

## ABSTRACT

### DIAGNOSIS DIABETES DARI REKAMAN MEDIS ELEKTRONIK MENGGUNAKAN DBSCAN, K-MEANS CLUSTERING, DAN NAIVE BAYES

Oleh

Bagaskara Putra Wibowo

18/425515/PA/18407

Pandemi COVID-19 membuat pemerintah menetapkan mandat karantina yang membatasi aktivitas di luar ruangan bagi warga. Perubahan gaya hidup di masa pandemi termasuk kurang berolahraga dan perubahan kebiasaan makan mengakibatkan peningkatan risiko terkena diabetes. Kunci untuk mengatasi diabetes adalah deteksi dini. Pendekatan pembelajaran mesin dapat digunakan untuk membantu dokter mengidentifikasi pasien diabetes melalui data rekam medis elektronik. Penelitian-penelitian yang ada dengan menggunakan metode *supervised learning* memiliki akurasi yang rendah yang disebabkan oleh kesalahan data.

Pendekatan *hybrid unsupervised-supervised learning* dapat mengatasi masalah ini dengan menggunakan pendekatan *unsupervised learning* untuk memproses data terlebih dahulu sehingga pengklasifikasi *supervised* dapat bekerja lebih baik. Penelitian ini menggunakan DBSCAN dan *k-means clustering* dengan tujuan untuk menghilangkan data *outlier* dan *noise* dari dataset Pima Indians Diabetes Database. DBSCAN mengelompokkan data mentah menggunakan *epsilon* dan titik minimum yang ditentukan untuk menghapus data apa pun yang tidak mengelompok. *K-means clustering* kemudian mengelompokkan data yang dihasilkan untuk mengelompokkan data serupa di mana data yang berbeda dari mayoritas dalam *cluster* dianggap sebagai *noise* dan dihapus. Data bersih diterapkan pada pengklasifikasi Naive Bayes.

Dari percobaan, metode yang diusulkan memperoleh akurasi rata-rata 98,0%, *precision* rata-rata 99,3%, *recall* rata-rata 97,9%, dan skor F1 98,6%. Metode yang diusulkan memiliki performa lebih baik dibanding metode-metode penelitian sebelumnya dengan pendekatan serupa.

**Keywords :** Diabetes, DBSCAN, *k-means clustering*, naive bayes

## ABSTRACT

### DIABETES DIAGNOSIS FROM ELECTRONIC MEDICAL RECORDS USING DBSCAN, K-MEANS CLUSTERING, AND NAIVE BAYES

By

Bagaskara Putra Wibowo

The COVID-19 pandemic has led to the government establishing a quarantine mandate that limits outdoor activities for citizens. The change in lifestyle during the pandemic era including less exercise and changes in eating habits resulted in a surge in the risk of people contracting diabetes. The key to treating diabetes is early detection. A machine learning approach can be used to help doctors identify patients having diabetes through electronic medical records data. Existing researches using supervised learning methods have low accuracy that are caused by errors in the data.

A hybrid unsupervised-supervised learning can solve this issue by using unsupervised learning approaches to preprocess the data so that the supervised classifier can perform better. This research attempts to use DBSCAN and K-means clustering for the purpose of removing outliers and noisy data from the Pima Indians Diabetes Database dataset. DBSCAN clusters the raw data using specified epsilon and minimum points to remove any data that are not clustered. K-means clustering then clusters the resulting data to group similar data together where data that are different from the majority within the clusters are considered noisy and removed. The clean dataset is applied on a Naive Bayes classifier.

From the experiment, the proposed method obtained an average accuracy of 98.0%, average precision of 99.3%, average recall of 97.9%, and an F1-score of 98.6%. The proposed method yields a better performance when compared to past researches with similar approaches.

**Keywords :** Diabetes, DBSCAN, k-means clustering, naive bayes