ABSTRACT

Diabetes is one of the chronic diseases whose prevalence is increasingly rising worldwide. Prevention and education regarding diabetes have become a crucial priority in the global health agenda. The implementation of technologies such as data mining and machine learning-based predictions in early diagnosis and management of diabetes is also a significant step in reducing the impact of this disease. Data mining builds a model that can predict whether someone has diabetes or not. The goodness or badness of a model in performing classification can be determined by its accuracy value. The results of classification model accuracy are uncertain, they are not always good or high. The improvement of accuracy values can be achieved through feature selection. Feature selection plays an important role in reducing model complexity by eliminating irrelevant features or those that are highly correlated with one another. The data used in this research is diabetes data consisting of several features that may influence the diagnosis of diabetes. The implementation was carried out using the Python 3 programming language in Google Colaboratory. After feature selection was performed, a Logistic Regression model was built and evaluated using accuracy, precision, recall, and F1-score metrics. The research results indicate that feature selection using Pearson Correlation successfully improved the accuracy performance of the diabetes prediction model by 1.03%, increasing from 89.74% without feature selection to 90.77% after feature selection was applied.

Keywords: Feature Selection, Pearson Correlation, Logistic Regression, Diabetes Predict