PROJECT REPORT

DIABETES PREDICTION USING NAÏVE BAYES, KNN, AND RANDOM FOREST ALGORITHMS

DAVID JU
18.K1.0022

Faculty of Computer Science
Soegijapranata Catholic University
2023

ABSTRACT

Diabetes is a serious disease in which sufferers have high blood sugar because the body is unable to produce the insulin the body needs to regulate glucose. This allows complications to occur or can cause the risk of developing other diseases such as infection, vision problems, nerve damage, kidney failure, heart failure, etc. One way to anticipate is with early diagnosis. The more data there is about a patient's medical record, the classification algorithm can learn and can have a big role in predicting whether someone has diabetes or not. The dataset used was obtained from Kaggle which was sourced from the National Institute of Diabetes and Digestive and Kidney Diseases and Donor of database: Vincent Sigillito (vgs@aplcen.apl.jhu.edu) Research Center, RMI Group Leader Applied Physics Laboratory The Johns Hopkins University. The dataset has 765 data and 8 attributes. Many classification algorithms are used in preventing diabetes, one of which is Naïve Bayes, which is used for main predictions and comparing prediction results with the KNN and random forest algorithms. All models were tested with the same ratio with training data of 90%, 80%, 70%, and 60%. The accuracy results were obtained from Naïve Bayes, with a percentage accuracy of 74-81 percent. Then after that, the data was tested again by eliminating one of the 8 attributes to see which factor had the biggest influence on diabetes. The data results will be tested one by one for the 8 attributes using the Naïve Bayes, KNN and Random Forest algorithms. Tested the same as the previous training data, namely with test sizes 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9. This comparison was carried out to see what things really influence it by looking at accuracy, precision and recall in table form.

Keyword: Naïve Bayes, KNN and Random Forest