PROSIDING Seminar Nasional SISFOTEK ISSN 2597-3584 (Media Online)
Sistem Informasi dan Teknologi
website : www.seminar.iaii.or.id | email : seminar@iaii.or.id

Performansi K-NN, J48, Naive Bayes dan Regresi Logistik Sebagai Algoritma Pengklasifikasi Diabetes

Agung Mulyo Widodo[1] Yanathifal Salsabila Anggraeni[2], Nizirwan Anwar[3], Arief Ichwani[4] Binastya Anggara Sekti[5]
[1,3] Program Studi Teknik Informatika, Fakultas Ilmu Komputer, Universitas Esa Unggul
[2,4,5] Program Studi Sistem Informasi, Fakultas Ilmu Komputer, Universitas Esa Unggul
nizirwan.anwar@esaunggul.ac.id

Abstract

Diabetes is a chronic disease characterized by high blood sugar (glucose) levels. This disease is often found in adults who are elderly, but this disease can also attack people who are still young. Along with advances in machine learning technology to support decision makers, many predictive models are made of whether a person can be classified as diabetic or not by using certain algorithms. In this study, a prediction model was made whether a person is classified as diabetic or not, based on parameters/variables, namely weight, height, cholesterol levels, fasting sugar, non-fasting sugar, uric acid levels and gender. Prediction model is made using K-NN, J48 (based on decision tree), Naive Bayes and logistic regression classification algorithms. Then a performance analysis was carried out on the testing results of each of these algorithms, and it was found that the K-NN algorithm produced a prediction model with the highest accuracy compared to the three algorithms used in this study.

Keywords: Diabetes, K-NN, J48, Naive Bayes, Logistics regression

Abstrak

Diabetes adalah penyakit kronis yang ditandai dengan ciri-ciri berupa tingginya kadar gula (glukosa) darah. Penyakit ini sering kali ditemukan pada orang dewasa yang sudah lanjut usia, namun penyakit ini juga dapat menyerang orang yang tergolong masih muda. Seiring dengan kemajuan teknologi machine learning sebagai pendukung pembuat keputusan, banyak dibuat model prediksi apakah seseorang dapat diklasifikasi sebagai penderita diabetes atau tidak menderita diabetes dengan menggunakan algoritma-algoritma tertentu. Pada penelitian dilakukan pembuatan model prediksi apakah seseorang terklasifikasi sebagai penderita diabetes atau tidak, berdasarkan paramater/ variabel yaitu berat badan, tinggi badan, kadar kolesterol, gula saat puasa, gula saat tidak puasa, kadar asam urat dan juga jenis kelamin. Model prediksi dibuat dengan menggunakan algoritma-algoritma pengklaisifkasi K-NN, J48 (dengan dasar decision tree), naive bayes dan regresi logistic. Kemudia dilakukan analisis performansi terhadap hasil-hasil testing dari masing-masing algoritma-algoritma tersebut, dan diperoleh bahwa algoritma K-NN menghasilkan model prediksi dengan akurasi yang tertinggi dibandingkan ketiga algoritma yang digunakan dalam penelitian ini.

Kata kunci: Diabetes, K-NN, J48, Naive Bayes, Regresi logistik

1. Pendahuluan

Diabetes mellitus (atau biasa disebut diabetes saja) adalah penyakit kronis yang ditandai dengan tingginya kadar gula (glukosa) di dalam darah. Kondisi ini juga sering disebut sebagai penyakit gula atau kencing manis.Gula yang berada di dalam darah seharusnya diserap oleh sel-sel tubuh untuk kemudian diubah menjadi energi. Insulin adalah hormon yang bertugas untuk membantu penyerapan glukosa dalam sel-sel tubuh untuk diolah menjadi energi, sekaligus menyimpan sebagian glukosa sebagai cadangan energi.

Apabila terjadi gangguan pada insulin, seseorang berisiko tinggi mengalami diabetes. Data laporan WHO tahun 2003 menunjukkan hanya 50% pasien DM di negara maju mematuhi pengobatan yang diberikan. Pada DM yang tidak terkendali dapat terjadi komplikasi. Timbulnya komplikasi mempengaruhi kualitas hidup dan mempengaruhi perekonomian. Prevalensi diabetes mellitus di Indonesia pada tahun 2013 adalah sebesar 2,1%. Angka tersebut lebih tinggi dibandingkan dengan tahun 2007 (1,1%). Sebanyak 31 provinsi (93,9%) menunjukkan kenaikan prevalensi diabetes mellitus yang cukup berarti. [1]

Tingginya angka kejadian diabetes yang tinggi, dengan keadaan negara Indonesia yang merupakan negara berkembang dan memiliki populasi penduduk tinggi, membuat sulitnya berkonsultasi dengan tenaga medis untuk melakukan pemeriksaan. Seiring dengan perkembangan ilmu mengenai *machine learning*, maka kesulitan yang dihadapi dapat menjadi sebuah dorongan bagi teknologi informasi untuk mencari

solusi dengan tujuan memudahkan proses pengecekan atau pemeriksaan oleh tenaga medis terhadap para pasien diabetes. M*achine learning* (ML) pertama kali dikemukakan oleh beberapa ilmuwan matematika seperti Adrien Marie Legendre, Thomas Bayes dan Andrey Markov pada tahun 1920-an dengan mengemukakan dasar-dasar , *machine learning* dan konsepnya [2].

Kemudian, untuk mendukung keputusan apakah seseorang itu bisa divonis sebagai penderita diabetes atau tidak, maka dapat dilakukan dengan membuat sebuah model prediksi berdasarkan paramater/variabel seperti berat badan, tinggi badan, kadar kolesterol, gula saat puasa, gula saat tidak puasa, kadar asam urat dan juga jenis kelamin. Dalam membuat model prediksi tersebut dapat dilakukan dengan menggunakan metode statistik tradisional yaitu regresi logistik. Namun seiring dengan perkembangan teknologi komputasi dapat pula dilakukan dengan menggunakan algoritma pengklasifikasi pada *data mining*. *Data mining* adalah proses mencari pola atau informasi menarik dalam data terpilih dengan menggunakan teknik atau metode tertentu.

Pada penelitian ini sendiri akan dipilih algoritma K-NN (*K-Nearest Neighbors*), J48 (dengan dasar *decision tree*), *Naive Bayes* dan Regresi Logistik. Dari beberapa algoritma yang digunakan, nantinya akan dibandingkan performansinya sehingga dapat diketahui mana algoritma yang akurasinya paling tepat dalam melakukan klasifikasi terhadap penentuan apakah seseorang dapat dinyatakan menderita diabetes mellitus atau tidak. Adapun output yang akan dihasilkan dari penelitian ini sendiri yaitu berupa sebuah model prediksi apakah sesorang terklasifikasi sebagai penderita diabetes atau tidak menderita diabetes.

Penelitian yang dilakukan oleh Sharavan dan Gayathri pada [3] menggunakan Algoritma J48 dengan dasar *decision tree* untuk melakukan klasifikasi. Adapun hasil yang didapatkan adalah bahwa algoritma KNJ48 adalah metode klasifikasi yang efisien untuk dataset demam berdarah dalam penelitian ini dan memiliki tingkat akurasi yang baik dan bekerja untuk waktu yang lebih rendah dibandingkan dengan algoritma J48 dan algoritma *random forest* lainnya. Banu, et al. [4] menggunakan *Algoritma zero R, one R, decision stump* dan J48 dalam penelitiannya untuk memprediksi kanker payudara dan ukuran kinerja dapat dianalisis melalui *confusion matrix*. Hasil akhir dari penelitian ini menunjukan bahwa kinerja pengklasifikasi dievaluasi melalui konfusi matriks dalam hal akurasi. Algoritma J48 memberikan hasil 75.52% yang memberikan akurasi yang lebih baik dibandingkan dengan Algoritma lain.

Keleş [5] dalam penelitiannya menggunakan algoritma Bagging, IBk, *random committee, random forest*, dan *simpleCART*. Hasil dari penelitiannya sendiri mengungkapkan bahwa tingkat akurasi yang tinggi dari algoritme tersebut menyarankan bahwa tumor kanker payudara memang dapat diidentifikasi secara non-invasif, dengan biaya rendah dan tanpa membuat pasien terpapar radiasi berbahaya, dengan menggunakan klasifikasi *data mining* algoritma, seperti Bagging, IBk, *random committee, random forest*, dan *simpleCART*, dengan tingkat akurasi lebih tinggi dari 90.0 persen.

Penelitian yang dilakukan oleh Hobeqi [6] menggunakan metode *data mining* dengan menggunakan algoritma 4.5. Di mana algoritma 4.5 digunakan sebagai teknik pembelajaran pada *data mining* tersebut. Dari hasil uji yang dilakukan, metode pohon keputusan (*decision tree*) yang diproses dengan *software Rapidminer* lebih efektif dan fleksibel jika digunakan pada proses pengklasifikasian calon nasabah pada PT. Cipta Daya *Resoure inhouse* BNI46. Disamping itu, pemilihan variabel (atribut kondisi dan attribut keputusan) yang akan digunakan dalam menentukan sebuah klasifikasi juga sangat mempengaruhi *rule* atau *knowledge* yang dihasilkan.

Yustanti [7] menggunakan metode gabungan antara tahapan data mining yang dikenal dengan istilah *Cross-Industry Standard Process for Data Mining* (CRISP-DM) dan metode pengembangan perangkat lunak *Waterfall Model.* Berdasarkan penelitian ini diperoleh hasil bahwa aplikasi yang dihasilkan memiliki tingkat pemrosesan yang cukup lama karena data testing (data baru) dibandingan satu persatu dengan data training (data lama). Tingkat presisi dari prediksi yang dihasilkan sekitar 80%.

2. Metode Penelitian

Dalam penelitian ini, penulis menggunakan objek penelitian yaitu penggunaan algoritma klasifikasi K-NN (*K-Nearest Neighbor*), J48, Naive Bayes dan Regresi Logistik pada sebuah *database* diabetes yang didapatkan dari internet yaitu website *kaggle.com.* Dalam proses pengolahan data, dalam penelitian ini akan dilakukan proses klasifikasi database dengan menggunakan tool yaitu aplikasi WEKA (*Waikato Environment for Knowledge Analysis*). Tool WEKA dianggap sebagai tool yang tepat dalam melakukan proses klasifikasi terhadap algoritma yang akan digunakan. Saat melakukan proses klasifikasi pada tool WEKA, dataset diabetes akan diubah menjadi format csv. yang mana merupakan format yang tersedia untuk dapat terbaca pada WEKA. Beberapa fitur unggulan yang dimiliki oleh WEKA yaitu [8] :

1. Classification
2. Regression
3. Clustering
4. Association Rules
5. Visualization
6. Data Preprocessing

2.1 Algoritma K-NN *(K-Nearest Neighbor)*

Algoritma *k-nearest neighbor* (K-NN atau KNN) adalah sebuah metode untuk melakukan klasifikasi terhadap objek berdasarkan data pembelajaran yang jaraknya paling dekat dengan objek tersebut, Ketepatan algoritma K-NN ini sangat dipengaruhi oleh ada atau tidaknya fitur-fitur yang tidak relevan, atau jika bobot fitur tersebut tidak setara dengan relevansinya terhadap klasifikasi. Riset terhadap algoritma ini sebagian besar membahas bagaimana memilih dan memberi bobot terhadap fitur, agar performa klasifikasi menjadi lebih baik, K-NN ini juga tergolong dalam contoh teknik *lazy learning* dimana merupakan sebuah tekning yang menunggu hingga pertanyaan datang sehingga sama dengan *data training*.

2.2 Algoritma J48 (Dengan Dasar Decision Tree)

Algoritma *Decision tree J48* merupakan implementasi dari algoritma C4.5 yang memproduksi *decision tree*. Ini merupakan standar algoritma yang digunakan dalam machine learning. *Decision tree* merupakan salah satu algoritma klasifikasi dalam data mining. Algoritma klasifikasi merupakan algoritma yang secara induktif dalam pembelajaran dalam mengkonstruksikan sebuah model dari dataset yang belum diklasifikasikan [9]. Setiap data dari item berdasarkan dari nilai dari setiap atribut. Klasifikasi dapat dilihat sebagai *mapping* dari sekelompok set dari atribut dari kelas tertentu.
Dasar dari algoritma ini adalah untuk membagi data ke dalam beberapa bagian berdasarkan nilai atribut dari item yang ada pada training dataset. Algoritma J48 dapat melakukan klasifikasi baik melalui decision tree ataupun rules yang diperoleh dari pohon tersebut [10].

Decision tree mengklasifikasikan data yang diberikan menggunakan nilai dari atribut. Dataset dengan atribut pilihan kemudian diklasifikasikan menggunakan *decision tree* J48. J48 adalah salah satu jenis *classifier* pada metode klasifikasi dalam *data mining* dan bagian dari C4.5 *decision tree* yang sederhana [11]. C4.5 membangun sebuah pohon keputusan berdasarkan pada seperangkat input data yang berlabel [12]. Pohon keputusan adalah model prediksi menggunakan struktur pohon atau struktur berhirarki. Secara umum algoritma C4.5 untuk membangun pohon keputusan adalah sebagai berikut:

1. Pilih atribut sebagai akar
2. Buat cabang untuk tiap-tiap nilai
3. Bagi kasus dalam cabang
4. Ulangi proses.

2.3 Algoritma Naive Bayes

Naïve Bayes Classifier merupakan sebuah metode klasifikasi yang berakar pada teorema Bayes. Metode pengklasifikasian dengan menggunakan metode probabilitas dan statistik yg dikemukakan oleh ilmuwan Inggris Thomas Bayes, yaitu memprediksi peluang di masa depan berdasarkan pengalaman di masa sebelumnya sehingga dikenal sebagai Teorema Bayes. Ciri utama dari *Naïve Bayes Classifier* ini adalah asumsi yang sangat kuat (naïf) akan independensi dari masing-masing kondisi / kejadian [13].

Pengklasifikasi Bayes didasari oleh teorema bayes yang ditemukan oleh Thomas Bayes pada abad ke-18. Dalam studi pembandingan algoritma klasifikasi telah ditemukan *simple bayesian* atau yang biasa dikenal dengan *Naïve Bayes classifier*. *Naïve Bayes classifier* menunjukkan akurasi dan kecepatan yang tinggi bila diterapkan pada *database* yang besar. Metode ini sering digunakan dalam menyelesaikan masalah dalam bidang mesin pembelajaran karena metode ini dikenal memiliki tingkat akurasi yang tinggi dengan perhitungan sederhana [14].

2.4 Metode Regresi Logistik

Regresi logistik (*logistic regression*) adalah bagian dari analisis regresi yang digunakan ketika variabel dependen (respon) merupakan variabel dikotomi yaitu yang terdiri dari dua nilai yang mewakili muncul atau tidaknya suatu kejadian yang biasanya diberi angka 0 atau 1. Tidak seperti regresi linier biasa, regresi logistik tidak mengasumsikan hubungan antara variabel independen dan dependen secara linier. Regresi logistik adalah sebuah pendekatan untuk membuat model prediksi seperti halnya regresi linear atau yang biasa disebut dengan istilah *Ordinary Least Squares* (OLS) *regression.* Perbedaannya adalah pada regresi logistik, peneliti memprediksi variabel terikat yang berskala dikotomi. Skala dikotomi yang dimaksud adalah skala data nominal dengan dua kategori, misalnya: ya dan tidak, baik dan buruk atau tinggi dan rendah [15]. Asumsi regresi logistik antara lain:

1. Regresi logistik tidak membutuhkan hubungan linier antara variabel independen dengan variabel dependen.
2. Variabel independen tidak memerlukan asumsi multivariate normality.
3. Asumsi homokedastisitas tidak diperlukan.
4. Variabel bebas tidak perlu diubah ke dalam bentuk metrik (interval atau skala ratio).
5. Variabel dependen harus bersifat dikotomi (2 kategori, misal: tinggi dan rendah atau baik dan buruk).
6. Variabel independen tidak harus memiliki keragaman yang sama antar kelompok variable.
7. Kategori dalam variabel independen harus terpisah satu sama lain atau bersifat eksklusif.
8. Sampel yang diperlukan dalam jumlah relatif besar, minimum dibutuhkan hingga 50 sampel data untuk sebuah variabel prediktor (independen).
9. Dapat menyeleksi hubungan karena menggunakan pendekatan non linier log transformasi untuk memprediksi odds ratio. Odd dalam regresi logistik sering dinyatakan sebagai probabilitas.

3. Hasil dan Pembahasan

Dalam penelitian ini, database diabetes yang didapatkan dari internet yaitu website *kaggle.com*. Data set ini berisi 520 instances dan 17 atribut. Adapun data set beserta atribut yang akan digunakan terdapat pada Tabel 1.

Dataset diabetes yang didapatkan kemudian dikonversi terlebih dahulu menjadi format csv sehingga dapat kompatibel dan terbaca oleh *tool* WEKA. Dalam memulai proses belajar pada WEKA, harus dipastikan terlebih dahulu bahwa atribut kelas sudah sesuai. Sebelum memulai pembelajaran, terdapat beberapa parameter tes sebagai berikut:

1) *Use training set*: model mengevaluasi seberapa baik memprediksi kelas dari semua contoh data. Ini agak berbahaya karena dapat menyebabkan akurasi menjadi tinggi secara artifisial.
2) *Supplied test set*: model mengevaluasi seberapa baik prediksi kelas instance data uji yang dimuat dari file terpisah. Data uji ini berupa ARFF dan tentunya harus sama atributnya dengan data latih.
3) Cross-validation: model dievaluasi dengan validasi silang, dengan jumlah lipatan sesuai dengan input pengguna (default: 10).

Percentage split: model mengevaluasi seberapa baik prediksi kelas instance yang merupakan% tertentu dari semua data (*hold-out*). Besar % menurut masukan pengguna (default: 66). Dalam penelitian ini, akan digunakan opsi tes *cross-validation* dengan *fold* 10.

Tabel 1.Data Set Diabetes

No	Attribute	Data Type	Data Mising
1	Age	Nominal	0
2	Gender	Nominal	0
3	Polyuria	Nominal	0
4	Polydipsia	Nominal	0
5	Sudden Weight Loss	Nominal	0
6	Weakness	Nominal	0
7	Polyphagia	Nominal	0
8	Genital Thrush	Nominal	0
9	Visual Blurring	Nominal	0
10	Itching	Nominal	0
11	Irritability	Nominal	0
12	Delayed Healing	Nominal	0
13	Partial Paresis	Nominal	0
14	Muscle Stiffness	Nominal	0
15	Alopecia	Nominal	0
16	Obesity	Nominal	0
17	Class	Nominal	0

Setelah memulai dengan menekan tombol Start, maka didapatkan hasil pembelajaran sebagai berikut:

1) *Run Information*: merupakan hasil pembelajaran dari algoritma yang digunakan yang berisi informasi mengenai jumlah *instances*, *attributes*, dan juga hasil *confusion matrix*.
2) *Classifier model (full training set):* merupakan model pembelajaran pada WEKA yang dihasilkan dari semua data *training set* dalam bentuk teks.
3) *Summary*: merupakan sebuah rangkuman hasil yang menunjukkan akurasi dari model pengklasifikasi untuk memprediksi kelas sebenarnya dari setiap *instance* menurut opsi pengujian. Untuk opsi uji validasi silang 10 kali lipat yang dipilih, ada 520 instans yang diuji dalam 10 iterasi (*fold*).
4) *Detailed accuracy by class*: merupakan ukuran kinerja yang lebih rinci di tingkat kelas.
5) *Confusion matrix*: merupakan hasil matriks yang menunjukkan berapa banyak instance yang diprediksi untuk setiap kelas dan jumlah instance yang sesuai dengan label aktual (per baris).

Penelitian ini akan mengevaluasi performance algoritma dari *Machine Learning (ML*) (khususnya *supervised learning*), maka kita menggunakan acuan *Confusion Matrix*. *Confusion Matrix* merepresentasikan prediksi dan kondisi sebenarnya (aktual) dari data yang dihasilkan oleh algoritma ML. Berdasarkan *Confusion Matrix*, kita bisa menentukan *Accuracy*, *Precission*, *Recall* dan Specificity. Tujuannya sendiri adalah untuk mengukur dan memperbandingkan dari masing-masing algoritma, algoritma mana kah yang paling baik performansi nya. Adapun cara perhitungan untuk akurasi (*accuracy*) adalah sebagai berikut:

$$\text{Akurasi} = \frac{TP+TN}{TP+TN+FP+FN} * 100\% \quad (1)$$

Perhitungan untuk mengukur perbandingan algoritma tersebut kemudian dapat dilanjutkan dengan mengukur presisi (precission) dengan rumus sebagai berikut:

$$\text{Presisi} = \frac{TP}{FP+TP} * 100\% \quad (2)$$

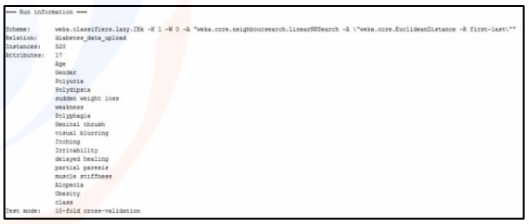
Setelah mengetahui cara perhitungan akurasi dan presisi, maka selanjutnya akan diketahui mengenai perhitungan *recall* yaitu sebagai berikut:

$$\text{Recall} = \frac{TP}{FN+TP} * 100\% \quad (3)$$

Pada cara perhitungan rumus di atas, terdapat beberapa singkatan yaitu TP, TN, FP dan FN dimana TP adalah singkatan dari True Positive, TN singkatan dari True Negative, FP singkatan dari False Positive dan FN singkatan dari False Negative. Langkah selanjutnya adalah dengan melakukan klasifikasi dataset yang ada dengan *tool* WEKA pada setiap algoritma dan hasil yang didapatkan adalah sebagai berikut:

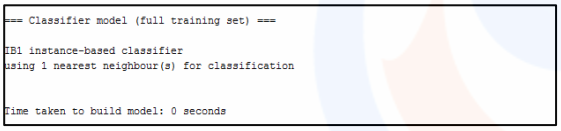
3.1 Algoritma K-NN (K-Nearest Neighbor)

Setelah melakukan proses pembelajaran menggunakan tool WEKA terhadap algoritma K-NN, maka didapatkan hasil berupa *run information* sebagai berikut:



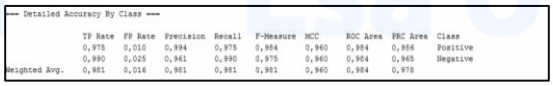
Gambar 1. Run Information K-NN

Selanjutnya, didapatkan pula hasil berupa *Classifier* Model sebagai berikut:



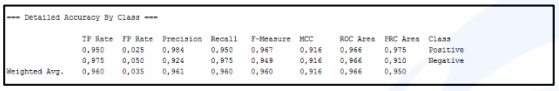
Gambar 2. Classifier Model *K-NN*

Hasil *summary* K-NN dibawah ini menunjukan bahwa terdapat 510 kejadian (98,07%) yang diprediksi dengan benar oleh kelas, dan 10 kejadian lainnya (1,92%) yang salah diprediksi.



Gambar 3. Summary Model *K-NN*

Dibawah ini merupakan hasil proses pembelajaran berupa *detailed accuracy* menggunakan algoritma K-NN:



Gambar 4. *Detailed Accuracy* Model *K-NN*

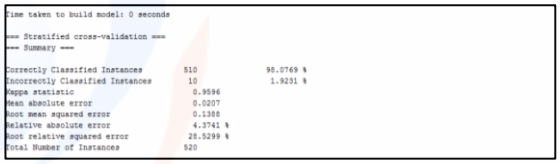
Hasil *confusion matrix* dari klasifikasi menggunakan algoritma K-NN dapat ditunjukan sebagai berikut:

Gambar 5. *Confusion Matrix* Model *K-NN*

Dari hasil klasifikasi WEKA dengan menggunakan Algoritma K-NN atau *K Nearest Neighbors*, didapatkan hasil *True Positive (*TP*)* yaitu 312, *True Negative* (TN) yaitu 198, *False Positive* (FP) yaitu 8 dan *False Negative* (FN) yaitu 2. Dari hasil klasifikasi tersebut, didapatkan pula hasil *Precision Positive* yaitu 0.994, *Precision Negative* yaitu 0.961 dan *Weighted Average* nya yaitu 0.981, sedangkan hasil *Recall Positive* didapatkan sebesar 0.975, *Recall Negative* yaitu 0.990, dan *Weighted Average* nya yaitu 0.981.
Selain didapatkan nilai *True Positive (TP), True Negative (TN), False Positive (FP), False Negative (FN), Precision Positive, Precision Negative, Recall Positive* dan *Recall Negative*, ada pun nilai *Accuracy, Precision, Recall* dan *F-measure* dari hasil klasifikasi tersebut dengan hasil *Accuracy 98%, Precision 97.5%, Recall 99.4%* dan *F-measure 0.985.*

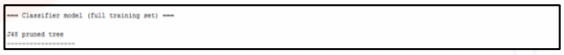
3.2 Algoritma J48

Klasifikasi selanjutnya dilakukan dengan menggunakan algoritma J48, dari proses pembelajaran, maka didapatkan hasil *run information* sebagai berikut:

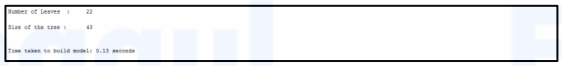


Gambar 6. *Run Information* Model *J48*

Hasil *classifier model* dari algoritma J48 adalah sebagai berikut:

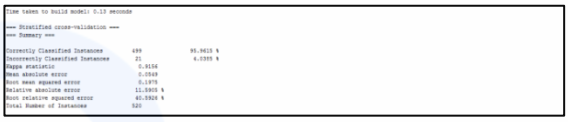


Gambar 7. *Classifier* Model *J48 (1)*



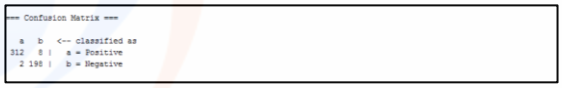
Gambar 8. . *Classifier* Model *J48 (2)*

Summary dari hasil klasifikasi algoritma J48 menunjukan bahwa terdapat 499 kejadian (95,96%) yang diprediksi dengan benar oleh kelas, dan 21 kejadian lainnya (4.03%) yang salah diprediksi.



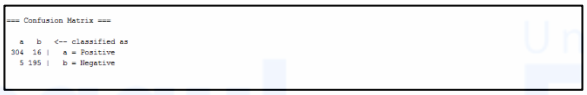
Gambar 9. Summary Model *J48 (2)*

Berikut merupakan hasil *detailed accuracy* dari klasifikasi algoritma J48:



Gambar 10. *Detailed Accuracy* Model *J48*

Selanjutnya didapatkan pula hasil *confusion matrix* sebagai berikut:



Gambar 11. *Confusion MAtrix* Model *J48*

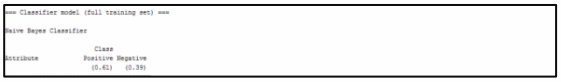
Dari hasil klasifikasi WEKA dengan menggunakan Algoritma J48 , didapatkan hasil *True Positive* (TP) yaitu 304, *True Negative* (TN) yaitu 195, *False Positive* (FP) yaitu 16 *dan False Negative* (FN) yaitu 5. Dari hasil klasifikasi tersebut, didapatkan pula hasil *Precision Positive* yaitu 0.984, *Precision Negative* yaitu 0.924 dan *Weighted Average* nya yaitu 0.961, sedangkan hasil *Recall Positive* didapatkan sebesar 0.950, *Recall Negative* yaitu 0.975, dan *Weighted Average* nya yaitu 0.960. Selain didapatkan nilai *True Positive (TP), True Negative (TN), False Positive (FP), False Negative (FN), Precision Positive, Precision Negative, Recall Positive* dan *Recall Negative*, ada pun nilai *Accuracy, Precision, Recall* dan *F-measure* dari hasil klasifikasi tersebut dengan hasil *Accuracy* 95.9%, *Precision* 95%, *Recall* 98.3%, dan *F-measure* 0.965.

3.3 Algoritma Naive Bayes

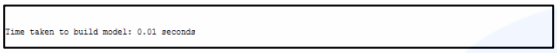
Proses klasifikasi yang ketiga dilanjutkan dengan menggunakan algoritma *naive bayes*, berdasarkan proses klasifikasi tersebut maka didapatkan hasil run information sebagai berikut:

Gambar 12. *Run Information* Model *Naïve Bayes*

Hasil berikutnya yang didapatkan adalah *classifier model* yang tertera sebagaimana gambar berikut:

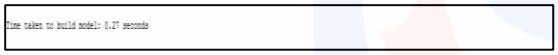


Gambar 13. *Classifier* Model *Naïve Bayes (1)*



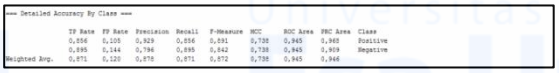
Gambar 14. *Classifier* Model *Naïve Bayes (2))*

Selanjutnya didapatkan hasil berupa *summary* dari proses klasifikasi algoritma *Naive Bayes* yang menunjukan bahwa terdapat 453 kejadian (87,11%) yang diprediksi dengan benar oleh kelas, dan 67 kejadian lainnya (12,88%) yang salah diprediksi.



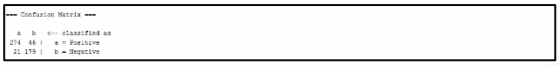
Gambar 15. *Summary* Model *Naïve Bayes*

Detailed *accuracy* yang didapatkan dari hasil klasifikasi algoritma *Naive Bayes* adalah sebagai berikut:



Gambar 16. *Detailed Accuracy* Model *Naïve Bayes*

Hasil *confusion matrix* dari proses klasifikasi algoritma *Naive Bayes* dapat ditunjukkan sebagai berikut:



Gambar 17. *Confusion Matrix* Model *Naïve Bayes*

Dari hasil klasifikasi WEKA dengan menggunakan Algoritma *Naive Bayes* , didapatkan hasil *True Positive* (TP) yaitu 274, *True Negative* (TN) yaitu 179, *False Positive* (FP) yaitu 46 dan *False Negative* (FN) yaitu 21. Dari hasil klasifikasi tersebut, didapatkan pula hasil *Precision Positive* yaitu 0.929, *Precision Negative* yaitu 0.796 dan *Weighted Average* nya yaitu 0.878, sedangkan hasil *Recall Positive* didapatkan sebesar 0.856, *Recall Negative* yaitu 0.895, dan *Weighted Average* nya yaitu 0.871.

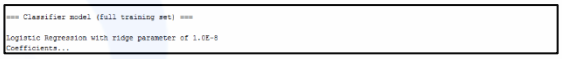
Selain didapatkan nilai *True Positive* (TP), *True Negative* (TN), *False Positive* (FP), *False Negative* (FN), *Precision Positive*, *Precision Negative*, *Recall Positive* dan *Recall Negative*, ada pun nilai *Accuracy*, *Precision*, *Recall* dan *F-measure* dari hasil klasifikasi tersebut dengan hasil *Accuracy* 87.11%, *Precision* 65.2%, *Recall* 92.88%, dan *F-measure* 0.766.

3.4 Algoritma Regresi Logistik

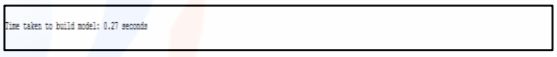
Klasifikasi yang terakhir dilakukan dengan menggunakan algoritma regresi logistik, dimana didapatkan hasil berupa *run information* sebagai berikut:

Gambar 18. *Run Information* Model *Regresi Logistik*

Kemudian didapatkan pula hasil *classifier model* dari klasifikasi menggunakan algoritma regresi logistik, yaitu:

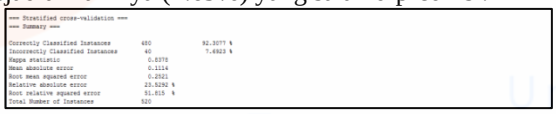


Gambar 19. *Classfier* Model *Regresi Logistik (1)*



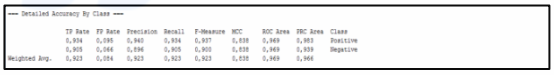
Gambar 20. *Classfier* Model *Regresi Logistik (2)*

Hasil *summary* klasifikasi algoritma regresi logistik menunjukan bahwa terdapat 480 kejadian (92.30%) yang diprediksi dengan benar oleh kelas, dan 40 kejadian lainnya (7.69%) yang salah diprediksi.



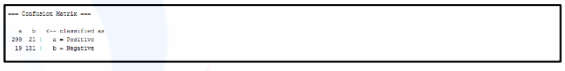
Gambar 21. *Summary* Model *Regresi Logistik*

Detailed *accuracy* dari algoritma regresi logistik dapat ditunjukan sebagai berikut:



Gambar 22. *Detailed Accuracy* Model *Regresi Logistik*

Hasil *confusion matrix* dari proses klasifikasi algoritma regresi logistik juga dapat ditunjukan pada gambar dibawah ini:



Gambar 23. *Confusion Matrix* Model *Regresi Logistik*

Dari hasil klasifikasi WEKA dengan menggunakan algoritma Regresi Logistik , didapatkan hasil *True Positive* (TP) yaitu 299, *True Negative* (TN) yaitu 181, *False Positive* (FP) yaitu 21 *dan False Negative* (FN) yaitu 19. Dari hasil klasifikasi tersebut, didapatkan pula hasil *Precision Positive* yaitu 0.940, *Precision Negative* yaitu 0.896 dan *Weighted Average* nya yaitu 0.923,

sedangkan hasil *Recall Positive* didapatkan sebesar 0.934, *Recall Negative* yaitu 0.905, dan *Weighted Average* nya yaitu 0.923. Selain didapatkan nilai *True Positive (*TP*), True Negative (*TN*), False Positive (*FP*), False Negative (*FN*), Precision Positive, Precision Negative, Recall Positive* dan *Recall Negative*, ada pun nilai *Accuracy*, *Precision*, *Recall* dan *F-measure* dari hasil klasifikasi tersebut dengan hasil *Accuracy* 92.3%, *Precision* 93.4%, *Recall* 94.3%, *dan F-measure* 0.939.

3.5 Hasil Penelitian

Menentukan diagnosa penyakit diabetes bukanlah hal yang mudah terutama dengan tingginya tingkat gejala yang di alami setiap pasien.

Tabel 2. Tabel Perbandingan

No	Algoritma	Accuracy	Precision	Recall	F-1 measure
1	*K-NN*	98%	97.5%	99.4%	98.5%
2	*J48*	95.5%	95%	98.3%	96.5%
3	*Naïve Bayes*	87.11%	65.2%	92.88%	76.6%
4	*Regresi Logistik*	92.3%	93.4%	94.3%	93.39%

Seiring berkembangnya waktu, maka didapatkan sebuah teknik *data mining* yang dapat digunakan dalam proses pengambilan keputusan apakah seorang pasien dapat di diagnosa mengidap diabetes atau tidak. Dalam penelitian ini, penulis telah menerapkan teknik klasifikasi *data mining* dengan algoritma K-NN, J48, *Naive Bayes*, dan Regresi Logistik yang digunakan untuk mengklasifikasikan penyakit diabetes. Kinerja pengklasifikasi dievaluasi melalui konfusi matriks. Berdasarkan perhitungan mengenai akurasi, *precision*, *recall* dan *f-measure* dari masing-masing algoritma dan metode, maka didapatkan hasil yang dinyatakan dalam tabel 2.

Berdasarkan hasil perbandingan performansi algoritma yang tertera pada tabel di atas, maka dapat disimpulkan bahwa Algoritma K-NN memberikan tingkat akurasi yang paling baik daripada Algoritma lainnya yaitu dengan akurasi sebesar 98%.

4. Kesimpulan

Diagnosis suatu penyakit bukanlah pekerjaan yang mudah dan cepat untuk dilakukan, salah satunya terhadap penyakit diabetes. Seiring perkembangan teknologi, diagnosis penyakit diabetes bisa lebih dimudahkan dengan adanya *machine learning* dengan memanfaatkan algoritma klasifikasi *machine learning* yang bisa membantu untuk menentukan diagnosis pasien diabetes. Berdasarkan proses klasifikasi dengan menggunakan algoritma K-NN, J48, *Naive bayes* dan Regresi logistik pada aplikasi WEKA dengan beberapa informasi dan hasil *confusion matrix* yang diperoleh, dari keempat algoritma K-NN, J48, *Naive Bayes* dan Regresi logistik, disimpulkan bahwa algoritma K-NN memiliki performansi yang paling baik dalam menentukan diagnosis penyakit diabetes yaitu dengan memberikan akurasi sebesar 98% dibandingkan algoritma lainnya.

Daftar Rujukan

[1] Hestiana, Dita Wahyu. 2017. Faktor-Faktor Yang Berhubungan Dengan Kepatuhan Dalam Pengelolaan Diet Pada Pasien Rawat Jalan Diabetes Mellitus Tipe 2 Di Kota Semarang. Journal Of Health Education, 2(2), 139-140.

[2] Takdirillah, Robby, 2020. Apa itu machine learning? beserta pengertian dan cara kerjanya. Dicoding. Tersedia di: https://www.dicoding.com/blog/machine-learning-adalah/. Diakses tanggal 26 Desember 2020.

[3] Saravana, N, V.Gayathri, 2018. *Performance and Classification Evaluation of J48 Algorithm and Kendall's Based J48 Algorithm (KNJ48)*. IJCTT Journal, 26, 75-80.

[4] Banu, G. Rasitha, Prakash, Illham Bashier, Summera, 2017. *Applications of Data Mining Classification Techniques on Predicting Breast Cancer Disease*. International Journal of Latest Trends in Engineering and Technology, 8, 322-325.

[5] Keleş, Mümine Kaya. 2019. *Breast Cancer Prediction and Detection Using Data Mining Classification Algorithms: A Comparative Study*. Tehnički vjesnik, 26, 150-154.

[6] Honesqi, Hanggi Dwifa. 2017. Klasifikasi *Data Mining* Untuk Menentukan Tingkat Persetujuan Kartu Kredit, Jurnal TEKNOIF, 5, 58-61.

[7] Yustanti, Wiyli, 2012. Algoritma *K-Nearest Neighbour* untuk Memprediksi Harga Jual Tanah. Jurnal Matematika, Statistika dan Komputasi, 9.

[8] Suhartono, Derwin, 2018. WEKA: Software untuk Memahami Konsep *Data Mining*. Binus University. https://socs.binus.ac.id/2018/11/29/weka-software-untuk-memahami-konsep-data-mining/. Diakses tanggal 15 November 2020.

[9] Situmorang, Liswati. 2019. Analisa Kelayakan Usaha Mikro Kecil Dan Menengah (UMKM) Untuk Produk Krasida Pada PT. Pegadaian Menggunakan Metode J48. KOMIK (Konferensi Nasional Teknologi Informasi dan Komputer), 3, 6-9.

[10] Kaunang, Fergie Joanda. 2018. Penerapan Algoritma J48 Decision Tree Untuk Analisis Tingkat Kemiskinan di Indonesia. Cogito Smart Journal, 4(2) 351.

[11] J N. S. Diwandari and N. A. Setiawan. 2019. Perbandingan Algoritma J48 dan NBTREE Untuk Klasiifikasi Diagnosa Penyakit Pada Soybean". *Seminar Nasional Teknologi Informasi dan Komunikasi (SENTIKA 2015)*. 208.

[12] Agustiani, Sarifah, Ali Mustopa, Andi Saryoko, Windu Gata, Siti Khotimatul Wildah. 2020. Penerapan Algoritma J48 Untuk Deteksi Penyakit Tiroid Paradigma. Jurnal Informatika dan Komputer, 22, 154-155.

[13] Widianto, Mochammad Haldi, 2019. Algoritma Naive Bayes. Binus University. Tersedia di: https://binus.ac.id/bandung/2019/12/algoritma-naive-bayes/. Diakses tanggal 22 Agustus 2021.

[14] Handayani, Fitri, Feddy Setio Pribadi. 2015. Implementasi Algoritma Naive Bayes Classifier dalam Pengklasifikasian Teks Otomatis Pengaduan dan Pelaporan Masyarakat melalui Layanan Call Center 110. Jurnal Teknik Elektro, 7, 20-22,

[15] Hidayat, Anwar. Regresi Logistik. Statistikan. 2015. https://www.statistikian.com/2015/02/regresi-logistik.html. Diakses tanggal 5 November 2020.