

Penerapan Metode Klasifikasi K-Nearest Neighbor pada Dataset Penderita Penyakit Diabetes

Andi Maulida Argina^{a,1}

^a Universitas Muslim Indonesia, Jl. Urip Sumoharjo KM.5, Makassar 90231, Indonesia

¹ andimaulidaargina@gmail.com

INFORMASI ARTIKEL	ABSTRAK
Diterima : 10 – 05 – 2020 Direvisi : 25 – 06 – 2020 Diterbitkan : 31 – 07 – 2020 Kata Kunci: klasifikasi k-nearest neighbor analisis performa diabetes dataset	Diabetes adalah penyakit yang berlangsung lama atau kronis serta ditandai dengan kadar gula (glukosa) darah yang tinggi atau di atas nilai normal. Jika diabetes tidak dikontrol dengan baik, Pengujiannya berbagai metode pada sebuah dataset merupakan salah satu cara dalam penetapan metode klasifikasi yang tepat, masalah yang diangkat pada penelitian ini adalah bagaimana mengukur performa metode klasifikasi dalam mengelola dataset penderita diabetes. Metode yang digunakan yaitu algoritma K-Nearest Neighbor (KNN), dimana merupakan sebuah metode untuk melakukan klasifikasi terhadap objek berdasarkan data pembelajaran yang jaraknya paling dekat dengan objek tersebut. Pada hasil akhir penelitian ini, telah dihitung akurasi tertinggi 39% pada K=3, presisi tertinggi 65% pada K=3 dan K=5, <i>recall</i> tertinggi 36% pada K=3, dan <i>F-Measure</i> tertinggi 46% pada K=3. 

I. Pendahuluan

Diabetes adalah suatu penyakit metabolik yang diakibatkan oleh meningkatnya kadar glukosa atau gula darah. Gula darah sangat vital bagi kesehatan karena merupakan sumber energi yang penting bagi sel-sel dan jaringan. Jika tidak dikelola dengan baik, diabetes dapat menyebabkan terjadinya berbagai komplikasi, seperti penyakit jantung koroner, stroke, obesitas, serta gangguan pada mata, ginjal, dan saraf.

Terdapat banyak metode klasifikasi dalam *supervised learning* pada *machine learning*, diantaranya adalah *K-Nearest Neighbor* (knn), *Naive Bayes Classifier* (nbc), *Support Vector Machine* (svm), *Neural Network* (nn), *Random Forest Classifier* (rfc), *Ada Boost Classifier* (abc), serta *Quadratic Discriminant Analysis* (qda). Metode tersebut memiliki kelebihan serta kekurangannya masing-masing, salah satu faktor menjadi keunggulan metode klasifikasi tersebut dilihat dari bagaimana metode tersebut mengolah objek dataset, K-Nearest Neighbor atau KNN adalah algoritma yang berfungsi untuk melakukan klasifikasi suatu data berdasarkan data pembelajaran (*train data sets*), yang diambil dari K tetangga terdekatnya (*nearest neighbors*).

Pada penelitian ini penulis menggunakan metode KNN untuk menghitung akurasi, presisi, *recall*, dan *F-Measure* berdasarkan nilai K. Tahap yang dilakukan pada penelitian ini adalah *splitting* data *training* dan data *testing*, menerapkan metode klasifikasi knn, serta menghitung performa metode yang akan diuji.

II. Metode

A. Data Mining

Data mining adalah proses yang menggunakan statistik, matematika, kecerdasan buatan, dan machine learning untuk mengekstraksi dan mengidentifikasi informasi yang bermanfaat.

Data Mining didefinisikan sebagai proses penemuan pola dalam data. Berdasarkan tugasnya, data mining dikelompokkan menjadi deskripsi, estimasi, prediksi, klasifikasi, clustering dan asosiasi. Proses dalam tahap data mining terdiri dari tiga langkah Utama, yaitu data Preparation Pada langkah ini, data dipilih, dibersihkan, dan dilakukan preprocessed mengikuti pedoman dan knowledge dari ahli domain yang menangkap dan mengintegrasikan data internal dan eksternal ke dalam tinjauan organisasi secara menyeluruh. Penggunaan algoritma data mining dilakukan pada langkah ini untuk menggali data yang terintegrasi untuk memudahkan identifikasi informasi bernilai. Namun semakin besar data yang diolah maka semakin besar pula waktu prosesnya[3][4].

B. K-Nearest Neighbor

K-Nearest Neighbor (K-NN) termasuk kelompok instance-based learning. Algoritma ini juga merupakan salah satu teknik lazy learning. kNN dilakukan dengan mencari kelompok k objek dalam data training yang paling dekat (mirip) dengan objek pada data baru atau data testing. diperlukan suatu sistem klasifikasi sebagai sebuah sistem yang mampu mencari informasi. Contoh kasus, misal diinginkan untuk mencari solusi terhadap masalah seorang pasien baru dengan menggunakan solusi dari pasien lama. Perhitungan jarak ketetanggaan menggunakan algoritma eucliden seperti yang ditunjukkan pada persamaan 1.

$$euc = \sqrt{((a_1 - b_1)^2 + \dots + (a_n - b_n)^2)}$$

Dimana $a = a_1, a_2, \dots, a_n$, dan $b = b_1, b_2, \dots, b_n$ mewakili n nilai atribut dari dua record. Untuk atribut dengan nilai kategori..

Algoritma K-Nearest Neighbor (K-NN) adalah sebuah metode klasifikasi terhadap sekumpulan data berdasarkan pembelajaran data yang sudah terklasifikasi sebelumnya. Termasuk dalam *supervised learning*, dimana hasil *query instance* yang baru diklasifikasikan berdasarkan mayoritas kedekatan jarak dari kategori yang ada dalam K-NN.

Algoritma ini bekerja dengan berdasarkan pada jarak terpendek dari sample uji ke sample latih untuk menentukan KNNnya. Setelah mengumpulkan KNN, kemudian diambil mayoritas dari KNN untuk dijadikan prediksi dari sample uji. Dekat atau jauhnya tetangga biasanya dihitung berdasarkan jarak Euclidean. Langkah-langkah untuk menghitung metode K-Nearest Neighbor antara lain:

1. Menentukan parameter K
2. Menghitung jarak antara *data training* dan *data testing*

Perhitungan jarak yang paling umum dipakai pada perhitungan pada algoritma KNN adalah menggunakan perhitungan jarak Euclidean. Rumusannya adalah sebagai berikut:

$$euc = \sqrt{\sum_{i=1}^n (p_i - q_i)^2}$$

dimana :

p_i = sample data / *data training*

q_i = data uji / *data testing*

i = variabel data

n = dimensi data

3. Mengurutkan jarak yang terbentuk
4. Menentukan jarak terdekat sampai urutan K
5. Memasangkan kelas yang bersesuaian
6. Mencari jumlah kelas dari tetangga yang terdekat dan tetapkan kelas tersebut sebagai kelas data yang akan dievaluasi

Akurasi

Akurasi didefinisikan sebagai tingkat kedekatan antara nilai prediksi dengan nilai actual [5]. rumus akurasi dipaparkan pada persamaan 2.

Presisi

Presisi didefinisikan sebagai rasio item relevan yang dipilih terhadap semua item yang terpilih [6]. Presisi dapat diartikan sebagai kecocokan antara permintaan informasi dengan jawaban terhadap permintaan tersebut. rumus presisi ditunjukkan pada persamaan 3.

Recall

Recall didefinisikan sebagai rasio dari item relevan yang dipilih terhadap total jumlah item relevan yang tersedia. Rumus Recall diuraikan pada persamaan 4.

F-Measure

Measure adalah harmonic mean antara nilai presisi dan recall, F-measure juga kadang disebut dengan nama F1-Score. Rumus F-Measure dijabarkan pada persamaan 5.

$$AKURASI = ((TP+TN))/((TP+TN+FP+FN))$$

$$PRESISI = TP/(TP+FP)$$

$$RECALL = TP/(TP+FN)$$

$$F\text{-Measure} = 2 \text{ (Presisi x Recall)}/(\text{Presisi} + \text{Recall})$$

Keterangan Variabel:

- TP : True Positive
 TN : True Negative
 FP : False Positive
 FN : False Negative.

III. Hasil dan Pembahasan

Seperti yang telah dipaparkan sebelumnya bahwa tahapan yang dilakukan pada penelitian ini adalah dengan melakukan pembagian data training dan data testing, data yang digunakan sebanyak 77 data, dengan pembagian sebesar 90% sebagai data training dan 10% sebagai data tesing. Tahapan selanjutnya adalah menerapkan metode KNN, pemilihan nilai K pada penelitian ini yaitu nilai K=3,4 dan 5. Tabel 1. Menunjukkan hasil percobaan metode knn pada k=3,4 dan 5

Tabel 1. Hasil Pemasangan Kelas sesuai K

K=3		K = 4		K = 5	
1	TP	1	TP	1	TP
0	FN	0	FN	0	FN
1	TP	1	TP	1	TP
0	FN	0	FN	0	FN
0	FP	1	TP	0	FP
...	...	...	...	...	...
0	FN	0	FN	0	FN
0	FN	0	FN	0	FN
0	FN	0	FN	0	FN

Berdasarkan Tabel 1. Tahap selanjutnya adalah meneruskan hasil tersebut ke dalam bentuk confusion matrix, Tabel 2 menunjukkan confusion matrix pada k=3.

Tabel 2. Confusion Matrix K = 3

n = 77	Predicted : 1	Predicted : 0
Actual : 1	TP = 20	FN = 36
Actual : 0	FP = 11	TN = 10

Setelah diterapkan kedalam confusion matrix, performa metode dapat diukur, Tabel 3. Menunjukkan performa metode K-nn pada nilai K=3, dimana performa yang diukur adalah akurasi, presisi serta recall

Tabel 3. Hasil KNN dimana K=3

Akurasi	39%
Presisi	65%
Recall	36%
F-Measure	46%

Berdasarkan Tabel 1. Tahap selanjutnya adalah meneruskan hasil tersebut ke dalam bentuk confusion matrix, Tabel 4 menunjukkan confusion matrix pada k=4.

Tabel 4. Confusion Matrix K = 4

n = 77	Predicted : 1	Predicted : 0
Actual : 1	TP = 18	FN = 37
Actual : 0	FP = 13	TN = 9

Setelah diterapkan kedalam confusion matrix, performa metode dapat diukur, Tabel 3. Menunjukkan performa metode K-NN pada nilai K=4, dimana performa yang diukur adalah akurasi, presisi serta recall

Tabel 5. Hasil KNN dimana K=4

Akurasi	35%
Presisi	58%
Recall	33%
F-Measure	42%

Berdasarkan Tabel 1. Tahap selanjutnya adalah meneruskan hasil tersebut ke dalam bentuk confusion matrix, Tabel 4 menunjukkan confusion matrix pada k=5.

Tabel 6. Confusion Matrix K = 5

n = 77	Predicted : 1	Predicted : 0
Actual : 1	TP = 20	FN = 39
Actual : 0	FP = 11	TN = 7

Setelah diterapkan kedalam confusion matrix, performa metode dapat diukur, Tabel 3. Menunjukkan performa metode K-nn pada nilai K=5, dimana performa yang diukur adalah akurasi, presisi serta recall

Tabel 7. Hasil KNN dimana K=5

Akurasi	35%
Presisi	65%
Recall	34%
F-Measure	44%

IV. Kesimpulan

Dari hasil perhitungan algoritma K-Nearest Neighbor (KNN) di atas, maka telah mendapatkan hasil akurasi tertinggi yaitu 39% pada K=3, presisi tertinggi yaitu 65% pada K=3 dan K=5, *recall* tertinggi yaitu 36% pada K=3, dan *F-Measure* tertinggi yaitu 46% pada K=3. Nilai yang diperoleh tidak cukup baik dikarenakan jumlah data yang digunakan cukup kecil. Saran untuk penelitian selanjutnya adalah melakukan percobaan yang sama dengan menambahkan jumlah data serta menerapkan crossvalidation.

Daftar Pustaka

- [1] N. Fadhilah, H. Azis, and D. Lantara, "Validasi Pencarian Kata Kunci Menggunakan Algoritma Levenshtein Distance Berdasarkan Metode Approximate String Matching," *Pros. Semin. Nas. Ilmu Komput. dan Teknol. Inf.*, vol. 3, no. 2, pp. 3–7, 2018.
- [2] Hasran, "Klasifikasi Penyakit Jantung Menggunakan Metode K-Nearest Neighbor," *Indones. J. Data Sci.*, vol. 1, no. 1, pp. 1–4, 2020.
- [3] M. M. Baharuddin, T. Hasanuddin, and H. Azis, "Analisis Performa Metode K-Nearest Neighbor untuk Identifikasi Jenis Kaca," *Ilk. J. Ilm.*, vol. 11, no. 28, pp. 269–274, 2019.
- [4] A. Ilham, "Komparasi Algoritma Klasifikasi Dengan Pendekatan Level Data Untuk Menangani Data Kelas Tidak Seimbang," *J. Ilm. Ilmu Komput.*, vol. 3, no. 1, pp. 9–14, 2017.
- [5] M. Yusa, E. Utami, and E. T. Luthfi, "Analisis Komparatif Evaluasi Performa Algoritma Klasifikasi pada Readmisi Pasien Diabetes," *J. Buana Inform.*, vol. 7, no. 4, pp. 293–302, 2016, doi: 10.24002/jbi.v7i4.770.
- [6] Rizky Ade Putranto, Triastiti Wuryandari, and Sudarno, "Perbandingan Analisis Klasifikasi Antara Decision Tree Dan Support Vector Machine Multiclass Untuk Penentuan Jurusan Pada Siswa Sma," *J. Gaussian*, vol. 4, no. 4, pp. 1007–1016, 2015.
- [7] Y. Lukito and A. R. Chrismanto, "Perbandingan Metode-Metode Klasifikasi untuk Indoor Positioning System," *J. Tek. Inform. dan Sist. Inf.*, vol. 1, no. 2, pp. 123–131, 2015, doi: 10.28932/jutisi.v1i2.373.
- [8] S. Niu, J. Yang, S. Wang, and G. Chen, "Improvement and parallel implementation of canny edge detection algorithm based on GPU," *Proc. Int. Conf. ASIC*, no. 6, pp. 641–644, 2011, doi:

- 10.1109/ASICON.2011.6157287.
- [9] W. Ye, Y. Xia, and Q. Wang, "An Improved Canny Algorithm for Edge Detection," *J. Comput. Inf. Syst.*, vol. 75, pp. 1516–1523, 2011, doi: 10.1109/WCSE.2009.718.
 - [10] T. F. Wu, C. J. Lin, and R. C. Weng, "Probability estimates for multi-class classification by pairwise coupling," *J. Mach. Learn. Res.*, vol. 5, pp. 975–1005, 2004.
 - [11] K. Crammer, "On the algorithmic implementation of multiclass kernel-based vector machines," *J. Mach. Learn. Res. - JMLR*, vol. 2, no. 2, pp. 265–292, 2002.
 - [12] M. J. Hartmann and G. Carleo, "Neural-Network Approach to Dissipative Quantum Many-Body Dynamics," *Phys. Rev. Lett.*, vol. 122, no. 25, p. 250502, Jun. 2019, doi: 10.1103/PhysRevLett.122.250502.
 - [13] B. Gao and L. Pavel, "On the Properties of the Softmax Function with Application in Game Theory and Reinforcement Learning," 2017.
 - [14] H. Zhang, "The optimality of Naive Bayes," *Proc. Seventeenth Int. Florida Artif. Intell. Res. Soc. Conf. FLAIRS 2004*, vol. 2, pp. 562–567, 2004.
 - [15] V. Metsis, I. Androutsopoulos, and G. Paliouras, "Spam filtering with Naive Bayes - Which Naive Bayes?," *3rd Conf. Email Anti-Spam - Proceedings, CEAS 2006*, 2006.
 - [16] M. Christopher, P. Raghavan, and H. Schutze, *An Introduction to Information Retrieval*. Cambridge University Press, 2009.
 - [17] Y. L. Pavlov, "Random forests," *Random For.*, pp. 1–122, 2019, doi: 10.1201/9780367816377-11.
 - [18] T. Hastie, S. Rosset, J. Zhu, and H. Zou, "Multi-class AdaBoost," *Stat. Interface*, vol. 2, no. 3, pp. 349–360, 2009, doi: 10.4310/sii.2009.v2.n3.a8.
 - [19] R. Puri and K. Khamrui, "Application of Quantitative Descriptive Analysis (QDA), Principal Component Analysis (PCA) and Response Surface Methodology (RSM) in standardization of cham-cham making," 2015.
 - [20] A. Tharwat, "Linear vs. quadratic discriminant analysis classifier: a tutorial," *Int. J. Appl. Pattern Recognit.*, vol. 3, no. 2, p. 145, 2016, doi: 10.1504/ijapr.2016.079050.
 - [21] A. Tharwat, "Classification assessment methods," *Appl. Comput. Informatics*, 2018, doi: 10.1016/j.aci.2018.08.003.
 - [22] P. A. Flach and M. Kull, "Precision-Recall-Gain curves: PR analysis done right," *Adv. Neural Inf. Process. Syst.*, vol. 2015-Janua, pp. 838–846, 2015.
 - [23] L. Nurhayati and H. Azis, "Perancangan Sistem Pendukung Keputusan Untuk Proses Kenaikan Jabatan Struktural Pada Biro Kepegawaian," *Semin. Nas. Teknol. Inf. dan Multimed.*, pp. 6–7, 2016.
 - [24] J. D. Kelleher, B. Mac Namee, and A. D. Arcy, *Fundamentals of Machine Learning For Predictive Data Analytics Algorithms, Worked Examples, and Case Studies*. London: The MIT Press, 2015.
 - [25] K. H. Brodersen, C. S. Ong, K. E. Stephan, and J. M. Buhmann, "The balanced accuracy and its posterior distribution," *Proc. - Int. Conf. Pattern Recognit.*, pp. 3121–3124, 2010, doi: 10.1109/ICPR.2010.764.
 - [26] A. A. Karim, H. Azis, and Y. Salim, "Kinerja Metode C4.5 dalam Penyaluran Bantuan Dana Bencana 1," *Pros. Semin. Nas. Ilmu Komput. dan Teknol. Inf.*, vol. 3, no. 2, pp. 84–87, 2018.
 - [27] A. Fitria and H. Azis, "Analisis Kinerja Sistem Klasifikasi Skripsi menggunakan Metode Naïve Bayes Classifier," *Pros. Semin. Nas. Ilmu Komput. dan Teknol. Inf.*, vol. 3, no. 2, pp. 102–106, 2018.