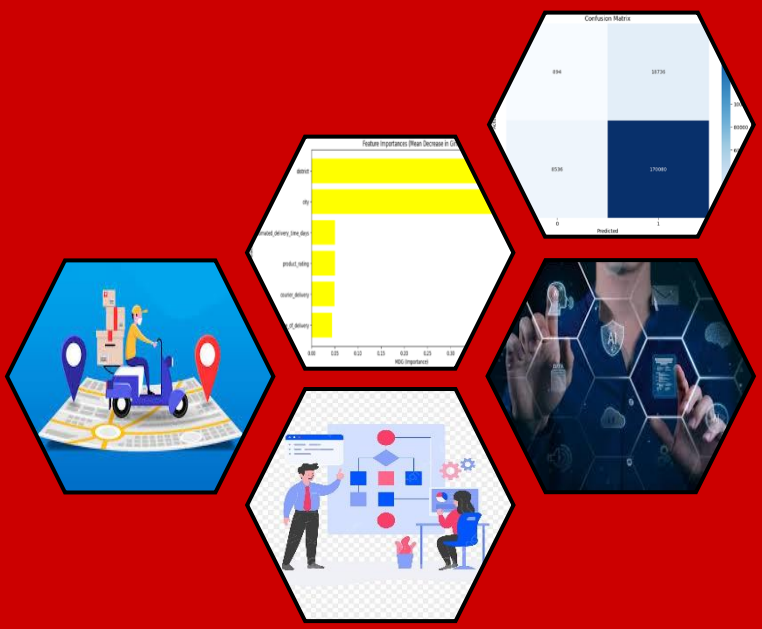
ANALISIS PERBANDINGAN KINERJA ALGORITMA *GRADIENT BOOSTING* DAN *BOOTSTRAP AGGREGATING (BAGGING)* UNTUK KLASIFIKASI KETEPATAN WAKTU SISTEM PENGIRIMAN PRODUK *E-COMMERCE*



ANNISA
H011201025

PROGRAM STUDI MATEMATIKA
DEPARTEMEN MATEMATIKA
FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
UNIVERSITAS HASANUDDIN
MAKASSAR
2024

ANALISIS PERBANDINGAN KINERJA ALGORITMA *GRADIENT BOOSTING* DAN *BOOTSTRAP AGGREGATING (BAGGING)* UNTUK KLASIFIKASI KETEPATAN WAKTU SISTEM PENGIRIMAN PRODUK *E-COMMERCE*

ANNISA
H011201025

PROGRAM STUDI MATEMATIKA
DEPARTEMEN MATEMATIKA
FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
UNIVERSITAS HASANUDDIN
MAKASSAR
2024

Analisis Perbandingan Kinerja Algoritma *Gradient Boosting* Dan *Bootstrap Aggregating (Bagging)* Untuk Klasifikasi Ketepatan Waktu Sistem Pengiriman Produk *E-Commerce*

ANNISA

H011201025

Skripsi

sebagai salah satu syarat untuk mencapai gelar sarjana

Program Studi Matematika

pada

PROGRAM STUDI MATEMATIKA DEPARTEMEN MATEMATIKA

FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM UNIVERSITAS HASANUDDIN

MAKASSAR

2024

SKRIPSI

Analisis Perbandingan Kinerja Algoritma *Gradient Boosting* dan *Bootstrap Aggregating (Bagging)* Untuk Klasifikasi Ketepatan Waktu Sistem Pengiriman Produk *E-Commerce*

ANNISA

H011201025

Skripsi,

telah dipertahankan di depan Panitia Ujian Sarjana pada 2024

dan dinyatakan telah memenuhi syarat kelulusan

pada

Program Studi Matematika

Departemen Matematika

Fakultas Matematika dan Ilmu Pengetahuan Alam

Universitas Hasanuddin

Mengesahkan:
Pembimbing Tugas Akhir,

Dr. Khaeruddin, M.Sc.
NIP.19650914199103100 3

Mengetahui:
Ketua Program Studi,

Dr. Firman, S.Si., M.Si.
NIP. 196804292002121001

PERNYATAAN KEASLIAN SKRIPSI DAN PELIMPAHAN HAK CIPTA

Dengan ini saya menyatakan bahwa, skripsi berjudul "Analisis Perbandingan Kinerja Algoritma *Gradient Boosting* Dan *Bootstrap Aggregating (Bagging)* Untuk Klasifikasi Ketepatan Waktu Sistem Pengiriman Produk *E-Commerce*" adalah benar karya saya dengan arahan dari Dr. Khaeruddin, M.Sc. sebagai pembimbing tugas akhir. Karya ilmiah ini belum diajukan dan tidak sedang diajukan dalam bentuk apa pun kepada perguruan tinggi mana pun. Sumber informasi yang berasal atau dikutip dari karya yang diterbitkan maupun tidak diterbitkan dari penulis lain telah disebutkan dalam teks dan dicantumkan dalam Daftar Pustaka skripsi ini. Apabila di kemudian hari terbukti atau dapat dibuktikan bahwa sebagian atau keseluruhan skripsi ini adalah karya orang lain, maka saya bersedia menerima sanksi atas perbuatan tersebut berdasarkan aturan yang berlaku.

Dengan ini saya melimpahkan hak cipta (hak ekonomis) dari karya tulis saya berupa skripsi ini kepada Universitas Hasanuddin.

Makassar, 11 Desember 2024

Annisa

H011201025

KATA PENGANTAR

Puji syukur kehadirat penulis haturkan kepada Allah SWT Alhamdulillah atas segala limpahan Rahmat dan hidayah-Nya sehingga penulis dapat menyelesaikan skripsi yang berjudul **"Analisis Perbandingan Kinerja Algoritma *Gradient Boosting* dan *Bootstrap Aggregating (BAGGING)* Untuk Klasifikasi Ketepatan Waktu Sistem Pengiriman Produk *E-Commerce*" .** Shalawat dan salam kepada Rasulullah Muhammad Shalallahu 'Alaihi Wasallam yang telah menjadi suri tauladan terbaik bagi penulis.

Skripsi ini dibuat sebagai salah satu syarat untuk memenuhi salah satu syarat dalam menempuh ujian Tingkat sarjana Strata 1 (S1) pada Program Studi Matematika, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Hasanuddin. Penulis menyadari dalam penyusunan skripsi ini tidak terlepas dari doa, dukungan, bantuan, bimbingan, dan semangat yang diberikan dari berbagai pihak. Untuk itu, dalam kesempatan ini penulis mengucapkan banyak terima kasih kepada :

1. Cinta pertama dan panutanku, Ayahanda Khaeruddin, Beliau memang tidak sempat merasakan Pendidikan sampai bangku perkuliahan, namun beliau mampu mendidik penulis, memotivasi, memberikan dukungan hingga penulis mampu menyelesaikan studinya sampai sarjana.
2. Pintu surgaku, Ibunda Sukarni, beliau sangat berperan penting dalam meyelesaikan program studi penulis, beliau memang tidak sempat menyelesaikan pendidikannya sampai sarjana, tapi semangat motivasi serta do'a yang selalu beliau berikan hingga penulis mampu menyelesaikan studinya sampai sarjana.
3. Bapak Dr. Khaeruddin, M.Sc. selaku dosen pembimbing, atas segala ilmu, saran, serta bimbingan kepada penulis selama proses penyusunan skripsi ini.
4. Bapak Prof. Dr. Budi Nurwahyu, MS. Selaku dosen penguji I dan Ibu Naimah Aris, S. Si., M. Math. Selaku dosen penguji II, yang senantiasa memberikan kritik dan saran yang membangun dalam menyelesaikan skripsi ini.
5. Seluruh Bapak dan Ibu Dosen Program Studi Matematika yang telah memberikan ilmu-ilmu yang bermanfaat.
6. Kakak saya Sukandi beserta istrinya dan adik saya M. Rifky, terimakasih sudah memberikan motivasi serta dukungan kepada penulis.
7. Sepupu saya tercinta Endang Hardianti, S.Kep,.Ners. beserta keluarga, terima kasih sudah menjadi sepupu, saudara Perempuan, teman yang selalu mendengar segala keluh kesah penulis serta memberikan dukungan sehingga penulis bisa menyelesaikan skripsi ini.
8. Sahabat saya Mega Putri Lestari, terimakasih selalu memberikan semangat dan dukangan serta selalu ada menemani penulis dalam proses penyelesaian skripsi ini.
9. Seseorang yang tak kalah penting kehadirannya Andi Adrian, terimakasih telah menjadi bagian dari perjalanan hidup penulis. Terimakasih telah mendukung, menghibur, mendengarkan keluh kesah, dan memberikan semangat untuk pantang menyerah.
10. Teman-teman KKNT Gelombang 110, terimakasih sudah memberi warna dalam perjalanan perkuliahan penulis.
11. Keluarga besar ANDES, yang sudah menjadi teman voli di masa perkuliah hingga seterusnya bagi penulis.

12. Semua pihak yang tidak dapat disebutkan satu persatu yang telah banyak membantu memberikan pemikiran demi kelancaran dan keberhasilan penyusunan skripsi ini.
13. Terakhir, terimakasih kepada diri saya sendiri Annisa, karena telah mampu berusaha keras dan berjuang sejauh ini. Terimakasih sudah menepikan ego dan memilih untuk Kembali bangkit dan menyelesaikan semua ini. Terimakasih telah mengendalikan diri dari berbagai tekanan diluar keadaan dan tidak pernah mau memutuskan untuk menyerah.

Makassar, 28 November 2024

Annisa

ABSTRAK

ANNISA. **Analisis Perbandingan Kinerja Algoritma *Gradient Boosting* Dan *Bootstrap Aggregating (Bagging)* Untuk Klasifikasi Ketepatan Waktu Sistem Pengiriman Produk *E-Commerce*** (dibimbing oleh Dr. Khaeruddin, M. Sc.)

E-commerce merupakan singkatan dari *electronic commerce*, yang berarti suatu aktivitas yang memanfatkan jaringan internet untuk proses jual beli barang atau jasa. Kepuasan konsumen mempunyai peranan yang cukup penting bagi pengguna *e-commerce*. Logistik atau jasa pengiriman memegang peran penting dalam pengantaran produk kepada pelanggan. Sistem pengiriman barang (logistik) menjadi faktor menarik yang dapat mendukung atau menghambat kemajuan *e-commerce*, terutama dari segi kualitas layanan terhadap konsumen. Penelitian ini bertujuan untuk membandingkan algoritma yang paling akurat dalam penentuan klasifikasi ketepatan waktu pengiriman produk *e-commerce* menggunakan algoritma *Gradient Boosting dan Bagging,* serta mengetahui variabel-variabel yang paling berpengaruh. Data yang digunakan dalam penelitian ini adalah data pengiriman *e-commerce* di Indonesia mulai pertengahan tahun 2022 sampai dengan pertengahan tahun 2023 dengan jumlah data sebanyak 991230. Penelitian ini menggunakan dua algoritma klasifikasi, yaitu algoritma *Gradient Boosting* dan *Boostrap Agregatting (Bagging)* serta dilakukan penangan data tidak seimbang dengan *Synthetic Minority Oversampling Technique (SMOTE)*. Berdasarkan hasil analisis prediksi diperoleh bahwa setelah dilakukan *SMOTE*, metode *Bagging* memiliki kinerja lebih baik dibandingkan dengan metode *Gradient Boosting* dalam meprediksi ketepatan waktu pengiriman produk *e-commerce* di Indonesia dengan nilai akurasi sebesar 85% dan mengalami peningkatan setelah dilakukan *cross-validation* dengan 10 fold sebesar 91%. Variabel Kota dan kecamatan tujuan pengiriman memiliki nilai *features importance* paling besar dibanding variabel lain dalam setiap model.

Kata Kunci : *E-commerce,* Klasifikasi, *Gradient Boosting, Bagging, SMOTE, Features Importance*

ABSTRACT

ANNISA. **Comparative Analysis of Gradient Boosting and Bootstrap Aggregating (Bagging) Algorithm Performance for Timeliness Classification of E-Commerce Product Delivery System** (supervised by Dr. Khaeruddin, M. Sc.).

E-commerce stands for electronic commerce, which means an activity that utilizes the internet network for the process of buying and selling goods or services. Customer satisfaction plays an important role for e-commerce users. Logistics or delivery services play an important role in delivering products to customers. The delivery system (logistics) is an interesting factor that can support or hinder the progress of e-commerce, especially in terms of service quality to consumers. This research aims to compare the most accurate algorithm in determining the classification of timeliness of e-commerce product delivery using the Gradient Boosting and Bagging algorithms, as well as knowing the most influential variables. The data used in this research is e-commerce delivery data in Indonesia from mid-2022 to mid-2023 with a total data of 991230. This research uses two classification algorithms, namely the Gradient Boosting and Boostrap Aggregatting (Bagging) algorithms and unbalanced data handling with the Synthetic Minority Oversampling Technique (SMOTE). Based on the results of prediction analysis, it is obtained that after SMOTE, the Bagging method has better performance than the Gradient Boosting method in predicting the timeliness of e-commerce product delivery in Indonesia with an accuracy value of 85% and has increased after cross-validation with 10 folds of 91%. City and sub-district delivery destination variables have the greatest feature importance value compared to other variables in each model.

Keywords: E-commerce, Classification, Gradient Boosting, Bagging, SMOTE, Features Importance.

DAFTAR ISI

HALAMAN JUDUL i

PERNYATAAN PENGAJUAN ii

HALAMAN PENGESAHAN iii

PERNYATAAN KEASLIAN SKRIPSI iv

KATA PENGANTAR v

ABSTRAK vii

ABSTRACT viii

DAFTAR ISI ix

DAFTAR GAMBAR xi

DAFTAR TABEL xii

BAB I PENDAHULUAN 1

PENDAHULUAN 1

1. 1 Latar Belakang 1

1. 2 Rumusan Masalah 3

1. 3 Tujuan Penelitian 3

1. 4 Batasan Masalah 3

1. 5 Manfaat penelitian 3

1. 6 Landasan Teori 3

BAB II METODE PENELITIAN 19

2. 1 Data dan Variabel Penelitian 19

2. 2 Langkah-langkah Penelitian 21

BAB III HASIL DAN PEMBAHASAN 24

3. 1 Preprocessing Data 24

3. 2 Pembagian Data 26

3. 3 Sampling Data 27

3. 4 Analisis Klasifikasi *Gradient Boosting* 27

3. 5 Analisis Klasifikasi *Boostrap Aggregating (Bagging)* ……35

3. 6 Perbandingan Kinerja Metode Klasifikasi ..... ……44

BAB IV PENUTUP ……46

4. 1 Kesimpulan ..... ……46

4. 2 Saran ……..... ……46

DAFTAR PUSTAKA .....……..47

LAMPIRAN ……..52

DAFTAR GAMBAR

Gambar 1. 1 *Training* Dataset dan *Testing* Dataset ……5
Gambar 1. 2 Ilustrasi Target Variabel dan Fitur Variabel ……6
Gambar 1. 3 Tipe Model klasifikasi ……7
Gambar 1. 4 Ilustrasi Algoritma *SMOTE* ……8
Gambar 1. 5 Ilustrasi *Decision Tree* ……...9
Gambar 1. 6 Pohon Klasifikasi CART ……11
Gambar 1. 7 Ilustrasi Algoritma *Gradient Boosting* ……12
Gambar 1. 8 Prinsip Kerja Algoritma *Bagging* ……14
Gambar 1. 9 Konstruksi Klasifikasi Bagging ……16
Gambar 1. 10 Ilustrasi dari *K-Fold Cross-Validation* dengan K=10 ……...17
Gambar 2. 1 *Flowchart* Penelitian ……22
Gambar 3. 1 Cek Nilai yang Hilang Pada Setiap Data ……...25
Gambar 3. 2 Cek Data Duplikat ……25
Gambar 3. 3 Data Setelah Transformasi Ke Numerik ……...27
Gambar 3. 4 Confusion Matrix Proporsi Data 90:10 Algoritma *Gradient Boosting*…….29
Gambar 3. 5 Confusion Matrix Proporsi Data 80:20 Algoritma *Gradient Boosting*…….31
Gambar 3. 6 Confusion Matrix Proporsi Data 70:30 Algoritma *Gradient Boosting*…….33
Gambar 3. 7 *Features Importance* Algoritma *Gradient Boosting* ……...35
Gambar 3. 8 *10-Fold Cross-Validation* Algoritma *Gradient Boosting* ……36
Gambar 3. 9 Confusion Matrix Proporsi Data 90:10 Algoritma Bagging……..…....37
Gambar 3. 10 Confusion Matrix Proporsi Data 80:20 Algoritma Bagging……40
Gambar 3. 11 Confusion Matrix Proporsi Data 70:30 Algoritma Bagging……...41
Gambar 3. 12 *Features Importance* Algoritma *Bagging* ……....43
Gambar 3. 13 *10-Fold Cross-Validation* Algoritma *Bagging* ……...44

DAFTAR TABEL

Tabel 1. 1 *Confusion Matrix* ……18
Tabel 2. 1 Struktur Data yang Digunakan ……21
Tabel 3. 1 Dataset Setelah Seleksi Data ……..26
Tabel 3. 2 Proporsi Data Latih dan Data Uji……..28
Tabel 3. 3 Data Sebelum dan Sesudah di *SMOTE* ……28
Tabel 3. 4 Parameter Algoritma *Gradient Boosting* ……..……29
Tabel 3. 5 Hasil Evaluasi *Confusion Matrix* Gradien Boosting Proporsi Data 90:10 …31
Tabel 3. 6 Hasil Evaluasi *Confusion Matrix* Gradien Boosting Proporsi Data 80:20 …33
Tabel 3. 7 Hasil Evaluasi *Confusion Matrix* Gradien Boosting Proporsi Data 70:30 …35
Tabel 3. 8 Nilai *Features Importance* Algoritma *Gradient Boosting* ……..……36
Tabel 3. 9 Parameter Algoritma *Bagging* ……..37
Tabel 3. 10 Hasil Evaluasi *Confusion Matrix* Bagging Proporsi Data 90:10 ……..39
Tabel 3. 11 Hasil Evaluasi *Confusion Matrix* Bagging Proporsi Data 80:20 ……..41
Tabel 3. 12 Hasil Evaluasi *Confusion Matrix* Bagging Proporsi Data 70:30 ……..43
Tabel 3. 13 Nilai *Features Importance* Algoritma *Bagging* ……...44
Tabel 3. 14 Perbandingan Kinerja Metode Klasifikasi ……...4

BAB I
PENDAHULUAN

1.1 Latar Belakang

Pesatnya kemajuan teknologi dan informasi di era sekarang ini menjadi salah faktor utama yang mendorong lahirnya *e-commerce*. *E-commerce* merupakan singkatan dari *electronic commerce*, yang berarti suatu aktivitas yang memanfatkan jaringan internet untuk proses jual beli barang atau jasa. Definisi *electronic commerce (e-commerce)* menurut Laudon dan Laudon (2009) adalah suatu proses membeli dan menjual produk-produk secara elektronik oleh konsumen dan dari perusahaan ke perusahaan dengan komputer sebagai perantara transaksi bisnis. Secara singkat, sejarah dari *e-commerce* bermula di awal tahun 1970an, dengan adanya inovasi semacam *electronic fund transfer* (EFT) (Widagdo, 2016)**.** Kemunculan e-*commerce* ini membuat masyarakat menjadi semakin mudah untuk membeli suatu produk. *E-commerce* memiliki peran penting dalam mendorong pertumbuhan ekonomi. Keberhasilan ini perlu terus didukung dengan peningkatan kinerja *e-commerce* dalam memberikan kepuasan, agar masyarakat tetap setia. Kesetian pelanggan ini secara langsung dapat dilakukan dengan meningkatkan kemudahan penggunaan, daya tarik, dan keamanan/ jaminan dari websites (Kassim dan Abdullah, 2010). Ruang lingkup *e-commerce* terdiri dari delapan yaitu teknologi, marketing *"new consumer processes", economic, electronic linkage, information value adding, market-making, service infrastructure, dan legal privacy public police* (Indrajit, 2002). Berkaitan dengan *service infrastructure*, salah satu yang menjadi bahasan saat ini adalah mengenai logistik.

Logistik atau jasa pengiriman memegang peran penting dalam pengantaran produk kepada pelanggan. Sistem pengiriman barang (logistik) menjadi faktor menarik yang dapat mendukung atau menghambat kemajuan *e-commerce*, terutama dari segi kualitas layanan terhadap konsumen. Layanan logistik menjadi sulit dikendalikan mengingat Indonesia adalah negara kepulauan. Menurut Aulia E. Marianto, Ketua Asosiasi *E-commerce* Indonesia (idEA), menyatakan pada tahun 2017 bahwa tantangan terbesar *e-commerce* di Indonesia adalah rendahnya ketepatan waktu pengiriman (cnnindonesia.com).

Data mining merupakan suatu rangkaian proses untuk menggali nilai tambah berupa informasi yang selama ini tidak diketahui secara manual dari suatu basis data dengan melakukan proses ekstraksi dan menggali pola penting dari data yang ada. Pada data mining, data yang berukuran besar diolah dengan menggunakan teknik-teknik tertentu untuk mendapatkan informasi baru mengenai data tersebut. Salah satu teknik yang biasa digunakan dalam data mining adalah klasifikasi (Azhari et al., 2021).

Klasifikasi adalah proses menemukan model atau fungsi yang menggambarkan dan membedakan kelas atau konsep data. Dalam konteks data pengiriman produk *e-commerce*, klasifikasi dapat digunakan untuk memprediksi apakah suatu pengiriman akan tiba tepat waktu *(ontime)* atau tidak tepat waktu *(delayed)* berdasarkan berbagai informasi yang tersedia. Klasifikasi juga merupakan proses pembelajaran sebuah fungsi atau model terhadap sekumpulan data latih, sehingga model tersebut dapat digunakan untuk memprediksi klasifikasi dari data uji (Ordonez et al., 2018).

Ensemble Learning adalah sebuah metode dalam *machine learning* yang melibatkan pelatihan beberapa model *(base-models)* untuk menyelesaikan masalah

yang sama dan menggabungkan hasilnya guna mencapai performa yang lebih baik. Dalam penelitian ini, mengusulkan model klasifikasi untuk prediksi klasifikasi ketepatan waktu pengriman produk *e-commerce*. Penelitian ini membandingkan dua algoritma klasifikasi yaitu *Gradient Boosting* dan *Bagging* .*Gradien Boosting* merupakan algoritma *machine learning* yang menggunakan teknin *ensamble* dari *decision tree,* algoritma ini mampu menyelesaikan persoalan klasifikasi dan prediksi data (Ismanto & Novalia, 2021). Sedangkan Bagging atau *Bootstrap Aggregation* adalah metode *ensemble* yang dapat meningkatkan kinerja klasifikasi dengan mengombinasikan beberapa model klasifikasi yang dilatih secara acak pada dataset pelatihan, sehingga dapat mengurangi variasi dan mencegah *overfitting* (Nugroho & Religia, 2021). Kedua metode ini sama-sama kuat dalam meningkatkan akurasi prediksi pada data yang kompleks, namun memiliki pendekatan yang berbeda. Gradient Boosting berfokus pada pengurangan bias dengan melatih model secara bertahap, di mana setiap model memperbaiki kesalahan model sebelumnya, menghasilkan akurasi tinggi tetapi rentan terhadap overfitting jika tidak diatur dengan baik. Di sisi lain, Bagging dapat mengurangi variansi dengan membuat beberapa model pohon yang digabungkan sehingga lebih stabil dan tahan terhadap overfitting, cocok untuk dataset besar.

Pengembangan algoritma *Gradien Boosting* dan *Bagging* merupakan algoritma yang banyak digunakan dalam menangani beberapa kasus klasifikasi dan prediksi. Pada tahun 2018 Rimbun Siringoringo dan Indra Kelana Jaya melakukan penelitian yang berjudul "*Ensemble Learning* Dengan Metode *Smote Bagging* Pada Klasifikasi Data Tidak Seimbang", Pada penelitian tersebut, teknik *ensemble* metode *SMOTE Bagging* diterapkan untuk mengklasifikasi 11 dataset tidak seimbang. Sebagai pembanding, dipilih tiga jenis algoritma klasifikasi yang konvensional yaitu *SVM, k-NN* dan *C4.5*. Dengan skema 5 *cross fold validation* diperoleh hasil bahwa nilai *AUC SMOTE Bagging* lebih tinggi pada 10 dataset. Nilai rerata *AUC* terendah sampai tertinggi diperoleh oleh algoritma *SVM, k-NN, C4.5* dan *SMOTE Bagging* dengan nilai masing- masing 0.638, 0.742, 0.770 dan 0.895. Pada tahun 2022 Muhamad Sopiyan, Fauziah, dan Yunan Fauzi Wijaya melakukan penelitian tentang deteksi penipuan pada kartu kredit, salah *satu* algoritma yang digunakan adalah *Gradient Boosting* yang menghasilkan nilai akurasi dengan presentase data latih 98,86% dan data uji 98,83%, serta evaluasi skor AUC sebagai hasil pengujian algoritma adalah 0,9994. Kemudian pada tahun 2023 Rofik, Dwika Ananda Agustina Pertiwi, dan Much Aziz Muslim melakukan penelitian yang berjudul "Peningkatan akurasi *gradient boosting* dalam prediksi peringkat aplikasi di google playstore" dengan tujuan untuk mengoptimalkan akurasi dalam memprediksi peringkat aplikasi di Google Playstore menggunakan algoritma *Gradient Boosting* , dengan menerapkan algoritma tersebut penelitian tersebut berhasil mencapai akurasi sebesar 92.69%, dengan MAE 0.311, RMSE 0.467, dan R-square 0.144 .

Berdasarkan penelitian terdahulu yang telah dibahas, penulis tertarik untuk menggunakan metode yang sama dengan membandingkan kedua metode tersebut dengan dataset yang berbeda yaitu dataset pengiriman produk *E-commerce*. Dengan demikian penulis memilih **"Analisi Perbandingan Kinerja Algoritma *Gradien Boosting* dan *Bootstrap Aggregating (Bagging)* Untuk Klasifikasi Ketepatan Waktu Sistem Pengiriman Produk *E-Commerce*"** sebagai judul dari penelitian ini.

1.2 Rumusan Masalah

Berdasarkan latar belakang yang telah di uraikan maka rumusan masalah penelitian ini adalah :

1. Bagamana hasil yang diperoleh dari perbandingan algoritma *Gradient Boosting dan Bagging* pada klasifikasi ketepatan waktu sistem pengiriman produk *E-commerce* ?
2. Bagaimana kinerja algoritma *Gradient Boosting* dan *Bagging* dalam hal akurasi, presisi, recall, dan F1-score pada klasifikasi ketepatan waktu pengiriman produk di E-commerce?
3. Bagaimana interpretasi terhadap variabel yang paling berpengaruh terhadap hasil klasifikasi ketepatan waktu sistem pengiriman produk *E-commerce* ?

1.3 Tujuan Penelitian

Tujuan dari penelitian ini adalah untuk menganalisis dan membandingkan algoritma Gradient Boosting dan Bagging dalam memprediksi ketepatan waktu pengiriman produk e-commerce. Serta mengidentifikasi variabel yang berpengaruh signifikan terhadap ketepatan waktu pengiriman produk e-commerce.

1.4 Batasan Masalah

Batasan masalah yang dapat diambil dari penelitian ini adalah sebagai berikut :

1. Sumber data untuk penelitian ini menggunakan data Pengiriman *e-commerce* pertengahan tahun 2022 sampai dengan pertengahan tahun 2023 yang diperoleh dari Kaggle (https://www.kaggle.com/datasets/adisaputra10/e-commerce-shipping-data).
2. *Machine Learning* yang digunakan dalam penelitian ini adalah *Gradient Boosting* dan *Bagging*.
3. Variabel yang digunakan terdiri dari 7 variabel dimana 6 diantaranya adalah variabel fitur dan satu variabel target.
4. Analisis klasifikasi dilakukan menggunakan metode *Gradient Boosting* dan *Bagging* dengan bantuan Phyton.

1.5 Manfaat Penelitian

Manfaat penelitian ini diharapkan dapat memberikan gambaran kepada masyarakat terkait faktor-faktor apa saja yang dapat mempengaruhi produk di kirim tepat waktu dan tidak tepat waktu pada pengiriman produk *e-commerce* serta memberikan pengetahuan mengenai implementasi algoritma klasifikasi *Gradient Boosting* dan *Bagging*.

1.6 Landasan Teori

1. 6. 1 *E-Commerce*

Saat ini, penjualan online menjadi salah satu langkah yang diambil oleh setiap perusahaan untuk meningkatkan penjualan mereka dengan memanfaatkan perangkat Elektronik seperti *gadget* atau komputer. *E-commerce* didefinisikan sebagai proses pembelian, penjualan, mentransfer atau bertukar produk, jasa atau informasi melalui jaringan komputer di Internet (Pradana, 2015). Penjualan online atau *e-commerce* adalah suatu aplikasi dan proses bisnis yang menghubungkan toko dan konsumen

melalui transaksi elektronik dan dapat membantu pihak toko dalam pemasaran produk yang maksimal (Susandi & Sukisno, 2017).

E-commerce semakin berkembang pesat seiring dengan kemajuan teknologi informasi dan komunikasi. Hal ini menyebabkan peningkatan konsumsi barang, dan sebagai akibatnya, meningkat pula permintaan terhadap layanan pengiriman. Faktor paling penting dalam mempertahankan keberlanjutan layanan pengiriman adalah kualitas pelayanan. Terdapat beberapa hal yang mempengaruhi kualitas pelayanan jasa pengiriman salah satu di antaranya yaitu ketepatan pengiriman. Ketepatan pengiriman adalah ketepatan waktu pengiriman dan jaminan barang pesanan konsumen selamat hingga sampai tujuan menjadi ujung tombak perusahan agar tetap dapat memperoleh kepercayaan dan kesetiaan konsumen (Zahra et al., 2022).

Menurut Pujawan (2017) sebagaimana dikutip dalam Aminah dkk, indikator ketepatan waktu pengiriman ada 3 :

1. Transportasi yang digunakan
   Transportasi yang digunakan adalah kendaraan yang digunakan oleh pengirim untuk mengirim barang yang dipesan pelanggan untuk sampai ke penerima.
2. Estimasi barang sampai ke penerima.
   Perkiraan waktu yang diperkirakan barang akan sampai ke penerima dengan ketentuan yang akan terjadi.
3. Jarak yang ditempuh
   Jarak merupakan Panjang keseluruhan lintasan yang ditempuh dari si pengirim hingga si penerima.

1. 6. 2 Data Mining

Data mining adalah proses menemukan pola dan pengetahuan yang menarik dari data dalam jumlah besar. Sumber data dapat berupa database, gudang data, *web*, repositori informasi lainnya, atau data yang dialirkan ke dalam sistem secara dinamis (Jiawei Han at all, 2016). Data mining juga didefinisikan sebagai proses untuk menemukan hubungan, pola, dan tren baru yang bermakna dengan menyaring data yang sangat besar yang tersimpan dalam penyimpanan, menggunakan teknik pengenalan pola seperti metode statistik dan matematika. Berawal dari beberapa disiplin ilmu, data mining bertujuan untuk memperbaiki teknik tradisional sehingga bisa menangani:

a. Jumlah data yang sangat besar,
b. Dimensi data yang tinggi,
c. Data yang heterogen dan berbeda sifat.

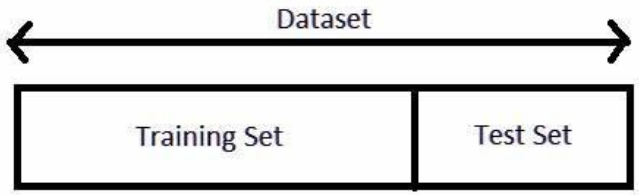
Menurut para ahli, data mining adalah analisis dari observasi data dalam jumlah besar untuk menemukan hubungan yang sebelumnya tidak diketahui dan metode baru untuk meringkas data agar lebih mudah dipahami dan berguna bagi pengguna data (Devi Heryana, 2019). Salah satu teknik yang dikembangkan dalam data mining adalah menelusuri data yang ada untuk membangun model, kemudian menggunakan model tersebut untuk mengenali pola data lain yang tidak terdapat dalam basis data yang tersimpan. Teknik ini juga dapat dimanfaatkan untuk kebutuhan prediksi. Data mining adalah pengembangan dari analisis data tradisional dan pendekatan statistik yang mengintegrasikan teknik analisis dari berbagai disiplin ilmu,

seperti analisis numerik, pencocokan pola, dan kecerdasan buatan termasuk pembelajaran mesin, jaringan saraf tiruan, dan algoritma genetika (Jackson, 2002).

1. 6. 3 Machine Learning

Machine Learning atau Pembelajaran Mesin adalah bidang ilmu yang mempelajari algoritma komputer yang mampu mengenali pola dalam data untuk mengubah berbagai jenis data menjadi tindakan nyata dengan meminimalkan keterlibatan manusia. Pembelajaran mesin bertujuan agar sistem komputer mampu belajar dari data dan mengekstrak pola yang bermanfaat dari data sehingga sistem tersebut dapat belajar tanpa harus diprogram secara spesifik (Hilman Ferdinandus Pardede, 2023). Kemampuan Pembelajaran Mesin untuk mengumpulkan data dan mempelajari informasi yang tersedia memungkinkannya melaksanakan berbagai tugas yang berbeda, tergantung pada konteks pembelajaran yang telah dilakukan.

Dengan menggunakan *machine learning*, kita dapat membuat komputer "belajar" dari data yang ada dan membuat keputusan sendiri. Proses belajar ini memerlukan dua langkah, yaitu penyusunan data latih (*training*) dan data uji (*testing*). Jumlah data uji biasanya lebih sedikit dibandingkan dengan data latih (Ibnu Daqiqil, 2021).



Gambar 1. 1 *Training* Dataset dan *Testing* Dataset
Sumber : (https://www.megabagus.id/training-set-test-set/#google_vignette)

Untuk tujuan penemuan pengetahuan, data latih digunakan untuk algoritma pembelajaran mesin untuk menemukan pola baru, dan data uji kemudian digunakan untuk memvalidasi sejauh mana pola tersebut benar-benar ada dan dapat dipercaya. Sedangkan untuk tujuan pemodelan prediktif, data latih digunakan untuk algoritma pembelajaran mesin untuk membangun model, dan data uji kemudian digunakan untuk mengevaluasi akurasi prediksi model (Liu & Cocea, 2017).

1. 6. 4 Klasifikasi

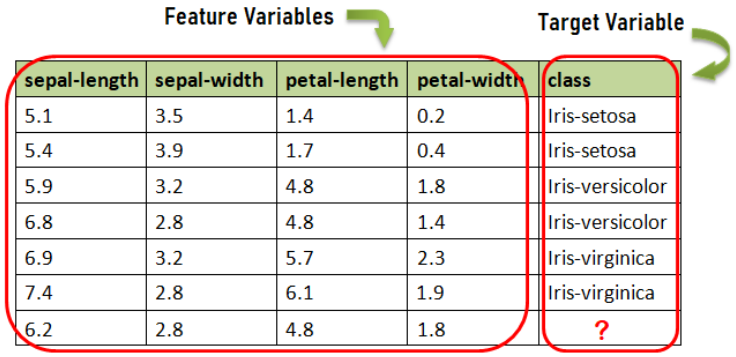
Klasifikasi merupakan aktivitas yang membutuhkan penggunaan *algoritma machine* learning yang mempelajari cara mengelompokkan kelas-kelas kedalam suatu kelas tertentu (Jasman dkk, 2022). Contoh sederhana yang mudah diketahui ialah mengklasifikasikan pasien ke dalam kelas “sembuh” atau “belum sembuh”. Klasifikasi juga merupakan bentuk analisis data yang bertujuan untuk menemukan model atau fungsi yang dapat membedakan atau menggambarkan suatu konsep atau kelas data . Berikut adalah beberapa elemen yang membentuk klasifikasi:

1. **Fitur (Features)**: Fitur adalah atribut atau karakteristik individu dari data yang digunakan oleh model untuk melakukan prediksi. Misalnya, dalam analisis teks,

fitur dapat berupa kata-kata atau frasa. Sedangkan dalam analisis gambar, fitur dapat berupa piksel atau atribut visual lainnya.

2. **Label (Labels)**: Label atau target adalah kategori atau kelas yang ditetapkan untuk setiap sampel data dalam dataset pelatihan. Label ini digunakan oleh model untuk belajar pola dan membuat prediksi. Misalnya, dalam klasifikasi spam email, label bisa berupa "spam" atau "bukan spam".
3. **Model Machine Learning**: Model ini merupakan kumpulan algoritma yang digunakan untuk mempelajari hubungan antara fitur-fitur dari data dan label-labelnya. Setelah dilatih dengan data pelatihan, model ini dapat digunakan untuk memprediksi label dari data baru yang belum dilihat sebelumnya.

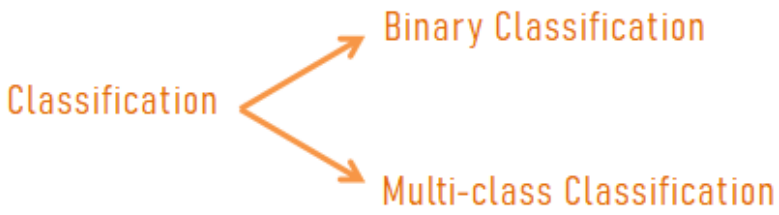
Klasifikasi berusaha memahami hubungan antara kumpulan variabel fitur dan variabel target. Dalam klasifikasi, variabel target bertipe kategori sedangkan variabel fitur bertipe numerik.



sepal-length	sepal-width	petal-length	petal-width	class
5.1	3.5	1.4	0.2	Iris-setosa
5.4	3.9	1.7	0.4	Iris-setosa
5.9	3.2	4.8	1.8	Iris-versicolor
6.8	2.8	4.8	1.4	Iris-versicolor
6.9	3.2	5.7	2.3	Iris-virginica
7.4	2.8	6.1	1.9	Iris-virginica
6.2	2.8	4.8	1.8	?

Gambar 1. 2 Ilustrasi Variabel Target dan Variabel Fitur
Sumber : (https://ilmudatapy.com/apa-itu-regresi-klasifikasi-dan-clustering-klasterisasi/)

Variabel target adalah kategori atau kelas yang ingin kita klasifikasi atau prediksi, sedangkan variabel fitur adalah informasi yang membantu model dalam menentukan kelas dari variabel target. Model klasifikasi menggunakan pola dalam variabel fitur untuk mengelompokkan data baru ke dalam kategori yang sesuai di variabel target. Di dalam *machine learning*, klasifikasi dibagi menjadi dua jenis yaitu *Binary classification* dan *Multi-class classification.*



Gambar 1. 3 Tipe Model Klasifikasi
Sumber : (https://ilmudatapy.com/apa-itu-regresi-klasifikasi-dan-clustering-klasterisasi/)

Binary classification terjadi ketika variabel target hanya memiliki dua kategori, seperti 0 dan 1 atau Ya dan Tidak. Contoh aplikasinya adalah dalam memprediksi apakah seorang nasabah bank akan mengambil pinjaman atau tidak. Sebaliknya, *multi-class classification* terjadi ketika variabel target memiliki lebih dari dua kategori. Sebagai contoh, dataset bunga adalah contoh *multi-class classification* karena mencakup tiga kategori: bunga-setosa, bunga-versicolor, dan bunga-virginica.

Menurut (Gupta dkk, 2015) Langkah-langkah untuk proses klasifikasi adalah sebagai berikut :

1) Tahap pertama membangun model dari data latih dengan nilai label kelas diketahui. Algoritma klasifikasi digunakan untuk membuat model dari dataset latih.
2) Langkah kedua melakukan pemeriksaan akurasi dari data latih, jika kebenaran model memuaskan maka model akan digunakan untuk mengklasifikasikan data dengan label kelas yang tidak diketahui.

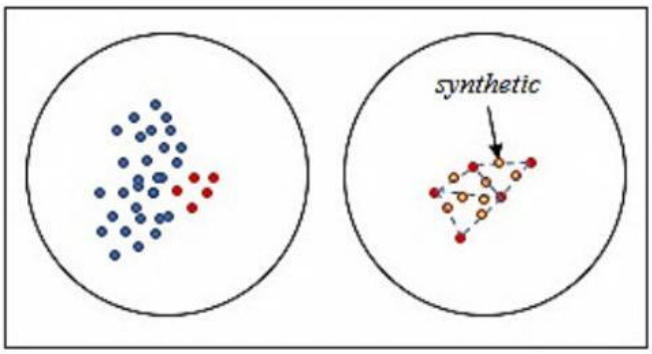
1. 6. 5 Data *Imbalanced*

Data *Imbalanced* atau data yang tak seimbang merupakan suatu kondisi dimana pada sebuah himpunan data terdapat satu kelas yang memiliki jumlah kasus yang kecil bila dibandingkan dengan kelas lainnya. *Imbalanced* Dataset merupakan data yang biasanya diolah secara klasifikasi dengan salah satu kelas pada datanya mempunyai nilai yang sangat jauh berbeda jumlahnya dari kelas lainnya. Kelas data yang memiliki jumlah yang banyak disebut sebagai kelas mayoritas, sementara kelas data yang memiliki jumlah yang sedikit disebut sebagai kelas minoritas. Misalnya, dalam sebuah dataset di mana hanya 1% dari instansinya termasuk dalam kelas minoritas, jika model mengklasifikasikan semua instansi sebagai kelas mayoritas, model tersebut masih akan mendapatkan akurasi keseluruhan sebesar 99%. Namun, instansi kelas minoritas, yang ingin kita klasifikasikan dengan tepat, semuanya salah diklasifikasikan oleh model ini meskipun memiliki akurasi yang sangat tinggi. Oleh karena itu, berbagai metode telah dikembangkan untuk mengatasi masalah klasifikasi yang tidak seimbang. Salah satu metode yang paling populer adalah dengan menggunakan *Sampling-based approaches* (Sun et al., 2015). Dengan adanya

penerapan *sampling* pada data *imbalanced*, tingkat *imbalanced* data semakin kecil dan klasifikasi dapat dilakukan dengan tepat. Pendekatan berbasis *sampling* merupakan modifikasi distribusi data latih *sehingga* kedua kelas data (mayoritas dan minoritas) lebih terwakili dalam data latih. *Sampling* dibagi menjadi dua jenis, yaitu *undersampling* dan *oversampling*. Metode *oversampling* digunakan untuk menyeimbangkan distribusi data dengan cara menambah jumlah data pada kelas minoritas, sedangkan metode *undersampling* dilakukan dengan mengurangi jumlah data pada kelas mayoritas. Dalam penelitian ini, digunakan metode oversampling menggunakan *Synthetic Minority Oversampling Technique* (SMOTE).

1. 6. 5. 1 *Synthetic Minority Oversampling Technique* (SMOTE)

Synthetic Minority Oversampling Technique, atau disingkat SMOTE merupakan salah satu metode *oversampling* dipaparkan oleh Nitesh Chawla dan rekan-rekannya dalam makalahnya tahun 2002. Metode SMOTE menciptakan data sintetis dengan menggunakan interpolasi linear antara titik kelas minoritas dan salah satu dari K tetangga terdekatnya. SMOTE adalah metode oversampling yang efektif dan telah banyak diterapkan dalam berbagai aplikasi (Elreedy et al., 2023). Teknik penambahan jumlah sampel pada kelas minoritas dilakukan dengan mereplikasi data kelas minoritas secara acak sehingga jumlah data menjadi sama dengan data kelas mayoritas. Data yang direplikasi berasal dari kelas minoritas. Metode yang digunakan dalam algoritma SMOTE adalah *k-nearest neighbors* (ketetanggaan data), yang termasuk dalam kelompok metode statistik nonparametrik. Jumlah replikasi yang dilakukan disesuaikan dengan jumlah anggota pada kelas mayoritas. Ilustrasi algoritma SMOTE ditunjukkan pada gambar berikut.



Gambar 1. 4 Ilustrasi Algoritma SMOTE

Sumber : (https://www.researchgate.net/figure/Process-of-Synthetic-Minority-Oversampling-Technique-SMOTE-Algorithm-27_fig1_368993605)

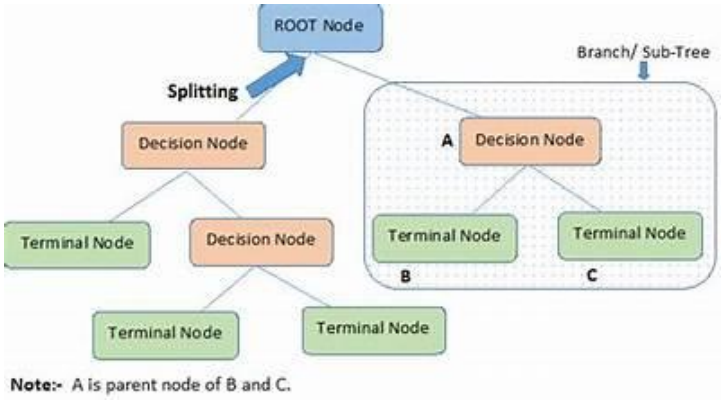
Berdasarkan Gambar 1.4 proses SMOTE dapat dilakukan pada pembuatan sintetik baru dengan tahapan algoritma sebagai berikut :

- Identifikasi fitur vektor dan tetangganya
- Mengambil perbedaan antara kedua hal tersebut.
- Menggandakan perbedaan dengan jumlah acak antara 0 dan 1.

- Identifikasi titik baru pada garis segmen dengan menambahkan jumlah acak ke fitur vektor.
- Mengulangi proses tersebut di atas untuk fitur yang diidentifikasi sampai dengan selesai.

1. 6. 6 *Decision Tree* (Pohon Keputusan)

Pohon keputusan adalah model yang banyak digunakan untuk tugas klasifikasi dan regresi. Pada dasarnya, mereka mempelajari hierarki pertanyaan *if/else*, yang mengarah pada keputusan. *Decision Tree* (Pohon Keputusan) adalah model prediksi yang menggunakan struktur pohon atau struktur hierarki, di mana setiap *node* internal menunjukkan pengujian pada atribut, setiap cabang (*branch*) menunjukkan hasil pengujian, dan setiap simpul daun (*leaf node*) mewakili label kelas. Bagian paling atas dari decision tree adalah akar pohon *(root node*) (Gourav & Kaur, 2020). *Decision Tree* digunakan untuk mempelajari klasifikasi dan prediksi pola dari data serta menggambarkan relasi antara variabel atribut x dan variabel target y dalam bentuk pohon (Nasrullah, 2021).



Gambar 1. 5 Ilustrasi *Decision Tree*
Sumber : (https://softscients.com/2020/05/02/buku-belajar-dasar-statistika-dengan-r-decision-tree/)

Berdasarkan Gambar 1. 5 di atas, simpul akar (root node) mewakili seluruh populasi atau sampel yang kemudian dibagi menjadi dua atau lebih. *Splitting* adalah proses membagi simpul (node) menjadi dua atau lebih. Simpul keputusan (decision node) adalah ketika sebuah sub-node dipecah menjadi sub-node lebih lanjut. Cabang (branch) adalah bagian dari keseluruhan pohon.

Metode pohon keputusan mengubah sejumlah besar fakta menjadi pohon keputusan yang mewakili aturan-aturan. Pohon keputusan juga berguna untuk mengeksplorasi data dan menemukan hubungan tersembunyi antara beberapa variabel fitur dengan variabel target. Data dalam pohon keputusan biasanya dinyatakan dalam bentuk tabel dengan atribut dan record. Atribut menyatakan

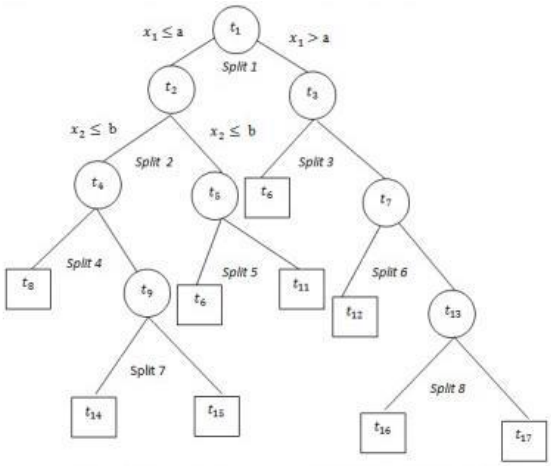
parameter yang disebut kriteria dalam pembentukan pohon. Misalnya, untuk menentukan apakah akan bermain tenis, kriteria yang diperhatikan adalah cuaca dan suhu. Salah satu atribut adalah atribut yang menyatakan data solusi untuk setiap item data, yang disebut atribut hasil (Khazari et al., 2017).

Decision Tree memiliki dua fungsi utama dalam data mining, yaitu klasifikasi dan regresi. Decision Tree membangun model regresi atau klasifikasi dalam bentuk struktur pohon. Dengan membagi kumpulan data menjadi bagian-bagian yang lebih kecil secara bertahap, pohon keputusan yang terkait juga berkembang. Pohon klasifikasi digunakan ketika variabel target adalah data kategorik, sedangkan pohon regresi digunakan ketika variabel target adalah data prediksi. Pada pohon keputusan, optimasi dilakukan dalam bentuk benar (true) atau salah (false).

1. 6. 7 *Classification and Regression tree* (CART)

Classification and Regression Trees adalah salah satu metode atau algoritma dari teknik eksplorasi data *decision tree*. Metode CART diperkenalkan oleh Leo Breiman, Jerome Friedman, Richard Olshen and Charles Stone pada tahun 1984 dalam buku “Classification and Regression Trees” (Moisen, 2008). Sejak diperkenalkan, CART telah menjadi salah satu metode yang prediktif dalam analisis data dan *machine learning*. Metode ini telah diterapkan dalam berbagai bidang, seperti ilmu data, analisis bisnis, dan ilmu prediktif. *Classification and Regression Trees* (CART) adalah metode analisis yang dipakai untuk membuat pohon keputusan melalui aturan pemisahan data. Pohon keputusan ini dapat digunakan untuk dua tujuan utama, yaitu klasifikasi dan regresi, bergantung pada jenis variabel dependen (variabel target) yang ada dalam data. Jika variabel dependen yang dimiliki bertipe kategorikal, maka CART akan menghasilkan pohon klasifikasi (classification tree). Sebaliknya, jika variabel dependen bertipe kontinu atau prediktif, maka CART akan menghasilkan pohon regresi (regression tree). Kelebihan CART adalah kemampuannya untuk menghasilkan model yang mudah diinterpretasikan dan dipahami oleh manusia. Namun, CART memiliki kekurangan yaitu pohon yang dihasilkan *overfitting* sehingga hasil prediksi model tidak akurat, varians kecil dalam data dapat menyebabkan varians yang sangat tinggi dalam prediksi, sehingga mempengaruhi stabilitas hasil CART mungkin tidak stabil dalam *decision tree* karena CART sangat sensitiv terhadap data baru. CART sangat bergantung pada jumlah sampel. Jika sampel data latih dan uji berubah maka pohon Keputusan yang dihasilkan juga ikut berubah. Oleh Karena itu, metode *ensamble* (gabungan) dikembangkan.

Menurut Breiman, et al., (1984), CART termasuk anggota analisis klasifikasi yang disebut decision trees, karena proses analisis dari CART digambarkan dalam bentuk atau struktur yang menyerupai sebuah pohon, lebih tepatnya pohon klasifikasi berbentuk biner.

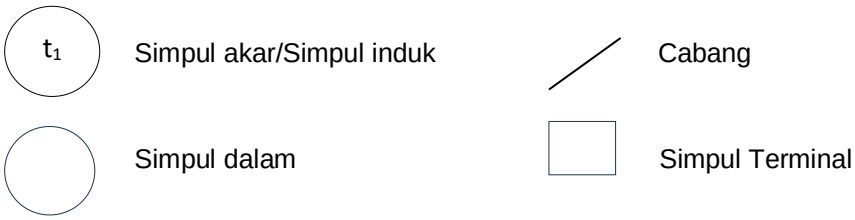


Gambar 1. 6 Gambar Pohon Klasifikasi CART

Sumber :

(https://ejournal3.undip.ac.id/index.php/gaussian/article/download/14715/14239)

Simbol pada pohon klasifikasi biner CART



Menurut (Mardika, Z. W. dkk, 2016) Pada dasarnya dalam membuat sebuah pohon klasifikasi, CART bekerja dalam lima langkah utama.

1. *Tree building* process yaitu proses pembentukan atau pembuatan sebuah pohon klasifikasi. Terdiri dari proses pemecahan simpul yaitu proses pemecahan simpul induk menjadi dua buah simpul anak melalui aturan pemecahan (splitting rules) tertentu dan dilakukan secara berulang (recursive).
2. Proses *class assignment* yaitu proses mengidentifikasikan simpul-simpul yang terbentuk pada suatu kelas tertentu melalui aturan pengidentifikasian (assignment rules).
3. Proses penghentian pembuatan atau pembentukan pohon klasifikasi (stopping the trees building process). Pada tahap ini pohon terakhir atau maximal tree (Tmax) telah terbentuk.
4. Pruning the tree yaitu proses pemangkasan atau pemotongan Tmax menjadi pohon yang lebih kecil (T).
5. Proses optimal tree selection yaitu pemilihan atau penentuan pohon klasifikasi yang optimal.

1. 6. 8 *Ensemble Learning*

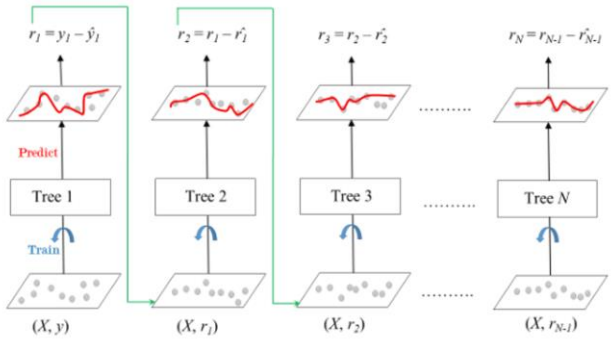
Ensemble Learning adalah suatu pendekatan dalam pembelajaran mesin di mana sejumlah model yang sering disebut "*weak learners*" dilatih untuk

menyelesaikan masalah yang sama dan kemudian digabungkan untuk meningkatkan hasil akhir. Ide dasarnya adalah bahwa dengan menggabungkan model-model yang lemah, kita dapat menciptakan model yang lebih akurat. Ada tiga jenis utama dari *Ensemble Learning*, yaitu *Boosting*, *Bagging*, dan *Stacking*. Ketiga metode ini memiliki kesamaan dalam hal penggunaan beberapa model dasar atau base model dari berbagai algoritma *machine learning* (Cendani dan Adi Wibowo, 2022).

1. 6. 9 Algoritma *Gradient Boosting*

Gradient boosting termasuk *supervised learning* berbasis *decision tree* yang dapat digunakan untuk klasifikasi dengan teknik boosting (Silvia Elsa Suryana .,dkk, 2021). Algoritma *Gradient Boosting* bekerja dengan cara menambahkan prediktor baru secara bertahap kedalam *ensemble*, menggantikan prediktor yang dinilai kurang akurat, dan kemudian mengevaluasi apakah kesalahan dari prediktor sebelumnya telah diperbaiki.

Pada Gradient Boosting, prediksi akhir dilakukan secara adiktif, Terdapat input yang di perlukan untuk Gradient Boosting yaitu fungsi kerugian (loss function) yang akan diminimumkan dan bersifat diturunkan $L(y_i, F(x))$, data peubah respon $\{(x_i, y_i\}_1^n$, banyaknya iterasi M, dan learning rate v. Banyaknya iterasi pada Gradient Boosting menyatakan banyaknya pohon yang di bentuk (Friedmand, 2002).



Gambar 1. 7 Ilustrasi Algoritma *Gradient Boosting*

Sumber : (https://pythongeeks.org/gradient-boosting-algorithm-in-machine-learning/)

Dalam *Gradient Boosting*, tahap awal dimulai dengan pembuatan pohon klasifikasi awal yang berupa nilai konstan F_0 berdasarkan persamaan, secara default F_0 adalah nilai yang meminimuman kerugian. Pada klasifikasi binner, nilai konstan awal dibentuk berdasarkan peluang atau probabilitas dari masing-masing kelas, dihitung berdasarkan log odds (logit). Kemudian dilakukan iterasi berulang untuk memperbaiki pohon klasifikasi tersebut dengan menambahkan pohon-pohon baru, sampai model memenuhi kriteria tertentu (Natekin dan Alois Knoll, 2013).

Menurut (Aryal dan Bikalpa Paudel, 2020) Langkah-langkah *Gradient Boosting* :

1) Membuat nilai prediksi awal $F_0(x)$.

$$F_0(x) = \log\left(\frac{x}{1-x}\right)$$

$F_0(x)$ = nilai prediksi kosntan awal

x =nilai target klasifikasi

2) Menghitung residual

Untuk $m = 1$ sampai M, hitung :

$$r_{im} = y_i - p$$

r_{im} = residual pada iterasi m untuk data ke-i

y_i = nilai variable Y pada data ke-i

p = nilai probabilitas prediksi

3) Membuat pohon klasifikasi

Kemudian sesuaikan dengan pohon klasifikasi dengan nilai r_{im} dan buat daerah simpul terminal R_{jm} untuk $j = 1, \dots, jm$. Dengan R_{jm} adalah himpunan sampel pada simpul terminal untuk iterasi ke-m, j adalah indeks simpul terminal dalam pohon Keputusan, dan jm adalah jumlah simpul terminal yang dihasilkan pohon Keputusan pada iterasi ke-m.

4) Setelah pohon terbentuk, kemudian menghitung pembaruan prediksi di setiap simpul terminal yang meminimalkan loss function. Pada kasus klasifikasi binner dengan los log-loss pembaruan nilai di setiap simpul terminal dihitung dengan rumus :

Untuk $j = 1, \dots, jm$. Hitung :

$$Y_{jm} = \frac{\sum_{j=1}^{jm}(y-p)}{\sum_{j=1}^{jm} p(1-p)}$$

Y_{jm} = adalah prediksi target pada iterasi ke-jm.

5) Langkah selanjutnya adalah memperbarui model $F_m(x)$ dalam bentuk *greedy function* untuk mengoptimalkan fungsi loss pada setiap iterasi. Update model :

$$F_m(x) = F_{m-1}(x) + v\sum_{j=1}^{jm} Y_{jm} J(x \in R_{jm})$$

$F_m(x)$ = nilai prediksi baru untuk masing-masing data training pada iterasi ke-m.

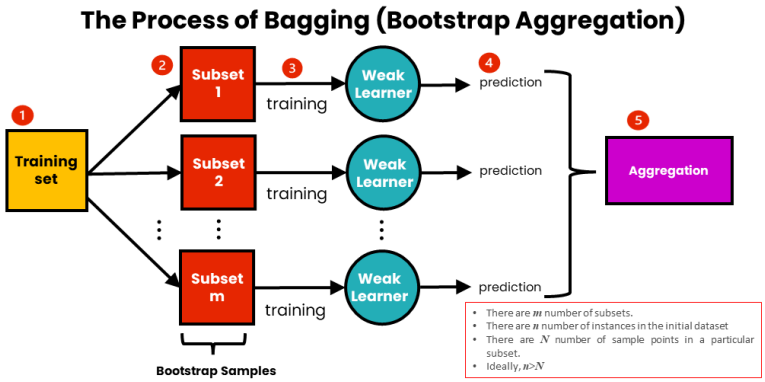
Setelah setiap iterasi, kita mengidentifikasi bagian di mana model melakukan kesalahan dan berusaha memperbaikinya. Setiap model yang ditambahkan mencoba memprediksi kesalahan residual (yaitu selisih antara prediksi dengan nilai sebenarnya). Setiap model baru yang ditambahkan dalam Gradient Boosting diberi bobot berdasarkan seberapa baik performanya dalam mengurangi kesalahan dari model sebelumnya. Model yang lebih baik akan mendapatkan bobot yang lebih besar, sehingga berkontribusi lebih banyak terhadap prediksi akhir.

Gradient Boosting adalah teknik yang sangat kuat, namun berisiko mengalami overfitting, yaitu ketika model terlalu menyesuaikan diri dengan data latih sehingga kinerjanya kurang optimal pada data baru. Untuk mengatasi ini, Gradient Boosting menerapkan regularisasi melalui beberapa teknik berikut:

1. **Learning Rate**: Mengatur seberapa besar langkah yang diambil pada setiap iterasi. Learning rate yang lebih kecil membuat proses belajar lebih lambat, namun mengurangi risiko overfitting.
2. **Jumlah Pohon**: Terlalu banyak pohon dapat meningkatkan risiko overfitting, sehingga jumlah pohon sering kali dibatasi untuk menjaga model tetap efisien.
3. **Kedalaman Pohon**: Dengan membatasi kedalaman pohon, model tetap sederhana, yang membantu mencegahnya dari mempelajari pola yang terlalu spesifik dari data latih.

1. 6. 10 Algoritma *Bagging Classifier*

Algoritma *Bagging* di usulkan pada tahun 1994 oleh Breiman (Leo Breiman, 1996) sebagai metode *ensemble* dalam pembelajaran mesin yang bertujuan untuk meningkatkan akurasi metode klasifikasi statistik. Algoritma *Bagging* berasal dari teknik *"Bootstrap"* dan telah banyak digunakan dalam bidang seperti biostatistik dan penginderaan jauh. Algoritma *Bagging* dioptimalkan untuk kumpulan data pelatihan dengan dimensi yang relatif kecil. Bagging membangun suatu model dengan menggunakan beberapa pohn Keputusan (*decision tree)* dan menggabungkan prediksinya untuk mendapatkan prediksi yang lebih akurat dan stabil daripada hanya mengandalkan satu pohon Keputusann *decision tree*. Algoritma Bagging terdiri dari dua jenis yaitu regresi dan klasifikasi. Dalam Bagging regresi hasilnya secara numerik atau kontinu, sedangkan hasil klasifikasi dalam kategori. Operasi dasar algoritma dapat dilihat pada gambar berikut.



Gambar 1. 8 Prinsip Kerja Algoritma *Bagging*
Sumber : (https://blog.stackademic.com/bagging-balancing-bias-and-variance-for-better-predictions-6b61ad1ba575)

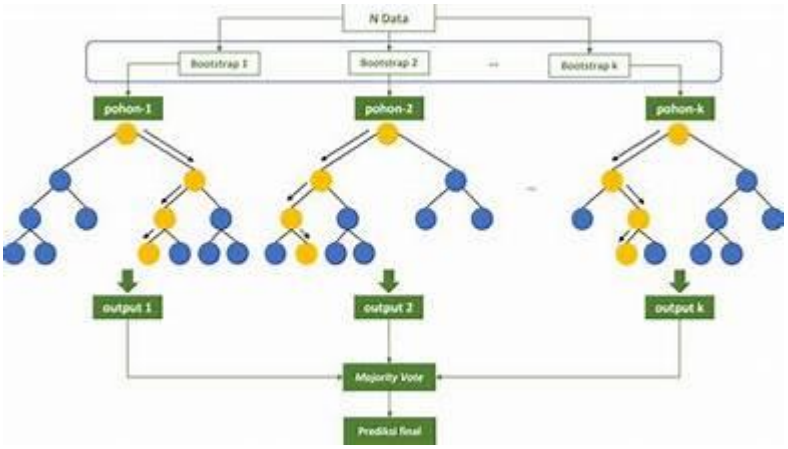
Set pelatihan awalnya dibagi menjadi N sub-cluster. Setiap sub-cluster ini digunakan sebagai set pelatihan untuk menghasilkan klasifikasi. Kemudian, klasifikasi yang dihasilkan digabungkan untuk membuat keputusan akhir. Oleh karena itu, prosedur ini dikenal sebagai *Bagging*. Singkatnya, jika data pelatihan terdiri dari N item, kumpulan data pelatihan dengan N sampel dihasilkan melalui pemilihan acak dari kumpulan pelatihan. Dalam proses ini, beberapa sampel mungkin tidak terpilih dalam set data pelatihan tertentu, sementara yang lain mungkin terpilih beberapa kali. Setiap pohon keputusan dilatih dengan data pelatihan yang terdiri dari sampel yang diambil secara acak, dan hasil akhir ditentukan oleh suara mayoritas (Widodo et al., 2021). Langkah-langkah algoritma *bagging* adalah sebagai berikut (Bisri & Rachmatika, 2019).

1) Ambil dataset awal
   Set data latih $D = \{(x_1, y_1), \dots (x_n, y_n)\}$,
2) Buat sampel dari dataset dengan penggantian
   Sampel T set elemen n dari D $(replacement)$
   $D_1, D_2, \dots D_T \rightarrow T$ set pelatihan

3) Latih model pada setiap sampel
   $Training$ di setiap $D_i, = 1, \dots, T$ dan urutan T $output$ $f_1(x), \dots, f_T(x)$
4) Gabungkan hasil dari semua model
   Aggregate classifier dapat digunakan untuk regresi maupun klasifikasi.
   Regresi
   $f(x) = \sum_{i=1}^{T} f_i(x)$
   Rata-rata $f_{iuntuk\ i=1,\dots,T}$
   Klasifikasi
   $f(x) = sign(\sum_{i=1}^{T} f_i(x))$
   Atau
   $f(x) = sign(\sum_{i=1}^{T} sign(f_i(x)))$

Proses pembentukan model melibatkan beberapa langkah utama yang memadukan teknik *ensemble learning* (khususnya bagging) dan pemilihan fitur secara acak. Proses ini dirancang untuk menciptakan beberapa *decision trees* independen, yang hasilnya kemudian digabungkan untuk meningkatkan akurasi dan stabilitas model. Langkah pertama adalah memilih beberapa subset acak dari dataset pelatihan dengan metode *bootstrap sampling*. Dalam metode ini, data diambil secara acak dari dataset asli dengan pengulangan (*with replacement*), sehingga setiap subset data bisa memiliki pengamatan yang sama berulang kali, dan beberapa pengamatan dari dataset asli mungkin tidak dipilih. Setelah subset acak dari data dibuat, pembentukan *decision tree* dimulai. Setiap *decision tree* dilatih menggunakan subset data yang berbeda.

Setelah semua *decision trees* terbentuk, model menggabungkan atau diagregasi (*aggregate*) prediksi dari setiap pohon untuk menghasilkan prediksi akhir. Untuk masalah klasifikasi, Bagging menggunakan mayoritas suara dari semua pohon. Setiap pohon memberikan prediksi kelas, dan kelas yang dipilih oleh mayoritas pohon akan menjadi prediksi akhir.



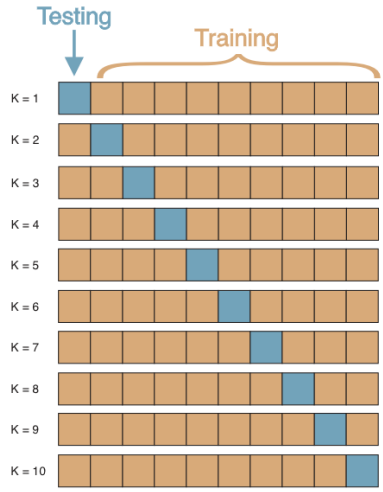
Gambar 1. 9 Kontruksi klasifikasi Bagging
Sumber : https://sainsdata.id/machine-learning/893/random-forest-untuk-model-klasifikasi-menggunakan-scikitlearn-python/

Proses pembentukan Bagging juga melibatkan pengaturan beberapa *hyperparameter* yang dapat mempengaruhi performa model. Hyperparameter utama dalam bagging meliputi:

1. **Jumlah pohon** dalam hutan (*n_estimators*): Semakin banyak pohon, semakin stabil prediksi, namun juga membutuhkan lebih banyak waktu komputasi.
2. **Random state** menentukan nilai seed untuk pengacakan. Jika diatur ke nilai tertentu, hasil pelatihan akan dapat direproduksi. Penggunaan *random_state* membantu dalam membuat eksperimen lebih konsisten, terutama saat membandingkan performa model di beberapa eksperimen.

1. 6. 11 *Cross- Validation*

Cross-Validation merupakan salah satu metode statistik yang digunakan untuk mengevaluasi kinerja suatu model. Metode *Cross-Validation* yang digunakan dalam penelitian ini adalah *k-fold Cross-Validation*. Proses kerja *k-fold Cross-Validation* melibatkan pembagian data menjadi k grup atau *fold* secara acak, dengan setiap grup memiliki ukuran yang sama atau hampir sama. Selanjutnya, dilakukan k kali validasi, di mana setiap kali validasi menggunakan satu grup sebagai data uji dan sisa grup lainnya sebagai data pelatihan. Tahap akhir adalah menghitung rata-rata akurasi dari semua k kali validasi, yang kemudian digunakan sebagai evaluasi akhir (Olson & Delen, 2008). Dalam data mining dan *machine learning 10-fold cross-validation* adalah yang paling umum digunakan (Refaeilzadeh, P., Tang, L., & Liu, 2018).



Gambar 1. 10 Ilustrasi dari *K-Fold Cross-Validation* dengan K=10
Sumber : (file:///C:/Users/ASUS/Downloads/767-Article%20Text-3270-2-10-20230206%20(2).pdf)

Berdasarkan gambar di atas yang mengilustrasikan skema 10-fold cross-validation, data dibagi menjadi 10 fold dengan ukuran yang sama, sehingga terdapat 10 subset data untuk mengevaluasi kinerja algoritma. Pada setiap subset data, cross-validation akan menggunakan 9 fold sebagai data pelatihan dan 1 fold sebagai data uji.

1. 6. 12 Evaluasi Model Klasifikasi

Setelah melakukan analisis pada metode klasifikasi, hasil prediksi didapatkan dengan menjalankan algoritma untuk menemukan nilai prediksi terbaik. Secara umum, pengukuran kinerja algoritma klasifikasi dilakukan dengan membandingkan nilai prediksi dari algoritma dengan nilai target variabel pada data uji sebagai data aktual.

1. 6. 12. 1 *Confusion Matrix*

Confusion matrix adalah alat untuk analisis prediktif dalam *machine learning*. *Confusion matrix* digunakan untuk memeriksa kinerja model *machine learning* berbasis klasifikasi. Confusion matrix adalah matriks N x N yang digunakan untuk mengevaluasi kinerja model klasifikasi, dimana N adalah jumlah kelas target (Ting, 2017). Matriks ini berisi 4 istilah penting yaitu *True Positive* adalah jumlah observasi kelas positif yang tepat diklasifikasikan pada kelas positif, *True Negative* adalah observasi kelas negatif yang tepat diklasifikasikan pada kelas negatif, *False Positive* adalah jumlah observasi negatif yang salah diklasifikasikan sebagai kelas positif, sedangkan *False Negative* adalah jumlah observasi positif yang salah diklasifikasikan sebagai kelas negatif f (Chawla *et al*., 2002). Dengan melihat confusion matrix, kita bisa mengetahui akurasi model dengan memperhatikan nilai-nilai diagonal yang menunjukkan jumlah klasifikasi yang benar.

Aktual	Prediksi	
	Negatif	Positif
Negatif	TN	FP
Positif	FN	TP

Tabel 1. 1 *Confusion Matrix*

Menurut (Doreswamy & Hemanth, 2011) dalam pengukuran kinerja *confusion matrix* terdapat empat bagian untuk mengidentifikasi suatu prediksi, diantaranya sebagai berikut :

1. TP (*True Positive*) adalah jumlah data dengan nilai aktual positif dan nilai prediksi positif
2. TN (*True Negative*) adalah jumlah data dengan nilai aktual positif dan nilai prediksi negative
3. FP (*False Positive*) adalah jumlah data dengan nilai aktual negatif dan nilai prediksi positif
4. FN (*False Negative*) adalah jumlah data dengan nilai aktual negatif dan nilai prediksi negatif

Pada klasifikasi biner terdapat beberapa nilai evaluasi yang sering digunakan. Dapat dilihat pada *confusion matrix* berikut (Sokolova & Lapalme, 2009) :

1. *Accuracy* (ACC) adalah presentase ukuran performa model yang diperoleh dengan membandingkan jumlah prediksi benar (positif dan negatif) dengan total keseluruhan data dalam dataset.

$$accuracy = \frac{(TP+TN)}{(TP+TN+FP+FN)} \times 100\% \quad (1)$$

2. *Precision* (PREC) adalah presentase prediksi positif yang benar, dihitung dengan membandingkan jumlah prediksi positif yang benar, dengan total prediksi positif.

$$Precision = \frac{TP}{(TP+FP)} \times 100\% \quad (2)$$

3. *Recall* atau *sensitivity* adalah presentase kasus positif yang berhasil terprediksi dengan benar, dihitung dengan membandingkan jumlah prediksi positif yang benar dengan total kasus positif sebenarnya.

$$Recall = \frac{TP}{(TP+FN)} \times 100\% \quad (3)$$

4. *F1-Score* adalah ukuran keseimbangan antara *presisi* dan *recall.*

$$F1 - Score = 2 \times \frac{(Recall \times Presisi)}{(Recall + Presisi)} \quad (4)$$

Dimana ,
TP = True Positif
TN = True Negatif
FP = False Positif
FN = False Negatif

1. 6. 12. 2 Variabel *Importance*

Variabel importance atau pentingnya variabel adalah ukuran yang digunakan untuk menentukan seberapa besar pengaruh suatu variabel terhadap hasil atau prediksi dalam suatu model statistik atau *machine learning*. *Variabel Importance* digunakan untuk mengetahui seberapa penting variabel prediktor dalam mempengaruhi variabel respon. Cara kerja *Variabel Importance* adalah dengan cara mengukur seberapa besar perubahan error suatu model sebelum dilakukan permutasi dan sesudah dilakukan permutasi (Mercadier & Lardy, 2019). Variabel *Importance* diperoleh berdasarkan pemecah terbaik yang diperoleh dari pemecahan masing-masing pohon dan diakumulasikan secara terpisah untuk setiap variabel. Proses ini dinamakan *Mean Decrease Gini* (MDG). Semakin besar nilai MDG maka peubah tersebut dianggap semakin penting.

BAB II
METODE PENELITIAN

2.1 Data dan Variabel Penelitian

Data yang digunakan dalam penelitian ini adalah data pengiriman *e-commerce* di Indonesia mulai pertengahan tahun 2022 sampai dengan pertengahan tahun 2023 dengan jumlah data sebanyak 991.230, yang mencakup informasi tentang *id* order, *id* produk, tanggal pemesanan, tujuan pengiriman, waktu estimasi pengiriman, penilaian produk, kurir yang digunakan, jenis pengiriman yang digunakan, serta status pengiriman. Namun, peneliti hanya menggunakan beberapa dari informasi di atas di antaranya tujuan pengiriman, waktu estimasi pengiriman, penilaian produk, kurir, jenis pengiriman dan status pengiriman. Data ini merupakan data sekunder yang diperoleh dari kaggle yang dapat di akses di https://www.kaggle.com/datasets/adisaputra10/e-commerce-shipping-data.

Variabel	Notasi	Tipe Data
Ontime	Y	Kategori : *Ontime* *Delayed*
Courier Delivery	X1	Kategori : SiCepat J&T Express JNE Ninja Xpress POS Indonesia
City	X2	Kategori : Bandar Lampung Denpasar Makassar Pekanbaru Tangerang, dst
District	X3	Kategori : Pontianak Kota Pontianak Barat Rumbai Beji Tapos, dst.
Type of delivery	X4	Kategori : Same Day

		Express Reguler Next Day
Estimated delivery time days	X5	Numerik
Product Rating	X6	Numerik

Tabel 2.1 Struktur Data yang Digunakan

Berdasarkan Tabel 2. 1 terdapat 7 variabel yang digunakan, dimana 7 variabel tersebut dibagi menjadi enam variabel fitur yang terdiri dari *courier delivery, city, district, tipe of delivery, estimated delivery time days, dan product rating* serta satu variabel target yaitu variabel *Ontime*.

Penjelasan dari variabel-variabel yang digunakan adalah sebagai berikut :

1. *Ontime* (Y)

Varibel ini mengacu pada status pengiriman pesanan, yang menunjukkan apakah suatu pesanan berhasil dikirim tepat waktu sesuai dengan estimasi waktu pengiriman yang diberikan pada saat pemesanan. Variabel ini adalah variabel yang digunakan sebagai variable dependen. Variabel ini memiliki dua kategori, yaitu : ontime (Tepat waktu) dan delayed (tidak tepat waktu).

2. *Courier Delivery* (X1)

Courier delivery adalah variabel kategorikal yang mengidentifikasi layanan pengiriman yang digunakan untuk mengirim pesanan. Layanan pengiriman ini berupa nama perusahaan kurir tempat bekerja. Variabel ini di wakili dalam bentuk kategori diantaranya SiCepat, J&T Express, JNE, Ninja Xpress, dan POS Indonesia.

3. *City* (X2)

Variabel *city* dalam penelitian ini mengacu pada kota tujuan pengiriman pesanan. Variabel ini merupakan variabel kategorikal yang menunjukkan kota tempat pesanan di kirim.

4. *District* (X3)

Variabel ini mengacu pada kecamatan atau subdistrik dari kota tempat pesanan dikirim. Variabel *district* adalah variabel kategorikal yang menunjukkan kecamatan atau subdistrik dalam kota tempat pesanan dikirim.

5. *Type of delivery* (X4)

Type of delivery dalam penelitian ini adalah variabel yang mengacu pada jenis layanan pengirimn yang digunakan untuk mengirim pesanan. Variabel ini merupakan variabel kategorikal yang menunjukkan jenis layanan pengiriman yang dipilih untuk pesanan. Jenis layan ini bervariasi di antaranya adalah Same Day, Express, Reguler, dan Next Day.

6. *Estimated delivery time days* (X5)

Dalam penelitian ini variabel tersebut menunjukkan perkiraan waktu pengiriman untuk pesanan dalam hitungan hari. Variabel tersebut bertipe numeric yang menunjukkan jumlah hari yang diestimasi untuk pengiriman pesanan dari titik pemesanan hingga pesanan tiba di alamat tujuan.

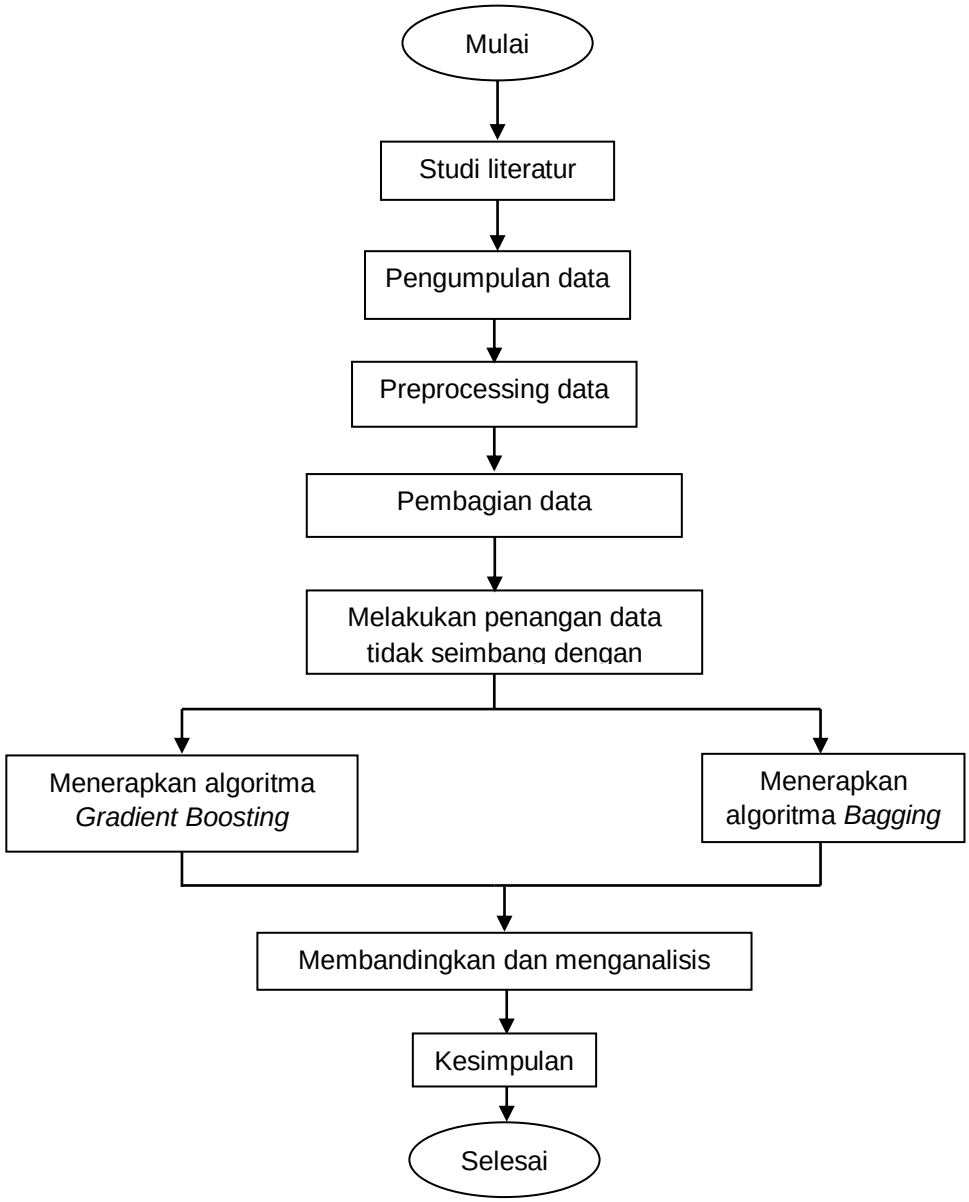
7. *Product Rating* (X6)

Variabel Product Rating dalam penelitian ini mengacu pada peringkat yang diberikan kepada produk dalam pesanan, biasanya oleh pelanggan setelah mereka menerima

produk. Variabel ini bertipe numerik pada skala 1 hingga 5 yang menunjukkan pringkat yang diberikan oleh pelanggan terhadap produk yang mereka beli. Peringkat ini mencerminkan tingkat kepuasan pelanggan terhadap produk tersebut.

2.2 Langkah-langkah Penelitian

Penelitian ini membahas penerapan algoritma Gradient Boosting dengan Bagging dalam klasifikasi sistem pengiriman produk e-commerce. Langkah-langkah penelitian ini dapat digambarkan secara skematis pada diagram alur berikut.



Gambar 2.1 Flowchart Penelitian

Penjelasan dari Flowchart :

1. Studi Literatur

   Penelitian dimulai dengan mempelajari tentang e-commerce, algoritma *Gradient Boosting*, dan algoritma *Bagging* serta perkembangannya untuk mencapai hasil yang optimal menggunakan referensi dari buku,jurnal, dan sumber lainnya.

2. Pengumpulan Data

   Melakukan pengumpulan data sekunder yang diperoleh dari kaggle, yaitu data pengiriman e-commerce di Indonesia pada pertengahan tahun 2022 sampai dengan pertengahan tahun 2023.

3. *Preprocesing* Data
   a. Transformasi Data
      Transformasi digunakan untuk mengubah atribut objek/string/kategorikal menjadi numerik karena *machine learning* hanya dapat memproses data numerik. Oleh karena itu, jika ada atribut yang masih berbentuk objek atau string, perlu dilakukan transformasi menjadi numerik. Dalam bidang *machine learning* terdapat banyak jenis *categorical encoding* dan dua yang paling popular adalah dengan menggunakan *label encoding* dan *one hot encoding.* Pada penelitian ini digunakan *label encoding pada python*. *Label encoding* mempunyai teknik mengubah kategori menjadi representasi numerik secara berurutan dari yang bernilai kecil ke besar. Sehingga proses *machine learning* bobot atau nilai yang besar akan berpengaruh.
   b. Pemberaihan Data, yaitu membersihkan data dari informasi yang tidak diperlukan. Proses ini memastikan bahwa tidak ada nilai yang hilang dalam data (*Missing value*). Jika ditemukan nilai yang hilang, pengisian bisa dilakukan berdasarkan informasi yang tersedia dalam dataset. Selain itu, pengecekan juga dilakukan untuk mendeteksi dan menghapus data duplikat karena data duplikat dapat mempengaruhi pembuatan model dan akurasi *machine learning*.
   c. Seleksi Data, yaitu proses memilih variabel data yang relevan dari dataset yang lebih besar untuk analisis atau model pembelajaran mesin. Tujuannya adalah untuk memastikan bahwa variabel yang digunakan benar-benar mendukung tujuan analisis atau pemodelan yang dilakukan, mengurangi kompleksitas data, dan meningkatkan efisiensi serta keakuratan hasil. Pada penelitian ini beberapa variabel di hapus karena tidak memberikan kontribusi signifikan terhadap hasil analisis atau model yang digunakan.

4. Pembagian Data
   a. Memilih Variabel Fitur dan Target
      Variabel fitur (Features) adalah atribut yang digunakan oleh model untuk membuat prediksi. Fitur ini terdiri dari kolom-kolom dalam dataset yang berisi informasi yang relevan untuk proses prediksi. Sedangkan variable target adalah variable yang ingin diprediksi oleh model. Dalam kolom ini biasanya berisi nilai yang ingin di prediksi, seperti kelas atau label. Pada penelitian ini variabel vitur terdiri dari kolom *courier delivery, city, district*, *Type of delivery, Estimated delivery time days,* dan *Product rating.* Sedangkan variable target yaitu kolom *ontime*.
   b. Membagi Data Menjadi Data Latih dan Uji
      Data Latih digunakan untuk melatih data dalam bentuk model, sedangkan data uji digunakan untuk memprediksi dan melihat keakuratan model yang terbentuk. Pada penelitian ini membagi data menjadi beberapa bagian yaitu 90:10, 80:20, dan 70:30, untuk membandingkan kinerja model pada berbagai skenario pelatihan dan pengujian. Dalam setiap perbandingan, angka pertama menunjukkan persentase data yang digunakan untuk melatih model, sementara angka kedua menunjukkan persentase data yang digunakan untuk menguji performa model tersebut.

5. Melakukan Penangan Data Tidak Seimbang

   Untuk mengatasi penangan data yang tidak seimbang dapat dilakukan dengan *Sythetic Minority Oversampling Technique* (SMOTE) pada *data training*.

6. Penerapan Algoritma

   Pada penelitian ini algoritma yang digunakan yaitu algoritma *Gradient boosting* dan *Bagging*.

7. Membandingkan dan Menganalisis Hasil

   Setelah mendapatkan hasil klasifikasi dari masing-masing metode langkah selanjutnya adalah membandingkan kinerja klasifikasi *Gradient Boosting* dan *Bagging* ditinjau dari nilai *accuracy, precision, sensitivity,* dan *specifity* serta *features impotance* untuk mengetahui variabel apa yang paling berpengaruh. Kemudian membandingkan hasil klasifikasi untuk menyatakan model yang paling baik berdasarakan kinerja yang ditunjukkan.