JURNAL ILMIAH BETRIK
Besemah Teknologi Informasi dan Komputer
Editor Office : LPPM Sekolah Tinggi Teknologi Pagar Alam, Jln. Masik Siagim No. 75
Simpang Mbacang, Pagar Alam, SUM-SEL, Indonesia
Phone : +62 852-7901-1390.
Email : betrik@sttpagaralam.ac.id | admin.jurnal@sttpagaralam.ac.id
Website : https://ejournal.sttpagaralam.ac.id/index.php/betrik/index
ISSN : 2339 - 1871

ANALISIS KOMPARASI ALGORITMA K-NEAREST NEIGHBOR DAN SUPPORT VECTOR MACHINE DENGAN PENDEKATAN MULTI DATASET

Julyan Adi Saputra[1], Syaeful Anas Aklani[2]
Program studi Sistem Informasi Universitas Internasional Batam[12]
Jl. Gajah Mada, Tiban Indah, Kec. Sekupang, Kota Batam, Kepulauan Riau
Sur-el : 1931060.julyan@uib.edu [1], syaeful.anas@uib.ac.id [2]

Abstrak*:* Data mining merupakan proses untuk mengidentifikasi data yang valid dan memiliki potensi yang berguna untuk pelaku yang melakukannya. Salah satu dari tujuan dilakukan data mining adalah untuk mempelajari data yang sudah ada sebelumnya yang menyusun pola tertentu dan digunakan untuk melakukan prediksi. Pembelajaran mesin menggunakan data dan algoritma untuk membuat model dari kumpulan data tersebut. Terdapat banyak sekali algortima yang dapat dipakai yaitu C4.5, K-Means, Support Vector Machine (SVM), K-Nearest Neighbor (K-NN), Naïve Bayes, dan lainnya. Dikarenakan banyaknya algoritma dalam data mining tentu masing - masing memiliki kelebihan dan kekurangan tersendiri. Dalam penelitian ini akan berfokus pada perbandingan antara algoritma Support Vector Machine dan algoritma K-Nearest Neighbor dalam hal akurasi, presisi dan waktu proses yang dihasilkan.

Kunci Utama: K-Nearest Neighbor; Pembelajaran Mesin; Perbandingan Algortima; Support Vector Machine

Abstract*: Data mining is a process of identifying data that is valid and has the potential to be useful to the person who did it. One of the purposes of data mining is to study previously existing data that composes certain patterns and is used to make predictions. Machine learning works by utilizing data and algorithms to create models with patterns from the data set. There are many algorithms that can be used, such as C4.5, K-Means, Support Vector Machine (SVM), K-Nearest Neighbor (K-NN), Naïve Bayes, and others. Since there are many algorithms in data mining, each has its own advantages and disadvantages. This research will focus on the comparison between the Support Vector Machine algorithm and the K-Nearest Neighbor algorithm in terms of accuracy, precision and processing time.*

Keywords *: Algorithm Comparison; Machine Learning; K-Nearest Neighbor; Support Vector Machine*

1. PENDAHULUAN

Pembelajaran mesin telah menjadi tren baru. Pembelajaran mesin menggunakan data dan algoritma untuk membuat model dari kumpulan data tersebut. Selain itu, machine learning juga mempelajari bagaimana model yang dihasilkan dapat memprediksi produksi berdasarkan model yang ada [1]. Machine learning merupakan cabang dari kecerdasan buatan yang berfokus pada pengembangan sebuah sistem yang dapat

belajar sendiri tanpa harus diprogram berulang kali.

Sebelum menghasilkan output dalam machine learning, diperlukan data untuk dilakukan pembelajaran. Untuk penerapanya juga tidak bisa secara umum atau untuk semua permasalahan. Data merupakan faktor terpenting dalam pemodelan machine learning yang digunakan untuk proses pembelajaran, terlebih lagi pemodelan yang berfokus pada teks [2]. Proses pembelajaran ini menggunakan algoritma yang juga merupakan aplikasi dari statistika. Data dibagi menjadi dua yaitu data untuk pembelajaran dan juga untuk uji, tujuan dari dibaginya data adalah untuk memastikan keefektifan model yang dibentuk.

Distribusi data yang digunakan bervariasi tergantung pada algoritma. Data adalah kumpulan informasi atau fakta tentang sesuatu, yang diperoleh melalui observasi atau dengan mencari sumber-sumber tertentu yang terdiri dari simbol, angka, kata-kata, maupun kalimat. Manfaat data juga beragam yaitu sebagai dasar perencanaan, dasar untuk membuat keputusan, bahan untuk evaluasi, dan lainnya. Penambangan data merupakan proses penemuan dalam dataset besar yang sebelumnya tidak diketahui [3].

Salah satu fungsi dari machine learning adalah untuk menganalisis data. Analisis data adalah suatu proses dimana data yang terkumpul diolah sedemikian rupa sehingga hasil dari proses tersebut dapat ditentukan dan dari situ dicari solusinya sebagai dasar permasalahan yang akan dipecahkan. Dalam melaksanakan analisis data, tahap pengumpulan data sangatlah penting [4]. Tujuan dari analisis data adalah untuk menginterpretasikan informasi dalam bentuk yang mudah dipahami dan menemukan atau menarik kesimpulan berdasarkan pengujian hipotesis.

Dalam dunia yang serba digital, analisis data memiliki peranan yang penting dalam menentukan kesuksesan sebuah perusahaan. Proses analisis menuntut perusahaan untuk menentukan hal penting pada waktu tertentu. Dan juga analisis data berperan dalam memperoleh informasi mengenai perkembangan bisnis [5]. Jika sebuah perusahaan dapat menganalisa data dengan akurat, maka perusahaan tersebut dapat mengambil tindakan yang diperlukan untuk meningkatkan kinerja maupun menghindari krisis dalam menghadapi permasalahan.

Dalam analisis data banyak sekali jenis algoritma yang dapat digunakan. Masing – masing algoritma memiliki keunggulan tersendiri dalam menganalisa. SVM adalah algoritma populer untuk memecahkan masalah klasifikasi. Data berdimensi tinggi dapat diklasifikasikan dengan benar menggunakan algoritma ini, memungkinkan SVM memberikan hasil prediksi dengan akurasi tinggi. Sedangkan K-NN dikenal sebagai algoritma lazy learning karena sebelum dijalankan, diperlukan pelatihan singkat [6]. Algoritma ini juga memiliki keuntungan karena kuat terhadap data training yang tedapat banyak noise serta dengan ukuran data training yang berdimensi besar [7].

Dalam penelitian ini akan berfokus pada perbandingan antara algoritma *Support Vector Machine* dan algoritma *K-Nearest Neighbor*. Tujuan utama dari penelitian ini adalah melakukan perbandingan dilakukan dengan tujuan menentukan algoritma manakah yang menghasilkan kinerja yang lebih maksimal, terutama antara algoritma *Support Vector Machine* dan *K-Nearest Neighbor.*

2. METODE PENELITIAN

2.1 Landasan Teori

A. Sistem Informasi

Sistem yang terdiri dari kombinasi manusia, fasilitas, teknologi, media, prosedur-prosedur, dan pengendalian

dalam suatu organisasi yang berfungsi untuk menyajikan suatu dasar informasi untuk pengambilan keputusan [8]. Menurut [9] dalam penelitiannya, ada enam komponen yang mencakup semua jenis sistem informasi, yaitu:

1. **Hardware**, merupakan perangkat keras berbasis komputer dan semua komponennya yang dapat digunakan secara fisik seperti monitor, keyboard, printer, central processing unit, dan lain-lain.
2. **Software**, merupakan perangkat lunak dari sistem informasi berupa operating system, aplikasi yang digunakan untuk mengolah, mengatur dan menganalisis data.
3. ***Database***, berupa kumpulan data di dalam sistem informasi yang tersusun dalam tabel atau file.
4. ***Network***, merupakan alat yang menghubungkan antar subsistem, sehingga memungkinkan adanya interaksi berupa pertukaran data dan informasi.
5. **Prosedur**, berupa gambaran bagaimana data tertentu diproses untuk menghasilkan sebuah output dari sistem informasi.
6. ***User***, merupakan pihak yang bertanggung jawab dalam penggunaan dan pengembangan sistem informasi.

B. Machine Learning

Dalam penelitian [10] menyebutkan bahwa pembelajaran mesin adalah cabang ilmu yang merupakan bagian dari kecerdasan buatan dan diprogram untuk mengaktifkan komputer cerdas untuk berperilaku seperti manusia dan meningkatkan pemahaman mereka melalui pembelajaran otomatis.

Pembelajaran mesin berfungsi saat data dimasukkan ke dalam analisis data besar untuk menemukan pola tertentu. Berdasarkan teknik pembelajarannya, machine learning dibedakan menjadi supervised learning, unsupervised learning, semi supervised learning, dan reinforcement learning.

C. Support Vector Machine

Support Vector Machine merupakan algoritma supervised learning yang mengelompokkan kasus linear yang mampu dipisahkan (dikelompokan) dan dibagi sesuai dengan kelas ataupun hubungan sebab akibat nya [11]. Dalam hal identifikasi hyperplane tertentu SVM mampu memaksimalkan margin antar kelas berbeda [12]

D. K-Nearest Neighbor

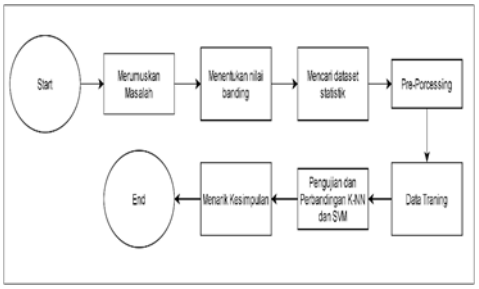
Menurut [13] algoritma *K-Nearest Neighbor (K-NN)* merupakan algoritma supervised learning untuk memecahkan sebuah kasus dengan menghitung antara kasus baru dengan kasus lama dengan memperhatikan jarak yang paling dekat.

E. Rapid Miner

RapidMiner adalah perangkat lunak data untuk persiapan data, pembelajaran mesin, penambangan teks, dan analitik prediktif. [14]. Terdapat koleksi dari algoritma pembelajaran mesin pada rapidminer yang digunakan untuk melakukan data mining. Tools yang ada pada rapidminer berfungsi untuk data pre-processing, regresi, clustering, klasifikasi, rule association dan visualisasi data tersebut kedalam bentuk yang mudah untuk dapat dipahami [15]. Salah satu manfaat dari melakukan data mining berupa klasifikasi data yang berupa teknik dalam mengelompokan data berdasarkan hubungan data terhadap data sampel.

2.2 Alur Penelitian

Alur penelitian dibuat dengan tujuan agar permasalahan dari topik yang diangkat penulis dapat terjawab secara sistematis.



Gambar 1. Alur Penelitian

2.3 Penentuan Nilai Banding

Penelitian ini membandingkan Support Vector Machine dan K-Nearest Neighbor dari segi:

A. Akurasi

Akurasi adalah total prediksi yang benar dari data yang ada dengan rumus :

$$\text{Akurasi} = \frac{TP+TN}{TP+TN+FP+FN}$$

B. Waktu Proses

Merupakan kecepatan dari algoritma dalam menyelesaikan prosesnya.

C. Presisi

Presisi adalah akurasi antar data yang diminta dan hasil prediksi yang diberikan, dengan rumus:

$$\text{Presisi} = \frac{TP}{TP+FP}$$

2.4 Mencari Dataset Statistik

Dalam penelitian ini akan digunakan total 2 dataset dari sumber yang berbeda. Dataset statistik yang digunakan pada penelitian ini didapatkan melalui aplikasi google chrome. Keyword yang digunakan adalah "dataset statistik" dan "open dataset".

Dataset yang didapat berjumlah dua yang berisikan data pemegang kartu kredit dengan format file berupa Comma Separated Value File (.csv) yang dapat dibuka menggunakan software Microsoft Excel. Terdapat perbedaan pada kedua dataset tersebut yaitu atribut dan juga jumlah data record.

2.5 Pre-Processing Data

A. Data Cleaning

Data yang diperoleh akan dilakukan proses pembersihan data atau *cleaning* data. Proses ini dilakukan pada aplikasi microsoft excel dengan menggunakan fitur yang bernama sort & filter untuk mensortir data.

Tujuan dilakukan sortir data ini adalah untuk mengurutkan data, sehingga mempermudah penulis dalam mencari data dengan atribut yang kosong dan setelah itu dilakukan penghapusan pada data tersebut. Setelah itu penulis juga menggunakan fitur conditional formatting untuk menyorot data duplikat.

B. Data Integration

Integrasi data merupakan penggabungan dataset dari berbagai sumber yang berbeda. Proses ini dilakukan langsung dalam software rapidminer dengan memanfaatkan tools yang ada. Tools yang digunakan berupa "Union", tools ini dapat digunakan untuk mengabungkan 2 dataset yang berbeda dengan kondisi atribut dari 2 dataset tersebut sama maupun berbeda.

C. Data Transformation

Transformasi data ini dilakukan dengan tujuan utamanya adalah mengubah skala pengukuran dari data asli ke dalam bentuk yang berbeda. Dalam tahap ini, penulis mengubah atribut "Income_Category" ke dalam bentuk yang lebih sederhana. Bentuk awalnya adalah campuran dari kata, simbol dan angka ditransformasikan ke dalam bentuk hanya angka saja. Contoh data yang telah ditransformasi dapat dilihat pada Tabel 1 berikut.

Tabel 1. Transformasi Data

Sebelum	Sesudah
$120K +	5
$80K - $120K	4
$60K - $80K	3
$40K - $60K	2
Less than $40K	1

D. Data Reduction

Tahapan data reduction atau pengurangan data ini dilakukan dengan tujuan untuk mengurangi data yang dikelola. Terdapat beberapa cara dalam tahapan ini seperti pengurangan dimensi, pengurangan jumlah data dan kompresi data. Penulis menggunakan pengurangan jumlah dalam tahapan ini. Jumlah data yang awalnya 10.787 menjadi 10.325.

3. HASIL DAN PEMBAHASAN

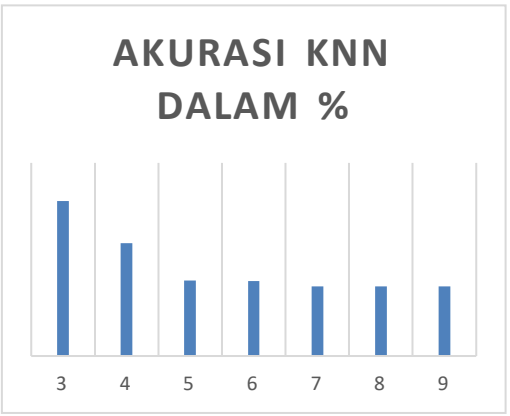
Setelah melalui tahap pre-processing, tahapan berikutnya adalah membagi data menjadi dua bagian dengan rasio 70:30. Dalam penelitian ini perangkat keras maupun lunak yang digunakan memiliki rincian spesifikasi seperti tabeldibawah ini.

Tabel 2. Spesifikasi Laptop

Perangkat	Spesifikasi
Sistem Operasi	Windows 11
Tipe Sistem	64-Bit
Prosesor	Intel® Core® i5 8th Gen
Memori (RAM)	12GB
Penyimpanan	-1 TB HDD -512 GB SSD M.2 Sata
Perangkat Lunak	-Ms. Excel -RapidMiner Studio

3.1 Pencarian Nilai K Terbaik padaAlgoritma K-NN

Evaluasi dilakukan untuk menentukan nilai K terbaik pada K-NN. Nilai K yang dinilai baik adalah nilai K yang memiliki tinggat akurasi tertinggi antara K=3 sampai dengan K=9. Dan nilai K yang tertinggi akan dipilih untuk dibandingkan dengan algoritma SVM. Selain itu, akan digunakan juga untuk menilai akurasi dengan cross validation. Nilai K yang telah diperoleh dapat dilihat pada gambar berikut ini dalam bentuk grafik batang.



Gambar 2. Grafik Persentase Akurasi K pada K-NN

Dapat dilihat bahwa akurasi dengan nilai tertinggi yang diperoleh yaitu sebesar 95,82% dengan nilaik K=3. Maka selanjutnya nilai tersebut akan digunakan sebagai pembanding dengan algoritma SVM.

3.2 Perbandingan Algortima K-NN dan SVM

Hasil evaluasi yang di dapat menunjukan bahwa algoritma K-NN dengan nilai K=3 lebih baik dari SVM dengan nilai akurasi sebesar 95,82% dalam waktu proeses selama 48,27 detik. Berikut berupa tabel yang menunjukan data hasil berupa akurasi dan juga waktu proses dari kedua algortima SVM dan K-NN.

Tabel 3. Hasil Akurasi dan Waktu Proses

Model	Akurasi (%)	Waktu (s)
K=3	95,82	48,27
K=4	94,51	47,86
K=5	93,24	48,55
K=6	93,33	48,13
K=7	93,17	48,52
K=8	93,16	48,43
K=9	93,16	48,18
SVM	93,21	49,71

Dan untuk hasil dari perhitungan recall, precision dan juga F1 score dapat dilihat dari tabel berikut ini.

Tabel 4. Hasil Recall, Precision dan F1 Score

Model	Recall (%)	Precison (%)	F1 Score (%)
K=3	100	95,70	97,80
K=4	99,96	94,49	97,15
K=5	99,94	93,40	96,56
K=6	99,99	93,32	96,54
K=7	99,99	93,18	96,54
K=8	100	93,16	96,54
K=9	100	93,16	96,54
SVM	99,96	93,24	96,48

3.3 Perbandingan Algoritma K-NN dan SVM dengan K-Fold Cross Validation

Dari hasil validasi k-Fold Cross Validation dengan nilai k=10 didapatkan bahwa algoritma SVM mendapatkan hasil akurasi yang lebih maksimal, dengan nilai akurasi sebesar 93,09%. Sementara K-NN dengan nilai K=3 sebesar 91,79%. Namun, untuk waktu proses K-NN lebih cepat diangka 16,11 detik sedangkan SVM ada pada 39,96 detik. Dan dari segi presisi, K-NN berada diangka 93,21% yang dimana lebih unggul dari SVM yang mendapat 93,15%.

Hasil dari validasi perbandingan SVM dan K-NN dapat dilihat pada tabel berikut.

Tabel 5. Hasil Perbandingan dengan Cross Valdiation

Model	Akurasi (%)	Precison (%)	Waktu (s)
K=3	91,79	93,21	16,11
SVM	93,09	93,15	39,96

4. SIMPULAN

Berdasarkan penelitian penulis berupa perbandingan algortima dalam data mining yaitu K-NN dan SVM kesimpulan yang diperoleh yaitu:

A. Didapat hasil bahwa dalam pengujian pada kedua algortima, K-NN memperoleh akurasi yang lebih tinggi dengan K=3 mendapat nilai akurasi sebesar 95,82% tanpa validasi K-Fold Cross Validation. Sedangkan dengan K-Fold Cross Validation SVM lebih unggul dengan akurasi 93.09%.

B. Dalam kecepatan proses, K-NN lebih cepat dengan waktu proses 48,27 detik tanpa K-Fold Cross Validation dan 16,11 detik dengan K-Fold Cross Validation.

C. Algoritma K-NN menghasilkan presisi yang lebih baik diangka 95,70% tanpa K-Fold Cross Validation dan 93,21% dengan K- Fold Cross Validation.

D. Pada penelitian ini, dapat disimpulkan bahwa algoritma K-Nearest Neighbor memiliki performa yang lebih baik dibandingkan dengan *Support Vector Machine.*

DAFTAR RUJUKAN

[1] M. R. A. Nasution and M. Hayaty, "Perbandingan Akurasi dan Waktu Proses Algoritma K-NN dan SVM dalam Analisis Sentimen Twitter," *J. Inform.*, vol. 6, no. 2, pp. 226–235, 2019, doi: 10.31311/ji.v6i2.5129.

[2] R. Ferdiana, F. Jatmiko, D. D. Purwanti, A. S. T. Ayu, and W. F. Dicka, "Dataset Indonesia untuk Analisis Sentimen," *J. Nas. Tek. Elektro dan Teknol. Inf.*, vol. 8, no. 4, pp. 334–339, 2019.

[3] H. Suroyo, "Penerapan Machine Learning dengan Aplikasi Orange Data Mining Untuk Menentukan Jenis Buah Mangga," *Semin. Nas. Teknol. Komput. Sains*, vol. 1, no. 1, pp. 343–347, 2019, [Online]. Available: https://prosiding.seminar-id.com/index.php/sainteks/article/view/177

[4] A. Rijali, "Analisis Data Kualitatif," *Alhadharah J. Ilmu Dakwah*, vol. 17, no. 33, p. 81, 2019.

[5] I. Setiawan, "Perbedaan Data Enginner, Data Scientist Dan Data Analyst," *Widya Accarya*, vol. 12, no. 2, pp. 306–309, 2021.

[6] D. Kusumaningrum and E. M. Imah, "Studi Komparasi Algoritma Klasifikasi Mental Workload Berdasarkan Sinyal EGG," *J. Sist. Cerdas*, vol. 3, no. 2, pp. 133–143, 2020.

[7] R. Rahmiati, D. Irfan, A. Agustin, and S. Hediyati, "Aplikasi Pengukur Tingkat Sentimen Pelanggan Berdasarkan Komplain Pelanggan PLN Menggunakan Algoritma K-Nearest Neighbor," *INOVTEK Polbeng - Seri Inform.*, vol. 5, no. 2, pp. 332–346, 2020.

[8] A. Maulana, M. Sadikin, and A. Izzuddin, "Implementasi Sistem Informasi Manajemen Inventaris Berbasis Web Di Pusat Teknologi Informasi Dan Komunikasi – BPPT," *Setrum Sist. Kendali-Tenaga-elektronika-telekomunikasi-komputer*, vol. 7, no. 1, p. 182, 2018.

[9] A. Rochman, T. Tullah, and A. Rahman, "Sistem Informasi Data Pasien di Klinik Aulia Medika Pasarkemis," *Sisfotek Glob.*, vol. 9, no. 1, pp. 1–6, 2019.

[10] E. Retnoningsih and R. Pramudita, "Mengenal Machine Learning Dengan Teknik Supervised Dan Unsupervised Learning Menggunakan Python," *Bina Insa. Ict J.*, vol. 7, no. 2, p. 156, 2020, doi: 10.51211/biict.v7i2.1422.

[11] A. Handayanto, K. Latifa, N. D. Saputro, and R. R. Waliansyah, "Analisis dan Penerapan Algoritma Support Vector Machine (SVM) dalam Data Mining untuk Menunjang Strategi Promosi," *JUITA J. Inform.*, vol. 7, no. 2, p. 71, 2019, doi: 10.30595/juita.v7i2.4378.

[12] M. Azhari, Z. Situmorang, and R. Rosnelly, "Perbandingan Akurasi, Recall, dan Presisi Klasifikasi pada Algoritma C4.5, Random Forest, SVM dan Naive Bayes," *J. Media Inform. Budidarma*, vol. 5, no. 2, p. 640, 2021, doi: 10.30865/mib.v5i2.2937.

[13] E. Nasri and A. S. AW, "Aplikasi Seleksi Penentuan Nasabah Untuk Penjualan Barang Secara Kredit Dengan Algoritma K-Nearest Neighbor," *J. Ilm. Sains Dan Teknol.*, vol. 4, no. 1, pp. 1–11, 2020.

[14] S. M. Dewi, A. P. Windarto, and D. Hartama, "Penerapan Datamining Dengan Metode Klasifikasi Untuk Strategi Penjualan Produk Di Ud.Selamat Selular," *KOMIK (Konferensi Nas. Teknol. Inf. dan Komputer)*, vol. 3, no. 1, pp. 617–621, 2019.

[15] G. Ramadhan, A. P. Windarto, E. Irawan, W. Saputra, and H. Okprana, "Penerapan Data Mining Menggunakan Algoritma C4.5 Dalam Mengukur Tingkat Kepuasan Pasien BPJS," *Pros. Semin. Nas. Ris. Dan Inf. Sci.*, vol. 2, pp. 376–385, 2020.