BAB II
LANDASAN TEORI

2.1 Data Mining

Data mining merupakan analisis peninjauan beberapa data untuk meringkas data agar dapat dipahami bagi pemilik data, dengan cara yang terstruktur (Anitha & Sridevi, 2019). Data mining adalah sebuah proses dengan mengggunakan teknik untuk mengolah data menggunakan statistik, kecerdasan buatan, perhitungan, dan *machine learning* untuk ekstraksi data (Utomo & Mesran, 2020). Data mining dapat digunakan untuk identifikasi sebuah pola dalam kumpulan data yang mencakup ilmu komputer dan statistik dengan menerapkan berbagai algoritme pembelajaran (Adekitan et al., 2019). Data mining memiliki kelebihan untuk menangkap data lebih cepat dan proses penguraian data, dan penyajian data dapat terstruktur dengan menampilkan hasil dengan teknologi (Xing & Bei, 2020).

Beberapa pengertian data mining tersebut dapat disimpulkan data mining merupakan analisis peninjauan beberapa dara dengan teknik pengolahan data menggunakan statistik, perhitungan, melibatkan kecerdasan buatan untuk menangkap data lebih cepat dan penyajian data ditampilkan dengan teknologi *machine learning*. Istilah data mining dapat disebut *Knowledge Discovery in Database (KDD)*, memiliki tujuan dalam data yang besar apakah ada pengetahuan yang berguna dengan ekstrak pola menggunakan algoritme (Ghazal & Hammad, 2022). Dalam beberapa pengertian data mining, melibatkan beberapa data latih dan data uji (*testing*) dengan memilih algoritme dalam pengelolaan data dapat diterapkan dalam penelitian ini dan memberikan hasil sebuah prediksi.

Dalam proses pengumpulan informasi, dan pengolahan data yang di uji, data mining menerapkan berbagai teknik (Adani et al., 2019) , antara lain :

a. *Association Discovery*, merupakan teknik yang mempelajari sekumpulan data untuk mengidentifikasi hubungan antar kemunculan beberapa atribut di dalam data. Teknik ini berusaha mengidentifikasi nilai yang sering muncul secara bersama di setiap baris data dan hasil disajikan dalam bentuk sebuah aturan.

b. *Clustering*, merupakan teknik yang mengidentifikasi sejumlah kelompok digunakan untuk input data, dalam *Clustering* kelompok kecil yang tersebar digabungkan menjadi kelompok besar berdasarkan kesamaan, dan dapat digunakan untuk deteksi cluster dari rekaman yang berdekatan menurut kriteria tertentu dari seluruh variabel.

c. *Sequintial Discovery*, merupakan teknik untuk menemukan pola diantara peristiwa yang terjadi rentang waktu tertentu. Metode ini digunakan untuk identifikasi pola komoditas yang berulang. Teknik ini fokus pada kebiasaan yang sering muncul di masa mendatang.

d. *Classification*, Penelitian lain klasifikasi dapat diartikan sebagai penentuan sebuah data baru ke beberapa kategori (kelas) yang telah didefinisikan. Dalam sebuah buku konsep data mining klasifikasi dapat digunakan untuk mengelompokkan data dalam kategori atau kelas berdasarkan atribut dan fitur tertentu (Witten et al., 2016).

e. *Neural Network,* merupakan metode khusus untuk identifikasi pola dan mengendalikan tren. Inti dari proses ini meniru fungsi sistem saraf manusia.

2.2 Klasifikasi

Klasifikasi merupakan pelabelan sebuah data dengan cara memisahkan sekumpulan data sesuai dengan jenis label atau kelas yang sudah di tentukan (Widaningsih, 2019). Klasifikasi dalam data mining digunakan untuk pengelompokan data dengan membangun model atau algoritme yang dapat memprediksi kelas dari data baru untuk mengambil keputusan berdasarkan pola yang teridentifikasi dalam data (Witten et al., 2016). Keunggulan penerapan klasifikasi menggunakan algoritme dapat dinilai dari kebenaran klasifikasi model terhadap data sebenarnya, dan tepatnya penerapan model dalam prediksi kelas klasifikasi. Pada penelitian sebelumnya terdapat beberapa metode klasifikasi diterapkan dengan algoritme *K-Nearest Neighbor (K-NN)*, *Naïve Bayesian, Neural Network*, *C4.5* (Dinata et al., 2020).

2.3 *Preprocessing*

Preprocessing merupakan teknik untuk mengubah data yang telah dikumpulkan dari beberapa sumber yang akan dilakukan pengolahan lebih lanjut. *Preprocessing* digunakan untuk mengatasi data yang tidak sesuai dengan sistem (Prasetyo et al., 2019). Dalam *preprocessing* data yang digunakan dalam penelitian ini menggunakan metode *MinMax Scaler.* Metode *MinMax Scaler* adalah *transformasi* fitur dengan menskala secara individual dengan rentang tertentu. *Preprocessing MinMax Scaler* melakukan proses pengurangan pada data yang akan dilakukan normalisasi dengan data terkecil pada fitur dan dilakukan pembagian dari hasil pengurangan nilai terbesar pada fitur dengan nilai terkecil pada fitur (Azzahra Nasution et al., 2019). Rumus *MinMax Scaler* dapat ditunjukkan pada persamaan 2.1 :

$$X_{new} = \frac{(X_{old} - X_{min})}{(X_{max} - X_{min})} \quad (2.1)$$

Keterangan :

X_{new} = Hasil normalisasi nilai

X_{old} = Nilai asli

X_{min} = Nilai minimal pada data usia

X_{max} = Nilai maksimal pada data usia

2.4 *K-Nearest Neighbor* (KNN)

K-Nearest Neighbor (K-NN) merupakan metode yang digunakan untuk klasifikasi suatu objek yang memiliki jarak terdekat, dengan mencoba data baru yang belum diketahui kelas atau label berdasarkan hasil *K* (Br Sinuhaji et al., 2024). Jarak pada metode KNN dapat dihitung menggunakan rumus *Manhattan* pada persamaan 2.2:

$$d = \sum_{i=1}^{p} |x_{1i} - x_{2i}| \quad (2.2)$$

Keterangan :

x_{1i} = Data *training*

x_{2i} = Data uji atau data *testing*

i = Variabel data

d = Jarak

p = Dimensi data

Langkah – langkah dalam metode *K-Nearest Neighbor (K-NN)* adalah :

a. Menentukan nilai parameter K
b. Menghitung jarak antara data *training* dan data *testing*.
c. Mengurutkan data dari terkecil ke terbesar berdasarkan jarak yang terbentuk.
d. Menetapkan kelas yang telah ditentukan dengan jumlah nilai K terbanyak pada data *testing*.

Kelebihan dalam penggunaan metode *K-Nearest Neighbor (K-NN)* adalah mudah dalam proses implementasi, dapat dilihat dari langkah-langkah metode *K-Nearest Neighbor (K-NN)* dalam prosesnya menggunakan dua parameter yaitu nilai K dan fungsi jarak. Kelebihan metode *K-Nearest Neighbor (K-NN)* lainnya adalah tidak memerlukan *training* sebelum prediksi sehingga saat ada data baru dapat dilakukan tanpa mengurangi nilai keakuratan. Adapun kekurangan pada metode *K-Nearest Neighbor (K-NN)* apabila terdapat data yang besar performa algoritme akan menurun dan diperlukan standarisasi terlebih dahulu dengan cara menormalisasikan data sebelum penerapan algoritme *K-Nearest Neighbor (K-NN)* (Kumar et al., 2018).

2.5 *Confusion Matrix*

Confusion Matrix merupakan metode yang digunakan untuk mengukur kinerja dari proses klasifikasi. Tabel *Confusion Matrix* ditunjukkan pada Gambar 2.1 :

True Label	*Negative*	*TN*	*FP*
	Positive	FN	*TP*
		Negative	*Positive*
		Predict Label	

Gambar 2.1 *Confusion Matrix*

Pada gambar 2.1 merupakan gambar tabel dari *Confusion Matrix*. TN (*True Negative)* dalam gambar tersebut berarti data negatif yang diprediksi benar*, TP*

(True Positive) berarti data positif di prediksi benar saat di uji , FN *(False Negative)* berarti data positif yang diprediksi negatif*, dan FP (False Positive)* berarti data negatif di prediksi positif. Pengukuran kinerja tersebut menggunakan akurasi, *recall*, dan *precission*. Pengujian akurasi adalah untuk identikasi sebuah rasio benar positif pada data yang telah di uji. Pengujian *recall* digunakan untuk identifikasi sebuah rasio pada data yang menunjukkan benar positif dengan membandingkan data keseluruhan yang benar positif, sedangkan pada pengujian *precission* digunakan untuk mengukur rasio yang benar positif terhadap semua hasil yang positif. Rumus akurasi, *recall*, dan *precission* ditunjukkan pada persamaan 2.3 , 2.4 , dan 2.5 :

$$\text{Akurasi} \quad \frac{TP+TN}{TP+TN+FP+FN} \text{ x } 100\% \qquad (2.3)$$

$$\textit{Recall} \quad \frac{TP}{TP+FN} \text{ x } 100\% \qquad (2.4)$$

$$\textit{Precission} \quad \frac{TP}{TP+FP} \text{ x } 100\% \qquad (2.5)$$

2.6 Penelitian Terkait

Sebagai penguatan topik penelitian, penulis melakukan analisis dari beberapa riset penelitian sebelumnya, sebagai berikut :

No	Landasan Literatur	Metode Penelitian	Permasalahan	Hasil Penelitian
1	Judul Penelitian : Klasifikasi Penerima Beasiswa Aceh Carong (Aceh Pintar) di Universitas Malikussaleh Menggunakan Algoritma KNN (*K-Nearest Neighbor*) Penulis : Ar Razi	*K-Nearest Neighbor*	Penentuan penerima beasiswa untuk putra-putri Aceh di Universitas Malikussaleh	Sistem Klasifikasi dapat membantu admin universitas dalam proses seleksi beasiswa dengan waktu yang efisien dan tepat sasaran.

	Tahun : 2022			
2	Judul Penelitian : Algoritme *K-Nearest Neighbor* Untuk Memprediksi Kelulusan Ujian Nasional Berbasis Desktop Pada SMAN 12 Tangerang Penulis : Pandu Yuli Santoso dan Dewi Kusumaningsih Tahun : 2018	*K-Nearest Neighbor*	Prediksi kesiapan siswa dalam Ujian Nasional di SMAN 12 Tangerang dalam Tryout Ujian. Pada sekolah tersebut memiliki banyak siswa dan pengolahan data masih secara manual, memiliki kendala efisiensi waktu dan prediksi kelulusan siswa yang secara manual.	Hasil prediksi kelulusan dengan 1140 data pada hasil nilai *Tryout* siswa data 5 tahun terakhir menghasilkan Sistem klasifikasi prediksi kelulusan yang dapat membantu pihak sekolah dalam pengambilan solusi bagi siswa dan siswi yang memperoleh nilai dibawah standar yang ditentukan.
3	Judul Penelitian : Aplikasi Klasifikasi Penerima Kartu Indonesia Sehat Menggunakan Algoritme *K-Nearest Neighbor* Penulis : Aninda Zulaifa Abidin dan Yogiek Indra Kurniawan Tahun : 2019	*K-Nearest Neighbor*	Penentuan pemberian Kartu Indonesia Sehat pada masyarakat yang berhak menerima atau yang tidak berhak menerima bantuan.	Hasil pengujian dengan metode KNN pada klasifikasi penerima KIS sebanyak 12 kali percobaan dengan data *testing* sebanyak 200 data yang diambil secara acak dapat membantu admin dalam menentukan masyarakat yang berhak mendapatkan Kartu Indonesia Sehat.

4	Judul Penelitian : Klasifikasi Penyakit Ginjal Kronis menggunakan *K-Nearest Neighbor* Penulis : Ardina Ariani dan Samsuryadi Tahun : 2019	*K-Nearest Neighbor*	Diagnosa pasien penyakit gagal ginjal kronis dengan 24 gejala.	Hasil penelitian dari dataset yang diperoleh dari *Repository University of California (*UCI *Repository Machine Learning Benchmark)* menghasilkan hasil klasifikasi cukup baik dalam klasifikasi penyakit gagal ginjal kronis.
5	Judul Penelitian : Penerapan Metode Klasifikasi *K-Nearest Neigbor* pada Dataset Penderita Penyakit Diabetes Penulis : Andi Maulida Argina Tahun : 2020	*K-Nearest Neighbor*	Pengukuran peforma metode klasifikasi dengan dataset penderita diabetes.	Penerapan metode KNN pada penderita diabetes sebanyak 77 dara dengan pembagian 90% data *training* dan 10% data *testing* menghasilkan kurang cukup dala proses identifikasi penderita diabetes
6	Judul Penelitian : Klasifikasi Kebakaran Hutan Menggunakan Algoritme *K-Nearest Neighbor* Penulis : Ari Rudiyan , Akhmad Erik	*K-Nearest Neighbor*	Klasifikasi daerah berpotensi kebakaran hutan Kalimantan Barat dan perancangan REST API untuk deteksi kebakaran hutan.	Hasil pengujian menggunakan data *testing* sejumlah 30% dari 14.201 data dan 70% data *training* dari 14.201 data menghasilkan penentuan pembuatan

	Dzulkifli, dan Khabib Munazar Tahun : 2022			REST API pada lokasi rawan terjadinya kebakaran hutan di daerah Kalimantan Barat.
7	Judul Penelitian : Klasifikasi Penerima Dana Bantuan Desa Menggunakan Metode KNN *(K-Nearest Neighbors)* Penulis : Riyan Latifahul Hasanah, Muhamad Hasan, Witriana Endah Pangesti, Fanny Fatma Wati, dan Windu Gata Tahun : 2019	*K-Nearest Neighbor*	Untuk mengklasifikasi data baru perihal kelayakan atau tidak layak menerima bantuan dana desa	Hasil klasifikasi pada dataset penerimaan bantuan dana desa dapat dikategorikan sesuai dengan layak atau tidak laya dari data desa yang di uji coba
8	Judul Penelitian : Penerapan Algoritma *K-Nearest Neighbor (KNN)* Untuk Klasifikasi Penyakit Diabetes Melitus Studi Kasus : Warga Desa Jatitengah Penulis : Happy Andrian Dwi Fasnuari, Haris Yuana, dan M. Taofik Chulkamdi	*K-Nearest Neighbor*	Tingkat kematian penderita diabetes tergolong tinggi dikarenakan penderita tidak merasakan gejala atau tidak memahami ciri-ciri penyakit tersebut. Penderita diabetes harus menjalani uji klinis kesehatan agar	Pengujian metode KNN dengan data warga penderita diabetes melitus sebanyak 108 data *training* dan 27 data *testing* dapat membantu pengenalan dini perihal penyakit diabetes pada seseorang yang belum

	Tahun : 2022		dapat di diagnosis dengan tepat.	mengetahui ciri penyakit tersebut.
9	Judul Penelitian : Algoritma K-Nearest Neighbor dengan Euclidean Distance dan Manhattan Distance untuk Klasifikasi Transportasi Bus Penulis : Rozzi Kesuma Dinataa, Hafizal Akbara, dan Novia Hasdyna Tahun : 2020	*K-Nearest Neighbor*	Memberikan rekomendasi terbaik dengan klasifikasi transportasi bus jalur Lhokseumawe-Medan.	Hasil penelitian transportasi bus jalur Lhokseumawe-Medan dengan metode KNN menggunakan rumus jarak *Manhattan* diperoleh hasil pengukuran jarak menggunakan rumus *Manhattan* lebih baik daripada rumus jarak *Euclidian*
10	Judul Penelitian : Penerapan Algoritma *K-Nearest Neighbor (KNN)* untuk Memprediksi Penyakit Stroke Penulis : Muhammad Naja Maskuri, Harliana, Kadek Sukerti, dan R.M. Herdian Bhakti Tahun : 2022	*K-Nearest Neighbor*	Upaya pencegahan penyakit stroke sejak dini , dikarenakan penyakit tersebut memiliki nilai kematian urutan nomor 3 tertinggi.	Hasil penelitian dari 100 dataset penyakit stroke dengan metode KNN, sebanyak 20 data yang di ujikan dari 80 data latih dapat berupaya mencegah penyakit stroke dari gejala yang dialami oleh penderita.