

INTISARI

ANALISIS PENGARUH FITUR DAN PENGEMBANGAN MODEL KLASIFIKASI *MACHINE LEARNING* UNTUK DETEKSI PENYAKIT DIABETES

Muhammad Aulia Alfarisi

19/441162/SV/16514

Diabetes mellitus adalah penyakit yang mengakibatkan kadar gula darah naik dan berpotensi menyebabkan komplikasi yang lebih serius. Skrining diabetes menjadi sangat penting agar tercegahnya terjadi masalah ini. Masalah tersebut dapat diatasi dengan perancangan sistem *machine learning* untuk memprediksi penyakit diabetes. Penelitian ini bertujuan untuk mengembangkan model klasifikasi machine learning untuk deteksi penyakit diabetes. Lima model klasifikasi seperti *Random Forest*, *Support Vector Machine*, *Logistic Regression*, *Linear Discriminant Analysis*, dan *Multi-layer Perceptron* dibandingkan untuk menentukan model terbaik dalam memprediksi diabetes. Dataset pertama diambil dari Kaggle dan dataset kedua dari *Vanderbilt Biostatistics Dataset* dari *Department of Medicine, University of Virginia School of Medicine*. Analisis kontribusi fitur menggunakan metode *Gini Importance* didapatkan hasil yaitu fitur *voltage_ratio*, *bmi*, dan *age* merupakan fitur yang paling berkontribusi dalam memprediksi diabetes. Hasil pelatihan dan tes menunjukkan bahwa model *Random Forest* memiliki akurasi tertinggi dengan nilai akurasi training 95% dan akurasi testing sebesar 87%. Model ini juga mempunyai presisi, recall, specificity, dan F1-score tertinggi dibandingkan model lainnya. Selain itu, hasil validasi model menggunakan dataset 2 menunjukkan bahwa kelima model klasifikasi memiliki akurasi di atas 90% dengan model *Logistic Regression* memiliki akurasi tertinggi dengan nilai 93% dan nilai AUC tertinggi dengan nilai 0,82. Hasil penelitian ini menunjukkan bahwa model *Random Forest* dan *Logistic Regression* merupakan model terbaik dalam memprediksi diabetes.

Kata Kunci: deteksi penyakit diabetes, *machine learning*, glukometer non-invasif, *feature importance*, klasifikasi

ABSTRACT

ANALYSIS OF THE CONTRIBUTION OF FEATURES AND DEVELOPMENT OF MACHINE LEARNING CLASSIFICATION MODELS FOR DIABETES DETECTION

Muhammad Aulia Alfarisi
19/441162/SV/16514

Diabetes mellitus is a disease that causes blood sugar levels to rise and has the potential to cause more serious complications. Diabetes screening is very important to prevent this problem. This issue can be overcome by developing a machine learning system to predict diabetes. This research focuses on developing a machine learning classification model for diabetes detection. Five classification models such as Random Forest, Support Vector Machine, Logistic Regression, Linear Discriminant Analysis, and Multi-layer Perceptron were evaluated to identify the most accurate model for predicting diabetes. The first dataset was taken from Kaggle and the second dataset from the Vanderbilt Biostatistics Dataset from the Department of Medicine, University of Virginia School of Medicine. Analysis of feature contribution using the Gini Importance method showed that the voltage_ratio, BMI and age features were the features that contributed most to predicting diabetes. The training and test results indicate that the Random Forest model has the highest accuracy with a training accuracy value of 95% and testing accuracy of 87%. This model also has the highest precision, recall, specificity and F1-score compared to other models. Apart from that, the results of model validation using dataset 2 show that the five classification models have an accuracy of above 90% with the Logistic Regression model having the highest accuracy with a value of 93% and the highest AUC value with a value of 0.82. The results of this study show that the Random Forest and Logistic Regression models are the best models in predicting diabetes.

Keywords: *diabetes detection, machine learning, non-invasive glucometer,, feature importance, classification*