BAB II

TINJAUAN PUSTAKA

2.1 *Rating*

Menurut Lackermair et al (2013) *rating* merupakan penilaian pelanggan pada skala tertentu. Sistem penilaian yang banyak digunakan untuk *rating* di Shopee adalah dengan memberikan bintang. Semakin banyak jumlah bintang yang diberikan menunjukkan peringkat penjual yang lebih tinggi. Tipe yang berbeda dari penilaian yang diberikan oleh banyak orang adalah *rating* rata-rata pembeli terhadap perbedaan fitur produk atau layanan yang ditawarkan oleh penjual (Filieri, 2014) dalam Nuraini Daulay (2020).

Dengan demikian, dapat disimpulkan bahwa *rating* yang diberikan para ahli di atas merupakan penilaian pelanggan terhadap kinerja toko *online*. Dari bintang satu hingga bintang lima, penjual *online* dapat menerima umpan balik menggunakan sistem penilaian ini. Meningkatnya jumlah bintang menunjukkan kemampuan penjual *online* untuk menawarkan layanan berkualitas tinggi kepada pelanggan. Hal ini sesuai dengan penelitian sebelumnya (Hulisi et al., 2012) yang menemukan bahwa peringkat pelanggan yang lebih baik sebanding dengan jumlah penjualan.

2.2 Aplikasi Shopee

Shopee merupakan salah satu *e-commerce* yang sedang berkembang di Indonesia dengan berbagai macam produk mulai dari pakaian hingga barang-barang untuk kebutuhan sehari-hari. Shopee juga menawarkan layanan pengiriman yang sudah terjamin dan metode pembayaran yang aman. PT Shopee Internasional Indonesia adalah pendiri Shopee, yang memulai debutnya di Indonesia pada bulan Desember 2015. Shopee merupakan anak perusahaan garena yang berpusat di Singapura dan hadir di Singapura, Malaysia, Thailand, Taiwan, Indonesia, Vietnam, dan Filipina.

Perusahaan *e-commerce* yang teliti dalam mempelajari karakteristik masyarakat Indonesia juga berkontribusi pada peningkatan kemampuan mereka untuk memasarkan produk mereka di pasar Indonesia. Menurut Handi Irawan (2008), penelitian yang dilakukan membagi sifat konsumen Indonesia menjadi sepuluh sifat yang berbeda. Sifat-sifat ini termasuk kecenderungan konsumen Indonesia untuk tidak memiliki perencanaan yang baik. Tingkat pembelian impulsif di Indonesia masih tinggi karena hal ini. Masyarakat Indonesia sering melakukan pembelian impulsif, yang biasanya dilakukan tanpa perencanaan. Pembelian *online* yang tiba-tiba dan langsung tanpa niat sebelumnya disebut sebagai pembelian implus *online* (Chan et al., 2017).

2.3 *Web Scraping*

Web scraping adalah teknik otomatis untuk mengekstrak dan mengumpulkan data dari situs *web*. Proses ini melibatkan pengambilan informasi terstruktur dari halaman *web*, biasanya melalui analisis kode HTML, dan menyimpannya dalam format yang dapat digunakan untuk analisis atau pengolahan lebih lanjut. Teknik ini memanfaatkan struktur dokumen HTML untuk mengidentifikasi dan mengekstrak data yang diinginkan. *Web scraping* memungkinkan pengumpulan informasi secara efisien dari berbagai sumber *online*, yang kemudian dapat disimpan dalam sistem file lokal atau *database* untuk penggunaan selanjutnya. Para ahli mendefinisikan *web scraping* sebagai metode untuk memperoleh data dari halaman *web* secara sistematis. Proses ini melibatkan pengambilan konten dari URL tertentu dan mengubahnya menjadi format yang dapat dianalisis. Tujuannya adalah untuk mengekstrak informasi yang relevan dari struktur HTML halaman *web*.

2.4 Analisis Sentimen

Analisis sentimen adalah jenis penelitian yang melihat opini, sentimen, evaluasi, penilaian, sikap, dan perasaan terhadap hal-hal seperti produk, layanan, organisasi, individu, masalah, peristiwa, topik, dan atributnya (Liu, 2012). Menurut Ling et al., (2014), tugas utama dalam analisis sentimen adalah menempatkan teks dalam dokumen, kalimat, atau pendapat ke dalam kelompok polaritas. Polaritas berarti bahwa teks dalam dokumen, kalimat, atau pendapat memiliki aspek positif atau negatif.

2.5 *Text Mining*

Text mining adalah proses penemuan informasi baru dengan mengekstrak pola secara otomatis dari berbagai sumber teks (Kini et al., 2015). Menurut Feldman dan Sanger (2007), *text mining* adalah proses mengekstraksi informasi berguna dari kumpulan dokumen dari waktu ke waktu melalui identifikasi pola yang menarik.

2.7 *Text Prerocessing*

Text preprocessing adalah tahap awal penerjemahan teks yang dimaksudkan untuk mempersiapkan teks untuk pemrosesan lebih lanjut (Feldman & Sanger, 2007). Untuk mencapai tujuan tersebut, karakter-karakter yang membentuk sebuah teks harus dipisahkan menjadi unsur-unsur yang memiliki makna tertentu.

2.7.1 *Case Folding*

Penghapusan karakter selain huruf, yaitu tanda baca dan angka merupakan ciri dari *case folding*. Huruf juga dapat diubah menjadi huruf kecil atau huruf besar (Rahman, 2017).

2.7.2 *Cleaning*

Menurut (Husada & Paramita, 2021) *cleaning* merupakan tahapan penghapusan karakter yang ada dalam teks pada dokumen seperti *hashtag*, URL, *emoticon*, *username*, dan lainnya dihapus. Tujuan tahapan ini adalah menghilangkan *noise* pada teks dokumen untuk memudahkan dalam pengolahan data.

2.7.3 *Stopword Removal*

Stopword Removal adalah proses penghapusan kata-kata yang tidak memiliki nilai penting dalam sebuah teks. Kata-kata tersebut seringkali digunakan secara umum dan tidak memberikan dampak signifikan pada pemrosesan dokumen seperti "atau", "dengan", "ke", "di", dan "tetapi" (Cahyono, 2017).

2.7.4 *Tokenizing*

Tokenizing adalah cara membagi suatu kalimat berdasarkan kata-kata penyusunnya (Rahman, 2017). Dapat digambarkan juga sebagai proses pemecah satu kalimat menjadi beberapa kata.

2.7.5 *Stemming*

Untuk mengelompokkan setiap kata yang memiliki kata dasar yang sama, dilakukan proses penghapusan imbuhan atau reduksi kata berimbuhan menjadi bentuk dasarnya kembali (Rahman, 2017). Tujuan dari proses ini adalah untuk memudahkan pengelompokkan dan klasifikasi data agar lebih terstruktur.

2.8 Pembobotan Kata

Fitur ekstraksi *Countvectorizer* pada penelitian kali ini, bekerja dengan cara merubah kumpulan teks data pada data set menjadi sebuah *matrix* dengan token jumlah kata yang ada pada sebuah *dataset*. *Countvectorizer* tidak hanya memberikan jalan yang mudah untuk mengubah kumpulan data teks dan membuat kosa kata dari kata kata berbeda yang diketahui, tapi juga *encode* menjadi dokumen baru menggunakan kosa kata tersebut (Irawaty, 2020).

Countvectorizer adalah fitur penghitungan numerik dan metode ekstraksi fitur teks yang umum digunakan. Pada proses data *training*, hanya mempertimbangkan setiap kata pada frekuensi data *training* dalam teks (Yang, 2020). *Vektor* dibangun berdasarkan dimensi ukuran kata pada *tweet*. Jumlahnya bertambah setiap kali satu kata ditemukan, dan dimensinya bertambah juga bertambah satu (Turki & Roy, 2022).

2.9 *Decision Tree*

Decision Tree adalah teknik yang populer dan kuat untuk klasifikasi dan prediksi. *Decision Tree* memiliki kemampuan untuk mengubah fakta menjadi sebuah pohon keputusan yang dapat menyampaikan aturan yang dapat dipahami dengan mudah. Metode *Decision Tree* mengubah data ke dalam bentuk pohon keputusan, mengubah pohon ke dalam bentuk peran, dan kemudian mengubah peran. Di dalam metode *Decision Tree*, *internal node, root node,* dan *terminal node* merupakan bagian dari pohon. Dalam klasifikasi, *root node* dan *internal node* dan label kelas adalah variabel atau fitur, dan data *query* akan mengikuti *root node* dan *internal node* hingga *terminal node.* Label kelas pada data *query* ini didasarkan pada label *internal node* sebelumnya (Christian, Yessica, & Indrastanti, 2021).

2.10 SMOTE

Nithes V. Chawla pertama kali memperkenalkan SMOTE, yang merupakan turunan dari *oversampling*. Untuk mengatasi masalah ketidakseimbangan kelas (CIP), SMOTE digunakan. Untuk meningkatkan kinerja metode klasifikasi, SMOTE memungkinkan *overfitting*, yaitu data pada kelas minoritas yang terduplikasi, melalui modifikasi *dataset* yang tidak seimbang.

Metode SMOTE berbeda dari metode *oversampling* yang digunakan sebelumnya dan menggunakan prinsip memperbanyak pengamatan secara acak untuk menyelesaikan masalah data tidak seimbang. Metode SMOTE menggunakan data buatan untuk membuat jumlah data kelas minor setara dengan kelas mayor.

2.11 Evaluasi

Pada tahapan evaluasi dilakukan dengan melakukan pengujian seberapa baik model klasifikasi yang telah dibuat. Evaluasi tersebut dilakukan dengan menghitung *precision, recall, f1-score,* dan *accuracy.* Untuk melakukan pengujian ini menggunakan *confusion matrix*. *Confusion matrix* adalah alat penting untuk evaluasi kinerja dalam klasifikasi data mining adalah matriks kekacauan, yang memberikan gambaran menyeluruh tentang hasil prediksi model. Dalam *matrix* ini, elemen-elemen utamanya yaitu *True Positive, True Negative, False Positive,* dan *False Negative.*

Tabel 2. 1 *Confusion Matrix*

	Predicted Positive	*Predicted Negative*
Actual Positive	*True Positive* (TP)	*False Negative* (FN)
Actual Negative	*False Positive* (FP)	*True Negative* (TN)

a. *True Positive* (TP) : menunjukkan data bernilai positif yang diprediksi benar sebagai positif.
b. *False Positive* (FP) : menunjukkan data bernilai negatif yang diprediksi sebagai positif.

c. *False Negative* (FN) : menunjukkan data bernilai positif yang diprediksi sebagai negatif.
d. *True Negative* (TN) : menunjukkan data bernilai negatif yang diprediksi benar sebagai negatif.

Perumusan perhitungan nilai *accuracy*, *precision*, *recall*, dan *f1-score* adalah sebagai berikut:

$$Accuracy = \frac{TP+TN}{(TP+TN)+(FP+FN)}$$

$$Precision = \frac{TP}{(TP+FP)}$$

$$Recall = \frac{TP}{(TP+FN)}$$

$$F1\text{-}score = \frac{(2\times Recall\times Precision)}{Recall+Precision}$$

2.12 Penelitian Terkait

Berikut beberapa penelitian terkait dengan permasalahan yang ada:

Tabel 2. 2 Penelitian Terkait

No	Nama/Tahun Penelitian Terkait	Judul Penelitian	Hasil Penelitian
1	Christian Cahyaningtyas, Yessica Nataliani, Indrastanti Ratna Widiasari (2021).	Analisis Sentimen pada *Rating* Aplikasi Shopee Menggunakan Metode *Decision Tree* Berbasis SMOTE.	Hasil penelitian menunjukkan bahwa algoritma *Decision Tree* dengan SMOTE menghasilkan nilai akurasi 99,91 persen, AUC (*Area Under Curve*) 0,999, *recall* 99,88 persen, dan nilai *precision* 99,98 persen. Hasil penelitian dengan

			algoritma *Decision Tree* tanpa SMOTE menunjukkan nilai akurasi 99,89 persen, AUC (*Area Under Curve*) 0,950, recall 99,88 persen, dan nilai *precision* 99,98 persen. Kesimpulan dari evaluasi yang ada adalah bahwa SMOTE dapat mempengaruhi nilai *precision* dan AUC (*Area Under Curve*). Nilai *recall* dan *precision* tidak mempengaruhi, atau hasilnya sama jika menggunakan atau tanpa SMOTE. Selisih nilai tepat adalah 0,02 persen dan AUC adalah 0,049.
2	Meishita Inelza Putri, Iqbal Kharisudin (2022).	Analisis Sentimen Pengguna Aplikasi *Marketplace* Tokopedia Pada Situs Google Play Menggunakan Metode *Support*	Karena memiliki nilai AUC paling tinggi, SMOTE adalah metode klasifikasi yang tepat untuk klasifikasi *review* pengguna Tokopedia karena dapat

		Vector Machine (SVM), *Naïve Bayes*, dan *Logistic Regression.*	meningkatkan kinerja data tidak seimbang. Ini menunjukkan bahwa *k-fold cross validation Support Vector Machine*-SMOTE lebih efektif dalam meningkatkan akurasi klasifikasi daripada non-SMOTE.
3	Rani Puspita, Agus Widodo (2021).	Perbandingan Metode KNN, *Decision Tree*, dan *Naïve Bayes* Terhadap Analisis Sentimen Pengguna Layanan BPJS.	Peneliti menggunakan alat *RapidMiner* versi 9.7.2. Hasil penelitian menunjukkan bahwa analisis sentimen data Twitter terhadap layanan BPJS dengan metode KNN mencapai tingkat akurasi 95.58% dengan *class precision* untuk prediktor *negative* 45.00%, predictor *positive* 0.00%, dan prediktor *neutral* 96.83%. Dengan metode *Decision Tree*, tingkat akurasi mencapai 96.13% dengan *class precision* untuk prediktor *negative* 55.00%,

			prediktor *positive* 0.00%, dan prediktor *neutral* 97.28%. Selanjutnya, metode *Naïve Bayes* mencapai akurasi 89.14% dengan *class precision* untuk prediktor *negative* 16.67%, prediktor *positive* 1.64%, dan prediktor *neutral* 98.40%.
4	Fely Dany Prasetya, Handoyo Widi Nugroho, Joko Triloka (2022).	Analisa Data Mining Untuk Prediksi Penyakit Hepatitis C Menggunakan Algoritma *Decision Tree* C.45 Dengan *Particle Swarm Optimization.*	Hepatitis adalah peradangan pada hati, atau hati, yang dapat disebabkan oleh genetika, infeksi virus, alkohol, atau obat-obatan berdasarkan laporan yang dibuat oleh organisasi di seluruh dunia. Hasil analisis data hepatitis dengan kategori *0=Donor Darah, 0s=dugaan Donor Darah, 1=Hepatitis, 2=Fibrosis, 3=Cirrhosis) pada penggunaan metode C4.5 Algoritma *Decision Tree* hanya pada klasifikasi

			untuk mendapatkan nilai akurasi yang tinggi, dengan nilai akurasi terbaik sebesar 99,35%. Kemudian, pengujian yang menggunakan optimasi algoritma PSO untuk algoritma *Decision Tree* C4.5 menghasilkan nilai akurasi terbaik sebesar 99,67%, yang menunjukkan bahwa optimasi algoritma PSO dapat meningkatkan akurasi algoritma *Decision Tree* C4.5 sehingga menghasilkan hasil yang lebih baik dengan perbedaan akurasi 0,32%.
5	Lutfiah Maharani Siniwi, Alan Prahutam, Arief Rachman Hakim (2021).	*QUERY EXPANSION RANKING* PADA ANALISIS SENTIMEN MENGGUNAKAN KLASIFIKASI *MULTINOMIAL NAÏVE BAYES.*	Hasil dari pengklasifikasian sentimen dapat digunakan sebagai referensi untuk pengembangan *e-commerce* di masa depan. Berdasarkan penjelasan tersebut,

			tujuan yang ingin dicapai adalah untuk mendapatkan nilai akurasi dan *kappa statistic* dari hasil kinerja klasifikasi sentimen yang terkait dengan ulasan aplikasi Shopee dengan menggunakan klasifikai *Multinomial Naïve Bayes* dan pilihan fitur *Query Expansion Ranking*. Diharapkan bahwa ulasan yang ditulis oleh pengguna aplikasi Shopee akan memberikan umpan balik untuk perbaikan di masa mendatang.
6	Agung Purwanto, Handoyo Widi Nugroho (2023).	ANALISA PERBANDINGAN KINERJA ALGORITMA C4.5 DAN ALGORITMA *K-NEAREST NEIGHBORS* UNTUK KLASIFIKASI PENERIMA BEASISWA.	Penugasan beasiswa adalah masalah manajemen operasi yang dihadapi administrator universitas, yang biasanya diselesaikan berdasarkan pengalaman pribadi *administrator*. Penelitian ini

			mengusulkan metode insentif yang terinspirasi oleh pemrograman dinamis untuk menggantikan proses pengambilan keputusan tradisional dalam penugasan beasiswa. Tujuannya adalah untuk menemukan skema penugasan beasiswa yang optimal dengan ekuitas tertinggi sambil memperhitungkan kendala praktis dan persyaratan ekuitas Metodologi yang digunakan dalam menentukan penerima beasiswa di Universitas Muhammdiyah Pringsewu adalah dengan membandingkan tahapan Algoritma C.45 dan Algoritma *K-Nearest Neighbors*. Dari beberapa data sampel calon penerima dari

			jurusan Algoritma *K-Nearest Neighbors* memiliki performansi yang lebih baik yaitu presisi 98,08%, akurasi 98,30% dan nilai *recall* 98,00%, dengan hasil AUC sebesar 1,000 sedangkan C4,5 algoritma mencapai 97,23% dengan nilai *precision* 94.43%, nilai *recall* 100,00% dan hasil AUC 0,956.
7	Aviv Fitria Yulia, Handoyo Widi Nugroho (2022).	Implementasi Algoritma *K-Means Classifier* Sebagai Pendukung Keputusan Penerima Dana Bantuan Siswa Miskin (Studi Kasus: SMKN Sukoharjo).	Manusia membutuhkan pendidikan dalam kehidupannya, Pendidikan merupakan usaha agar manusia dapat mengembangkan potensi dirinya melalui proses pembelajaran. Kemiskinan merupakan sebuah kondisi yang berada di bawah garis nilai standar kebutuhan minimum, baik untuk makanan dan non makanan, Program

			Bantuan Siswa Miskin (BSM) adalah program yang bertujuan untuk menghilangkan halangan siswa miskin untuk bersekolah,masalah yangterjadi di SMKN SUKOHARJO yaitu pihak sekolah mengalami kesulitan dalam penentuan penerima Bantuan Siswa Miskin (BSM), hal ini dikarenakan banyaknya kriteria yang harus di pertimbangkan dalam menentukan penerima bantuan seperti: jumlah siswa yaitu berjumlah 1044 siswa, penghasilan orang tua, beban orang tua, jarak tempuh, dan nilai siswa. membangun sistem seleksi penerima beasiswa di SMKN SUKOHARJO dengan proses selektif dan tepat sasaran teknik pemanfaatan data disebut

			juga *Data Mining*. Salah satu metode *Data mining* yang cukup populer yaitu *clustering* dengan menggunakan algoritma *K-Means. K-Means* dapat mengolah data tanpa diberitahu lebih dahulu label kelasnya. Penelitian ini akan menghasilkan tiga kelompok: layak menerima bantuan, dipertimbangkan menerima bantuan, tidak layak menerima bantuan. Pengolahan data seleksi penerimaan dana bantuan menggunakan algoritma *K-Means* mendapatkan hasil *Davies bouldin* indeks sebesar 0,262. Hasil tersebut dinilai cukup baik sebab semakin dekat hasil yang diperoleh dengan angka nol, maka kemiripan data

			anggota antar cluster semakin baik.
8	Muhammad Nur Ikhsanto, Handoyo Widi Nugroho (2015).	ANALISIS PERFORMA DAN DESAIN JARINGAN KOMPUTER MENGGUNAKAN *TOP-DOWN NETWORK* DESAIN STUDI KASUS PADA CV. MERAH PUTIH.	Analisis jaringan komputer sangat penting dan dapat membantu meningkatkan performa jaringan. Banyak perusahaan yang mendesain jaringan tidak sesuai dengan tujuan bisnis mereka sehingga performa jaringan yang ada pada perusahaan tidak sesuai dengan yang diharapkan. Dengan demikian diperlukan analisis pada jaringan komputer yang ada dalam perusahaan, baik dari sisi performa dan desain jaringan. Desain jaringan yang ada sangat berkaitan erat akan performa jaringan. Dalam penelitian ini parameter-parameter yang ada dalam jaringan komputer seperti *delay, jitter, bandwidth,*

			utilization, paket *loss* dan *throughput* akan diukur untuk menentukan performa jaringan dan kemudian parameter tersebut digunakan sebagai informasi untuk mendesain ulang jaringan agar performa jaringan menjadi baik serta menghasilkan desain jaringan yang lebih terstruktur sesuai akan kebutuhan perusahaan.
9	Muhamad Septa Utama, Handoyo Widi Nugroho (2023).	Kajian Algoritma C4.5 dan K-NN Untuk Memprediksi Penduduk Miskin.	Kemiskinan adalah isu serius yang mempengaruhi masyarakat di seluruh dunia. Perserikatan Bangsa-Bangsa (PBB) telah mengakui kemiskinan sebagai prioritas utama dalam Tujuan Pembangunan Berkelanjutan. Meskipun ada penurunan tingkat kemiskinan dalam

			beberapa tahun terakhir, masih banyak individu yang berjuang untuk memenuhi kebutuhan dasar mereka. Oleh karena itu, diperlukan upaya yang lebih efektif dalam mengidentifikasi penduduk miskin agar program-program bantuan dapat tepat sasaran. Penelitian ini bertujuan untuk membandingkan dua metode klasifikasi, yaitu C4.5 dan *K-Nearest Neighbors* (KNN), dalam memprediksi tingkat kemiskinan. Metode C4.5 menggunakan *Decision Tree* untuk mengklasifikasikan data, sementara KNN menggunakan jarak terdekat untuk melakukan klasifikasi. Data yang digunakan dalam penelitian ini adalah data kemiskinan

			di Indonesia. Metodologi penelitian ini melibatkan tahapan pra-pemrosesan data, termasuk pembersihan data, seleksi fitur, eksplorasi data, dan *balancing* data. Selanjutnya, dilakukan pelatihan dan pengujian model menggunakan algoritma C4.5 dan KNN. Hasil evaluasi kinerja model menggunakan metrik seperti akurasi, *recall*, presisi, F1 *measure*, dan *Area Under Curve* (AUC). Penelitian ini masih berada pada tahap desain model, dan tindak lanjut yang akan dilakukan adalah melanjutkan penelitian hingga hasil evaluasi algoritma. Dengan menggunakan *confusion matrix*, akan dipilih algoritme terbaik yang

			dapat mendeteksi penduduk miskin dengan akurasi yang tinggi. Hasil penelitian ini diharapkan dapat memberikan wawasan yang berguna dalam mengembangkan program-program bantuan yang efektif untuk pengentasan kemiskinan di Indonesia.