

## **BAB I**

### **PENDAHULUAN**

#### **1.1 Latar Belakang**

Perkembangan teknologi informasi mendorong perusahaan ritel untuk memahami perilaku pelanggan sebagai dasar dalam menyusun strategi pemasaran yang efektif di tengah persaingan bisnis yang semakin ketat [1]. Dalam kondisi persaingan bisnis ritel yang kompetitif, pelanggan menjadi faktor penting dalam menentukan keberhasilan perusahaan. Pelanggan merupakan aset utama perusahaan sehingga pemahaman terhadap kebutuhan, preferensi dan perilaku pelanggan menjadi fondasi dalam penyusunan strategi bisnis yang adaptif dan efektif. Perusahaan yang tidak mampu memahami perilaku konsumennya berisiko menerapkan strategi pemasaran yang kurang relevan dan kehilangan daya saing [2].

Salah satu pendekatan yang digunakan untuk memahami perilaku pelanggan adalah segmentasi pelanggan. Melalui segmentasi pelanggan, perusahaan dapat mengetahui karakteristik masing-masing kelompok pelanggan sehingga strategi pemasaran dapat dilakukan secara lebih terarah dan personal [3]. Segmentasi pelanggan membantu perusahaan menghindari pendekatan pemasaran yang bersifat umum dan memungkinkan perusahaan menerapkan strategi yang lebih efektif sesuai karakteristik pelanggan [2]. Model RFM telah lama digunakan untuk segmentasi pelanggan berdasarkan *recency*, *frequency* dan *monetary* [3]. Faktor *recency* menilai seberapa baru pelanggan melakukan transaksi, *frequency* menilai seberapa sering pelanggan bertransaksi, sedangkan *monetary* menilai seberapa besar total pengeluaran yang dilakukan oleh pelanggan saat bertransaksi [3], [4]. Pendekatan RFM dinilai lebih unggul dibandingkan metode segmentasi tradisional karena lebih berfokus pada perilaku pelanggan secara nyata serta memberikan gambaran yang lebih dinamis terkait nilai pelanggan bagi perusahaan. Pendekatan RFM memungkinkan segmentasi yang lebih tanggap terhadap perubahan perilaku pelanggan dan lebih relevan untuk strategi pemasaran berbasis data [5].

Salah satu pendekatan yang banyak digunakan untuk segmentasi pelanggan adalah analisis RFM yang dikombinasikan dengan algoritma K-Means. Kombinasi antara RFM dengan algoritma K-Means efektif dalam mengelompokkan pelanggan berdasarkan karakteristik pembelian [1]. Hasil kombinasi antara RFM dengan metode klasterisasi, algoritma K-Means lebih unggul dibanding algoritma K-Medoids dan Fuzzy C Means (FCM), membuktikan bahwa kombinasi tersebut efektif dalam mengelompokkan pelanggan ke dalam segmen yang relevan [4]. Keunggulan algoritma K-Means adalah kesederhanaan, efisiensi dan kemampuannya dalam membentuk klaster yang jelas [2]. Akan tetapi, algoritma K-Means memiliki kelemahan, yaitu memerlukan penentuan jumlah awal klaster dan inisialisasi *centroid* secara acak, sehingga menyebabkan penurunan akurasi dan performa serta membuat proses evaluasi cenderung lebih lama saat digunakan pada *dataset* yang besar [6]. keterbatasan berikutnya yaitu sensitivitas algoritma K-means terhadap *outlier* dalam *dataset*. *Outlier* yang secara signifikan menyimpang dari mayoritas, dapat mendistorsi perhitungan *centroid* dan merusak akurasi pengelompokan [7]. Berikutnya, algoritma K-Means memiliki masalah sensitivitas terhadap pemilihan atribut dan skala data yang digunakan sehingga dapat memengaruhi kualitas klasterisasi [8]. Pada data RFM, kondisi tersebut dapat berkaitan dengan kompleksitas data dan korelasi antar fitur.

Untuk mengatasi keterbatasan dalam penentuan jumlah klaster, diatasi menggunakan metrik evaluasi Silhouette Score, Davies-Bouldin Index dan Calinski-Harabasz Index. Ketiga metrik tersebut dipilih karena mampu memberikan perspektif yang berbeda dalam menilai kualitas klaster yang terbentuk. Metriks Silhouette Score memberikan ukuran seberapa baik setiap titik data ditempatkan dalam klaster mereka sendiri dibandingkan dengan klaster lain, sementara Davies-Bouldin Index mengevaluasi seberapa jauh klaster tersebut berada dari yang lain [9], metriks Calinski-Harabasz Index memaksimalkan dispersi antar klaster sekaligus meminimalkannya di dalam klaster [10]. Untuk mengatasi keterbatasan dalam inisialisasi *centroid*, digunakan metode klasterisasi algoritma K-Means++ yang menghasilkan *centroid* awal yang terpisah satu sama lain dan mampu meningkatkan konvergensi dan kualitas klasterisasi [11]. metode

initialization seperti K-Means++ terbukti meningkatkan stabilitas kluster dan kualitas klusterisasi dibanding random initialization, tanpa mengubah mekanisme dasar algoritma K-Means [12], [13].

Untuk penanganan data *outlier*, penelitian sebelumnya menggunakan beberapa metode untuk deteksi outlier seperti Z-Score [14] dan Interquartile Range (IQR) [15] untuk mengidentifikasi data outlier serta Local Outlier Factor (LOF) [16] digunakan dalam mendeteksi anomali data. Pada penelitian ini, metode DBSCAN digunakan karena mampu mendeteksi noise dan outlier berdasarkan kepadatan data serta memiliki fleksibilitas dalam menangani distribusi data yang kompleks dan bervariasi [17]. Sedangkan untuk penanganan sensitivitas terhadap pemilihan atribut dan skala data, terdapat beberapa metode seleksi atribut yang dapat digunakan untuk meningkatkan performa algoritma K-Means antara lain PCA, Random Forest Feature Importance, Chi-Square Test dan XGBoost [8]. Metode-metode tersebut memiliki pendekatan yang berbeda dalam pemilihan fitur, baik melalui reduksi dimensi maupun pemberian skor kontribusi atribut. Dalam penelitian ini, PCA dipilih sebagai metode reduksi dimensi karena sesuai digunakan pada data numerik seperti RFM yang memiliki potensi korelasi antar variabel. PCA digunakan untuk menyederhanakan jumlah fitur tanpa menghilangkan informasi penting secara signifikan. PCA mentransformasikan data ke dalam beberapa komponen utama yang merupakan kombinasi linier dari fitur-fitur asli [18].

Penelitian ini mengangkat judul “Optimasi Klustering K-Means untuk Segmentasi Pelanggan dengan Integrasi PCA, DBSCAN dan K-Means++” dengan harapan dapat menghasilkan segmentasi pelanggan yang lebih optimal dan akurat dibandingkan metode klusterisasi konvensional.

## 1.2 Rumusan Masalah

Berdasarkan latar belakang yang telah diuraikan, permasalahan dalam penelitian ini yaitu algoritma K-Means memiliki keterbatasan dalam penentuan jumlah kluster ( $K$ ), inisialisasi *centroid* secara acak serta sensitivitas terhadap *outlier*, skala data dan kompleksitas data yang ditandai dengan korelasi antar fitur. Kondisi tersebut dapat memengaruhi kualitas hasil klusterisasi pada data RFM sehingga hasil segmentasi pelanggan menjadi kurang optimal. Diperlukan pendekatan untuk meningkatkan kualitas klusterisasi pada data RFM melalui proses reduksi dimensi, penanganan *outlier* serta optimasi inisialisasi *centroid* menggunakan PCA, DBSCAN dan K-Means++.

## 1.3 Batasan Masalah

Supaya penelitian lebih terarah dan tidak meluas, maka dalam penelitian ini diberikan batasan masalah sebagai berikut:

- Penelitian ini menggunakan *dataset* yang berasal dari UCL *Machine learning Repository* yaitu Online Retail.
- Penelitian ini menggunakan pendekatan RFM (recency, frequency, monetary).
- Metode klusterisasi yang digunakan adalah algoritma K-Means, K-Means++ dan hybrid yang menggabungkan PCA, DBSCAN dan K-Means++.
- Evaluasi menggunakan Silhouette Score, Davies-Bouldin Index dan Calinski-Harabasz Index.

## 1.4 Tujuan Penelitian

Berdasarkan rumusan masalah yang telah ditetapkan, tujuan dari penelitian ini adalah sebagai berikut:

- Meningkatkan kualitas klusterisasi pada data RFM dengan mengatasi keterbatasan algoritma K-Means dalam inisialisasi *centroid*, sensitivitas terhadap *outlier*, serta kompleksitas data yang ditandai dengan korelasi antar fitur.

- b. Menerapkan pendekatan hybrid yang mengintegrasikan PCA untuk reduksi dimensi, DBSCAN untuk penanganan *outlier*, dan K-Means++ untuk optimasi inisialisasi centroid pada segmentasi pelanggan berbasis RFM.
- c. Mengevaluasi kualitas klusterisasi yang dihasilkan menggunakan metrik Silhouette Score, Davies-Bouldin Index dan Calinski-Harabasz Index.

### **1.5 Manfaat Penelitian**

Penelitian ini diharapkan dapat memberikan manfaat yang dikelompokkan ke dalam dua aspek, yaitu manfaat teoritis dan manfaat praktis sebagai berikut:

#### **1.5.1 Manfaat Teoritis**

Memberikan kontribusi dalam pengembangan metode klusterisasi melalui integrasi beberapa algoritma.

#### **1.5.2 Manfaat Praktis**

Membantu perusahaan ritel dalam memahami perilaku pelanggan untuk mendukung strategi pemasaran dan manajemen pelanggan.

### **1.6 Sistematika Penulisan**

Sistematika penulisan dalam penelitian ini disusun untuk memberikan gambaran yang terstruktur mengenai isi setiap bab. Adapun sistematika penulisan penelitian ini adalah sebagai berikut:

## **BAB I PENDAHULUAN**

Bab ini berisi latar belakang masalah, rumusan masalah, tujuan penelitian, manfaat penelitian, batasan masalah serta sistematika penulisan.

## **BAB II TINJAUAN PUSTAKA**

Bab ini berisi studi literatur terhadap penelitian terdahulu terkait dengan optimasi klusterisasi K-Means dan dasar teori mengenai konsep segmentasi pelanggan, metode klusterisasi, algoritma K-Means, K-Means++, PCA, DBSCAN, metrik evaluasi klusterisasi, serta dasar teori yang relevan lainnya.

### **BAB III METODE PENELITIAN**

Bab ini menjelaskan objek penelitian serta alur penelitian dengan pendekatan CRISP-DM, dengan tahapan *business understanding*, *data understanding*, *data preparation*, *modeling* dan *evaluation*.

### **BAB IV HASIL DAN PEMBAHASAN**

Bab ini berisi implementasi dari metodologi CRISP-DM yang mencakup pemahaman bisnis.

### **BAB V PENUTUP**

Bab ini berisi tentang kesimpulan hasil penelitian yang sesuai dengan rumusan masalah dan saran yang dapat diberikan untuk melakukan pengembangan penelitian selanjutnya.

