CHAPTER 4
ANALYSIS AND DESIGN

4.1. Design

The use of flowchart aims to know the processes or procedures of a program that makes it easier to understand the program to be built. Flowchart system can be described as here:

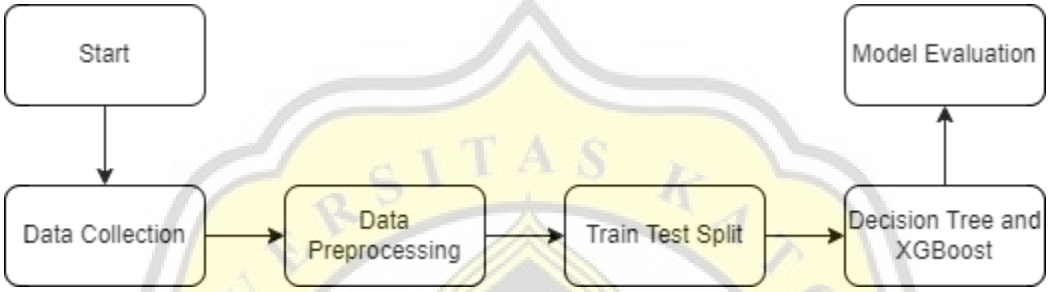

Figure 4.1 : Flowchart of Diabetes Prediction using Decision Tree and XGBoost

The Flowchart above is used as an overview or flow of the system itself. Starting from data collection, then the dataset will be processing in the data preprocessing part, such as : filling null value with median, balancing the class output using synthetic minority oversampling technique (SMOTE) and removing outlier using Z-Score. After data processing, the data will be divided with 4 categories : 90% train 10% test, 80% train and 20% test. 70% train 30% test, and 60% train and 40% test. The data will be trained and tested using Decision Tree and XGBoost Algorithm. Lastly, we have to evaluate the model using Accuracy, Precision, Recall and F-1 Score.

In general, Decision Tree and XGBoost Algorithm system are predicting the data whether the patient is diabetic or not. We have to fit the train data into the machine, so the machine can learn the data. Test data are used to evaluate the model that has been fitted with train data. The stages in the process are:

1. **Data Collection :** Collecting data from kaggle, the dataset named : Pima Indian Diabetes Database. The dataset contain 9 column and 768 rows, column Pregnancies to Age are attributes and Output is class whether the patient is

diabetic or not. All patients on the dataset are females at least 21 years old of Pima Indian heritage.

2. **Data Preprocessing :** Cleaning the dataset, such as : Balancing the output class, fill null value with median and remove the outlier using Z-Score
3. **Train Test Split :** Splitting the train and test data into 4 categories : 90% train 10% test, 80% train 20% test, 70% train 30% test, and 60% train 40% test
4. **Decision Tree and XGBoost :** Fitting the train data into Decision Tree and XGboost algorithm, In this process, the algorithms used play a role in maximizing the dataset and determining the accuracy level of the program itself.
5. **Model Evaluation :** Evaluate the Decision Tree and XGBoost model with Accuracy, Precision, Recall, and F-1 Score.