0	13	2	0	26
	Σ	124	288	28	115	55	0	610

Figure 4.3 1 Predicted SVM Confusion Matrix

Figure 4.3.1 describes predictions with the SVM algorithm which are then displayed in the confusion matrix. The Confusion Matrix table explains if the information in the category table is accurate. For instance, the Coffee data comprises 171 data, of which 75 data demonstrate the accuracy of the data, while the other data is regarded as being incorrect. Energy Drinks then contains 223 data, 179 of which demonstrate the accuracy of the data, while the other data is regarded as being incorrect. Energy Shots then contains 36 data, of which 27 data demonstrate the accuracy of the data, while the other data is regarded as being incorrect. The next category, Soft Drinks, includes 91 data, 84 of which demonstrate the accuracy of the data; the other data is regarded as incorrect. Tea then contains 63 data, of which 25 data demonstrate the accuracy of the data, while the other data is regarded as incorrect. And finally, when the data is being tested in the water data section, it shows that there is no data in the water category.

4.4. Desain

The design of this approach for processing data is essentially the same. The main difference between the K-NN and SVM algorithms, though, is how they are calculated. The Orange design, shown as a flowchart, is seen below.

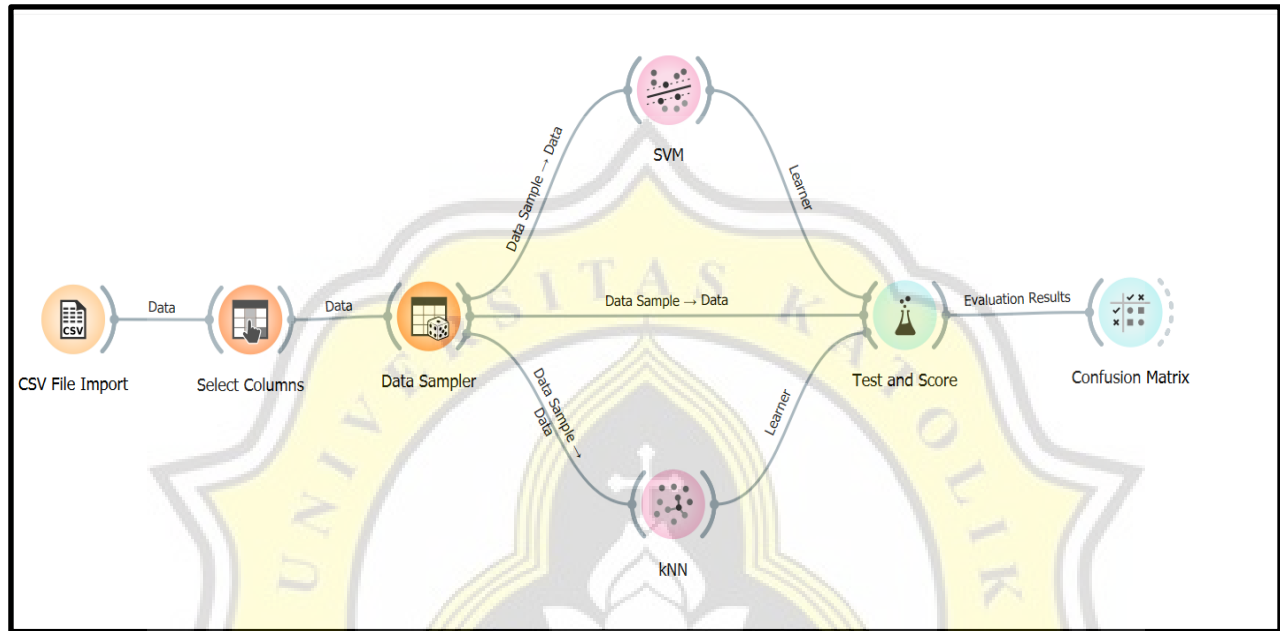


Figure 4.4. 1 Flowchart Orange Data Mining

In Figure 4.4.1 explains about the flowcharts for these two algorithms are identical. beginning with inserting data in a dataset using the.csv format. If you want to focus on a certain category, enter "Select Columns" next. The kind of algorithm is then entered, followed by the amount of data to be processed. K-NN and SVM are to be employed as algorithms. The findings and predictions are automatically shown in the "Test and Score" section after the data has been analyzed by the algorithm. The Confusion Matrix is used to display the accurate and wrong data after the prediction results have been assessed.