CHAPTER 1
INTRODUCTION

1.1 Background

Diabetes is a chronic disease characterized by high blood sugar (glucose) levels that occurs as the result of the pancreas not producing enough insulin or the body cannot effectively use the insulin. According to WHO international in 2014, 8.5% of adults aged 18 years and older had diabetes. In 2016, diabetes was the direct cause of 1.6 million deaths and in 2012 high blood glucose was the cause of another 2.2 million deaths. It is a very serious problem that needs to be solved, the cure has not been found yet, but we still could prevent it by diagnosing people who have diabetes disease. If people positively have a diabetes disease, they can quickly prevent it from a young age by watching their diet. So, it is an important thing to predict the patient who has diabetes disease.

Nowadays, this problem could be solved using machine learning by training the dataset of patients with and without the symptoms of diabetes. In this case, the dataset will be trained with two different algorithms to predict the onset of diabetes disease, one using the SciKit-Learn library and the other using the TensorFlow library. The algorithm that will be used is logistic regression and Naïve Bayes classifier, as the result of the output is a probability between 0 and 1, with the sum of one whether the patient has diabetes disease or not, based on certain diagnostic measurements included in the dataset. The datasets itself are saved as CSV or comma-separated values, which consist of many medical predictor variables such as BMI, insulin level, age, and so on. With this kind of data structure, it is easy for the IDE compiler to read the data, just need to clean the invalid data inside.

There are also many ways to solve this problem, but this project will find the best way to solve the problem both with and without a library. It will find out how a library can be so effective in solving the problem, since this problem can be

solved even without using a library, but a simple logistic regression code.The conclusion of this project is to predict the onset of diabetes based on diagnostic measures using machine learning and logistic regression algorithms and analyzing the differences of Logistic Regression code and Naïve Bayes from SciKit-Learn Library. From that we can improve health and education and implement it into our daily life so it can help a lot of people.

1.2 Problem Formulation

1. How can we predict the onset of diabetes disease based on diagnostic measures using machine learning?
2. What is the difference between Naïve Bayes with SciKit-Learn library and Logistic Regression code?
3. Which library works better to predict the onset of diabetes disease?

1.3 Scope

The scope of this project is to get an output prediction and accuracy of the onset of diabetes based on diagnostic measurement by using two different libraries yet using the same algorithm. After getting two outputs from two different libraries then compares between two outputs and analyzes the differences.

1.4 Objective

1. Predict the onset of diabetes disease based on diagnostic measures using machine learning.
2. Show the differences between Naïve Bayes from SciKit-Learn Library and Logistic Regression code.
3. Analyze the differences between Naïve Bayes from SciKit-Learn Library and Logistic Regression code, which gives the better result.