CHAPTER 4
ANALYSIS AND DESIGN

4.1. Analysis

In this subchapter, the analysis of the study is explained. First, the algorithms will be trained and predict the test set. Second, calculate the accuracy, precision , recall, and F1 score from the predicted result, and compare the result of both algorithms.

4.2. Design

In this subchapter, the design of the study is explained. First, load the dataset, the dataset was checked to see if there were any missing values, because it can lead to inaccurate results or reduce the models accuracy [9]. The next step is data preprocessing, one hot encoding is applied where changing positive, male, and yes into 1 and negative, female, and no into 2. After data preprocessing is done, the dataset was splitted into X and Y, where Y is the label/target which contains the attribute “class” that indicates whether the person suffers diabetes or not. And next, split the dataset into 2 sets, namely training and test set , with ratio 80:20, 75:25, 70:30, 60:40, 50:50.

Second step, we train the dataset with both algorithms. First the Artificial Neural Network, the ANN will be built in three models. The first model consists of input layers, hidden layers and output layers. The input layer has 16 nodes indicating each feature or column of the dataset. Then three hidden layers are used, with 200 neurons for the first and second layer, and 150 neurons for the third layer. Rectified Linear Unit (ReLu) activation function is used for each layer. The model output layer has two nodes with Sigmoid activation function to determine either the person predicted with diabetes or not. binary cross entropy (BCE) loss is used to calculate the cost function. The neural network’s parameters are updated using Adam optimizer with the learning rate of 0.001, and a batch size of 500. As for the used number of training or epochs is 500 [15]. After training the dataset with the ANN, then make the prediction with the test set and calculate for the analysis. The second model was built with 4 hidden layers and 200, 200, 200, 150 neurons respectively. The third model was built with 5 hidden layers and 200, 200, 150, 150, 150 neurons respectively. And the last one is 6 hidden layers with 200, 200, 200, 150, 150. 150 neurons respectively. All four models have the same Adam optimizer, learning rate, activation function, epochs, batch, input layer and output layers as the first ANN model.

The second algorithm is the Extreme Gradient Boosting. First, the dataset was trained with the XGBoost without hyperparameters tuning, then tuning the hyperparameters using Bayesian Optimizer by applying 10-fold validation to the XGBoost and analyzing the result to find the best hyperparameters. The hyperparameters that tuned are learning_rate, min_child_weight, colsample_bytree, max_depth, gamma, n_estimators or the number of boosting rounds with ranged as following Table 4.1 below [16]:

Table 4.1 Hyperparameters Search Spaces

Hyperparameters	range
max_depth	$3 - 11$
gamma	$1 \times 10^{-9} - 0.5$
n_estimators	$100 - 250$
learning_rate	$0.01 - 0.5$
min_child_weight	$1 - 10$
colsample_bytree	$0.1 - 0.8$

The used evaluation metric is Area Under Curve (AUC), because it was one of the best ways to get the classification model performance result generally, the unbalanced class, where the higher AUC score shows the best performance. [16]. After getting the best parameters, then the dataset is trained with the tuned parameters XGBoost , and makes a prediction with the test set. Lastly calculate the prediction result for the analysis.

After training , makes predictions and calculates for the analysis with both algorithms. We compare both algorithms accuracy, precision , recall, and f1 scores. From the comparison then we can find out which algorithms have the better performance or results on diabetes prediction.