CHAPTER 1
INTRODUCTION

1.1. Background

Diabetes is considered one of the deadliest and most chronic diseases that causes elevated blood sugar levels. Many complications develop when diabetes is left untreated. A lengthy identification process leads the patient to a diagnostic center or consultation doctor. However, the rise of machine learning approaches is solving this critical problem. The motivation for this study is to design a model that can predict a patient's likelihood of diabetes with maximum accuracy. Therefore, in this experiment, we use two machine learning classification algorithms, Decision Tree and XGBoost, to detect diabetes early.

Experiments are run on the Pima Indians Diabetes Database (PIDD) obtained from Kaggle. The performance of both algorithms are evaluated according to various criteria such as Precision, Accuracy, F-1 Score, and Recall. Accuracy is measured by correctly and incorrectly classified instances.

Results obtained show Decision Tree outperforms with the highest accuracy of 88.88% with 80% train data and 20% test data, balanced dataset and random state = 0

1.2. Problem Formulation

With reference to the above background, some problem formulations have been obtained. I used to get an answer from a project I did and depending on my conclusions about it. For a thesis, the following points are required:

1. Which algorithm has better accuracy for predicting Diabetes ?

1.3. Scope

The main limitations contained in this project are as follows :

1. This project use Decision Tree and XGBoost to predict the data
2. Using precision, recall, F-measure, and accuracy for evaluation metric

3. Using random state = 0 and 42 parameter to controls the shuffling applied to the data
4. The dataset only contains 768 data to predict whether its dibetic or not
5. The dataset only contain all of female patient with age older than 21

1.4. Objective

Goals to be achieved in this project is to predict the diabetic data and compare between Decision Tree and XGBoost algorithm that has the best accuracy. The best algorithm will be obtained that will be used to make a diabetes prediction system.