IMPLEMENTASI ALGORITMA *DECISION TREE* C4.5 UNTUK PREDIKSI PENYAKIT DIABETES

Noviandi
Program Studi Rekam Medis dan Informasi Kesehatan, Fakultas Ilmu-Ilmu Kesehatan,
Universitas Esa Unggul, Jakarta
Jalan Arjuna Utara No. 9, kebon Jeruk, Jakarta 11510
noviandi@esaunggul.ac.id

Abstract

Diabetes mellitus (DM) is a chronic disease that causes death. Uncontrolled, identified and unpredictable increases in blood sugar quickly lead to complications. In data mining, many have used approaches to predict the disease, one of which is the use of algortima decison tree C4.5. The motive of this study is to build a predictive model of the likelihood of diabetic patients with the C4.5 algorithm and see the akurasi of the resulting model. Prediction models are made using Pima Indians Diabetes Databases (PPID) data sourced from the UCI Machine Learning Repository. Prediction model with C4.5 decision tree algorithm has 70.32% akurasi by producing 9 rules, with the number of classes "not" as many as 4 rules and classes "yes" as many as 5 rule to predict DM disease.

Keyword: *diabetes, decision tree C4.5, Accuracy*

Abstrak

Diabetes Melitus (DM) adalah salah satu penyakit penyakit kronis yang menyebabkan kematian. Peningkatan gula darah yang tidak terkontrol, teridentifikasi dan tidak terprediksi dengan cepat mengakibatkan terjadinya komplikasi. Dalam data mining telah banyak menggunakan pendekatan-pendekatan dalam melakukan prediksi penyakit salah satu nya penggunaan algortima decison tree C4.5. Motif dari penelitian ini adalah membangun sebuah model prediksi kemungkinan diabetes pasien dengan algoritma C4.5 dan melihat akurasi dari model yang dihasilkan. Model prediksi dibuat dengan menggunakan data Pima Indians Diabetes Databases (PPID) yang bersumber dari UCI Machine Learning Repository. Model prediksi dengan algoritma *decision tree* C4.5 memiliki akurasi 70.32% dengan menghasilkan 9 rule, dengan jumlah class tidak sebanyak 4 rule dan 5 rule class iya untuk melakukan prediksi penyakit DM.

Kata kunci*:* Diabetes, C4.5 *decision tree*, Akurasi

Pendahuluan

Diabetes mellitus (DM) merupakan penyakit gangguan metabolik akibat pankreas tidak memproduksi cukup insulin atau tubuh tidak dapat menggunakan insulin yang diproduksi secara efektif sehingga terjadi peningkatan konsentrasi glukosa dalam darah yang dikenal dengan *hiperglikemia.* DM dapat dibagi atas diabetes tipe 1, yang terjadi karena produksi insulin yang tidak memadai oleh pankreas dan diabetes tipe 2, terjadi karena kegagalan sel dalam respon efektif terhadap insulin yang di produksi oleh pancreas(1). Peningkatan penyakit DM setiap tahun selalu terjadi. (1) mengatakan bahwa penderita penyakit DM telah mencapai 382 juta pada tahun 2013 yang membawa 6,6% dari total populasi orang dewasa didunia. Riskesdas 2013 penyakit DM di Indonesia adalah 6.9% dari total populasi orang diatas 15 tahun.

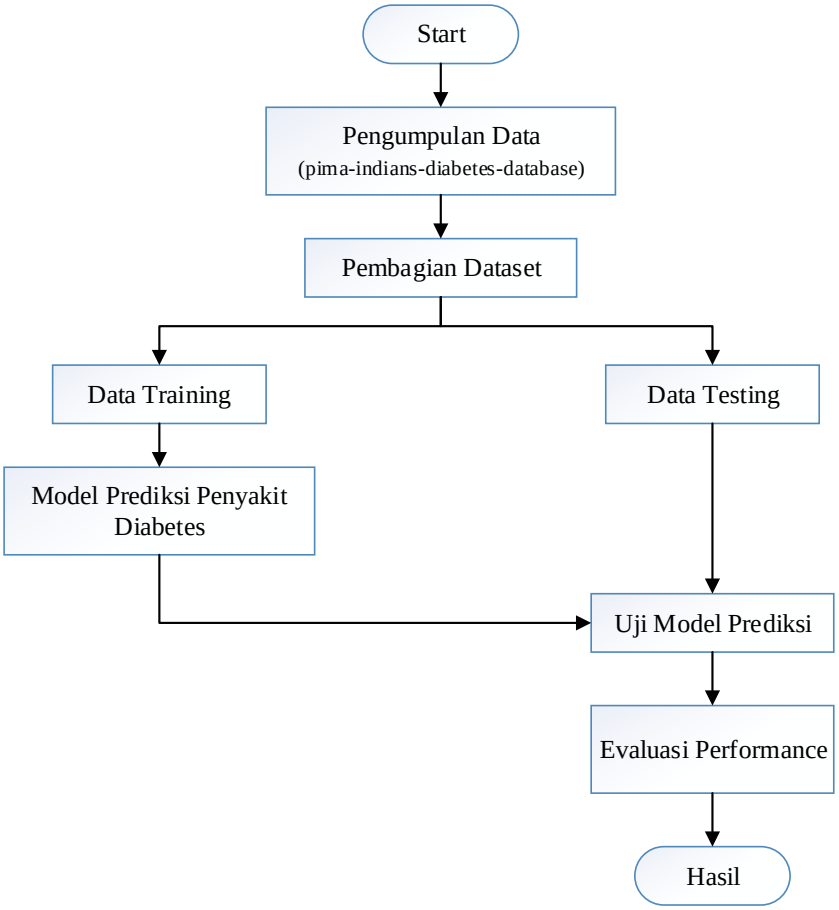
Diabetes adalah faktor resiko penyebab terjadinya komplikasi mikrovaskuler. Penderita DM mengakibatkan penyakit cardio vascular dua hingga empat kali lebih banyak dibandingakan dengan yang bukan penderita DM. Kerusakan vascular mikro dan akibat penyakit vascular kardio adalah retinopati dan neuropati. Dampak yang diakibatkan penyakit DM tesebut, maka dilakukan prediksi awal.

Kegiatan dalam melakukan prediksi terhadap berbagai penyakit telah banyak dilakukan dalam berbagai bidang keilmuan salah satunya bidang ilmu komputer science. Banyak percobaan yang di lakukan peneliti dengan menggunakan teknik data mining untuk memprediksi penyakit menggunakan berbagai algoritma seperti Naïve Bayes, Decission Tree, SVM, J48 dan lain-lain. Penelitian yang dilakukan (2) melakukan peningkatan terhadap algoritma K-means dan algoritma regresi logistik untuk memprediksi

penyakit diabetes tipe 2. Dataset yang diukur terdiri atas beberapa variabel, yaitu *number of times pregnant, plasma glucose concentration at 2h in an oral glucose tolerance test, diastolic blood pressure, triceps skin fold thickness, 2-h serum insulin, BMI, diabetes pedigree function* dan *age.* Model yang dihasilkan memiliki akurasi prediksi 3,04% sehingga model tersebut bermanfaat untuk manajemen kesehatan. (3) melakukan deteksi penyakit diabetes dengan menggunakan data *Pima Indians Diabetes Databases* (PIDD). Algoritma Decission Tree, SVM dan Naïve Bayes digunakan untuk memprediksi penyakit diabetes. Kinerja dari ketiga algoritma dievaluasi pada berbagai ukuran. Algoritma Naïve Bayes memiliki nilai akurasi tertinggi 76,30% dibandingkan dengan dua algortima lainnya. Penerapan algoritma C4.5 dilakukan (4) dengan menggabukan C4.5 dan adaboost untuk prediksi penyakit jantung. Model prediksi penyakit jantung dengan algoritma C4.5 memiliki nilai akurasi 86.59%.

Choubey et al dalam(5) menggunakan Genetic Algoritma (GA) dan Naïve Bayes untuk klasifikasi penyakit diabetes. Algoritma GA digunakan dalam menyeleksi atribut, sedangkan algoritma naïve bayes sebagai metode dalam klasifikasi. Hasil eksperimen menunjukkan bahwa kedua algoritma tersebut memberikan klasifikasi yang lebih baik terhadap dataset PIDD untuk diagnosia penyakit diabetes.

Penelitian ini melakukan prediksi secara diagnostik apakah pasien menderita penyakit diabetes dengan meng-implementasi algoritma C4.5, terhadap wanita yang telah melahirkan dengan melihat beberapa faktor lainnya.



Gambar 1
Diagram Model yang Diusulkan

Metode Penelitian

Dataset diabetes yang digunakan adalah data sekunder dari database kesehatan *Pima Indians Diabetes Dataset* (PPID) yang dapat diakses melalui https://www.kaggle.com/uciml/pima-indians-diabetes-database. Data terdiri atas 768 record dengan beberapa variabel prediktor medis (jumlah kehamilan pasien, kadar gula darah, tekanan darah, tingkat insulin, BMI, usia) dan satu variabel target. Data kadar gula darah diperiksa pada pasien yang telah berpuasa selama 8 sampai 10 jam dan kemudian diberi beban glukosa sebanyak 75 gram. Kategori kadar gula darah normal adalah ≤ 139, prediabetes 140 sampai

dengan 199 dan di kategorikan diabetes ≥ 200. Kategori data insulin adalah, 0 menyatakan bahwa tidak ada insulin dalam serum darah pasien, > 41μU/ml adalah resistensi insulin. Kategori BMI menurut WHO, kategori kurus adalah < 18.5, kategori normal 18.5 sampai dengan 25.0 dan kategori gemuk adalah > 25.0.

Tabel 1
Data Penyakit Diabetes
National Institute of Diabetes and Digestive and Kidney Diseases

Jumlah Wanita Melahirkan	Kadar Gula Darah	Tekanan Darah	Insulin	Body Mash Index	Usia	Outcome
6	148	72	0	33.6	50	1
1	85	66	0	26.6	31	0
8	183	64	0	23.3	32	1
1	89	66	94	28.1	21	0
0	137	40	168	43.1	33	1
5	116	74	0	25.6	30	0
3	78	50	88	31	26	1
10	115	0	0	35.3	29	0
2	197	70	543	30.5	53	1
8	125	96	0	0	54	1
4	110	92	0	37.6	30	0
10	168	74	0	0	38	0.537
10	139	80	0	0	27.1	1.441
1	189	60	23	846	30.1	0.398
5	166	72	19	175	25.8	0.587
7	100	0	0	0	30	0.484
0	118	84	47	230	45.8	0.551
7	107	74	0	0	29.6	0.254
1	103	30	38	83	43.3	0.183
1	115	70	30	96	34.6	0.529
3	126	88	41	235	39.3	0.704
8	99	84	0	0	35.4	0.388
7	196	90	0	0	39.8	0.451
9	119	80	35	0	29	0.263
11	143	94	33	146	36.6	0.254

Sumber: www.kaggle.com/uciml/pima-indians-diabetes-database

Pembagian Dataset

Dataset diabetes dibagi menjadi data training dan data testing. Data training digunakan untuk menghasilkan model prediksi dengan menggunakan algoritma *decision tree C4.5* dan data testing digunakan untuk melihat performa model prediksi yang dihasilkan. Pembagian data training dan data testing dalam penelitian ini menggunakan metode k fold cross validation dengan jumlah fold = 10

Decision tree

Decision tree adalah algoritma *supervised machine learning* yang digunakan untuk memecahkan masalah klasifikasi. Tujuan utama menggunakan algoritma *decision tree,* karena algoritma C4.5 mampu menghasilkan model prediksi secara spesifik dalam bentuk aturan yang mudah untuk diimplementasikan. Dalam *decision tree* memiliki root node dan internode untuk melakukan prediksi dan klasifikasi. Ada beberapa tahap dalam membuat *decision tree* c4.5(6) yaitu:

1. Tentukan atribut yang akan dijadikan akar dengan menentukan nilai entropy terendah dan nilai gain tertinggi. Menentukan nilai entropy dengan rumus:

$$Entropy(y) = \sum_{i=1}^{n} -p_i \log_2 p_i \qquad (1)$$

2. Tentukan nilai gain, dengan rumus:

$$Gain\ (y, A) = Entropy\ (y) - \sum_{c\ \in nilai\ (A)} \frac{y_c}{y}\ entropy\ (y_c) \qquad (2)$$

3. Membuat cabang untuk masing-masing nilai.

4. Membagi setiap kasus menjadi cabang.
5. Mengulang proses untuk masing-masing cabang, sehingga semua kasus memiliki kelas yang samA

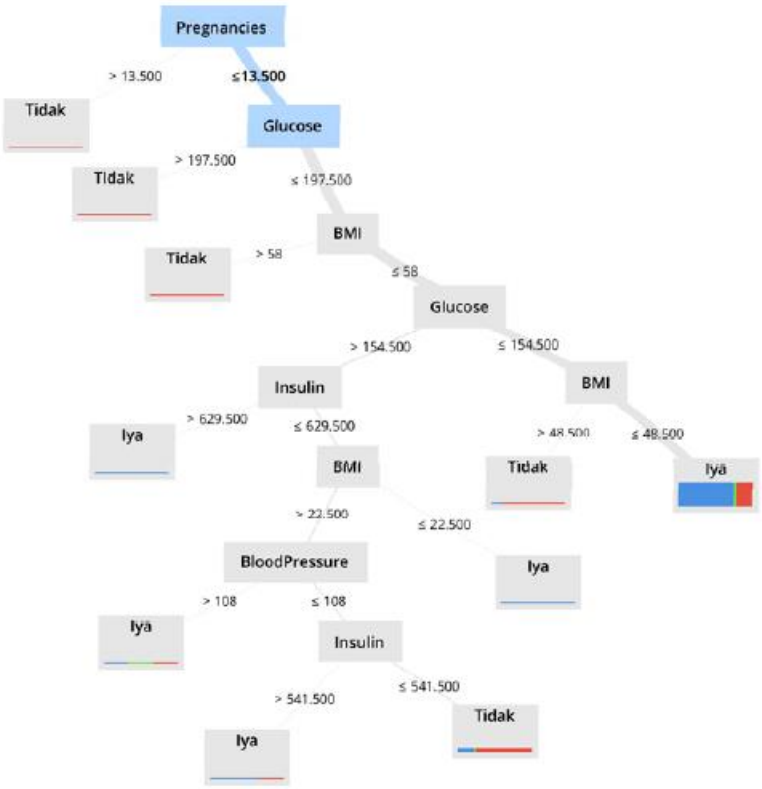
Evaluasi Performance

Evaluasi kinerja model klasifikasi yang dihasilkan, dilihat berdasarkan pada hasil pengujian objek yang diprediksi dengan benar dan salah. Model klasifikasi akan ditentukan nilai akurasi. Akurasi dalam klasifikasi menentukan persentase ketepatan record data yang di klasifikasikan secara benar(7). Hasil hitungan, ditabulasi kedalam bentuk confusion matrix yang memiliki jumlah nilai *true positif* (TP), *false negative* (FN) dan *true negative* (TN).

Instrumen Penelitian

Perangkat lunak yang digunakan untuk membuat model prediksi penyakit diabetes adalah Rapidminer.

Hasil dan Pembahasan



Gambar 2.
***Decision tree* klasifikasi penyakit diabetes menggunakan algoritma C4.5**

Decision tree klasifikasi penyakit diabetes pada gambar 2 dengan menggunakan algoritma C4.5 terdapat 9 rule, dengan jumlah class tidak sebanyak 4 rule dan 5 rule class iya dikategorikan penyakit diabetes.

Confusion Matrix *Decision tree* Algoritma C4.5

Tabel 2.
Confusion Matrik *Decision tree* Algoritma C4.5

	True (Iya)	True (Tidak)	True (Tidak)	Class Precision
Pred. Iya	480	25	174	70.69%
Pred. Tidak	0	0	0	0.00%
Pred. Tidak	25	4	60	67.42%
Class Recall	95.05%	0.00%	25.64%	

Pada Tabel 2 menunjukkan bahwa nilai *true positif* mengidap penyakit DM dengan nilai *precision* adalah 70.69% dan false negative 67.42% dari 768 wanita melahirkan. Metode algoritma C4.5 yang digunakan, maka model prediksi yang dihasilkan memiliki *class recall true negative* 95.05% dan *false negative* 25.64%.

Nilai akurasi dari model *decision tree* algoritma C4.5 untuk prediksi penyakit DM adalah 70.32%. Hasil akurasi yang diperoleh dengan menggunakan algoritma C4.5 dapat digunakan untuk membuat GUI atau aplikasi yang dapat digunakan untuk membantu dokter dan pasien DM dalam mendiagnosa penyakit DM.

Kesimpulan

Penedeteksian penyakit diabetes menjadi salah satu hal yang sangat penting dalam dunia medis. Dalam penelitian yang dilakukan adalah membuat model prediksi penyakit diabetes. Prediksi penyakit diabetes dilakukan dengan menggunakan algoritma *decision tree* C4.5. Akurasi model prediksi adalah 70.32% dengan 9 rule dengan 4 rule dengan class tidak dan 5 rule dengan class tidak. Penelitian yang dilakukan nantinya, dari hasil rule yang telah diperoleh dapat digunakan untuk perancangan aplikasi berbasis android untuk deteksi penyakit diabetes.

Daftar Pustaka

1. Perveen S, Shahbaz M, Guergachi A, Keshavjee K. Performance Analysis of Data Mining Classification Techniques to Predict Diabetes. Procedia Comput Sci [Internet]. 2016;82(March):115–21. Available from: http://dx.doi.org/10.1016/j.procs.2016.04.016
2. Wu H, Yang S, Huang Z, He J, Wang X. Type 2 diabetes mellitus prediction model based on data mining. Informatics Med Unlocked [Internet]. 2018;10(August 2017):100–7. Available from: https://doi.org/10.1016/j.imu.2017.12.006
3. Sisodia D, Sisodia DS. Prediction of Diabetes using Classification Algorithms. Procedia Comput Sci [Internet]. 2018;132(Iccids):1578–85. Available from: https://doi.org/10.1016/j.procs.2018.05.122
4. Rohman A, Suhartono V, Supriyanto C. Penerapan algoritma c4.5 berbasis. 2017;13:13–9.
5. Choubey D, Paul S, Kumar S, Kumar S. Classification of Pima indian diabetes dataset using naive bayes with genetic algorithm as an attribute selection. Commun Comput Syst [Internet]. 2017;451–5. Available from: http://www.crcnetbase.com/doi/10.1201/9781315364094-82
6. Santosa B, Umam A. Data Mining dan Big Data Analytics: Teori dan Implementasi Menggunakan Python & Apache Spark. 1st ed. Isa, editor. Yogyakarta: Penebar Media Pustaka; 2018.
7. Mochammad Yusa, Utami E, Luthfi ET. Evaluasi Performa Algoritma Klasifikasi *Decision tree* ID3 , C4,5 dan Cart pada Dataset Readmisi Pasien Diabetes. InfoSys J. 2016;4(1):23–34.