

ABSTRACT

Diabetes is a non-communicable disease that has a death rate of 70% in the world. Majority of diabetes cases, 90%-95%, are of diabetes cases are type 2 diabetes which is caused by an unhealthy lifestyle. Type 2 diabetes can be detected earlier by using an examination that contains diabetes-related parameters. However, the dataset does not always contain complete information, the distribution between positive and negative classes is mostly imbalanced, and some parameters have low importance to the decision class. To overcome the problems, preprocessing needs to be carried out to improve detection precision and recall. This paper proposes an approach on dataset preprocessing, which is applied to diabetes prediction. The preprocessing approach consists of the following process: missing value process, imbalanced data process, feature importance process, and data augmentation process. The data preprocessing process uses the median for missing value, random oversampling for imbalanced data, the Gini score in the random forest for feature importance, and posterior distribution for data augmentation. Random forest and logistic regression were used as classification algorithms. The experimental results show that the classification increased by 20% precision and 24% recall by applying our preprocessing proposed method with random forest classifier compared to without preprocessing proposed method with random forest classifier.

Keywords: Diabetes Mellitus Type 2, Data Preprocessing, Data Augmentation, Random Forest, Classification