CHAPTER 3
RESEARCH METHODOLOGY

3.1. Data Collection

The dataset used for this project is taken from Kaggle, this dataset is in CSV 38.01 KB format consisting of 614 rows and 13 columns as factors namely loan id, gender, marital status, dependents, education, self-employed, applicant income, applicant income, total loan, loan term, credit history, property area, loan status. This dataset is divided into two ratios 70% for training, 30% for testing, and 60% for training, 40% for testing. Training datasets to train models and test datasets to test datasets. And each test data will be divided into three data to see the consistency of its accuracy, as well as testing with several parameters.

3.2. Algorithms

This project uses Logistic Regression and Extreme Gradient Boosting algorithms. Logistics Regression is used because it is efficient to train, does not have assumptions so there is no need to test assumptions, has high accuracy results, and the algorithm can be used to predict decisions. XGBoost or called Extreme Gradient Boosting is one of the powerful algorithms which has steps with trimming to increase the generalizability of the model, newton boosting which works to prevent gradient descent, and extra randomization parameters to reduce tree correlation increasing the strength of the algorithm ensemble.

3.3. Design

The dataset is collected and preprocessed, the data will be cleaned by filling in the missing variables and handling noise, then data transformation is carried out so that the dataset becomes more effective and efficient. After splitting the dataset, the model uses Logistic Regression, and Extreme Gradient Boosting, then the results are used to test the data.

3.4. Coding

Here, this project uses Python 3.6.9 as the programming language. In addition to its popularity, Python is used because this programming language is an interpreted language so that it is easy to understand, has a simple syntax, Python has an extensive standard library, and is open source. And I use google colaboratory because google colaboratory uses the cloud to execute code so that the performance of the local machine will not decrease in code execution. This project uses pandas library to read datasets and numpy which is useful when doing calculations.

3.5. Analysis

In this project, the data taken from filling out online application forms from companies that want to automate the credit application process are 615 rows and 13 columns as parameters. Furthermore, the data will be broken down into training datasets and test datasets, which later the model will implement the Logistic Regression and Extreme Gradient Boosting algorithms to analyze which algorithm has the highest accuracy, then whether the two algorithms are suitable to be applied in this case or not, and which variables affect the outcome decision that a loan is rejected and accepted.