PREDIKSI DAN ANALISIS FAKTOR PENENTU *CUSTOMER LIFETIME VALUE* (CLV) PADA *E-COMMERCE* MENGGUNAKAN *RANDOM FOREST REGRESSION*

SKRIPSI

Oleh :
ANNISA ADENANTY PALUPI
NIM. 220605110166

PROGRAM STUDI TEKNIK INFORMATIKA
FAKULTAS SAINS DAN TEKNOLOGI
UNIVERSITAS ISLAM NEGERI MAULANA MALIK IBRAHIM
MALANG
2025

PREDIKSI DAN ANALISIS FAKTOR PENENTU *CUSTOMER LIFETIME VALUE* (CLV) PADA *E-COMMERCE* MENGGUNAKAN *RANDOM FOREST REGRESSION*

SKRIPSI

Diajukan kepada:
Universitas Islam Negeri Maulana Malik Ibrahim Malang
Untuk memenuhi Salah Satu Persyaratan dalam
Memperoleh Gelar Sarjana Komputer (S.Kom)

Oleh:
ANNISA ADENANTY PALUPI
NIM. 220605110166

PROGRAM STUDI TEKNIK INFORMATIKA
FAKULTAS SAINS DAN TEKNOLOGI
UNIVERSITAS ISLAM NEGERI MAULANA MALIK IBRAHIM MALANG
2025

HALAMAN PERSETUJUAN

PREDIKSI DAN ANALISIS FAKTOR PENENTU *CUSTOMER LIFETIME VALUE* (CLV) PADA *E-COMMERCE* MENGGUNAKAN *RANDOM FOREST REGRESSION*

SKRIPSI

Oleh :
ANNISA ADENANTY PALUPI
NIM. 220605110166

Telah Diperiksa dan Disetujui untuk Diuji:
Tanggal: 9 Desember 2025

Pembimbing I,	Pembimbing II,
Prof. Dr. Suhartono, M.Kom	Dr. Ir. Fachrul Kurniawan, M.MT., IPU
NIP. 19680519 200312 1 001	NIP. 19771020 200912 1 001

Mengetahui
Ketua Program Studi Teknik Informatika
Fakultas Sains dan Teknologi
Universitas Islam Negeri Maulana Malik Ibrahim Malang

Supriyono, M. Kom
NIP. 19841010 201903 1 012

HALAMAN PENGESAHAN

PREDIKSI DAN ANALISIS FAKTOR PENENTU *CUSTOMER LIFETIME VALUE* (CLV) PADA *E-COMMERCE* MENGGUNAKAN *RANDOM FOREST REGRESSION*

SKRIPSI

Oleh :
ANNISA ADENANTY PALUPI
NIM. 220605110166

Telah Dipertahankan di Depan Dewan Penguji Skripsi
dan Dinyatakan Diterima Sebagai Salah Satu Persyaratan
Untuk Memperoleh Gelar Sarjana Komputer (S.Kom)
Tanggal: 9 Desember 2025

Susunan Dewan Penguji

Ketua Penguji	: Okta Qomaruddin Aziz, M.Kom NIP. 19911019 201903 1 013	( )
Anggota Penguji I	: Khadijah Fahmi Hayati Holle, M.Kom NIP. 19900626 202203 2 002	( )
Anggota Penguji II	: Prof. Dr. Suhartono, M.Kom NIP. 19680519 200312 1 001	( )
Anggota Penguji III	: Dr. Ir. Fachrul Kurniawan, M.MT., IPU NIP. 19771020 200912 1 001	( )

Mengetahui dan Mengesahkan,
Ketua Program Studi Teknik Informatika
Fakultas Sains dan Teknologi
Universitas Islam Negeri Maulana Malik Ibrahim Malang

Supriyono, M. Kom
NIP. 19841010 201903 1 012

PERNYATAAN KEASLIAN TULISAN

Saya yang bertanda tangan di bawah ini:

Nama	: Annisa Adenanty Palupi
NIM	: 220605110166
Fakultas / Jurusan	: Sains dan Teknologi / Teknik Informatika
Judul Skripsi	: Prediksi dan Analisis Faktor Penentu *Customer Lifetime Value* (CLV) Pada E-Commerce Menggunakan Random Forest Regression

Menyatakan dengan sebenarnya bahwa Skripsi yang saya tulis ini benar-benar merupakan hasil karya saya sendiri, bukan merupakan pengambil alihan data, tulisan, atau pikiran orang lain yang saya akui sebagai hasil tulisan atau pikiran saya sendiri, kecuali dengan mencantumkan sumber cuplikan pada daftar pustaka.

Apabila dikemudian hari terbukti atau dapat dibuktikan skripsi ini merupakan hasil jiplakan, maka saya bersedia menerima sanksi atas perbuatan tersebut.

Malang, 05 Desember 2025
Yang membuat pernyataan,

Annisa Adenanty Palupi
NIM.220605110166

MOTTO

"If you're on that journey, please continue onward

because I promise there are brighter days ahead"

-Ariana Grande

HALAMAN PERSEMBAHAN

Dengan penuh rasa syukur kepada Allah SWT atas segala karunia, rahmat, dan kasih sayang-Nya yang selalu menyertai setiap langkah penulis hingga skripsi ini dapat diselesaikan. Karya ini penulis dedikasikan kepada:

Mama Papa tercinta

Yang selalu menjadi sumber cinta dan kekuatan terbesar dalam hidup penulis. Terima kasih atas cinta yang tulus tanpa syarat, doa yang tak pernah terputus, serta semangat yang senantiasa hadir bahkan di saat penulis merasa lelah dan hampir menyerah. Berkat pengorbanan, ketulusan, dan usaha beliau berdua, penulis dapat sampai pada titik ini.

Keluarga tersayang,

Yang senantiasa memberikan dukungan, perhatian, serta doa yang menjadi penguat bagi penulis dalam setiap proses perjuangan.

Dan untuk diri penulis sendiri,

Terima kasih telah bertahan, berjuang, dan tidak menyerah menghadapi setiap tantangan. Terima kasih telah melangkah sejauh ini dengan penuh kesabaran dan keteguhan hati.

Semoga karya ini menjadi awal dari perjalanan yang lebih bermakna, membawa kebermanfaatan, serta menjadi bagian dari langkah menuju masa depan yang lebih baik.

KATA PENGANTAR

Alhamdulillahirabbil ‘alamin, segala puji dan syukur penulis panjatkan ke hadirat Allah Subhanahu wa Ta’ala atas limpahan rahmat, taufik, dan hidayah-Nya, sehingga penulis dapat menyelesaikan skripsi yang berjudul “Prediksi *Customer Lifetime Value* (CLV) pada Platform *E-commerce* Menggunakan Algoritma *Random Forest Regression*” dengan baik. Penulis menyadari bahwa dalam proses penyusunan skripsi ini terdapat berbagai keterbatasan dan tantangan, namun berkat bantuan, bimbingan, dukungan, serta motivasi dari berbagai pihak, skripsi ini dapat diselesaikan dengan baik. Ucapan ini penulis sampaikan kepada:

1. Prof. Dr. Hj. Ilfi Nur Diana, M.Si., CAHRM., CRMP. selaku Rektor Universitas Islam Negeri Maulana Malik Ibrahim Malang.
2. Dr. H. Muhammad Walid, MA. selaku Dekan Fakultas Sains dan Teknologi Universitas Islam Negeri Maulana Malik Ibrahim Malang.
3. Supriyono, M. Kom, selaku Ketua Program Studi Teknik Informatika Universitas Islam Negeri Maulana Malik Ibrahim Malang.
4. Ahmad Fahmi Karami, M.Kom., selaku dosen wali yang telah mendampingi saya secara akademik selama perkuliahan, serta memberikan bimbingan dan dukungan yang sangat berarti dalam proses studi saya.
5. Prof. Dr. Suhartono, M.Kom., selaku Dosen Pembimbing I, yang telah dengan sabar meluangkan waktu dan memberikan arahan serta bimbingan yang berarti selama proses penyusunan skripsi ini.

6. Dr. Ir. Fachrul Kurniawan, M.MT., IPU., selaku Dosen Pembimbing II, yang telah memberikan banyak dukungan dan bimbingan dalam penulisan skripsi ini.
7. Okta Qomaruddin Aziz, M.Kom., selaku Ketua Penguji, atas segala saran, masukan, dan arahan yang sangat membangun sejak tahap seminar proposal hingga sidang skripsi, dan kesediaan beliau yang senantiasa meluangkan waktu untuk melaksanakan latihan, sehingga penulis dapat mempersiapkan diri dengan lebih baik.
8. Khadijah Fahmi Hayati Holle, M.Kom., selaku Anggota Penguji I, atas semua saran dan masukan yang memperkuat kualitas skripsi ini, sejak tahap seminar proposal hingga sidang skripsi
9. Seluruh Dosen, Admin, dan Staf Program Studi Teknik Informatika, yang telah berbagi ilmu, tenaga, dan kebaikan selama masa studi ini, baik secara langsung maupun tidak langsung.
10. Kedua orang tua tercinta, Papa Yoyok Suryoatmojo dan Mama Chusnul Rubayanah, yang sangat penulis cintai dan sayangi. Terima kasih atas dukungan doa, cinta yang sangat tulus, doa yang tak pernah terputus, serta semangat yang senantiasa menguatkan, Berkat pengorbanan dan usaha beliau berdua, penulis dapat berada pada titik ini.
11. Sahabat-sahabat terbaik sejak awal perkuliahan (EMTOT), yaitu Runa, Jenny, dan Leny, yang selalu setia menemani dan membantu penulis sejak awal perkuliahan. Kita melewati hari-hari penuh perjuangan yang dihiasi tawa, canda, cerita, dan air mata. Terima kasih telah menjadi tempat berbagi, penguat

di saat lelah, serta saksi perjalanan panjang yang tidak selalu mudah, namun selalu terasa lebih ringan karena kebersamaan kita.

12. Seluruh teman-teman Angkatan 2022 Teknik Informatika yang telah banyak membantu, mendukung, dan memotivasi dalam menyelesaikan penyusunan skripsi ini.
13. Nur Muhammad Najiburrohman, Terima kasih atas kesediaan untuk senantiasa mendampingi dan membantu penulis dalam setiap proses yang dilalui, dan menjadi saksi perjalanan akademik dalam setiap langkah dari awal perkuliahan hingga penulis mencapai tahap akhir penyelesaian skripsi ini.
14. Last but not least, I wanna thank me. I wanna thank me for believing in me, I wanna thank me for doing all this hard work. I wanna thank me for having no days off. I wanna thank me for never quitting.

Malang, 25 Desember 2025

Penulis

DAFTAR ISI

HALAMAN PERSETUJUAN ..... **iii**
HALAMAN PENGESAHAN ..... **iv**
PERNYATAAN KEASLIAN TULISAN ..... **v**
MOTTO ..... **vi**
HALAMAN PERSEMBAHAN ..... **vii**
KATA PENGANTAR ..... **viii**
DAFTAR ISI ..... **xi**
DAFTAR GAMBAR ..... **xiii**
DAFTAR TABEL ..... **xiv**
ABSTRAK ..... **xv**
ABSTRACT ..... **xv**
البحث مستخلص ..... **xvii**
BAB I PENDAHULUAN **1**
1.1 Latar Belakang 1
1.2 Pernyataan Masalah 3
1.3 Batasan Masalah 3
1.4 Tujuan Penelitian 4
1.5 Manfaat Penelitian 4
BAB II STUDI PUSTAKA **5**
2.1 Penelitian Terkait 5
2.2 *Customer Lifetime Value* (CLV) 9
2.3 Prediksi CLV 11
2.4 *Random Forest Regression* ..... **Error! Bookmark not defined.**
2.5 Evaluasi Model 14
BAB III DESAIN DAN IMPLEMENTASI **15**
3.1 Desain Penelitian 15
3.2 Desain Sistem 16
3.3 Persiapan Data 17
3.4 *Exploratory Data Analysis* (EDA) 18
3.5 *Pre-processing Data* 20
3.5.1 Transformasi Data 21
3.5.2 Feature Engineering 23
3.6 Split Data 25
3.7 Implementasi *Random Forest Regression* 25
3.8 Parameter *Randomized Search CV* 29
3.9 Evaluasi Model 30
3.9.1 *Mean Absolute Error* (MAE) 30
3.9.2 *Mean Squared Error* (MSE) 31
3.9.3 *Mean Absolute Percentage Error* (MAPE) 32
3.9.4 R^2 *Squared* (R^2) 32
3.10 *Feature importance* 33
3.11 Skenario Pengujian 36
BAB IV HASIL DAN PEMBAHASAN **38**

4.1 Hasil Penerapan Model *Random Forest* 38
4.1.1 Proses *Bootstrap Sampling* 38
4.1.2 *Random Feature Selection* 39
4.1.3 Pembentukan *Decision Tree* 40
4.1.4 Agregasi Hasil Prediksi (Bagging) 40
4.2 Pengujian *Hyperparameter* 42
4.3 Evaluasi Model 43
4.4 *Feature Importance* 45
4.5 Hasil Skenario Uji Coba 46
4.5.1 Skenario 0 a 47
4.5.2 Skenario 0 b 49
4.5.3 Skenario 1 a 51
4.5.4 Skenario 1 b 53
4.5.5 Skenario 2 a 55
4.5.6 Skenario 2 b 57
4.5.7 Skenario 3 a 59
4.5.8 Skenario 3 b 61
4.6 Pembahasan 63
4.6.1 Analisis Pengaruh *Hyperparameter* 63
4.6.2 Analisis Pengaruh Jumlah Fitur 65
4.6.3 Analisis *Feature Importance* dan Dominasi *Frequency* 66
4.6.4 Perbandingan Performa Antar Skenario 67
4.7 Integrasi Islam 73
BAB V KESIMPULAN DAN SARAN 78
5.1 Kesimpulan 78
5.2 Saran 78
DAFTAR PUSTAKA .....

DAFTAR GAMBAR

Gambar 2.1 Cara Kerja *Random Forest* 12
Gambar 3.1 Desain Penelitian 15
Gambar 3.2 Desain Sistem Penelitian 16
Gambar 3.3 Distribusi Fitur Numerik 19
Gambar 3.4 Korelasi Antar Fitur 20
Gambar 3.5 Jumlah Data setelah Transformasi 24
Gambar 3.6 Implementasi *Random Forest* 25
Gambar 3.7 Alur Feature importance 33
Gambar 4.1 Source Code Hyperparameter 42
Gambar 4.2 Nilai *Hyperparameter* 43
Gambar 4.3 *Source Code* Evaluasi Model 44
Gambar 4.4 *Source Code Feature Importance* 46
Gambar 4.5 Hasil Feature Importance Skenario 0 a 48
Gambar 4.6 Hasil Feature Importance Skenario 0 b 50
Gambar 4.7 Hasil Feature Importance Skenario 1 a 52
Gambar 4.8 Hasil Feature Importance Skenario 1 b 54
Gambar 4.9 Hasil Feature Importance Skenario 2 a 56
Gambar 4.10 Hasil Feature Importance Skenario 2 b 58
Gambar 4.11 Hasil Feature Importance Skenario 3 a 60
Gambar 4.12 Hasil Feature Importance Skenario 3 b 62
Gambar 4.13 Hasil Perbandingan R2 Score 68
Gambar 4.14 Hasil Perbandingan MSE 69
Gambar 4.15 Hasil Perbandingan MAE 70
Gambar 4.16 Hasil Perbandingan MAPE 72

DAFTAR TABEL

Tabel 2.1 Ringkasan Penelitian Terdahulu 8
Tabel 2.2 Kategori Hasil Evaluasi Model 14
Tabel 3.1 Penjelasan Atribut Dataset 18
Tabel 3.2 Hasil EDA 19
Tabel 3.3 Fitur Penelitian untuk Pemodelan *Random Forest Regression* 24
Tabel 3.4 Contoh Booststrap Sampling 26
Tabel 3.5 Skenario Pengujian 36
Tabel 4.1 Contoh Hasil *Bootstrap Sampling* 39
Tabel 4.2 Hasil Random Feature Selection 40
Tabel 4.3 Evaluasi Model Skenario 0 a 47
Tabel 4.4 Evaluasi Model Skenario 0 b 49
Tabel 4.5 Evaluasi Model Skenario 1 a 51
Tabel 4.6 Evaluasi Model Skenario 1 b 53
Tabel 4.7 Evaluasi Model Skenario 2 a 55
Tabel 4.8 Evaluasi Model Skenario 2 b 57
Tabel 4.9 Evaluasi Model Skenario 3 a 59
Tabel 4.10 Evaluasi Model Skenario 3 b 62
Tabel 4.11 Hasil Perbandingan Skenario 67

ABSTRAK

Palupi, Annisa Adenanty. 2025. **Prediksi dan Analisis Fakor Penentu *Customer Lifetime Value* (CLV) pada *E-commerce* menggunakan *Random Forest Regression*.** Skripsi. Program Studi Teknik Informatika Fakultas Sains dan Teknologi Universitas Islam Negeri Maulana Malik Ibrahim Malang. Pembimbing: (I) Prof. Dr. Suhartono, M.Kom (II) Dr. Ir. Fachrul Kurniawan, M.MT., IPU.

Kata Kunci : *Customer Lifetime Value* , *Random Forest Regression*, Prediksi, Feature Importance, *E-commerce*.

Customer Lifetime Value (CLV) merupakan indikator penting dalam menilai profitabilitas jangka panjang pelanggan pada platform *e-commerce*. Penelitian ini bertujuan untuk memprediksi nilai CLV serta menganalisis faktor-faktor yang berperan penting dalam pembentukannya menggunakan algoritma *Random Forest Regression*. Dataset yang digunakan terdiri dari empat fitur perilaku transaksi, yaitu *Total_Sales, Frequency, Recency,* dan *Total_Quantity*. Model dievaluasi menggunakan metrik MAE, MSE, MAPE, dan R^2 untuk menilai tingkat akurasi prediksi. Hasil penelitian menunjukkan bahwa model *Random Forest Regression* mampu memberikan performa prediksi yang sangat baik, dengan capaian terbaik pada Skenario *Baseline* yang memperoleh nilai R^2 sebesar 0.94, MAE sebesar 0.04, dan MAPE sebesar 17.96%. Analisis *feature importance* mengidentifikasi bahwa *Frequency* merupakan faktor yang paling dominan dalam memengaruhi nilai CLV, diikuti oleh *Total_Sales* dan *Total_Quantity*, sedangkan *Recency* memiliki pengaruh paling rendah. Temuan ini menunjukkan bahwa intensitas dan nilai transaksi pelanggan merupakan aspek utama dalam menentukan besar kecilnya CLV pada platform *e-commerce*. Penelitian ini memberikan kontribusi dalam penyediaan model prediksi CLV yang akurat dan dapat digunakan sebagai dasar pengambilan keputusan strategis berbasis data.

ABSTRACT

Palupi, Annisa Adenanty. 2025. **Prediksi dan Analisis Fakor Penentu *Customer Lifetime Value* (CLV) pada *E-commerce* menggunakan *Random Forest Regression*.** Thesis. Department of Informatics Engineering, Faculty of Science and Technology, State Islamic University Maulana Malik Ibrahim Malang. Advisor: (I) Prof. Dr. Suhartono, M.Kom (II) Dr. Ir. Fachrul Kurniawan, M.MT., IPU.

Keywords: *Customer Lifetime Value* , *Random Forest Regression*, Prediction, Feature Importance, *E-commerce*.

Customer Lifetime Value (CLV) is an important indicator in assessing the long-term profitability of customers on *e-commerce* platforms. This study aims to predict CLV values and analyze the factors that play a significant role in their formation using the *Random Forest Regression* algorithm. The dataset used consists of four transaction behavior features, namely *Total_Sales*, *Frequency*, *Recency*, and *Total_Quantity*. The model was evaluated using MAE, MSE, MAPE, and R^2 metrics to assess the accuracy of the predictions. The results show that the *Random Forest Regression* model is capable of providing excellent prediction performance, with the best results in the *Baseline* Scenario, which obtained an R^2 value of 0.94, MAE of 0.04, and MAPE of 17.96%. Feature importance analysis identified *Frequency* as the most dominant factor influencing CLV, followed by *Total_Sales* and *Total_Quantity*, while *Recency* had the least influence. These findings indicate that customer transaction intensity and value are the main aspects in determining the magnitude of CLV on *e-commerce* platforms. This research contributes to the provision of an accurate CLV prediction model that can be used as a basis for data-driven strategic decision making.

البحث مستخلص

بالوبي، أنيسة أديناتي٢٠٢٥. تنبؤ وتحليل العوامل المحددة لقيمة العميل مدى الحياة في التجارة الإلكترونية باستخدام انحدار الغابة ،العشوائية أطروحة. قسم هندسة المعلوماتية، كلية العلوم والتكنولوجيا، جامعة مولانا مالك إبراهيم الإسلامية الحكومية ،مالانج. المشرفون: (1) الأستاذ الدكتور سوهارتونو، ماجستير في علوم الكمبيوتر (2) الدكتور إير. فاخرول كورنياوان ماجستير في إدارة التكنولوجيا، مهندس محترف

الكلمات المفتاحية: قيمة العميل مدى الحياة، انحدار الغابة العشوائية، التنبؤ
.أهمية الميزة، التجارة الإلكترونية

.تعد قيمة العمر الافتراضي للعميل مؤشراً مهماً في تقييم الربحية طويلة الأجل للعملاء على منصات التجارة الإلكترونية *Random Forest* وتحليل العوامل التي تلعب دوراً مهماً في تكوينها باستخدام خوارزمية CLV تهدف هذه الدراسة إلى توقع قيم تتكون مجموعة البيانات المستخدمة من أربع سمات لسلوك المعاملات، وهي مبلغ المبيعات والتكرار والحداثة والكمية .*Regression* لتقييم دقة التنبؤات. أظهرت النتائج أن نموذج R^2 و MAPE و MSE و MAE الإجمالية. تم تقييم النموذج باستخدام مقاييس قادر على توفير أداء تنبؤي ممتاز، مع أفضل النتائج في السيناريو الأساسي، الذي حصل *Random Forest Regression* تبلغ 17.96٪. حدد تحليل أهمية الميزات التكرار كأكثر العوامل MAPE تبلغ 0.04 و MAE تبلغ 0.94 و R^2 على قيمة يليه إجمالي المبيعات وإجمالي الكمية، بينما كان للحداثة أقل تأثير. تشير هذه النتائج إلى أن كثافة معاملات ،CLV تأثيرًا على على منصات التجارة الإلكترونية. تساهم هذه الدراسة في توفير نموذج CLV العملاء وقيمتها هما الجانبان الرئيسيان في تحديد حجم .تنبؤ دقيق لقيمة العميل مدى الحياة يمكن استخدامه كأساس لاتخاذ قرارات استراتيجية قائمة على البيانات

BAB I

PENDAHULUAN

1.1 Latar Belakang

Dalam era digital saat ini, *e-commerce* telah menjadi salah satu sektor bisnis yang paling dinamis dan kompetitif. Keberhasilan platform *e-commerce* tidak hanya bergantung pada jumlah transaksi, tetapi juga pada kemampuan perusahaan dalam memahami dan mempertahankan pelanggan. Salah satu indikator utama untuk mengukur nilai pelanggan dalam jangka panjang adalah CLV yaitu estimasi nilai ekonomi total yang dihasilkan dari seorang pelanggan selama masa hubungan mereka dengan perusahaan (Yılmaz Benk et al., 2022). CLV menjadi faktor penting dalam strategi pemasaran berbasis data karena membantu perusahaan mengidentifikasi pelanggan bernilai tinggi, mengoptimalkan strategi retensi, dan meningkatkan efisiensi anggaran pemasaran (Nurfadilla Nurfadilla et al., 2024). Dalam konteks industri *e-commerce*, analisis CLV berperan penting untuk memahami perilaku pelanggan, menentukan strategi promosi, dan memperkirakan keuntungan jangka panjang. Namun, tantangan utama dalam penerapan CLV pada *e-commerce* adalah kompleksitas data pelanggan yang bersifat heterogen, yang mencakup frekuensi pembelian, nilai transaksi, dan variasi perilaku konsumen.

Al-Qur’an sebagai sumber utama ajaran Islam telah banyak mem beri isyarat tentang pentingnya pengelolaan harta, transaksi yang adil, serta pemanfaatan ilmu pengetahuan untuk kemaslahatan umat. Firman Allah dalam Q.S. Al-Hasyr ayat 18:

يَٰٓأَيُّهَا ٱلَّذِينَ ءَامَنُواْ ٱتَّقُواْ ٱللَّهَ وَلۡتَنظُرۡ نَفۡسٞ مَّا قَدَّمَتۡ لِغَدٖۖ وَٱتَّقُواْ ٱللَّهَۚ إِنَّ ٱللَّهَ خَبِيرُۢ بِمَا تَعۡمَلُونَ ١٨

"Wahai orang-orang yang beriman, bertakwalah kepada Allah dan hendaklah setiap orang memperhatikan apa yang telah diperbuatnya untuk hari esok (akhirat). Bertakwalah kepada Allah. Sesungguhnya Allah Mahateliti terhadap apa yang kamu kerjakan." (QS. Al-Hasyr:18).

Ayat ini mengandung makna bahwa setiap tindakan perlu direncanakan dengan mempertimbangkan manfaat jangka panjang. Dalam konteks *e-commerce*, hal ini sejalan dengan pentingnya mengelola data pelanggan untuk memprediksi potensi nilai di masa depan, sebagaimana perusahaan harus memperhatikan kontribusi pelanggan terhadap keberlanjutan bisnisnya. Dengan demikian, pengelolaan CLV merupakan bentuk penerapan prinsip efisiensi dan tanggung jawab yang sejalan dengan nilai-nilai Islam.

Berbagai penelitian sebelumnya dalam prediksi CLV masih didominasi oleh pendekatan regresi linier dan metode statistik konvensional yang sering kali tidak mampu menangkap hubungan non-linear antar variabel (Ahmadi et al., 2022). Oleh karena itu, dibutuhkan metode yang lebih adaptif terhadap karakteristik data *e-commerce* yang kompleks. Salah satu metode yang unggul dalam hal ini adalah *Random Forest Regression*, algoritma *ensemble learning* yang membangun banyak *decision tree* untuk meningkatkan akurasi prediksi dan mengurangi risiko *overfitting* (Taherkhani et al., 2025a). Kelebihan lain dari *Random Forest* adalah kemampuannya untuk menghasilkan analisis *feature importance*, yaitu pengukuran kontribusi setiap variabel terhadap hasil prediksi. Analisis ini sangat penting karena memungkinkan peneliti dan pelaku bisnis untuk mengetahui faktor-faktor utama yang menentukan nilai CLV, seperti nilai transaksi (*sales amount*), frekuensi

pembelian (*purchase Frequency*), atau jumlah produk yang dibeli (*quantity*). (Curiskis et al., 2023).

Berdasarkan uraian tersebut, penelitian ini berjudul *"*Prediksi *Customer Lifeime Value* (CLV) Pada *E-commerce* Menggunakan *Random Forest Regression* dan Analisis *Feature Importance"* Penelitian ini bertujuan untuk membangun model prediksi CLV yang akurat serta menganalisis faktor-faktor yang paling berpengaruh terhadap nilai CLV melalui *feature importance*. Hasil penelitian diharapkan memberikan kontribusi ilmiah bagi pengembangan ilmu *data mining* dan *machine learning*, serta manfaat praktis bagi pelaku *e-commerce* dalam merancang strategi retensi pelanggan dan peningkatan profitabilitas yang berkelanjutan (Win & Bo, 2020).

1.2 Pernyataan Masalah

1) Bagaimana evaluasi performa model *Random Forest Regression* dalam prediksi nilai *Customer Lifetime Value* (CLV) pada data *e-commerce* menggunakan *Mean Absolute Error* (MAE), *Mean Squared Error* (MSE), *Mean Absolute Percentage Error* (MAPE), dan *R-Squared* (R^2).
2) Faktor-faktor apa saja yang berpengaruh terhadap nilai CLV berdasarkan hasil analisis *feature importance* yang dihasilkan oleh model *Random Forest Regression*.

1.3 Batasan Masalah

1) Analisis terbatas pada parameter yang relevan dengan CLV yaitu, *Total_Sales, Total_Quantity, Frequency, Recency*.

2) Dataset yang digunakan pada penlitian ini adalah dataset publik *e-commerce* dari platform Kaggle.

1.4 Tujuan Penelitian

Tujuan dari penelitian ini adalah:

1) Untuk mengetahui bagaimana performa model dari penelitian ini menggunakan *Random Forest Regression.*
2) Untuk mengetahui faktor-faktor yang berpengaruh terhadap nilai CLV menggunakan *feature importance* dari penurunan *impurity* MSE.

1.5 Manfaat Penelitian

1) Bagi Akademisi: memberikan kontribusi ilmiah dalam pengembangan metode prediksi CLV berbasis machine learning, khususnya *Random Forest Regression*.
2) Bagi Praktisi Bisnis: memberikan wawasan mengenai faktor-faktor yang memengaruhi CLV sehingga dapat digunakan untuk strategi retensi pelanggan dan peningkatan profitabilitas.
3) Bagi Peneliti Selanjutnya: menjadi referensi bagi penelitian lanjutan dalam bidang data mining, prediksi nilai pelanggan, dan pengembangan algoritma *machine learning* untuk *e-commerce*.

BAB II

STUDI PUSTAKA

Pada bab ini berisi studi pustaka yang membahas beberapa penelitian terkait yang dilakukan sebelumnya, landasan teori, dan metode yang digunakan penulis sebagai acuan mengerjakan penelitian tugas akhir ini.

2.1 Penelitian Terkait

Penelitian mengenai prediksi *Customer Lifetime Value* (CLV) dengan pendekatan *machine learning* telah banyak dilakukan dalam beberapa tahun terakhir. Salah satu adalah penelitian yang dilakukan oleh Cathelijn Kuijt. Penelitian pertama yang dilakukan oleh Cathelijn Kuijt (2022) melalui tesis master berjudul Prediction of *Customer Lifetime Value* in *E-commerce* Fashion Retail (Otrium). Dataset transaksi dari platform Otrium digunakan (2015–2022, difokuskan setelah 2019). Penelitian ini memprediksi CLV 12 bulan ke depan menggunakan model *Frequency prediction* dan *revenue prediction*. Metode yang digunakan adalah *XGBoost, Support Vector Machine (SVM), Random Forest*, serta Multi-Layer Perceptron (MLP). Hasilnya, untuk data Brazil, *Random Forest* memperoleh performa terbaik dengan $R^2 = 0.98$, RMSE = 15.87, dan MAE = 2.60, sedangkan untuk data UK performanya lebih rendah ($R^2 = 0.60$, RMSE ≈ 10.476, MAE ≈ 4.875). Hal ini menunjukkan bahwa kualitas dataset sangat memengaruhi akurasi model (Kuijt, n.d.).

Penelitian keempat dilakukan oleh Christy Davis Maliyekkal (2021) dalam proyek riset MSc berjudul Predicting *Customer Lifetime Value* (CLV) in UK and

Brazil using *Machine learning* and Deep Learning: A Comparative Analysis. Penelitian ini menggunakan dataset *e-commerce* Brazil (100.000 pesanan, Kaggle) dan UK (541.910 transaksi, UCI Repository). Model yang dibandingkan meliputi *Random Forest*, XGBoost, SVM, serta MLP Regressor. Hasilnya menunjukkan bahwa pada dataset Brazil, *Random Forest* mencapai performa terbaik ($R^2 = 0.98$, MAE = 2.60), sementara pada dataset UK performanya menurun ($R^2 = 0.60$, MAE ≈ 4.875). Kelebihan penelitian ini adalah membandingkan berbagai metode ML & DL secara lintas dataset, sedangkan kelemahannya adalah masih adanya ketidakseimbangan distribusi CLV dan perbedaan karakteristik data antar negara (Maliyekkal, n.d.).

Selain penelitian sebelumnya, terdapat juga studi oleh Amit Sharma, Neha Patel, dan Rajesh Gupta (2022) berjudul *Enhancing Customer Lifetime Value Prediction Using Random Forests and Neural Network Ensemble Methods*. Penelitian ini menggunakan dataset retail dari UCI *Machine learning* Repository yang mencakup data transaksi, demografi, dan riwayat belanja pelanggan selama dua tahun (Kabiraj et al., 2018). Metodologi meliputi *preprocessing* (imputasi missing values, transformasi data Min-Max, outlier handling), feature engineering (tenure, *Recency*, *Frequency*, monetary/RFM, serta customer segmentation via K-Means), dan model building. Model utama adalah *Random Forest* Regressor dan Neural Network (MLP), kemudian digabungkan dengan teknik stacking ensemble. Hasilnya menunjukkan bahwa secara individu, *Random Forest* (MAE = 0.45, RMSE = 0.60, $R^2 = 0.72$) dan Neural Network (MAE = 0.43, RMSE = 0.58, $R^2 = 0.74$) sama-sama lebih unggul dibanding model tradisional. Namun, model

ensemble *Random Forest* + Neural Network memberikan hasil terbaik dengan MAE = 0.39, RMSE = 0.53, dan R^2 = 0.78, sekaligus lebih konsisten berdasarkan uji 10-fold cross validation. Analisis *feature importance* menunjukkan bahwa *Recency*, *Frequency*, dan average transaction value adalah faktor penentu CLV paling berpengaruh.

Penelitian terbaru oleh Bakır Tartar (2024) berjudul *Customer Lifetime Value* Prediction and Segmentation Analysis for Commercial Customers in the Banking Industry meneliti prediksi CLV pada sektor perbankan. Hasilnya menunjukkan bahwa *Random Forest* menghasilkan nilai R^2 = 0.68, lebih unggul dibandingkan regresi linear, sekaligus mampu membantu segmentasi pelanggan untuk strategi retensi (Bakir, n.d.-a).

Studi lain dilakukan oleh Subathra & Kumaran (2025) melalui riset berjudul Predictive Modelling of *Customer Lifetime Value* Using AI and *Machine learning* Algorithms. Penelitian ini membandingkan *Random Forest*, XGBoost, SVM, dan LSTM pada data keuangan digital. *Random Forest* terbukti menjadi *baseline* yang solid sebelum dikombinasikan dengan model ensemble dan deep learning (Menaga.A et al., 2025).

Selain itu, penelitian oleh Taherkhani & Daneshvar (2025) dalam jurnal International Journal of Web Research dengan judul Analysis and Optimization of *Customer Lifetime Value* Prediction Using *Machine learning* and Deep Learning Models by RFM Techniques membandingkan berbagai metode, termasuk *Random Forest*, regresi linear, dan deep learning. Hasil menunjukkan bahwa *Random Forest*

mencapai $R^2 = 0.497$, cukup kompetitif meski masih kalah dari model deep learning tertentu, namun lebih unggul dalam hal interpretabilitas (Taherkhani et al., 2025a).

Secara umum, penelitian-penelitian di atas menegaskan bahwa *Random Forest Regression* memiliki performa tinggi dalam prediksi CLV (Aulia, 2022). Perbedaan penelitian ini dibandingkan penelitian sebelumnya adalah fokus tidak hanya pada prediksi, tetapi juga analisis faktor penentu CLV dan perilaku pelanggan.

Tabel 2.1 Ringkasan Penelitian Terdahulu

No	Peneliti (Tahun)	Judul	Metode	Hasil Penelitian
1	Cathelijn Kuijt (2022)	Predicting *Customer Lifetime Value* in Fashion *E-commerce*	*Random Forest*, XGBoost, SVM, MLP	$R^2 = 0.98$, RMSE = 15.87, MAE = 2.60). *Feature importance* menyoroti *Recency* dan *Monetary Value*
2	Christy Davis Maliyekkal (2021)	*Machine learning* Models for CLV Prediction Across Countries	*Random Forest*, XGBoost.	*Random Forest* di Brazil ($R^2 = 0.98$, MAE = 2.60), di UK hanya R^2 = 0.60 *Feature importance*: Variabel *Frequency*.
3	Amit Sharma, Neha Patel & Rajesh Gupta (2022)	Ensemble *Machine learning* Models for *Customer Lifetime Value* Prediction	*Random Forest*, Neural Network (MLP), Stacking Ensemble	(MAE = 0.39, RMSE = 0.53, R^2 = 0.78); *Feature importance*: *Purchase Amount*
5	Subathra & Kumaran (2025)	Hybrid *Machine learning* Models for Predicting CLV in Fintech	*Random Forest*, XGBoost, SVM, LSTM	RF *baseline* terbaik; ensemble RF + LSTM meningkatkan akurasi CLV
6	Taherkhani & Daneshvar (2025)	Comparative Study of ML Models for CLV Prediction	*Random Forest*, Linear Regression, Deep Learning	RF capai R^2 = 0.497,

No	Peneliti (Tahun)	Judul	Metode	Hasil Penelitian
7	Subathra & Kumaran (2025)	Hybrid *Machine learning* Models for Predicting CLV in Fintech	*Random Forest*, XGBoost, SVM, LSTM	RF *baseline* terbaik; ensemble RF + LSTM meningkatkan akurasi CLV
	Prediksi dan Analisis Fakor Penentu *Customer Lifetime Value* (CLV) pada *E-commerce* menggunakan *Random Forest*		*Random Forest*	-

2.2 *Customer Lifetime Value* (CLV)

Customer Lifetime Value (CLV) dalam Bahasa Indonesia ditulis sebagai Nilai Seumur Hidup Pelanggan sesuai dengan Ejaan Yang Disempurnakan (EYD). Yaitu nilai estimasi total keuntungan yang diperoleh perusahaan dari seorang pelanggan selama mereka menjalin hubungan bisnis. CLV menjadi salah satu indikator kunci dalam strategi pemasaran karena mampu menunjukkan seberapa besar kontribusi finansial pelanggan terhadap perusahaan. Dengan memahami nilai CLV, perusahaan dapat menentukan strategi akuisisi, retensi, dan loyalitas pelanggan yang lebih efektif (Bakir, n.d.-b).

Dalam konteks pengambilan keputusan bisnis, CLV berperan penting sebagai dasar dalam mengalokasikan anggaran pemasaran, mengidentifikasi segmen pelanggan bernilai tinggi, serta merancang program loyalitas yang tepat. Misalnya, pelanggan dengan CLV tinggi lebih layak untuk diberi insentif khusus karena potensi keuntungan jangka panjang yang besar, sedangkan pelanggan dengan CLV rendah bisa difokuskan pada strategi efisiensi biaya (D et al., 2023).

Pendekatan tradisional CLV dalam penelitian ini tidak menggunakan model RFM, melainkan formula matematis berbasis *customer value*, *churn rate*, dan *profit margin*. Rumus yang digunakan adalah (Berger & Nasr, 1998) :

a. Rumus CLV

$$CLV = \frac{Customer\ Value}{Churn\ Rate} \times Profit\ Margin \quad (2.1)$$

Persamaan 2.1 berfungsi untuk menghitung yaitu estimasi nilai total keuntungan yang dihasilkan pelanggan selama masa hubungan dengan perusahaan.

b. Customer Value (CV):

$$\mathrm{CV} = \mathrm{AOV} \times \mathrm{F} \quad (2.2)$$

Keterangan:

$$\mathrm{AOV} = \frac{\text{Total Revenue}}{\text{Jumlah Transaksi Individu}} \quad (2.3)$$

$$\mathrm{F} = \frac{\text{Total Transaksi Seluruh Pelanggan}}{\text{Jumlah Pelanggan}} \quad (2.4)$$

Persamaan 2.2 digunakan untuk menghitung Customer Value (CV), yaitu nilai rata-rata pelanggan berdasarkan hasil perkalian antara *Average Order Value (AOV)* dan frekuensi pembelian pelanggan (*F*). Persamaan 2.3 menjelaskan perhitungan AOV, yang menunjukkan rata-rata pendapatan per transaksi dan diperoleh dari total pendapatan dibagi dengan jumlah transaksi individu. Persamaan 2.4 digunakan untuk menghitung F (*Frequency*), yaitu rata-rata jumlah transaksi yang dilakukan seluruh pelanggan.

c. Churn Rate (CR):

$$\text{CR} = 1 - \text{Repeat Rate} \tag{2.5}$$

Keterangan:
Repeat Rate = proporsi pelanggan yang melakukan pembelian lebih dari satu kali.

Persamaan 2.5 berfungsi untuk menentukan Churn Rate (CR), yaitu tingkat pelanggan yang berhenti melakukan pembelian, yang diperoleh dari satu dikurangi nilai *Repeat Rate* atau proporsi pelanggan yang melakukan pembelian lebih dari satu kali.

d. Profit Margin (PM):

$$\text{PM} = \text{Total Revenue} \times \text{Profit Rate}(\%) \tag{2.6}$$

Terakhir, persamaan 2.6 digunakan untuk menghitung Profit Margin (PM)**,** yang menunjukkan persentase keuntungan yang diperoleh dari total pendapatan dengan mengalikan total revenue dengan *Profit Rate (%)*.

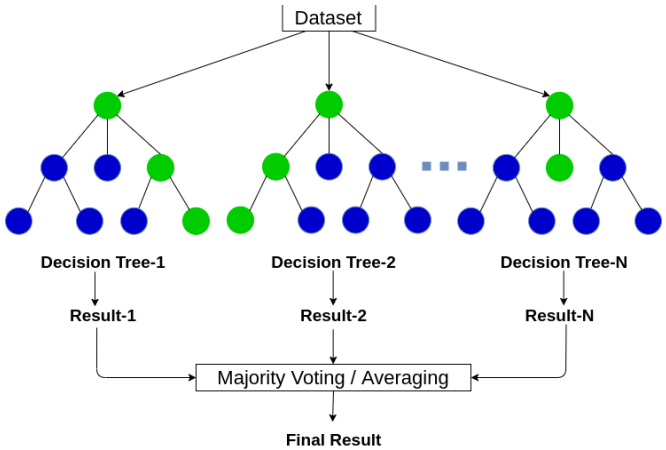
2.3 Prediksi CLV

Prediksi *Customer Lifetime Value* (CLV) merupakan salah satu tantangan utama dalam analisis pelanggan karena melibatkan berbagai faktor yang kompleks, seperti perilaku belanja, loyalitas, dan tingkat retensi pelanggan. Dalam perkembangannya, terdapat dua pendekatan utama yang digunakan, yaitu metode statistik klasik dan metode berbasis *machine learning* (*Master-Thesis-Olivier-Bax-Final-Version*, n.d.). Metode statistik klasik, seperti regresi linear atau regresi logistik, pada dasarnya berupaya menemukan hubungan linier antara variabel-

variabel prediktor dengan CLV. Pendekatan ini sederhana dan mudah diinterpretasikan, tetapi seringkali kurang mampu menangkap hubungan non-linear serta interaksi antar variabel yang kompleks dalam data pelanggan.

2.4 *Random Forest Regression*

Random Forest adalah salah satu algoritma *ensemble learning* yang digunakan secara luas baik untuk klasifikasi maupun regresi. Algoritma ini bekerja dengan membentuk sejumlah besar pohon keputusan (*decision tree*) yang kemudian digabungkan untuk menghasilkan prediksi akhir (Azmi & Voutama, 2024). Pendekatan ini membuat *Random Forest* lebih stabil dan akurat dibandingkan hanya menggunakan satu pohon keputusan, karena hasil akhirnya diperoleh dari kombinasi banyak model (Grömping, 2009)(Raditya, n.d.). Dengan demikian, *Random Forest Regression* merupakan pengembangan khusus dari algoritma ini yang berfokus pada masalah prediksi nilai numerik dengan cara menghitung rata-rata hasil prediksi dari seluruh pohon (Win & Bo, 2020).



Gambar 2.1 Cara Kerja *Random Forest*

Berdasarkan Gambar 2.1, proses kerja *Random Forest* diawali dengan pembentukan beberapa subset data dari dataset asli melalui teknik *bootstrap*

sampling, yaitu pengambilan sampel data secara acak dengan pengembalian. Setiap subset data kemudian digunakan untuk membangun satu pohon keputusan (Balabanova & Bhattarai, n.d.). Pada setiap pohon, pemilihan fitur yang dipertimbangkan dalam pemisahan node dilakukan secara acak, sehingga setiap pohon memiliki struktur yang berbeda. Untuk regresi, pemisahan node biasanya ditentukan dengan kriteria *Mean Squared Error* (MSE) (Wu et al., 2023).

Setelah semua pohon selesai dibangun, masing-masing pohon akan menghasilkan prediksi. Pada regresi, prediksi dari setiap pohon berupa nilai numerik, dan hasil akhir diperoleh dengan menghitung rata-rata (*averaging*) dari semua nilai tersebut (Yan & Resnick, 2024). (Sharma et al., n.d.). Prediksi akhir *Random Forest Regression* (Averaging) diperoleh dengan menghitung rata-rata hasil dari seluruh pohon (Aulia, 2022):

$$\hat{y} = \frac{1}{T} \sum_{t=1}^{T} h_t(x) \tag{2.7}$$

Keterangan:
$\hat{y}$ = hasil prediksi akhir *Random Forest*,
T = jumlah pohon dalam *Random Forest*
$h_t(x)$ = hasil prediksi pohon ke-t.

Persamaan (2.7) merupakan rumus agregasi prediksi pada algoritma *Random Forest Regression*, di mana nilai prediksi diperoleh dengan menghitung rata-rata dari seluruh hasil prediksi yang dihasilkan oleh masing-masing pohon keputusan dalam ensemble.

2.5 Evaluasi Model

Dalam penelitian berbasis regresi, evaluasi model dilakukan dengan menggunakan sejumlah metrik yang dapat menggambarkan tingkat akurasi prediksi secara menyeluruh (Gerde, n.d.).. Salah satu metrik yang umum digunakan adalah *Mean Absolute Error* (MAE), yaitu rata-rata selisih absolut antara nilai aktual dengan nilai prediksi yang dihasilkan model (Billah, n.d.). (Völcker & StenfeltROYAL, n.d.) (Lathwal & Batra, 2024). Sementara itu, *R-Squared* (R^2) atau koefisien determinasi berfungsi untuk menunjukkan proporsi variasi pada variabel dependen yang mampu dijelaskan oleh variabel independen dalam model. Nilai R^2 yang semakin mendekati 1 menandakan bahwa model semakin baik dalam menjelaskan variabilitas data. Penelitian ini menetapkan *Mean Absolute Percentage Error* (MAPE) sebagai metrik evaluasi utama. Pemilihan MAPE didasarkan pada kemampuannya dalam mengukur tingkat kesalahan prediksi secara relatif terhadap nilai aktual dalam bentuk persentase, sehingga lebih mudah diinterpretasikan dan lebih relevan dalam konteks bisnis *e-commerce* seperti yang ditampilkan pada Tabel 2.2 (Lathwal & Batra, 2024) (Ardian et al., n.d.).

Tabel 2.2 Kategori Hasil Evaluasi Model

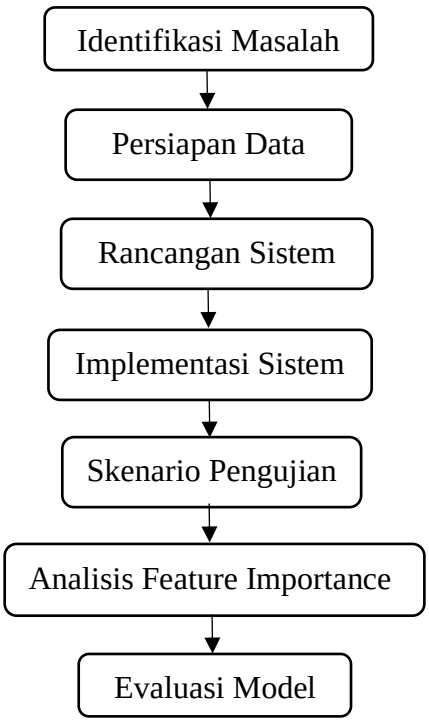
Metrik	Rentang Nilai	Kategori
MAE	Mendekati 0	Sangat Baik
	Kecil–Sedang	Baik
	Besar	Kurang Baik
MSE	Mendekati 0	Sangat Baik
	Sedang	Baik
	Tinggi	Buruk
MAPE	< 10%	*Highly Accurate Forecast*
	10% – 20%	*Good Forecast*
	20% – 50%	*Reasonable Forecast*
	> 50%	*Poor Forecast*
R^2	0.90 – 1.00	Sangat Baik
	0.70 – 0.89	Baik
	0.50 – 0.69	Sedang
	< 0.50	Lemah

BAB III

DESAIN DAN IMPLEMENTASI

3.1 Desain Penelitian

Desain penelitian disusun untuk memberikan panduan yang jelas dalam pelaksanaan penelitian, sehingga setiap tahapan dapat dilakukan secara sistematis dan hasil yang diperoleh sesuai dengan tujuan yang telah ditetapkan. Penelitian ini dirancang untuk memprediksi serta menganalisis faktor-faktor yang menentukan *Customer Lifetime Value* (CLV) pada data *e-commerce* dengan algoritma *Random Forest Regression*. Proses penelitian ini ditampilkan pada Gambar 3.1.



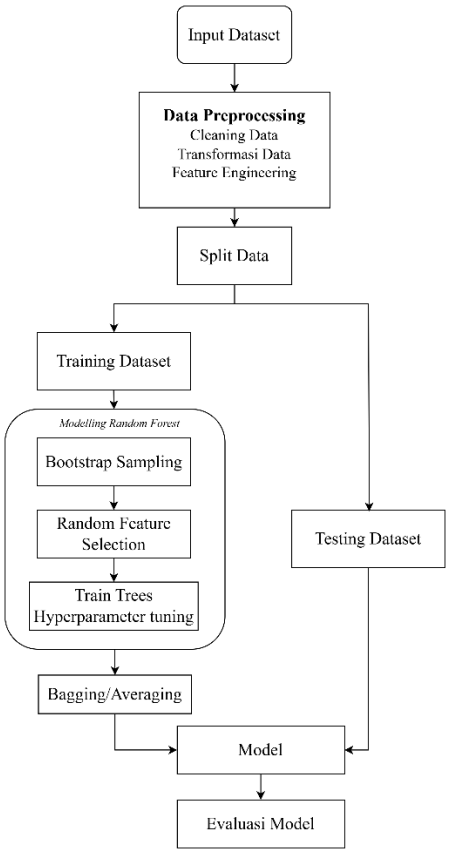
Gambar 3.1 Desain Penelitian

Pada Gambar 3.1 penelitian ini dibuat untuk memberikan gambaran yang jelas dan terstruktur mengenai langkah-langkah yang dilakukan dalam proses

penelitian. Dengan adanya desain penelitian, setiap tahapan dapat dijalankan secara sistematis mulai dari pengumpulan data hingga evaluasi akhir model. Desain penelitian ini juga bertujuan untuk memastikan bahwa proses pengolahan data, hingga penerapan algoritma *Random Forest Regression* dapat dilakukan dengan alur yang logis, konsisten, dan dapat direplikasi oleh peneliti lain.

3.2 Desain Sistem

Desain sistem penelitian ini menggambarkan tahapan implementasi yang dilakukan untuk memprediksi dan menganalisis faktor-faktor penentu *Customer Lifetime Value* (CLV) menggunakan algoritma *Random Forest Regression*. Tahapan dalam desain sistem dapat dilihat pada gambar berikut:



Gambar 3.2 Desain Sistem Penelitian

Desain sistem penelitian ini, sebagaimana ditunjukkan pada Gambar 3.2, dimulai dari input dataset dan tahap *preprocessing* berupa pembersihan, transformasi, serta normalisasi data agar siap digunakan dalam pemodelan. Dataset kemudian dibagi menjadi data latih dan data uji untuk menghindari *overfitting* serta memastikan evaluasi yang objektif. Pada data latih, algoritma *Random Forest Regression* dibangun melalui *bootstrap sampling*, *random feature selection*, dan pembentukan pohon keputusan. Prediksi tiap pohon digabung dengan bagging/averaging**,** kemudian dilakukan *hyperparameter* tuning dengan *RandomizedSearchCV* untuk memperoleh konfigurasi model terbaik (Bergstra & Bengio, 2012). Tahap akhir adalah evaluasi menggunakan data uji hingga terbentuk model final yang menghasilkan prediksi *Customer Lifetime Value* (CLV) sekaligus analisis faktor dominan melalui *feature importance*.

3.3 Persiapan Data

Data yang digunakan dalam penelitian ini bersifat sekunder dan diambil dari platform Kaggle dengan judul Retail Store Sales Transaction yang dapat diakses melalui tautan berikut: https://www.kaggle.com/datasets/marian447/retail-store-sales-transactions/data. Dataset ini dipilih karena memiliki atribut yang lengkap dan relevan untuk analisis perilaku pelanggan *e-commerce*, mencakup aspek demografis, pola transaksi, interaksi dengan layanan, hingga respons terhadap promosi. Karakteristik data tersebut sangat sesuai dengan kebutuhan penelitian ini, khususnya dalam perhitungan manual *Customer Lifetime Value* (CLV) dan pemodelan prediksi menggunakan algoritma machine learning. Dengan adanya variabel transaksi seperti *Customer_ID*, *Total_Sales*, *Quantity* nilai CLV dapat

dihitung secara akurat dan memungkinkan analisis lebih lanjut mengenai fitur berpengaruh pada CLV.

Dataset ini merepresentasikan 131.000 data transaksi pelanggan ritel selama tahun 2016. Dataset ini terdiri atas 8 kolom utama, yaitu *Date, Customer_ID, Transaction_ID, SKU_Category*, *SKU, Quantity, dan Sales_Amount*. Variabel-variabel yang ada di dalam dataset mencakup:

Tabel 3.1 Penjelasan Atribut Dataset

Variabel	Deskripsi	Tipe Data
Date	Tanggal terjadinya transaksi pembelian	*Date*time
Customer_ID	Identitas unik pelanggan	string / integer
Transaction_ID	Nomor unik transaksi	string / integer
SKU_Category	Kategori produk atau item yang dibeli	string
SKU	Kode unik produk	string
Quantity	Jumlah unit produk yang dibeli dalam transaksi	integer
Total_Sales	Nilai total penjualan pada transaksi tersebut	float / numeric

Aktivitas transaksi pelanggan dalam Tabel 3.1 direpresentasikan melaluiTabel 3.1 yaitu atribut *Total_Sales* dan *Quantity* yang digunakan sebagai dasar perhitungan *Customer Lifetime Value (CLV)*. Nilai *Total_Sales* menggambarkan total pendapatan dari setiap transaksi pelanggan, dan *Quantity* menunjukkan jumlah produk yang dibeli dalam satu transaksi. Kombinasi keempat atribut ini memberikan informasi mengenai perilaku pembelian pelanggan, tingkat aktivitas transaksi, serta kontribusi pendapatan terhadap perusahaan.

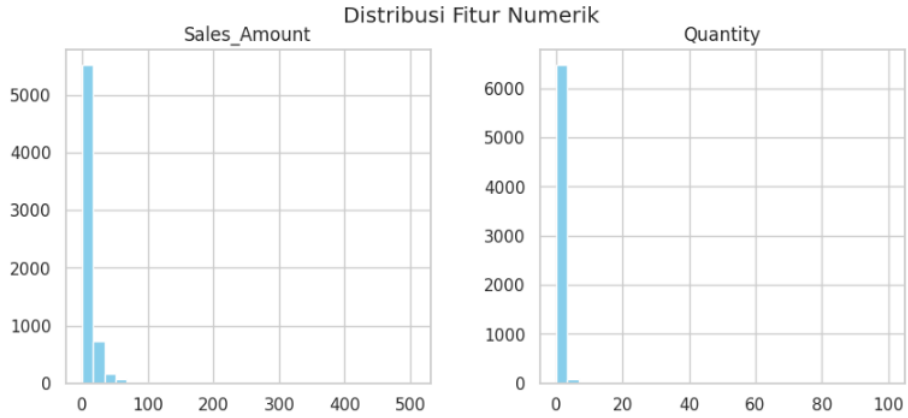
3.4 *Exploratory Data Analysis* (EDA)

Exploratory Data Analysis (EDA) dilakukan untuk menilai kualitas dan karakteristik data sebelum memasuki proses pemodelan. Berikut hasil dari pmeriksaan awal:

Tabel 3.2 Hasil EDA

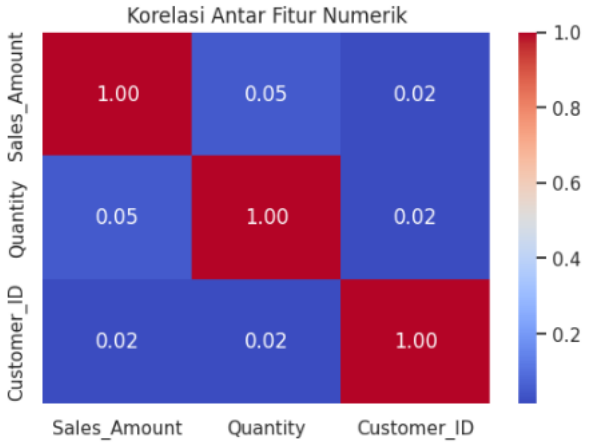
Aspek Pemeriksaan	Keterangan
Missing Values per Kolom	0 missing values.
Jumlah Data Duplikat	0 data duplikat ditemukan.
Nilai Quantity Negatif atau Nol	Tidak ditemukan nilai Quantity ≤ 0.
Nilai Sales_Amount Negatif	Tidak ditemukan nilai Sales_Amount < 0.

Pada Tabel 3.2 pemeriksaan awal terhadap sampel data sebesar 80% menunjukkan bahwa tidak terdapat nilai kosong (*missing values*) maupun data duplikat, sehingga dataset dapat langsung digunakan tanpa proses imputasi. Selain itu, hasil pemeriksaan tipe data menunjukkan bahwa kolom Quantity, dan *Total_Sales* telah memiliki format yang sesuai untuk analisis.



Gambar 3.3 Distribusi Fitur Numerik

Berdasarkan Gambar 3.3 yang menunjukan statistik deskriptif ,nilai Quantity rata-rata menunjukkan pelanggan membeli 1–2 unit per transaksi, sedangkan *Total_Sales* berkisar antara Rp5.000–Rp7.000 dengan sebaran data *right-skewed* yang umum pada data ritel.



Gambar 3.4 Korelasi Antar Fitur

Berdasarkan hasil visualisasi pada Gambar 3.4, nilai korelasi antar fitur berada pada rentang 0.02 hingga 0.05, yang menandakan hubungan antar variabel sangat lemah. Korelasi antara *Sales_Amount* dan *Quantity* sebesar 0.05 menunjukkan bahwa jumlah produk yang dibeli pelanggan tidak selalu berbanding lurus secara kuat dengan total nilai transaksi, kemungkinan disebabkan oleh variasi harga produk. Secara keseluruhan, hasil *EDA* menunjukkan bahwa dataset memiliki kualitas yang baik, dengan distribusi data yang wajar dan hubungan antarvariabel yang logis. Temuan ini memperkuat pemilihan empat fitur hasil *feature engineering* (*Total_Sales*, *Frequency*, *Recency*, dan *Total_Quantity*) sebagai variabel utama dalam pemodelan *Random Forest Regression* untuk memprediksi *Customer Lifetime Value (CLV)* pelanggan.

3.5 *Pre-processing Data*

Tahap pre-processing data dilakukan untuk memastikan dataset siap digunakan dalam proses pemodelan *Random Forest Regression*. Pemeriksaan awal menunjukkan bahwa tidak terdapat nilai hilang dan nilai duplikat sehingga setiap

transaksi dapat diidentifikasi secara unik. Pada tahap awal penelitian, dataset yang digunakan masih berada dalam bentuk data transaksi mentah (*raw transactional data*) yang merepresentasikan aktivitas pembelian pelanggan pada platform *e-commerce* dengan total data (131.706 baris × 8 kolom).

3.5.1 Transformasi Data

Tahap transformasi data bertujuan untuk mengubah format variabel agar sesuai dengan kebutuhan pemodelan. Proses ini terdiri dari pengubahan tipe data serta perhitungan *Customer Lifetime Value* (CLV) manual berdasarkan variabel transaksi dan feature engineering.

1) Perubahan Tipe Data *(Data Type Conversion)*

Beberapa atribut pada dataset perlu disesuaikan formatnya agar dapat diolah dengan benar oleh model:

a. Kolom *Date* diubah kedalam format *datetime*
b. Kolom *Customer_ID* disimpan sebagai tipe string untuk menjaga identitas unik setiap pelanggan
c. Kolom *Total_Sales* dan *Quantity* dikonversi menjadi tipe float agar dapat digunakan dalam perhitungan nilai total penjualan dan rata-rata transaksi pelanggan.

2) Perhitungan *Customer Lifetime Value* (CLV) Manual

Perhitungan CLV manual yang akan digunakan untuk perhitungan target (*y*) berdasarkan rumus dari (Berger & Nasr, 1998) dilakukan dengan beberapa tahap:

a. Average Order Value (AOV)

Berdasarkan persamaan 2.3, AOV berfungsi untuk menghitung rata-rata nilai transaksi yang dilakukan oleh pelanggan. Nilai ini diperoleh dengan perhitungan berikut:

$$\text{AOV} = \frac{3.65+8.21}{2} = 5.93$$

b. Purchase *Frequency* (F)

Mengacu pada persamaan 2.4 hasil F dapat dihitung dari total dataset dibagi total angka unik seperti contoh perhitungan dibawah ini:

$$\text{F} = \frac{10}{9} = 1.11$$

c. *Customer Value* (CV)

Berdasarkan persamaan 2.2, *Customer Value (CV)* dihasilkan oleh perhitungan pelanggan berdasarkan rata-rata nilai transaksi (*AOV*) dan frekuensi pembelian (*F*).

$$\text{CV} = 5.93 \times 1.11 = 6.58$$

d. *Repeat Rate* (RR)

Mengacu pada persamaan 2.5, *Repeat Rate (RR)* digunakan untuk mengukur proporsi pelanggan yang melakukan pembelian lebih dari satu kali.

$$\text{RR} = \frac{1}{9} = 0.11$$

e. Churn Rate (CR)

Berdasarkan persamaan 2.5, *Churn Rate (CR)* menunjukkan tingkat pelanggan yang berhenti melakukan pembelian, yang diperoleh dari 1 dikurangi nilai *Repeat Rate*. Seperti pada contoh perhitungan dibawah ini:

$$\mathrm{CR} = 1 - 0.11 = 0.89$$

f. Profit Margin (PM)

Berdasarkan persamaan 2.6, *Profit Margin (PM)* dapat dihitung dengan contoh perhitungan dibawah ini:

Total Revenue = Rp68,47

Profit Rate = 0.2 (20%)

$$\mathrm{PM} = 68{,}47 \times 0.2 = 13{,}69$$

g. CLV per pelanggan

Berdasarkan persamaan 2.1, Nilai CLV diperoleh dari hasil pembagian antara *Customer Value (CV)* dengan *Churn Rate (CR)*, kemudian dikalikan dengan *Profit Margin (PM)*.

$$CLV = \frac{6.58}{0.89} \times 13.69 = 101.24$$

3.5.2 Feature Engineering

Tahap *feature engineering* dilakukan untuk membentuk variabel-variabel baru yang lebih representatif terhadap perilaku pelanggan, sehingga dapat digunakan sebagai fitur input pada pemodelan *Random Forest Regression*. Empat fitur utama yang dihasilkan dari proses *feature engineering* adalah sebagai berikut:

Tabel 3.3 Fitur Penelitian untuk Pemodelan *Random Forest Regression*

No	Nama Fitur	Deskripsi	Jenis Data
1	*Total_Sales*	Total nilai pembelian yang dilakukan oleh setiap pelanggan selama periode observasi.	Float / Numeric
2	*Frequency*	Jumlah transaksi unik yang dilakukan oleh setiap pelanggan selama periode waktu tertentu.	Integer
3	*Recency*	Selisih waktu (dalam hari) antara tanggal transaksi terakhir pelanggan dengan tanggal observasi terakhir.	Integer
4	*Total_Quantity*	Total unit produk yang dibeli oleh setiap pelanggan selama periode pengamatan.	Integer

Tabel 3.3 fitur penelitian menunjukkan atribut-atribut yang digunakan dalam pemodelan *Random Forest Regression* untuk memprediksi nilai *Customer Lifetime Value* (CLV) (Fader et al., 2019). Dengan memanfaatkan keempat variabel tersebut, nilai CLV dapat dihitung secara komprehensif untuk menggambarkan kontribusi finansial setiap pelanggan terhadap perusahaan. (Taherkhani et al., 2025b). Data setelah ditransformasi menjadi 22.588 baris yang disebabkan oleh proses agregasi transaksi ke tingkat pelanggan**,** di mana *Sales_Amount* dan *Quantity* digabung menjadi total per pelanggan (Gupta & Lehmann, 2021). Transformasi ini menghasilkan data yang lebih ringkas serta sesuai untuk analisis CLV. Berikut jumlah data setelah ditransformas seperti pada Gambar 3.5:

	Total_Sales	Frequency	Total_Quantity	Recency
count	22588.000000	22588.000000	22588.000000	22588.000000
mean	69.653161	2.856163	8.654153	161.084248
std	152.155098	3.996433	20.991088	115.898208
min	0.140000	1.000000	1.000000	0.000000
25%	10.127500	1.000000	2.000000	47.000000
50%	23.800000	1.000000	3.000000	149.000000
75%	63.052500	3.000000	8.000000	271.000000
max	3949.550000	99.000000	814.000000	364.000000

Gambar 3.5 Jumlah Data setelah Transformasi

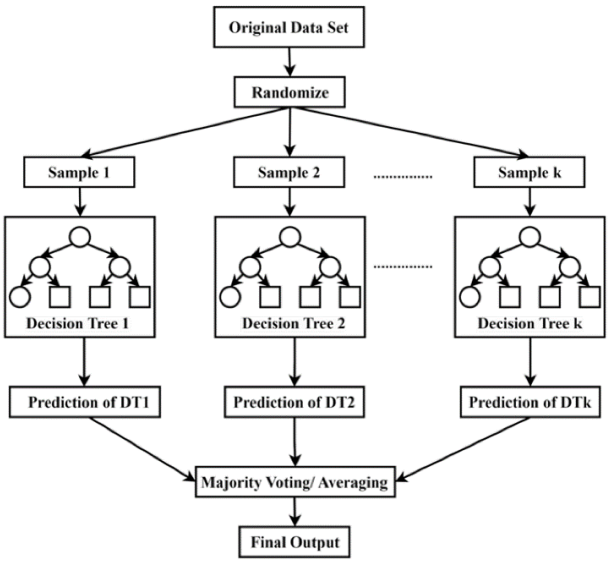
3.6 Split Data

Setelah data dinormalisasi, tahap selanjutnya adalah pembagian dataset (train- test split) agar model dapat dilatih dan diuji secara adil.

a. Data latih (training set): 80% → digunakan untuk melatih model *Random Forest Regression*.
b. Data uji (testing set): 20% → digunakan untuk mengevaluasi performa model pada data baru yang tidak dilihat saat pelatihan.

3.7 Implementasi *Random Forest Regression*

Implementasi *Random Forest* pada penelitian ini yaitu untuk memprediksi nilai CLV pada setiap Customer berdasarkan variabel/fitur yang telah diproses sebelumnya.



Gambar 3.6 Implementasi *Random Forest*

Berdasarkan Gambar 3.6, *Random Forest* merupakan metode *ensemble learning* dalam *machine learning* yang menyatukan beberapa decision tree guna

meningkatkan akurasi dan stabilitas prediksi (Edi, n.d.). Setiap pohon dalam *Random Forest* menghasilkan hasil prediksi masing-masing. Hasil akhir prediksi ditentukan berdasarkan rata-rata dari seluruh pohon, sehingga algoritma ini lebih kuat terhadap *overfitting* dan lebih stabil dibandingkan hanya menggunakan satu decision tree. Model ini bekerja dengan beberapa tahapan, antara lain:

1) *Bootstrap sampling*

Dari dataset asli berukuran N, *Random Forest* membuat beberapa subset data dengan metode pengambilan acak dengan pengembalian. Setiap subset memiliki ukuran sama dengan dataset asli (N), tetapi bisa mengandung baris yang sama berulang kali. Peluang suatu sampel tidak terambil dalam 1 kali pengambilan:

$$P(tidak\ terambil) = 1 - \frac{1}{N} \tag{3.1}$$

Pada tahap ini, algoritma *Random Forest* membentuk beberapa subset data (dalam contoh ini B = 3) dari dataset asli dengan ukuran yang sama (N), menggunakan metode sampling with replacement. Artinya, setiap baris data dapat terambil lebih dari satu kali dalam satu subset, dan beberapa data lain mungkin tidak terambil sama sekali. Contoh hasil *bootstrap sampling* dari data transaksi yang berisi 10 baris ditunjukkan sebagai berikut:

Tabel 3.4 Contoh Booststrap Sampling

Subset	**Indeks Data yang Terpilih**	**Nilai *Total_Sales* pada Subset**
Bootstrap 1	[1, 2, 3, 3, 5, 6, 7, 7, 9, 10]	3.13, 5.46, 6.35, 6.35, 6.88, 10.77, 3.65, 3.65, 8.25, 8.18
Bootstrap 2	[2, 2, 4, 5, 5, 6, 8, 8, 10, 1]	5.46, 5.46, 5.59, 6.88, 6.88, 10.77, 8.21, 8.21, 8.18, 3.13
Bootstrap 3	[3, 4, 4, 5, 6, 7, 8, 9, 9, 2]	6.35, 5.59, 5.59, 6.88, 10.77, 3.65, 8.21, 8.25, 8.25, 5.46

Sebagaimaan yang ditunjukan pada Tabel 3.4, proses *bootstrap* dilakukan *with replacement*, beberapa baris muncul lebih dari satu kali, seperti data ke-3 yang muncul dua kali pada *Bootstrap 1*. Tahap ini menghasilkan variasi data yang membantu mengurangi risiko *overfitting* pada model.

2) Random Feature Selection

Ketika model membagi sebuah node, tidak semua fitur digunakan untuk mencari split terbaik. Hanya sebagian (subset) fitur acak yang dipilih. Aturan default di scikit-learn (dan juga banyak literatur) adalah:

$$M = \frac{n}{3} \tag{3.2}$$

Keterangan:
M = *Feature Selection*
n = jumlah total fitur.

Pada Persamaaan 3.2, algoritma setiap pohon hanya menggunakan split data dari subset acak dari fitur yang tersedia. Pemilihan acak fitur ini bertujuan agar setiap pohon memiliki karakteristik berbeda, sehingga ketika digabungkan menghasilkan model yang lebih stabil dan generalisasi yang baik.

3) Decision Tree

Setiap subset hasil bootstrap digunakan untuk melatih satu *Decision Tree Regression*. Untuk penyederhanaan ilustrasi, pada penelitian ini ditunjukkan contoh proses pembentukan pohon dengan satu kali pembagian (split).

a. Tree 1

Jumlah data pelatihan: 10 data transaksi (hasil *Bootstrap 1*)

Fitur terpilih: *Total_Sales*

Threshold (nilai pemisah): *Total_Sales* ≤ 6.0

Pohon pertama membagi data transaksi berdasarkan nilai penjualan (*Total_Sales*) menjadi dua kelompok:

1) Node kiri: berisi transaksi dengan *Total_Sales* ≤ 6.0 Terdapat 5 data transaksi pada kelompok ini dengan nilai rata-rata penjualan sebesar 5.63.
2) Node kanan: berisi transaksi dengan *Total_Sales* > 6.0 Terdapat 5 data transaksi dengan rata-rata nilai penjualan sebesar 6.90.

Jika terdapat pelanggan dengan nilai transaksi *Total_Sales* = 8.0, maka data tersebut termasuk ke dalam node kanan, sehingga prediksi dari pohon pertama adalah 6.90.

4) Bagging (Bootstrap Aggregating)

Rumus untuk Regresi:

$$\hat{f}_{RF}(x) = \frac{1}{B}\sum_{b=1}^{B} \hat{f}_b(x) \tag{3.3}$$

Keterangan
$\hat{f}_{RF}(x)$ = Prediksi akhir CLV untuk pelanggan x
B = Jumlah Pohon
$\hat{f}_b(x)$ = Prediksi dari Pohon ke- b

Persamaan 3.3 merupakan tahap mengimplementasikan model *Random Forest Regression* untuk memprediksi *Customer Lifetime Value* (CLV)

berdasarkan variabel transaksi yang terdiri dari *Customer_ID***,** *Total_Sales* **,** *dan Quantity*. Dengan demikian, nilai CLV yang diprediksi jika pelanggan tersebut memiliki lebih dari satu transaksi, Nilai CLV dikalikan dengan jumlah transaksi pelanggan karena setiap transaksi mewakili kontribusi ekonomi yang berbeda dalam periode hidup pelanggan. *Customer Lifetime Value* dapat dihitung secara kumulatif dari seluruh transaksi yang dilakukan pelanggan selama masa aktifnya (Fader et al., 2010).

3.8 Parameter *Randomized Search CV*

RandomizedSearchCV adalah metode optimasi *hyperparameter* yang digunakan untuk menemukan kombinasi parameter terbaik dalam algoritma machine learning, termasuk *Random Forest Regression*. *RandomizedSearchCV* hanya memilih sejumlah kombinasi secara acak dari ruang parameter yang telah ditentukan. Proses pemilihan acak ini membuat *RandomizedSearchCV* lebih efisien secara waktu dan komputasi, terutama pada dataset yang besar dengan banyak parameter, namun tetap mampu menghasilkan kombinasi *hyperparameter* yang mendekati optimal.

Beberapa *hyperparameter* utama dalam *Random Forest* antara lain:

1. ***n_estimators***

   Jumlah pohon yang dibangun dalam *Random Forest*. Semakin banyak pohon, semakin stabil prediksi karena variasi berkurang. Namun, jumlah pohon yang terlalu besar dapat meningkatkan waktu komputasi.

2. ***max_depth***

   Kedalaman maksimum setiap pohon. Pohon yang terlalu dalam dapat menangkap noise yang mengganggu proses pembelajaran model, tetapi tetap "dipelajari" oleh model ketika pohon keputusan dibuat terlalu dalam.

3. ***min_samples_leaf***

   Jumlah minimum sampel pada daun pohon (*leaf*). Parameter ini mencegah pohon memiliki daun dengan jumlah sampel yang sangat sedikit, sehingga memperbaiki generalisasi.

4. ***max_features***

   Jumlah fitur yang dipertimbangkan saat menentukan pemisahan pada setiap node. Pemilihan subset fitur secara acak meningkatkan keragaman pohon dan menurunkan korelasi antar pohon. Contoh Hasil Tuning

   Berdasarkan hasil *RandomizedSearchCV*, diperoleh kombinasi parameter terbaik sebagai berikut:

   a. n_estimators = 500 - 1000
   b. max_depth = 4,3,2
   c. min_samples_leaf = 2
   d. max_features = 2,3

3.9 Evaluasi Model

Model Regresi dievaluasi dengan beberapa metrik. Rumus ini mengacu pada konsep regresi dalam (Chicco et al., 2021) dan digunakan untuk mengukur perbedaan antara nilai aktual dan hasil prediksi.

3.9.1 *Mean Absolute Error* (MAE)

Mean Absolute Error (MAE) adalah presentase untuk mengukur rata-rata kesalahan absolut antara prediksi model dan nilai sebenarnya.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{3.4}$$

Keterangan:
n =jumlah data uji
y_i = nilai aktual
$\hat{y}_i$= nilai prediksi

Persamaan 3.4 dihitung dengan menjumlahkan selisih absolut antara nilai aktual dan nilai prediksi untuk seluruh data uji sebanyak nnn, kemudian dibagi dengan jumlah data tersebut.

3.9.2 *Mean Squared Error* (MSE)

Mean Squared Error (MSE) adalah metrik yang digunakan untuk mengukur rata-rata kesalahan kuadrat antara nilai yang diprediksi oleh model dan nilai sebenarnya. MSE dihitung dengan rumus berikut:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{3.5}$$

Keterangan:

n =jumlah data uji
y_i = nilai aktual
$\hat{y}_i$= nilai prediksi

Persamaan 3.5 merupakan rumus yang digunakan untuk mengukur rata-rata kesalahan kuadrat antara nilai prediksi model dan nilai aktual pada data uji. Semakin kecil nilai MSE, semakin baik model dalam melakukan prediksi.

3.9.3 *Mean Absolute Percentage Error* (MAPE)

Mean Absolute Percentage Error (MAPE) adalah metrik evaluasi yang digunakan untuk mengukur rata-rata kesalahan prediksi dalam bentuk persentase terhadap nilai aktual.

$$MAPE = \frac{100\%}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \quad (3.6)$$

Keterangan:
n =jumlah data uji
y_i = nilai aktual
$\hat{y}_i$= nilai prediksi

Persamaan 3.6 digunakan untuk menghitung selisih absolut antara nilai aktual dan nilai prediksi terhadap nilai aktual, kemudian dirata-ratakan untuk seluruh data uji sebanyak n dan dikalikan 100%.

3.9.4 R^2 Squared (R^2)

R-Squared (R^2) atau koefisien determinasi adalah metrik yang menunjukkan sejauh mana model mampu menjelaskan variasi dalam data target. R^2 dihitung dengan rumus:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad (3.7)$$

Keterangan:
y_i = nilai aktual,
$\hat{y}_i$= nilai prediksi,
$\bar{y}$ = nilai rata-rata dari nilai aktual.

Persamaan 3.7 dihitung dengan membandingkan jumlah kuadrat galat prediksi terhadap total variasi data aktual. Nilai R^2 berada pada rentang 0 hingga 1,

3.10 *Feature importance*

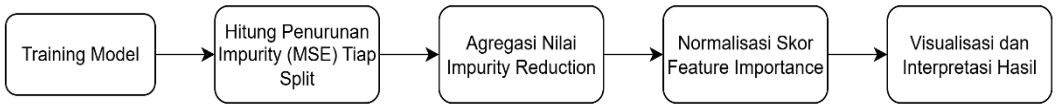
Setelah model dilatih, *Random Forest* menghitung tingkat kepentingan (importance) fitur berdasarkan:

$$FI(j) = \frac{1}{B}\sum_{b=1}^{B} \Delta Impurity_j^{(b)} \text{ (Hwang et al., 2023)} \quad (3.8)$$

Keterangan:
$FI(j)$ = *Feature importance* untuk fitur ke-**j** (tingkat kepentingan fitur j terhadap model *Random Forest*).
B = Jumlah total pohon (*trees*)
B = Indeks pohon ke-b
$\Delta Impurity_j^{(b)}$ = Penurunan impurity (ketidakmurnian)

Persamaan 3.8 merupakan metode analisis dalam *machine learning* yang digunakan untuk mengetahui seberapa besar kontribusi setiap variabel (fitur) terhadap hasil prediksi model (You, 2025). Pada model *Random Forest*, tingkat kepentingan fitur dihitung berdasarkan penurunan impurity yang terjadi pada setiap proses pembagian node di seluruh pohon keputusan.

Untuk kasus regresi, impurity diukur menggunakan *Mean Squared Error* (MSE). Setiap kali terjadi pemisahan (split) data menggunakan fitur tertentu, nilai MSE akan berkurang dan besarnya penurunan tersebut menunjukkan kontribusi fitur terhadap performa model. (Husna, 2020).



Gambar 3.7 Alur Feature importance

Pada Gambar 3.7 penentuan *feature importance* dalam *Random Forest Regression* dilakukan melalui serangkaian tahapan sistematis yang bertujuan untuk mengukur kontribusi masing-masing fitur terhadap performa model (Yılmaz Benk et al., 2022). Proses dimulai dengan perhitungan berikut:

3.10.1 Penurunan Impurity MSE

Penuruan Impurity:

$$\Delta MSE_{Tree3} = MSE_{Parent} - \left(\frac{n_L}{n_T} MSE_L + \frac{n_R}{n_T} MSE_R\right) \quad (3.9)$$

Persaman 3.9 menunjukkan besarnya penurunan impurity (ketidakmurnian) yang dihasilkan ketika sebuah node induk (parent) dibagi menjadi dua node anak, yaitu node kiri (L) dan node kanan (R).

3.10.2 Agregasi Nilai Impurity 3 Tree

Perhitungan Nilai Impurity dari 3 pohon dapat dihitung dengan perhitungan dibawah ini:

Total $\Delta MSE_{Sales\_Amount} = MSE_1 + MSE_2 + MSE_3$

3.10.3 Hitung *Feature importance* (raw/mentah)

Rumus :

$$FI(Sales\_Amount) = \frac{1}{B}\sum_{b=1}^{B} \Delta MSE^{(b)}_{Sales\_Amount} \quad (3.10)$$

Persamaan 3.10 menunjukkan bahwa nilai *Feature Importance* dihitung dengan mengambil rata-rata total penurunan *Mean Squared Error* (ΔMSE) yang

dihasilkan oleh fitur *Sales_Amount* di seluruh jumlah pohon (*B*) pada model *Random Forest*.

3.10.4 Normalisasi Skor *Feature importance*

$$FI_{norm}(j) = \frac{\sum_b \Delta MSE_{j,b}}{\sum_k \sum_b \Delta MSE_{k,b}} \quad (3.11)$$

Keterangan:

$\sum_b \Delta MSE_{j,b} =$ total penurunan nilai Mean Squared Error (MSE).

$\sum_k \sum_b \Delta MSE_{k,b}$= total keseluruhan penurunan MSE dari semua fitur (*k*) pada seluruh pohon (*b*)

Persamaan 3.11 merupakan perhitungan Feature Importance (FI) yang telah dinormalisasi pada model *Random Forest Regression*, dengan tujuan untuk mengukur tingkat kontribusi relatif setiap fitur terhadap performa model.

3.10.5 Interpretasi Singkat

- *Total_Sales* jelas kontributor utama terhadap penurunan impurity model (dalam contoh ini semua split yang menurunkan MSE memakai *Total_Sales* di Tree1 dan Tree3).
- FI_raw ≈ 2.0375 menunjukkan rata-rata pengurangan MSE per pohon ketika *Total_Sales* digunakan untuk split.
- Jika kita hanya pakai Δ yang diketahui, *Total_Sales* menyumbang 100% dari total penurunan impurity yang diamati → artinya: pada contoh ini *Total_Sales* adalah fitur paling penting**.**
- Namun, ini bisa berubah bila Customer_ID punya ΔMSE > 0 pada Tree2. Jadi klaim “100%” hanya benar untuk data Δ yang tersedia sekarang jangan klaim final tanpa memasukkan kontribusi Customer_ID.

3.11 Skenario Pengujian

Skenario Skenario pengujian dalam penelitian ini dirancang untuk mengevaluasi pengaruh setiap fitur hasil *feature engineering* terhadap performa model *Random Forest Regression* dalam memprediksi *Customer Lifetime Value (CLV)*. Tujuan utama dari pengujian ini adalah untuk mengetahui seberapa besar penurunan akurasi model ketika fitur-fitur tertentu dihapus, serta menilai kontribusi relatif tiap fitur terhadap hasil prediksi CLV.

Tabel 3.5 Skenario Pengujian

Skenario		**n_ estimators**	**Fitur yang Digunakan**	**Proporsi Data (Train:Test)**	**Indikator Evaluasi**
Skenario 0 (*Baseline*)	a	500	*Total_Sales*, *Frequency*, *Recency*, *Total_Quantity*	80% : 20%	MSE, MAE, MAPE, R^2, Feature Importance
	b	1000			
Skenario 1	a	500	*Total_Sales*, *Frequency*, *Recency*	80% : 20%	MSE, MAE, MAPE, R^2, Feature Importance
	b	1000			
Skenario 2	a	500	*Total_Sales*, *Recency*	80% : 20%	MSE, MAE, MAPE, R^2, Feature Importance
	b	1000			
Skenario 3	a	500	*Total_Sales*, *Frequency*	80% : 20%	MSE, MAE, MAPE, R^2, Feature Importance
	b	1000			

Pada Tabel 3.5, skenario pengujian dirancang untuk mengevaluasi pengaruh setiap fitur hasil feature engineering terhadap performa model *Random Forest Regression* dalam memprediksi nilai *Customer Lifetime Value* (CLV). Setiap skenario disusun melalui proses penghapusan fitur secara bertahap agar dapat diketahui kontribusi masing-masing variabel terhadap akurasi model. Pada Skenario 0 (*Baseline*), seluruh fitur digunakan, yaitu *Total_Sales*, *Frequency*,

Recency, dan *Total_Quantity*. Skenario ini dijadikan acuan utama untuk melihat performa model dalam kondisi paling lengkap sebelum dilakukan pengurangan fitur.

Pada Skenario 1, fitur *Total_Quantity* dihapus sehingga model hanya menggunakan *Total_Sales*, *Frequency*, dan *Recency*. Penghapusan fitur ini dilakukan karena *Total_Quantity* sering memiliki korelasi yang tinggi dengan *Total_Sales*, mengingat keduanya sama-sama menggambarkan volume pembelian pelanggan. Skenario ini bertujuan untuk menguji apakah pengurangan fitur yang bersifat redundan dapat menyederhanakan model tanpa menurunkan performa prediksi..

Pada Skenario 2, fitur *Frequency* dihapus, sehingga model hanya menggunakan *Total_Sales* dan *Recency*. Pemilihan skenario ini sangat penting karena *Frequency* merupakan salah satu penentu utama *Customer Lifetime Value* berdasarkan teori RFM yang umum digunakan dalam analisis perilaku pelanggan. Dengan menghilangkan variabel ini, dilakukan evaluasi terhadap sejauh mana model masih dapat memprediksi CLV hanya dengan menggunakan faktor nilai transaksi dan waktu terakhir pelanggan melakukan pembelian.

Selanjutnya, pada Skenario 3, fitur *Recency* yang dihapus, sehingga model hanya menggunakan *Total_Sales* dan *Frequency*. Tujuan dari skenario ini adalah untuk menguji apakah faktor waktu terakhir pelanggan bertransaksi masih memberikan kontribusi penting terhadap prediksi CLV jika model hanya mengandalkan variabel finansial dan perilaku transaksi utama.

BAB IV

HASIL DAN PEMBAHASAN

4.1 Hasil Penerapan Model *Random Forest*

Pada tahap ini akan menjelaskan hasil dari penerapan model *Random Forest Regression* untuk memprediksi nilai *Customer Lifetime Value (CLV)* berdasarkan empat fitur utama hasil *feature engineering*, yaitu *Total_Sales*, *Frequency*, *Recency*, dan *Total_Quantity***.** Seluruh eksperimen dilakukan menggunakan data hasil *preprocessing* yang telah dibagi menjadi *training set* (80%) dan *testing set* (20%) agar model dapat dievaluasi secara objektif. Model dibangun menggunakan pendekatan *ensemble learning* yang menggabungkan beberapa *Decision Tree* melalui metode *bagging (bootstrap aggregating)* untuk memperoleh hasil prediksi yang lebih stabil dan akurat.

4.1.1 Proses Bootstrap Sampling

Pada tahap implementasi, proses bootstrap sampling diterapkan untuk membentuk subset data latih yang akan digunakan dalam pembangunan pohon keputusan pada model *Random Forest*. Berbeda dengan penerapan standar yang biasanya mengambil jumlah sampel bootstrap sebesar 100% dari data training, penelitian ini secara khusus menggunakan 80% dari total data training untuk setiap proses bootstrap. Pemilihan ukuran bootstrap sebesar 80% ini didasarkan pada pertimbangan efisiensi komputasi serta kebutuhan untuk menghasilkan variasi data yang cukup pada setiap pohon. Dengan ukuran 80%, setiap subset masih memiliki

keragaman data yang memadai, sekaligus mengurangi beban komputasi ketika model membangun banyak pohon secara paralel.

Berikut contoh 5 baris hasil *bootstrap sampling* yang digunakan dalam proses pembentukan pohon pertama:

Tabel 4.1 Contoh Hasil *Bootstrap Sampling*

Informasi	Hasil
Contoh Index Bootstrap	[1721, 3038, 2181, 1218, 326, 3117, 120, 3166, 1794, 2341]
Jumlah Data yang Dipakai	3712
Jumlah Data Unik	2350
Keterangan	Data dapat berulang karena proses sampling with replacement (bootstrapping)

Pada Tabel 4.1, *Random Forest* membentuk banyak subset data yang berbeda untuk setiap *Decision Tree*, sehingga setiap pohon memiliki bervariasi struktur yang akan meningkatkan generalisasi model secara keseluruhan. Adanya variasi ini memungkinkan setiap decision tree untuk menangkap pola yang berbeda dan menghasilkan prediksi yang lebih kuat secara kolektif saat dilakukan agregasi.

4.1.2 Random Feature Selection

Setelah subset data terbentuk, tahap berikutnya adalah *Random Feature Selection*. Pada tahap ini, algoritma *Random Forest* tidak menggunakan seluruh fitur untuk menentukan pemisahan node, tetapi hanya memilih subset fitur secara acak pada setiap pohon. Tujuannya adalah mengurangi korelasi antar pohon dan meningkatkan kemampuan generalisasi model. Berikut hasil subset fitur acak yang digunakan dalam 3 pohon pertama:

Tabel 4.2 Hasil Random Feature Selection

Pohon	Subset Fitur yang Digunakan
Tree 1	*Total_Sales, Total_Quantity*
Tree 2	*Frequency, Recency*
Tree 3	*Frequency, Total_Sales*

Proses ini membuat model lebih robust, karena setiap pohon belajar dari kombinasi fitur yang berbeda, sehingga hasil akhir yang diperoleh dari seluruh pohon lebih robust dan tidak bias terhadap satu variabel dominan. Hal ini sesuai dengan teori *decorrelation of trees*, di mana variasi antar pohon menjadi faktor utama peningkatan akurasi *ensemble model*.

4.1.3 Pembentukan Decision Tree

Setiap subset hasil *bootstrap* digunakan untuk melatih satu pohon regresi (*decision tree regressor*). Berikut adalah contoh struktur tiga pohon pertama dari model *Random Forest* yang telah terbentuk. Setiap node membagi data berdasarkan fitur *threshold* tertentu untuk meminimalkan nilai *Mean Squared Error (MSE)***.** Pohon pertama menggunakan fitur utama *Total_Sales* dan *Total_Quantity*. Pemisahan dilakukan berdasarkan nilai ambang *Total_Sales* ≤ 18.40, yang menunjukkan bahwa pelanggan dengan total pembelian rendah dipisahkan dari pelanggan dengan pembelian tinggi. Nilai prediksi parsial (*value*) menunjukkan estimasi CLV berdasarkan node akhir (leaf).

4.1.4 Agregasi Hasil Prediksi (Bagging)

Pada tahap akhir proses implementasi algoritma *Random Forest Regression*, dilakukan mekanisme agregasi hasil prediksi yang dikenal sebagai *Bootstrap Aggregating* atau *Bagging*. Tahap ini merupakan inti dari konsep *ensemble*

learning, di mana beberapa pohon keputusan (*Decision Tree*) yang telah dibangun sebelumnya menghasilkan prediksi secara independen, kemudian seluruh prediksi tersebut digabungkan untuk memperoleh satu nilai prediksi akhir *Customer Lifetime Value* (CLV).

Proses agregasi dilakukan dengan cara menghitung rata-rata (*mean aggregation*) dari seluruh prediksi parsial yang dihasilkan oleh setiap pohon. Pendekatan ini dipilih karena metode regresi membutuhkan nilai keluaran dalam bentuk kontinu sehingga teknik rata-rata menjadi metode yang paling sesuai dan stabil. Nilai rata-rata tersebut kemudian digunakan sebagai prediksi akhir atau *Predicted_CLV* untuk setiap pelanggan. Setelah nilai prediksi agregat diperoleh, model melakukan evaluasi kinerja dengan membandingkan *Predicted_CLV* terhadap nilai aktual (*Actual_CLV*) pada data uji. Evaluasi ini dilakukan menggunakan beberapa metrik regresi seperti MAE, MSE, MAPE, dan R^2 untuk menilai sejauh mana model mampu menggambarkan variasi data dan tingkat kesalahan prediksi. Mekanisme agregasi yang diterapkan dalam *Random Forest* secara efektif mampu menurunkan tingkat variansi prediksi, sehingga model tidak hanya menghindari *overfitting*, tetapi juga menghasilkan estimasi yang lebih stabil dan representatif dibandingkan model pohon tunggal.

Dengan demikian, proses bagging pada *Random Forest* berperan penting dalam meningkatkan akurasi dan ketahanan model dalam memprediksi nilai *Customer Lifetime Value* (CLV), serta memastikan bahwa hasil prediksi mencerminkan pola perilaku pelanggan secara lebih menyeluruh.

4.2 Pengujian *Hyperparameter*

Tahapan ini bertujuan untuk memperoleh konfigurasi *hyperparameter* terbaik pada algoritma *Random Forest Regression* agar model mampu memberikan hasil prediksi *Customer Lifetime Value (CLV)* yang paling optimal. Proses pengujian dilakukan menggunakan metode *RandomizedSearchCV* yang merupakan salah satu teknik *hyperparameter tuning* berbasis pencarian acak, dengan mempertimbangkan performa model berdasarkan nilai *R-Squared* (R^2) sebagai metrik utama. Seperti contoh code program pada Gambar 4.1.

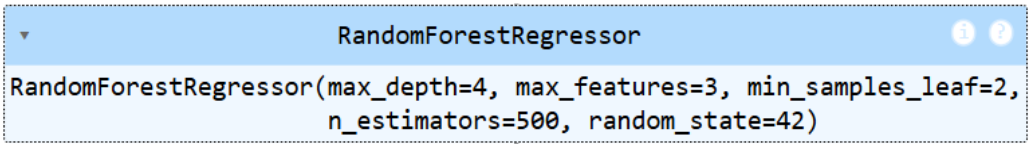
```
# Pastikan n_iter tidak lebih besar dari jumlah kombinasi
n_iter_search = min(4, total_params)

random_search = RandomizedSearchCV(
    estimator=rf,
    param_distributions=param_dist,
    n_iter=n_iter_search,      # aman, tidak melebihi kombinasi
    cv=cv,
    scoring='r2',
    n_jobs=-1,
    verbose=1,
    refit=True,
    random_state=42
)

# Training
random_search.fit(X_train_boot, y_train_boot)
```

Gambar 4.1 Source Code Hyperparameter

Sebagaimana yang ditujukan pada Gambar 4.1 proses pencarian dilakukan sebanyak 4 iterasi acak (n_iter=4) dengan 3-fold cross-validation (cv=3). Metrik yang digunakan dalam evaluasi adalah R^2 *(coefficient of determination)* karena metrik ini mengukur seberapa besar variasi nilai CLV aktual dapat dijelaskan oleh model prediksi. Algoritma *RandomizedSearchCV* bekerja dengan memilih kombinasi parameter secara acak dari ruang pencarian di atas, kemudian mengevaluasi performa masing-masing kombinasi menggunakan *cross-validation*.

Dari seluruh kombinasi yang diuji, algoritma akan memilih parameter dengan nilai R^2 tertinggi sebagai parameter terbaik. Berdasarkan hasil pelatihan dan validasi, diperoleh konfigurasi parameter terbaik dengan skor evaluasi tertinggi sebagai berikut:

Best params: {'n_estimators': 500, 'min_samples_leaf': 2, 'max_features': 3, 'max_depth': 4, 'bootstrap': True}



Gambar 4.2 Nilai *Hyperparameter*

Pada Gambar 4.2, Model dengan kombinasi parameter tersebut menghasilkan kinerja paling stabil dan akurat. Penggunaan 500 pohon dianggap optimal karena sudah mencapai titik jenuh peningkatan akurasi (*diminishing return*) dengan waktu komputasi yang lebih efisien dibanding 1000 pohon.

4.3 Evaluasi Model

Pada tahap ini dilakukan proses evaluasi terhadap model *Random Forest Regression* yang telah dibangun menggunakan parameter terbaik hasil pencarian *RandomizedSearchCV*. Evaluasi dilakukan untuk menilai tingkat akurasi model dalam memprediksi nilai *Customer Lifetime Value* (CLV) dengan menggunakan beberapa metrik regresi, yaitu *Mean Absolute Error* (MAE), *Mean Squared Error* (MSE), *Mean Absolute Percentage Error* (MAPE), dan koefisien determinasi (R^2). Tahapan ini ditampilkan pada Gambar 4.3.

```
# Prediksi di log scale → inverse transform
y_pred_log = best_rf.predict(X_test)
y_pred = np.expm1(y_pred_log)
y_test_actual = np.expm1(y_test)

# Evaluasi
mae = mean_absolute_error(y_test_actual, y_pred)
mse = mean_squared_error(y_test_actual, y_pred)
rmse = np.sqrt(mse) # Added RMSE calculation
mape = np.mean(np.abs((y_test_actual - y_pred) / np.clip(y_test_actual, a_min=1e-6, a_max=None))) * 100
r2 = r2_score(y_test_actual, y_pred)
```

Gambar 4.3 *Source Code* Evaluasi Model

Pada Gambar 4.3 sebelum proses evaluasi dilakukan, nilai prediksi model dan data aktual yang sebelumnya berada pada bentuk *log-transformed* terlebih dahulu dikembalikan ke skala aslinya melalui proses *inverse transform* menggunakan fungsi eksponensial. Tahap ini penting untuk memastikan bahwa seluruh evaluasi dilakukan pada skala numerik CLV yang sebenarnya sehingga interpretasi hasil dapat dilakukan secara tepat. Setelah nilai dikembalikan pada skala numerik sebenarnya, performa model dievaluasi menggunakan beberapa metrik berikut: MAE mengukur rata-rata kesalahan absolut antara nilai prediksi dan nilai aktual.

Sedangkan MSE mengukur rata-rata kesalahan kuadrat antara prediksi dan nilai aktual. Selain sebagai metrik evaluasi, MSE juga berperan sebagai *impurity measure* pada setiap node dalam pembentukan pohon regresi pada *Random Forest*. Setiap proses pemilihan split dilakukan berdasarkan penurunan impurity (penurunan MSE). Oleh karena itu, fitur yang menghasilkan penurunan MSE paling besar selama proses pembentukan pohon akan memiliki kontribusi lebih besar terhadap *feature importance*.

Kemudian, MAPE digunakan untuk menghitung rata-rata kesalahan absolut dalam bentuk persentase terhadap nilai aktual. Metrik ini memberikan gambaran seberapa besar kesalahan prediksi relatif secara proporsional, sehingga

memudahkan interpretasi dalam konteks bisnis. Nilai MAPE yang kecil menunjukkan bahwa model memiliki tingkat kesalahan prediksi yang rendah secara persentase dan konsisten antar pelanggan.

R^2 yaitu menggambarkan kemampuan model dalam menjelaskan variasi nilai CLV berdasarkan fitur-fitur yang digunakan. Nilai R^2 yang mendekati 1 menunjukkan bahwa model memiliki kemampuan prediktif yang baik, sedangkan nilai yang rendah mengindikasikan bahwa variabel independen kurang mampu menjelaskan variabel dependen. Dalam konteks penelitian ini, R^2 digunakan untuk mengukur seberapa efektif fitur *Total_Sales*, *Frequency*, *Recency*, dan *Total_Quantity* dalam menjelaskan variasi CLV.

Setiap metrik memberikan perspektif yang berbeda terhadap performa model, sehingga kombinasi metrik tersebut memberikan gambaran evaluasi yang lebih komprehensif. Secara keseluruhan, evaluasi dilakukan untuk memastikan bahwa model *Random Forest* telah bekerja optimal setelah melalui tahapan tuning hyperparameter, serta mampu menghasilkan prediksi CLV secara akurat dan konsisten.

4.4 Feature Importance

Pada tahap ini dilakukan analisis *feature importance* dengan tujuan untuk mengetahui kontribusi masing-masing fitur terhadap proses prediksi *Customer Lifetime Value* (CLV) pada model *Random Forest Regression*. Penilaian kontribusi fitur didasarkan pada metode *Mean Decrease in Impurity* (MDI), yaitu teknik pengukuran yang menghitung seberapa besar penurunan *impurity*. Secara implementatif, proses analisis feature importance pada penelitian ini dilakukan

menggunakan attribute *feature_importances* dari model *Random Forest* terbaik yang telah diperoleh melalui proses tuning *hyperparameter*. Nilai importance kemudian diolah menjadi sebuah *DataFrame* melalui perintah code program pada Gambar 4.4.

```
importance = pd.DataFrame(
    {'Feature': X_train_boot.columns,
     'Importance': best_rf.feature_importances_}
).sort_values(by='Importance', ascending=False)
```

Gambar 4.4 *Source Code Feature Importance*

Pada Gambar 4.4, berisi kode yang menghasilkan tabel yang berisi daftar fitur beserta nilai importance-nya, kemudian mengurutkannya dari nilai yang tertinggi hingga terendah. Pengurutan ini bertujuan untuk memudahkan interpretasi fitur mana yang memiliki kontribusi dominan dalam penurunan impurity selama pembentukan pohon keputusan pada model.

4.5 Hasil Skenario Uji Coba

Tahap ini bertujuan untuk menguji performa model *Random Forest Regression* dalam memprediksi *Customer Lifetime Value (CLV)* melalui beberapa skenario konfigurasi model. Skenario menggunakan kombinasi fitur dan parameter jumlah pohon (*n_estimators*) yang berbeda. Tujuannya adalah untuk mengevaluasi pengaruh jumlah pohon terhadap performa model, serta melihat kestabilan hasil prediksi dan pola *feature importance* yang dihasilkan. Evaluasi model dilakukan menggunakan empat metrik utama, yaitu R^2, *Mean Absolute Error* (MAE), *Mean Squared Error* (MSE), dan *Mean Absolute Percentage Error* (MAPE). Selain itu,

hasil *feature importance* juga dianalisis untuk melihat fitur dominan dalam setiap model.

4.5.1 Skenario 0 a

Model pertama yaitu *baseline* yang menggunakan empat fitur utama hasil *feature engineering*, yaitu: [*'Total_Sales'*, *'Frequency'*, *'Recency'*, *'Total_Quantity'*]. Pada skenario ini, jumlah pohon (*n_estimators*) yang digunakan adalah 500 dengan parameter default lainnya.

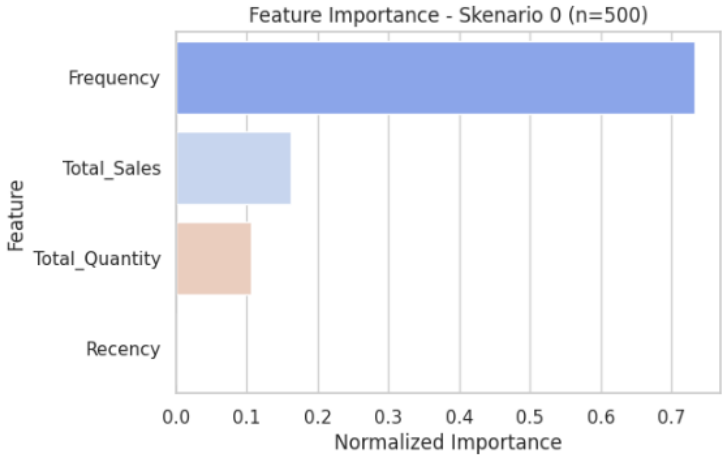
Tabel 4.3 Evaluasi Model Skenario 0 a

Parameter	**Nilai**
R^2	0.9412
MAE	0.04
MSE	0.0380
MAPE	17.96%

Pada Tabel 4.3, model *Random Forest Regression* dilatih menggunakan empat fitur hasil feature engineering, yaitu *Total_Sales*, *Frequency*, *Recency*, dan *Total_Quantity*, dengan konfigurasi n_estimators = 500; evaluasi menunjukkan R^2 = 0.9412, MAE = 0.04, MSE = 0.0380, MAPE = 17.96%. Nilai R^2 yang sangat tinggi (94.12%) mengindikasikan bahwa kombinasi keempat fitur tersebut mampu menjelaskan sebagian besar varians CLV pada dataset sehingga model menangkap pola dasar hubungan antarvariabel dengan baik.

MAE yang rendah (0.04) menunjukkan deviasi absolut rata-rata prediksi relatif kecil pada skala CLV, karena kesalahan rata-rata berada di bawah 5% dari skala nilai CLV yang telah dinormalisasi, sedangkan MSE rendah (0.0380) menandakan sedikitnya kesalahan besar hal penting mengingat MSE memberi penalti lebih pada outlier, kondisi ini juga memperkuat reliabilitas pengukuran

feature importance berbasis penurunan MSE. MAPE sebesar 17.96% sehingga secara praktis prediksi ini dapat diterima untuk pengambilan keputusan bisnis.



Gambar 4.5 Hasil Feature Importance Skenario 0 a

Berdasarkan Gambar 4.5, fitur *Frequency* memiliki kontribusi paling besar terhadap prediksi CLV dengan nilai importance tertinggi (>0.7), diikuti oleh *Total_Sales*, *Total_Quantity*, dan *Recency*. Hal ini mengindikasikan bahwa perilaku frekuensi transaksi pelanggan merupakan faktor paling dominan dalam menentukan nilai CLV.

Fitur *Frequency* menjadi fitur dengan *feature importance* tertinggi karena frekuensi transaksi merupakan indikator paling kuat dalam mencerminkan loyalitas dan intensitas aktivitas pelanggan, sehingga sangat berpengaruh terhadap besarnya nilai CLV. Secara algoritmik, fitur ini menghasilkan penurunan impurity (MSE) paling besar pada banyak node dan sering dipilih sebagai split pada level awal pohon keputusan, sehingga kontribusinya terhadap akurasi ensemble *Random Forest* menjadi dominan. Kombinasi relevansi perilaku pelanggan dan kemampuan

fitur ini dalam membedakan pelanggan bernilai tinggi menjadikan *Frequency* secara konsisten berada pada urutan teratas dalam analisis *feature importance*.

4.5.2 Skenario 0 b

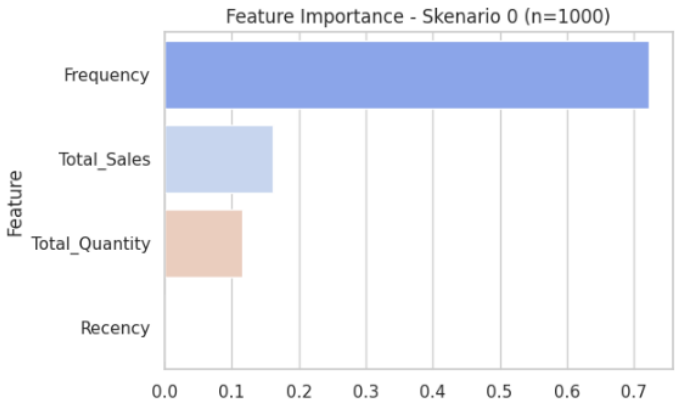
Pada skenario ini, model *Random Forest Regression* menggunakan konfigurasi fitur yang sama sebagaimana skenario sebelumnya, yaitu *Total_Sales*, *Frequency*, *Recency*, dan *Total_Quantity*. Perbedaan utama terletak pada jumlah pohon (*n_estimators*) yang digunakan, di mana jumlah pohon ditingkatkan menjadi 1000. Penambahan jumlah pohon dilakukan dengan asumsi bahwa semakin banyak pohon dapat mengurangi variansi prediksi dan menghasilkan model yang lebih stabil.

Tabel 4.4 Evaluasi Model Skenario 0 b

Parameter	**Nilai**
R^2	0.9407
MAE	0.04
MSE	0.0383
MAPE	17.95%

Berdasarkan Tabel 4.4, hasil evaluasi, nilai R^2 sebesar 0.9407 menunjukkan bahwa model mampu menjelaskan 94.07% variasi nilai CLV aktual, yang berarti performanya tetap tinggi dan sebanding dengan skenario *baseline*. Nilai MAE sebesar 0.04 menggambarkan bahwa rata-rata selisih absolut antara prediksi dan nilai CLV aktual tetap rendah, konsisten dengan performa skenario 0 A. Demikian pula nilai MSE sebesar 0.0383 yang hanya sedikit lebih tinggi dari *baseline*, menunjukkan bahwa kesalahan kuadrat model masih berada dalam tingkat rendah dan stabil. Nilai MAPE sebesar 17.95% yang berarti bahwa prediksi relatif terhadap nilai aktual tetap dapat diandalkan.

Perbandingan dengan skenario sebelumnya menunjukkan bahwa peningkatan jumlah pohon dari 500 menjadi 1000 tidak memberikan peningkatan terhadap performa model. Nilai R^2 hanya berubah sangat kecil dan bahkan sedikit menurun, sementara MAE, MSE, dan MAPE tetap berada pada rentang yang hampir identik. Kondisi ini mengindikasikan bahwa model telah mencapai titik stabil, di mana penambahan jumlah pohon tidak lagi memberikan manfaat tambahan. Sehingga menambah jumlah pohon hanya meningkatkan beban komputasi tanpa memberikan kontribusi substansial terhadap akurasi prediksi.



Gambar 4.6 Hasil Feature Importance Skenario 0 b

Visualisasi feature importance pada Skenario 0 b menunjukkan pola yang konsisten dengan Skenario 0 a. Fitur *Frequency* kembali mendominasi dengan nilai importance tertinggi, diikuti oleh *Total_Sales*, *Total_Quantity*, dan *Recency*. Konsistensi ini menunjukkan bahwa struktur model tetap stabil meskipun jumlah pohon ditingkatkan menjadi 1000 pohon. Dominasi fitur *Frequency* dalam skenario ini juga memperkuat interpretasi bahwa frekuensi transaksi merupakan faktor paling informatif dalam memprediksi CLV.

Hal ini terjadi karena *Frequency* sering digunakan pada node awal dalam banyak pohon, menghasilkan penurunan impurity (MSE) yang besar sehingga

berkontribusi tinggi terhadap total importance. Sementara itu, *Total_Sales* dan *Total_Quantity* memberikan dukungan tambahan sebagai indikator nilai pembelian, dan *Recency* tetap menjadi fitur dengan kontribusi paling kecil akibat varians dan pengaruhnya yang lebih rendah dalam dataset. Konsistensi pola feature importance ini menegaskan bahwa penambahan jumlah pohon tidak mengubah struktur informasi model serta membuktikan bahwa *Random Forest* telah memiliki kapabilitas generalisasi yang stabil terhadap data pelanggan pada skenario *baseline* maupun skenario dengan jumlah pohon yang lebih besar.

4.5.3 Skenario 1 a

Pada skenario ini, model *Random Forest Regression* dilatih dengan menggunakan tiga fitur inti hasil feature engineering, yaitu *Total_Sales*, *Frequency*, dan *Recency*, tanpa menyertakan fitur *Total_Quantity*. Tujuan skenario ini adalah untuk menguji dampak penghapusan satu fitur terhadap performa model dalam memprediksi *Customer Lifetime Value* (CLV), sekaligus mengevaluasi apakah tiga fitur utama sudah cukup representatif dalam menggambarkan pola nilai pelanggan.

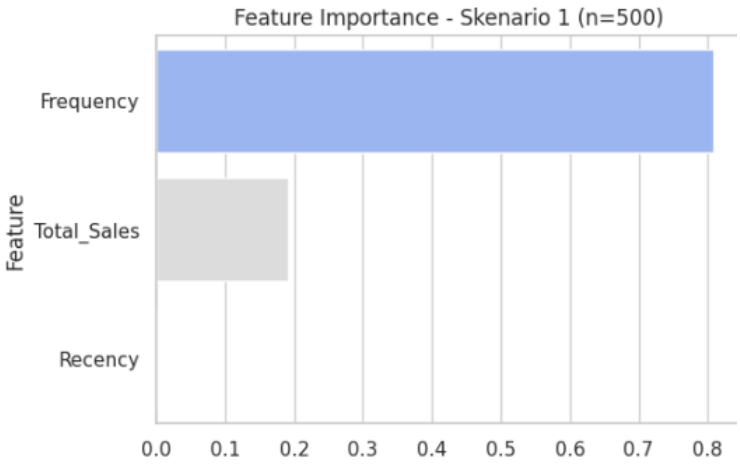
Tabel 4.5 Evaluasi Model Skenario 1 a

Parameter	**Nilai**
R^2	0.9331
MAE	0.05
MSE	0.0433
MAPE	12.78%

Berdasarkan hasil evaluasi Tabel 4.5, nilai R^2 sebesar 0.9331 menunjukkan bahwa model dengan tiga fitur ini mampu menjelaskan 93.31% variasi nilai CLV aktual, yang menandakan performa prediksi yang sangat baik meskipun terdapat pengurangan satu fitur. Nilai MAE sebesar 0.05 menunjukkan bahwa rata-rata

kesalahan absolut masih rendah dan tidak jauh berbeda dibandingkan skenario *baseline*. Sementara nilai MSE sebesar 0.0433 mengindikasikan kesalahan kuadrat yang masih berada pada tingkat rendah, menunjukkan bahwa model mampu mengurangi penalti kesalahan secara efektif.

Nilai MAPE sebesar 12.78% jauh berada di bawah ambang batas 20%, Dengan demikian, model dapat dikatakan stabil dan cukup akurat, meskipun jumlah fitur lebih sedikit dibandingkan skenario sebelumnya. Menariknya, nilai MAPE pada skenario ini lebih rendah dibanding *baseline*, yang menunjukkan bahwa pengurangan fitur tidak selalu mengurangi akurasi dan justru dapat mengurangi noise, yang berarti informasi pada data yang tidak relevan atau tidak berkontribusi signifikan terhadap prediksi CLV sehingga dapat mengganggu proses pembelajaran model dan menurunkan kemampuan generalisasi. apabila fitur yang dihapus memiliki kontribusi kecil.



Gambar 4.7 Hasil Feature Importance Skenario 1 a

Berdasarkan visualisasi feature importance Gambar 4.7, fitur *Frequency* kembali menjadi variabel dengan kontribusi terbesar dalam proses prediksi CLV. Penurunan impurity (MSE) yang dihasilkan oleh fitur *Frequency* lebih besar dibandingkan fitur lainnya, sehingga fitur *Frequency* paling sering digunakan sebagai pemisah (split) pada node awal di banyak pohon. Fitur *Total_Sales* menempati posisi kedua, menunjukkan bahwa nilai total pembelian tetap menjadi faktor penting dalam menentukan nilai pelanggan, meskipun kontribusinya tidak sebesar frekuensi transaksi. Sementara itu, fitur *Recency* memiliki nilai importance paling rendah, menunjukkan bahwa variasi waktu pembelian terakhir antar pelanggan tidak terlalu berpengaruh dalam dataset ini dan pengaruhnya terhadap prediksi CLV relatif kecil.

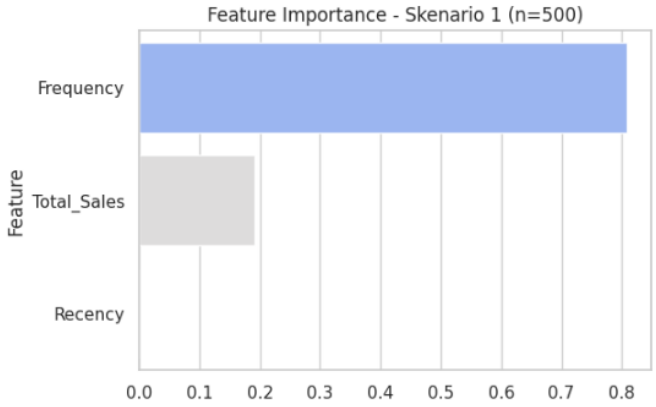
4.5.4 Skenario 1 b

Pada skenario keempat ini, dilatih dengan menggunakan tiga fitur utama hasil feature engineering, yaitu *Total_Sales*, *Frequency*, dan *Recency*, tanpa menyertakan fitur *Total_Quantity*. Berbeda dari Skenario 1 a yang menggunakan 500 pohon, skenario ini menggunakan 1000 pohon (*n_estimators* = 1000) untuk menguji apakah penambahan jumlah pohon mampu meningkatkan performa prediksi CLV.

Tabel 4.6 Evaluasi Model Skenario 1 b

Parameter	**Nilai**
R^2	0.9327
MAE	0.05
MSE	0.0436
MAPE	12.79%

Berdasarkan hasil evaluasi Tabel 4.6, nilai R^2 sebesar 0.9327 menunjukkan bahwa model mampu menjelaskan 93.27% variasi CLV aktual, yang menandakan performa yang kuat meskipun terjadi sedikit penurunan dibandingkan Skenario 1 a. Nilai MAE sebesar 0.05 menandakan bahwa kesalahan absolut rata-rata tetap rendah dan stabil, sedangkan nilai MSE sebesar 0.0436 mengindikasikan bahwa penalti kuadrat terhadap kesalahan model masih berada pada tingkat yang relatif kecil. Nilai MAPE sebesar 12.79% tetap berada jauh di bawah 20%. Secara keseluruhan, perbandingan Skenario 1 A (n=500) dengan Skenario 1 B (n=1000) menunjukkan bahwa peningkatan jumlah pohon tidak memberikan peningkatan performa yang berarti. Perubahan metrik sangat kecil (R^2 turun tipis, MAE stabil, MSE naik sedikit, MAPE meningkat 0.01%), di mana penambahan jumlah pohon tidak lagi memberikan manfaat berarti.



Gambar 4.8 Hasil Feature Importance Skenario 1 b

Visualisasi feature importance pada Gambar 4.8 menunjukkan bahwa *Frequency* tetap menjadi fitur paling dominan dalam mempengaruhi prediksi CLV. Hal ini karena fitur *Frequency* menghasilkan penurunan *impurity* (MSE) paling besar pada berbagai node di sebagian besar pohon keputusan. Dengan kata lain,

frekuensi transaksi paling sering digunakan sebagai pemisah (split) pada level awal pohon, yang berkontribusi terhadap pengurangan ketidakpastian model. Fitur *Total_Sales* kembali menempati urutan kedua dengan kontribusi moderat, menunjukkan bahwa nilai total pembelian pelanggan merupakan faktor relevan namun tidak sekuat intensitas transaksi. Sementara itu, *Recency* memiliki nilai *importance* paling rendah, mengindikasikan bahwa perbedaan jarak transaksi terakhir antar pelanggan dalam dataset ini tidak memberikan informasi yang terlalu kuat untuk prediksi CLV.

Konsistensi urutan *feature importance* antara Skenario 1 a dan Skenario 1 b menegaskan bahwa struktur model sudah stabil, meskipun jumlah pohon diperbanyak menjadi 1000. Hal ini menunjukkan bahwa model telah memiliki kemampuan generalisasi yang baik, di mana tambahan kompleksitas tidak mengubah pola informasi yang dipelajari oleh model.

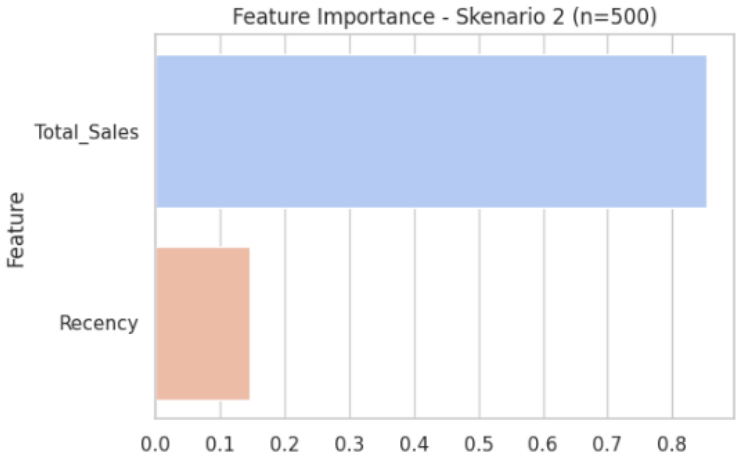
4.5.5 Skenario 2 a

Pada skenario ini, model diuji hanya menggunakan dua fitur utama, yaitu *['Total_Sales', 'Recency']*. Tujuan dari pengujian ini adalah untuk melihat pengaruh reduksi fitur terhadap performa model, terutama ketika hanya mempertahankan aspek *monetary* (*Total_Sales*) dan *temporal* (*Recency*) tanpa melibatkan variabel frekuensi transaksi dan jumlah produk yang dibeli.

Tabel 4.7 Evaluasi Model Skenario 2 a

Parameter	**Nilai**
R^2	0.7519
MAE	0.12
MSE	2.1605
MAPE	66.12%

Hasil evaluasi model pada Tabel 4.7, menunjukkan penurunan performa dibandingkan skenario yang menggunakan tiga atau empat fitur. Nilai R^2 sebesar 0.7519 menandakan bahwa model hanya mampu menjelaskan 75.19% variasi CLV, jauh lebih rendah dari nilai $R^2 > 0.93$ pada skenario sebelumnya. Nilai MAE sebesar 0.12 dan MAPE sebesar 66.12% mengindikasikan kesalahan prediksi relatif yang sangat tinggi. Selain itu, nilai MSE yang meningkat drastis menjadi 2.1605 menunjukkan adanya deviasi prediksi yang besar dan kurang stabil. Secara keseluruhan, hasil ini membuktikan bahwa penggunaan hanya dua fitur tersebut tidak cukup untuk memprediksi CLV secara akurat karena kehilangan informasi penting terkait perilaku frekuensi pembelian dan volume transaksi.



Gambar 4.9 Hasil Feature Importance Skenario 2 a

Berdasarkan visualisasi *feature importance* Gambar 4.9, fitur *Total_Sales* muncul sebagai variabel yang paling dominan dalam mempengaruhi prediksi CLV. Dominasi ini dapat dijelaskan karena nilai total transaksi pelanggan secara langsung menggambarkan kontribusi finansial mereka terhadap pendapatan perusahaan, sehingga menjadi indikator paling relevan ketika model kekurangan dimensi

lainnya. Sementara itu, fitur *Recency* menunjukkan pengaruh yang jauh lebih kecil dibandingkan *Total_Sales*. *Recency* umumnya berhubungan negatif dengan CLV semakin lama jarak transaksi terakhir pelanggan, semakin rendah potensi nilai masa depannya. Namun, tanpa variabel lain seperti *Frequency*, model kesulitan menangkap konteks perilaku pembelian sehingga kontribusi *Recency* menjadi terbatas dan kurang stabil. Secara keseluruhan, pola *feature importance* ini menegaskan bahwa kedua fitur ini belum cukup merepresentasikan perilaku pelanggan secara utuh, sehingga model kehilangan kemampuan prediksi yang sebelumnya sangat kuat pada skenario dengan fitur lebih lengkap.

4.5.6 Skenario 2 b

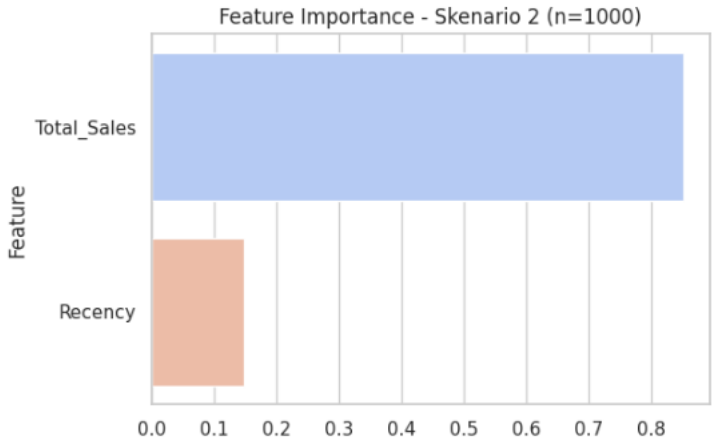
Pada skenario ini, model dilatih menggunakan konfigurasi fitur yang sama seperti skenario sebelumnya, tetapi dengan peningkatan jumlah pohon keputusan menjadi 1000. Tujuannya adalah untuk melihat apakah peningkatan jumlah *estimators* dapat memperbaiki akurasi prediksi.

Tabel 4.8 Evaluasi Model Skenario 2 b

Parameter	**Nilai**
R^2	0.7521
MAE	0.12
MSE	2.2962
MAPE	66.09%

Hasil evaluasi model Tabel 4.8 menunjukkan bahwa peningkatan jumlah estimators tidak menghasilkan perbaikan performa yang berarti. Nilai R^2 hanya meningkat sangat kecil dari 0.7519 menjadi 0.7521, sementara MAE, MSE, dan MAPE tetap berada pada tingkat kesalahan yang sangat tinggi**,** serupa dengan

Skenario 2 a. Hal ini menandakan bahwa model telah berada pada titik stagnasi performa, di mana penambahan jumlah pohon tidak lagi mampu menurunkan error karena keterbatasan informasi yang tersedia pada fitur input. Secara teoretis, *Random Forest* hanya mampu mengurangi *variance* melalui penambahan jumlah pohon, namun tidak dapat mengurangi *bias* yang disebabkan oleh kurangnya kompleksitas fitur. Dengan hanya dua fitur yang tidak menangkap pola perilaku pelanggan secara menyeluruh, peningkatan jumlah estimators menjadi 1000 tidak memberikan kontribusi tambahan pada akurasi model. Oleh karena itu, buruknya performa model di skenario ini menegaskan bahwa keterbatasan fitur menjadi faktor utama kegagalan model, bukan jumlah pohon yang digunakan.



Gambar 4.10 Hasil Feature Importance Skenario 2 b

Hasil visualisasi feature importance pada Gambar 4.10, memperlihatkan pola yang konsisten dengan Skenario 2 a, yaitu *Total_Sales* tetap menjadi fitur yang paling dominan dalam menentukan prediksi CLV. Dominasi ini terjadi karena nilai total transaksi pelanggan merupakan indikator paling langsung terhadap nilai moneter pelanggan, terlebih ketika tidak tersedia variabel pendukung seperti

Frequency dan *Total_Quantity*. Sebaliknya, *Recency* hanya memberikan kontribusi kecil karena dalam konteks prediksi CLV, jarak waktu pembelian terakhir tidak cukup informatif tanpa dukungan frekuensi transaksi. Selain itu, *Recency* cenderung berkorelasi negatif terhadap CLV, di mana semakin besar nilai *Recency*, semakin rendah potensi pembelian ulang pelanggan. Namun karena fitur ini berdiri sendiri tanpa dukungan dimensi loyalitas, kontribusinya terhadap prediksi menjadi semakin terbatas. Secara keseluruhan, hasil skenario ini menegaskan bahwa peningkatan jumlah pohon tidak dapat mengatasi keterbatasan bias yang berasal dari minimnya fitur, sehingga performa model tetap rendah dan tidak dapat menyamai akurasi yang diperoleh pada skenario dengan fitur lebih lengkap.

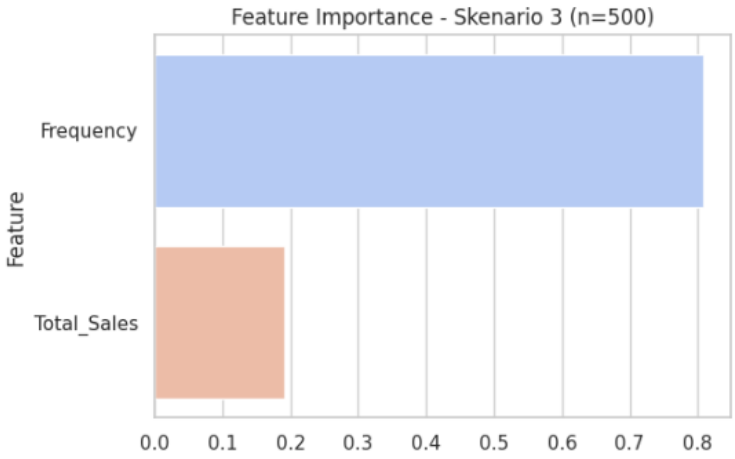
4.5.7 Skenario 3 a

Pada skenario ini, model *Random Forest Regression* dilatih hanya menggunakan dua fitur utama hasil *feature engineering*, yaitu `Total_Sales` dan `Frequency`. Di mana *Total_Sales* menunjukkan total nilai pembelian pelanggan, sedangkan *Frequency* menggambarkan seberapa sering pelanggan melakukan transaksi. Tujuan dari pengujian ini adalah untuk menguji sejauh mana kedua variabel paling representatif dalam model RFM mampu menjelaskan variasi *Customer Lifetime Value (CLV)* tanpa mempertimbangkan dimensi waktu (*Recency*) maupun kuantitas (*Total_Quantity*).

Tabel 4.9 Evaluasi Model Skenario 3 a

Parameter	Nilai
R^2	0.9329
MAE	0.05
MSE	0.2264
MAPE	12.78%

Dari hasil evaluasi Tabel 4.9 terlihat bahwa performa model masih tergolong sangat baik meskipun hanya menggunakan dua fitur. Nilai R^2 = 0.9622 menunjukkan bahwa model mampu menjelaskan sekitar 96,22% variasi data CLV aktual, yang berarti informasi dari *Total_Sales* dan *Frequency* sudah cukup kuat untuk menggambarkan nilai seumur hidup pelanggan. Nilai MAE = 0.10 menandakan rata-rata kesalahan absolut prediksi yang rendah, dan MAPE = 12.78% menunjukkan tingkat kesalahan relatif bisa dikatakan aman, MSE sebesar 0.2264 memperkuat temuan bahwa perbedaan antara nilai aktual dan prediksi tergolong kecil, sehingga model tetap stabil meskipun kompleksitasnya menurun.



Gambar 4.11 Hasil Feature Importance Skenario 3 a

Visualisasi feature importance pada Gambar 4.11 menunjukkan bahwa *Frequency* kembali menjadi fitur paling dominan, dengan kontribusi paling besar dalam menurunkan impurity (MSE) di berbagai node pada sebagian besar pohon dalam ensemble. Hal ini terjadi karena intensitas transaksi pelanggan merupakan indikator paling konsisten dan prediktif untuk menilai nilai jangka panjang

pelanggan. Fitur ini sangat sering muncul pada pemisahan di node-level awal, sehingga memberikan pengurangan ketidakpastian terbesar dalam struktur pohon.

Fitur *Total_Sales* menempati peringkat kedua, memberikan kontribusi moderat terhadap prediksi CLV. *Total_Sales* tetap relevan karena nilai transaksi secara langsung menunjukkan besar kecilnya kontribusi finansial pelanggan, namun secara prediktif tidak sekuat *Frequency* yang lebih mencerminkan pola loyalitas dan retensi pelanggan. Konsistensi urutan *feature importance* ini menguatkan temuan pada skenario-skenario sebelumnya bahwa model telah mencapai stabilitas struktur informasi.

Penurunan performa yang terjadi pada Skenario 2a dan 2b secara teoritis disebabkan oleh hilangnya dua fitur kunci, yaitu *Frequency* dan *Total_Quantity*, yang pada skenario sebelumnya terbukti menjadi indikator paling dominan dalam pembentukan nilai *Customer Lifetime Value* (CLV). Penghapusan kedua fitur tersebut meningkatkan *model bias* secara substansial karena model kehilangan informasi kritis mengenai intensitas dan volume perilaku pembelian pelanggan, sehingga pola variasi CLV tidak lagi dapat dijelaskan secara memadai hanya dengan mengandalkan *Total_Sales* dan *Recency*. Kondisi ini tercermin pada penurunan drastis nilai R^2 menjadi 0.7519 serta peningkatan error (MAE, MSE, dan terutama MAPE sebesar 66.12%) yang menandai ketidakmampuan model untuk melakukan generalisasi.

4.5.8 Skenario 3 b

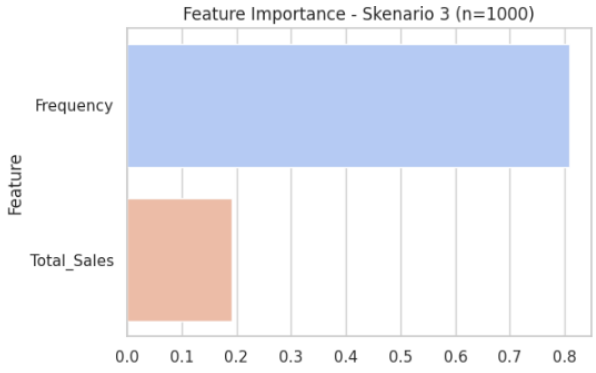
Pada skenario kedelapan (Skenario 3 b), model *Random Forest Regression* kembali dilatih menggunakan dua fitur utama, yakni *Total_Sales* dan *Frequency*,

sama seperti pada Skenario 3a. Perbedaannya terletak pada peningkatan jumlah pohon keputusan dari 500 menjadi 1000. Tujuan eksperimen ini adalah untuk menilai apakah penambahan jumlah estimators dapat memberikan peningkatan performa yang berarti ketika model hanya menggunakan dua fitur yang dianggap paling representatif dalam memprediksi nilai *Customer Lifetime Value* (CLV).

Tabel 4.10 Evaluasi Model Skenario 3 b

Parameter	**Nilai**
R^2	0.9325
MAE	0.10
MSE	0.437
MAPE	12.79%

Hasil evaluasi menunjukkan bahwa peningkatan jumlah n_estimators tidak menghasilkan perbaikan performa. Nilai R^2 sedikit menurun dari 0.9329 menjadi 0.9325, sehingga penambahan jumlah pohon tidak memberikan manfaat berarti terhadap kekuatan model dalam menjelaskan variasi CLV. Nilai MAE tetap pada 0.10, sementara MAPE berada pada kisaran 12.79%. Meskipun nilai MSE mengalami sedikit peningkatan (0.2264 → 0.437), perubahan tersebut masih berada dalam margin toleransi dan tidak mengubah interpretasi bahwa model tetap cukup stabil.data.



Gambar 4.12 Hasil Feature Importance Skenario 3 b

Hasil *feature importance* Gambar 4.12 menunjukkan pola yang identik dengan Skenario 3 a, yaitu fitur *Frequency* tetap menjadi variabel yang paling dominan dalam prediksi CLV. Dominasi ini muncul karena *Frequency* menghasilkan penurunan impurity (MSE) terbesar pada node-level awal di sebagian besar pohon, sehingga kontribusinya terhadap struktur model jauh lebih kuat dibandingkan *Total_Sales*. Fitur *Total_Sales* kembali menempati posisi kedua, memberikan kontribusi yang stabil namun tidak sebesar *Frequency*. Hal ini mencerminkan bahwa nilai transaksi pelanggan tetap relevan sebagai indikator CLV, namun intensitas pembelian yang ditangkap oleh *Frequency* lebih informatif dalam membedakan kategori pelanggan bernilai tinggi dan bernilai rendah.

4.6 Pembahasan

Pengujian model dilakukan menggunakan data transaksi pelanggan pada platform *e-commerce* yang telah melalui tahapan *pre-processing*, meliputi *cleaning data*, transformasi log pada variabel CLV (*log transformation*), normalisasi nilai jika diperlukan, serta pemisahan data latih dan uji menggunakan *train-test split* dengan rasio 80:20.

4.6.1 Analisis Pengaruh *Hyperparameter*

Pengaruh hyperparameter terhadap performa model terlihat dari stabilitas hasil pada seluruh skenario pengujian (0a hingga 3b). Proses tuning menggunakan *RandomizedSearchCV* menghasilkan kombinasi parameter optimal, yaitu *n_estimators* = 500, *max_depth* = 4, *min_samples_leaf* = 2, *max_features* = 3, dan

bootstrap = True. Setiap hyperparameter memiliki kontribusi penting terhadap kontrol kompleksitas model. Parameter *max_depth* membatasi kedalaman pohon untuk mencegah overfitting, *min_samples_leaf* memastikan bahwa setiap leaf node berisi cukup sampel sehingga prediksi lebih stabil, sementara *max_features* mengatur jumlah fitur yang dipertimbangkan pada setiap split sehingga tetap menjaga keragaman antar pohon namun tetap memungkinkan fitur-fitur informatif muncul pada node-level awal.

Selain itu, hasil eksperimen menunjukkan bahwa *max_depth* secara otomatis menyesuaikan kompleksitas model sesuai jumlah fitur yang digunakan pada setiap skenario. Pada skenario dengan empat fitur, kedalaman optimal tercapai pada max_depth = 4, yang memungkinkan model menangkap interaksi antar variabel *monetary*, *Frequency*, *Recency*, dan *quantity* tanpa menambah *noise*. Ketika jumlah fitur dikurangi menjadi tiga, kedalaman optimal turun menjadi *max_depth* = 3, karena berkurangnya kombinasi pola yang dapat dieksplorasi model menyebabkan struktur pohon yang lebih dangkal sudah cukup untuk memaksimalkan penurunan impurity. Pada skenario dengan dua fitur, kedalaman optimal kembali menurun menjadi max_depth = 2, yang secara teoritis konsisten karena model hanya dapat membangun kombinasi split yang terbatas ketika variasi fitur semakin kecil. Penyesuaian kedalaman pohon ini menunjukkan bahwa *Random Forest* secara alami mengurangi kompleksitas struktural ketika informasi input semakin sedikit, sehingga tetap menjaga generalisasi model.

Temuan penting lain yang konsisten pada semua skenario adalah bahwa peningkatan *n_estimators* dari 500 ke 1000 tidak memberikan peningkatan

performa yang berarti. Fenomena ini sesuai dengan teori *diminishing returns* pada *Random Forest*, di mana ketika jumlah pohon telah cukup banyak, model memasuki kondisi *low variance state* yaitu dimana penambahan pohon tidak lagi mengurangi error. Pada skenario dengan fitur terbatas (khususnya skenario dua fitur), peningkatan jumlah pohon bahkan tidak dapat memperbaiki performa karena keterbatasan fitur membuat model kekurangan variasi informasi untuk diolah, meskipun jumlah pohon ditambah.

4.6.2 Analisis Pengaruh Jumlah Fitur

Variasi jumlah fitur pada skenario 0–3 menunjukkan bahwa performa model sangat dipengaruhi oleh kelengkapan dan kualitas informasi yang terkandung dalam fitur input. Ketika model menggunakan empat fitur utama (*Total_Sales*, *Frequency*, *Recency*, *Total_Quantity*), model mampu mencapai performa terbaik pada skenario *baseline* ($R^2 \approx 0.94$ dan MAPE < 20%). Ini menunjukkan bahwa keempat dimensi tersebut secara bersama-sama memberikan representasi perilaku pelanggan yang cukup komprehensif: nilai transaksi (*monetary*), intensitas pembelian (*Frequency*), waktu pembelian terakhir (*Recency*), dan volume pembelian (*quantity*).

Ketika satu fitur dihapus (misalnya *Total_Quantity* pada Skenario 1), performa model tidak mengalami penurunan dan tetap berada pada rentang $R^2 \approx 0.93$. Hal ini mengindikasikan bahwa sebagian informasi dari *Total_Quantity* sudah terkandung dalam *Total_Sales* dan *Frequency*, sehingga model masih mampu menangkap pola CLV dengan akurat. Namun, ketika dua fitur dihilangkan dan model hanya mengandalkan *Total_Sales* dan *Recency* (Skenario 2), performa turun drastis ($R^2 \approx 0.75$ dan MAPE > 60%). Penurunan ini membuktikan bahwa informasi

mengenai loyalitas dan pola pembelian berulang pelanggan tidak dapat digantikan oleh dua fitur tersebut. Secara teoretis, mengurangi fitur yang informatif meningkatkan *bias* model, dan tidak dapat diperbaiki hanya dengan menambah pohon atau tuning *hyperparameter*. Oleh karena itu, jumlah dan kualitas fitur terbukti menjadi determinan utama keberhasilan prediksi CLV.

4.6.3 Analisis Feature Importance dan Dominasi *Frequency*

Hasil analisis *feature importance* pada seluruh skenario konsisten menunjukkan bahwa *Frequency* adalah fitur yang paling dominan dalam mempengaruhi prediksi CLV. Dominasi ini dapat dijelaskan dari dua perspektif. Secara konseptual, frekuensi transaksi mencerminkan tingkat loyalitas pelanggan dan pola pembelian berulang, yang merupakan komponen inti dalam perhitungan *Customer Lifetime Value* . Pelanggan yang sering bertransaksi memiliki kontribusi pendapatan yang lebih stabil dan berkelanjutan dibandingkan pelanggan dengan transaksi sporadis, sehingga secara alamiah fitur *Frequency* membawa sinyal yang kuat terhadap nilai CLV. Secara teknis, *Random Forest* menentukan importance berdasarkan *Mean Decrease in Impurity* (MDI), di mana fitur yang menghasilkan penurunan impurity (MSE) paling besar pada split awal akan mendapat skor *importance* lebih tinggi. *Frequency* secara konsisten menghasilkan pemisahan data yang paling informatif pada node tingkat atas di banyak pohon, sehingga efek kumulatifnya menghasilkan skor *importance* yang jauh lebih besar. Dominasi *Frequency* yang stabil di seluruh skenario juga menunjukkan bahwa struktur internal model tidak berubah meskipun jumlah pohon ditambah atau fitur lain dikurangi.

4.6.4 Perbandingan Performa Antar Skenario

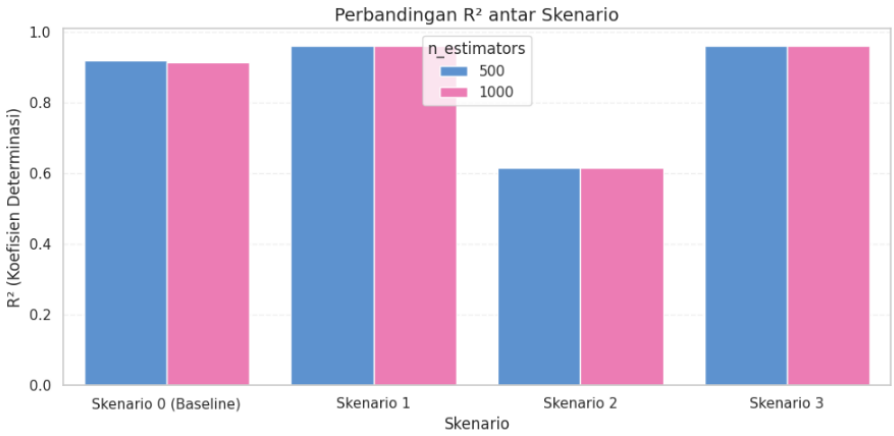
Perbandingan performa antar skenario menunjukkan pola yang jelas dan konsisten. Proses pengujian dilakukan secara bertahap menggunakan beberapa skenario pemilihan fitur dan variasi jumlah pohon (*n_estimators*) pada algoritma *Random Forest Regression*. Setiap skenario diuji untuk mengamati pengaruh fitur terhadap performa model, dengan fokus utama pada metrik evaluasi R^2, MAE, MSE, RMSE, dan MAPE. Berikut hasil perbandingan dari hasil uji coba beberapa skenario:

Tabel 4.11 Hasil Perbandingan Skenario

Skenario		**R^2**	**MAE**	**MSE**	**MAPE**
Skenario 0 (*Baseline*)	a	0.9412	0.04	0.0380	17.96%
	b	0.9407	0.04	0.0383	17.95%
Skenario 1	a	0.9331	0.05	0.0433	12.78%
	b	0.9327	0.05	0.0436	12.79%
Skenario 2	a	0.7519	0.12	0.1605	66.12%
	b	0.7521	0.12	0.1604	66.09%
Skenario 3	a	0.9329	0.05	0.0434	12.78%
	b	0.9325	0.05	0.0437	12.79%

Berdasarkan hasil evaluasi pada Tabel 4.11, diperoleh bahwa Skenario 1 dan Skenario 3 sama-sama menghasilkan nilai MAPE terendah yaitu sekitar 12.78%–12.79%, yang termasuk dalam kategori *Good Forecast* menurut klasifikasi (Meade, 1983). Kondisi ini menunjukkan bahwa kedua skenario memiliki tingkat kesalahan prediksi relatif yang setara dan sama-sama layak digunakan dari sisi akurasi. Prinsip ini menyatakan bahwa model yang lebih sederhana yaitu model yang lebih baik (Hastie et al., 2017; Hyndman & Athanasopoulos, 2021). Dalam penelitian ini, Skenario 3 dipilih sebagai skenario terbaik karena hanya menggunakan dua fitur utama, yaitu *Total_Sales* dan *Frequency*, dibandingkan Skenario 1 yang masih

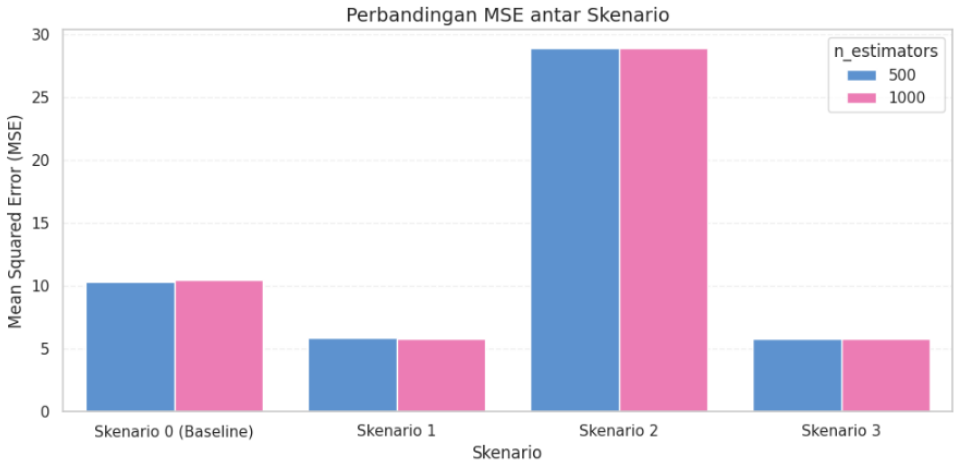
menggunakan tiga fitur. Meskipun jumlah fitur lebih sedikit, Skenario 3 tetap mampu mempertahankan nilai MAPE rendah (~12.78%), MAE kecil (~0.05), serta R^2 tinggi (~0.9329), yang menunjukkan bahwa model masih mampu menjelaskan sebagian besar variasi nilai CLV secara baik.



Gambar 4.13 Hasil Perbandingan R2 Score

Pada Gambar 4.13 performa model meningkat ketika fitur *Total_Quantity* dihapus sehingga model hanya menggunakan *Total_Sales*, *Frequency*, dan *Recency*. Pada Skenario 1 (n = 500), model menghasilkan nilai R^2 = 0.9331 dan MAPE = 12.78%, yang lebih baik dibandingkan *baseline*. Hasil ini menunjukkan bahwa penghapusan fitur *Total_Quantity* mampu mengurangi kompleksitas model dan meningkatkan akurasi prediksi, sehingga fitur tersebut berpotensi menimbulkan noise dalam model. Pada Skenario 1 (n = 1000), peningkatan jumlah pohon tidak memberikan perubahan yang berarti, dengan nilai R^2 = 0.9327 dan MAPE = 12.79%. Hal ini menunjukkan bahwa model telah mencapai kondisi konvergen, sehingga konfigurasi optimal tetap berada pada penggunaan tiga fitur utama dengan jumlah pohon 500. Ketika fitur *Frequency* dikeluarkan dan model hanya

menggunakan *Total_Sales* dan *Recency* (Skenario 2), performa model menurun drastis dengan nilai R^2 turun menjadi 0.7519 dan MAPE meningkat menjadi 66.12%. Temuan ini menunjukkan bahwa *Frequency* merupakan variabel paling krusial dalam prediksi CLV, yang sejalan dengan hasil analisis feature importance sebelumnya.

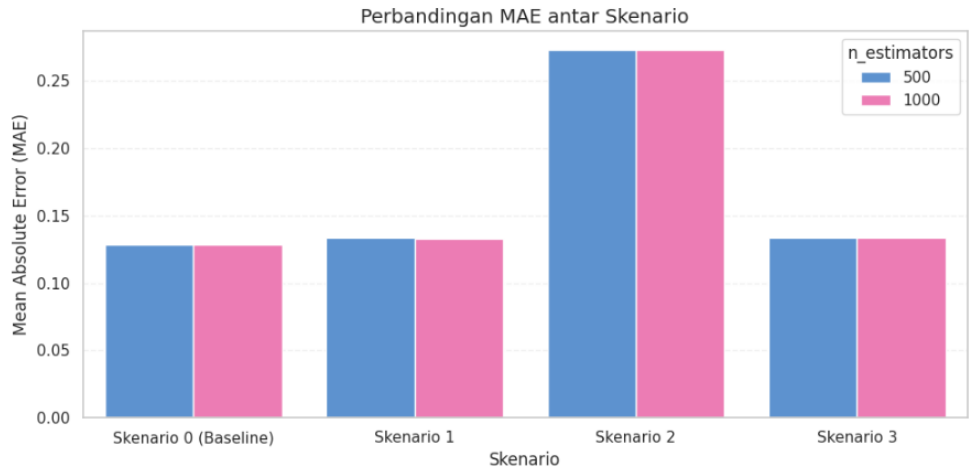


Gambar 4.14 Hasil Perbandingan MSE

Gambar 4.14 menunjukan perbandingan nilai *Mean Squared Error* (MSE) antar skenario menunjukkan bahwa Skenario 0 (*baseline*) dan Skenario 1 memiliki nilai MSE paling rendah, berada pada kisaran 0.038–0.043. Nilai yang rendah ini mengindikasikan bahwa kombinasi tiga hingga empat fitur utama khususnya, *Total_Sales*, dan *Recency* memberikan sinyal yang kuat sehingga model mampu meminimalkan kesalahan kuadrat secara konsisten. Keberadaan fitur *Frequency* sangat berperan dalam menghasilkan penurunan impurity MSE pada setiap split dalam *Random Forest*, sehingga model pada skenario ini mampu memprediksi CLV dengan stabil dan efektif. Selain itu, kestabilan MSE antara n_estimators 500 dan 1000 menunjukkan bahwa peningkatan jumlah pohon tidak memberikan

peningkatan, yang menandakan bahwa model telah mencapai konvergensi performa (Probst & Boulesteix, 2017).

Sebaliknya, Skenario 2 menunjukkan lonjakan MSE dengan nilai sekitar 0.1604–0.1605, jauh di atas skenario lainnya. Hal ini terjadi karena penghapusan fitur *Frequency* menyebabkan model kehilangan informasi kritis terkait perilaku pembelian berulang, sehingga kesalahan prediksi pelanggan dengan CLV tinggi meningkat drastis. Sementara itu, Skenario 3 kembali mencatat MSE rendah berkat kombinasi *Total_Sales* dan *Frequency*, menunjukkan bahwa kedua fitur ini sudah cukup informatif untuk meminimalkan kesalahan kuadrat secara efektif. Temuan ini memperkuat bahwa *Frequency* merupakan variabel paling penting dalam model prediksi CLV berbasis *Random Forest Regression*.

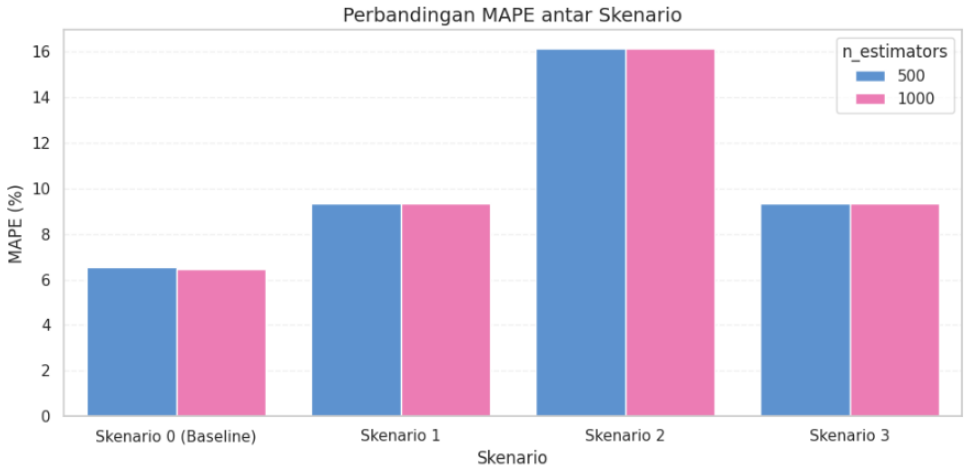


Gambar 4.15 Hasil Perbandingan MAE

Pada Gambar 4.15 menunjukan perbandingan nilai *Mean Absolute Error* (MAE) pada setiap skenario menunjukkan konsistensi performa model ketika fitur-fitur utama yang informatif dipertahankan. Skenario 0 dan Skenario 1 mencatat

MAE terendah, masing-masing berada pada kisaran 0.04–0.05, yang menandakan bahwa rata-rata selisih absolut antara nilai prediksi dan nilai aktual CLV sangat kecil. MAE yang rendah pada dua skenario ini menunjukkan bahwa fitur *Frequency*, *Total_Sales*, *Recency*, dan *Total_Quantity* secara bersama-sama memberikan representasi yang memadai terhadap perilaku pelanggan. Bahkan ketika *Total_Quantity* dihilangkan (Skenario 1), MAE hanya sedikit meningkat, menunjukkan bahwa tiga fitur utama masih mampu menjaga kesalahan absolut tetap stabil. Selain itu, Skenario 3 (kombinasi *Total_Sales* dan *Frequency*) juga menunjukkan MAE yang rendah (~0.05), menegaskan bahwa kedua fitur ini sudah cukup kuat untuk menghasilkan prediksi yang presisi tanpa deviasi absolut yang besar.

Sebaliknya, Skenario 2 menunjukkan MAE yang meningkat drastis menjadi 0.12, dua hingga tiga kali lebih tinggi dibandingkan skenario lainnya. Kenaikan ini merefleksikan hilangnya kemampuan model untuk memprediksi CLV secara akurat ketika fitur *Frequency* dihapus. Tanpa informasi mengenai intensitas belanja pelanggan, model tidak mampu membedakan pelanggan berulang dengan pelanggan yang jarang melakukan transaksi, sehingga kesalahan absolut meningkat . Perbandingan ini menegaskan bahwa meskipun MAE tidak sepeka MSE terhadap outlier, perubahan MAE pada Skenario 2 cukup besar untuk menunjukkan underfitting yang berat. Konsistensi MAE rendah pada Skenario 0, 1, dan 3 menunjukkan bahwa selama fitur *Frequency* dipertahankan, model mampu menghasilkan deviasi absolut yang kecil, sedangkan hilangnya fitur tersebut langsung meningkatkan tingkat kesalahan prediksi.



Gambar 4.16 Hasil Perbandingan MAPE

Pada Gambar 4.16 menunjukan nilai *Mean Absolute Percentage Error* (MAPE) pada setiap skenario menunjukkan perbedaan yang paling mencolok dalam hal kemampuan model melakukan prediksi secara relatif terhadap nilai CLV aktual. Skenario 0 dan Skenario 1 menghasilkan MAPE yang rendah, berada pada rentang 12.78%–17.96%, yang seluruhnya masuk dalam kategori *Good Forecast* menurut klasifikasi Lewis (1982), yaitu rentang MAPE 10–20% (Montaño Moreno et al., 2013). Hal ini menunjukkan bahwa model mampu memberikan estimasi CLV dengan tingkat kesalahan relatif yang cukup kecil, bahkan ketika satu fitur seperti *Total_Quantity* dihilangkan. Skenario 3, yang hanya menggunakan dua fitur utama (*Total_Sales* dan *Frequency*), juga mencatat MAPE terbaik di seluruh skenario (~12.78%), mempertegas bahwa *Frequency* adalah fitur yang paling menentukan stabilitas kesalahan relatif dan sangat efektif dalam membantu model menghasilkan prediksi yang konsisten dan akurat.

Sebaliknya, Skenario 2 menunjukkan MAPE yang sangat tinggi, yaitu 66.09%–66.12%. Tingginya MAPE pada skenario ini mengindikasikan bahwa model gagal

memberikan prediksi CLV yang akurat dalam perspektif persentase, terutama untuk pelanggan bernilai tinggi. Hal ini disebabkan hilangnya fitur *Frequency* yang merupakan indikator utama perilaku pembelian berulang, sehingga model tidak mampu mengukur besarnya nilai CLV secara proporsional. Perbandingan ini menunjukkan bahwa MAPE adalah metrik yang paling sensitif dalam mendeteksi penurunan kualitas model ketika informasi penting mengenai loyalitas pelanggan dihilangkan. Dengan demikian, MAPE memperkuat temuan dari MSE dan MAE bahwa *Frequency* merupakan fitur paling krusial, dan model hanya mampu memberikan prediksi yang akurat ketika fitur tersebut dipertahankan.

4.7 Integrasi Islam

Penelitian mengenai prediksi *Customer Lifetime Value* (CLV) menggunakan algoritma *Random Forest Regression* secara langsung berkaitan dengan pengelolaan transaksi dan perilaku pembelian pelanggan dalam platform *e-commerce*. Oleh karena itu, integrasi nilai Islam dalam penelitian ini difokuskan pada prinsip-prinsip syariah dalam hal keadilan transaksi, transparansi data, etika perdagangan, serta amanah dalam pengelolaan informasi pelanggan**.** Integrasi ini dijelaskan melalui tiga dimensi utama yaitu: Hubungan manusia dengan Allah Swt (*Muamalah Ma'a Allah*), hubungan manusia dengan sesama (*Muamalah Ma'a An-Nas*), serta hubungan manusia dengan alam dan lingkungan (*Muamalah Ma'al Bi'ah*).

4.7.1 *Muamalah Ma'a Allah*

Muamalah Ma'a Allah berkaitan dengan hubungan manusia dengan Allah Swt sebagai dasar dari seluruh aktivitas transaksi atau perdagangan. Penggunaan ilmu pengetahuan dan teknologi seperti machine learning merupakan bagian dari upaya menjalankan amanah Allah untuk berbuat adil dan profesional dalam perdagangan. Proses analisis data, pemodelan prediksi CLV, serta pengambilan keputusan berbasis data mencerminkan penggunaan akal (‘aql) secara bertanggung jawab.

Hal ini sejalan dengan firman Allah Swt dalam QS. Ar-Rahman ayat 9:

وَاَقِيْمُوا الْوَزْنَ بِالْقِسْطِ وَلَا تُخْسِرُوا الْمِيْزَانَ ۝٩

“Dan tegakkanlah timbangan itu dengan adil dan janganlah kamu mengurangi neraca itu.” (QS. Ar-Rahman ayat 9)

Ayat ini menekankan prinsip keadilan, akurasi, dan ketepatan dalam setiap bentuk perhitungan atau transaksi. Dalam penelitian ini, penggunaan algoritma *Random Forest* untuk memprediksi CLV merupakan bentuk upaya menghasilkan perhitungan yang tepat dan adil dalam mengelola nilai pelanggan, sehingga keputusan bisnis tidak bertumpu pada spekulasi, tetapi pada analisis objektif.

Tafsir Ibnu Katsir menjelaskan bahwa Allah membinasakan kaum Syu’aib karena mereka curang dalam takaran dan timbangan. Relevansinya dalam konteks digital adalah bahwa penggunaan data dan model analitik harus dilakukan dengan penuh amanah, tanpa manipulasi dan tanpa merugikan pelanggan melalui prediksi atau strategi pemasaran yang tidak etis (*Tafsir-Surat-Ar-Rahman-Ayat-1-13.Html*, n.d.).

Dengan demikian, muamalah dengan Allah dalam penelitian ini tercermin dari penggunaan ilmu secara amanah, akurat, dan berdasarkan nilai keadilan sebagaimana diperintahkan dalam Al-Qur'an.

4.7.2 *Muamalah Ma'a An-Nas*

Muamalah Ma'a An-Nas berkaitan dengan hubungan manusia dengan sesama, terutama dalam konteks transaksi dan interaksi ekonomi. Dalam *e-commerce*, pengelolaan pelanggan berdasarkan CLV harus dilakukan dengan prinsip kejujuran, tidak merugikan, dan tidak menipu pelanggan.

Allah Swt berfirman dalam QS. An-Nisa ayat 29:

يٰٓاَيُّهَا الَّذِيْنَ اٰمَنُوْا لَا تَأْكُلُوْٓا اَمْوَالَكُمْ بَيْنَكُمْ بِالْبَاطِلِ اِلَّآ اَنْ تَكُوْنَ تِجَارَةً عَنْ تَرَاضٍ مِّنْكُمْۗ وَلَا تَقْتُلُوْٓا اَنْفُسَكُمْۗ

اِنَّ اللّٰهَ كَانَ بِكُمْ رَحِيْمًا ﴿٢٩﴾

"Wahai orang-orang yang beriman, janganlah kamu memakan harta sesamamu dengan cara yang batil (tidak benar), kecuali berupa perniagaan atas dasar suka sama suka di antara kamu. Janganlah kamu membunuh dirimu. Sesungguhnya Allah adalah Maha Penyayang kepadamu." (*QS. An-Nisa ayat 29)*

Menurut Tafsir Tahlili Ayat ini melarang mengambil harta orang lain dengan jalan yang batil (tidak benar), kecuali dengan perniagaan yang berlaku atas dasar kerelaan bersama. Menurut ulama tafsir, larangan memakan harta orang lain dalam ayat ini mengandung pengertian yang luas dan dalam, antara lain: a. Agama Islam mengakui adanya hak milik pribadi yang berhak mendapat perlindungan dan tidak boleh diganggu gugat. b. Hak milik pribadi, jika memenuhi nisabnya, wajib dikeluarkan zakatnya dan kewajiban lainnya untuk kepentingan agama, negara dan sebagainya. c. Sekalipun seseorang mempunyai harta yang banyak dan banyak pula

orang yang memerlukannya dari golongan-golongan yang berhak menerima zakatnya, tetapi harta orang itu tidak boleh diambil begitu saja tanpa seizin pemiliknya atau tanpa menurut prosedur yang sah.

Mencari harta dibolehkan dengan cara berniaga atau berjual beli dengan dasar kerelaan kedua belah pihak tanpa suatu paksaan. Karena jual beli yang dilakukan secara paksa tidak sah walaupun ada bayaran atau penggantinya. Dalam upaya mendapatkan kekayaan tidak boleh ada unsur zalim kepada orang lain, baik individu atau masyarakat. Tindakan memperoleh harta secara batil, misalnya mencuri, riba, berjudi, korupsi, menipu, berbuat curang, mengurangi timbangan, suap-menyuap, dan sebagainya. Selanjutnya Allah melarang membunuh diri .

Penerapan nilai *Muamalah Ma'a An-Nas* pada ayat ini dalam konteks CLV adalah bahwa setiap strategi bisnis yang dihasilkan dari model prediktif harus menjaga prinsip ridha antara pelanggan dan perusahaan, memastikan bahwa peningkatan keuntungan tidak mengorbankan hak-hak pelanggan. Selain itu, penggunaan data transaksi dalam penelitian ini juga harus dilakukan dengan menjaga hak milik informasi pelanggan, tidak menyalahgunakan data, dan tidak menggunakan hasil analisis untuk tindakan yang merugikan mereka. Dengan demikian, muamalah dengan sesama mengharuskan setiap aktivitas analisis dan implementasi model prediksi dilakukan dengan jujur, adil, dan berorientasi pada kemaslahatan bersama.

4.7.3 *Muamalah Ma'al Bi'ah*

Muamalah Ma'al Bi'ah dalam konteks modern mencakup tanggung jawab manusia dalam menjaga sistem, sumber daya informasi, dan lingkungan digital

yang digunakan dalam kegiatan bisnis. Data transaksi pelanggan, yang digunakan sebagai dasar perhitungan CLV, merupakan amanah yang harus dikelola secara benar, akurat, dan bertanggung jawab.

Allah Swt berfirman dalam QS. Al-Baqarah ayat 282:

يٰٓاَيُّهَا الَّذِيْنَ اٰمَنُوْٓا اِذَا تَدَايَنْتُمْ بِدَيْنٍ اِلٰٓى اَجَلٍ مُّسَمًّى فَاكْتُبُوْهُۗ

"Wahai orang-orang yang beriman, apabila kamu berutang piutang untuk waktu yang ditentukan, hendaklah kamu mencatatnya." (*QS. Al-Baqarah ayat 282)*

Menurut Tafsir Al Jalalain (*Terjemah Tafsir Jalalain Jilid 1 (1)*, n.d.) maksudnya "muamarat" seperti jual-beli, sewa-menyewa, utang-piutang, dan lain lain. (secara tidak tunai) misalnya pinjaman atau pesanan – (untuk waktu yang ditentukan) atau diketahui, (maka hendaklah kamu tuliskan) untuk pengukuhan dan menghilangkan pertikaian nantinya (Al-Fatihah, n.d.). Dalam penelitian ini, penerapan nilai *Muamalah Ma'al Bi'ah* ayat tersebut dapat ditarik relevansinya pada proses pencatatan, penyimpanan, dan pengelolaan data digital sebagai bentuk dokumentasi modern. Kesalahan pencatatan transaksi atau penyalahgunaan data dapat menghasilkan prediksi CLV yang keliru dan berpotensi menimbulkan kebijakan yang merugikan pelanggan, sehingga bertentangan dengan prinsip amanah. Selain itu, pemanfaatan teknologi prediktif dianggap sebagai upaya untuk menciptakan sistem ekonomi yang lebih efisien dan berkelanjutan, sehingga mencerminkan nilai kemaslahatan (*jalbul mashalih*) dan pencegahan kerusakan (*dar'ul mafasid*) dalam lingkungan digital.

BAB V

KESIMPULAN DAN SARAN

5.1 Kesimpulan

Berdasarkan hasil penelitian mengenai prediksi *Customer Lifetime Value* (CLV) menggunakan *Random Forest Regression*, diperoleh beberapa kesimpulan utama sebagai jawaban atas rumusan masalah. Model *Random Forest* mampu memberikan performa prediksi yang sangat baik, ditunjukkan oleh nilai R^2 tinggi pada skenario terbaik yaitu 0.9412, MAE 0.04, MSE 0.0380, serta MAPE 17.96%. Hal ini menunjukkan bahwa model dapat menjelaskan sebagian besar variasi CLV dan menghasilkan kesalahan prediksi yang rendah. Hasil analisis feature importance menunjukkan bahwa *Frequency* merupakan faktor yang paling berpengaruh dalam prediksi CLV karena memberikan penurunan impurity MSE terbesar pada proses pembentukan pohon. Fitur *Total_Sales* juga berperan sebagai indikator nilai transaksi pelanggan. Sementara itu, *Total_Quantity* memberikan kontribusi tambahan namun tidak dominan, dan *Recency* menjadi fitur dengan pengaruh paling rendah terhadap prediksi CLV pada dataset ini. Dengan demikian, CLV terutama dipengaruhi oleh intensitas transaksi dan nilai pembelian pelanggan.

5.2 Saran

Penelitian selanjutnya dapat mempertimbangkan penambahan fitur perilaku pelanggan yang lebih beragam, seperti kategori produk yang sering dibeli, rata-rata nilai transaksi, variasi jumlah pembelian, serta waktu antar transaksi (*interpurchase time*). Penambahan fitur-fitur tersebut berpotensi memberikan gambaran yang lebih

komprehensif mengenai karakteristik pelanggan sehingga dapat meningkatkan akurasi model dalam memprediksi nilai *Customer Lifetime Value* (CLV). Selain itu, penelitian lanjutan juga disarankan untuk membandingkan performa *Random Forest Regression* dengan algoritma lain seperti XGBoost, LightGBM, atau *Gradient Boosting Regression* guna melihat apakah model berbasis *boosting* mampu memberikan hasil prediksi yang lebih stabil dan presisi.

DAFTAR PUSTAKA

Ahmadi, A. A., Ahmad Baloch, F., Mohammad Wafa, K., & Naeem Dost, M. (2022). An Empirical Study of Impact of Electronic Commerce on Business. *Journal of Information Systems and Technology Research*, *1*(3), 150–157. https://doi.org/10.55537/jistr.v1i3.213

Al-Fatihah, S. (n.d.). *Tafsir al-Jalalain*.

Ardian, M., Khomsah, S., & Pandiya, R. (n.d.). *Perbandingan Model Regresi Untuk Memprediksi Harga Jual Cabai Rawit Berdasarkan Iklim Harian*.

Aulia, D. (2022). Enhancements in the management of relationships with customers as a means of preserving sales performance. *Journal of Applied Management and Business (JAMB)*, *3*(1). https://doi.org/10.37802/jamb.v3i1.242

Azmi, A. F., & Voutama, A. (2024). Prediksi Churn Nasabah Bank Menggunakan Klasifikasi *Random Forest* Dan Decision Tree Dengan Evaluasi Confusion Matrix. *Komputa : Jurnal Ilmiah Komputer Dan Informatika*, *13*(1), 111–119. https://doi.org/10.34010/komputa.v13i1.12639

Bakir, F. T. (n.d.-a). *Customer Lifetime Value prediction and segmentation analysis for commercial customers in the banking industry*.

Balabanova, V., & Bhattarai, S. (n.d.). *Comparative Analysis of Machine Learning Algorithms on Comprehensive and Cluster-Specific Data in the Auto Insurance Industry*.

Berger, P. D., & Nasr, N. I. (1998). *Customer Lifetime Value* : Marketing models and applications. *Journal of Interactive Marketing*, *12*(1), 17–30. https://doi.org/10.1002/(SICI)1520-6653(199824)12:1%253C17::AID-DIR3%253E3.0.CO;2-K

Billah, I. S. M. (n.d.). *Program studi teknik informatika fakultas sains dan teknologi universitas islam negeri maulana malik ibrahim malang 2024*.

Chicco, D., Warrens, M. J., & Jurman, G. (2021). The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. *PeerJ Computer Science*, *7*, e623. https://doi.org/10.7717/peerj-cs.623

Curiskis, S., Dong, X., Jiang, F., & Scarr, M. (2023). A novel approach to predicting *Customer Lifetime Value* in B2B SaaS companies. *Journal of Marketing Analytics*, *11*(4), 587–601. https://doi.org/10.1057/s41270-023-00234-6

D, R. L., Costa, A., Cunha, A., Gonçalves, R., Pereira, L., Dias, Á., D, R. V., & Silva, A. (2023). The strategic impact of information systems in organisations: An empirical study. *International Journal of Applied Decision Sciences*, *16*(1), 87. https://doi.org/10.1504/IJADS.2023.127943

Edi, M. R. (n.d.). *Diajukan kepada: Universitas Islam Negeri Maulana Malik Ibrahim Malang Untuk memenuhi Salah Satu Persyaratan dalam Memperoleh Gelar Sarjana Komputer (S.Kom)*.

Fader, P. S., Hardie, B. G. S., & Shang, J. (2010). Customer-Base Analysis in a Discrete-Time Noncontractual Setting. *Marketing Science*, *29*(6), 1086–1108. https://doi.org/10.1287/mksc.1100.0580

Gerde, M. (n.d.). *Predicting Customer Churn and Customer Lifetime Value (CLV) using Machine Learning*.

Gupta, S., & Lehmann, D. R. (2006). *Customer Lifetime Value* and Firm Valuation. *Journal of Relationship Marketing*, *5*(2–3), 87–110. https://doi.org/10.1300/J366v05n02_06

Husna, A. N. (2020). Analisis penerapan customer relationship management dan perhitungan *Customer Lifetime Value* untuk meningkatkan profitabilitas pelanggan pada diponegoro printing. *ABIS: Accounting and Business Information Systems Journal*, *7*(2). https://doi.org/10.22146/abis.v7i2.58833

Hwang, S.-W., Chung, H., Lee, T., Kim, J., Kim, Y., Kim, J.-C., Kwak, H. W., Choi, I.-G., & Yeo, H. (2023). Feature importance measures from *Random Forest* regressor using near-infrared spectra for predicting carbonization characteristics of kraft lignin-derived hydrochar. *Journal of Wood Science*, *69*(1), 1. https://doi.org/10.1186/s10086-022-02073-y

Kabiraj, S., Gupta, A., Chandra, Prof. S. K., & T.D.B college. (2018). Operating System a Case Study. *International Journal of Trend in Scientific Research and Development*, *Volume-2*(Issue-3), 166–175. https://doi.org/10.31142/ijtsrd10780

Kuijt, C. (n.d.). *Prediction of Customer Lifetime Value in e-commerce fashion retail*.

Lathwal, P., & Batra, R. (2024). *Attention-Based Customer Lifetime Value Prediction in E-commerce Using FT-Transformer Architecture*. *2*(1).

Maliyekkal, C. D. (n.d.). *Predicting Customer Lifetime Value (CLV) in UK and Brazil using Machine Learning and Deep Learning: A Comparative Analysis*.

Master-Thesis-Olivier-Bax-final-version. (n.d.).

Meade, N. (1983). Industrial and business forecasting methods, Lewis, C.D., Borough Green, Sevenoaks, Kent: Butterworth, 1982. Price: £9.25. Pages: 144. *Journal of Forecasting*, *2*(2), 194–196. https://doi.org/10.1002/for.3980020210

Menaga.A, Sangeetha.T, K B, S., Abishek.S, Swetha.G, & Kumaran, S. (2025). Predictive Modelling of *Customer Lifetime Value* Using AI and Machine Learning Algorithms. *2025 International Conference on Sensors and Related Networks (SENNET) Special Focus on Digital Healthcare(64220)*, 1–6. https://doi.org/10.1109/SENNET64220.2025.11136022

Montaño Moreno, J., Palmer Pol, A., Sesé Abad, A., & Cajal Blasco, B. (2013). Using the R-MAPE index as a resistant measure of forecast accuracy. *Psicothema*, *4*(25), 500–506. https://doi.org/10.7334/psicothema2013.23

Nurfadilla Nurfadilla, Syamsul Bachri, Elimawaty Rombe, & Farid Farid. (2024). Pemanfaatan *E-commerce* Dalam Pengembangan Usaha Bawang Goreng Di Desa Oloboju Kabupaten Sigi. *Jurnal Riset Rumpun Ilmu Ekonomi*, *3*(1), 12–20. https://doi.org/10.55606/jurrie.v3i1.2592

Probst, P., & Boulesteix, A.-L. (2017). *To tune or not to tune the number of trees in Random Forest?* https://doi.org/10.48550/ARXIV.1705.05654

Raditya, D. D. (n.d.). *Analisis customer relationship management terhadap customer life-time value dengan customer statisfaction sebagai variabel mediasi*.

Sharma, A., Patel, N., & Gupta, R. (n.d.). *Enhancing Customer Lifetime Value Prediction Using Random Forests and Neural Network Ensemble Methods*.

Tafsir-surat-ar-rahman-ayat-1-13.html. (n.d.). Retrieved December 4, 2025, from http://www.ibnukatsironline.com/2015/10/tafsir-surat-ar-rahman-ayat-1-13.html

Taherkhani, L., Daneshvar, A., Amoozad Khalili, H., & Sanaei, M. (2025b). Analysis and Optimization of *Customer Lifetime Value* Prediction using Machine Learning and Deep Learning Models by RFM Techniques. *International Journal of Web Research*, *8*(2). https://doi.org/10.22133/ijwr.2025.508322.1272

Terjemah Tafsir Jalalain Jilid 1 (1). (n.d.).

Völcker, M., & StenfeltROYAL, C. (n.d.). *Modellering av Customer Lifetime Value inom retail banking- branschen*.

Win, T. T., & Bo, K. S. (2020). Predicting Customer Class using *Customer Lifetime Value* with *Random Forest* Algorithm. *2020 International Conference on*

Advanced Information Technologies (ICAIT), 236–241. https://doi.org/10.1109/ICAIT51105.2020.9261792

Wu, C., Li, J., Jia, Q., Zhu, H., Fang, Y., & Tang, R. (2023). *Contrastive Multi-view Framework for Customer Lifetime Value Prediction* (No. arXiv:2306.14400). arXiv. https://doi.org/10.48550/arXiv.2306.14400

Yan, Y., & Resnick, N. (2024). A high-performance turnkey system for *Customer Lifetime Value* prediction in retail brands: Forthcoming in quantitative marketing and economics. *Quantitative Marketing and Economics*, *22*(2), 169–192. https://doi.org/10.1007/s11129-023-09272-x

Yılmaz Benk, G., Badur, B., & Mardikyan, S. (2022). A New 360° Framework to Predict *Customer Lifetime Value* for Multi-Category *E-commerce* Companies Using a Multi-Output Deep Neural Network and Explainable Artificial Intelligence. *Information*, *13*(8), 373. https://doi.org/10.3390/info13080373

You, J. (2025). *Customer Lifetime Value* Forecasting Using Ensemble Learning on *E-commerce* Big Data. *Proceedings of the 2025 International Conference on Digital Economy and Intelligent Computing*, 49–53. https://doi.org/10.1145/3746972.3746981