SNESTIK

Seminar Nasional Teknik Elektro, Sistem Informasi, dan Teknik Informatika

https://ejurnal.itats.ac.id/snestik dan https://snestik.itats.ac.id

Informasi Pelaksanaan :
SNESTIK II - Surabaya, 26 Maret 2022
Ruang Seminar Gedung A, Kampus Institut Teknologi Adhi Tama Surabaya

Informasi Artikel:
DOI : 10.31284/p.snestik.2022.2750

Prosiding ISSN 2775-5126

Fakultas Teknik Elektro dan Teknologi Informasi-Institut Teknologi Adhi Tama Surabaya
Gedung A-ITATS, Jl. Arief Rachman Hakim 100 Surabaya 60117 Telp. (031) 5945043
Email : snestik@itats.ac.id

Metode K-Nearest Neighbor (KNN) dalam Memprediksi Curah Hujan di Kota Bandung

Deden Martia Nanda[1], Tacbir Hendro Pudjiantoro[2], Puspita Nurul Sabrina[3]

Program Studi Teknik Informatika, Universitas Jenderal Achmad Yani[1,2,3]

e-mail: dedenmartia17@if.unjani.ac.id

ABSTRACT

Rainfall has an erratic pattern so it is difficult to predict manually. The amount of rainfall that is quite large cannot be determined with certainty but it can be estimated. Thus, the existence of Data Mining allows machines to recognize and learn complex data patterns. Therefore machine learning can study rainfall data patterns to make predictions, so this research performs a data mining process on rainfall data calculated from January 1, 2015 to December 31, 2020 using the K-Nearest Neighbor method according to the criteria of average temperature, humidity mean air, average wind speed, and rainfall. This classification process has the aim of getting the best accuracy in predicting daily rainfall in the city of Bandung, as well as providing knowledge of the K-Nearest Neighbor algorithm regarding rainfall. The results of this study indicate that when the K value is 5, then the accuracy results are 86,199% with the accuracy test results using the Confusion Matrix resulting in an accuracy of 84.38%.

Keywords: *Algoritma; K-Nearest Neighbor (KNN); Rainfall; Data Mining; Confusion Matrix.*

ABSTRAK

Curah hujan memiliki pola yang tidak menentu sehingga sulit dilakukan prediksi dengan cara manual. Curah hujan yang cukup besar tidak dapat ditentukan secara pasti namun hal ini dapat diperkirakan. Dengan demikian, adanya Data Mining memungkinkan mesin mengenali dan mempelajari pola data yang rumit. Maka dari itu pembelajaran mesin dapat mempelajari pola data curah hujan untuk melakukan prediksi, maka penelitian ini melakukan proses data mining pada data curah hujan terbilang dari 1 januari 2015 sampai 31 desember 2020 dengan menggunakan metode K-Nearest Neighbor sesuai kriteria temperature rata-rata, kelembapan udara rata-rata, kecepatan angin rata-rata, dan curah hujan. Proses klasifikasi ini memiliki tujuan untuk mendapatkan akurasi terbaik dalam memprediksi curah hujan harian di Kota Bandung, serta memberi

pengetahuan algoritma K-Nearest Neighbor mengenai curah hujan. Hasil dari penelitian ini menunjukan bahwa ketika nilai K sebesar 5, maka didapatkan hasil akurasi sebesar 86.199% dengan hasil pengujian akurasi menggunakan Confusion Matrix dihasilkan akuarasi sebesar 84.38%.

Keywords: Algoritma; K-Nearest Neighbor (KNN); Curah Hujan; Data Mining; Confusion Matrix.

PENDAHULUAN

Iklim salah satu hal yang cukup krusial di dalam dunia ini. Garis lintang,ketinggian, lereng, jarak dari perairan, dan kondisi arus air laut merupakan pengaruh dari pembentukan iklim di suatu tempat. Namun setiap daerah memiliki iklim yang berbeda-beda. Setiap daerah memiliki jenis iklim yang dipengaruhi oleh garis lintang, karakteristik pola iklim global dipelajari melalui Klimatologi. Karakteristik pola iklim global juga mempertimbangkan kondisi seperti hujan, suhu, angin atau penguapan. Iklim di permukaan bumi dapat dibedakan berdasarkan garis lintang menjadi iklim kutub, sedang, tropis, subtropis dan khatulistiwa. Indonesia adalah negara yang beriklim tropis. Iklim berpengaruh sangat besar bagi kelangsungan hidup manusia [1]. Daerah tropis seperti Indonesia, anomali iklim ini menimbulkan pergeseran pola curah hujan, perubahan besar-besaran curah hujan dan perubahan temperatur udara yang mengakibatkan timbulnya musim kemarau, dan kekeringan [2].

Curah hujan memiliki pola yang tidak menentu sehingga sulit dilakukan prediksi dengan cara manual. Namun demikian, dengan adanya *Data Mining* memungkinkan mesin mengenali dan mempelajari pola data yang rumit [3]. Sehingga mesin dapat mempelajari pola data cuaca untuk melakukan prediksi [4].

Memprediksi curah hujan berdasarkan dataset yang di dapatkan sebelumnya. Penelitian ini menggunakan metode *K-Nearest Neighbor (KNN). KNN* termasuk kedalam bagian *Data Mining* yang mampu mengenali suatu pola data sequensial. Metoda klasifikasi ini menggunakan konsep perhitungan jarak terdekat dengan sebuah titik [5]. Penetuan jarak menggunakan rumus *euclidian distance* yang menghasilkan jarak antara data baru dengan seluruh data pada dataset yang sudah memiliki kelas. Penelitian sebelumnya yang dilakukan Riza Indriani Rakhmalia di Universitas Islam Indonesia Yogyakarta dengan judul *"Perbandingan Hasil Metode Naïve Bayes Classifier dan Support Vector Machine dalam Klasifikasi Curah Hujan".* Pada penelitian sebelumnya metode *Naïve Bayes* mampu memberikan klasifikasi dengan akurasi 78%. Sedangkan dengan menggunakan metode SVM mendapatkan akurasi 80% [6].

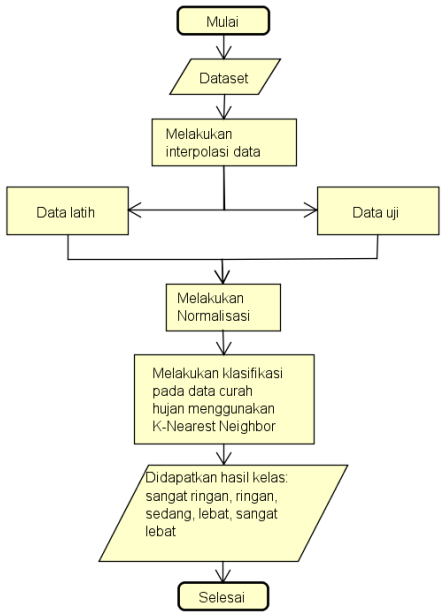
Memprediksi curah hujan diperlukan beberapa parameter yang terdiri dari temperature udara, kelembaban udara dan kecepatan angin yang dapat digunakan untuk melihat kecenderungan turunnya hujan yang akan datang. Pada penelitian sebelumnya, dalam memprediksi curah hujan dibutuhkan 3 variabel iklim diantaranya kecepatan angin, suhu dan kelembaban. Namun saat ini, curah hujan sudah semakin sulit untuk di prediksi [7]. Maka dari itu diperlukan model ataupun sistem yang dapat memprediksi curah hujan dengan akurat berdasarkan data terdahulu berdasarkan sumber *BMKG* (Badan Meteorologi dan Geofisika) [8].

Maka dari itu penelitian ini telah memprediksi curah hujan dengan menggunakan metode *K-Nearest Neighbor*. Sebelumnya data dilakukan praproses untuk melengkapi data yang hilang menggunakan teknik interpolasi data, sehingga didapatkan data yang lengkap, selanjutnya dilakukan normalisasi terhadap data sehingga di dapatkan fitur yang dapat diproses menggunakan metode *KNN* untuk mengklasifikasikan data tersebut kedalam 5 kelas yaitu sangat ringan, ringan, sedang, lebat, dan sangat lebat.

METODE

Metode penelitian ini dilakukan beberapa tahapan seperti perolehan data, kemudian dilanjutkan pra proses interpolasi. Data dibagi menjadi dua yaitu data latih 80% sebanyak 1.519 *record* dan data uji 20% sebanyak 379 *record*, setelah itu data akan di normalisasi, dan dilakukan prediksi

menggunakan KNN sehingga menghasilkan salah satu dari lima kelas. Metode penelitian ditunjukan pada Gambar 1.



Gambar 1. Metode Penelitian

Perolehan Data

Data diambil dari *website* BMKG selama enam tahun terahkir (2015-2020) dengan format (.csv). Data tersebut merupakan data harian sebanyak 1.899 x 4 parameter iklim = 7.569 data. Set data dibagi kedalam data latih dan data uji, untuk data latih yang digunakan adalah 80% dan 20% untuk data uji. Parameter diantaranya Temperatur(ºC), Kelembapan Udara(%), Kecepatan Angin(m/s), Curah Hujan(mm). Akan dilakukan pra proses yaitu data curah hujan ditunjukan pada Tabel 1.

Tabel 1. Data Curah Hujan

Hari ke	Tanggal	(ºC)	(%)	(m/s)	(mm)
1	01/01/2015	23,2	79	2	0,5
2	02/01/2015	23	84	2	1,5
….	…..	…	…	…	…
1898	30/12/2020	22,7	76,0	3,0	0,9
1899	31/12/2020	23,8	74,0	3,0	7,6

Pra Proses

Pra proses ini dilakukan untuk menyiapkan data sebelum masuk ke dalam pembelajaran. Dilakukan dengan dua tahapan yaitu Interpolasi Data dan Normalisasi.

1. Interpolasi Data

   Beberapa data cuaca dari website BMKG terdapat data yang tidak terukur nilainya dan tidak terbaca. Teknik interpolasi data ini digunakan untuk mengisi data yang kosong. Untuk mencari data yang kosong yaitu mencari titik tengah tiap variable data cuaca yang berada diantara dua nilai [11].

2. Normalisasi

Proses normalisasi data untuk setiap parameter iklim berfungsi untuk mengubah data menjadi bentuk normal, dikarenakan data yang diperoleh dari BMKG terdapat data yang bernilai sangat besar, sangat kecil. Maka dilakukan proses normalisasi *Min-Max* dengan penskalaan dalam rentang 0 sampai dengan 1 [12].

Prediksi Menggunakan *K-Neareast Neighbor*

Data Mining adalah proses yang melakukan pengolahan data histori kejadian-kejadian sebelumnya dan digunakan sebagai dasar untuk membangun sebuah pengetahuan sebuah sistem prediksi dari sebuah kasus terdahulu [9]. serangkaian proses untuk menggali nilai tambah dari suatu kumpulan data berupa pengetahuan yang selama ini tidak diketahui secara manual [10]. Pada tahap ini merupakan proses perhitungan *K-Nearest Neighbor*. Dengan menentukan atribut yang akan digunakan untuk proses perhitungan secara manual. Tahap klasifikasi sebagai berikut :

1. *K-Cross Validation*

K-Cross validation dapat disebut juga sebagai estimasi rotasi merupakan sebuah teknik validasi model yang digunakan untuk menilai hasil statistic analisis yang menggeneralisasi kumpulan data independent[13][14]. Data set yang telah diperoleh dibagi menjadi dua data latih sebanyak 80% dan data uji sebanyak 20%. Proses ini berfungsi untuk mengurangi waktu komputasi dengan menjaga keakuratan estimasi.

2. Klasifikasi Menggunakan *K-Neareast Neighbor*

Kemudian proses perhitungan metode KNN untuk melakukan klasifikasi terhadap objek berdasarkan dari data pembelajaran yang jaraknya paling dekat dengan objek tersebut [15]. Dimana kelas yang paling banyak muncul yang nantinya akan menjadi kelas hasil dari klasifikasi [16]. *KNN* dilakukan dengan mencari kelompok *k* objek yang paling dekat (mirip) dengan objek pada dataset sebelumnya [17]. Perhitungan menggunakan KNN ini adalah 10 data latih dan satu data uji. Berikut adalah Langkah-langkah dalam menghitung menggunakan metode KNN.

- Menentukan nilai *K*
- Menghitung jarak antara data latih dan data uji

Setelah diketahui nilai *K* terdekat maka untuk mencari jarak terdekat dengan data uji ini menggunakan rumus *Euclidian Distance* ditunjukan pada Persamaan (1).

$$d(x,y) = \sqrt{\sum_{i=1}^{m}(Xi - Yi)^2} \quad (1)$$

Maka akan diketahui nilai X_1, X_2, X_3, X_4. Selanjutnya melakukan perhitungan jarak menggunakan *Euclidian Distance*.

- Setelah melakukan perhitungan jarak antara data latih dan data uji, maka akan didapatkan hasil jarak menggunakan rumus *Euclidian Distance.*
- Urutan hasil perhitungan jarak

Setelah mengurutkan atau meranking hasil perhitungan dengan rumus *Euclidian distance* dari yang terkecil hingga terbesar. Maka ditarik kesimpulan data uji yang masuk kedalam label adalah *K* yang terdekat.

Pengujian Metode Menggunakan *Confuision Matrix*

Metode ini digunakan untuk mengukur performa masalah klasifikasi dimana keluaran dapat berupa dua akelas atau lebih. *Confusion Matrix* merupakan tabel dengan empat kombinasi berbeda dari nilai prediksi dan nilai aktual. Terdapat empat istilah yang merupakan representasi hasil proses klasifikasi pada *confusion matrix* yaitu *True Positif, True Negatif, False Positif* dan *False Negatif.* Untuk mengukur performa dari algoritma *K-Nearest Neighbor* menggunakan empat metode evaluasi yaitu :

- *Accuracy*, menggambarkan persentase jumlah record data yang terklasifikasi dengan benar oleh sistem menggunakan Persamaan (2).

$$Accuracy = \frac{TP}{Jumlah\ Total\ Dataset} \quad (2)$$

- *Precision*, menggambarkan presentase akurasi antara data yang diminta dengan hasil klasifikasi *KNN* menggunakan Persamaan (3).

$$Precision = \frac{TP}{TP+FP} \quad (3)$$

- *Recall*, menggambarkan presentase akurasi antara data yang diminta dengan hasil KNN menggunakan Persamaan (4).

$$Recall = \frac{TP}{TP+FN} \quad (4)$$

- *F1-Score*, menggambarkan suatu perbandingan nilai rata-rata dari *precision* dan *recall* menggunakan Persamaan (5).

$$F1 - Score = \frac{2\ x\ Recall\ x\ Precision}{Recall+Precision} \quad (5)$$

HASIL DAN PEMBAHASAN

Pembahasan I

Pembahasan ini berupa hasil peneliti yang telah dilakukan menggunakan metode *K-Nearest Neighbor* sebelumnya data telah dilakukan pra proses seperti data hilang dan normalisasi kemudian diklasifikasi dengan metode yang digunakan dengan tahapan sebagai berikut :

K-Cross Validation Proses ini model dilatih oleh subset pembelajaran dan divalidasi oleh subset validasi. Data set ini dibagi menjadi dua data yaitu data latih sebanyak 80% dan data uji sebanyak 20%.

Kemudian klasifikasi menggunakan *KNN* dengan 10 data latih dan 1 data uji. Dengan menentukan nilai *K* terlebih dahulu, setelah itu menghitung jarak antara data latih dan data uji. Setelah diketahui nilai *K* tetangga terdekat yaitu sebanyak 5. Maka untuk mencari jarak terdekat dengan data uji ini menggunakan rumus *Euclidian distance.* Berikut Data Latih dan Data Uji yang akan dilakukan perhitungan jarak dengan rumus *Euclidian Distance,* dan *Manhattan Distance.* Data latih ditujukan pada Tabel 3. Dan Data Uji ditujukan pada Tabel 4.

Tabel 2. Data Latih

Hari ke	(ºC)	(%)	(m/s)	(mm)	Label
1	23,2	79	2	0,5	Sangat Ringan
2	23	84	2	1,5	Sangat Ringan
3	23,3	81	2	23,2	Ringan
4	23,2	82	1	1,5	Sangat Ringan
5	23,7	78	2	42,9	Sedang
6	22,2	83	2	2,1	Sangat Ringan
7	23,2	80	2	4,2	Sangat Ringan
8	22,7	82	2	0	Sangat Ringan

Tabel 3. Data Uji

Hari ke	(ºC)	(%)	(m/s)	(mm)	Label
2539	23,2	81	4	5,3	?

Setelah mengurutkan atau me-ranking hasil perhitungan dengan rumus *Euclidian Distance* dari yang terkcil hingga yang terbesar maka untuk nilai *K* terdekat yang sebelumnya telah ditentukan *K- Cross Validation* yaitu *K* = 5 hasil yang didapatkan dapat dilihat pada Tabel 5, dan 6.

Tabel 4. Hasil Perhitungan Euclidean Distance dengan K = 5

Hari ke	(ºC)	(%)	(m/s)	(mm)	Euclidian Distance	Label
4	23,2	84	1	1,5	5.242136969	Sangat Ringan
5	23,7	79	2	0,5	5.571355311	Sangat Ringan
6	22,7	82	2	0	5.774080013	Sangat Ringan
7	23,3	81	2	23,2	18.25979189	Ringan
8	23,7	78	2	42,9	37.7757859	Sedang

Tabel 5. Hasil Perhitungan Manhattan Distance dengan K = 5

Hari ke	(ºC)	(%)	(m/s)	(mm)	Manhattan Distance	Label
4	23,2	84	1	1,5	8,8	Sangat Ringan
5	23,7	79	2	0,5	8,8	Sangat Ringan
6	22,7	82	2	0	9	Sangat Ringan
7	23,3	81	2	23,2	20	Ringan
8	23,7	78	2	42,9	43,1	Sedang

Maka dapat ditarik kesimpulan data uji ini masuk kedalam label Sangat Ringan, dikarenakan *K* terdekat sebanyak 5 menyatakan label Sangat Ringan = 3, Ringan 1, dan Sedang = 1. Data uji yang sudah berlabel ditujukan pada Tabel 7.

Tabel 6. Data Uji yang Sudah Berlabel

Hari ke	(ºC)	(%)	(m/s)	(mm)	Manhattan Distance	Label
2539	23,2	81	4	5,3	Sangat Ringan	2539

Pembahasan II

Pengujian Metode Menggunakan *Confuision Matrix* Setelah dilakukan perhitungan *Accuracy, Precision, Recall, dan F1-Score*. Dapat disimpulkan hasil yang didapatkan seperti pada Tabel 8. Dengan nilai akurasi sebesar 85.78%.

Tabel 7. Pengujian Metode Menggunakan Confuision Matrix

Kelas	*Accuracy*	*Precision*	*Recall*	F1-*Score*
Sangat Ringan	90.05%	90%	86%	87%
Ringan	90.05%	80%	85%	82%
Sedang	94%	77%	85%	80%
Lebat	95.84%	89%	86%	87%
Sangat Lebat	96.63%	88%	85%	86%

KESIMPULAN

Berdasarkan hasil dari penelitian ini adalah memprediksi curah hujan menggunakan metode *K-Nearest Neighbor* dengan parameter temperatur rata-rata, kelembapan rata-rata, kecepatan angin rata-rata, dan curah hujan. Menghasilkan suatu prediksi curah hujan harian dengan melihat hasil prediksi curah hujan menggunakan metode *K-Nearest Neighbor*. Parameter yang paling berpengaruh dalam hasil akurasi yaitu kecepatan angin rata-rata dikarenakan menurut BMKG angin berperan dalam memindahkan awan dari satu tempat ke tempat yang lainnya. Wilayah yang memiliki angin yang lemah memiliki kemungkinan memiliki intensitas curah hujan yang kecil. Berdasarkan nilai *K* yang telah dilakukan pelatihan *K-Cross Validation* maka didapatkan hasil terbaik sebanyak 5 dengan nilai akurasi sebesar 86.199%. Hasil pengujian metode dengan menggunakan *Confusion Matrix* menghasilkan akurasi sebesar 85.78%.

REFERENSI

[1] B. Tjasyono H. K., *Meteorologi Indonesia Volume I -Karakteristik dan Sirkulasi Atmosfer*, vol. I. 2012.

[2] B. Irawan, "Fenomena Anomali Iklim El Nino dan La Nina: Kecenderungan Jangka Panjang dan Pengaruhnya terhadap Produksi Pangan," *Forum Penelit. Agro Ekon.*, vol. 24, no. 1, p. 28, 2016.

[3] F. R. Lumbanraja, R. S. Sani, D. Kurniawan, and A. R. Irawati, "Implementasi Metode Support Vector Machine Dalam Prediksi Persebaran Demam Berdarah Di Kota Bandar Lampung," *J. Komputasi*, vol. 7, no. 2, 2019.

[4] S. B. Navathe, W. Wu, S. Shekhar, X. Du, X. Sean Wang, and H. Xiong, "Database Systems for Advanced Applications: 21st International Conference, DASFAA 2016 Dallas, TX, USA, April 16–19, 2016 Proceedings, Part I," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 9642, pp. 214–228, 2016.

[5] A. Bode, "K-Nearest Neighbor Dengan Feature Selection Menggunakan Backward Elimination Untuk Prediksi Harga Komoditi Kopi Arabika," *Ilk. J. Ilm.*, vol. 9, no. 2, pp. 188–195, 2017.

[6] P. S. Statistika, F. Matematika, D. A. N. Ilmu, P. Alam, and U. I. Indonesia, "Perbandingan Hasil Metode Naïve Bayes Classifier dan Support Vector Machine Dalam," 2018.

[7] D. Howard and B. Mark, "Neural Network Toolbox Documentation," *Neural Netw. Tool*, p. 846, 2004.

[8] N. Ritha, M. Bettiza, and A. Dufan, "Prediksi Curah Hujan dengan Menggunakan Algoritma Levenberg-Marquardt dan Backpropagation," *J. Sustain.*, vol. 5, no. 2, pp. 11–16, 2016.

[9] F. Gorunescu, *Intelligent System Reference Library, Volume 12*. 2005.

[10] M. Bramer, *Principles of Data Mining*, no. January 2007. 2007.

[11] A. Anjomshoaa and M. Salmanzadeh, "Filling missing meteorological data in heating and cooling seasons separately," *Int. J. Climatol.*, vol. 39, no. 2, pp. 701–710, 2019.

[12] H. Leidiyana, "Penerapan Algoritma K-Nearest Neighbor Untuk Penentuan Resiko Kredit Kepemilikan Kendaraan Bermotor," *J. Penelit. Ilmu Komputer, Syst. Embed. Log.*, vol. 1, no. 1, pp. 65–76, 2013.

[13] M. J. Hartmann and G. Carleo, "Neural-Network Approach to Dissipative Quantum Many-Body Dynamics," *Phys. Rev. Lett.*, vol. 122, no. 25, p. 250502, 2019.

[14] 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.

[15] A. J. T, D. Yanosma, and K. Anggriani, "Implementasi Metode K-Nearest Neighbor (Knn) Dan Simple Additive Weighting (Saw) Dalam Pengambilan Keputusan Seleksi Penerimaan Anggota Paskibraka," *Pseudocode*, vol. 3, no. 2, pp. 98–112, 2017

[16] M. Fansyuri, "Analisa algoritma klasifikasi k-nearest neighbor dalam menentukan nilai akurasi terhadap kepuasan pelanggan (study kasus pt. Trigatra komunikatama)," *Humanika J. Ilmu Sos. Pendidikan, dan Hum.*, vol. 3, no. 1, pp. 29–33, 2020.

[17] N. Reflan, A. Aflahah, Kusrini, and Juwari, "Implementasi Metode K-Nearest Neighbor (Knn) Untuk Memprediksi Varietas Padi Yang Cocok Untuk Lahan Pertanian," *J. Inf. Politek. Indonusa Surakarta*, vol. 4, pp. 2–8, 2018.