ABSTRACT

Diabetes is a disease that continues to increase, and the number of deaths is increasing. This serious chronic disease is caused by metabolic disorders that occur because the pancreas cannot produce or produce enough insulin (a hormone that regulates glucose). According to the International of Diabetic Federation (IDF), the global prevalence rate of people with diabetes continues to increase every year. Diabetes is one of the most common diseases and the biggest cause of death in the world. Detection of diabetes can be done with data mining techniques. Data Mining is a process of collecting important information from big data in an expertise related to informatics. Data Mining can also be used in research that is engaged in other aspects, one of which is in the health sector to predict Diabetes in a group of individuals using the classification method. In this study, the dataset used came from the Pima Indians Diabetes Databases (PPID). This study aims to compare the classification performance of the Supervised Learning Algorithm, namely Naïve Bayes and XGBoost. This study will also handle missing values on the dataset and discuss the Grid Search method as an optimization based on the performance of diabetes classification accuracy on the Naïve Bayes Algorithm and XGBoost. Accuracy results are evaluated by using a confusion matrix and calculating the AUC value. So, from the classification results, the Naïve Bayes Algorithm classification model is obtained with a model accuracy value of 79.68% and the XGBoost Algorithm has a better performance with an accuracy value of 90.10%.

Keywords— ***Data Mining, Classification, Diabetes, Naive Bayes, XGBoost***