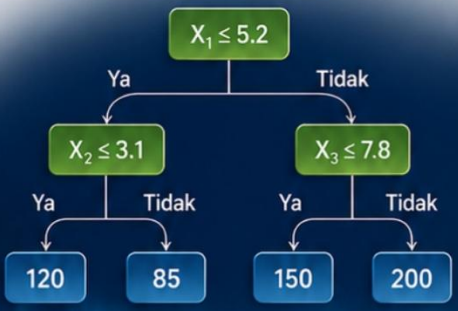
PENERBIT YAYASAN MMI

Decision Tree Regression dan Regresi Linear

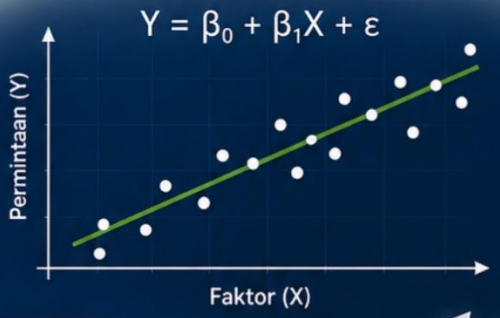
— Teori, Metode, dan Implementasi —

dalam Prediksi Permintaan Usaha

DECISION TREE REGRESSION



REGRESI LINEAR



Pendekatan Teoritis dan Praktis Menggunakan Data Nyata

Mutiara Choirunnisah
Angga Putra Juledi
Iwan Purnama
Marnis Nasution

Decision Tree Regression dan Regresi Linear — Teori, Metode, dan Implementasi dalam Prediksi Permintaan Usaha

Decision Tree Regression dan *Regresi Linear*: Teori, Metode, dan Implementasi dalam Prediksi Permintaan Usaha

PENERBIT YAYASAN MMI

UU No 28 tahun 2014 tentang Hak Cipta

Fungsi dan sifat hak cipta Pasal 4
Hak Cipta sebagaimana dimaksud dalam Pasal 3 huruf a merupakan hak eksklusif yang terdiri atas hak moral dan hak ekonomi.

Pembatasan Pelindungan Pasal 26
Ketentuan sebagaimana dimaksud dalam Pasal 23, Pasal 24, dan Pasal 25 tidak berlaku terhadap:

i. penggunaan kutipan singkat Ciptaan dan/atau produk Hak Terkait untuk pelaporan peristiwa aktual yang ditujukan hanya untuk keperluan penyediaan informasi aktual;
ii. Penggandaan Ciptaan dan/atau produk Hak Terkait hanya untuk kepentingan penelitian ilmu pengetahuan;
iii. Penggandaan Ciptaan dan/atau produk Hak Terkait hanya untuk keperluan pengajaran, kecuali pertunjukan dan Fonogram yang telah dilakukan Pengumuman sebagai bahan ajar; dan
iv. penggunaan untuk kepentingan pendidikan dan pengembangan ilmu pengetahuan yang memungkinkan suatu Ciptaan dan/atau produk Hak Terkait dapat digunakan tanpa izin Pelaku Pertunjukan, Produser Fonogram, atau Lembaga Penyiaran.

Sanksi Pelanggaran Pasal 113

1. Setiap Orang yang dengan tanpa hak melakukan pelanggaran hak ekonomi sebagaimana dimaksud dalam Pasal 9 ayat (1) huruf i untuk Penggunaan Secara Komersial dipidana dengan pidana penjara paling lama 1 (satu) tahun dan/atau pidana denda paling banyak Rp100.000.000 (seratus juta rupiah).
2. Setiap Orang yang dengan tanpa hak dan/atau tanpa izin Pencipta atau pemegang Hak Cipta melakukan pelanggaran hak ekonomi Pencipta sebagaimana dimaksud dalam Pasal 9 ayat (1) huruf c, huruf d, huruf f, dan/atau huruf h untuk Penggunaan Secara Komersial dipidana dengan pidana penjara paling lama 3 (tiga) tahun dan/atau pidana denda paling banyak Rp500.000.000,00 (lima ratus juta rupiah).

Decision Tree Regression dan *Regresi Linear*: Teori, Metode, dan Implementasi dalam Prediksi Permintaan Usaha

Mutiara Choirunnisah;
Angga Putra Juledi;
Iwan Purnama;
Marnis Nasution;

PENERBIT YAYASAN MMI

Decision Tree Regression dan *Regresi Linear*: Teori, Metode, dan Implementasi dalam Prediksi Permintaan Usaha

Mutiara Choirunnisah; Angga Putra Juledi; Iwan Purnama;

Marnis Nasution;

Desain Cover :
Angga Putra Juledi

Sumber :
https://munandar.yayasanmmi.com/decision-tree-regression-dan-regresi-linear

Tata Letak :
Iwan Purnama

Proofreader :
Marnis Nasution

Ukuran :
Jml hal judul: 1, Jml hal isi naskah: 76, Uk: 15x23 cm

ISBN :
978-634-05-3363-7

Cetakan Pertama :
Juli 2026

PENERBIT YAYASAN MUNANDAR MEMBANGUN INDONESIA
Jl. Pasar Banjar Dusun XVI Desa Simpang Empat, Asahan, Sumatera Utara 21271
Telp/Wa : 082363080181
https://munandar.yayasanmmi.com/
E-mail: mmipublisher@yayasanmmi.com

PRAKATA

Puji syukur ke hadirat Tuhan Yang Maha Esa atas segala rahmat, karunia, dan anugerah-Nya sehingga buku *Decision Tree Regression* dan *Regresi Linear*: Teori, Metode, dan Implementasi dalam Prediksi Permintaan Usaha ini dapat diselesaikan dengan baik. Buku ini disusun sebagai salah satu referensi akademik yang bertujuan memberikan pemahaman yang komprehensif mengenai konsep, metode, serta implementasi algoritma *Decision Tree Regression* dan *Regresi Linear* dalam membangun model prediksi permintaan usaha berbasis data.

Melalui buku ini, pembaca diajak memahami dasar-dasar regresi, proses pengolahan data, pembentukan model prediksi, evaluasi kinerja model, hingga penerapannya pada berbagai studi kasus yang relevan dengan dunia usaha.

Penulis menyadari bahwa buku ini masih memiliki keterbatasan. Oleh karena itu, kritik dan saran yang bersifat membangun sangat diharapkan sebagai bahan evaluasi untuk penyempurnaan pada edisi berikutnya. Semoga buku ini dapat memberikan manfaat, memperkaya khazanah ilmu pengetahuan, serta menjadi referensi yang mendukung pembelajaran, penelitian, dan pengembangan solusi prediktif dalam meningkatkan kualitas pengambilan keputusan di berbagai bidang usaha.

Hormat Kami,

Penulis

DAFTAR ISI

BAB 1

KONSEP DASAR ANALISIS DATA

1.1 Data dan Informasi

Data merupakan sekumpulan fakta, angka, simbol, atau hasil pengamatan yang belum diolah sehingga belum memiliki makna yang utuh. Data dapat berasal dari berbagai sumber, seperti transaksi bisnis, survei, sensor, media sosial, maupun aktivitas operasional suatu organisasi. Dalam konteks analisis data, data menjadi komponen utama yang digunakan untuk menghasilkan pengetahuan yang dapat mendukung proses pengambilan keputusan. Semakin baik kualitas data yang dimiliki, semakin besar peluang untuk memperoleh hasil analisis yang akurat dan dapat dipercaya.

Informasi merupakan hasil pengolahan data melalui proses pengelompokan, perhitungan, analisis, maupun interpretasi sehingga memiliki makna dan dapat digunakan sebagai dasar dalam pengambilan keputusan. Berbeda dengan data yang masih bersifat mentah (*raw data*), informasi telah melalui proses transformasi sehingga mampu menjawab kebutuhan pengguna. Sebagai contoh, data penjualan harian suatu toko belum memberikan gambaran mengenai kondisi bisnis. Namun, setelah diolah menjadi informasi berupa tren penjualan bulanan atau prediksi permintaan produk, data tersebut menjadi lebih bernilai bagi manajemen.

Dalam era transformasi digital, data sering disebut sebagai aset strategis organisasi karena mampu menghasilkan informasi yang mendukung peningkatan efisiensi, produktivitas, serta daya saing perusahaan. Oleh

sebab itu, pemahaman mengenai perbedaan antara data dan informasi menjadi dasar penting sebelum mempelajari teknik analisis data, data mining, maupun machine learning. Pengelolaan data yang baik akan menghasilkan informasi berkualitas sehingga organisasi dapat menyusun strategi bisnis secara lebih efektif dan berbasis bukti (*evidence-based decision making*).

1.2 Jenis Data

Dalam analisis data, pemahaman mengenai jenis data sangat penting karena menentukan metode analisis, teknik visualisasi, maupun algoritma yang akan digunakan. Secara umum, data dapat diklasifikasikan berdasarkan sifat, sumber, maupun bentuk penyajiannya. Berdasarkan sifatnya, data dibedakan menjadi data kualitatif dan data kuantitatif. Data kualitatif merupakan data yang berbentuk kategori atau karakteristik tertentu, seperti jenis kelamin, kategori produk, dan tingkat kepuasan pelanggan. Sebaliknya, data kuantitatif berupa nilai numerik yang dapat diukur atau dihitung, seperti jumlah penjualan, pendapatan, atau volume produksi.

Data kuantitatif masih dibedakan menjadi data diskrit dan data kontinu. Data diskrit diperoleh melalui proses perhitungan sehingga menghasilkan nilai bilangan bulat, misalnya jumlah pelanggan atau jumlah transaksi. Sementara itu, data kontinu diperoleh melalui proses pengukuran sehingga dapat memiliki nilai pecahan, seperti berat badan, suhu, atau waktu produksi.

Berdasarkan sumbernya, data dibedakan menjadi data primer dan data sekunder. Data primer diperoleh secara langsung melalui observasi, wawancara, eksperimen, atau kuesioner, sedangkan data sekunder berasal dari dokumen, laporan perusahaan, publikasi pemerintah,

maupun basis data daring. Selain itu, berdasarkan struktur penyimpanannya, data dibagi menjadi data terstruktur, semi-terstruktur, dan tidak terstruktur. Data terstruktur umumnya tersimpan dalam basis data relasional, sedangkan data tidak terstruktur meliputi gambar, video, audio, maupun teks bebas.

1.3 Sumber Data

Sumber data merupakan asal diperolehnya data yang digunakan dalam proses analisis. Pemilihan sumber data yang tepat sangat menentukan kualitas hasil penelitian maupun model prediksi yang akan dibangun. Dalam analisis data modern, sumber data tidak hanya berasal dari aktivitas internal organisasi, tetapi juga dari berbagai sumber eksternal yang tersedia secara luas melalui teknologi digital.

Sumber data internal meliputi seluruh data yang dihasilkan oleh aktivitas operasional organisasi, seperti data transaksi penjualan, data persediaan, data pelanggan, laporan keuangan, maupun sistem Enterprise Resource Planning (ERP). Data internal umumnya memiliki tingkat akurasi yang tinggi karena berasal dari proses bisnis yang berlangsung secara langsung. Sementara itu, sumber data eksternal dapat diperoleh dari Badan Pusat Statistik (BPS), kementerian, lembaga pemerintah, media sosial, situs e-commerce, laporan industri, maupun penyedia data terbuka (*open data*). Data eksternal sering dimanfaatkan untuk melengkapi analisis sehingga mampu menggambarkan kondisi lingkungan bisnis secara lebih komprehensif. Dalam pengembangan model prediksi permintaan usaha, kombinasi antara data internal dan data eksternal akan menghasilkan model yang lebih baik. Sebagai contoh, data historis penjualan dapat dipadukan dengan data musim, hari libur nasional, kondisi ekonomi, maupun tren

perilaku konsumen. Integrasi berbagai sumber data memungkinkan algoritma machine learning mengenali pola yang lebih kompleks sehingga meningkatkan tingkat akurasi prediksi.

Oleh karena itu, peneliti perlu memastikan bahwa sumber data yang digunakan memiliki kualitas yang baik, relevan dengan tujuan penelitian, mutakhir, serta memenuhi aspek validitas dan reliabilitas. Pengelolaan sumber data yang tepat akan menjadi fondasi utama dalam menghasilkan informasi yang akurat serta mendukung pengambilan keputusan berbasis data.

1.4 Variabel Dependen dan Independen

Variabel merupakan karakteristik, atribut, atau nilai yang menjadi objek pengamatan dalam suatu penelitian atau analisis data. Dalam pemodelan statistik maupun machine learning, variabel dibedakan menjadi dua kelompok utama, yaitu variabel dependen (*dependent variable*) dan variabel independen (*independent variable*). Pemahaman terhadap kedua jenis variabel ini sangat penting karena menjadi dasar dalam membangun model prediksi dan menjelaskan hubungan antarvariabel.

Variabel dependen adalah variabel yang menjadi target atau hasil yang ingin diprediksi maupun dijelaskan. Nilai variabel dependen dipengaruhi oleh satu atau lebih variabel independen. Dalam konteks prediksi permintaan usaha, jumlah permintaan produk merupakan variabel dependen karena nilainya diperkirakan berdasarkan berbagai faktor yang memengaruhinya. Sebaliknya, variabel independen merupakan variabel yang berperan sebagai faktor penyebab atau prediktor. Contohnya meliputi harga produk, jumlah promosi, musim, tingkat pendapatan masyarakat, jumlah pelanggan, maupun kondisi

ekonomi.

Pada metode Regresi Linear, hubungan antara variabel independen dan dependen diasumsikan bersifat linier sehingga perubahan pada variabel independen akan memengaruhi nilai variabel dependen secara proporsional. Sementara itu, pada Decision Tree Regression hubungan tersebut tidak harus linier karena algoritma mampu membentuk aturan keputusan berdasarkan pola data yang kompleks. Oleh sebab itu, Decision Tree Regression lebih fleksibel dalam menangani hubungan nonlinier dibandingkan Regresi Linear.

Pemilihan variabel yang relevan menjadi salah satu faktor utama dalam meningkatkan performa model prediksi. Variabel yang tidak relevan dapat menurunkan akurasi model, sedangkan variabel yang memiliki hubungan kuat dengan target mampu meningkatkan kemampuan model dalam menghasilkan prediksi yang lebih akurat. Oleh karena itu, proses seleksi variabel (*feature selection*) sering dilakukan sebelum pembangunan model untuk memperoleh hasil analisis yang optimal.

1.5 Statistik Deskriptif

Statistik deskriptif merupakan cabang statistika yang digunakan untuk menggambarkan karakteristik suatu kumpulan data tanpa melakukan generalisasi terhadap populasi. Statistik deskriptif bertujuan menyajikan data dalam bentuk yang lebih ringkas sehingga memudahkan peneliti dalam memahami pola, distribusi, dan karakteristik data sebelum dilakukan analisis lebih lanjut. Tahap ini menjadi langkah awal yang penting dalam proses analisis data maupun pembangunan model machine learning.

Ukuran statistik deskriptif terdiri atas ukuran pemusatan, ukuran penyebaran, dan ukuran posisi. Ukuran pemusatan meliputi nilai rata-

rata (*mean*), median, dan modus yang digunakan untuk mengetahui nilai representatif suatu data. Sementara itu, ukuran penyebaran meliputi varians, simpangan baku (*standard deviation*), rentang (*range*), serta interquartile range (IQR) yang menggambarkan tingkat variasi data. Selain itu, ukuran posisi seperti kuartil, desil, dan persentil digunakan untuk mengetahui posisi suatu nilai dalam distribusi data.

Dalam analisis prediksi permintaan usaha, statistik deskriptif membantu peneliti mengidentifikasi pola penjualan, fluktuasi permintaan, maupun keberadaan data ekstrem (*outlier*). Informasi tersebut sangat berguna dalam menentukan strategi pembersihan data (*data cleaning*) serta memilih metode analisis yang sesuai. Misalnya, distribusi data yang sangat tidak seimbang dapat memengaruhi performa model prediksi sehingga memerlukan proses transformasi data terlebih dahulu.

Selain digunakan dalam penelitian, statistik deskriptif juga dimanfaatkan oleh organisasi untuk menyusun laporan kinerja, mengevaluasi tren penjualan, serta mendukung pengambilan keputusan berbasis data. Oleh karena itu, penguasaan konsep statistik deskriptif menjadi fondasi penting sebelum mempelajari teknik analisis yang lebih kompleks seperti regresi, data mining, maupun machine learning.

1.6 Visualisasi Data

Visualisasi data merupakan proses menyajikan data dalam bentuk grafis sehingga informasi yang terkandung di dalamnya dapat dipahami dengan lebih mudah. Penyajian data melalui grafik, diagram, maupun peta memungkinkan pengguna mengenali pola, tren, hubungan antarvariabel, serta anomali yang sulit ditemukan apabila

data hanya disajikan dalam bentuk tabel. Oleh karena itu, visualisasi data menjadi bagian penting dalam analisis data modern dan pengambilan keputusan berbasis bukti.

Berbagai jenis visualisasi dapat digunakan sesuai dengan karakteristik data dan tujuan analisis. Diagram batang (*bar chart*) digunakan untuk membandingkan nilai antar kategori, sedangkan diagram garis (*line chart*) digunakan untuk menunjukkan perubahan nilai dari waktu ke waktu. Diagram lingkaran (*pie chart*) menggambarkan proporsi suatu kategori terhadap keseluruhan data, sedangkan histogram digunakan untuk melihat distribusi frekuensi data. Selain itu, *scatter plot* sering digunakan untuk mengidentifikasi hubungan antara dua variabel numerik, sedangkan *box plot* membantu mendeteksi data ekstrem (*outlier*) dan variasi distribusi.

Dalam prediksi permintaan usaha, visualisasi data membantu peneliti memahami pola historis penjualan, tren musiman, pengaruh promosi, maupun perubahan perilaku konsumen. Informasi tersebut menjadi dasar dalam pemilihan variabel prediktor dan pembangunan model prediksi yang lebih akurat. Visualisasi juga memudahkan komunikasi hasil analisis kepada pihak manajemen karena informasi dapat dipahami secara cepat tanpa harus membaca data numerik secara rinci. Perkembangan perangkat lunak seperti Microsoft Power BI, Tableau, Python Matplotlib, Seaborn, dan Plotly semakin mempermudah proses visualisasi data secara interaktif. Dengan visualisasi yang baik, organisasi mampu memperoleh wawasan yang lebih mendalam sehingga proses pengambilan keputusan menjadi lebih cepat, akurat, dan efektif.

1.7 Data Mining

Data mining merupakan proses menemukan pola, hubungan, tren, maupun pengetahuan baru dari sekumpulan data berukuran besar menggunakan teknik statistik, kecerdasan buatan (*Artificial Intelligence*), dan pembelajaran mesin (*Machine Learning*). Data mining menjadi salah satu disiplin ilmu yang berkembang pesat seiring meningkatnya volume data yang dihasilkan oleh organisasi. Tujuan utama data mining adalah mengubah data mentah menjadi informasi yang bernilai sehingga dapat digunakan sebagai dasar pengambilan keputusan yang lebih efektif dan efisien.

Secara umum, data mining merupakan bagian dari proses *Knowledge Discovery in Databases* (KDD), yaitu serangkaian tahapan yang meliputi pemilihan data, pembersihan data, transformasi data, proses penambangan data, evaluasi pola, dan penyajian pengetahuan. Pada tahap penambangan, berbagai algoritma diterapkan untuk menemukan pola tertentu sesuai dengan tujuan analisis. Teknik yang umum digunakan meliputi klasifikasi (*classification*), regresi (*regression*), klasterisasi (*clustering*), asosiasi (*association rule mining*), dan deteksi anomali (*anomaly detection*).

Dalam dunia usaha, data mining dimanfaatkan untuk memprediksi permintaan produk, menganalisis perilaku pelanggan, mendeteksi kecurangan transaksi, mengoptimalkan persediaan, serta menyusun strategi pemasaran. Sebagai contoh, perusahaan ritel dapat menganalisis data penjualan historis untuk mengidentifikasi pola pembelian pelanggan dan memperkirakan kebutuhan stok pada periode tertentu. Dengan demikian, risiko kekurangan maupun kelebihan persediaan dapat diminimalkan.

Keberhasilan penerapan data mining sangat bergantung pada kualitas

data yang digunakan. Data yang lengkap, konsisten, dan relevan akan menghasilkan pola yang lebih akurat dibandingkan data yang mengandung banyak kesalahan atau kehilangan nilai (*missing values*). Oleh karena itu, proses prapengolahan data (*data preprocessing*) menjadi tahap yang sangat penting sebelum algoritma diterapkan. Dengan dukungan data yang berkualitas dan algoritma yang tepat, data mining mampu menghasilkan informasi strategis yang memberikan nilai tambah bagi organisasi dalam menghadapi persaingan bisnis yang semakin dinamis.

1.8 *Machine Learning*

Machine Learning merupakan cabang dari kecerdasan buatan (*Artificial Intelligence*) yang memungkinkan komputer mempelajari pola dari data dan menghasilkan prediksi atau keputusan tanpa diprogram secara eksplisit untuk setiap kondisi. Berbeda dengan pemrograman konvensional yang bergantung pada aturan yang dibuat oleh manusia, machine learning membangun model berdasarkan pengalaman yang diperoleh dari data sehingga mampu meningkatkan performa seiring bertambahnya jumlah data yang dipelajari.

Secara umum, machine learning dibagi menjadi tiga pendekatan utama, yaitu *supervised learning*, *unsupervised learning*, dan *reinforcement learning*. *Supervised learning* menggunakan data yang telah memiliki label sehingga model belajar memetakan hubungan antara variabel masukan dan keluaran. Metode Regresi Linear dan Decision Tree Regression termasuk dalam kategori ini karena memanfaatkan data historis untuk memprediksi nilai numerik di masa mendatang. Sebaliknya, *unsupervised learning* digunakan untuk menemukan pola tersembunyi pada data tanpa label, sedangkan

reinforcement learning memungkinkan sistem belajar melalui mekanisme penghargaan (*reward*) dan hukuman (*penalty*).

Dalam dunia usaha, machine learning telah diterapkan pada berbagai bidang, seperti prediksi permintaan produk, rekomendasi produk, analisis risiko kredit, deteksi penipuan, pemeliharaan prediktif (*predictive maintenance*), hingga analisis perilaku pelanggan. Kemampuan machine learning dalam mengolah data berukuran besar menjadikannya sebagai teknologi penting dalam era transformasi digital dan Industri 4.0.

Meskipun memiliki kemampuan prediksi yang tinggi, keberhasilan machine learning tetap dipengaruhi oleh kualitas data, pemilihan algoritma, proses pelatihan model, serta evaluasi performa menggunakan metrik yang sesuai. Oleh karena itu, pemahaman terhadap konsep dasar machine learning menjadi bekal penting bagi peneliti dan praktisi dalam mengembangkan solusi analitik yang efektif untuk mendukung pengambilan keputusan berbasis data.

1.9 Tahapan Analisis Data

Analisis data merupakan proses sistematis untuk mengolah, mengevaluasi, dan menginterpretasikan data sehingga menghasilkan informasi yang dapat digunakan sebagai dasar pengambilan keputusan. Dalam penelitian maupun pengembangan model prediksi, analisis data dilakukan melalui serangkaian tahapan yang saling berkaitan. Pelaksanaan setiap tahapan secara tepat akan meningkatkan kualitas hasil analisis dan akurasi model yang dihasilkan.

Tahap pertama adalah identifikasi permasalahan, yaitu menentukan tujuan analisis dan pertanyaan penelitian yang akan dijawab. Selanjutnya dilakukan pengumpulan data dari berbagai sumber yang

relevan, baik data primer maupun data sekunder. Setelah data terkumpul, dilakukan prapengolahan data (*data preprocessing*), yang meliputi pembersihan data (*data cleaning*), penanganan nilai hilang (*missing values*), penghapusan data duplikat, transformasi data, normalisasi, serta pemilihan variabel (*feature selection*).

Tahap berikutnya adalah eksplorasi data (*Exploratory Data Analysis/EDA*), yaitu memahami karakteristik data melalui statistik deskriptif dan visualisasi. Hasil eksplorasi menjadi dasar dalam menentukan metode analisis yang paling sesuai. Selanjutnya dilakukan pembangunan model, misalnya menggunakan Regresi Linear atau Decision Tree Regression. Model kemudian dievaluasi menggunakan metrik seperti *Mean Absolute Error* (MAE), *Mean Squared Error* (MSE), *Root Mean Squared Error* (RMSE), *Mean Absolute Percentage Error* (MAPE), dan koefisien determinasi (R^2) untuk mengukur tingkat akurasi prediksi.

Tahap terakhir adalah interpretasi dan implementasi hasil, yaitu menjelaskan makna hasil analisis serta memanfaatkannya dalam proses pengambilan keputusan. Dalam konteks prediksi permintaan usaha, hasil analisis dapat digunakan untuk menyusun strategi produksi, mengoptimalkan persediaan, meningkatkan efisiensi distribusi, dan merencanakan aktivitas pemasaran.

BAB 2
REGRESI LINEAR

2.1 Konsep Regresi

Regresi merupakan salah satu metode statistik yang digunakan untuk menganalisis hubungan antara satu atau lebih variabel independen (*independent variables*) dengan variabel dependen (*dependent variable*). Tujuan utama analisis regresi adalah membangun suatu model matematis yang mampu menjelaskan hubungan antarvariabel sekaligus memprediksi nilai variabel dependen berdasarkan perubahan variabel independen. Dalam bidang ilmu data, ekonomi, bisnis, teknik, kesehatan, hingga ilmu sosial, regresi menjadi metode yang paling banyak digunakan karena mampu memberikan informasi kuantitatif mengenai pengaruh suatu variabel terhadap variabel lainnya.

Konsep dasar regresi berangkat dari asumsi bahwa perubahan pada suatu variabel dapat dipengaruhi oleh faktor-faktor tertentu yang dapat diukur. Sebagai contoh, dalam prediksi permintaan usaha, jumlah permintaan produk dipengaruhi oleh harga, promosi, musim, pendapatan masyarakat, maupun jumlah pelanggan. Hubungan tersebut kemudian dimodelkan dalam bentuk persamaan matematis sehingga memungkinkan peneliti melakukan estimasi terhadap nilai permintaan di masa mendatang. Model regresi tidak hanya berfungsi sebagai alat prediksi, tetapi juga sebagai alat analisis untuk mengetahui besarnya pengaruh masing-masing variabel independen terhadap variabel dependen.

Dalam praktiknya, analisis regresi terdiri atas beberapa jenis, seperti

regresi linear sederhana, regresi linear berganda, regresi logistik, regresi polinomial, dan regresi nonlinier. Buku ini berfokus pada regresi linear karena metode ini menjadi dasar bagi berbagai algoritma prediktif yang lebih kompleks. Regresi linear memiliki keunggulan berupa kemudahan interpretasi, proses komputasi yang relatif sederhana, serta kemampuan menjelaskan hubungan sebab-akibat secara kuantitatif.

Perkembangan teknologi informasi telah menjadikan regresi linear sebagai salah satu metode penting dalam bidang *data science* dan *machine learning*. Berbagai perangkat lunak seperti Python, R, SPSS, SAS, dan MATLAB menyediakan fasilitas untuk membangun model regresi secara cepat dan akurat. Oleh karena itu, pemahaman mengenai konsep regresi menjadi fondasi utama sebelum mempelajari metode prediksi yang lebih kompleks seperti Decision Tree Regression, Random Forest, maupun Neural Network.

2.2 Sejarah Regresi Linear

Regresi linear merupakan salah satu metode statistik tertua yang hingga saat ini masih banyak digunakan dalam penelitian maupun analisis data. Istilah *regression* pertama kali diperkenalkan oleh ilmuwan Inggris, Sir Francis Galton, pada tahun 1886 melalui penelitiannya mengenai hubungan tinggi badan antara orang tua dan anak. Galton menemukan bahwa tinggi badan anak cenderung mendekati rata-rata populasi meskipun orang tuanya memiliki tinggi badan yang sangat tinggi atau sangat rendah. Fenomena tersebut dikenal sebagai regression toward the mean atau regresi menuju rata-rata, yang kemudian menjadi dasar pengembangan analisis regresi modern.

Perkembangan teori regresi semakin pesat setelah Karl Pearson memperkenalkan konsep korelasi untuk mengukur kekuatan hubungan

antarvariabel. Selanjutnya, metode estimasi parameter melalui Ordinary Least Squares (OLS) dikembangkan berdasarkan konsep yang diperkenalkan oleh Adrien-Marie Legendre pada tahun 1805 dan disempurnakan oleh Carl Friedrich Gauss. Metode OLS menjadi teknik utama dalam menentukan garis regresi yang meminimalkan jumlah kuadrat kesalahan prediksi.

Memasuki abad ke-20, regresi linear berkembang menjadi salah satu teknik analisis utama dalam berbagai disiplin ilmu, seperti ekonomi, psikologi, pertanian, teknik, kedokteran, dan bisnis. Kemajuan teknologi komputer memungkinkan perhitungan regresi dilakukan secara otomatis menggunakan perangkat lunak statistik sehingga penerapannya semakin luas.

Saat ini, regresi linear tidak hanya digunakan sebagai metode statistik klasik, tetapi juga menjadi bagian penting dalam bidang *machine learning*. Banyak algoritma modern dikembangkan berdasarkan prinsip dasar regresi linear, baik untuk prediksi numerik maupun analisis hubungan antarvariabel. Meskipun telah muncul berbagai metode berbasis kecerdasan buatan, regresi linear tetap menjadi pilihan utama karena sederhana, mudah diinterpretasikan, dan memiliki tingkat akurasi yang baik ketika hubungan antarvariabel bersifat linier.

2.3 Regresi Linear Sederhana

Regresi linear sederhana (*Simple Linear Regression*) merupakan metode statistik yang digunakan untuk menganalisis hubungan antara satu variabel independen dengan satu variabel dependen. Tujuan utama metode ini adalah membentuk persamaan linear yang dapat digunakan untuk menjelaskan hubungan kedua variabel sekaligus memprediksi nilai variabel dependen berdasarkan perubahan variabel independen. Regresi

linear sederhana banyak digunakan karena konsepnya mudah dipahami dan implementasinya relatif sederhana.

Model regresi linear sederhana dinyatakan dalam bentuk persamaan:

$Y = a + bX$

dengan:

Y = variabel dependen;

X = variabel independen;

a = konstanta (*intercept*), yaitu nilai Y ketika X bernilai nol; dan

b = koefisien regresi (*slope*), yang menunjukkan besarnya perubahan Y akibat perubahan satu satuan pada X.

Dalam konteks prediksi permintaan usaha, variabel X dapat berupa harga produk, sedangkan variabel Y adalah jumlah permintaan. Apabila koefisien regresi bernilai negatif, maka kenaikan harga akan menurunkan permintaan. Sebaliknya, apabila koefisien bernilai positif, kenaikan nilai variabel independen akan meningkatkan permintaan.

Pembangunan model regresi linear sederhana dilakukan menggunakan metode Ordinary Least Squares (OLS), yaitu teknik yang mencari garis terbaik dengan meminimalkan jumlah kuadrat selisih antara nilai aktual dan nilai prediksi. Setelah model terbentuk, langkah berikutnya adalah mengevaluasi kualitas model menggunakan koefisien determinasi (R^2), nilai signifikansi (*p-value*), dan analisis residual.

Meskipun hanya melibatkan satu variabel independen, regresi linear sederhana tetap memberikan informasi yang sangat bermanfaat, terutama pada tahap eksplorasi data. Model ini sering digunakan sebagai dasar untuk memahami hubungan antarvariabel sebelum dikembangkan menjadi regresi linear berganda yang melibatkan lebih dari satu variabel prediktor.

Dalam era *data science*, regresi linear sederhana masih menjadi salah satu algoritma dasar yang dipelajari karena memberikan pemahaman mengenai hubungan sebab-akibat, estimasi parameter, dan interpretasi model secara kuantitatif. Oleh karena itu, penguasaan metode ini menjadi langkah awal yang penting sebelum mempelajari teknik prediksi yang lebih kompleks.

2.4 Regresi Linear Berganda

Regresi linear berganda (*Multiple Linear Regression*) merupakan pengembangan dari regresi linear sederhana yang digunakan untuk menganalisis hubungan antara satu variabel dependen dengan dua atau lebih variabel independen. Metode ini banyak diterapkan dalam penelitian ilmiah, analisis bisnis, ekonomi, dan data science karena mampu menjelaskan pengaruh beberapa faktor secara simultan terhadap suatu variabel target. Dalam konteks prediksi permintaan usaha, penggunaan lebih dari satu variabel prediktor umumnya menghasilkan model yang lebih representatif dibandingkan hanya menggunakan satu variabel.

Model regresi linear berganda dinyatakan dalam bentuk persamaan:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n + \varepsilon$$

dengan Y sebagai variabel dependen, β_0 sebagai konstanta, β_1 hingga β_n sebagai koefisien regresi masing-masing variabel independen, X_1 hingga X_n sebagai variabel prediktor, dan ε sebagai galat (*error term*). Setiap koefisien menunjukkan besarnya perubahan nilai variabel dependen apabila salah satu variabel independen berubah satu satuan, dengan asumsi variabel lainnya tetap.

Dalam prediksi permintaan usaha, variabel independen dapat berupa harga produk, jumlah promosi, musim, tingkat pendapatan masyarakat,

diskon, jumlah pelanggan, maupun kondisi ekonomi. Dengan menggabungkan beberapa variabel tersebut, model mampu menangkap hubungan yang lebih kompleks sehingga menghasilkan prediksi yang lebih akurat dibandingkan regresi linear sederhana.

Tahapan pembangunan model meliputi pengumpulan data, eksplorasi data, pemilihan variabel, estimasi parameter menggunakan metode *Ordinary Least Squares* (OLS), pengujian signifikansi model melalui uji F dan uji t, serta evaluasi menggunakan koefisien determinasi (R^2). Selain itu, perlu dilakukan pemeriksaan terhadap multikolinearitas, heteroskedastisitas, dan normalitas residual agar model memenuhi asumsi statistik.

Keunggulan regresi linear berganda terletak pada kemampuannya mengukur kontribusi masing-masing variabel terhadap variabel dependen. Informasi tersebut sangat bermanfaat dalam penyusunan strategi bisnis karena perusahaan dapat mengetahui faktor mana yang paling berpengaruh terhadap permintaan produk. Oleh sebab itu, regresi linear berganda menjadi salah satu metode yang paling banyak digunakan dalam analisis prediktif berbasis data.

2.5 Asumsi Regresi Linear

Keakuratan hasil analisis regresi linear sangat dipengaruhi oleh terpenuhinya sejumlah asumsi statistik. Asumsi-asumsi tersebut diperlukan agar estimasi parameter yang dihasilkan bersifat tidak bias (*unbiased*), efisien, dan dapat digunakan untuk melakukan inferensi maupun prediksi secara tepat. Apabila salah satu asumsi dilanggar, maka hasil analisis dapat menjadi kurang akurat dan interpretasi model menjadi tidak valid.

Asumsi pertama adalah linearitas, yaitu hubungan antara variabel

independen dan dependen harus bersifat linear. Hubungan ini dapat diperiksa melalui grafik *scatter plot* atau analisis residual. Asumsi kedua adalah independensi residual, yang menyatakan bahwa kesalahan prediksi antarobservasi tidak saling berkorelasi. Pada data runtun waktu (*time series*), independensi residual umumnya diuji menggunakan uji Durbin-Watson.

Asumsi ketiga adalah homoskedastisitas, yaitu varians residual harus konstan pada seluruh rentang nilai variabel independen. Apabila varians residual berubah-ubah, kondisi tersebut disebut heteroskedastisitas dan dapat mengurangi ketepatan estimasi parameter. Asumsi keempat adalah normalitas residual, yang menyatakan bahwa residual mengikuti distribusi normal. Pengujian normalitas dapat dilakukan menggunakan uji Shapiro-Wilk, Kolmogorov-Smirnov, maupun grafik *Normal Probability Plot (P-P Plot)*.

Pada regresi linear berganda terdapat asumsi tambahan, yaitu tidak terjadi multikolinearitas antarvariabel independen. Multikolinearitas menyebabkan koefisien regresi menjadi tidak stabil sehingga sulit diinterpretasikan. Kondisi ini biasanya diperiksa menggunakan nilai *Variance Inflation Factor* (VIF) atau nilai toleransi.

Pemenuhan seluruh asumsi regresi linear akan menghasilkan model yang memiliki validitas statistik tinggi. Oleh karena itu, sebelum melakukan interpretasi maupun prediksi, peneliti wajib melakukan serangkaian uji asumsi agar model yang dibangun benar-benar layak digunakan sebagai dasar pengambilan keputusan.

2.6 Estimasi Parameter

Estimasi parameter merupakan proses menentukan nilai koefisien regresi yang paling sesuai untuk menggambarkan hubungan antara variabel

independen dan variabel dependen. Dalam regresi linear, parameter yang diestimasi meliputi konstanta (*intercept*) dan koefisien regresi (*slope*). Nilai parameter tersebut menjadi dasar dalam membangun persamaan regresi yang digunakan untuk menjelaskan hubungan antarvariabel serta melakukan prediksi terhadap data baru.

Metode estimasi parameter yang paling umum digunakan adalah Ordinary Least Squares (OLS). Prinsip utama OLS adalah mencari garis regresi yang menghasilkan jumlah kuadrat residual (*Sum of Squared Errors/SSE*) paling kecil. Residual merupakan selisih antara nilai aktual dengan nilai prediksi. Dengan meminimalkan jumlah kuadrat residual, model diharapkan mampu memberikan estimasi yang paling mendekati kondisi sebenarnya.

Selain menghasilkan nilai koefisien regresi, proses estimasi juga menghasilkan ukuran statistik lain, seperti *standard error*, nilai t, *p-value*, dan interval kepercayaan. Informasi tersebut digunakan untuk menguji apakah masing-masing variabel independen memiliki pengaruh yang signifikan terhadap variabel dependen. Variabel dengan nilai *p-value* yang lebih kecil dari tingkat signifikansi (misalnya 0,05) dianggap memiliki pengaruh yang signifikan terhadap model.

Dalam aplikasi prediksi permintaan usaha, estimasi parameter memungkinkan perusahaan mengetahui seberapa besar pengaruh harga, promosi, musim, atau faktor lainnya terhadap jumlah permintaan. Informasi ini menjadi dasar dalam menyusun strategi bisnis, seperti penentuan harga, perencanaan produksi, dan pengelolaan persediaan.

Saat ini, proses estimasi parameter dapat dilakukan dengan mudah menggunakan perangkat lunak statistik seperti SPSS, R, Python (Scikit-learn dan Statsmodels), SAS, maupun MATLAB. Meskipun demikian, pemahaman terhadap konsep estimasi tetap penting agar peneliti mampu

mengevaluasi hasil analisis secara kritis dan menghindari kesalahan interpretasi. Dengan estimasi parameter yang tepat, model regresi linear akan memiliki kemampuan prediksi yang lebih baik dan dapat digunakan sebagai dasar pengambilan keputusan berbasis data.

2.7 Interpretasi Koefisien

Interpretasi koefisien merupakan tahap penting dalam analisis regresi linear karena memberikan pemahaman mengenai hubungan antara variabel independen dan variabel dependen. Setelah model regresi dibangun dan parameter diestimasi, setiap koefisien regresi harus diinterpretasikan secara tepat agar hasil analisis dapat digunakan sebagai dasar pengambilan keputusan. Interpretasi ini tidak hanya menjelaskan arah hubungan antarvariabel, tetapi juga menunjukkan besarnya pengaruh masing-masing variabel terhadap variabel dependen.

Dalam persamaan regresi linear:

$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n + \varepsilon$

nilai β_0 (intercept) menunjukkan nilai variabel dependen ketika seluruh variabel independen bernilai nol. Meskipun dalam beberapa kasus nilai tersebut tidak memiliki makna praktis, konstanta tetap diperlukan untuk membentuk persamaan regresi. Sementara itu, β_1, β_2, hingga β_n merupakan koefisien regresi yang menunjukkan perubahan rata-rata pada variabel dependen akibat kenaikan satu satuan pada variabel independen, dengan asumsi variabel lainnya tetap (*ceteris paribus*).

Sebagai ilustrasi, misalkan diperoleh model prediksi permintaan usaha:

Permintaan = 500 − 8(Harga) + 12(Promosi)

Koefisien −8 pada variabel harga menunjukkan bahwa setiap kenaikan harga sebesar satu satuan akan menurunkan permintaan rata-rata sebanyak delapan unit, dengan asumsi faktor lain tetap. Sebaliknya,

koefisien 12 pada variabel promosi menunjukkan bahwa setiap peningkatan satu satuan aktivitas promosi akan meningkatkan permintaan rata-rata sebanyak dua belas unit.

Selain memperhatikan nilai koefisien, peneliti juga harus mengevaluasi signifikansi statistik melalui nilai t-hitung, p-value, dan interval kepercayaan. Koefisien yang signifikan secara statistik menunjukkan bahwa variabel tersebut benar-benar memberikan pengaruh terhadap variabel dependen. Dengan demikian, interpretasi koefisien tidak hanya memberikan pemahaman matematis mengenai model, tetapi juga menghasilkan informasi strategis yang dapat dimanfaatkan dalam penyusunan kebijakan bisnis, perencanaan pemasaran, dan pengambilan keputusan berbasis data.

2.8 Kelebihan dan Kekurangan

Regresi linear merupakan salah satu metode analisis prediktif yang paling banyak digunakan karena memiliki konsep yang sederhana dan mudah dipahami. Meskipun demikian, seperti metode statistik lainnya, regresi linear memiliki sejumlah kelebihan dan keterbatasan yang perlu dipahami sebelum diterapkan dalam suatu penelitian maupun pengembangan model prediksi.

Salah satu kelebihan utama regresi linear adalah kemudahan interpretasi. Hubungan antara variabel independen dan dependen dapat dijelaskan secara matematis melalui koefisien regresi sehingga pengguna dapat mengetahui besarnya pengaruh masing-masing variabel terhadap hasil prediksi. Selain itu, proses komputasi regresi linear relatif cepat dan tidak memerlukan sumber daya komputasi yang besar, sehingga sangat sesuai untuk dataset berukuran kecil hingga menengah. Regresi linear juga mudah diimplementasikan menggunakan berbagai perangkat lunak

statistik seperti SPSS, R, Python, SAS, maupun Microsoft Excel.

Di sisi lain, regresi linear memiliki beberapa keterbatasan. Metode ini mengasumsikan bahwa hubungan antara variabel bersifat linear. Apabila hubungan yang terjadi bersifat nonlinier, maka tingkat akurasi model dapat menurun secara signifikan. Regresi linear juga sensitif terhadap keberadaan data pencilan (*outlier*) karena data ekstrem dapat memengaruhi nilai koefisien regresi. Selain itu, model ini mengharuskan terpenuhinya beberapa asumsi statistik, seperti normalitas residual, homoskedastisitas, independensi residual, dan tidak adanya multikolinearitas.

Dalam perkembangan *machine learning*, regresi linear sering dijadikan sebagai model dasar (*baseline model*) untuk membandingkan performa algoritma lain, seperti Decision Tree Regression, Random Forest Regression, maupun Gradient Boosting Regression. Oleh karena itu, pemahaman mengenai kelebihan dan keterbatasan regresi linear akan membantu peneliti menentukan metode yang paling sesuai dengan karakteristik data dan tujuan analisis yang ingin dicapai.

2.9 Studi Kasus Regresi Linear

Untuk memahami penerapan regresi linear secara praktis, berikut disajikan contoh studi kasus mengenai prediksi permintaan usaha pada sebuah perusahaan ritel yang menjual produk kebutuhan rumah tangga. Perusahaan tersebut ingin mengetahui pengaruh harga dan promosi terhadap jumlah permintaan produk guna mendukung perencanaan produksi dan pengelolaan persediaan.

Data yang digunakan merupakan data historis penjualan selama dua belas bulan yang mencakup variabel harga produk, biaya promosi, dan jumlah permintaan. Setelah dilakukan proses pembersihan data (*data cleaning*)

dan analisis eksploratif, model regresi linear berganda dibangun menggunakan metode *Ordinary Least Squares* (OLS). Hasil analisis menghasilkan persamaan sebagai berikut:

$$\text{Permintaan} = 620 - 9{,}15(\text{Harga}) + 15{,}80\ (\text{Promosi})$$

Persamaan tersebut menunjukkan bahwa kenaikan harga sebesar satu satuan diperkirakan menurunkan permintaan rata-rata sebesar 9,15 unit, sedangkan peningkatan promosi sebesar satu satuan diperkirakan meningkatkan permintaan rata-rata sebesar 15,80 unit. Hasil uji signifikansi menunjukkan bahwa kedua variabel memiliki nilai *p-value* kurang dari 0,05 sehingga berpengaruh signifikan terhadap permintaan.

Evaluasi model dilakukan menggunakan beberapa metrik, antara lain Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), dan koefisien determinasi (R^2). Misalnya, diperoleh nilai R^2 sebesar 0,87, yang menunjukkan bahwa sekitar 87% variasi permintaan dapat dijelaskan oleh variabel harga dan promosi. Nilai tersebut mengindikasikan bahwa model memiliki kemampuan prediksi yang cukup baik.

Hasil prediksi kemudian dimanfaatkan oleh manajemen untuk menyusun strategi bisnis, seperti menentukan kebijakan harga, merencanakan program promosi, memperkirakan kebutuhan stok, dan mengoptimalkan jadwal produksi. Studi kasus ini menunjukkan bahwa regresi linear mampu memberikan informasi yang mudah dipahami sekaligus mendukung pengambilan keputusan berbasis data. Namun demikian, apabila pola hubungan antarvariabel menjadi lebih kompleks atau bersifat nonlinier, metode seperti Decision Tree Regression dapat menjadi alternatif yang lebih fleksibel dan akan dibahas pada bab berikutnya.

BAB 3
DECISION TREE REGRESSION

3.1 Pengantar Decision Tree

Decision Tree atau pohon keputusan merupakan salah satu metode *supervised learning* yang digunakan untuk proses klasifikasi maupun regresi. Algoritma ini bekerja dengan cara membagi data ke dalam beberapa kelompok berdasarkan nilai variabel tertentu sehingga membentuk struktur menyerupai pohon. Setiap percabangan (*branch*) menggambarkan aturan keputusan (*decision rule*), sedangkan setiap simpul akhir (*leaf node*) menunjukkan hasil prediksi. Karena menghasilkan aturan yang mudah dipahami, Decision Tree menjadi salah satu algoritma yang paling populer dalam bidang data mining dan machine learning.

Pada awal perkembangannya, Decision Tree lebih banyak digunakan untuk masalah klasifikasi, seperti klasifikasi penyakit, deteksi penipuan, dan segmentasi pelanggan. Seiring berkembangnya penelitian di bidang machine learning, metode ini juga dikembangkan untuk menyelesaikan masalah regresi yang menghasilkan keluaran berupa nilai kontinu. Pendekatan tersebut dikenal sebagai Decision Tree Regression, yaitu algoritma yang memprediksi nilai numerik berdasarkan proses pembagian data secara berulang hingga diperoleh kelompok data yang memiliki tingkat homogenitas tinggi.

Keunggulan utama Decision Tree terletak pada kemampuannya menangani hubungan yang kompleks tanpa memerlukan asumsi linearitas seperti pada regresi linear. Algoritma ini mampu mengenali

interaksi antarvariabel secara otomatis sehingga sangat efektif digunakan pada data yang memiliki pola nonlinier. Selain itu, Decision Tree dapat menangani data numerik maupun kategorikal, serta relatif tahan terhadap keberadaan data yang hilang (*missing values*) maupun pencilan (*outlier*) dibandingkan beberapa metode statistik konvensional.

Dalam dunia usaha, Decision Tree Regression banyak diterapkan untuk memprediksi permintaan produk, menentukan harga optimal, memperkirakan tingkat penjualan, mengelola persediaan, hingga melakukan analisis risiko. Kemudahan interpretasi model juga menjadi alasan utama mengapa algoritma ini sering digunakan sebagai dasar pengambilan keputusan oleh manajemen. Dengan memahami konsep dasar Decision Tree, pembaca akan lebih mudah mengikuti pembahasan mengenai mekanisme pembentukan pohon keputusan dan algoritma CART pada subbab berikutnya.

3.2 Struktur Decision Tree

Decision Tree memiliki struktur hierarkis yang menyerupai pohon, terdiri atas beberapa komponen utama yang saling berhubungan dalam membentuk proses pengambilan keputusan. Struktur tersebut meliputi root node, internal node, branch, dan leaf node. Setiap komponen memiliki fungsi yang berbeda namun saling melengkapi dalam menghasilkan model prediksi yang optimal.

Root node merupakan simpul pertama yang mewakili seluruh dataset sebelum dilakukan proses pembagian (*splitting*). Pada tahap ini algoritma akan memilih variabel terbaik sebagai dasar pemisahan data berdasarkan kriteria tertentu, seperti pengurangan varians (*variance reduction*) pada Decision Tree Regression. Selanjutnya, data dibagi ke

dalam beberapa kelompok sehingga menghasilkan internal node. Internal node merupakan simpul yang masih memiliki percabangan dan akan terus diproses hingga memenuhi kondisi penghentian.

Hubungan antar simpul direpresentasikan oleh branch atau cabang. Setiap cabang menunjukkan aturan keputusan (*decision rule*) berdasarkan nilai suatu variabel. Misalnya, apabila harga produk lebih kecil dari nilai tertentu maka data akan diarahkan ke cabang kiri, sedangkan apabila lebih besar maka akan menuju cabang kanan. Proses ini berlangsung secara berulang sehingga membentuk struktur pohon yang semakin rinci.

Bagian terakhir adalah leaf node atau simpul daun. Pada Decision Tree Regression, leaf node tidak menghasilkan kategori sebagaimana pada klasifikasi, melainkan menghasilkan nilai prediksi berupa rata-rata data yang berada pada simpul tersebut. Nilai inilah yang digunakan sebagai hasil akhir prediksi ketika terdapat data baru.

Semakin dalam struktur pohon, semakin spesifik aturan keputusan yang dihasilkan. Namun, pohon yang terlalu dalam berpotensi mengalami overfitting, yaitu kondisi ketika model sangat baik terhadap data pelatihan tetapi kurang mampu melakukan prediksi pada data baru. Oleh karena itu, struktur Decision Tree harus dibangun secara seimbang agar mampu menghasilkan model yang sederhana namun tetap memiliki tingkat akurasi tinggi.

3.3 Decision Tree untuk Regresi

Decision Tree Regression merupakan pengembangan algoritma Decision Tree yang digunakan untuk memprediksi variabel target berupa nilai kontinu atau numerik. Berbeda dengan Decision Tree Classification yang menghasilkan kelas atau kategori tertentu, Decision

Tree Regression menghasilkan nilai prediksi berdasarkan rata-rata data pada setiap simpul daun (*leaf node*). Oleh karena itu, metode ini sangat sesuai digunakan dalam berbagai kasus prediksi seperti permintaan usaha, harga rumah, penjualan, produksi, maupun konsumsi energi.

Prinsip kerja Decision Tree Regression dimulai dengan memilih variabel terbaik yang mampu membagi data menjadi beberapa kelompok dengan tingkat homogenitas yang tinggi. Algoritma kemudian menghitung kualitas setiap kemungkinan pembagian menggunakan ukuran seperti Mean Squared Error (MSE) atau Variance Reduction. Variabel yang menghasilkan penurunan kesalahan terbesar dipilih sebagai dasar pembentukan cabang pertama. Proses tersebut dilakukan secara rekursif hingga seluruh kondisi penghentian terpenuhi, misalnya jumlah data minimum pada simpul atau kedalaman pohon maksimum.

Salah satu keunggulan utama Decision Tree Regression adalah kemampuannya menangani hubungan nonlinier tanpa memerlukan transformasi data. Model juga mampu mengenali interaksi antarvariabel secara otomatis sehingga tidak diperlukan asumsi linearitas seperti pada regresi linear. Selain itu, Decision Tree Regression relatif mudah diinterpretasikan karena setiap keputusan dapat ditelusuri melalui aturan yang terbentuk pada struktur pohon.

Dalam prediksi permintaan usaha, misalnya, algoritma dapat menggunakan variabel harga, promosi, musim, hari libur, dan jumlah pelanggan untuk membentuk aturan keputusan. Sebagai contoh, apabila harga rendah dan promosi tinggi, maka model dapat memprediksi permintaan yang lebih besar dibandingkan kondisi lainnya. Pola-pola tersebut diperoleh secara otomatis dari data tanpa harus ditentukan terlebih dahulu oleh peneliti.

Karena fleksibilitas dan kemampuannya menangani pola data yang kompleks, Decision Tree Regression menjadi salah satu algoritma yang banyak digunakan dalam bidang data science dan machine learning. Algoritma ini juga menjadi dasar bagi metode ensemble yang lebih canggih, seperti Random Forest dan Gradient Boosting Regression, yang akan dibahas pada bab selanjutnya.

3.4 Algoritma CART

Classification and Regression Trees (CART) merupakan salah satu algoritma Decision Tree yang paling banyak digunakan dalam analisis data dan machine learning. Algoritma ini diperkenalkan oleh Leo Breiman, Jerome Friedman, Richard Olshen, dan Charles Stone pada tahun 1984 melalui buku *Classification and Regression Trees*. CART dirancang untuk menangani dua jenis permasalahan, yaitu klasifikasi dan regresi. Untuk klasifikasi, algoritma menghasilkan keluaran berupa kategori atau kelas, sedangkan pada regresi menghasilkan nilai kontinu seperti jumlah penjualan, harga, atau permintaan usaha.

Karakteristik utama CART adalah penggunaan binary split, yaitu setiap simpul (*node*) hanya dibagi menjadi dua cabang. Pendekatan ini menghasilkan struktur pohon yang lebih sederhana sehingga mudah dipahami dan diinterpretasikan. Pada kasus regresi, CART memilih variabel dan titik pemisah (*split point*) yang mampu meminimalkan variasi data dalam setiap kelompok. Proses tersebut dilakukan menggunakan ukuran seperti Variance Reduction atau Mean Squared Error (MSE).

Tahapan kerja algoritma CART dimulai dengan mengevaluasi seluruh variabel independen yang tersedia. Selanjutnya, algoritma menghitung seluruh kemungkinan titik pemisah untuk setiap variabel dan memilih

kombinasi yang menghasilkan penurunan kesalahan prediksi terbesar. Setelah pembagian dilakukan, setiap kelompok akan diproses kembali menggunakan prosedur yang sama hingga memenuhi kondisi penghentian, misalnya jumlah data minimum pada suatu simpul, kedalaman maksimum pohon (*maximum depth*), atau nilai penurunan kesalahan yang sudah sangat kecil.

Dalam prediksi permintaan usaha, CART dapat memanfaatkan variabel seperti harga produk, jumlah promosi, musim, hari libur, lokasi penjualan, maupun tingkat pendapatan masyarakat untuk membentuk aturan keputusan. Karena tidak memerlukan asumsi linearitas, algoritma ini mampu mengenali hubungan yang kompleks antarvariabel dan menghasilkan prediksi yang lebih fleksibel dibandingkan regresi linear.

Keunggulan lain dari CART adalah kemampuannya menangani data numerik maupun kategorikal, tidak sensitif terhadap skala data, serta relatif tahan terhadap data pencilan (*outlier*). Oleh karena itu, CART menjadi dasar bagi berbagai algoritma ensemble modern seperti Random Forest dan Gradient Boosting yang saat ini banyak digunakan dalam pengembangan sistem prediksi berbasis machine learning.

3.5 Pembentukan Pohon Keputusan

Pembentukan pohon keputusan (*Decision Tree Construction*) merupakan proses utama dalam algoritma Decision Tree Regression. Tujuan proses ini adalah menghasilkan struktur pohon yang mampu membagi data ke dalam kelompok-kelompok yang homogen sehingga setiap simpul daun (*leaf node*) memiliki tingkat variasi data yang rendah. Semakin homogen data pada setiap simpul, semakin baik kemampuan model dalam menghasilkan prediksi yang akurat.

Proses pembentukan pohon dimulai dari root node, yaitu simpul yang berisi seluruh data pelatihan (*training data*). Algoritma kemudian mengevaluasi seluruh variabel independen beserta kemungkinan nilai pemisahnya. Setiap kemungkinan pembagian dihitung menggunakan fungsi evaluasi, seperti Mean Squared Error (MSE) atau Variance Reduction, untuk menentukan kualitas pembagian data.

Setelah diperoleh variabel terbaik, data dibagi menjadi dua kelompok sesuai aturan yang terbentuk. Masing-masing kelompok selanjutnya diperlakukan sebagai simpul baru (*child node*) yang kembali dievaluasi menggunakan prosedur yang sama. Proses ini berlangsung secara rekursif, yaitu berulang hingga salah satu kondisi penghentian (*stopping criteria*) tercapai.

Beberapa kondisi penghentian yang umum digunakan antara lain:

- jumlah data pada simpul telah mencapai batas minimum (*minimum samples split*);
- kedalaman pohon telah mencapai batas maksimum (*maximum depth*);
- nilai penurunan kesalahan (*impurity reduction*) sudah sangat kecil; atau
- seluruh data pada simpul memiliki nilai target yang relatif sama.

Setelah pohon selesai dibangun, setiap simpul daun akan menyimpan nilai prediksi berupa rata-rata variabel target dari seluruh data yang berada pada simpul tersebut. Ketika terdapat data baru, proses prediksi dilakukan dengan mengikuti aturan percabangan dari root node hingga mencapai leaf node yang sesuai.

Pada prediksi permintaan usaha, proses pembentukan pohon memungkinkan algoritma menemukan pola keputusan, misalnya

kombinasi harga rendah, promosi tinggi, dan musim tertentu yang menghasilkan tingkat permintaan paling besar. Kemampuan ini menjadikan Decision Tree Regression sangat efektif dalam memodelkan hubungan yang kompleks tanpa memerlukan transformasi matematis sebagaimana pada regresi linear.

3.6 Kriteria Split

Keberhasilan Decision Tree Regression sangat ditentukan oleh pemilihan kriteria split, yaitu metode yang digunakan untuk memilih variabel dan titik pemisah terbaik pada setiap simpul pohon. Tujuan utama proses split adalah menghasilkan kelompok data yang memiliki tingkat homogenitas tinggi sehingga kesalahan prediksi menjadi sekecil mungkin. Pemilihan split yang tepat akan menghasilkan struktur pohon yang lebih efisien dan memiliki kemampuan generalisasi yang lebih baik.

Pada Decision Tree Regression, kriteria split yang paling umum digunakan adalah Variance Reduction dan Mean Squared Error (MSE). Variance Reduction mengukur seberapa besar penurunan variasi data setelah dilakukan pembagian. Algoritma akan memilih variabel yang menghasilkan penurunan varians terbesar karena menunjukkan bahwa data pada masing-masing kelompok menjadi lebih seragam.

Sementara itu, Mean Squared Error (MSE) digunakan untuk mengukur rata-rata kuadrat selisih antara nilai aktual dengan nilai rata-rata pada setiap simpul. Nilai MSE yang lebih kecil menunjukkan bahwa data dalam kelompok memiliki tingkat penyimpangan yang rendah sehingga prediksi menjadi lebih akurat. Oleh karena itu, algoritma selalu memilih split yang menghasilkan nilai MSE paling kecil.

Selain kedua ukuran tersebut, beberapa implementasi Decision Tree juga menggunakan ukuran lain, seperti Mean Absolute Error (MAE) atau fungsi kerugian (*loss function*) tertentu sesuai karakteristik data. Namun, pada implementasi CART yang terdapat dalam pustaka Scikit-learn, MSE masih menjadi pilihan utama karena memberikan hasil yang stabil untuk sebagian besar kasus regresi.

Sebagai contoh, pada prediksi permintaan usaha, algoritma dapat mengevaluasi berbagai titik pemisah pada variabel harga. Misalnya, algoritma membandingkan pembagian data pada harga ≤ Rp20.000 dan harga > Rp20.000, kemudian menghitung MSE masing-masing kelompok. Split yang menghasilkan penurunan MSE terbesar akan dipilih sebagai percabangan berikutnya.

Pemilihan kriteria split yang optimal tidak hanya meningkatkan akurasi prediksi, tetapi juga menghasilkan struktur pohon yang lebih sederhana sehingga mudah diinterpretasikan oleh pengguna. Oleh sebab itu, proses ini menjadi inti dari pembangunan model Decision Tree Regression dan sangat menentukan kualitas hasil prediksi yang dihasilkan.

BAB 4
PERSIAPAN DATA

4.1 Pengumpulan Data

Pengumpulan data merupakan tahapan awal yang sangat menentukan keberhasilan proses analisis dan pembangunan model prediksi. Kualitas model Regresi Linear maupun Decision Tree Regression sangat dipengaruhi oleh kualitas data yang digunakan. Data yang lengkap, akurat, relevan, dan representatif akan menghasilkan model dengan tingkat akurasi yang lebih tinggi dibandingkan data yang tidak konsisten atau mengandung banyak kesalahan. Oleh karena itu, proses pengumpulan data harus dilakukan secara sistematis sesuai dengan tujuan penelitian dan karakteristik permasalahan yang akan diselesaikan. Dalam penelitian prediksi permintaan usaha, data dapat diperoleh dari berbagai sumber, baik internal maupun eksternal. Sumber internal meliputi sistem informasi penjualan, sistem inventori, laporan transaksi, data pelanggan, dan laporan keuangan perusahaan. Sementara itu, sumber eksternal dapat berasal dari Badan Pusat Statistik (BPS), Bank Indonesia, laporan industri, media sosial, marketplace, data cuaca, maupun kalender hari besar nasional yang dapat memengaruhi pola permintaan.

Tahapan pengumpulan data dimulai dengan mengidentifikasi variabel yang relevan, menentukan periode pengamatan, memilih sumber data, dan memastikan legalitas serta keandalan data. Selanjutnya dilakukan proses validasi awal untuk memeriksa kelengkapan data, konsistensi format, dan kemungkinan adanya data duplikat. Pada tahap ini,

dokumentasi mengenai asal data, waktu pengambilan, serta metode pengumpulan juga perlu dilakukan agar penelitian dapat direplikasi dan dipertanggungjawabkan secara ilmiah.

Dalam praktiknya, pengumpulan data sering melibatkan integrasi beberapa sumber data. Sebagai contoh, data historis penjualan perusahaan dapat dikombinasikan dengan data promosi, kondisi cuaca, hari libur, dan indikator ekonomi untuk menghasilkan model prediksi yang lebih komprehensif. Integrasi berbagai sumber data memungkinkan algoritma mengenali pola yang lebih kompleks sehingga meningkatkan kemampuan prediksi terhadap permintaan usaha.

4.2 Data Permintaan Usaha

Data permintaan usaha merupakan variabel utama dalam penelitian yang bertujuan membangun model prediksi menggunakan Regresi Linear maupun Decision Tree Regression. Data ini menggambarkan jumlah produk atau jasa yang diminta oleh konsumen dalam periode tertentu. Informasi mengenai permintaan sangat penting karena menjadi dasar dalam penyusunan rencana produksi, pengendalian persediaan, strategi pemasaran, serta pengambilan keputusan operasional perusahaan.

Karakteristik data permintaan sangat dipengaruhi oleh jenis usaha yang dijalankan. Pada sektor ritel, data permintaan biasanya dinyatakan dalam jumlah unit produk yang terjual setiap hari, minggu, atau bulan. Pada sektor manufaktur, data dapat berupa jumlah pesanan pelanggan, sedangkan pada sektor jasa dapat berupa jumlah pelanggan yang menggunakan layanan dalam periode tertentu. Oleh karena itu, pemilihan satuan pengukuran harus disesuaikan dengan karakteristik bisnis yang dianalisis.

Dalam penelitian ini, data permintaan usaha berperan sebagai variabel

dependen (target) yang akan diprediksi berdasarkan sejumlah variabel independen, seperti harga, promosi, musim, hari libur, dan faktor ekonomi. Agar model menghasilkan prediksi yang akurat, data permintaan harus memiliki cakupan waktu yang cukup panjang sehingga mampu merepresentasikan berbagai kondisi bisnis. Data dengan periode pengamatan yang lebih panjang umumnya memberikan informasi yang lebih lengkap mengenai pola tren, fluktuasi, dan perubahan musiman.

Sebelum digunakan dalam proses pemodelan, data permintaan perlu melalui tahap validasi untuk memastikan tidak terdapat data ganda, nilai yang tidak logis, atau kesalahan pencatatan. Selain itu, analisis statistik deskriptif juga dilakukan untuk memahami karakteristik data, seperti nilai minimum, maksimum, rata-rata, median, simpangan baku, serta distribusi data.

4.3 Data Historis Penjualan

Data historis penjualan merupakan kumpulan informasi mengenai aktivitas penjualan yang telah terjadi pada masa lalu dan menjadi salah satu sumber data terpenting dalam pengembangan model prediksi permintaan usaha. Data historis mencerminkan perilaku konsumen, pola pembelian, tren pasar, serta perubahan permintaan dari waktu ke waktu. Oleh karena itu, hampir seluruh metode peramalan modern memanfaatkan data historis sebagai dasar dalam membangun model prediksi.

Komponen data historis penjualan umumnya meliputi tanggal transaksi, kode produk, nama produk, jumlah unit terjual, harga jual, nilai transaksi, diskon, lokasi penjualan, serta identitas pelanggan apabila tersedia. Semakin lengkap atribut yang dimiliki, semakin besar peluang model machine learning untuk mengenali pola hubungan antarvariabel yang

memengaruhi permintaan.

Dalam penelitian prediksi permintaan usaha, data historis penjualan berfungsi sebagai sumber utama untuk melatih (*training*) model Regresi Linear maupun Decision Tree Regression. Algoritma akan mempelajari hubungan antara variabel-variabel prediktor dengan jumlah penjualan yang terjadi pada periode sebelumnya. Setelah pola tersebut dipahami, model dapat digunakan untuk memprediksi permintaan pada periode mendatang.

Sebelum digunakan, data historis harus melalui proses pemeriksaan kualitas. Data yang tidak lengkap, transaksi ganda, kesalahan pencatatan, maupun perubahan format pencatatan harus diperbaiki melalui proses *data cleaning*. Selain itu, analisis tren dan pola musiman juga perlu dilakukan agar karakteristik penjualan dapat dipahami secara menyeluruh. Misalnya, peningkatan penjualan menjelang hari raya atau penurunan permintaan pada musim tertentu dapat menjadi informasi penting dalam pembangunan model.

Data historis yang memiliki rentang waktu panjang dan kualitas tinggi akan menghasilkan model prediksi yang lebih stabil dan akurat. Oleh karena itu, organisasi perlu menerapkan sistem pencatatan transaksi yang terintegrasi sehingga data penjualan dapat dimanfaatkan secara optimal dalam mendukung pengambilan keputusan berbasis data.

4.4 Data Musiman

Data musiman (*seasonal data*) merupakan data yang menunjukkan pola perubahan permintaan yang terjadi secara berulang pada periode tertentu, seperti harian, mingguan, bulanan, atau tahunan. Dalam dunia usaha, pola musiman menjadi salah satu faktor yang sangat memengaruhi tingkat permintaan suatu produk atau jasa. Oleh karena itu, data musiman perlu

diperhatikan dalam pembangunan model prediksi agar hasil yang diperoleh lebih akurat dan mampu mencerminkan kondisi bisnis yang sebenarnya.

Pola musiman dapat dipengaruhi oleh berbagai faktor, seperti pergantian musim, hari besar keagamaan, libur nasional, awal tahun ajaran baru, promosi tahunan, maupun kebiasaan masyarakat. Sebagai contoh, permintaan pakaian muslim cenderung meningkat menjelang Hari Raya Idulfitri, sedangkan penjualan alat tulis mengalami peningkatan menjelang tahun ajaran baru. Produk makanan dan minuman juga sering menunjukkan peningkatan permintaan pada akhir pekan atau hari libur panjang.

Dalam proses analisis data, pola musiman dapat diidentifikasi melalui visualisasi data menggunakan grafik deret waktu (*time series*), analisis tren, maupun teknik dekomposisi data. Hasil analisis tersebut membantu peneliti memahami kapan terjadi peningkatan atau penurunan permintaan sehingga pola tersebut dapat dimasukkan sebagai variabel prediktor dalam model.

Baik Regresi Linear maupun Decision Tree Regression dapat memanfaatkan data musiman sebagai variabel independen. Pada Regresi Linear, musim biasanya direpresentasikan dalam bentuk variabel dummy, sedangkan pada Decision Tree Regression, algoritma dapat secara otomatis mengenali pola musiman melalui proses pembentukan percabangan. Dengan demikian, model mampu menghasilkan prediksi yang lebih sesuai dengan kondisi nyata.

Penggunaan data musiman juga memberikan manfaat strategis bagi perusahaan dalam menyusun jadwal produksi, mengelola persediaan, mengoptimalkan distribusi, dan merancang program promosi. Oleh sebab itu, identifikasi pola musiman menjadi salah satu tahap penting

dalam proses persiapan data untuk mendukung keberhasilan sistem prediksi permintaan usaha.

4.5 Data Eksternal

Selain data internal perusahaan, pembangunan model prediksi permintaan usaha juga memerlukan data eksternal sebagai sumber informasi tambahan yang mampu menjelaskan perubahan kondisi lingkungan bisnis. Data eksternal merupakan data yang diperoleh dari luar organisasi dan sering kali memiliki pengaruh signifikan terhadap tingkat permintaan produk atau jasa. Integrasi data internal dan eksternal memungkinkan model prediksi menghasilkan estimasi yang lebih akurat dibandingkan hanya menggunakan data historis penjualan.

Jenis data eksternal yang umum digunakan meliputi indikator ekonomi, tingkat inflasi, suku bunga, nilai tukar mata uang, pertumbuhan pendapatan masyarakat, data cuaca, kepadatan lalu lintas, kalender hari besar nasional, tren media sosial, hingga indeks kepercayaan konsumen. Di Indonesia, data tersebut dapat diperoleh dari berbagai lembaga resmi seperti Badan Pusat Statistik (BPS), Bank Indonesia (BI), Badan Meteorologi, Klimatologi, dan Geofisika (BMKG), serta kementerian terkait.

Sebagai contoh, permintaan payung cenderung meningkat pada musim hujan, sedangkan permintaan minuman dingin meningkat ketika suhu udara tinggi. Demikian pula, perubahan tingkat inflasi dapat memengaruhi daya beli masyarakat sehingga berdampak pada jumlah permintaan berbagai produk. Oleh karena itu, memasukkan variabel-variabel eksternal ke dalam model dapat meningkatkan kemampuan algoritma dalam mengenali pola yang tidak dapat dijelaskan hanya oleh data internal.

Sebelum digunakan, data eksternal perlu disesuaikan dengan periode pengamatan dan format data internal agar dapat diintegrasikan dengan baik. Proses sinkronisasi waktu, penyamaan satuan pengukuran, serta validasi kualitas data menjadi langkah penting untuk menghindari kesalahan analisis.

Pemanfaatan data eksternal menunjukkan bahwa prediksi permintaan usaha tidak hanya dipengaruhi oleh faktor internal perusahaan, tetapi juga oleh dinamika lingkungan bisnis. Dengan mengintegrasikan kedua sumber data tersebut, perusahaan dapat menyusun strategi yang lebih adaptif terhadap perubahan pasar dan meningkatkan kualitas pengambilan keputusan.

4.6 Data Cleaning

Data cleaning atau pembersihan data merupakan proses memperbaiki kualitas data sebelum digunakan dalam analisis maupun pembangunan model machine learning. Tahapan ini bertujuan mengidentifikasi dan memperbaiki berbagai permasalahan pada dataset, seperti data duplikat, kesalahan penulisan, inkonsistensi format, nilai yang hilang (*missing values*), serta data yang tidak sesuai dengan aturan bisnis. Data yang bersih akan meningkatkan akurasi model dan mengurangi risiko kesalahan interpretasi hasil analisis.

Proses data cleaning diawali dengan pemeriksaan struktur dataset untuk memastikan seluruh variabel memiliki tipe data yang sesuai, seperti numerik, kategorikal, tanggal, atau teks. Selanjutnya dilakukan identifikasi data duplikat yang dapat menyebabkan hasil analisis menjadi bias. Data yang memiliki nilai tidak logis, misalnya jumlah penjualan bernilai negatif atau tanggal transaksi yang tidak valid, juga harus diperbaiki atau dihapus.

Tahap berikutnya adalah standardisasi format data. Sebagai contoh, format tanggal harus diseragamkan, penulisan nama produk harus konsisten, dan satuan pengukuran harus menggunakan standar yang sama. Pada dataset yang berasal dari berbagai sumber, proses integrasi sering kali menghasilkan perbedaan format sehingga standardisasi menjadi sangat penting.

Dalam penelitian prediksi permintaan usaha, data cleaning memberikan dampak langsung terhadap performa Regresi Linear maupun Decision Tree Regression. Regresi Linear sangat sensitif terhadap kesalahan data karena dapat memengaruhi estimasi parameter, sedangkan Decision Tree Regression dapat menghasilkan struktur pohon yang kurang optimal apabila dataset mengandung banyak data tidak konsisten.

Saat ini, proses data cleaning dapat dilakukan menggunakan berbagai perangkat lunak seperti Microsoft Excel, OpenRefine, Python (Pandas), maupun R. Meskipun sebagian besar proses dapat diotomatisasi, validasi manual tetap diperlukan untuk memastikan bahwa perubahan yang dilakukan tidak menghilangkan informasi penting. Dengan demikian, data cleaning menjadi salah satu tahapan paling krusial dalam persiapan data karena menentukan kualitas model prediksi yang akan dibangun.

BAB 5
PEMBANGUNAN MODEL PREDIKSI

5.1 Tahapan Machine Learning

Pembangunan model prediksi menggunakan *Machine Learning* merupakan proses sistematis yang terdiri atas beberapa tahapan yang saling berkaitan. Setiap tahapan memiliki peran penting dalam menghasilkan model yang mampu melakukan prediksi secara akurat, efisien, dan memiliki kemampuan generalisasi terhadap data baru. Dalam penelitian prediksi permintaan usaha, tahapan ini diterapkan untuk membangun model menggunakan metode Regresi Linear dan Decision Tree Regression sehingga mampu memperkirakan permintaan produk berdasarkan data historis dan faktor-faktor yang memengaruhinya.

Tahap pertama adalah identifikasi permasalahan (problem definition), yaitu menentukan tujuan analisis, variabel target, serta ruang lingkup penelitian. Pada tahap ini peneliti juga menentukan jenis pendekatan machine learning yang digunakan, misalnya *supervised learning* untuk kasus prediksi nilai numerik. Tahap berikutnya adalah pengumpulan data, yang mencakup pengambilan data dari berbagai sumber internal maupun eksternal sesuai dengan kebutuhan penelitian.

Setelah data terkumpul, dilakukan prapemrosesan data (data preprocessing) yang meliputi pembersihan data, penanganan *missing value*, penghapusan data duplikat, transformasi data, normalisasi, serta *feature engineering*. Tahapan ini bertujuan menghasilkan dataset yang bersih dan siap digunakan dalam proses pelatihan model.

Selanjutnya dilakukan pembagian dataset menjadi data pelatihan

(*training data*) dan data pengujian (*testing data*). Model kemudian dibangun menggunakan algoritma Regresi Linear maupun Decision Tree Regression, dilanjutkan dengan proses pelatihan (*training*) agar algoritma mampu mempelajari pola hubungan antarvariabel. Setelah model selesai dilatih, dilakukan evaluasi menggunakan metrik seperti Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Percentage Error (MAPE), dan koefisien determinasi (R^2).

Tahap terakhir adalah implementasi dan pemantauan model. Model yang telah memenuhi kriteria performa dapat digunakan untuk memprediksi permintaan usaha dan mendukung pengambilan keputusan. Pemantauan berkala tetap diperlukan karena pola data dapat berubah seiring waktu, sehingga model perlu diperbarui (*retraining*) agar tetap memberikan hasil prediksi yang optimal.

5.2 Training Data

Training data atau data pelatihan merupakan sekumpulan data yang digunakan untuk melatih algoritma machine learning agar mampu mengenali pola hubungan antara variabel independen dan variabel dependen. Dalam pembelajaran terawasi (*supervised learning*), training data terdiri atas data masukan (*features*) dan data target (*label*) yang telah diketahui nilainya. Algoritma akan memanfaatkan data tersebut untuk membangun model matematis yang dapat digunakan dalam proses prediksi.

Kualitas training data sangat memengaruhi performa model. Dataset yang lengkap, representatif, dan memiliki distribusi yang baik akan menghasilkan model dengan kemampuan generalisasi yang lebih tinggi. Sebaliknya, data yang terlalu sedikit atau tidak mencerminkan kondisi

sebenarnya dapat menyebabkan model mengalami *underfitting* atau *overfitting*. Oleh karena itu, sebelum digunakan sebagai data pelatihan, dataset harus melalui proses pembersihan dan validasi sebagaimana telah dibahas pada bab sebelumnya.

Dalam penelitian prediksi permintaan usaha, training data biasanya terdiri atas data historis penjualan beserta variabel-variabel yang memengaruhi permintaan, seperti harga produk, jumlah promosi, musim, hari libur, kondisi ekonomi, dan faktor lainnya. Algoritma Regresi Linear akan mempelajari hubungan matematis antarvariabel, sedangkan Decision Tree Regression akan membangun aturan keputusan berdasarkan proses pembagian data secara bertahap.

Proporsi pembagian data pelatihan dapat bervariasi, misalnya 70:30, 80:20, atau 90:10 antara training data dan testing data. Pemilihan proporsi bergantung pada jumlah data yang tersedia. Untuk dataset berukuran besar, penggunaan 80% sebagai data pelatihan umumnya sudah mampu menghasilkan model yang stabil.

Training data yang berkualitas akan meningkatkan kemampuan model dalam mengenali pola historis dan menghasilkan prediksi yang lebih akurat. Oleh karena itu, penyusunan training data menjadi salah satu langkah paling penting dalam proses pembangunan model machine learning.

5.3 Testing Data

Testing data atau data pengujian merupakan bagian dari dataset yang tidak digunakan selama proses pelatihan model, tetapi dimanfaatkan untuk mengukur kemampuan model dalam melakukan prediksi terhadap data baru. Tujuan utama penggunaan testing data adalah mengevaluasi kemampuan generalisasi model sehingga dapat diketahui apakah model

hanya menghafal data pelatihan atau benar-benar mampu mengenali pola yang berlaku secara umum.

Dalam praktik machine learning, testing data harus benar-benar dipisahkan dari training data sebelum proses pelatihan dimulai. Pemisahan ini bertujuan menghindari *data leakage*, yaitu kondisi ketika informasi dari data pengujian secara tidak sengaja digunakan selama pelatihan sehingga menghasilkan evaluasi yang bias. Oleh karena itu, seluruh proses pelatihan, pemilihan fitur, maupun optimasi parameter harus dilakukan hanya menggunakan training data.

Pada penelitian prediksi permintaan usaha, testing data berisi data penjualan yang belum pernah dipelajari oleh model. Setelah model selesai dibangun, data tersebut digunakan untuk membandingkan nilai prediksi dengan nilai aktual. Selisih antara kedua nilai tersebut kemudian dihitung menggunakan berbagai metrik evaluasi, seperti MAE, MSE, RMSE, MAPE, dan R^2.

Sebagai contoh, apabila perusahaan memiliki data penjualan selama lima tahun, maka sekitar 80% data dapat digunakan sebagai training data, sedangkan 20% sisanya dijadikan testing data. Pendekatan ini memungkinkan peneliti memperoleh gambaran mengenai kemampuan model dalam memprediksi kondisi yang belum pernah diamati sebelumnya.

Keberhasilan suatu model tidak hanya ditentukan oleh tingginya akurasi pada training data, tetapi juga oleh performanya pada testing data. Model yang menunjukkan hasil baik pada kedua dataset memiliki kemampuan generalisasi yang baik dan lebih layak diterapkan dalam lingkungan bisnis. Oleh karena itu, testing data menjadi komponen penting dalam proses validasi model sebelum diimplementasikan sebagai sistem pendukung keputusan.

5.4 Cross Validation

Cross validation merupakan teknik evaluasi model dalam machine learning yang digunakan untuk mengukur kemampuan generalisasi suatu model terhadap data yang belum pernah dipelajari sebelumnya. Berbeda dengan metode pembagian data sederhana (*hold-out validation*), cross validation memanfaatkan seluruh dataset secara bergantian sebagai data pelatihan dan data pengujian sehingga hasil evaluasi menjadi lebih stabil dan representatif. Teknik ini sangat penting ketika jumlah data yang tersedia relatif terbatas karena memungkinkan setiap data berkontribusi dalam proses pelatihan maupun pengujian.

Metode yang paling umum digunakan adalah K-Fold Cross Validation. Pada metode ini, dataset dibagi menjadi *k* bagian (*fold*) dengan ukuran yang sama. Salah satu fold digunakan sebagai data pengujian, sedangkan fold lainnya digunakan sebagai data pelatihan. Proses tersebut diulang sebanyak *k* kali hingga setiap fold memperoleh kesempatan menjadi data pengujian. Nilai performa model kemudian dihitung sebagai rata-rata dari seluruh proses pengujian. Nilai *k* yang sering digunakan adalah 5 atau 10 karena memberikan keseimbangan antara akurasi evaluasi dan efisiensi komputasi.

Dalam penelitian prediksi permintaan usaha, cross validation memberikan gambaran yang lebih objektif mengenai kemampuan Regresi Linear dan Decision Tree Regression dalam memprediksi permintaan pada berbagai kondisi data. Teknik ini juga membantu mengurangi risiko evaluasi yang dipengaruhi oleh pembagian dataset secara acak.

Implementasi cross validation pada Python umumnya menggunakan pustaka Scikit-learn, khususnya kelas KFold atau cross_val_score. Melalui pendekatan ini, peneliti dapat memperoleh estimasi performa

model yang lebih andal sebelum model diterapkan pada lingkungan operasional.

Penggunaan cross validation menjadi praktik yang direkomendasikan dalam machine learning karena mampu meningkatkan kepercayaan terhadap hasil evaluasi model. Dengan demikian, model yang dipilih benar-benar memiliki kemampuan prediksi yang baik dan tidak hanya sesuai terhadap data pelatihan.

5.5 Pemodelan Regresi Linear

Pemodelan Regresi Linear merupakan proses membangun persamaan matematis yang menggambarkan hubungan antara variabel independen dan variabel dependen. Dalam penelitian prediksi permintaan usaha, metode ini digunakan untuk memperkirakan jumlah permintaan berdasarkan faktor-faktor seperti harga produk, promosi, musim, hari libur, maupun variabel ekonomi lainnya. Regresi Linear dipilih karena memiliki konsep yang sederhana, mudah diinterpretasikan, serta mampu memberikan informasi mengenai pengaruh masing-masing variabel terhadap permintaan.

Tahapan pemodelan diawali dengan pemilihan variabel prediktor (*feature selection*) berdasarkan hasil analisis eksploratif dan relevansi terhadap tujuan penelitian. Selanjutnya dilakukan pembagian dataset menjadi data pelatihan dan data pengujian. Model kemudian dilatih menggunakan metode Ordinary Least Squares (OLS) yang bertujuan memperoleh koefisien regresi dengan meminimalkan jumlah kuadrat galat (*Sum of Squared Errors*).

Setelah model terbentuk, dilakukan pengujian terhadap asumsi-asumsi regresi, seperti linearitas, normalitas residual, homoskedastisitas, independensi residual, dan multikolinearitas. Apabila seluruh asumsi

terpenuhi, model dapat digunakan untuk melakukan prediksi terhadap data baru. Selanjutnya performa model dievaluasi menggunakan metrik seperti MAE, MSE, RMSE, MAPE, dan koefisien determinasi (R^2).

Dalam implementasi menggunakan Python, pembangunan model dapat dilakukan melalui pustaka Scikit-learn menggunakan kelas LinearRegression atau pustaka Statsmodels apabila diperlukan analisis statistik yang lebih rinci, seperti nilai *p-value*, uji F, dan interval kepercayaan. Hasil pemodelan tidak hanya menghasilkan nilai prediksi, tetapi juga memberikan informasi mengenai kontribusi setiap variabel terhadap perubahan permintaan usaha.

Meskipun memiliki keterbatasan dalam menangani hubungan nonlinier, Regresi Linear tetap menjadi model dasar (*baseline model*) yang penting untuk dibandingkan dengan algoritma machine learning lainnya. Oleh karena itu, metode ini masih banyak digunakan dalam penelitian akademik maupun aplikasi bisnis sebagai acuan dalam mengevaluasi performa model prediksi.

5.6 Pemodelan Decision Tree Regression

Pemodelan Decision Tree Regression bertujuan membangun model prediksi yang mampu menghasilkan nilai kontinu melalui struktur pohon keputusan. Berbeda dengan Regresi Linear yang membentuk persamaan matematis, Decision Tree Regression membangun serangkaian aturan keputusan berdasarkan proses pembagian data secara bertahap. Pendekatan ini memungkinkan algoritma mengenali hubungan nonlinier serta interaksi antarvariabel tanpa memerlukan asumsi statistik tertentu.

Proses pemodelan dimulai dengan menentukan variabel target dan variabel prediktor yang akan digunakan. Dataset kemudian dibagi menjadi data pelatihan dan data pengujian. Selanjutnya algoritma CART

(*Classification and Regression Trees*) mengevaluasi seluruh variabel independen untuk mencari titik pemisah (*split point*) terbaik berdasarkan penurunan varians (*Variance Reduction*) atau Mean Squared Error (MSE). Variabel yang menghasilkan penurunan kesalahan terbesar akan dipilih sebagai dasar pembentukan cabang pertama.

Proses pembagian data dilakukan secara rekursif hingga memenuhi kondisi penghentian (*stopping criteria*), seperti kedalaman maksimum pohon (*maximum depth*), jumlah minimum data pada simpul (*minimum samples split*), atau jumlah minimum data pada simpul daun (*minimum samples leaf*). Setiap simpul daun kemudian menghasilkan nilai prediksi berupa rata-rata variabel target dari seluruh data yang berada pada simpul tersebut.

Implementasi Decision Tree Regression menggunakan Python umumnya memanfaatkan pustaka Scikit-learn, khususnya kelas DecisionTreeRegressor. Beberapa parameter penting yang dapat dikonfigurasi meliputi criterion, max_depth, min_samples_split, min_samples_leaf, dan random_state. Pengaturan parameter tersebut sangat berpengaruh terhadap kompleksitas struktur pohon dan tingkat akurasi model.

Setelah model selesai dibangun, evaluasi dilakukan menggunakan data pengujian dengan menghitung nilai MAE, MSE, RMSE, MAPE, dan R^2. Selain itu, struktur pohon juga dapat divisualisasikan sehingga peneliti dapat memahami aturan keputusan yang dihasilkan. Kemudahan interpretasi tersebut menjadi salah satu keunggulan utama Decision Tree Regression dibandingkan beberapa algoritma machine learning lainnya.

Dalam prediksi permintaan usaha, Decision Tree Regression mampu mengidentifikasi kombinasi faktor seperti harga, promosi, musim, dan kondisi ekonomi yang menghasilkan tingkat permintaan tertentu.

Fleksibilitas ini menjadikan Decision Tree Regression sangat sesuai untuk menangani data bisnis yang memiliki pola kompleks dan hubungan nonlinier.

5.7 Optimasi Parameter

Optimasi parameter (*hyperparameter tuning*) merupakan tahapan penting dalam pembangunan model machine learning yang bertujuan memperoleh konfigurasi parameter terbaik sehingga model memiliki tingkat akurasi yang optimal. Berbeda dengan parameter model yang dipelajari secara otomatis selama proses pelatihan, *hyperparameter* ditentukan oleh pengguna sebelum proses pelatihan dimulai. Pemilihan nilai *hyperparameter* yang tepat dapat meningkatkan kemampuan model dalam mengenali pola data sekaligus mengurangi risiko *overfitting* maupun *underfitting*.

Pada model Regresi Linear, proses optimasi parameter relatif sederhana karena algoritma ini memiliki sedikit *hyperparameter*. Namun demikian, optimasi masih dapat dilakukan melalui proses seleksi variabel (*feature selection*), penerapan teknik regularisasi seperti Ridge Regression, Lasso Regression, atau Elastic Net, serta standardisasi data apabila diperlukan. Teknik-teknik tersebut membantu meningkatkan stabilitas model dan mengurangi pengaruh multikolinearitas antarvariabel independen.

Sebaliknya, Decision Tree Regression memiliki beberapa *hyperparameter* yang sangat memengaruhi bentuk pohon keputusan dan performa model. Parameter yang umum dioptimalkan meliputi max_depth, min_samples_split, min_samples_leaf, max_features, dan criterion. Parameter max_depth mengatur kedalaman maksimum pohon, sedangkan min_samples_split menentukan jumlah minimum data yang diperlukan sebelum suatu simpul dapat dibagi. Sementara itu,

min_samples_leaf mengatur jumlah minimum data pada setiap simpul daun agar model tidak terlalu spesifik terhadap data pelatihan.

Beberapa metode optimasi yang banyak digunakan antara lain Grid Search, Random Search, dan Bayesian Optimization. Grid Search mengevaluasi seluruh kombinasi parameter yang telah ditentukan, sedangkan Random Search memilih kombinasi parameter secara acak sehingga lebih efisien pada ruang pencarian yang besar. Dalam implementasi menggunakan Python, proses optimasi dapat dilakukan menggunakan kelas GridSearchCV dan RandomizedSearchCV dari pustaka Scikit-learn, yang umumnya dikombinasikan dengan teknik K-Fold Cross Validation untuk memperoleh hasil evaluasi yang lebih objektif.

Optimasi parameter yang dilakukan secara sistematis akan menghasilkan model dengan tingkat akurasi yang lebih tinggi, kemampuan generalisasi yang lebih baik, serta risiko kesalahan prediksi yang lebih rendah. Oleh karena itu, tahapan ini menjadi bagian yang tidak terpisahkan dalam pembangunan sistem prediksi permintaan usaha berbasis machine learning.

5.8 Interpretasi Model

Interpretasi model merupakan proses memahami dan menjelaskan bagaimana suatu model machine learning menghasilkan prediksi. Tahapan ini sangat penting karena hasil prediksi tidak hanya harus memiliki tingkat akurasi yang tinggi, tetapi juga mampu memberikan informasi yang dapat dipahami oleh pengguna sebagai dasar dalam pengambilan keputusan. Dalam penelitian prediksi permintaan usaha, interpretasi model membantu menjelaskan faktor-faktor yang paling berpengaruh terhadap perubahan permintaan sehingga hasil analisis

dapat dimanfaatkan secara optimal oleh pihak manajemen.

Pada Regresi Linear, interpretasi model dilakukan melalui analisis koefisien regresi. Setiap koefisien menunjukkan arah dan besarnya pengaruh suatu variabel independen terhadap variabel dependen. Koefisien bernilai positif menunjukkan hubungan searah, sedangkan koefisien bernilai negatif menunjukkan hubungan berlawanan. Selain itu, nilai p-value, uji t, uji F, dan koefisien determinasi (R^2) digunakan untuk mengevaluasi signifikansi model dan kontribusi masing-masing variabel terhadap kemampuan prediksi.

Berbeda dengan Regresi Linear, Decision Tree Regression diinterpretasikan melalui struktur pohon keputusan. Setiap percabangan menggambarkan aturan keputusan (*decision rule*) yang digunakan algoritma dalam menghasilkan prediksi. Dengan menelusuri jalur dari root node hingga leaf node, pengguna dapat memahami bagaimana kombinasi variabel tertentu menghasilkan nilai prediksi tertentu. Selain itu, Decision Tree juga menyediakan informasi mengenai feature importance, yaitu tingkat kontribusi setiap variabel dalam pembentukan model.

Sebagai contoh, hasil analisis dapat menunjukkan bahwa variabel harga memiliki tingkat kepentingan sebesar 45%, promosi 30%, musim 15%, dan hari libur 10%. Informasi tersebut memberikan gambaran bahwa harga merupakan faktor dominan yang memengaruhi permintaan usaha. Hasil ini dapat digunakan oleh perusahaan dalam menentukan strategi penetapan harga, penyusunan program promosi, maupun perencanaan persediaan.

Dalam beberapa tahun terakhir, teknik interpretasi model semakin berkembang melalui pendekatan Explainable Artificial Intelligence (XAI), seperti SHAP (SHapley Additive exPlanations) dan LIME (Local

Interpretable Model-agnostic Explanations). Teknik tersebut mampu menjelaskan kontribusi setiap variabel terhadap hasil prediksi secara lebih rinci, bahkan pada model machine learning yang kompleks. Dengan demikian, interpretasi model tidak hanya meningkatkan transparansi, tetapi juga meningkatkan kepercayaan pengguna terhadap sistem prediksi yang dibangun.

BAB 6
EVALUASI MODEL

6.1 Konsep Evaluasi

Evaluasi model merupakan tahapan penting dalam proses pengembangan *machine learning* yang bertujuan untuk mengukur kemampuan suatu model dalam menghasilkan prediksi yang akurat, konsisten, dan dapat digeneralisasikan terhadap data baru. Setelah model dibangun menggunakan algoritma seperti Regresi Linear atau Decision Tree Regression, diperlukan suatu mekanisme untuk mengetahui apakah model tersebut telah memenuhi tujuan penelitian dan layak diterapkan pada kondisi nyata. Tanpa proses evaluasi yang baik, model berisiko memberikan prediksi yang kurang akurat sehingga dapat menyebabkan kesalahan dalam pengambilan keputusan bisnis.

Dalam konteks prediksi permintaan usaha, evaluasi model dilakukan dengan membandingkan nilai hasil prediksi (*predicted value*) terhadap nilai aktual (*actual value*). Selisih antara kedua nilai tersebut disebut sebagai *error* atau galat prediksi. Semakin kecil nilai galat yang dihasilkan, semakin baik kemampuan model dalam melakukan prediksi. Oleh karena itu, berbagai metrik evaluasi dikembangkan untuk mengukur tingkat kesalahan model dari berbagai sudut pandang. Pemilihan metrik evaluasi harus disesuaikan dengan karakteristik permasalahan yang dihadapi. Pada kasus regresi, metrik yang paling banyak digunakan meliputi Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Percentage Error (MAPE), dan Coefficient of Determination

(R^2). Masing-masing metrik memiliki karakteristik yang berbeda sehingga sering digunakan secara bersamaan agar evaluasi model menjadi lebih komprehensif. Misalnya, MAE memberikan gambaran rata-rata kesalahan absolut, sedangkan RMSE memberikan penalti yang lebih besar terhadap kesalahan prediksi yang ekstrem.

Selain mengukur tingkat kesalahan, evaluasi model juga bertujuan mendeteksi kondisi overfitting dan underfitting. Model yang terlalu kompleks biasanya memiliki performa sangat baik pada data pelatihan tetapi buruk pada data pengujian, sedangkan model yang terlalu sederhana tidak mampu menangkap pola data secara optimal. Oleh karena itu, evaluasi tidak hanya dilakukan pada training data, tetapi juga pada testing data melalui teknik seperti K-Fold Cross Validation.

Dalam implementasi menggunakan Python, proses evaluasi dapat dilakukan menggunakan pustaka Scikit-learn, yang menyediakan berbagai fungsi untuk menghitung metrik evaluasi secara otomatis. Hasil evaluasi kemudian menjadi dasar dalam membandingkan performa Regresi Linear dan Decision Tree Regression sehingga peneliti dapat menentukan algoritma yang paling sesuai untuk memprediksi permintaan usaha.

Evaluasi model bukan sekadar tahap akhir dalam pembangunan model, melainkan proses yang memastikan bahwa sistem prediksi memiliki tingkat akurasi, reliabilitas, dan kemampuan generalisasi yang memadai untuk diterapkan dalam pengambilan keputusan berbasis data.

6.2 Mean Absolute Error (MAE)

Mean Absolute Error (MAE) merupakan salah satu metrik evaluasi yang paling sederhana dan paling banyak digunakan untuk mengukur tingkat kesalahan model regresi. MAE menghitung rata-rata nilai

absolut dari selisih antara nilai aktual dengan nilai hasil prediksi. Karena menggunakan nilai absolut, seluruh kesalahan dianggap bernilai positif sehingga tidak terjadi saling meniadakan antara kesalahan positif dan negatif. Karakteristik ini menjadikan MAE mudah dipahami dan diinterpretasikan oleh peneliti maupun praktisi bisnis.

Secara matematis, MAE dirumuskan sebagai:

$$MAE = \frac{1}{n}\sum_{i=1}^{n} | y_i - \hat{y}_i |$$

dengan:

- y_i= nilai aktual,
- $\hat{y}_i$= nilai prediksi,
- n= jumlah data.

Nilai MAE menunjukkan rata-rata besar kesalahan prediksi dalam satuan yang sama dengan data asli. Sebagai contoh, apabila model prediksi permintaan usaha menghasilkan nilai MAE sebesar 12 unit, maka dapat diartikan bahwa rata-rata kesalahan prediksi adalah sekitar 12 produk pada setiap periode pengamatan. Semakin kecil nilai MAE, semakin tinggi tingkat akurasi model.

Salah satu keunggulan MAE adalah kemudahannya dalam interpretasi. Nilai yang dihasilkan langsung menunjukkan besarnya kesalahan rata-rata tanpa melalui proses kuadrat ataupun akar kuadrat sebagaimana pada MSE dan RMSE. Oleh karena itu, MAE sering digunakan ketika peneliti ingin mengetahui besarnya kesalahan prediksi dalam satuan yang mudah dipahami oleh pengguna.

Namun demikian, MAE juga memiliki keterbatasan. Metrik ini memberikan bobot yang sama terhadap seluruh kesalahan sehingga tidak mampu memberikan penalti lebih besar terhadap kesalahan

ekstrem (*outlier*). Dalam beberapa kasus bisnis, kesalahan prediksi yang sangat besar dapat menimbulkan kerugian yang jauh lebih tinggi dibandingkan kesalahan kecil. Oleh sebab itu, MAE umumnya digunakan bersama metrik lain seperti RMSE agar evaluasi menjadi lebih komprehensif.

Pada implementasi menggunakan Scikit-learn, MAE dapat dihitung menggunakan fungsi mean_absolute_error(). Nilai MAE yang rendah menunjukkan bahwa model Regresi Linear maupun Decision Tree Regression mampu menghasilkan prediksi yang mendekati kondisi sebenarnya. Oleh karena itu, MAE menjadi salah satu indikator utama dalam membandingkan performa kedua algoritma pada penelitian prediksi permintaan usaha.

6.3 Mean Squared Error (MSE)

Mean Squared Error (MSE) merupakan metrik evaluasi yang mengukur rata-rata kuadrat selisih antara nilai aktual dan nilai prediksi. Berbeda dengan MAE yang menggunakan nilai absolut, MSE mengkuadratkan setiap kesalahan sehingga kesalahan yang lebih besar akan memperoleh penalti yang lebih tinggi. Karakteristik tersebut membuat MSE sangat sensitif terhadap keberadaan *outlier* dan sering digunakan ketika kesalahan prediksi yang besar harus diminimalkan.

Secara matematis, MSE dirumuskan sebagai:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y_i})^2$$

dengan:

- y_i= nilai aktual,
- $\widehat{y_i}$= nilai prediksi,

- n= jumlah data.

Nilai MSE selalu bernilai positif atau nol. Semakin kecil nilai MSE, semakin baik kualitas model dalam melakukan prediksi. Nilai nol menunjukkan bahwa seluruh hasil prediksi sama persis dengan nilai aktual, meskipun kondisi tersebut hampir tidak pernah terjadi pada data nyata.

Keunggulan utama MSE adalah kemampuannya memberikan penalti lebih besar terhadap kesalahan prediksi yang ekstrem. Sebagai contoh, apabila suatu model melakukan kesalahan prediksi sebesar 50 unit, maka dampaknya terhadap nilai MSE akan jauh lebih besar dibandingkan lima kesalahan masing-masing sebesar 10 unit. Hal ini menjadikan MSE sangat sesuai digunakan pada kasus di mana kesalahan besar dapat menyebabkan kerugian operasional yang signifikan, seperti perencanaan produksi, pengendalian persediaan, atau estimasi permintaan usaha.

Meskipun demikian, MSE memiliki kelemahan karena hasil perhitungannya menggunakan satuan kuadrat sehingga tidak mudah diinterpretasikan secara langsung. Oleh sebab itu, dalam praktiknya MSE sering dikombinasikan dengan RMSE yang mengembalikan nilai kesalahan ke satuan asli data.

Dalam implementasi menggunakan Python, nilai MSE dapat dihitung melalui fungsi mean_squared_error() pada pustaka Scikit-learn. Hasil evaluasi ini sering digunakan sebagai dasar dalam proses optimasi parameter, pemilihan model terbaik, serta perbandingan performa antara Regresi Linear dan Decision Tree Regression.

6.4 Root Mean Squared Error (RMSE)

Root Mean Squared Error (RMSE) merupakan salah satu metrik

evaluasi yang paling banyak digunakan dalam analisis regresi dan machine learning. RMSE diperoleh dengan menghitung akar kuadrat dari Mean Squared Error (MSE) sehingga hasil akhirnya kembali ke satuan yang sama dengan variabel target. Karakteristik ini menjadikan RMSE lebih mudah diinterpretasikan dibandingkan MSE, karena pengguna dapat langsung memahami besarnya kesalahan prediksi dalam satuan data asli, misalnya unit produk, kilogram, liter, atau rupiah.

Secara matematis, RMSE dirumuskan sebagai:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

dengan:

- y_i= nilai aktual,
- $\hat{y}_i$= nilai prediksi,
- n= jumlah data.

RMSE memberikan penalti yang lebih besar terhadap kesalahan prediksi yang ekstrem karena menggunakan kuadrat kesalahan sebelum dihitung akar kuadratnya. Oleh sebab itu, metrik ini sangat sesuai digunakan pada kasus-kasus yang memerlukan tingkat ketelitian tinggi, seperti prediksi permintaan usaha, estimasi produksi, perencanaan stok, maupun analisis keuangan. Semakin kecil nilai RMSE, semakin baik kemampuan model dalam menghasilkan prediksi yang mendekati kondisi sebenarnya.

Sebagai contoh, apabila model Regresi Linear menghasilkan RMSE sebesar 18 unit, sedangkan Decision Tree Regression menghasilkan RMSE sebesar 10 unit, maka dapat disimpulkan bahwa Decision Tree Regression memiliki tingkat kesalahan prediksi yang lebih rendah.

Dengan demikian, model tersebut lebih layak digunakan sebagai dasar dalam penyusunan strategi bisnis.

Salah satu kelebihan RMSE adalah kemampuannya membedakan model yang memiliki kesalahan prediksi besar meskipun nilai MAE kedua model relatif sama. Namun demikian, RMSE juga memiliki kelemahan karena sangat sensitif terhadap *outlier*. Kesalahan prediksi yang sangat besar dapat meningkatkan nilai RMSE secara signifikan sehingga hasil evaluasi menjadi lebih dipengaruhi oleh sejumlah kecil data ekstrem.

Dalam implementasi menggunakan Scikit-learn, RMSE dapat dihitung melalui fungsi mean_squared_error() yang kemudian diikuti dengan proses akar kuadrat menggunakan pustaka NumPy. Saat ini beberapa versi Scikit-learn juga telah menyediakan parameter untuk memperoleh RMSE secara langsung.

Karena mempertimbangkan besarnya kesalahan secara lebih ketat, RMSE menjadi salah satu indikator utama dalam membandingkan performa berbagai model regresi. Dalam penelitian prediksi permintaan usaha, penggunaan RMSE bersama MAE dan MSE memberikan gambaran yang lebih komprehensif mengenai kualitas model yang dibangun.

6.5 Mean Absolute Percentage Error (MAPE)

Mean Absolute Percentage Error (MAPE) merupakan metrik evaluasi yang mengukur rata-rata persentase kesalahan prediksi terhadap nilai aktual. Berbeda dengan MAE dan RMSE yang menghasilkan nilai dalam satuan asli data, MAPE menyajikan tingkat kesalahan dalam bentuk persentase sehingga lebih mudah dipahami oleh pengguna, khususnya dalam konteks bisnis dan manajemen. Oleh karena itu,

MAPE menjadi salah satu ukuran yang paling populer dalam evaluasi model peramalan dan prediksi permintaan.

Secara matematis, MAPE dirumuskan sebagai:

$$MAPE = \frac{100\%}{n} \sum_{i=1}^{n} \left| \frac{y_i - \widehat{y_i}}{y_i} \right|$$

dengan:

- y_i= nilai aktual,
- $\widehat{y_i}$= nilai prediksi,
- n= jumlah data.

Nilai MAPE menunjukkan rata-rata persentase penyimpangan hasil prediksi terhadap data aktual. Sebagai contoh, apabila model menghasilkan MAPE sebesar 6%, maka rata-rata kesalahan prediksi adalah sekitar enam persen dari nilai sebenarnya. Semakin kecil nilai MAPE, semakin tinggi tingkat akurasi model.

Dalam praktik bisnis, terdapat beberapa pedoman umum untuk menginterpretasikan nilai MAPE. Nilai MAPE kurang dari 10% biasanya dikategorikan sebagai model dengan akurasi sangat baik. Nilai antara 10% hingga 20% dianggap memiliki akurasi baik, sedangkan nilai 20% hingga 50% menunjukkan tingkat akurasi yang cukup. Apabila nilai MAPE melebihi 50%, model umumnya dianggap kurang layak digunakan karena menghasilkan tingkat kesalahan yang tinggi.

Meskipun mudah dipahami, MAPE memiliki beberapa keterbatasan. Metrik ini tidak dapat digunakan secara optimal apabila dataset mengandung nilai aktual nol atau sangat mendekati nol karena akan menyebabkan pembagian dengan nol atau menghasilkan persentase kesalahan yang sangat besar. Oleh sebab itu, pada beberapa kasus digunakan metrik alternatif seperti Symmetric Mean Absolute

Percentage Error (SMAPE).

Dalam implementasi menggunakan Python, MAPE dapat dihitung menggunakan fungsi mean_absolute_percentage_error() pada pustaka Scikit-learn. Hasil evaluasi MAPE sangat bermanfaat bagi manajemen karena lebih mudah dikomunikasikan kepada pihak nonteknis. Misalnya, perusahaan dapat menyampaikan bahwa model memiliki tingkat kesalahan prediksi rata-rata hanya lima persen, yang lebih mudah dipahami dibandingkan penyajian nilai MAE atau RMSE.

6.6 R-Squared (R^2)

Coefficient of Determination (R-Squared atau R^2) merupakan metrik evaluasi yang digunakan untuk mengukur kemampuan model regresi dalam menjelaskan variasi data pada variabel dependen. Berbeda dengan MAE, MSE, RMSE, dan MAPE yang mengukur besarnya kesalahan prediksi, R^2 mengukur seberapa besar proporsi variasi data yang berhasil dijelaskan oleh model. Oleh karena itu, R^2 menjadi salah satu indikator utama dalam mengevaluasi kualitas model regresi.

Secara matematis, R^2 dirumuskan sebagai:

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$$

dengan:

- SS_{res}= jumlah kuadrat residual (*Residual Sum of Squares*),
- SS_{tot}= jumlah kuadrat total (*Total Sum of Squares*).

Nilai R^2 berada pada rentang 0 hingga 1, meskipun dalam kondisi tertentu dapat bernilai negatif apabila model memiliki performa yang sangat buruk. Nilai yang semakin mendekati 1 menunjukkan bahwa model mampu menjelaskan sebagian besar variasi data, sedangkan nilai yang mendekati 0 menunjukkan bahwa model hanya mampu

menjelaskan sebagian kecil variasi tersebut.

Sebagai contoh, apabila model Regresi Linear menghasilkan nilai R^2 = 0,89, maka dapat diartikan bahwa sekitar 89% variasi permintaan usaha dapat dijelaskan oleh variabel-variabel independen yang digunakan dalam model, sedangkan sisanya sebesar 11% dipengaruhi oleh faktor lain di luar model.

Meskipun R^2 memberikan informasi mengenai kemampuan model dalam menjelaskan variasi data, metrik ini tidak menunjukkan besarnya kesalahan prediksi. Oleh karena itu, R^2 tidak boleh digunakan sebagai satu-satunya indikator evaluasi. Model dengan nilai R^2 tinggi belum tentu memiliki nilai RMSE atau MAE yang rendah. Sebaliknya, model dengan R^2 sedikit lebih rendah dapat menghasilkan kesalahan prediksi yang lebih kecil.

Pada implementasi menggunakan Scikit-learn, nilai R^2 dapat dihitung menggunakan fungsi r2_score(). Dalam penelitian prediksi permintaan usaha, nilai R^2 biasanya digunakan bersama MAE, MSE, RMSE, dan MAPE agar evaluasi model menjadi lebih lengkap dan objektif.

DAFTAR PUSTAKA

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (2017). *Classification and Regression Trees*. Routledge.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer.

Géron, A. (2023). *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow* (3rd ed.). O'Reilly Media.

Han, J., Kamber, M., & Pei, J. (2022). *Data Mining: Concepts and Techniques* (4th ed.). Morgan Kaufmann.

Hastie, T., Tibshirani, R., & Friedman, J. (2021). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (2nd ed., corrected). Springer.

Hyndman, R. J., & Athanasopoulos, G. (2021). *Forecasting: Principles and Practice* (3rd ed.). OTexts.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2023). *An Introduction to Statistical Learning* (2nd ed.). Springer.

Nofriansyah, D., & Defit, S. (2017). *Multicriteria Decision Making (MCDM)*. Deepublish.

Prasetyo, E. (2014). *Data Mining: Mengolah Data Menjadi Informasi Menggunakan MATLAB*. Andi Offset.

Santosa, B. (2018). *Data Mining: Teknik Pemanfaatan Data untuk Keperluan Bisnis*. Graha Ilmu.

Santoso, S. (2023). *Menguasai SPSS 29*. Elex Media Komputindo.

Sugiyono. (2023). *Metode Penelitian Kuantitatif, Kualitatif, dan R&D* (Edisi 3). Alfabeta.

Suyanto. (2019). *Data Mining untuk Klasifikasi dan Klasterisasi Data*. Informatika.

Turban, E., Sharda, R., & Delen, D. (Edisi Indonesia). (2021). *Business*

Intelligence dan Analytics. Salemba Empat.

Widodo, P. P., & Handayanto, R. T. (2020). *Penerapan Data Mining dengan Python*. Informatika.

BIOGRAFI PENULIS

Mutiara Choirunnisah, Lahir di Kisaran,Kabupaten Asahan 02 Juni 2004. Menempuh Pendidikan Sekolah Dasar di SDN 12195 Kampung Padang,Kemudian Sekolah Menengah Pertama di MTs Al-Ittihad Aek Nabara, Selanjutnya Sekolah Menengah Atas di SMKN 1 PANGKATAN. Tahun 2022 melanjutkan kuliah di Fakultas Sains dan Teknologi Mengambil jurusan Sistem Informasi(SI) di Univertsitas Labuhanbatu sejak Tahun 2022-2026.

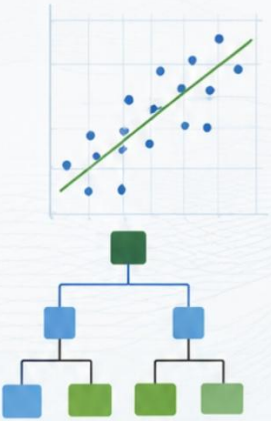
Angga Putra Juledi, Lahir di Kota Padang pada tanggal 19 Juli 1994. Dalam menempuh Pendidikan dimulai dari Sekolah Dasar SDN 19 Padang tamat tahun 2006, SMPN 3 Padang tamat pada tahun 2009, dan di SMA Pertiwi 2 Padang tamat pada tahun 2012. Lalu melanjutkan ke pendidikan perguruan tinggi swasta yaitu S1 (Sarjana) Universitas Putra Indonesia "YPTK" Padang lulus pada tahun 2018 dengan jurusan Sistem Informasi, Dan melanjutkan Program Pascasarjana (S2) di Universitas Putra Indonesia "YPTK" Padang pada tahun 2019 Program Studi Teknik Informatika Konsentrasi Sistem Informasi. Saya mengabdikan diri sebagai salah satu Dosen di bidang Ilmu Komputer pada Fakultas Sains Dan Teknologi dengan Program Studi Sistem Informasi di Universitas LabuhanBatu dan menjadi dosen tetap pada tahun 2020 pada kampus tersebut.

Iwan Purnama, Adalah Pendiri dan pendidik LKP Ibay Komputer, akademisi bidang komputer, serta konsultan digital yang berdedikasi pada penguatan kompetensi generasi muda dan pemberdayaan ekonomi berbasis teknologi. Beliau saat ini menjabat sebagai Dekan Fakultas

Sains dan Teknologi Universitas Labuhanbatu, dan Dosen pada Universitas Labuhanbatu, dengan fokus riset pada integrasi AI dalam pendidikan dan Pengembangan sistem informasi untuk UMKM. Beliau juga aktif pada Organiasi Ketua Forum Komunikasi Dosen Wilayah Sumatera Utara, Pengurus APTIKOM Wilayah Sumatera Utara, Himpunan Dosen Gemilang Indonesia, Penerbit Buku dan Pengelola Jurnal, Pembina Yayasan Labuhanbatu Berbagi Gemilang dan Aktivitas Lainnya.

Marnis Nasution, Lahir di Bengkulu 30 Maret 1990.Selama sekolah dasar sampai menengah di tempuh di kota Bengkulu. Melanjutkan Pendidikan tinggi strata-1 dan strata-2 di Universitas Putra Indonesia "YPTK" Padang dari tahun 2008-2024 dengan jurusan Sistem Informasi. Saat ini aktif menjadi Dosen di Universitas Labuhanbatu,Sumatera Utara

DECISION TREE REGRESSION DAN REGRESI LINEAR



Teori, Metode, dan Implementasi dalam Prediksi Permintaan Usaha

SINOPSIS

Perkembangan teknologi informasi dan kecerdasan buatan telah mendorong pemanfaatan analisis data sebagai dasar pengambilan keputusan di berbagai bidang usaha. Kemampuan memprediksi permintaan secara akurat menjadi salah satu faktor penting dalam meningkatkan efisiensi operasional, mengoptimalkan persediaan, mengurangi biaya, serta meningkatkan kepuasan pelanggan. Oleh karena itu, pemahaman terhadap metode analisis prediktif menjadi kebutuhan yang semakin relevan bagi mahasiswa, akademisi, peneliti, maupun praktisi bisnis.

Buku Decision Tree Regression dan Regresi Linear: Teori, Metode, dan Implementasi dalam Prediksi Permintaan Usaha membahas secara komprehensif konsep, teori, dan implementasi dua metode prediksi yang banyak digunakan dalam statistika dan machine learning, yaitu Regresi Linear dan Decision Tree Regression. Pembahasan disusun secara sistematis, dimulai dari konsep dasar analisis data, regresi linear, dan decision tree, hingga tahapan persiapan data, pembangunan model, evaluasi performa menggunakan berbagai metrik seperti MAE, MSE, RMSE, MAPE, dan R^2, serta analisis hasil prediksi dalam konteks dunia usaha.

ISBN 978-634-05-3363-7 (PDF)
9 786340 533637

PENERBIT YAYASAN MMI