Indonesian Journal of Data and Science ISSN: 2715-9930
Vol 1, No 2, Juli 2020, pp. 39-43

Analisis performa metode Knn pada Dataset pasien pengidap Kanker Payudara

Dewi Cahyanti[a,1], Alifah Rahmayani[a,2], Syafira Ainy Husniar[a,3]

[a] *Universitas Muslim Indonesia, Jl. Urip Sumoharjo KM.5, Makassar 90231, Indonesia*
[1] *dewicahyanti751@gmail.com;* [2] *Emailtugasku18@gmail.com;* [3] *syafira.ainyhusniar@gmail.com;*

INFORMASI ARTIKEL

Diterima : 11 – 04 – 2020
Direvisi : 29 – 06 – 2020
Diterbitkan : 31 – 07 – 2020

Kata Kunci:
k-nearest neigbor
analysis performa
kanker payudara
crossvalidation
klasifikasi

ABSTRAK

Abstrak-Kanker payudara adalah penyakit non kulit yang berasal dari sel kelenjar, saluran kelenjar, dan jaringan penunjang payudara. Paper ini menggunakan metode K Nearest Neighbor untuk mengklasifikasi dataset. K-Nearest Neighbor adalah metode untuk melakukan klasifikasi terhadap objek berdasarkan data pembelajaran yang jaraknya paling dekat dengan objek tersebut. Penelitian ini mencoba menerapkan metode knn pada dataset pasien pengidap penyakit kanker payudara, k yg diterapkan adalah k=3 hingga k=5 serta menerapkan crossvalidation dengan kfold=5, setelah dilakukan pengujian maka dengan metode KNN diperoleh hasil tertinggi untuk Akurasi dengan nilai 0,93 pada 20% keempat (K3), 20% Pertama(K4) dan 20% pertama(K5), untuk Presisi dengan nilai 0,97 pada 20% keempat(K3), untuk Recall dengan nilai 0,98 pada 20% ketiga (K3) dan F-measure dengan nilai 0,94 pada 20% keempat(K3) dan 20% ketiga(K5).

I. Pendahuluan

Kanker payudara adalah penyakit non kulit berbahaya yang paling umum dialami oleh wanita, penyakit tersebut disebabkan oleh beberapa faktor yaitu dari sel dan saluran kelenjar hingga jaringan penopang payudara, kecuali kulit dari payudara. Kanker payudara juga termasuk penyebab nomor dua kematian terbanyak akibat kanker pada wanita setelah kanker serviks, dan cenderung terus meningkat setiap tahunnya. Kanker payudara ini secara umum dibagi menjadi 2,yaitu benign atau biasa disebut jinak dan malignant atau biasa disebut juga ganas, biasanya kanker payudara jinak ditandai dengan berbentuk benjolan kecil bulat, dan lembut. Kanker payudara dalam tingkat jinak biasanya akan mempunyai keadaan dan pertumbuhan yang tidak bersifat kanker. Kanker ini bisa terdeteksi tetapi tidak akan menjalar dan merusak jaringan di dekatnya. Pada kanker payudara dalam tingkat ganas ditandai dengan bentuk yang tidak simetris, kasar, terasa nyeri, dan lainnya biasanya kanker payudara menjalar dan merusak jaringan dan organ lain yang ada di dekatnya.

Pada penelitian ini, kami melakukan klasifikasi penyakit kanker payudara dengan metode KNN. Metode KNN memiliki beberapa keunggulan, yaitu pelatihan yang sederhana, cepat, mudah dimengerti, dan efektif apabila ukuran data pelatihan besar. Namun, KNN ini juga terdapat kelemahan, yaitu nilai k bias.

Crossvalidation adalah salah satu metode yang digunakan agar dapat mensimulasikan semua data agar setiap data dapat berkesempatan menjadi data training dan data tesing. Pada penelitian ini crossvalidation dibagi menjadi kfold = 5.

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 ketetanggan menggunakan algoritma euclidien seperti yang ditunjukkan pada persamaan 1.

$$\text{euc} = \sqrt{((a\_1 - b\_1)\text{^}2 + \cdots + (a\_n - b\_n)\text{^}2)}$$

Dimana a = a1,a2, …, an, dan b = b1, b2, …, bn 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 terklasifikasikan sebelumya. 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 Eucledian. 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:

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

   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.

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

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

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

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

Keterangan Variabel:

TP : True Positive

TN : True Negative

FP : False Positive

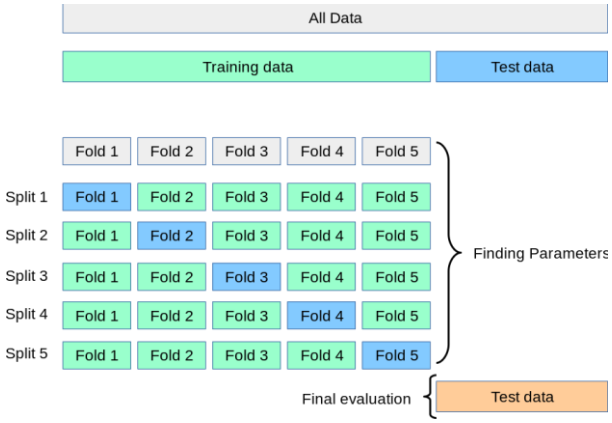
FN : False Negative.

C. K-Fold Cross Validation

K-Fold Cross Validation adalah salah satu dari jenis pengujian cross validation yang berfungsi untuk menilai kinerja proses sebuah metode algoritme dengan membagi sampel data secara acak dan mengelompokkan data tersebut sebanyak nilai K k-fold. Kemudian salah satu kelompok k-fold tersebut akan dijadikan sebagai data uji sedangkan sisa kelompok yang lain akan dijadikan sebagai data latih.

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 500 data, dengan pembagian crossvalidation sebesar kfold=5, sejingga menjadi 80% data training dan 20% data testing disetiap tahapan. Gambar 2 mengilustrasikan crossvalidation pada penelitian ini



Gambar 2. Penerapan crossvalidationHasil Akhir 20% Pertama

Proses data mining disini diawali membuat model dengan data training kemudian mengimplementasikannya ke data testing dengan menggunakan algoritma K-Nearest Neighbor, sehingga didapatkan hasil akhir berupa nilai akurasi dan hasil klasifikasi sebagai berikut.

Tabel 1. Hasil pengujian performa metode KNN

	Crossvalidation	Akurasi	Presisi	Recall	F-measure
K3	20% Pertama	0,90	0,91	0,92	0,90
	20% Kedua	0,89	0,91	0,91	0,90
	20% Ketiga	0,92	0,94	0,98	0,93
	20% Keempat	0,93	0,97	0,93	0,94
	20% Kelima	0,42	0,84	0,26	0,38
	20% Pertama	0,93	0,93	0,97	0,85

	Crossvalidation	Akurasi	Presisi	Recall	F-measure
K4	20% Kedua	0,87	0,88	0.91	0,89
	20% Ketiga	0,91	0,91	0,94	0,92
	20% Keempat	0,90	0,94	0,90	0,91
	20% Kelima	0,42	0,83	0,24	0,35
K5	20% Pertama	0,93	0,93	0,97	0,85
	20% Kedua	0,92	0,94	0,93	0,93
	20% Ketiga	0,93	0,94	0,95	0,94
	20% Keempat	0,92	0,94	0,93	0,93
	20% Kelima	0,41	0,8	0,25	0,38

IV. Kesimpulan

Jadi, dapat di simpulkan dari perhitungan 569 data yang di bagi menjadi 20% training dan 80% testing dengan K = 3,4 dan 5 mendapat nilai tertinggi untuk akurasi adalah 0,93 pada 20% keempat(K3), 20% Pertama(K4) dan 20% pertama(K5), untuk Presisi dengan nilai 0,97 pada 20% keempat(K3), untuk Recall dengan nilai 0,98 pada 20% ketiga(K3) dan F-measure dengan nilai 0,94 pada 20% keempat(K3) dan 20% ketiga(K5).

Daftar Pustaka

[1] 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.

[2] 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.

[3] 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.

[4] 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.

[5] H. Azis, R. D. Mallongi, D. Lantara, and Y. Salim, "Comparison of Floyd-Warshall Algorithm and Greedy Algorithm in Determining the Shortest Route," *Proc. - 2nd East Indones. Conf. Comput. Inf. Technol. Internet Things Ind. EIConCIT 2018*, pp. 294–298, 2018.

[6] N. Fadhillah, Huzain 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.

[7] S. Chugh, K. Arivu Selvan, and R. K. Nadesh, "Prediction of heart disease using apache spark analysing decision trees and gradient boosting algorithm," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 263, no. 4, pp. 0–10, 2017.

[8] M. Lestari, "Penerapan Algoritma Klasifikasi Nearest Neighbor (K-NN) Untuk Mendeteksi Penyakit Jantung," *Fakt. Exacta*, vol. 7, no. September 2010, pp. 366–371, 2014.

[9] V. Chaurasia, "Early Prediction of Heart Diseases Using Data Mining," *Caribb. J. Sci. Technol.*, vol. 1, no. December, pp. 208–217, 2013.

[10] Rosmasari *et al.*, "Usability Study of Student Academic Portal from a User's Perspective," *Proc. - 2nd East Indones. Conf. Comput. Inf. Technol. Internet Things Ind. EIConCIT 2018*, pp. 108–113, 2018.

[11] Hasran, "Klasifikasi Penyakit Jantung Menggunakan Metode K-Nearest Neighbor," *Indones. J. Data Sci.*, vol. 1, no. 1, pp. 1–4, 2020

[12] A. Tharwat, "Classification assessment methods," *Appl. Comput. Informatics*, 2018, doi: 10.1016/j.aci.2018.08.003.

[13] 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.

[14] 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.

[15] 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.

[16] 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.