CHAPTER 3
RESEARCH METHODOLOGY

3.1 Literature Study

To start the comparison between two algorithms, the first step begins with collecting the data, the next step is to find the best algorithm to solve it with and without a library. In this case it uses the Sci-Kit Learn library to obtain the result, the other algorithm is using a program without a library. And the last part is to analyze the difference between the result of two different algorithms.

3.2 Collecting Data

The dataset contains nine columns and it was taken from kaggle's website (https://www.kaggle.com/uciml/pima-indians-diabetes-database). The data from kaggle is originally from the National Institute of Diabetes and Digestive and Kidney Diseases and consists of 768 records of patients.. That nine columns contains eight columns of variables from the patient and one column of the result whether the patient has diabetes or not. From here the data was used to train the machine to predict the other patient which has a bigger risk with diabetes. This Data type from kaggle is CSV format so it can be easily processed. The attributes of the dataset are shown on the table.

Table 1. *Attributes of the dataset*

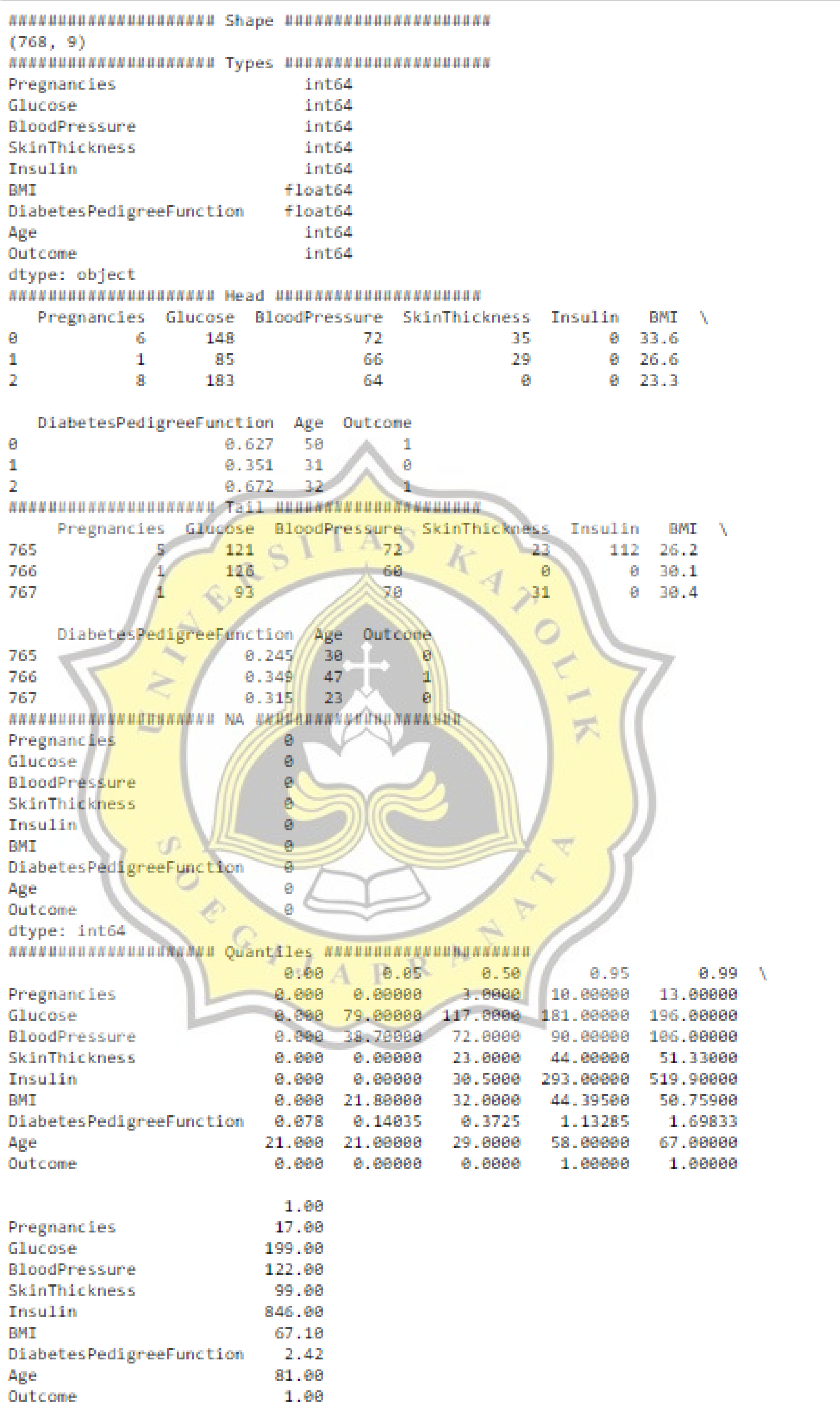
No.	**Attributes**	**Explanation**
1.	Pregnancies	Number of times pregnant
2.	Glucose	Plasma glucose concentration a 2 hours in an oral glucose tolerance test
3.	Blood Pressure	Diastolic blood pressure (mm Hg)
4.	Skin Thickness	Triceps skin fold thickness (mm)

5.	Insulin	2-Hour serum insulin (mu U/ml)
6.	BMI	Body mass index (weight in kg/(height in m)^2)
7.	Diabetes Pedigree	Diabetes pedigree function
8.	Age	Age (years)
9.	Outcome	Class variable (0 or 1) 268 of 768 are 1, the others are 0

3.3 Analysis

After the data has been collected, there are some steps that need to be done to make the machine easier to read and process the data. These steps include: data preparation, data cleaning, feature engineering, and evaluation. All of these steps will be explained in the next chapter. The main concern of this project is to compare the difference between two libraries and algorithms, what are the challenges if we don't use the library. In order to do that, the best method of each library must be found first, then the comparison will be efficient.

These are complete information of the data from the Kaggle, then shown in the program to check the data.



```
##################### Shape #####################
(768, 9)
##################### Types #####################
Pregnancies                   int64
Glucose                       int64
BloodPressure                 int64
SkinThickness                 int64
Insulin                       int64
BMI                         float64
DiabetesPedigreeFunction    float64
Age                           int64
Outcome                       int64
dtype: object
##################### Head #####################
   Pregnancies  Glucose  BloodPressure  SkinThickness  Insulin   BMI  \
0            6      148             72             35        0  33.6
1            1       85             66             29        0  26.6
2            8      183             64              0        0  23.3

   DiabetesPedigreeFunction  Age  Outcome
0                     0.627   50        1
1                     0.351   31        0
2                     0.672   32        1
##################### Tail #####################
     Pregnancies  Glucose  BloodPressure  SkinThickness  Insulin   BMI  \
765            5      121             72             23      112  26.2
766            1      126             60              0        0  30.1
767            1       93             70             31        0  30.4

     DiabetesPedigreeFunction  Age  Outcome
765                     0.245   30        0
766                     0.349   47        1
767                     0.315   23        0
##################### NA #####################
Pregnancies                 0
Glucose                     0
BloodPressure               0
SkinThickness               0
Insulin                     0
BMI                         0
DiabetesPedigreeFunction    0
Age                         0
Outcome                     0
dtype: int64
##################### Quantiles #####################
                            0.00      0.05      0.50       0.95       0.99  \
Pregnancies                0.000   0.00000    3.0000   10.00000   13.00000
Glucose                    0.000  79.00000  117.0000  181.00000  196.00000
BloodPressure              0.000  38.70000   72.0000   90.00000  106.00000
SkinThickness              0.000   0.00000   23.0000   44.00000   51.33000
Insulin                    0.000   0.00000   30.5000  293.00000  519.90000
BMI                        0.000  21.80000   32.0000   44.39500   50.75900
DiabetesPedigreeFunction   0.078   0.14035    0.3725    1.13285    1.69833
Age                       21.000  21.00000   29.0000   58.00000   67.00000
Outcome                    0.000   0.00000    0.0000    1.00000    1.00000

                            1.00
Pregnancies                17.00
Glucose                   199.00
BloodPressure             122.00
SkinThickness              99.00
Insulin                   846.00
BMI                        67.10
DiabetesPedigreeFunction    2.42
Age                        81.00
Outcome                     1.00
```

***Illustration 4.1.** Information about the Dataset*

3.4 Implementation and Testing

This step will discuss how the algorithm will be implemented in the programming code (python). This process is to implement the data to be trained in the SciKit library and Logistic Regression Algorithm. In this case the Logistic Regression and Naïve Bayes algorithm is used because the result needs to be returned in 0 or 1 to represent the answer where 0 equals “No” and 1 equals “Yes”. In the implementation, the data will also be tested to check the accuracy of each model. The output data will be:

Table 2. *Output variable of the result*

No	Output Variable	Diabetes
1	Healthy (“0”)	The person does not have diabetes
2	Sick (“1”)	The person has diabetes

3.5 Conclusion

The conclusion from this dataset is that machine learning could be greatly helpful to solve the medical issues with training the data to make a prediction. The conclusion will also be an explanation of what makes the difference between two libraries and algorithms, the analysis about which algorithm works better and why there are gaps between two libraries.