PENGGABUNGAN ALGORITMA *FORWARD SELECTION* DAN *K-NEAREST NEIGHBOR* UNTUK MENDIAGNOSIS PENYAKIT DIABETES DI KOTA SEMARANG

Laily Hermawanti[1]* dan Achmad Nuruddin Safriandono[2]

[1] Program Studi Teknik Informatika, Fakultas Teknik, Universitas Sultan Fatah

[2] Program Studi Sistem Komputer, Fakultas Teknik, Universitas Sultan Fatah

Jl. Diponegoro 1A Jogoloyo Demak

*Email: lailyhermawanti@gmail.com

Abstrak

K-Nearest Neighbor merupakan salah satu algoritma yang diusulkan oleh para peneliti data mining di bidang kesehatan seperti diabetes. Diabetes adalah perubahan menetap dalam sistem kimiawi tubuh yang mengakibatkan darah mengandung terlalu banyak garam. Penyebab penyakit diabetes adalah kekurangan hormon insulin. Hormon adalah unsur kimia yang dibuat oleh tubuh (dalam hal ini pankreas) dan dilepas ke dalam aliran darah untuk digunakan oleh bagian tubuh yang membutuhkan. Akibat salah satu penyakit diabetes adalah buang air kecil lebih sering, sebab kelebihan gula dalam darah disaring keluar oleh ginjal dengan mengeluarkan lebih banyak garam dan air. Maka dari itu, penyakit diabetes perlu didiagnosis. Algoritma yang digunakan dalam penelitian ini adalah penggabungan algoritma Forward Selection dan K-Nearest Neighbor. Hasil penelitian menunjukkan bahwa penggabungan algoritma Forward Selection dan K-Nearest Neighbor memiliki akurasi yang lebih baik dari pada algoritma K-Nearest Neighbor. Penelitian ini menghasilkan akurasi yang lebih tinggi daripada penelitian-penelitian sebelumnya.

***Kata kunci**: data mining, Forward Selection, K-Nearest Neighbor, diabetes*

PENDAHULUAN

Diabetes adalah perubahan menetap dalam sistem kimiawi tubuh yang mengakibatkan darah mengandung terlalu banyak garam (http://www.atasidiabetes. com/2012/04/arti-diabetes-dan-penyebab-diabetes.html). Salah satu penyebab penyakit diabetes adalah kekurangan hormon insulin. Hormon adalah unsur kimia yang dibuat oleh tubuh (dalam hal ini pankreas) dan dilepas ke dalam aliran darah untuk digunakan oleh bagian tubuh yang membutuhkan. Salah satu akibat penyakit diabetes adalah buang air kecil lebih sering, sebab kelebihan gula dalam darah disaring keluar oleh ginjal dengan mengeluarkan lebih banyak garam dan air (http://www. atasidiabetes.com/2012/04/arti-diabetes-dan-penyebab diabetes.html). Maka dari itu, penyakit diabetes perlu di diagnosis.

Data mining dapat diaplikasikan di bidang kesehatan misalnya mendiagnosis penyakit kanker payudara, penyakit jantung, penyakit diabetes dan lain-lain (D. T. Larose, 2005). Terdapat beberapa metode dalam mendiagnosis penyakit diabetes misalnya *K-Nearest Neighbor* (Asha et al, 2012), dan *Naïve Bayes* (Jia Wu dan Zhihua Cai, 2011) dan lain-lain.

Penelitian yang dilakukan oleh Jia Wu dan Zhihua Cai menjelaskan bahwa para peneliti telah mengusulkan banyak metode efektif untuk meningkatkan performa *Naive Bayes* seperti metode *backward sequential elimination, lazy elimination* dan sebagainya. Pembahasan ini mengevaluasi performa konfigurasi baru (DE-WNB) pada 36 UCI seluruh standar *data set* dalam sistem Weka. Hasil eksperimen menunjukkan akurasi klasifikasi algoritma baru DE-WNB lebih tinggi dari algoritma lain yang digunakan untuk membandingkan. Algoritma *Differential Evolution Weighted Naïve Bayes* (DE-WNB) menghasilkan keakuratan sebesar 75.40±4.65 untuk *dataset* diabetes. Algoritma *Naïve Bayes* (NB) menghasilkan keakuratan sebesar 75.68%±4.85%. Algoritma *Gain Ratio-Weighted Naïve Bayes* (GR-WNB) menghasilkan keakuratan sebesar 65.11%±0.34%. Algoritma *Correlation-based Feature Selection-Weighted Naïve Bayes* (CFS-WNB) menghasilkan keakuratan sebesar 77.02%±4.87%. Algoritma *Mutual Information-Weighted Naïve Bayes* (MI-WNB) menghasilkan keakuratan sebesar

65.37%±0.90%. Algoritma *Tree- Weighted Naïve Bayes* (*Tree*-WNB) menghasilkan keakuratan sebesar 76.91%± 5.07% (Jia Wu dan Zhihua Cai, 2011).

Penelitian yang dilakukan oleh Asha et al, menjelaskan tentang penyakit diabetes dan menggunakan Pima Indian *dataset* diabetes. Penelitian ini menggunakan algoritma *K-Nearest Neighbor* (K-NN). Algoritma *K-Nearest Neighbor*(K-NN) menghasilkan akurasi sebesar 74,7826% (Asha et al, 2012).

Dari penelitian yang pernah dilakukan untuk diagnosis penyakit diabetes terutama yang menggunakan algoritma *K-Nearest Neighbor*, akurasinya belum tinggi.Penelitian ini menggunakan penggabungan algoritma *Forward Selection* dan *K-Nearest Neighbor* untuk mendiagnosis penyakit diabetes. *Forward Selection* digunakan untuk mereduksi ukuran *data set* (J. Han dan M. Kamber, 2006) dan diharapkan dapat meningkatkan hasil akurasi pada *K-Nearest Neighbor*.

Tujuan penelitian ini adalah untuk menerapkan penggabungan algoritma *K-Nearest Neighbor* dan *Forward Selection* untuk meningkatkan akurasi dalam mendiagnosis penyakit diabetes.

Urgensi penelitian ini adalah dapat memberikan kontribusi keilmuan tentang penerapan penggabungan algoritma *Forward Selection* dan *K-Nearest Neighbor* untuk meningkatkan akurasi dalam diagnosis penyakitdiabetes. Selain itu, hasil dari penelitian ini dapat digunakan oleh pihak rumah sakit sebagai analisa diagnosis penyakit diabetes dengan menggunakan penggabungan algoritma *Forward Selection* dan *K-Nearest Neighbor.*

METODE

Penelitian ini menggunakan proses *Cross-Standard Industry-Data Mining* (CRISP-DM) dengan tahap-tahap penelitian meliputi pemahaman bisnis, pemahaman data, pengolahan data, pemodelan dan evaluasi (Larose, 2005).

Tahap Pemahaman Bisnis

Penelitian ini dilakukan untuk menerapkan penggabungan algoritma *Forward Selection* dan *K-Nearest Neighbor* untuk meningkatkan akurasi dalam mendiagnosis penyakit diabetes.

Tahap Pemahaman Data

Penelitian ini mengambil *dataset* diabetes dari laboratorium.

Tahap Pengolahan Data

Teknik-teknik pengolahan data awal (data *pre-processing*) yang digunakan pada penelitian ini adalah (J. Han dan M. Kamber, 2006) :

1. Data *cleaning* dapat digunakan untuk data yang *missing value*. Karena ditemukan adanya data yang terlewat tidak terisi (*missing value*) pada data. Pengolahan data awal dilakukan untuk mengisi nilai yang *missing value* dengan pekerjaan *replace missing value* dilakukan.
2. Data *reduction* digunakan untuk menghasilkan *data set* yang volumenya lebih kecil. Salah satu strategi *data reduction* yang digunakan pada penelitian ini adalah *attribute subset selection. Attribute subset selection* digunakan untuk mereduksi ukuran *data set* dengan menghilangkan atribut-atribut yang tidak relevan atau *redudant*. Salah satu teknik *attribute subset selection* yang digunakan pada penelitian ini adalah *Forward Selection.*

Tahap Pemodelan

Model yang digunakan dalam tahap ini menggunakan penggabungan algoritma *Forward Selection* dan *K-Nearest Neighbor*.

Algoritma *K-Nearest Neighbor*

Algoritma *K-Nearest Neighbor* merupakan salah satu algoritma yang digunakan untuk klasifikasi, meskipun juga dapat digunakan untuk estimasi dan prediksi (D.T. Larose, 2005). *K-Nearest Neighbor* adalah contoh algoritma berbasis pembelajaran, di mana *data set* pelatihan (*training*) disimpan, sehingga klasifikasi untuk *record* baru yang tidak diklasifikasi didapatkan dengan membandingkannya dengan *record* yang paling mirip dengan *training set* (D.T. Larose, 2005).

Langkah-langkah algoritma *K-Nearest Neighbor* adalah (D.T. Larose, 2005):

1. Menentukan parameter k, misal k = 5.
2. Menghitung jarak (*similarity*) di antara semua *training records* dan objek baru.
3. Pengurutan data berdasarkan nilai jarak dari nilai yang terkecil sampai terbesar.
4. Pengambilan data sejumlah nilai k (misal k=5).

5. Menentukan label yang frekuensinya paling sering di antara *k training records* yang paling dekat dengan objek.

Algoritma *Forward Selection*

Forward Selection menghilangkan atribut-atribut yang tidak relevan (J. Han dan M. Kamber, 2006). Algoritma *Forward Selection* didasarkan pada model regresi linear (R. Noori, dkk., 2011).

Langkah-langkah *Forward Selection* adalah (D. T. Larose, 2007) :

1. Mulai dengan tidak ada variabel-variabel dalam model.
2. Variabel yang paling berkorelasi dengan *hemoglobin* sebagai variabel dependen dipilih dan jika signifikan dimasukkan ke dalam model.
3. Menentukan prediktor-prediktor yang dimasukkan ke dalam model.
4. Prosedur yang variabel-variabelnya tidak signifikan maka masuk ke dalam model dan model regresi berganda (*multiple regression*) untuk *hemoglobin* sebagai variabel dependen:
   $y = \beta_0 + \beta_1(jenis\ kelamin) + \beta_2(trigliserid) + \beta_3(creatinin) + \beta_4(kolesterol\ total) + \varepsilon$
5. Menghitung *F*-statistik sekuensial pada variabel-variabel.

Algoritma *Forward Selection – K-Nearest Neighbor*

Langkah-langkah algoritma *Forward Selection - K-Nearest Neighbor* (KNN) adalah sebagai berikut :

1. Mulai dengan tidak ada variabel-variabel dalam model.
2. Variabel yang paling berkorelasi dengan variabel *dependent* maka dipilih dan jika signifikan dimasukkan ke dalam model.
3. Menentukan prediktor-prediktor yang dimasukkan ke dalam model.
4. Prosedur yang variabel-variabelnya tidak signifikan maka masuk ke dalam model dan model regresi berganda (*multiple regression*) untuk variabel *dependent* contoh *hemoglobin*:
   $y = \beta_0 + \beta_1(jenis\ kelamin) + \beta_2(trigliserid) + \beta_3(creatinin) + \beta_4(kolesterol\ total) + \varepsilon$
5. Menentukan atribut-atribut yang dipilih oleh *Forward Selection*.
6. Menentukan parameter k, misal k = 5.
7. Menghitung jarak (*similarity*) di antara semua *training records* dan objek baru.
8. Pengurutan data berdasarkan nilai jarak dari nilai yang terkecil sampai terbesar.
9. Pengambilan data sejumlah nilai k (misal k=5).
10. Menentukan label yang frekuensinya paling sering di antara *k training records* yang paling dekat dengan objek.

Tahap Evaluasi

Evaluasi pada penelitian ini menggunakan *confusion matrix* (*accuracy*).

HASIL DAN DISKUSI

Akurasi *dataset* diabetes dapat dilihat pada Tabel 1. Pada tabel 1, algoritma *K-Nearest Neighbor* menggunakan *dataset* diabetes menghasilkan akurasi sebesar 95.29%, sedangkan algoritma *Forward Selection-K-Nearest Neighbor* menghasilkan akurasi sebesar 96.08% sehingga mengalami peningkatan akurasi. Hasil pada tabel 1, akurasi metode *Forward Selection - K-Nearest Neighbor* lebih tinggi dari algoritma *K-Nearest Neighbor*.

Tabel 1. Akurasi Dataset Diabetes

Algoritma	Akurasi (%)
K-Nearest Neighbor	95.29%
Forward Selection–KNN	96.08%

Hasil menunjukkan metode *Forward Selection-K-Nearest Neighbor* dapat mencapai akurasi yang tinggi dalam mendiagnosis penyakit diabetes. Percobaan ini dilakukan untuk menunjukkan peningkatan akurasi dari algoritma *K-Nearest Neighbor* menjadi *Forward Selection -K-Nearest Neighbor*.

KESIMPULAN

Penggabungan algoritma *Forward Selection* dan *K-Nearest Neighbor* tingkat akurasinya lebih tinggi dari pada algoritma *K-Nearest Neighbor* dalam mendiagnosis penyakit diabetes. Penelitian ini menunjukkan penggabungan algoritma *Forward Selection* dan *K-Nearest Neighbor* merupakan salah satu algoritma yang tepat dalam mendiagnosis penyakit diabetes. Penelitian ini menghasilkan akurasi yang lebih tinggi daripada penelitian-penelitian sebelumnya.

DAFTAR PUSTAKA

http://www.atasidiabetes.com/2012/04/arti-diabetes-dan-penyebab diabetes.html

Asha Gowda Karegowda, M.A. Jayaram, A.S. Manjunath, 2012, Cascading K-means Clustering and K-Nearest Neighbor Classifier for Categorization of Diabetic Patients, *International Journal of Engineering and Advanced Technology (IJEAT).*

D. T. Larose, 2005, *Discovering Knowledge in Data: An Introduction to Data Mining*, United States of America: John Wiley & Sons, Inc.

D. T. Larose, 2007, *Data Mining Methods and Models*. New Jersey, Canada: John Wiley& Sons, Inc,.

Jia Wu and Zhihua Cai, 2011, Attribute Weighting via Differential Evolution Algorithm for Attribute Weighted Naive Bayes (WNB), *Journal of Computational Information Systems*, pp. 1672-1679,.

J. Han and M. Kamber, 2006, *Data Mining Concept dan Techniques*, 2nd ed. United States of America: Diane Cerra.

R. Noori, A. R. Karbassi, A. Moghaddamnia, D. Han, M. H. Zokaei-ashtiani, and Farokhnia, ,2011, Assessment of input variables determination on the SVM model performance using PCA, Gamma test , and forward selection techniques for monthly stream flow prediction, *Journal of Hydrology*, vol. 401, no. 3–4, pp. 177–189.