

BAB II

TINJAUAN PUSTAKA

Untuk menambah referensi dalam penelitian ini, penulis mencari berbagai penelitian terdahulu yang memiliki keterkaitan dengan topik yang penulis ambil. Studi-studi sebelumnya akan dijadikan sebagai dasar acuan serta bahan perbandingan dalam pelaksanaan dan penyelesaian penelitian ini.

Zulrahmadi melakukan penelitian yang bertujuan untuk mengetahui informasi terkait wilayah yang potensial sebagai penjualan sepeda motor Honda di Kabupaten Indragiri Hilir sehingga dapat digunakan oleh PT [12]. Capella Dinamik Nusantara untuk pengambilan keputusan. Penelitian tersebut dilakukan dengan clustering data menggunakan metode K-Means. Data dikumpulkan dari penjualan motor Honda pada PT. Capella Dinamik Nusantara dari tahun 2017 sampai dengan 2019 dengan menjadi 3 cluster, yaitu cluster C0 untuk wilayah kurang berpotensi, cluster C1 untuk wilayah cukup berpotensi, dan cluster C2 untuk wilayah sangat berpotensi. Hasil penelitian menghasilkan tiga cluster yaitu Cluster 0 (C0) yang dideskripsikan sebagai kurang berpotensi sebanyak lima wilayah desa, Cluster 1 (C1) yang dideskripsikan cukup berpotensi sebanyak tiga wilayah desa, Cluster 2 (C2) yang dideskripsikan sangat berpotensi sebanyak dua wilayah desa untuk penjualan motor. Iterasi terjadi sebanyak dua kali. Hasil cluster tersebut diharapkan bisa digunakan untuk mengambil keputusan terbaik sehingga keuntungan dari PT. Capella Dinamik Nusantara dapat meningkat dan dapat melakukan promosi yang tepat sasaran.

Parlina dkk melakukan penelitian dengan memanfaatkan algoritma K-Means untuk menentukan pegawai yang layak mengikuti *assessment center* [13]. Penelitian ini bertujuan untuk mengelompokkan data pegawai yang terdaftar dalam Program SDP (Staff Development Program) di PT. Bank Syariah Mandiri tahun 2017 menggunakan algoritma K-Means. Tujuan akhir dari pengelompokan ini adalah untuk mengetahui pegawai yang layak untuk lolos dalam program SPD sehingga dapat melakukan registrasi Assessment Center berdasarkan tiga kriteria utama, yaitu Kedisiplinan Pegawai, Target Kerja Pegawai, dan Kepatuhan terhadap Program SDP. Hasil dari penelitian ini menunjukkan bahwa tiga cluster berhasil terbentuk, masing-masing menunjukkan kelompok pegawai berdasarkan tingkat kecocokan mereka terhadap kriteria Program SDP. Pegawai dalam Cluster 1 dianggap Lolos dan siap mengikuti Assessment Center. Pegawai dalam Cluster 2 dikategorikan sebagai Hampir Lolos, tetapi memerlukan perbaikan administratif seperti kedisiplinan dari Juni hingga Oktober agar dapat mengikuti Assessment Center lanjutan. Pegawai dalam Cluster 3 dikategorikan sebagai Tidak Lolos, dan perlu memperbaiki kedisiplinan selama satu tahun penuh agar dapat dipertimbangkan untuk mengikuti program Assessment Center ini. Iterasi terjadi sebanyak tiga kali.

Selanjutnya Elin dkk melakukan penelitian dengan tujuan untuk memberikan kemudahan bagi PT XYZ, sebuah perusahaan distribusi minuman ringan di Lampung, dalam mengelola data transaksi penjualannya untuk mengetahui tingkat penjualan produk pada setiap cabangnya [7]. Penelitian ini juga bertujuan untuk membantu perusahaan meningkatkan akurasi peramalan penjualan serta merencanakan strategi bisnis yang sesuai, yang pada akhirnya akan memudahkan analisis dan pengambilan keputusan. Dengan demikian, perusahaan dapat mengarahkan upaya pemasaran dan sumber daya dengan lebih efektif untuk

dapat meningkatkan keuntungan dan pertumbuhan bisnis. Metode yang digunakan dalam penelitian ini adalah teknik data mining dengan algoritma K-means clustering menggunakan bahasa pemrograman Python. Data penjualan akan dikelompokkan ke dalam 3 cluster yaitu penjualan tertinggi, penjualan sedang, dan penjualan terendah. Penelitian ini menganalisis data penjualan minuman dari PT XYZ selama periode tahun 2022 yang terdiri dari 14 atribut dengan total 712 record yang tersebar di 12 wilayah Sumatera bagian selatan. Proses klasterisasi dilakukan dengan tahapan penentuan jumlah klaster, inisialisasi nilai centroid, perhitungan jarak terdekat menggunakan Euclidean Distance, pengelompokkan objek berdasarkan jarak terdekat dengan centroid, dan iterasi hingga nilai optimal tercapai. Untuk memvalidasi hasil clustering, penelitian ini menggunakan metode Silhouette Score untuk mengetahui jumlah klaster yang optimal. Hasil penelitian menunjukkan bahwa berdasarkan klasterisasi data penjualan berdasarkan region, penjualan dengan nilai tertinggi berada pada region 3 yang ditandai dengan warna hitam (tinggi), penjualan dengan nilai sedang berada pada region 11 yang ditandai dengan warna merah (sedang), dan penjualan dengan nilai terendah berada pada region 4 yang ditandai dengan warna hijau (rendah). Sementara itu, hasil klasterisasi untuk data penjualan berdasarkan outlet menunjukkan penjualan dengan klaster tinggi ditandai dengan warna hijau, klaster sedang ditandai dengan warna hitam, dan klaster rendah ditandai dengan warna merah. Pengujian menggunakan metode Silhouette Score menghasilkan skor 0,78 untuk klaster data penjualan berdasarkan region dan 0,58 untuk klaster data penjualan berdasarkan outlet. Kedua skor tersebut menunjukkan bahwa jumlah klaster yang dihasilkan masuk ke dalam kategori baik karena tidak mendekati -1, sehingga hasil klasterisasi dapat dijadikan rekomendasi bagi PT XYZ dalam menentukan strategi penjualan untuk meningkatkan keuntungan perusahaan.

Selanjutnya Eriskiannisa dkk melakukan penelitian dengan tujuan utama dari penelitian ini adalah menerapkan segmentasi pelanggan pada data transaksi *online retail* menggunakan algoritma K-Means clustering untuk mengidentifikasi karakteristik pelanggan dan menentukan strategi pemasaran yang lebih tepat sasaran [32]. Penelitian ini juga berfokus pada penentuan jumlah klaster optimal dengan metode Elbow, sehingga perusahaan dapat lebih memahami perilaku pelanggan dan meningkatkan efektivitas strategi pemasarannya. Metode penelitian yang digunakan meliputi studi literatur terkait segmentasi pelanggan dan K-Means pada rentang tahun 2018–2022, pengumpulan data transaksi *online retail* sebanyak 54.910 baris data dengan 8 kolom atribut (Invoice No, Stock Code, Description, Quantity, Invoice Date, Unit Price, Customer ID, Country), *exploratory data analysis* dan *data pre-processing* (penanganan *missing value*, *outlier*, transformasi data, penambahan kolom penjualan, serta normalisasi dan standarisasi data), implementasi K-Means dengan bahasa pemrograman Python melalui Google Colab, serta penentuan jumlah klaster optimal menggunakan metode Elbow. Hasil analisis menunjukkan bahwa jumlah klaster optimal adalah empat, dengan karakteristik sebagai berikut: 1) Klaster pertama terdiri dari pelanggan dengan jumlah pembelian rendah dan harga satuan tinggi, direkomendasikan diberikan diskon atau penawaran eksklusif, 2) Klaster kedua memiliki pembelian menengah dan harga cukup rendah, direkomendasikan diberi promo pembelian dalam jumlah tertentu, 3) Klaster ketiga memiliki pembelian tinggi dan harga rendah, direkomendasikan mendapatkan program hadiah atau undian, 4) Klaster keempat memiliki pembelian rendah dan harga tinggi, direkomendasikan diberi diskon berkala atau penawaran khusus. Temuan ini menegaskan bahwa kuantitas pembelian dan harga satuan berperan penting dalam memengaruhi perilaku pelanggan, serta menunjukkan bahwa K-Means dengan metode Elbow dapat memberikan hasil segmentasi yang efektif untuk mendukung strategi pemasaran.

Kemudian Nisa dkk melakukan penelitian yang bertujuan untuk melakukan analisis karakteristik pelanggan sebagai dasar penetapan segmentasi pelanggan dan *customer profiling* pada produk *digital service* Indihome milik PT. Telekomunikasi Indonesia [33]. Tujuan utama penelitian adalah untuk mengetahui perilaku pelanggan dan menerapkan strategi pemasaran yang tepat sehingga dapat mendatangkan keuntungan optimal bagi perusahaan, mengingat sebelumnya belum ada informasi yang memadai tentang karakteristik pelanggan. Dalam aspek metodologi, penelitian ini menggunakan kerangka kerja CRISP-DM (*Cross Standard Industries for Data Mining*) sebagai panduan sistematis dalam proses *data mining*. Algoritma K-Means dipilih sebagai metode clustering utama karena kemampuannya dalam mempartisi data ke dalam kelompok-kelompok berdasarkan kemiripan karakteristik. Untuk menentukan jumlah cluster yang optimal, peneliti menggunakan metode Elbow yang menganalisis nilai Sum Square. Data yang digunakan dalam penelitian ini terdiri dari 1.392 record yang merupakan 30% dari total 4.640 record data pelanggan yang tersedia. Data tersebut terdiri dari tiga jenis yaitu data pelanggan Indihome, data pelanggan *add-on*, dan data *churn* pelanggan. Atribut yang digunakan untuk proses clustering meliputi lama berlangganan, jumlah layanan yang diambil, dan total tagihan pelanggan. Melalui metode Elbow, ditentukan bahwa jumlah cluster optimal adalah $K = 3$, sehingga seluruh data pelanggan dikelompokkan ke dalam tiga segmen yang berbeda berdasarkan karakteristik tersebut. Hasil clustering pada simulasi terbaik menunjukkan bahwa Cluster 0 terdiri dari 396 pelanggan dengan kategori pelanggan yang memberikan keuntungan terbesar bagi perusahaan, memiliki waktu berlangganan 7-11 bulan, menggunakan 1-4 layanan, dan total tagihan berkisar 209-286 ribu rupiah. Cluster 1 berisi 286 pelanggan yang tanpa disadari memiliki potensi besar dalam memberikan keuntungan bagi perusahaan, dengan waktu berlangganan 7-11 bulan, menggunakan 1 layanan, dan total tagihan 275 ribu hingga 1,144 ribu rupiah. Sementara itu, Cluster 2 merupakan kelompok terkecil dengan 14 pelanggan yang dikategorikan sebagai pelanggan yang memberikan keuntungan lebih sedikit daripada biaya untuk memberikan pelayanan, memiliki waktu berlangganan 1-6 bulan, menggunakan 1-3 layanan, dan total tagihan 286-907,5 ribu rupiah.

Kemudian Dacwanda dan Nataliani melakukan penelitian dengan tujuan untuk menganalisis prestasi akademik siswa dengan mengelompokkan tingkat pencapaian mereka menggunakan metode k-means clustering [37]. Tujuan utama dari pengelompokan ini adalah membantu pihak sekolah memahami distribusi kemampuan siswa sehingga strategi pembelajaran dapat disesuaikan untuk meningkatkan kualitas pendidikan. Sampel penelitian melibatkan 266 siswa kelas IXA hingga IXH di SMP Negeri 2 Salatiga dengan data nilai dari 12 mata pelajaran sesuai kurikulum 2013, yang masing-masing memiliki dua komponen penilaian, yaitu pengetahuan dan keterampilan. Total terdapat 24 atribut yang dianalisis untuk menghasilkan pengelompokan yang lebih akurat dan representatif terhadap kemampuan akademik siswa selama satu semester pembelajaran. Proses analisis menggunakan perangkat lunak RapidMiner untuk mengolah data dan menjalankan algoritma k-means sekaligus menghitung validitas cluster menggunakan indeks Davies-Bouldin, yang menunjukkan bahwa tiga cluster merupakan jumlah yang paling optimal. Hasil analisis mengelompokkan siswa ke dalam tiga kategori, yaitu cluster pintar sebanyak 71 siswa, cluster sedang sebanyak 92 siswa, dan cluster cukup sebanyak 103 siswa. Temuan ini memberikan informasi yang sangat berguna bagi guru dan pihak sekolah dalam merancang metode pembelajaran yang lebih tepat sasaran, termasuk memberikan perhatian khusus bagi siswa yang membutuhkan pendampingan tambahan. Selain itu, hasil penelitian ini juga dapat menjadi dasar bagi pengembangan analisis lanjutan dengan menambahkan aspek non-akademik atau membandingkan metode k-means dengan algoritma clustering lainnya untuk evaluasi yang lebih komprehensif.

Lalu Saputra dan Nataliani melakukan penelitian yang bertujuan untuk mengelompokkan prestasi siswa berdasarkan nilai mata pelajaran dan absensi menggunakan metode k-means clustering [38]. Tujuannya adalah untuk membantu sekolah mengidentifikasi

siswa dengan kategori prestasi rendah, cukup, dan tinggi, sekaligus menentukan siswa top rank yang berpotensi menjadi teladan akademik. Objek penelitian ini adalah siswa kelas X–XII jurusan Bahasa SMAN 1 Tenganan pada tahun ajaran 2014–2017, dengan data yang mencakup rata-rata nilai rapor dari semester satu hingga lima, serta data kehadiran siswa. Sampel yang digunakan berjumlah 25 siswa, yang mewakili data dari angkatan tersebut, sehingga analisis dapat memberikan gambaran umum mengenai distribusi prestasi akademik dan tingkat kehadiran siswa. Proses analisis dilakukan dengan menggunakan perangkat lunak RStudio, yang diintegrasikan dengan algoritma k-means untuk mengelompokkan data ke dalam tiga cluster optimal berdasarkan metode Elbow dan Silhouette. Hasil analisis menunjukkan bahwa terdapat cluster rendah dengan 12 siswa, cluster cukup dengan 5 siswa, dan cluster tinggi dengan 8 siswa. Untuk mengidentifikasi siswa berprestasi terbaik dari cluster tinggi, metode *Simple Additive Weighting* (SAW) digunakan sehingga diperoleh siswa dengan peringkat teratas yang layak mendapatkan penghargaan atau perhatian khusus. Hasil penelitian ini diharapkan dapat digunakan pihak sekolah untuk mengevaluasi kualitas pembelajaran, memberikan motivasi bagi siswa, serta mengembangkan strategi pengajaran yang lebih tepat sasaran guna meningkatkan prestasi akademik di masa mendatang.

Dari berbagai penelitian terdahulu, penelitian yang akan dilakukan oleh penulis berfokus pada analisis data hasil Tantangan Bebras Indonesia tahun 2021 hingga 2024 dengan dua pendekatan utama, yaitu Exploratory Data Analysis (EDA) dan clustering menggunakan algoritma K-Means. Data yang akan dianalisis mencakup 156.753 entri peserta, yang selanjutnya akan diolah untuk menemukan pola dan informasi penting. Melalui EDA, penulis berencana memetakan sebaran nilai peserta berdasarkan provinsi serta menganalisis dimensi berpikir komputasional yang diukur dalam tantangan tersebut, seperti *Abstraction*, *Algorithmic Thinking*, *Decomposition*, *Pattern Recognition*, *Evaluation*, *Generalization*, *Logic*, dan *Data Analysis*. Pada tahap selanjutnya, metode K-Means Clustering akan diterapkan untuk mengelompokkan provinsi berdasarkan pola nilai, dengan pemilihan jumlah kluster optimal menggunakan metode Elbow dan Silhouette Score. Tujuan dari penelitian ini adalah untuk mengidentifikasi pola tersembunyi dalam data, memahami variasi kemampuan berpikir komputasional siswa secara geografis, serta memberikan rekomendasi berbasis data bagi pihak penyelenggara, pendidik, maupun pembuat kebijakan. Dengan pendekatan ini, diharapkan dapat teridentifikasi daerah-daerah yang memerlukan intervensi lebih lanjut maupun daerah dengan potensi yang dapat dijadikan acuan keberhasilan.

Tabel 2.1. Perbandingan Penelitian

Pembanding	Tujuan	Metode Algoritma	Jumlah Sampel	Jumlah Cluster	Objek	Tools
[12]	Mengetahui informasi terkait wilayah yang potensial sebagai	Clustering K-Means	1.213	3	Desa	-

	penjualan sepeda motor Honda di Kabupaten Indragiri Hilir					
[13]	Mengelompokkan data pegawai yang layak untuk lolos dalam program SPD (Staff Development Program)	Clustering K-Means	27	3	Pegawai	-
[7]	Memberikan kemudahan pada PT. XYZ untuk mengelola data transaksi penjualannya serta mengetahui tingkat penjualan produk pada setiap cabangnya di Sumatera	Clustering K-Means	712	3	Data Penjualan	Library Python: pandas, numpy, scikit-learn, matplotlib. Visualisasi: Scatter plot menggunakan matplotlib
[32]	Meneliti dan mengimplementasikan segmentasi pelanggan pada data transaksi <i>online retail</i> untuk mengidentifikasi karakteristik pelanggan dan menentukan strategi pemasaran yang lebih tepat.	Clustering K-Means	54.910	4	Data Penjualan	Library Python: Pandas, Numpy, Seaborn, Scikit-learn (K-Means, MinMaxScaler, StandardScaler, Metrics). Visualisasi: Scatter plot menggunakan matplotlib
[33]	Melakukan customer profiling pelanggan produk digital service Indihome menggunakan Algoritma K-Means serta mengetahui perilaku pelanggan dan menerapkan strategi pemasaran yang tepat untuk mendatangkan keuntungan bagi perusahaan	Clustering K-Means	1.392	3	Data Pelanggan	Library Python: pandas, numpy, Scikit-learn (K-Means, MinMaxScaler) Visualisasi: Scatter plot menggunakan matplotlib
[37]	Menganalisis prestasi akademik siswa dengan mengelompokkan tingkat pencapaian mereka menggunakan metode k-means clustering	Clustering K-Means	266	3	Siswa	RapidMiner untuk analisis data, penerapan algoritma K-Means, dan perhitungan validitas cluster menggunakan

						indeks Davies-Bouldin.
[38]	Mengelompokkan prestasi siswa berdasarkan nilai mata pelajaran dan absensi menggunakan metode k-means clustering	Clustering K-Means	25	3	Siswa	RStudio untuk analisis clustering K-Means dan metode <i>Simple Additive Weighting</i> (SAW) untuk menentukan siswa dengan peringkat tertinggi pada cluster tinggi.
Klasterisasi Kemampuan Berpikir Komputasional Siswa Berdasarkan Data Tantangan Bebras Menggunakan Algoritma K-Means	Tujuan penelitian ini adalah untuk menganalisis dan mengelompokkan data peserta Tantangan Bebras 2024 menggunakan EDA dan metode clustering.	EDA dan Clustering K-Means	156.753	3-4	Siswa	Microsoft SQL Server, Library Python: pandas, scikit-learn, matplotlib