CHAPTER 3
RESEARCH METHODOLOGY

This chapter describes in detail the steps taken on this project until later in the end find results that match what is done. This research stage discusses the workings of the system developed in this project. Here are some steps to take find the right and correct results.

3.1. Collecting Dataset

The dataset in this project, We took from Kaggle and named diabetes.csv that you can access it from here: https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database This dataset have 9 columns and 768 rows. The target column will be the output column on the 9th column.

3.2. Data Cleaning

The first thing we have to do is check the null value on Glucose, Blood Pressure, SkinThickness, Insulin, and BMI. Thus, all the zero values were replaced with the median value of that attribute.[1]

The next step is checking if the data is balanced or not. We have to compare the 0 and 1 values in Outcome column. The result is the data is unbalanced and we have to balancing the Outcome data by Oversampling the Minority values, which is the minority values is 1. The reason I use the Oversampling technique is to increase the sample of the minority class and to minimize errors in the main class. [3]

After doing oversampling the minority class, we have to check if there is outlier or not using boxplot and remove the outlier using Z-Score, We used z-score normalization on all the features to restrict the range of values between 3 to - 3. [15]

3.3. Data Correlation using Heatmap

Heatmaps are a fundamental visualization method that is broadly used to unravel patterns hidden in genomic data. They are especially popular for gene expression analysis (Eisen et al., 1998) and methylation profiling (Sturm et al., 2012). With the increasing

availability of genomic datasets, visualization methods that effectively show relations within multidimensional data are urgently needed. [8]

3.4. Fit the Train and test data

After that, we have to splitting the dataset into train and test data. In this case, we are splitting the dataset into 90% in the train data 10% in the test data, 80% in the train data 20% in the test data, 70% in the train data 30% ini the test data, and 60% in the train data 40% in the test data. We have to fit the train and test dataset into Decision Tree and XGBoost algorithm.

3.5. Decision Tree Algorithm

Based on research Data Mining Models Comparison for Diabetes Prediction, "After the implementations of these algorithms it can be said that for PID dataset Decision Tree gives best accuracy ". These algorithm means : Naive Bayes and KNN. [17]

The reason I use Decision Tree is because based on this research, the Decision Tree algorithm can provide higher accuracy than the Naive Bayes and KNN algorithms.

Decision Tree is a white box model and also an active learning scheme. Decision Tree comprise of several leaf nodes, some internal nodes and a single root node. A decision tree is shown in the Figure below, with its root at the top. In the figure square shape shows condition or interior node, in view of which the tree parts into diferent branches or edges. The end of the branch or edge that does not split any longer is the decision or leaf and is shown using oval shape. Every leaf node possesses a class label and is connected to the root node via internal nodes. The starting node of a Decision Tree is the root node and the path from this node to the leaf nodes yields the classifcation rules. System operators can use these rules as guidelines to assess and monitor real-time voltage stability. In this work, we use C4.5 Decision Tree algorithms that make use of information gain ratio for attribute selection. The employed C4.5 algorithm solves the over-fitted problem and is capable of efectively handling continuous attributes [14]

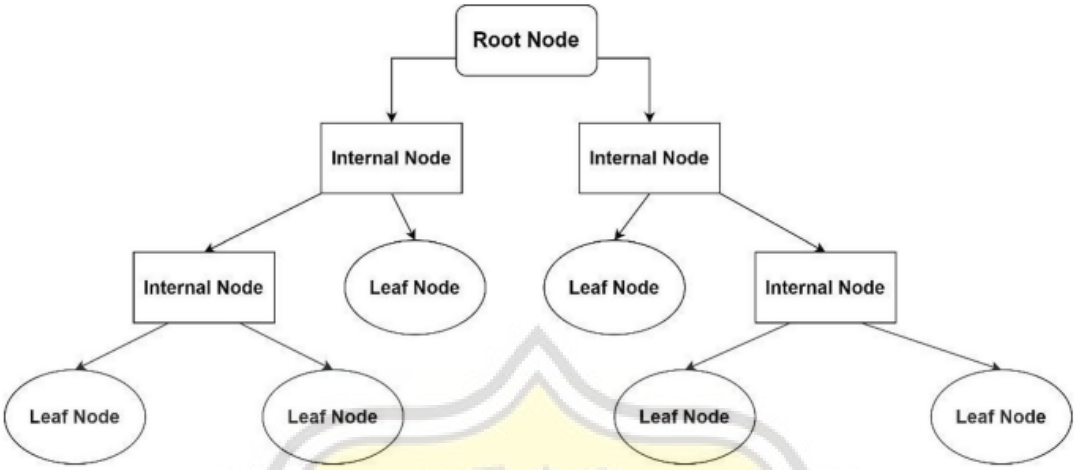

Figure 3.1 Decision Tree Architecture

3.6. XGBoost Algorithm

Based on the Research Analysis of Machine Learning Approaches in Diabetes Prediction, "Our Result shows that XG Boost achieved higher accuracy compared to other machine learning techniques."[16].

The reason I use XGBoost is because based on this research, XGBoost has the highest accuracy compared to other algorithms such as: K-Nearest Neighbor (KNN), Decision Tree (DT), Random Forest (RF), Ada Boost, Naive Bayie

XGBoost is a kind of ensemble learning algorithm based on Boosting. It is a general Gradient Boosting library developed by Dr. Chen of Washington University. The principle of XGBoost is to achieve accurate prediction results efficiently through iterative calculation of CART decision tree classifier. XGBoost is an optimization model that improves the existing gradient boost algorithm, which combines linear model and tree learning model. To avoid overfitting, XGBoost reduces the variance of the model by adding regular terms to the cost function, thus controlling the complexity of the model. The most important feature of this algorithm is that it can automatically use CPU multi-threads for parallel computing, and improve the accuracy by optimizing the algorithm. XGBoost algorithm has been widely used in the fields of artificial intelligence, data analysis, data mining and statistics. It is used to solve various practical problems and

has achieved very high accuracy. XGBoost is an improvement of boosting algorithm based on Gradient Boosting Decision Tree (GBDT). The idea of this algorithm is to construct multiple CART trees based on feature splitting nodes. Each time a CART tree is constructed, the residual predicted by the last model is fitted, so that the objective function is reduced. Finally, many CART weak classifiers are integrated into a strong classifier, and each leaf node of each tree corresponds to a score. When a sample is predicted, the model will find the corresponding leaf nodes in each tree according to the characteristics of the sample. The predicted value of the sample is the sum of the score of all leaf nodes. In the process of constructing XGBoost model, the optimal parameters of the model are obtained by training samples according to the principle of minimizing the objective function, and then the new sample is predicted by the optimal parameters and the prediction function. [18]

3.7. Evaluate the model with Precision, Recall, F-1 Score and, Confusion Matrix

The evaluation model is using Recall, Precision, F-1 Score and, Confusion Matrix. Precision evaluates the fraction of correctly classified instances or samples among the ones classified as positives. Recall is used as a placeholder for the name of the function in which it is called. F Score is used to evaluate binary classification systems, which classify examples into 'positive' or 'negative'.

Confusion matrix is used to visualize the performance of the algorithms which cross tabulates the observed and predicted classes with associated statistics, evaluation metrics like sensitivity, specificity, precision and accuracy are used to evaluate the performance of the method. The confusion matrix consists of four basic characteristics (numbers) that are used to define the measurement metrics of the classifier. These four numbers are:

1. TP (True Positive): TP represents the number of patients who have been properly classified to have malignant nodes, meaning they have the disease.
2. TN (True Negative): TN represents the number of correctly classified patients who are healthy.
3. FP (False Positive): FP represents the number of misclassified patients with the disease but actually they are healthy. FP is also known as a *Type I error*.

4. FN (False Negative): FN represents the number of patients misclassified as healthy but actually they are suffering from the disease. FN is also known as a *Type II error*.