BAB 1
PENDAHULUAN

1.1. Latar Belakang

Di era digital yang terus mengalami perkembangan pesat, data telah menjadi aset strategis bagi perusahaan, termasuk dalam industri ritel. Kemajuan Teknologi Informasi dan Komunikasi (TIK) memungkinkan organisasi untuk menghimpun, memproses, serta menganalisis data bisnis dalam skala besar yang dikenal sebagai Big Data. Data tersebut diperoleh dari berbagai sumber, antara lain sistem Point of Sale (POS), program kartu loyalitas, platform belanja daring, serta survei pelanggan. Di antara berbagai jenis data tersebut, data perilaku pembelian pelanggan memiliki nilai yang sangat penting karena menggambarkan pola konsumsi pelanggan, seperti frekuensi transaksi, kategori produk yang dibeli, waktu pembelian, total pengeluaran, serta preferensi terhadap merek tertentu [3], [5], [9]. Pemanfaatan analisis data perilaku pelanggan secara optimal dapat membantu perusahaan dalam memahami kebutuhan konsumen, merancang strategi pemasaran yang lebih tepat sasaran, serta meningkatkan daya saing dan kinerja pendapatan [5].

Meskipun demikian, pengolahan data perilaku pembelian pelanggan tidak terlepas dari sejumlah permasalahan, salah satunya adalah keberadaan data outlier. Data outlier merupakan observasi yang menyimpang secara signifikan dari pola mayoritas data dan berpotensi memengaruhi akurasi hasil analisis serta kualitas pengambilan keputusan bisnis [1]. Di sisi lain, kemunculan outlier juga dapat mencerminkan kondisi tertentu, seperti indikasi kecurangan transaksi, perubahan perilaku pelanggan yang tidak lazim, maupun kesalahan dalam proses pencatatan data [2], [9]. Oleh sebab itu, deteksi outlier menjadi tahapan penting dalam proses analisis data ritel agar informasi yang dihasilkan tetap valid dan dapat dipercaya.

Proses identifikasi outlier menjadi semakin menantang karena data perilaku pembelian pelanggan umumnya memiliki karakteristik non-linear, berdimensi tinggi, serta tidak selalu mengikuti distribusi normal [1], [7]. Metode deteksi outlier konvensional, seperti z-score, Interquartile Range (IQR), dan regresi linear, umumnya bergantung pada asumsi statistik tertentu serta pola data yang relatif

sederhana. Akibatnya, metode tersebut kurang efektif ketika diterapkan pada data berskala besar dengan struktur yang kompleks [1], [6]. Kondisi ini dapat menyebabkan outlier yang penting tidak teridentifikasi atau justru mengklasifikasikan data normal sebagai anomali, sehingga menurunkan kualitas hasil analisis secara keseluruhan [1].

Sejalan dengan perkembangan machine learning, pendekatan deteksi outlier berbasis unsupervised learning semakin banyak diadopsi karena kemampuannya dalam mengenali pola kompleks tanpa memerlukan data berlabel maupun asumsi distribusi tertentu [7], [10]. Pendekatan ini dinilai lebih adaptif dan fleksibel untuk diterapkan pada data ritel yang bersifat dinamis dan heterogen. Dalam penelitian ini, dilakukan analisis perbandingan terhadap dua algoritma deteksi outlier, yaitu Local Outlier Factor (LOF) dan Elliptic Envelope (EE). Algoritma LOF merupakan metode berbasis kepadatan lokal (density-based) yang mengevaluasi tingkat keanehan suatu data dengan membandingkan kepadatan lokalnya terhadap data di sekitarnya, sehingga efektif dalam mengidentifikasi anomali pada klaster dengan kepadatan yang bervariasi [6], [10].

Di sisi lain, algoritma Elliptic Envelope menerapkan pendekatan statistik berbasis robust covariance untuk membentuk batas elips multidimensi yang merepresentasikan distribusi data normal, di mana data yang berada di luar batas tersebut diklasifikasikan sebagai outlier [8]. Perbandingan antara LOF dan Elliptic Envelope dilakukan karena kedua algoritma tersebut merepresentasikan pendekatan yang berbeda, yaitu metode non-parametrik berbasis kepadatan lokal dan metode statistik yang mengandalkan asumsi distribusi data [6], [8].

Metode deteksi outlier lain, seperti Isolation Forest, One-Class SVM, dan Autoencoder, tidak menjadi fokus dalam penelitian ini karena masing-masing memiliki keterbatasan, antara lain sensitivitas terhadap pemilihan parameter, kurangnya perhatian terhadap struktur lokal data, serta kebutuhan komputasi yang relatif tinggi [4], [10]. Dengan demikian, penelitian ini difokuskan pada perbandingan algoritma LOF dan Elliptic Envelope sebagai pendekatan awal yang lebih efisien, interpretatif, dan sesuai untuk mendeteksi outlier pada data perilaku pembelian pelanggan di sektor ritel.

1.2. Rumusan Masalah

Rumusan masalah pada penelitian ini yaitu:

1. Bagaimana perbedaan hasil deteksi *outlier* antara algoritma LOF dan EE pada data perilaku pembelian pelanggan, ditinjau dari jumlah *outlier* yang teridentifikasi, tingkat kesesuaian *(overlap)* hasil deteksi, serta pola pemisahan data normal dan *outlier*?
2. Bagaimana perbedaan karakteristik perilaku pembelian pelanggan yang teridentifikasi sebagai outlier oleh algoritma LOF dan EE berdasarkan fitur-fitur pembelian yang digunakan, sehingga dapat mendukung pemahaman pola anomali pada tahap analisis awal data?

1.3. Batasan Masalah

Batasan masalah pada penelitian ini yaitu:

1. Dataset yang digunakan diambil dari Kaggle yaitu *Consumer Buying Analysis* sebanyak 2.240 data.
2. Visualisasi data outlier menggunakan metode PCA sehingga pola distribusi dan titik penyimpangan dapat diamati secara lebih jelas.
3. *Output* penelitian ini berupa perbandingan jumlah data outlier yang terdeteksi oleh algoritma *Local Outlier Factor* (LOF), *Elliptic Envelope* (EE), serta data yang secara bersamaan teridentifikasi sebagai *outlier* oleh kedua metode *(overlap)*. Selain itu, penelitian menampilkan visualisasi hasil deteksi outlier menggunakan *Principal Component Analysis* (PCA) serta analisis karakteristik *outlier* berdasarkan nilai rata-rata fitur utama untuk menggambarkan perbedaan pola deteksi antara metode LOF dan EE.

1.4. Tujuan Penelitian

Tujuan yang diharapkan dari penelitian ini yaitu:

1. Untuk menganalisis dan membandingkan hasil deteksi *outlier* yang dihasilkan oleh algoritma LOF dan EE pada data perilaku pembelian pelanggan, ditinjau dari jumlah outlier yang teridentifikasi, tingkat kesesuaian *(overlap)* hasil deteksi, serta pola pemisahan antara data normal dan outlier.

2. Untuk mengidentifikasi dan membandingkan karakteristik perilaku pembelian pelanggan yang terdeteksi sebagai *outlier* oleh algoritma LOF dan EE berdasarkan fitur-fitur pembelian yang digunakan, guna mendukung pemahaman pola anomali sebagai dasar analisis awal data dan pengambilan keputusan

1.5. Manfaat Penelitian

Manfaat yang diharapkan dari penelitian ini yaitu:

1. Memberikan kontribusi ilmiah dalam bidang data *mining* dan *machine learning*, khususnya pada deteksi *outlier* secara *unsupervised* pada data perilaku pembelian pelanggan yang kompleks dan berdimensi tinggi.
2. Memberikan gambaran komparatif bagi perusahaan ritel mengenai karakteristik dan efisiensi algoritma LOF dan EE sebagai pendukung analisis awal dan pengambilan keputusan berbasis data.
3. Menjadi referensi bagi penelitian selanjutnya dalam pemilihan dan pengembangan metode deteksi *outlier* yang sesuai dengan karakteristik data ritel dan perilaku pelanggan.

1.6. Keterbaruan Penelitian

Berikut ini akan diuraikan keterbaruan dari penelitian ini melalui *literature review* penelitian terdahulu yang berkaitan dengan topik penelitian antara lain:

1. Hodge dan Austin (2018) mengevaluasi berbagai algoritma untuk klasifikasi dan deteksi *outlier* pada data temporal. Hasil penelitian menunjukkan bahwa tidak terdapat satu algoritma yang unggul secara mutlak, namun *Random Forest* dan *Gradient Boosting Machines* termasuk metode dengan performa yang baik dalam deteksi *outlier* [11].
2. Alfian, G., et al., (2023) memanfaatkan teknologi RFID dan algoritma machine learning seperti *Isolation Forest*, ADASYN, dan *Multilayer Perceptron* untuk menganalisis kebiasaan belanja pelanggan. Penelitian ini menghasilkan akurasi tinggi hingga 97,78% dan diimplementasikan dalam aplikasi *web* untuk mendukung pemahaman preferensi konsumen serta strategi pemasaran [12].

3. Deniz dan Bülbül (2024) membandingkan beberapa algoritma prediktif, termasuk *Random Forest*, SVM, *Logistic Regression*, KNN, dan *XGBoost*, dan menyimpulkan bahwa metode ensemble, khususnya Random Forest dan *XGBoost*, memiliki performa terbaik dalam memprediksi perilaku pembelian pelanggan [13].
4. Dalbouh, et al., (2025) menganalisis pengaruh faktor demografis dan keterlibatan media sosial terhadap pengambilan keputusan konsumen menggunakan Random Forest dan metode statistik, dengan hasil akurasi mencapai 88%, presisi 0,90, dan recall 0,95 [14].

Keterbaruan penelitian ini terletak pada komparasi algoritma LOF dan EE dalam mendeteksi *outlier* pada data perilaku konsumen. Berbeda dari penelitian sebelumnya yang berfokus pada pemodelan prediktif dengan evaluasi *supervised*, studi ini menitikberatkan pada tahap pra-pemodelan menggunakan pendekatan *unsupervised*. Analisis dilakukan melalui perbandingan karakteristik hasil deteksi, jumlah dan tingkat kesesuaian *outlier*, serta pola pemisahan data normal dan anomali, sehingga memberikan kontribusi orisinal dalam memahami perbedaan paradigma deteksi berbasis kepadatan lokal dan statistik distribusi.