CHAPTER 4
ANALYSIS AND DESIGN

4.1. Analysis

The data was collected from public data *Wisconsin Breast Cancer Database UCI Repository Machine Learning* (*http://archive.ics. uci.edu/ml/machine-learning-databases/breast-cancer-wisconsin/*). The following is an image snippet of the raw data that will be used. Before the dataset is processed it will look like this.

```
1000025,5,1,1,1,2,1,3,1,1,2
1002945,5,4,4,5,7,10,3,2,1,2
1015425,3,1,1,1,2,2,3,1,1,2
1016277,6,8,8,1,3,4,3,7,1,2
1017023,4,1,1,3,2,1,3,1,1,2
1017122,8,10,10,8,7,10,9,7,1,4
1018099,1,1,1,1,2,10,3,1,1,2
1018561,2,1,2,1,2,1,3,1,1,2
1033078,2,1,1,1,2,1,1,1,5,2
1033078,4,2,1,1,2,1,2,1,1,2
1035283,1,1,1,1,1,1,3,1,1,2
1036172,2,1,1,1,2,1,2,1,1,2
1041801,5,3,3,3,2,3,4,4,1,4
1043999,1,1,1,1,2,3,3,1,1,2
1044572,8,7,5,10,7,9,5,5,4,4
1047630,7,4,6,4,6,1,4,3,1,4
1048672,4,1,1,1,2,1,2,1,1,2
1049815,4,1,1,1,2,1,3,1,1,2
1050670,10,7,7,6,4,10,4,1,2,4
1050718,6,1,1,1,2,1,3,1,1,2
1054590,7,3,2,10,5,10,5,4,4,4
1054593,10,5,5,3,6,7,7,10,1,4
1056784,3,1,1,1,2,1,2,1,1,2
1057013,8,4,5,1,2,?,7,3,1,4
1059552,1,1,1,1,2,1,3,1,1,2
1065726,5,2,3,4,2,7,3,6,1,4
```

Figure 4.1 Example of Original Dataset Snippet

Starting from the left, the attributes in the database consist of: *Sample code number, clump thickness, uniformity cell size, uniformity cell shape, marginal adhesion, single epithelial cell size, bare nucleoli, bland chromatin, normal nucleoli, mitoses,* and *class*.

To simplify data processing, we need to extract the .DATA file as shown in *Illustration 4* into CSV. After the conversion, it will look like this.

	Sample_code_number	Clump_Thickness	Uniformity_of_Cell_Size	Uniformity_of_Cell_Shape	Marginal_Adhesion	Single_Epithelial_Cell_Size	Bare_Nuclei	Bl
0	1000025	5	1	1	1	2	1	
1	1002945	5	4	4	5	7	10	
2	1015425	3	1	1	1	2	2	
3	1016277	6	8	8	1	3	4	
4	1017023	4	1	1	3	2	1	
...	...	...	...	...	...	...	...	
694	776715	3	1	1	1	3	2	
695	841769	2	1	1	1	2	1	
696	888820	5	10	10	3	7	3	
697	897471	4	8	6	4	3	4	
698	897471	4	8	8	5	4	5	

699 rows × 11 columns

Figure 4.2 After Convertion Dataset

At this stage, pre processing is carried out on the data. First, we do data cleaning for the incomplete, empty or null data, as well as deleting attributes that are not used in the classification process. This process is carried out after the data selection process and will reduce the amount of data. The amount of data before cleaning was 699, 16 data deleted along with 1 attributes, resulting in 683 final data. The omitted attribute was the Sample_code_number.

Below are some of the results of the independent variables in the training data. The value on the independent variable is the cause of the value on the dependent variable. The independent variable in this study is all the attribute except sample_code_number and class. The dependent variable in this study is the "Class" column. After we define it, we can implement it into a train/test split function.

Table 4.1. Data Training

	Clump_Thickness	Uniformity_of_Cell_Size	Uniformity_of_Cell_Shape	Marginal_Adhesion	Single_Epithelial_Cell_Size	Bare_Nuclei	Bland_Chromatin	Normal_Nucleoli	Mitoses
293	10	4	4	6	2	10	2	3	1
651	1	2	1	3	2	1	2	1	1
687	3	1	1	1	2	1	2	3	1
659	1	1	1	1	2	1	1	1	1
197	5	1	1	4	2	1	3	1	1
370	4	3	2	1	3	1	2	1	1
573	1	1	1	1	2	1	2	1	1
215	8	7	8	7	5	5	5	10	2
527	4	1	1	1	2	1	3	1	1
627	2	1	1	1	2	5	1	1	1
622	7	1	2	3	2	1	2	1	1

593	5	1	2	1	2	1	1	1	1
129	1	1	1	1	10	1	1	1	1
95	1	1	1	1	2	1	3	1	1
76	1	1	4	1	2	1	2	1	1
34	3	1	2	1	2	1	2	1	1
193	1	1	1	1	2	1	3	1	1
525	3	1	1	2	2	1	1	1	1
394	1	2	3	1	2	1	2	1	1
358	8	10	5	3	8	4	4	10	3

Then, we calculate the probability of each class in the population in the training data. In this stage, the data is divided using the train test split function with the distribution of data: 205 for testing data and 478 for training data. Produces the probability calculation value as follows:

- **Benign Probability**

$$\frac{Benign}{(Benign+Malignant)} \tag{2}$$

$$= \frac{314}{478} = 0{,}656904$$

- **Malignant Probability**

$$\frac{Malignant}{(Benign+Malignant)} \tag{3}$$

$$= \frac{164}{478} = 0{,}343096$$

After getting the prior probability, next step is to calculate the mean of each attribute. For the example we calculate the mean of Clump Thickness :

- **Mean for Benign**

$$\mu = \frac{Total\ values\ of\ the\ Clump\ Thickness\ attribute\ in\ the\ Benign\ Class}{Total\ Benign\ Class} \tag{4}$$

$$\mu = \frac{Value\ 1 + Value\ 2 + \cdots + Value\ 315}{Total\ Benign\ Class}$$

$$\mu = \frac{4+5+\cdots+3}{314}$$

$$\text{Mean} = \frac{955}{314} = \mathbf{3{,}04140}$$

- **Mean for Malignant**

$$\mu = \frac{Total\ values\ of\ the\ Clump\ Thickness\ attribute\ in\ the\ Malignant\ Class}{Total\ Malignant\ Class} \quad (5)$$

$$\mu = \frac{Value\ 1 + Value\ 2 + \cdots + Value\ 315}{Total\ Malignant\ Class}$$

$$\mu = \frac{8+8+\cdots+10}{164}$$

$$\text{Mean} = \frac{1178}{164} = \mathbf{7{,}18292}$$

The mean value of each attribute can be seen in this table:

Table 4.2. Mean of Each Attribute

Clump Thickness		
	Benign	Malignant
Mean	3,04140	7,18292

Uniformity of Cell Size		
	Benign	Malignant
Mean	1,33439	6,64634

Uniformity of Cell Shape		
	Benign	Malignant
Mean	1,41401	6,53048

Marginal Adhesion		
	Benign	Malignant
Mean	1,35987	5,60975

Single Epithelial Cell Size		
	Benign	Malignant
Mean	2,07324	5,48780

Bare Nuclei		
	Benign	Malignant
Mean	1,30891	7,54268

Bland Chromatin		
	Benign	Malignant
Mean	2,09235	6,03658

Normal Nucleoli		
	Benign	Malignant
Mean	1,24840	6,07317

Mitosis		
	Benign	Malignant
Mean	1,07324	2,53048

4.2. Desain

. The following flow is used to carry out a data cleaning procedure using jupyter notebook tools as a medium for data processing.

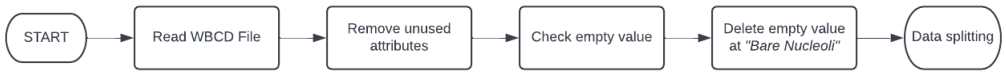

Figure 4.1 Flowchart of Data Cleaning

From the results of data cleaning, then the dependent and independent variables are determined. Then the data separation is done to get training data and test/validation data. Then, the model is evaluated by measuring accuracy using a confusion matrix.

Table 4.3. Data Testing

	Clump Thickness	Uniformity of Cell Size	Uniformity of Cell Shape	Marginal Adhesion	Single Epithelial Cell Size	Bare Nuclei	Bland Chromatin	Normal Nucleoli	Mitoses
465	10	9	8	7	6	4	7	10	3
633	8	7	4	4	5	3	5	10	1
513	3	1	1	1	1	1	2	1	1
677	5	1	1	1	2	1	1	1	1
67	5	3	4	1	8	10	4	9	1
270	8	4	7	1	3	10	3	9	2
356	5	3	3	1	3	3	3	3	3
340	10	3	3	1	2	10	7	6	1
156	1	2	2	1	2	1	2	1	1
550	3	1	1	1	2	1	2	1	1
38	5	4	4	9	2	10	5	6	1
182	6	1	1	1	2	1	3	1	1
557	5	1	1	3	2	1	1	1	1
509	2	1	1	1	2	1	1	1	1
655	3	1	1	1	2	1	2	1	1
17	4	1	1	1	2	1	3	1	1
146	3	4	5	2	6	8	4	1	1
674	1	1	1	1	2	1	2	1	1
230	7	4	7	4	3	7	7	6	1
468	4	1	1	1	2	1	1	1	1

Confusion matrix is described as a table with four different combinations of the actual value and the predicted value. From the results of the classification can be represented in the Confusion Matrix table, as True Positive, True Negative, False Positive, False Negative.

Table 4.4. Confusion Matrix Rules

	FALSE	TRUE
FALSE	True Negative (TN)	False Positif (FP)
TRUE	False Negative (FN)	True Positive(TP)

From the Table 4.2 it can explained that True Positive is positive data that is predicted to be true, True Negative is negative data that is predicted to be negative, False Positive is negative data that is predicted to be positive, False Negative is positive data that is predicted as negative data. From the confusion matrix table, we can use to measure the level of accuracy, precision, recall, and the value of the F1 score. Equation to calculate it as:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{6}$$

$$Precision = \frac{TP}{TP + FP} \tag{7}$$

$$Recall = \frac{TP}{TP + FP} \tag{8}$$

Accuracy is decribing how accurate the model is in classifying correctly. Precision describes the accuracy between the requested data and the prediction results. Recall / sensitivity describes the success of the model in retrieving information.