

BAB I

PENDAHULUAN

1.1 Latar Belakang

Transformasi digital yang masif telah meredefinisi lanskap ekonomi global, menempatkan perdagangan elektronik (*e-commerce*) sebagai pilar utama pertumbuhan ekonomi modern, khususnya di wilayah Asia-Pasifik yang kini mendominasi pangsa pasar dunia. Di Indonesia, fenomena ini tercermin dari lonjakan volume transaksi *e-commerce* yang diproyeksikan mencapai USD 75 miliar pada tahun 2024, dengan tingkat pertumbuhan tahunan gabungan (*CAGR*) sebesar 19% menuju tahun 2027 [1]. Pertumbuhan ini didorong oleh adopsi internet yang meluas dan perubahan perilaku konsumen pasca pandemi *COVID-19* yang memprioritaskan kenyamanan serta efisiensi, menjadikan belanja daring sebagai norma budaya baru. Namun, di balik angka pertumbuhan yang mengesankan ini, industri *e-commerce* menghadapi tantangan yang intens seiring dengan meningkatnya kemudahan konsumen dalam memilih maupun meninggalkan suatu *platform*. Fenomena ini menciptakan lingkungan bisnis yang kompetitif, di mana loyalitas pelanggan menjadi aset yang semakin langka dan berharga.

Di tengah dinamika kompetisi yang ketat tersebut, urgensi untuk mempertahankan pelanggan (*customer retention*) menjadi jauh lebih kritikal dibandingkan akuisisi pelanggan baru. Data statistik industri secara konsisten menunjukkan bahwa Biaya Akuisisi Pelanggan (*Customer Acquisition Cost - CAC*) di sektor *e-commerce* terus meningkat, mencapai rata-rata USD 84 untuk *B2B* dan USD 68 untuk *B2C* per pelanggan baru pada tahun 2025, yang nilainya dapat mencapai 5 hingga 25 kali lebih besar dibandingkan biaya mempertahankan pelanggan lama. Lebih jauh lagi, peningkatan tingkat retensi sebesar 5% saja berpotensi meningkatkan profitabilitas perusahaan antara 25% hingga 95% [2]. Kondisi ini menegaskan bahwa *customer churn* atau kondisi ketika pelanggan berhenti melakukan transaksi bukan sekadar permasalahan operasional, melainkan ancaman strategis yang berdampak langsung terhadap pendapatan dan profitabilitas perusahaan. Oleh karena itu, kemampuan untuk mengidentifikasi pelanggan yang berpotensi *churn* secara dini melalui pemanfaatan data analitik menjadi kebutuhan yang tidak terelakkan bagi bisnis *e-commerce* agar dapat melakukan intervensi proaktif sebelum pelanggan benar-benar meninggalkan *platform*.

Dalam konteks *e-commerce*, berbagai atribut perilaku pelanggan dapat dimanfaatkan untuk menganalisis kemungkinan terjadinya *churn*, seperti frekuensi pembelian (*frequency*),

waktu pembelian terakhir (*recency*), jumlah pembatalan pesanan, serta karakteristik demografis pelanggan. Atribut-atribut tersebut digunakan dalam beberapa penelitian analisis data pelanggan untuk mengidentifikasi pola pelanggan yang berpotensi melakukan *churn* [3,4,5]. Penelitian ini menggunakan *dataset E-Commerce Churn Dataset 2025* yang diperoleh dari *platform Kaggle*, dengan 2000 observasi dan 17 variabel yang mencakup aspek demografis serta perilaku transaksi pelanggan. *Dataset* ini bersifat sintesis dan dirancang untuk menggambarkan karakteristik umum perilaku pelanggan dalam ekosistem *e-commerce*, sehingga dapat digunakan sebagai data eksperimen dalam pengembangan dan evaluasi model prediksi *churn*. Jumlah observasi dan keberagaman variabel yang digunakan dinilai memadai untuk mendukung analisis prediksi *churn* pelanggan secara objektif. Meskipun demikian, berbagai pendekatan *machine learning* yang telah diterapkan dalam prediksi *churn* pelanggan sebagian besar masih menghadapi permasalahan ketidakseimbangan data (*imbalanced data*), yaitu kondisi ketika jumlah data pada satu kelas lebih dominan dibandingkan kelas lainnya. Dalam konteks prediksi *churn*, pelanggan yang tidak melakukan *churn* umumnya memiliki jumlah data lebih banyak sehingga disebut sebagai kelas mayoritas, sedangkan pelanggan yang melakukan *churn* memiliki jumlah data lebih sedikit dan dikategorikan sebagai kelas minoritas. Kondisi tersebut dapat menyebabkan model cenderung menghasilkan bias prediksi terhadap kelas mayoritas dan kurang sensitif dalam mendeteksi pelanggan *churn* [6,7,8]. Selain itu, sebagian penelitian masih bergantung pada metrik evaluasi yang kurang representatif. Kondisi ini menyebabkan model dengan akurasi tinggi belum tentu efektif dalam mendeteksi pelanggan *churn* secara tepat, sehingga menimbulkan kesenjangan antara performa statistik dan nilai strategis bagi bisnis [9,10].

Untuk menangani permasalahan tersebut, beberapa penelitian mengaplikasikan pendekatan *ensemble learning*, yaitu metode yang menggabungkan beberapa algoritma pembelajaran yang lebih stabil dalam menghadapi variasi data dan distribusi kelas yang tidak seimbang serta mampu meningkatkan sensitivitas terhadap kelas minoritas ketika dikombinasikan dengan teknik penanganan *class imbalance* seperti *Synthetic Minority Over-sampling Technique (SMOTE)*. Teknik *SMOTE* diterapkan setelah proses pembagian data dan hanya pada data *training* untuk menghindari terjadinya data *leakage* serta menjaga validitas evaluasi model terhadap data yang belum pernah dilihat sebelumnya [11]. Beberapa metode *ensemble learning* yang banyak diterapkan dalam prediksi *churn* meliputi *Random Forest*, *Gradient Boosting Machine (GBM)*, dan *Extreme Gradient Boosting (XGBoost)* karena memiliki kemampuan yang baik dalam menangani kompleksitas pola data dan meningkatkan performa klasifikasi. Implementasi model *XGBoost* pada layanan

berlangganan *B2B* perbankan mencapai *recall* 0,9881, menunjukkan efektivitas tinggi dalam mengidentifikasi nasabah berisiko *churn* [12]. Selain itu, pendekatan *voting classifier* berbasis *ensemble* yang dikombinasikan dengan *SMOTE*, menghasilkan *recall* sebesar 0,90, yang menunjukkan bahwa teknik tersebut mampu menghasilkan sistem prediksi *churn* yang lebih seimbang [13]. Dalam penelitian lain, model *Gradient Boosting* mampu mengidentifikasi kasus *lapse* (ketidakaktifan polis) pada layanan asuransi dengan tingkat *recall* sebesar 76,89% berdasarkan *confusion matrix*, yang menunjukkan kemampuan model dalam mendeteksi *lapse* dengan baik [11]. Khusus pada sektor *e-commerce*, kombinasi *Random Forest* dan *SMOTE* meningkatkan *recall* hingga 87,1% serta mencapai *AUC* sebesar 0,93, yang menunjukkan kemampuan model dalam mengenali pelanggan berpotensi *churn* dan membedakan secara efektif antara pelanggan *churn* dan non-*churn* [14]. Hasil-hasil tersebut menunjukkan bahwa kombinasi *ensemble learning* dan teknik *SMOTE* memberikan peningkatan performa yang lebih andal dalam permasalahan prediksi *churn*. Meskipun mayoritas penelitian tersebut dilakukan di luar konteks *e-commerce*, karakteristik permasalahan *churn* yang dikaji memiliki kesamaan mendasar, terutama pada aspek ketidakseimbangan kelas dan kompleksitas perilaku pelanggan. Selain itu, hasil performa antar model *ensemble learning* masih menunjukkan variasi, sehingga diperlukan analisis lebih lanjut untuk memahami efektivitas masing-masing model dalam mendeteksi pelanggan *churn* pada kondisi data yang tidak seimbang.

Dengan demikian, penerapan metode *ensemble learning* yang dikombinasikan dengan teknik penanganan *class imbalance* dinilai relevan untuk diteliti lebih lanjut dalam konteks pelanggan *e-commerce*. Penelitian ini bertujuan untuk menganalisis kinerja beberapa model *ensemble learning* dalam memprediksi *churn* pelanggan *e-commerce* dengan mempertimbangkan karakteristik data yang tidak seimbang. Selain itu, penelitian ini juga mengkaji pengaruh penerapan teknik penanganan *class imbalance* terhadap kemampuan model dalam mengidentifikasi pelanggan *churn* secara lebih akurat dan representatif. Dengan pendekatan tersebut, penelitian ini diharapkan mampu memberikan pemahaman yang lebih komprehensif mengenai efektivitas *ensemble learning* dalam prediksi *churn e-commerce*, serta menjadi referensi dalam pengembangan strategi retensi pelanggan pada industri *e-commerce*, sehingga diangkat dengan judul: **“Prediksi Churn Pelanggan E-Commerce Menggunakan Ensemble Learning”**.

1.2 Rumusan Masalah

Berdasarkan uraian latar belakang di atas, maka masalah yang ingin diselesaikan oleh penulis yaitu:

1. Bagaimana membangun model prediksi *churn* pelanggan *e-commerce* menggunakan metode *ensemble learning*?
2. Bagaimana aktivitas penanganan *class imbalance* terhadap peningkatan performa model prediksi *churn*?
3. Sejauh mana kinerja model *ensemble learning*, yaitu *Random Forest*, *Gradient Boosting*, dan *XGBoost* dalam memprediksi *churn* pelanggan *e-commerce* berdasarkan *Confusion Matrix*, *Recall*, dan *Area Under the Curve (AUC)*?

1.3 Tujuan

Untuk menjawab permasalahan yang telah dirumuskan, adapun tujuan dari penelitian ini adalah sebagai berikut:

1. Membangun model prediksi *churn* pelanggan *e-commerce* menggunakan metode *ensemble learning*.
2. Menganalisis penerapan teknik penanganan *class imbalance* terhadap peningkatan kemampuan model dalam mendeteksi kelas minoritas (pelanggan *churn*).
3. Membandingkan kinerja beberapa model *ensemble learning*, yaitu *Random Forest*, *Gradient Boosting*, dan *XGBoost* menggunakan *Confusion Matrix*, *Recall*, dan *AUC* sebagai metrik evaluasi.

1.4 Manfaat

Manfaat yang dapat diperoleh dari penelitian ini antara lain:

1. Memberikan kontribusi teoretis dalam memperkaya kajian akademik di bidang data analitik dan *machine learning*, khususnya terkait penerapan *ensemble learning* dan penanganan *class imbalance* pada permasalahan *churn prediction*.
2. Membantu perusahaan *e-commerce* dalam mengidentifikasi pelanggan berisiko *churn* dengan lebih akurat, sehingga strategi retensi pelanggan dapat dirancang secara lebih efektif dan berbasis data.
3. Dapat digunakan sebagai referensi bagi penulis lain yang ingin mengembangkan topik serupa dalam penelitian lanjutan.

1.5 Ruang Lingkup

Agar tujuan dan manfaat dapat tercapai, berikut adalah batasan-batasan yang ditetapkan dalam penelitian ini:

- a. Penelitian ini menggunakan data sekunder, yaitu *dataset churn* pelanggan *e-commerce* “*E-Commerce Churn Dataset 2025*”, berisi 17 variabel dan 2000 baris yang diperoleh dari *Kaggle* dan dapat diakses pada link <https://www.kaggle.com/datasets/nabihazahid/e-commerce-customer-insights-and-churn-dataset>.
- b. Pendekatan yang digunakan dalam penelitian ini adalah metode klasifikasi berbasis *ensemble learning*, yang meliputi *Random Forest*, *Gradient Boosting*, dan *XGBoost*.
- c. Untuk mengatasi permasalahan ketidakseimbangan distribusi kelas, penelitian ini menerapkan teknik *Synthetic Minority Over-sampling Technique (SMOTE)* pada tahap pemodelan untuk meningkatkan performa model dalam mendeteksi pelanggan *churn*.
- d. Proses analisis data dilakukan menggunakan *framework OSEMN (Obtain, Scrub, Explore, Model, iNterpret)*.
- e. Evaluasi kinerja model dilakukan menggunakan *Confusion Matrix*, *Recall*, dan *AUC* guna menilai kemampuan model dalam mengidentifikasi pelanggan *churn* secara representatif.
- f. Implementasi dan pengolahan data dilakukan menggunakan *software RStudio* dengan bahasa pemrograman *R* dan didukung oleh *library* utama seperti *randomForest*, *gbm*, *xgboost*, dan *caret*.