ABSTRACT

The focus of the study is to build a classification model that predicts diabetes mellitus type 2 from patient's Electronic Health Records (EHR) data using ensemble learning weighted voting that combines Naive Bayes, K Nearest Neighbor and Random Forest. This study aims to build a classification model to predict patients from existing data, not to predict future patient that have not yet received medical diagnosis. To do so, several phases were conducted; data collection, feature processing, dimensionality reduction (feature selection and extraction), data training and data testing.

The dataset were Electronic Health Records obtained from Practice Fusion in Kaggle that consists of 9,943 patient data with 17 different tables that were connected to each other. The data contains information about patient's general information (age, state, year of birth, etc.) with anonymous personal information, physical information (BMI, height, weight, blood pressure, etc.), past diagnosis, smoking status, medication received, lab observation, prescription records, immunization and allergy. Features was chosen based on domain knowledge from World Health Organization (1999), past research from Gilies et al., (2012) about diagnoses of diabetes mellitus type 2 and data completeness condition in each feature after preprocessing.

Experiment shows that ensemble learning weighted voting does not produce a better result in terms of performance in precision (75%), recall (71%) and accuracy (91%). This can be concluded from the experiment result that finds the highest precision and accuracy was obtained by Random Forest (92%), while the highest recall is obtained by Naive Bayes (91%).

Keywords: machine learning, classification, electronic health records, ensemble learning weighted voting.

INTISARI

Fokus pada studi ini adalah untuk membuat sebuah model klasifikasi untuk memprediksi pasien yang memiliki penyakit diabetes mellitus tipe 2 dari data *Electronic Health Records (EHR)* setiap pasien. Metode klasifikasi ini akan menggunkan *ensemble learning weighted voting* yang mengkombinasikan algoritma *Naïve Bayes, K Nearest Neighbor* dan *Random Forest*. Tujuan dari studi ini adalah untuk membangun model klasifikasi yang dapat memprediksi pasien dari data yang sudah ada, bukan untuk memprediksi pasien yang akan datang yang belum memiliki diagnosa kesehatan. Untuk melakukannya, terdapat beberapa proses yang dilakukan; mulai dari pengumpulan data, praproses, pengurangan jumlah dimensi data (terdiri dari pemilihan fitur dan ekstraksi fitur), pelatihan data dan pengujian data.

Data yang digunakan berasal dari Practice Fusion yang memiliki 9,943 data pasien dengan 17 tabel yang berbeda. Setiap table mewakili diagnosa kesehatan dari beberapa (tidak semua) pasien yang terhubung antara satu sama lain. Data yang diperolah terdiri informasi umum mengenai pasien (umur, asal, tahun lahir, dsb.) dengan data pribadi yang disamarkan, informasi fisik (BMI, tinggi badan, berat badan, tekanan darah, dsb.), *track record* diagnosa yang pernah terdeteksi, status perokok, obat yang pernah dikonsumsi, observasi laboratorium, data resep obat, imunisasi dan alergi. Fitur dipilih bedasarkan pengetahuan umum mengenai penyakit terkait dari World Health Organization (1999), penelitian Gilies *dkk* (2012) dan kondisi kelengkapan data setelah di praproses.

Penelitian ini menyimpulkan bahwa metode *ensemble learning weighted voting* tidak memiliki hasil yang lebih baik dari segi performa *precision* (75%), *recall* (71%) dan *accuracy* (91%). Kesimpulan ini dapat dilihat dari hasil *precision* dan *accuracy* terbaik dihasilkan oleh *Random Forest* (92%), serta *recall* terbaik diperoleh *Naive Bayes* (91%).

Kata kunci: *machine learning*, klasifikasi, *electronic health records*, *ensemble learning weighted voting*.