



Pemanfaatan Statistika dalam Bidang Informatika

Fadla Anas Tapiani^{1*}, Pratiara Tri Anggraeni², Purwadi³

¹⁻³Universitas Amikom Purwokerto, Indonesia

E-mail: ¹fadlaanas580@gmail.com, ²pratiaratrianggraeni@gmail.com,

³purwadi@amikompurwokerto.ac.id

Article Info

Article history:

Received June 03, 2026

Revised June 22, 2026

Accepted June 27, 2026

Keywords:

Statistics; Informatics;
Classification; Regression;
Machine Learning.

ABSTRACT

The very rapid development of information technology in the last ten years has led to a rapid increase in the amount of data in various areas of life, such as social media, digital financial services, online shopping, and population and public health data. The enormous growth of data requires data processing methods that can extract meaningful information quickly, precisely and accurately. Statistics, as a science that studies how to collect, process, analyze, interpret and present data, plays an important role in supporting the development of the field of informatics, especially in the fields of data mining and machine learning. This article aims to explore the use of various statistical and machine learning methods in the field of informatics, based on a review of eleven previous studies published between 2023 and 2024. This research uses a literature review method by grouping references based on the type of statistical method used, namely classification methods and regression methods. Furthermore, the research analyzes application patterns, research objects, and the level of accuracy produced in each study. The classification methods analyzed include K-Nearest Neighbors, Convolutional Neural Network, Naive Bayes Classifier, and Classification and Regression Tree, while the regression methods analyzed are simple linear regression and multiple linear regression. The research results show that classification methods are usually used to solve problems that produce categories, such as sentiment analysis and business feasibility assessments, with accuracy levels ranging from 68 percent to 92 percent. Meanwhile, regression methods are used more to predict numbers such as the price of goods, population, and number of infectious disease cases, with accuracy levels that tend to be higher. This research shows that the combination of classical statistical concepts and machine learning algorithms provides great benefits in helping decision making based on data in various areas of life, so that statistics cannot be separated from the development of informatics in the current digital era.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



Article Info

Article history:

Received June 03, 2026

Revised June 22, 2026

Accepted June 27, 2026

ABSTRACT

Perkembangan teknologi informasi yang sangat cepat dalam sepuluh tahun terakhir menyebabkan peningkatan pesat jumlah data di berbagai bidang kehidupan, seperti media sosial, layanan keuangan digital, belanja online, hingga data penduduk dan kesehatan masyarakat. Pertumbuhan data yang sangat besar membutuhkan cara pengolahan data yang bisa mengekstrak informasi berarti secara cepat, tepat, dan akurat. Statistik, sebagai ilmu yang mempelajari cara mengumpulkan, mengolah, menganalisis, mengartikan, dan menyajikan data, memainkan peran penting dalam mendukung

**Keywords:**

Statistika; Informatika;
Klasifikasi; Regresi; *Machine Learning*.

perkembangan bidang informatika, terutama di bidang data mining dan machine learning. Artikel ini bertujuan untuk mengeksplorasi penggunaan berbagai metode statistika dan pembelajaran mesin dalam bidang informatika, berdasarkan tinjauan terhadap sebelas penelitian sebelumnya yang diterbitkan antara tahun 2023 hingga 2024. Penelitian ini menggunakan metode kajian pustaka dengan mengelompokkan referensi berdasarkan jenis metode statistika yang digunakan, yaitu metode klasifikasi dan metode regresi. Selanjutnya, penelitian menganalisis pola penerapan, objek penelitian, serta tingkat akurasi yang dihasilkan dalam setiap studi. Metode-metode klasifikasi yang dianalisis meliputi *K-Nearest Neighbors*, *Convolutional Neural Network*, *Naive Bayes Classifier*, dan *Classification and Regression Tree*, sedangkan metode regresi yang dianalisis adalah regresi linear sederhana serta regresi linear berganda. Hasil penelitian menunjukkan bahwa metode klasifikasi biasanya digunakan untuk menyelesaikan masalah yang menghasilkan kategori, seperti analisis sentimen dan penilaian kelayakan usaha, dengan tingkat keakuratan berkisar antara 68 persen hingga 92 persen. Sementara itu, metode regresi digunakan lebih banyak untuk memprediksi angka-angka seperti harga barang, jumlah penduduk, dan jumlah kasus penyakit menular, dengan tingkat keakuratan yang cenderung lebih tinggi. Penelitian ini menunjukkan bahwa gabungan antara konsep statistika klasik dan algoritma pembelajaran mesin memberikan manfaat besar dalam membantu pengambilan keputusan berdasarkan data di berbagai bidang kehidupan, sehingga statistika tidak bisa dipisahkan dari perkembangan ilmu informatika di era digital saat ini.

This is an open access article under the [CC BY-SA](#) license.

**Corresponding Author:**

Fadla Anas Tapiani^{1*}
Universitas Amikom Purwokerto, Indonesia
E-mail: 1fadlaanas580@gmail.com

PENDAHULUAN

Dengan munculnya revolusi industri 4.0, dunia sedang mengalami perubahan digital yang sangat besar di hampir semua bagian kehidupan. Aktivitas yang dulu dilakukan dengan cara tradisional kini semakin banyak dilakukan melalui platform digital, seperti berinteraksi dalam sosial media, melakukan transaksi uang menggunakan dompet digital, dan juga melakukan pembelian atau penjualan barang melalui situs *e-commerce*. Perubahan dalam pola aktivitas tersebut menghasilkan jumlah data yang sangat besar dan terus bertambah setiap hari, baik dalam bentuk data yang sudah terorganisir seperti database transaksi maupun data yang belum terstruktur seperti komentar pengguna dan postingan di media sosial. Fenomena ini disebut big data, yang memiliki ciri-ciri seperti jumlah data yang sangat banyak, jenis data yang beragam, serta pertumbuhan data yang sangat cepat.

Volume data yang besar tidak akan berguna jika tidak diolah dan dianalisis dengan cara yang benar. Di sinilah peran statistika menjadi sangat penting. Statistik adalah ilmu yang belajar tentang cara mengumpulkan, mengolah, menampilkan, menganalisis, dan membuat kesimpulan dari data. Dalam bidang informatika, statistika tidak hanya digunakan untuk menjelaskan data secara dasar, tetapi sudah menjadi dasar penting bagi berbagai algoritma data mining dan machine learning. Algoritma-algoritma tersebut, seperti *Naive Bayes*, *K-Nearest Neighbors*, *regresi linear*, dan *Classification and Regression Tree*, semuanya didasarkan pada



prinsip-prinsip statistika seperti probabilitas, distribusi data, pengukuran jarak, serta penelitian mengenai hubungan antar variabel.

Penelitian sebelumnya menunjukkan cara metode statistika digunakan untuk menyelesaikan masalah nyata di bidang informatika. Misalnya, metode seperti *K-Nearest Neighbors* dan *Convolutional Neural Network* sudah digunakan untuk menganalisis perasaan masyarakat terhadap suatu topik atau layanan digital, berdasarkan data yang dikumpulkan dari media sosial dan ulasan aplikasi. Metode *Naive Bayes Classifier* digunakan untuk menentukan apakah usaha mikro, kecil, dan menengah (UMKM) layak dalam proses sertifikasi halal, serta untuk menentukan tingkat kemiskinan suatu daerah. Di sisi lain, metode regresi linear, baik yang sederhana maupun yang berganda, sering digunakan untuk memperkirakan nilai-nilai berupa angka, seperti harga bahan pangan, harga laptop, harga rumah, jumlah orang yang pindah tempat tinggal, hingga jumlah kasus penyakit yang menular. Di tingkat yang lebih luas, pendekatan analisis big data berbasis machine learning juga digunakan untuk memahami dan memprediksi cara orang membelanjakan uang di platform bisnis online.

Meskipun sudah banyak penelitian yang menggunakan metode statistika dalam berbagai kasus studi di bidang informatika, penelitian yang secara khusus mengelompokkan dan membandingkan penerapan metode tersebut berdasarkan jenisnya, yaitu metode klasifikasi dan metode regresi, masih belum banyak. Meski demikian, pemahaman yang dalam tentang bagaimana masing-masing metode statistika diterapkan, kelebihanannya, kekurangannya, serta tingkat keakuratannya sangat penting, baik bagi para akademisi maupun praktisi, agar bisa memilih metode yang paling sesuai dengan jenis data dan masalah yang dihadapi. Oleh karena itu, artikel ini dibuat untuk mempelajari secara terstruktur penggunaan statistika dalam bidang informatika berdasarkan sebelas penelitian sebelumnya yang terkait.

Berdasarkan penjelasan latar belakang di atas, perumusan masalah dalam penelitian ini dibuat seperti berikut: 1.) Bagaimana cara menggunakan metode statistika klasifikasi, seperti *K-Nearest Neighbors*, *Convolutional Neural Network*, *Naive Bayes Classifier*, dan *Classification and Regression Tree*, untuk menyelesaikan berbagai masalah dalam bidang informatika. 2.) Bagaimana cara menggunakan metode statistika regresi, seperti regresi linear sederhana dan regresi linear berganda, untuk memperkirakan angka-angka tertentu dalam berbagai contoh kasus di bidang informatika. 3.) Apa saja faktor dan kendala yang mempengaruhi tingkat kesesuaian dalam menerapkan metode statistika tersebut pada setiap studi kasus. 4.) Bagaimana penggunaan statistika membantu dalam pengambilan keputusan berdasarkan data di berbagai sektor yang diteliti.

Berdasarkan perumusan masalah tersebut, tujuan penelitian ini dibuat sebagai berikut: 1.) Mengenali dan memahami cara penggunaan metode statistika klasifikasi dalam berbagai contoh kasus di bidang informatika berdasarkan delapan belas penelitian sebelumnya. 2.) Menganalisis penggunaan metode statistik regresi dalam memperkirakan angka-angka pada berbagai contoh kasus di bidang informatika. 3.) Mencari tahu faktor-faktor dan hambatan yang mempengaruhi tingkat keakuratan model statistika yang digunakan dalam setiap penelitian. 4.) Menceritakan bagaimana penggunaan statistika membantu dalam mengambil keputusan yang didasarkan pada data di berbagai bidang, seperti kesehatan masyarakat, ekonomi, data penduduk, dan perdagangan online.

METODE

Tahapan yang dilakukan dalam penelitian ini terdiri dari empat langkah utama. Pertama, tentukan sumber dengan mengumpulkan sebelas artikel jurnal dan prosiding nasional yang diterbitkan pada periode tahun 2023 hingga 2024, yang membahas penggunaan metode statistika dan pembelajaran mesin dalam berbagai kasus studi di bidang informatika. Kedua,



kategorisasi, yaitu mengelompokkan kedua belas referensi tersebut berdasarkan jenis metode statistika yang digunakan, yaitu kelompok metode Penelitian ini dibuat dengan cara melihat dan mengumpulkan informasi dari berbagai sumber yang sudah ada, serta menjelaskan secara rinci dan berdasarkan kualitas informasi tersebut. Pendekatan ini dipilih karena tujuan dari penelitian ini adalah untuk mempelajari dan menggabungkan penggunaan metode statistika dalam bidang informatika berdasarkan pada penelitian yang sudah diterbitkan sebelumnya, bukan untuk menguji model atau mengumpulkan data baru.

Klasifikasi, kelompok metode regresi, dan kelompok pendekatan analisis *big data*. Ketiga, analisis, yaitu mempelajari secara dalam objek penelitian, metode yang diterapkan, langkah-langkah penelitian, serta hasil dan tingkat kebenaran yang disampaikan dalam setiap referensi. Keempat, sintesis, yaitu merangkum pola, kesamaan, perbedaan, serta tantangan dalam penerapan metode statistika pada kesebelas referensi tersebut untuk menjawab rumusan masalah penelitian.

Data yang digunakan dalam penelitian ini adalah data sekunder, berupa informasi mengenai metode penelitian, objek yang diteliti, tahapan yang dilakukan, serta hasil yang dilaporkan dalam setiap artikel rujukan. Analisis dilakukan dengan cara kualitatif, yakni dengan membandingkan dan mengelompokkan hasil temuan dari setiap sumber berdasarkan jenis metode statistika yang digunakan, sehingga bisa ditemukan pola umum mengenai bagaimana statistika diterapkan dalam bidang informatika.

HASIL DAN PEMBAHASAN

A. Statistika untuk Klasifikasi dalam Bidang Informatika

1. *K-Nearest Neighbors* (KNN)

K-Nearest Neighbors (KNN) adalah salah satu metode klasifikasi dalam statistika yang tidak memerlukan asumsi tertentu tentang bentuk distribusi data. Algoritma ini bekerja dengan membandingkan suatu data dengan data-data lainnya berdasarkan tingkat kesamaan atau kedekatannya. Algoritma ini mengelompokkan data baru berdasarkan kelas yang paling banyak ditemukan di antara K data terdekat, yang diukur jaraknya menggunakan metode tertentu, seperti jarak *Euclidean*. Apriliani dan timnya (2023) menggunakan algoritma KNN untuk menganalisis perasaan masyarakat terhadap program vaksinasi Covid-19. Mereka mengumpulkan data dari platform *Twitter* sebanyak 2.241 buah. Data tersebut dibagi menjadi dua kategori, yaitu sentimen positif dan negatif, melalui proses pelabelan data. Selama pra-proses teks, digunakan bahasa pemrograman *Python*. Untuk menentukan pentingnya kata, dilakukan pembobotan menggunakan metode TF-IDF. Selanjutnya, dibuat model, diuji, dan dinilai performanya. Hasil penelitian tersebut menunjukkan bahwa menggunakan parameter *K-Fold cross validation* sebanyak 10 dan nilai K sebesar 5 dalam algoritma KNN dapat memberikan tingkat akurasi klasifikasi sentimen sebesar 79,25 persen. Penerapan ini menunjukkan bahwa konsep statistika seperti ukuran jarak dan pembobotan frekuensi kata sangat berpengaruh dalam membuat model klasifikasi berbasis teks di bidang informatika.

2. *Convolutional Neural Network* (CNN)

Jaringan Saraf Tiruan Berbasis Konvolusi (CNN) adalah salah satu bentuk arsitektur pembelajaran mendalam yang juga menggunakan prinsip-prinsip statistika, seperti operasi konvolusi yang melibatkan perkalian dan penjumlahan bobot secara matematis, serta fungsi aktivasi yang didasarkan pada konsep probabilitas. Achmad dkk (2023) menggunakan algoritma CNN untuk menganalisis klasifikasi sentimen berdasarkan topik dari ulasan pengguna di dua aplikasi dompet digital, yaitu Dana dan Sakuku. Prosesnya meliputi tahapan



pra-pemrosesan data, pemrosesan data, klasifikasi, serta evaluasi model. Penelitian menunjukkan bahwa pengguna *Twitter* di Indonesia lebih sering memberikan ulasan negatif terhadap kedua layanan dompet digital tersebut. Dari hasilnya, sekitar 66 persen pengguna menyampaikan sentimen negatif terhadap layanan Dana, sementara 73 persen mengeluhkan layanan Sakuku. Dalam hal performa model, algoritma CNN yang digunakan dalam penelitian ini berhasil mencapai tingkat akurasi klasifikasi sebesar 81 persen pada data uji. Temuan ini menunjukkan bahwa pendekatan *deep learning*, meskipun membutuhkan komputasi yang lebih rumit dibandingkan metode statistika tradisional, tetap memerlukan dasar statistika yang baik dalam proses menentukan bobot fitur dan mengevaluasi kinerja model.

3. Naive Bayes Classifier dan Kombinasinya dengan Metode Lain

Naive Bayes Classifier adalah metode pengklasifikasian statistik yang menggunakan *Teorema Bayes* dan mengasumsikan bahwa setiap atribut atau fitur bersifat independen satu sama lain. Wijaya (2023) menggunakan metode *Naive Bayes* dalam proses *Knowledge Discovery in Database* (KDD) untuk mengklasifikasikan tingkat kemampuan usaha mikro, kecil, dan menengah (UMKM) dalam mengurus sertifikasi halal. Data yang digunakan berasal dari ikm-halal-2022, yang didapat dari mitra Ruang Halal di seluruh wilayah Pulau Jawa. Proses penelitian mencakup menginput data latih ke dalam perangkat lunak *RapidMiner*, memilih data dan membersihkannya, mengubah bentuk data, mengambil informasi dari data, serta mengevaluasi model yang dibuat. Penelitian menunjukkan bahwa model *Naive Bayes* yang dibuat berhasil mencapai tingkat akurasi klasifikasi sebesar 92,41 persen. Tingkat kesalahan dalam mengklasifikasikan UMKM yang belum memiliki sertifikat halal relatif kecil.

Metode *Naive Bayes* juga sering digunakan bersama dengan algoritma klasifikasi lainnya agar hasil klasifikasi lebih baik. Setiawan dan temannya (2024) menggunakan gabungan metode *Naive Bayes Classifier* dan *Classification and Regression Tree* (CART) untuk mengklasifikasikan tingkat kemiskinan di kabupaten dan kota di Indonesia. *Classification and Regression Tree* adalah metode statistik yang menggunakan pohon keputusan untuk membagi data secara berulang kali dengan kriteria tertentu sampai terbentuk beberapa kelompok data yang memiliki karakteristik serupa. Menggunakan kedua metode tersebut bersamaan memungkinkan membandingkan hasilnya antara pendekatan probabilistik (*Naive Bayes*) dan pendekatan berbasis pohon keputusan (CART) dalam mengklasifikasikan tingkat kemiskinan suatu daerah berdasarkan indikator sosial dan ekonomi, sehingga bisa membantu pemerintah dalam membuat kebijakan pengentasan kemiskinan yang lebih tepat sasaran.

Dalam penelitian lain, Damanhuri dan Husein (2024) membandingkan hasil kerja algoritma *Naive Bayes* dengan algoritma *Logistic Regression* dalam menganalisis sentimen dari ulasan aplikasi *Access by KAI* di *Google Play Store*, menggunakan teknik *Word2Vec* untuk mengekstrak fitur. Hasil penelitian menunjukkan bahwa metode *Logistic Regression* memberikan hasil yang lebih baik dibandingkan metode *Naive Bayes* dalam hal akurasi dan presisi, masing-masing mencapai 68,83 persen dan 75,49 persen. Sementara itu, metode *Naive Bayes* memiliki kelebihan pada nilai recall, yaitu sebesar 45,07 persen. Selain itu, hasil analisis sentimen menunjukkan bahwa mayoritas ulasan tentang aplikasi *Access by KAI* bersifat negatif, dengan 334 dari 400 data uji yang menunjukkan sentimen negatif tersebut. Temuan ini menunjukkan bahwa cara mengklasifikasikan data perlu disesuaikan dengan jenis penilaian yang paling penting, karena suatu cara bisa bagus pada satu jenis penilaian tapi kurang bagus pada jenis penilaian lainnya.

B. Statistika untuk Regresi dalam Prediksi Numerik

1. Regresi Linear Sederhana



Regresi linear sederhana adalah salah satu teknik statistik yang digunakan untuk mengetahui hubungan lurus antara satu variabel yang dipengaruhi dengan satu variabel yang memengaruhinya. Model ini sering digunakan dalam bidang informatika untuk memperkirakan atau menaksir nilai angka berdasarkan data yang sudah terjadi sebelumnya.

Harahap dkk. (2023) menggunakan metode regresi linear sederhana untuk memperkirakan harga beras di Kota Medan. Mereka memilih beras karena merupakan salah satu jenis komoditas pangan utama yang harganya sering berubah-ubah, sehingga memengaruhi kebutuhan pokok masyarakat. Model regresi yang dibuat dalam penelitian tersebut kemudian diaplikasikan ke dalam sebuah sistem berbasis web agar mempermudah proses prediksi. Evaluasi model dengan metode *Root Mean Square Error* (RMSE) menunjukkan kesalahan yang sangat kecil, yaitu sebesar 0,5599. Hal ini menunjukkan bahwa model regresi linear sederhana yang digunakan memiliki kemampuan memprediksi harga beras dengan sangat baik.

Darmadi dkk. (2024) juga menerapkan pendekatan serupa, yaitu dengan menggunakan algoritma regresi linear sederhana dalam kerangka kerja data mining untuk memperkirakan jumlah penduduk yang datang dan pindah di Kabupaten Simalungun. Penelitian itu dilakukan karena jumlah orang yang pindah di wilayah tersebut sangat tinggi, yaitu terdapat 14.998 orang pendatang baru dan 16.795 orang yang pindah pada tahun 2022. Penelitian ini menggunakan metode regresi linear sederhana untuk memahami hubungan antara berbagai variabel yang memengaruhi pergerakan penduduk, sehingga bisa diestimasi jumlah orang yang datang dan pindah pada bulan-bulan yang akan datang. Penelitian ini menunjukkan bahwa regresi linear sederhana bisa digunakan bukan hanya untuk data ekonomi seperti harga barang, tetapi juga untuk data penduduk dalam membantu merencanakan dan mengelola jumlah penduduk di suatu wilayah.

2. Regresi Linear Berganda

Berbeda dengan regresi linear sederhana, regresi linear berganda digunakan ketika satu variabel dependen dipengaruhi oleh beberapa variabel independen sekaligus. Metode ini sering digunakan dalam bidang informatika untuk memperkirakan angka-angka yang dipengaruhi oleh beberapa faktor sekaligus.

Wardani dan Nurmalitasari (2023) menggunakan metode regresi linear berganda untuk memperkirakan harga laptop. Mereka memanfaatkan perangkat lunak Python dan mengambil data dari 65 unit laptop yang didapat dari situs *ELS Computer*. Dalam penelitian itu, harga laptop dianggap sebagai variabel yang bergantung, sedangkan ukuran layar, kapasitas RAM dan penyimpanan, serta berat laptop dianggap sebagai variabel yang memengaruhi. Evaluasi model dilakukan dengan menggunakan beberapa metrik, yaitu *Mean Absolute Error* (MAE), *Mean Squared Error* (MSE), *Root Mean Square Error* (RMSE), dan *Mean Absolute Percentage Error* (MAPE), yang memberikan nilai akurasi prediksi sebesar 0,34. Nilai tersebut menunjukkan bahwa model regresi berganda yang telah dibuat masih bisa dikembangkan lebih lanjut, seperti dengan menambah jumlah data latih atau memperhatikan variabel independen lainnya, seperti merek dan jenis prosesor laptop, agar prediksi lebih akurat.

Penelitian dengan topik yang sama juga dilakukan oleh Khoiriyah dan Fatah tahun 2024, mereka menggunakan algoritma regresi linear dengan bantuan perangkat lunak RapidMiner untuk memperkirakan harga rumah. Penelitian ini menemukan beberapa faktor yang memengaruhi harga rumah secara nyata, yaitu luas tanah dengan koefisien sebesar 242.186, jumlah kamar mandi dengan koefisien 987.650, jumlah lantai dengan koefisien 542.411, serta fasilitas parkir dengan koefisien 408.438. Semua faktor tersebut memiliki tingkat signifikansi (*p-value*) di bawah batas 0,05. Faktor-faktor seperti akses ke jalan utama,



jumlah kamar tamu, dan adanya pendingin udara sebenarnya tidak berdampak besar terhadap harga rumah. Temuan ini menunjukkan betapa pentingnya pengujian signifikansi statistik, khususnya uji *p-value* pada koefisien regresi, dalam menentukan variabel mana yang benar-benar penting dan seharusnya dipertahankan dalam model prediksi.

3. Regresi untuk Prediksi Data Time Series pada Bidang Kesehatan

Regresi bisa digunakan bukan hanya untuk data ekonomi atau properti, tetapi juga untuk data deret waktu di bidang kesehatan masyarakat. Wahyuni dan timnya (2024) menggunakan metode Regresi Linear Berganda untuk memprediksi data waktu deret penyakit menular, terutama kasus Covid-19 di Indonesia, dengan memanfaatkan data yang dikumpulkan oleh Badan Nasional Penanggulangan Bencana (BNPB) mulai dari 2 Maret 2020 hingga 8 November 2020. Model tersebut menggunakan beberapa variabel seperti jumlah kasus baru, jumlah pasien yang sembuh, dan jumlah kematian. Prediksi dilakukan selama empat hari ke depan untuk memperkirakan tren peningkatan kasus dalam waktu singkat. Hasil uji model menunjukkan nilai R^2 yang hampir mencapai 100 persen, dan akurasi model mencapai 95,9 persen. Tingginya tingkat akurasi itu menunjukkan bahwa metode regresi linear berganda sangat efektif digunakan untuk memprediksi data deret waktu jangka pendek dalam kasus penyebaran penyakit menular.

C. Big Data Analytics dan Machine Learning untuk Analisis Perilaku Konsumen

Selain metode klasifikasi dan regresi yang spesifik, penggunaan statistika dalam bidang informatika juga berkembang ke arah pendekatan *big data analytics* yang digabungkan dengan berbagai algoritma *machine learning* secara lebih luas. Mubarok dkk. (2024) melakukan penelitian tentang penggunaan big data analytics dan machine learning untuk memprediksi perilaku konsumen di platform *e-commerce*. Penelitian ini menggunakan data primer yang dikumpulkan lewat survei kepada 1.000 orang responden serta data sekunder yang diperoleh dengan cara mengambil informasi dari internet dan menganalisis datanya. Penelitian ini menunjukkan bahwa harga, kecepatan pengiriman, dan pengalaman pengguna memiliki dampak besar terhadap kepercayaan pelanggan terhadap platform belanja online. Selain itu, penggunaan analisis prediktif berbasis machine learning dalam penelitian itu berhasil meningkatkan tingkat keakuratan memprediksi perilaku konsumen hingga 85 persen. Temuan ini menunjukkan bahwa penggunaan statistika pada data dalam skala besar memberikan manfaat besar bagi para pengusaha dalam menyusun strategi pemasaran dan pelayanan pelanggan yang lebih efektif serta didasarkan pada data.

D. Rangkuman Penerapan Metode Statistika

Tabel 1 berikut ini merupakan ringkasan tentang penggunaan metode statistika pada sebelas penelitian sebelumnya yang dibahas dalam artikel ini, mencakup nama penulis dan tahun terbit, metode statistika yang digunakan, objek atau bidang penerapannya, serta hasil utama yang ditemukan.

Tabel 1. Penggunaan Metode Statistika

Penulis (Tahun)	Metode Statistika	Objek/Bidang Penerapan	Hasil Utama
Apriliani dkk. (2023)	K-Nearest Neighbors	Analisis sentimen vaksinas di Twitter Covid-19	Akurasi 79,25% (K-Fold 10, K=5)
Achmad dkk. (2023)	<i>Convolutional Neural Network</i>	Analisis ulasan dompet	Akurasi 81%; sentimen negatif dominan



			digital Dana & Sakuku		
Wijaya (2023)	Naive Bayes		Klasifikasi kelayakan sertifikasi halal UMKM	Akurasi 92,41%	
Harahap dkk. (2023)	Regresi Sederhana	Linear	Prediksi harga beras di Kota Medan	RMSE 0,5599 (akurasi tinggi)	
Wardani & Nurmalitasari (2023)	Regresi Berganda	Linear	Prediksi harga laptop (65 data)	Nilai akurasi 0,34 (MAE/MSE/RMSE/MAPE)	
Setiawan & rekan (2024)	Naive Bayes & CART		Klasifikasi status kemiskinan kab/kota di Indonesia	Perbandingan dua model klasifikasi	
Darmadi dkk. (2024)	Regresi Sederhana	Linear	Estimasi migrasi penduduk Kab. Simalungun	Estimasi jumlah kedatangan & kepindahan	
Damanhuri & Husein (2024)	Naive Bayes & Logistic Regression		Analisis sentimen ulasan aplikasi Access by KAI	Akurasi LR 68,83%; recall NB 45,07%	
Khoiriyah & Fatah (2024)	Regresi (RapidMiner)	Linear	Prediksi harga rumah	4 variabel signifikan ($p < 0,05$)	
Wahyuni dkk. (2024)	Multiple Regression	Linear	Prediksi <i>time series</i> penyakit menular (Covid-19)	Akurasi 95,9% ($R^2 \approx 100\%$)	
Mubarok dkk. (2024)	Big Data Analytics & Machine Learning		Prediksi perilaku konsumen e-commerce	Akurasi prediksi hingga 85%	

E. Pembahasan Umum

Dari penelitian sebelas yang telah dipelajari, terlihat beberapa pola bagaimana statistika digunakan dalam bidang informatika. Pertama, metode klasifikasi seperti *K-Nearest Neighbors*, *Convolutional Neural Network*, dan *Naive Bayes Classifier* biasanya digunakan untuk menyelesaikan masalah yang hasilnya berupa kategori atau kelas, seperti sentimen positif atau negatif serta status layak atau tidak layak. Akurasi dari metode klasifikasi yang digunakan bervariasi antara 68 persen hingga 92 persen. Dalam penelitian Wijaya (2023), metode *Naive Bayes* memberikan akurasi tertinggi. Sementara itu, metode *Logistic Regression* dan *Naive Bayes* dalam penelitian Damanhuri dan Husein (2024) memiliki akurasi yang lebih rendah, kemungkinan karena bahasa alami itu sendiri lebih rumit dan maknanya bisa berbeda tergantung konteks, terutama pada data ulasan pengguna.

Kedua, metode regresi seperti regresi linear sederhana dan regresi linear berganda biasanya digunakan untuk menyelesaikan masalah yang menghasilkan nilai angka kontinu, seperti harga, jumlah penduduk, dan jumlah kasus penyakit. Akurasi yang dicapai oleh metode regresi dalam penelitian ini biasanya cukup baik, seperti terlihat dari nilai RMSE yang kecil dalam studi Harahap dkk. (2023) serta tingkat akurasi yang hampir mencapai 96 persen dalam studi Wahyuni dkk. (2024). Namun, ada juga penelitian yang menunjukkan akurasi yang tidak terlalu tinggi, seperti pada penelitian Wardani dan Nurmalitasari (2023). Kemungkinan hal ini



terjadi karena jumlah data yang digunakan terbatas, hanya 65 data, sehingga model regresi belum bisa memahami variasi data secara sempurna.

Ketiga, beberapa faktor yang terus-menerus mempengaruhi tingkat akurasi model statistika dalam kedua belas studi tersebut adalah: (1) jumlah dan kualitas data yang digunakan, di mana studi dengan data lebih banyak, seperti studi Apriliani dkk.(2023) yang menggunakan 2.241 data dan studi Mubarak dkk.(2024) dengan 1.000 responden, cenderung menghasilkan model yang lebih stabil; (2) kemampuan dalam memilih variabel independen yang tepat, seperti yang dilakukan oleh studi Khoiriyah dan Fatah (2024) yang melakukan uji signifikansi untuk menyaring variabel yang benar-benar penting; (3) tingkat kompleksitas dan keragaman data, terutama pada data teks yang tidak terstruktur seperti pada studi analisis sentimen, yang lebih sulit dibandingkan data numerik yang terstruktur; serta (4) ketepatan dalam menentukan parameter model, seperti nilai K pada algoritma KNN dan kedalaman pohon pada algoritma CART.

Keempat, secara umum, penggunaan statistika di bidang informatika memberikan pengaruh penting terhadap cara pengambilan keputusan berdasarkan data di berbagai bidang. Mulai dari bidang kesehatan masyarakat yang memantau perasaan masyarakat terhadap program vaksinasi dan memperkirakan penyebaran penyakit menular, hingga bidang ekonomi dan perdagangan yang memperkirakan harga barang dan memahami cara masyarakat membeli. Selain itu, di bidang kependudukan, statistika digunakan untuk memperkirakan pergerakan jumlah penduduk, dan di bidang sosial, digunakan untuk mengelompokkan tingkat kemiskinan di suatu daerah. Oleh karena itu, statistika tidak bisa dipisahkan dari perkembangan bidang informatika, terutama dalam membantu pengembangan sistem cerdas yang berbasis data.

KESIMPULAN DAN SARAN

Berdasarkan penelitian sebelumnya yang membahas penggunaan statistika dalam bidang informatika, terdapat beberapa kesimpulan yang dapat ditarik. Pertama, metode statistika klasifikasi seperti *K-Nearest Neighbors*, *Convolutional Neural Network*, dan *Naive Bayes Classifier* sering digunakan untuk menyelesaikan masalah klasifikasi di bidang informatika, seperti analisis sentimen dan penilaian kelayakan usaha. Tingkat akurasi metode ini berkisar antara 68 persen hingga 92 persen, tergantung pada jenis data dan seberapa tepat parameter model dipilih. Kedua, metode statistika regresi, seperti regresi linear sederhana dan regresi linear berganda, sering digunakan untuk memperkirakan angka-angka dalam berbagai kasus penelitian, seperti harga barang, harga rumah, pergerakan penduduk, dan jumlah orang yang terkena penyakit menular. Akurasi hasilnya biasanya cukup baik selama data yang digunakan memadai dan variabel-variabel yang dipilih tepat. Ketiga, faktor-faktor seperti jumlah dan kualitas data, pemilihan variabel yang tepat, tingkat kompleksitas data, serta penentuan parameter model yang akurat ternyata sangat memengaruhi tingkat keakuratan model statistika yang dibuat. Keempat, penggunaan statistika, baik berupa metode tradisional maupun yang diaplikasikan bersama dengan pendekatan *big data analytics* dan *machine learning*, memberikan manfaat besar dalam membantu pengambilan keputusan berdasarkan data di berbagai bidang kehidupan.

Berdasarkan hasil penelitian tersebut, beberapa rekomendasi bisa diajukan untuk penelitian lebih lanjut. Penelitian selanjutnya disarankan untuk memperhatikan penggunaan data yang lebih banyak dan beragam agar model lebih bisa diterapkan di berbagai situasi, melakukan penyesuaian parameter model dengan cara yang lebih terencana, seperti menggunakan teknik penyetelan *hyperparameter*, serta mencoba gabungan antara metode



statistik tradisional dengan pendekatan deep learning yang lebih canggih agar hasilnya lebih akurat, terutama untuk masalah yang melibatkan data tidak terstruktur seperti teks dan gambar.

DAFTAR PUSTAKA

- [1] Covid, V., Neighbors, M. K., Apriliani, D., Susanto, A., Hidayattullah, M. F., & Sasmito, G. W. (2023). *Sentimen Analisis Pandangan Masyarakat Terhadap*. 8(1), 34–37.
- [2] Damanhuri, R., & Husein, V. A. (2024). *Analisis Sentimen pada Ulasan Aplikasi Access by KAI Berbahasa Indonesia Menggunakan Word-Embedding dan Classical Machine Learning*. 15(September). <https://doi.org/10.14710/jmasif.15.2.62383>
- [3] Darmadi, R. A., Darma, S., & Safii, M. (2023). *Algoritma Regresi Linear Sederhana Dalam Mengestimasi Kedatangan dan Kepindahan Penduduk Di Kabupaten*.
- [4] E-commerce, P. K. D. I. (2025). *BIg Data Analytics Dan Machine Learning Untuk Memprediksi*. 8(1), 159–167.
- [5] Harahap, I. R., & Santoso, H. (2023). *Implementasi Metode Regresi Linear Sederhana Untuk Prediksi Harga Beras Di Kota Medan*. 267–273.
- [6] Informatika, J., & Informasi, S. (2023). *INFORMASI (Jurnal Informatika dan Sistem Informasi) Volume 15 No.2 / Nov / 2023*. 15(2), 225–236.
- [7] Khoiriyah, S., & Fatah, Z. (2024). *Penerapan Algoritma Linear Regression dalam Memprediksi Harga Rumah Menggunakan RapidMiner*. 3(2), 107–115.
- [8] Seran, K. M., Setyawan, Y., Jatipaningrum, M. T., & Astuti, F. (2024). *Klasifikasi Status Kemiskinan Kabupaten / Kota Di Indonesia Menggunakan Naïve Bayes Classifier Dan Classification And Regression Tree membandingkan algoritma klasifikasi data kesejahteraan sosial kabupaten Bantul menggunakan Naive Bayes Classifier dan Decision Tree Pada penelitian ini menggunakan 30 kali pengacakan data dengan fungsi set . seed ()*. *Penelitian*. 09(01), 1–9.
- [9] Wahyuni, S. N., Astuti, Y., Maulina, D., Yogyakarta, U. A., Catur, C., Informatika, J. M., Komputer, F. I., & Yogyakarta, U. A. (2024). *Implementasi Multi Linear Regression Untuk Prediksi Data Runtun Waktu Pada Penyakit Menular Menggunakan Pendekatan Machine Learning*. 11(2).
- [10] Wardani, S. N. (2023). *Penerapan Regresi Linier Berganda Untuk Memprediksi Harga Laptop Dengan Menggunakan Software Python*. 377–382.
- [11] Wijaya, G., & Anam, K. (2023). *Klasifikasi UMKM Menggunakan Algoritma Naive Bayes Berdasarkan Sudah Pernah Mempunyai Atau Mengurus Sertifikat Halal*. 4(1), 36–45.