ABSTRACT

Diabetes or often referred to as diabetes is a disease due to metabolic disorders caused by high blood glucose levels in the body for a long time. If not treated early, diabetes can cause complications to other diseases. The International Diabetes Federation (IDF) estimates that there are at least 463 million people worldwide in the age range of 20-79 years suffering from diabetes in 2019. This number is predicted to continue to increase to reach 578 million people in 2030 and 700 million people in 2020. 2045. The state of Indonesia itself is in 7th place out of a total of 10 countries with the highest number of people with diabetes, which is 10.7 million. This number is predicted to continue to increase to 13.7 million people in 2030 and 16.6 million people in 2045. Many people are diagnosed with diabetes after experiencing complications. Therefore, practitioners and researchers focus their attention on detecting diabetes by using data mining techniques in extracting valuable information from diabetes data sets. The dataset used in this study is a dataset from the National Institute of Diabetes and Digestive and Kidney Diseases and can be accessed at the UCI Machine Learning Repository: Pima Indians Diabetes Database. This dataset contains 768 female patients with 8 different medical condition diagnosis attributes and 1 goal attribute or label attribute. This study compares the K-Nearest Neighbor and Logistic Regression algorithms for data classification of the Pima Indians Diabetes Database. In this study, the authors handle missing values on the data and use the Grid Search method to find models with optimal accuracy results. Accuracy results were evaluated by using a confusion matrix and calculating the AUC value. From the classification results, the results of the K-Nearest Neighbor algorithm with an accuracy value of 85.06% and the Logistic Regression algorithm with an accuracy of 77.92% are obtained.

*Keywords***— *diabetes, data mining, classification, k-nearest neighbor, logistic regression***