Journal of Information Systems Management and Digital Business (JISMDB)
Volume 1, No 2 – Januari 2024
ISSN 3026-1449

PERBANDINGAN METODE PENGUKURAN JARAK PADA K-NEAREST NEIGHBOUR DALAM KLASIFIKASI DATA TEKS CARDIOVASKULAR

Daffa Ardiyansyah[1*], Nurfaida Oktafiani[2]
[1,2]Universitas Trunojoyo Madura, Indonesia
[1]daffaardiyansyah99@gmail.com, [2]nf.oktafiani@gmail.com

Received: 04-12-2023	Revised: 02-01-2024	Approved: 14-01-2024

ABSTRAK

Di era serba teknologi saat ini, pemanfaatan pengolahan data menjadi suatu kebutuhan yang tidak dapat dipungkiri, terutama dalam konteks kesehatan. Salah satu pemanfaatan dalam pengelolaan data Kesehatan adalah klasifikasi data teks penyakit, seperti penyakit cardiovascular, oleh karena itu penelitian ini ada dengan tujuan untuk mengevaluasi kinerja metode KNN dengan menggunakan 3 jenis pengukuran jarak, yaitu Euclidean, Manhattan/City Block, dan Mahalanobis. Meskipun sederhana, algoritma ini telah berhasil menghasilkan kinerja yang tinggi dalam beberapa kasus. Pendekatan KNN ini melakukan klasifikasi terhadap suatu objek berdasarkan pada kesamaan atau jarak terdekat dengan objek-objek dalam data latih (training). Penelitian menghasilkan nilai akurasi yang diperoleh dari variasi nilai k dari angka 1 hingga 31 dengan keliapatan ganjil. Hasil dari perhitungan jarak Euclidien, Manhattan, Dan Mahalanobis terhadap akurasi K-NN. Akurasinya apabila semakin banyak nilai k menyebabkan kenaikan akurasi, walaupun ada kondisi dimana nilai K tertentu akurasi yang menurun walau tidak siknifikan. total Akurasi terendah diperoleh dari jarak Euclidien, sedangkan total akurasi tertinggi diraih oleh jarak Manhattan dan disusul oleh jarak Mahalanobis. Jarak Manhattan dan Mahalanobis menghasilkan total akurasi terbaik terbanyak serta menghasilkan akurasi terbaik pada Sebagian besar ukuran K, sedangkan Euclidien menjadi perhitungan dengan total akurasi terbaik. Meningkatkan jumlah K pada setiap perhitungan jarak dapat meningkatkan akurasi klasifikasi. Jumlah K yang optimal dalam eksperimen ini adalah 29, menunjukkan efektivitas tertinggi. Penting untuk dicatat bahwa pemilihan jumlah K yang sesuai dapat memberikan dampak signifikan pada performa classifier, dan hasil eksperimen mendukung kesimpulan bahwa dengan menggunakan nilai K 29, kita dapat mencapai akurasi tertinggi dalam penyesuaian model terhadap data yang diberikan.

Kata Kunci: *Pembelajaran mesin, K-NN, Rumus jarak, Jarak Euclidien, Jarak Manhattan, Jarak Mahalanobis, Variasi nilai k, Akurasi*

PENDAHULUAN

Di era serba teknologi saat ini, pemanfaatan pengolahan data menjadi suatu kebutuhan yang tidak dapat dipungkiri, terutama dalam konteks kesehatan. Salah satu pemanfaatan dalam pengelolaan data Kesehatan adalah klasifikasi data teks penyakit, seperti penyakit cardiovascular. Penyakit ini merupakan salah satu penyebab utama kematian diseluruh dunia [1], Sehingga upaya indentifikasi,memprediksi atau klasifikasi kemungkinan terjadinya penyakit kardiovaskular sejak dini menjadi Langkah awal yang dapat mengurangi angka kematian dikarenakan penyakit Cardiovaskular.

Pengelompokan data dalam klasifikasi ditentukan oleh adanya label kelas atau target pada setiap titik data. Masalah klasifikasi dapat diselesaikan dengan beberapa jenis algoritma pembelajaran, salah satunya adalah supervisory learning. Data atau target yang diberi label penting dalam mencapai tingkat akurasi atau presisi yang diinginkan selama proses pembelajaran. Untuk klasifikasi, beberapa pilihan antara lain metode K-Nearest Neighbor (KNN).[2].

Algoritma K-Nearest Neighbor merupakan metode dalam supervised

learning yang beroperasi dengan menggunakan data yang sudah diberi label. Meskipun sederhana, algoritma ini telah berhasil menghasilkan kinerja yang tinggi dalam beberapa kasus. Pendekatan KNN ini melakukan klasifikasi terhadap suatu objek berdasarkan pada kesamaan atau jarak terdekat dengan objek-objek dalam data latih (training).

Pada tahap pembelajaran, algoritma ini menyimpan fitur-fitur vektor dan klasifikasi dari data latih. Sedangkan pada tahap klasifikasi, fitur-fitur yang sama dihitung untuk data yang akan diuji (yang klasifikasinya tidak diketahui). Langkah selanjutnya adalah menghitung jarak antara vektor baru dengan seluruh vektor dalam data latih, kemudian memilih sejumlah K vektor terdekat untuk menentukan klasifikasi dari titik-titik tersebut.

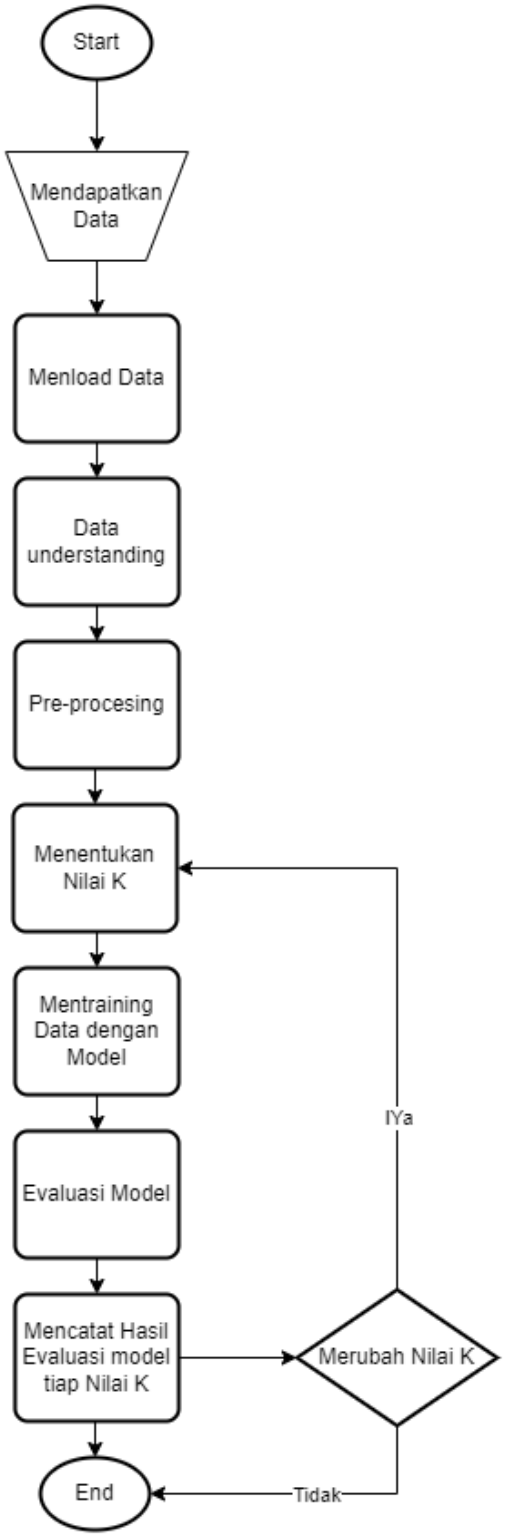
Perhitungan jarak antar data menjadi komponen penting dalam algoritma ini. KNN membandingkan data pelatihan dengan data baru untuk menentukan kesamaannya, yang sangat memengaruhi akurasi klasifikasi. Sebelum melakukan pengelompokan data untuk proses deteksi, ditetapkan terlebih dahulu ukuran jarak antar elemen data. Dalam berbagai aplikasi, beragam metode pengukuran jarak digunakan untuk menilai tingkat kemiripan data, seperti jarak Euclidean, Manhattan (City Block Distance), Mahalanobis, Korelasi, Berbasis Sudut, Minkowski, dan Squared Euclidean [3].

Banyak penelitian telah dilakukan untuk membandingkan peforma dari variasi teknik pengukuran jarak. Penelitian perbandingan akurasi dalam menerapkan metode pengukuran jarak (euclidean, manhattan, dan minkowski) pada pelabelan klaster status disparitas kebutuhan Guu menunjukkan hasil tingkat akurasi yang signifikan. Metode euclidean distance mencapai tingkat akurasi tertinggi sebesar 84.47%, diikuti oleh metode manhattan distance dan metode minkowski, keduanya dengan tingkat akurasi yang sama, yaitu 83.85%. Oleh karena itu, dapat disimpulkan bahwa metode euclidean menjadi pilihan terbaik yang dapat diimplementasikan dalam algoritma K-Means Clustering[4]. Hasil perhitungan jarak antar Kabupaten dan Kota di Sumatera Barat menggunakan persamaan Haversine dan Euclidean Distance menunjukkan bahwa tidak ada perbedaan yang signifikan antara keduanya. Perbedaan hasil antara kedua rumus tersebut tidak terlalu besar. Dengan demikian, matriks pembobot dari kedua persamaan tersebut memberikan hasil yang hampir sama, dan ketetanggaan antar wilayahnya juga serupa. Penggunaan persamaan Haversine dan Euclidean Distance dalam pengukuran jarak tidak menghasilkan perbedaan yang mencolok, sehingga keduanya dapat dianggap setara dalam konteks ini [5]. Presisi, F1 scoredan sensitifitas tertinggi adalah *distance metric euclidean*. *Distance metric chebyshev* memiliki nilai akurasi, presisi dan sensitivitas terendah sedangkan distance metric mahalanobis memiliki nilai F1 score terendah [6].

Berdasarkan penelitian yang dilakukan, terdapat variasi dalam kinerja metode pengukuran jarak dalam menggunakan KNN untuk berbagai aplikasi. Tujuan dari penelitian ini adalah untuk mengevaluasi kinerja metode KNN dengan menggunakan tiga jenis pengukuran jarak, yaitu Euclidean, Manhattan/City Block, dan Mahalanobis. Fokus utama penelitian ini adalah menganalisis akurasi data tekstual. Setiap metode jarak dibandingkan dengan berbagai nilai K mulai dari 1 hingga 31 dengan kelipatan ganjil. Evaluasi kinerja KNN dilakukan dengan mempertimbangkan akurasi dari setiap kombinasi metode jarak dan nilai K.

METODE PENELITIAAN

Pada penelitian ini ada beberapa tahap yang dilakukan dalam prosesnya, berikut beberapa proses yang digambarkan dengan flowchart.



Gambar 1. Flowchart Proses Penelitian

1. Mendapatkan Data

Pada penelitian ini menggunakan Dataset yang diperoleh dari Kaggle, yang berisi 68.2 ribu Data. Dengan judul data yaitu Cardiovascular Disease. Dataset ini

dapat di akses dari :

https://www.kaggle.com/datasets/colewelkins/cardiovascular-disease

Pada dataset tersebut terdapat beberapa variabel diantaranya adalah, id pasien, umur pasien, gender, berat badan, tinggi badan, tekanan darah systole, tekanan darah diastole, kadar kolesterol yang terdiri dari (normal, diatas normal, jauh diatas normal), kadar gula yang terdikiri dari 3 kategori (normal, diatas normal, jauh diatas normal), status pasien merokok atau tidak, status konsumsi alcohol, aktivitas fisik, memiliki penyakit kardiovaskular atau tidak, Body Mass Index (BMI), kategori tekanan darah (normal, hipertensi level 1, hypertensi level 2, dan hipertensi parah).

2. Menload Data

Proses memuat data merupakan langkah kritis dalam perjalanan analisis data atau pengembangan model, di mana kita membaca dan mengimpor dengan cermat data yang telah diunduh ke dalam lingkungan pemrograman atau alat analisis yang kita gunakan. Langkah awal ini sangat vital, karena memungkinkan kita untuk mengelola dataset dengan lebih efisien dan menjalankan berbagai manipulasi atau analisis yang dibutuhkan. Dalam konteks penelitian ini, kita membuat penggunaan bijak dari library Pandas untuk menyederhanakan proses ini, dengan mengimpor data ke dalam kode menggunakan Pandas di dalam text editor Visual Studio Code.

Dengan mengandalkan kehandalan Pandas, data yang sudah diunduh dapat dengan mudah diolah dan dimanipulasi. Penggunaan struktur data tabular yang dikenal sebagai DataFrame memungkinkan kita untuk melakukan berbagai operasi, seperti menyaring data, melakukan pengindeksan, dan transformasi data, semuanya dengan keterbacaan kode yang tinggi.

Inisiasi proses load data ini bukan hanya sekadar langkah teknis, tetapi juga fondasi yang kokoh untuk pengembangan model atau analisis lebih lanjut. Dengan menggunakan Pandas di dalam lingkungan Visual Studio Code, kita membangun dasar yang solid untuk menjalankan penelitian, memastikan bahwa data yang kita miliki dapat diakses dan dimanipulasi sesuai kebutuhan analisis atau pengembangan model yang tengah dilakukan.

3. Data Understanding

Tahap Data Understanding memegang peran sentral dalam proses analisis data di ranah ilmu data dan statistika. Tujuannya adalah meraih wawasan yang mendalam terkait data yang akan menjadi fokus analisis. Dalam konteks ini, beberapa kegiatan esensial dilakukan untuk memastikan pemahaman komprehensif dan dokumentasi karakteristik dataset. Eksplorasi dataset menjadi kegiatan utama, diikuti oleh pemahaman mendalam terhadap atribut, hubungan antar atribut, identifikasi nilai yang hilang, penanganan nilai ekstrem, dan sebagainya.

Pada tahap ini, upaya dilakukan untuk mendapatkan pemahaman awal mengenai dataset yang digunakan. Proses ini mencakup eksekusi beberapa langkah kunci, seperti menampilkan dan mengevaluasi tipe data setiap kolom, melakukan pemeriksaan terhadap keberadaan data yang kosong dalam setiap kolom, meninjau penyebaran data untuk mengidentifikasi tren atau pola yang mungkin ada, mengevaluasi keseimbangan data, dan menganalisis korelasi antar data.

Langkah-langkah ini membantu menciptakan fondasi yang kokoh untuk analisis data yang lebih mendalam. Dengan menyajikan tipe data, memeriksa integritas data, dan mengidentifikasi pola korelasi, tahap Data Understanding ini

memberikan landasan yang kaya informasi untuk perencanaan langkah-langkah analisis selanjutnya. Dengan demikian, fokus pada pemahaman mendalam data menjadi kunci untuk memastikan analisis yang tepat dan hasil yang lebih akurat dalam pengembangan model atau interpretasi dataset.

4. Preprocesing data

Preprocessing data adalah langkah-langkah yang dilakukan untuk membersihkan dan menyiapkan data sebelum dilakukan analisis atau pembuatan model. Tujuan dari preprocessing adalah untuk mengoptimalkan kualitas data dan meningkatkan kinerja model yang akan dibangun. Beberapa teknik preprocessing data umum meliputi: Mengidentifikasi dan menangani nilai yang hilang, seperti menghapus baris atau kolom yang memiliki nilai yang hilang, atau mengisi nilai yang hilang dengan nilai rata-rata, median, atau modus (handling missing values), Mendeteksi dan mengelola nilai-nilai ekstrem atau outlier yang dapat mempengaruhi hasil analisis atau model. Teknik yang umum digunakan termasuk penghapusan, penggantian, atau transformasi nilai outlier (handling outlier), Memastikan bahwa variabel-variabel numerik memiliki skala yang seragam, agar tidak ada variabel yang mendominasi yang lainnya. Ini sering dilakukan dengan penskalaan (scaling) atau normalisasi (penskalaan dan normalisasi), Mengubah variabel kategorikal menjadi bentuk numerik, misalnya dengan menggunakan metode one-hot encoding atau label encoding (Pengkodean Variabel Kategorikal), pengembangan fitur, data splitting, handling imbalance data, dan text cleaning.

Pada tahap ini, Data di proses agar sesuai dengan kebutuhan pembelajaran mesin, seperti Menghapus kolom data yang tidak dibutuhkan atau tidak berguna seperti data ID, Kemudian Data transformasi yaitu merubah data objek menjadi type data numerik agar nantinya dapat dilakukan perhitungan matematis pada rumus jarak, kemudian memisahkan data feature dan data label, kemudian melakukan oversampling data untuk menyeimbangkan data tiap label, Kemudian memisahkan data training dan data testing dengan proporsi data testing 20% dan data training 80% untuk melakukan pembelajaran mesin serta untuk evaluasi model nantinya, Dan melakukan normalisasi data agar data menjadi standart dan nilai tiap data tidak memberatkan satu dan lain pada perhitungan matematis jarak nanti.

5. Menentukan Nilai K

Menentukan nilai optimal untuk parameter k pada algoritma K-Nearest Neighbors (KNN) adalah bagian yang penting dalam membangun model yang baik. Parameter k menentukan jumlah tetangga terdekat yang akan diambil dalam memprediksi label atau nilai target suatu observasi.

Pada tahap ini melakukan inisialisasi nilai dimulai dari angka 1 hingga nantinya angka 31 dengan kelipatan ganjil, nantinya setiap nilai K akan di evaluasi hasil akurasi dari tiap modelnya.

6. Mentraining data dengan model

Training model untuk algoritma K-Nearest Neighbors (KNN) melibatkan proses memasukkan data pelatihan ke dalam model KNN. Dalam KNN, model hanya "menghafal" atau menyimpan data pelatihan dalam memori, dan prediksi dilakukan dengan melihat k-nearest neighbors dari data uji.

Pada tahap ini tiap model KNN dengan rumus jarak yaitu Euclidien, Manhattan, Dan Mahalanobis melakukan training dengan data training serta dengan K yang telah ditentukan sebelumnya.

7. Evaluasi Model

Evaluasi model K-Nearest Neighbors (KNN) melibatkan pengukuran kinerja model terhadap data yang tidak digunakan selama pelatihan, biasanya disebut sebagai data uji. Beberapa metrik evaluasi umum yang dapat digunakan untuk mengukur performa model KNN termasuk akurasi (accuracy), presisi (precision), recall, F1-score, dan matriks konfusi.

Pada tahap ini model dilakukan evaluasi dengan melihat hasil akurasi yang didapat dengan data testing serta menggunakan library sklearn yaitu accuracy_score.

8. Mencatat Hasil Evaluasi tiap Model

Pada tahap ini, Setiap hasil akurasi yang didapat dari 3 model berdasarkan nilai K tertentu dicatat dan ditentukan mana akurasi terbaik dari 3 model tersebut, kemudian Kembali lagi kelangkan menentukan K selanjutnya dan mentraining lagi serta melakukan evaluasi lagi, begitu seterusnya hingga nilai K yaitu 31.

HASIL DAN PEMBAHASAN

Dari penelitian mendapatkan hasil nilai akurasi pada setiap algoritme jarak yang digunakan dengan inputan data testing. Hasil akurasi yang diperoleh didapatkan dari variasi nilai k dari angka 1 hingga 31 dengan keliapatan ganjil.

Rumus/K	Euclidien	Manhattan	Mahalanobis
1	0.6436948023744028	0.6442015346749674	0.6455769509193572
3	0.683219921818445	0.6807586506442739	0.6788765020993195
5	0.6945852034168235	0.6933545678297379	0.6948747647314318
7	0.702837700883162	0.6985666714926886	0.7035616041696829
9	0.7085565368466773	0.7053713623859852	0.7053713623859852
11	0.71384103083828	0.7109454176921963	0.7108730273635442
13	0.7142753728101926	0.7166642536557116	0.7126103952511944
15	0.7158679600405385	0.716157521355147	0.7165194729984075
17	0.7163746923411032	0.7168814246416678	0.7171709859562763
19	0.7178224989141451	0.7179672795714492	0.7163023020124512
21	0.7163746923411032	0.7181844505574055	0.7180396699001014
23	0.7160851310264948	0.7182568408860576	0.7179672795714492
25	0.718546402200666	0.7187635731866223	0.7178948892427972
27	0.718618792529318	0.7190531345012307	0.7187635731866223
29	0.7194874764731432	0.7196322571304473	0.7204285507456204
31	0.7180396699001014	0.7188359635152743	0.7186911828579702
Total Akurasi Terbaik	3	7	6

Table 1. Akurasi tiap model terhadap nilai K

Hasil dari perhitungan jarak Euclidien, Manhattan, Dan Mahalanobis terhadap akurasi K-NN pada table, menunjukkan hasi; akurasi yang apabila semakin banyak nilai k menyebabkan kenaikan akurasi, walaupun ada kondisi dimana nilai K tertentu akurasi yang menurun walau tidak siknifikan. total Akurasi terendah diperoleh dari jarak Euclidien, sedangkan total akurasi tertinggi diraih oleh jarak Manhattan dan disusul oleh jarak Mahalanobis.

KESIMPULAN

Jarak Manhattan dan Mahalanobis menghasilkan total akurasi terbaik terbanyak serta menghasilkan akurasi terbaik pada Sebagian besar ukuran K, sedangkan Euclidien pada penelitian ini menjadi perhitungan dengan total akurasi

terbaik yang sedikit. Meningkatkan jumlah K pada setiap perhitungan jarak dapat meningkatkan akurasi klasifikasi. Jumlah K yang optimal dalam eksperimen ini adalah 29, menunjukkan efektivitas tertinggi. Penting untuk dicatat bahwa pemilihan jumlah K yang sesuai dapat memberikan dampak signifikan pada performa classifier, dan hasil eksperimen mendukung kesimpulan bahwa dengan menggunakan nilai K=29, kita dapat mencapai akurasi tertinggi dalam penyesuaian model terhadap data yang diberikan.

DAFTAR PUSTAKA

[1] S. Fadlilah, A. Sucipto, and T. Amestiasih, "Usia, Jenis Kelamin, Perilaku Merokok, dan IMT Berhubungan dengan Resiko Penyakit Kardiovaskuler," *Jurnal Keperawatan*, vol. 11, no. 4, pp. 261–268, Dec. 2019, doi: 10.32583/keperawatan.v11i4.558.

[2] A. Roihan, P. A. Sunarya, and A. S. Rafika, "Pemanfaatan Machine Learning dalam Berbagai Bidang," *Jurnal Khatulistiwa Informatika*, vol. 5, no. 1, p. 490845, 2020.

[3] W. Wahyono, I. N. P. Trisna, S. L. Sariwening, M. Fajar, and D. Wijayanto, "Comparison of distance measurement on k-nearest neighbour in textual data classification," *Jurnal Teknologi dan Sistem Komputer*, vol. 8, no. 1, pp. 54–58, Jan. 2020, doi: 10.14710/jtsiskom.8.1.2020.54-58.

[4] M. Nishom, "Perbandingan Akurasi Euclidean Distance, Minkowski Distance, dan Manhattan Distance pada Algoritma K-Means Clustering berbasis Chi-Square," *Jurnal Informatika: Jurnal Pengembangan IT*, vol. 4, no. 1, pp. 20–24, Jan. 2019, doi: 10.30591/jpit.v4i1.1253.

[5] V. H. Nabilla, Indonesia, Dony Permana, and Fadhilah Fitri, "Comparison of Haversine and Euclidean Distance Formula for Calculating Distance Between Regencies in West Sumatra," *UNP Journal of Statistics and Data Science*, vol. 1, no. 3, pp. 120–125, May 2023, doi: 10.24036/ujsds/vol1-iss3/39.

[6] K. F. Margolang, M. M. Siregar, S. Riyadi, and Z. Situmorang, "Analisa Distance Metric Algoritma K-Nearest Neighbor Pada Klasifikasi Kredit Macet," *Journal of Information System Research (JOSH)*, vol. 3, no. 2, pp. 118–124, Feb. 2022, doi: 10.47065/josh.v3i2.1262.