BAB 2
TINJAUAN PUSTAKA

2.1. DATA MINING

Data mining merupakan analisa yang dilakukan secara *automatic* (otomatis) pada data yang berjumlah besar dan kompleks yang bertujuan untuk mendapatkan nilai kecenderungan atau pola yang keberadaannya tidak disadari. Data mining merupakan proses menemukan sesuatu yang bermakna oleh suatu korelasi baru, pola dan juga tren yang terdapat dengan cara memilah-milah data yang berukuran besar, dimana data tersebut disimpan dalam *repository*, menggunakan teknologi sosialisasi pola serta stratistik dan teknik matematika (Larose, 2006). Menurut (Turban, 2005), *data mining* adalah proses yang memakai teknik statistik, teknik matematika, kecerdasan protesis, *machine learning* dalam melakukan ekstraksi dan mengidentifikasi informasi yang bermanfaat serta pengetahuan yang terkait oleh database yang besar.

Beberapa teknik dan sifat *data mining* adalah sebagai berikut :

1. Klasterisasi, adalah mempartisi *data-set* menjadi beberapa *sub-net* atau kelompok sedemikian rupa sehingga elemen-elemen dari suatu kelompok tertentu memiliki *set property* yang di *share* bersama, dengan tingkat similaritas tinggi dalam suatu kelompok yang rendah. Disebut juga dengan "*unsupervised learning*".
2. Regresi, adalah memprediksi nilai dari suatu variabel kontinyu yang diberikan berdasarkan nilai dari variabel yang lain, dengan mengasumsikan sebuah model ketergantungan linier atau nonlinier.
3. Klasifikasi. Adalah menemukan sebuah *record* data baru ke salah satu dari beberapa kategori (kelas) yang telah didefinisikan sebelumnya dan disebut dengan "*supervised learning*".
4. Kaidah Asosisasi (*association rule*), adalah mendeteksi kumpulan atribut-atribut yang muncul bersamaan (*co-occur*) dalam frekuensi yang sering dan membentuk sejumlah kaidah dari kumpulan-kumpulan tersebut. (Zulaifa Abidin and Kurniawan 2019).

2.2. KLASIFIKASI

Klasifikasi merupakan proses penemuan model membedakan kelas data, atau dengan cara mengklasifikasi data kedalam satu atau beberapa kelas yang sudah didefinisikan sebelumnya (Mustafa and Simpen 2019). Menurut Hermawati, Klasifikasi merupakan proses pembelajaran suatu fungsi tujuan (target) f yang memetakan tiap himpunan label kelas yang telah terdefinisi sebelumnya. Klasifikasi digunakan untuk memprediksi kelas dari objek yang kelasnya belum diketahui. Metode klasifikasi yang umum digunakan diantaranya adalah *Decision Tree, K-Nearest Neighbor, Naïve bayes, Neural Network, C4.5*, dan *Support Vector Machine* (Diansyah 2022).

Di dalam klasifikasi diberikan sejumlah *record* yang dinamakan *training set*, yang terdiri dari beberapa atribut, atribut dapat berupa kontinyu ataupun kategoris, salah satu atribut menunjukkan kelas *record*. Berikut adalah model klasifikasi seperti yang ditunjukkan **Gambar 2.1**

Gambar 2.1 Model Klasifikasi

Ada dua jenis model klasifikasi, yaitu :

1. Pemodelan deskriptif (*descriptive modelling*), yaitu model klasifikasi yang dapat berfungsi sebagai suatu alat penjelasan untuk membedakan objek-objek dalam kelas-kelas yang berbeda.
2. Pemodelan prediktif (*predictive modelling*), yaitu klasifikasi yang dapat digunakan untuk memprediksi label kelas *record* yang tidak diketahui.

   Pada proses klasifikasi didasarkan pada 4 (empat) komponen, yaitu :

   1. *Class*

      Variabel independen berupa kategori yang mempresentasikan "label" yang terdapat pada objek.
   2. *Predictor*

Variabel independen yang dipresentasikan oleh karakteristik data.

3. *Training dataset*

   Satu set data yang mempunyai nilai dari kedua komponen yang digunakan untuk menentukan kelas yang cocok berdasarkan predictor.

4. *Testing dataset*

   Berupa data baru yang diklasifikasikan oleh model data yang telah dibuat dan Accuracy klasifikasi evaluasi.

2.3. ALGORITMA *K-NEAREST NEIGHBOR*

Menurut Prasetyo, Purbaningtyas, dan Adityo tahun 2020, Algoritme *K-Nearest Neighbor* merupakan metode klasifikasi berdasarkan tetangga terdekat dengan konsep sederhana, kuat pada data non-linier, dan dapat digunakan dalam kasus multi-kelas(Prasetyo, Purbaningtyas, and Adityo 2020). Algoritme *K-Nearest Neighbor* merupakan metode klasifikasi terhadap sekumpulan data berdasarkan mayoritas, yang bertujuan untuk mengklasifikasikan objek baru berdasarkan atribut dan kategori yang sama dari sampel data training (Nikmatun et al. 2019) *K-Nearest Neighbor* termasuk dalam *supervised learning*, yang mana hasil dari *query instance* baru, diklasifikasikan berdasarkan mayoritas dari kategori pada KNN. Hasil dari klasifikasi diambil dari kelas yang paling banyak muncul, yang menjadi kelas hasil klasifikasi (Gorunescu, 2011). Rumus perhitungan jarak dengan *Euclidean* seperti dibawah ini :

$$d(x,y) = \sqrt{\sum_{i=1}^{n} (X_{training} - Y_{testing})^2} \qquad 2.1$$

Keterangan :

d (x,y)	: jarak *Euclidean*
Xtraining	: data training ke-*i*
Ytesting	: data testing

i : record (baris) ke-*i* dari tabel

n : jumlah data training

Langkah – langkah dalam menghitung Algoritme KNN :

1. Menentukan nilai K.
2. Menghitung kuadrat jarak *Euclidean* masing-masing label dari data training terhadap data testing yang diberikan.
3. Mengurutkan nilai dari hasil perhitungan jarak *Euclidean* data training terhadap data testing mulai dari nilai yang terkecil.
4. Melihat hasil kategori *nearest neighbor* dengan label kelas mayoritas terbanyak dari tetangga terdekat untuk dijadikan label kelas hasil klasifikasi.

2.4. ACCURACY

Algoritma Accuracy merupakan nilai atau ukuran dari suatu objek yang menentukan tingkat kemiripan dari objek tersebut kepada nilai objek aslinya. Nilai dari sebuah Accuracy dalam penelitian dirasa penting karena menjadi ukuran seberapa kuat metode tersebut digunakan dalam penelitian. Sebuah penelitian dapat dikatakan baik apabila memiliki nilai Accuracy yang tinggi, jika nilai Accuracy yang didapat dirasa kurang penelitian tersebut masih dapat dilanjutkan dengan cara mengubah atau menambahkan metode yang digunakan dengan harapan mendapat nilai Accuracy yang lebih baik, dimana nilai tersebut dapat menjadi acuan dalam melakukan penelitian selanjutnya.

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \; x \; 100\% \qquad (22.2)$$

Keterangan :

TP : Hasil positif yang diklasifikasikan dengan benar

TN : Hasil negatif yang diklasifikasikan dengan benar

FP : Hasil positif yang diklasifikasikan dengan salah

FN : Hasil negatif yang diklasifikasikan dengan salah

2.5. *RECALL*

Recall yang juga dikenal sebagai sensitivitas atau *true positive rate*, adalah metrik evaluasi yang digunakan untuk mengukur sejauh mana model mampu mengidentifikasi data positif secara benar dari keseluruhan data yang sebenarnya positif. Rumus untuk menghitung recall adalah sebagai berikut:

$$\text{Recall} = \frac{TP}{TP + FN} x100\% \quad (2.3)$$

- ***True Positive* (TP)**: Jumlah data positif yang berhasil diklasifikasikan dengan benar.
- ***False Negative* (FN)**: Jumlah data positif yang salah diklasifikasikan sebagai negatif.
- ***True Positive* + *False Negative***: Merupakan total jumlah data yang sebenarnya positif, baik yang terdeteksi dengan benar maupun yang terdeteksi salah.

Recall berfokus pada kemampuan model untuk menangkap semua kasus positif yang ada, sehingga metrik ini sangat penting dalam situasi di mana mengidentifikasi data positif lebih diprioritaskan, seperti dalam prediksi penyakit jantung. Nilai recall berkisar antara 0 hingga 1, di mana nilai yang lebih tinggi menunjukkan performa yang lebih baik dalam mendeteksi data positif.

2.6. *PRECISION*

Precision adalah metrik evaluasi yang digunakan untuk mengukur sejauh mana model mampu memberikan hasil prediksi positif yang benar dari keseluruhan data yang diprediksi sebagai positif. Precision sangat penting dalam situasi di mana Accuracy hasil positif menjadi prioritas, seperti dalam pengujian medis atau deteksi penipuan. Rumus untuk menghitung *precision* adalah sebagai berikut:

$$\text{Precision} = \frac{TP}{TP + FP} x100\% \quad (2.4)$$

- ***True Positive (TP)****:* Jumlah data positif yang benar-benar diklasifikasikan sebagai positif oleh model.
- ***False Positive (FP)****:* Jumlah data negatif yang salah diklasifikasikan sebagai positif oleh model.
- ***True Positive*** **+** ***False Positive***: Merupakan total jumlah data yang diprediksi sebagai positif, baik yang benar maupun yang salah.

Precision menunjukkan proporsi hasil prediksi positif yang benar-benar relevan. Nilai *precision* yang tinggi menunjukkan bahwa model memiliki tingkat keakuratan yang baik dalam menghasilkan prediksi positif, sehingga sangat membantu dalam mengurangi hasil positif palsu (*false positives*). Nilai *precision* juga berkisar antara 0 hingga 1, di mana nilai yang mendekati 1 mencerminkan model yang sangat andal dalam menghasilkan prediksi positif yang benar.

2.7. PENELITIAN TERKAIT

Sebagai upaya penguatan topik penilitan, penulis melakukan analisis dari hasil riset penelitian sebelumnya yang berkaitan dengan topik penelitian. Berikut ini beberapa hasil dari penelitian sebelumnya :

Tabel 2.1 Penelitian Terkait

Peneliti	Judul	Metode	Hasil
Widhi Ramdhani, David Bona, Rafi Bagus Musyaffa dan Chaerur Rozikin (2022)	Klasifikasi Penyakit Kangker Payudara Menggunakan Algoritma *K-Nearest Neighbor*	Algoritma *K-Nearest Neighbor*	Nilai persentasi pengklasifikasian penyakit kangker payudara dengan persentase 62,7% dalam kategori kangker jinak dan 37,3% dalam kategori kangker ganas.

Muhammad Naja Maskuri1, Harliana, Kadek Sukerti dan R.M. Herdian Bhakti (2022)	Penerapan Algoritma *K-Nearest Neighbor* (KNN) untuk Memprediksi Penyakit Stroke	Algoritma *K-Nearest Neighbor* (KNN)	Nilai k yang digunakan untuk menguji kinerja algoritma KNN dalam meprediksi penyakit stroke yaitu k-9, dimana didapatkan nilai Accuracy sebesar 95%. Berdasarkan hasil tersebut algoritma KNN memiliki tingkat Accuracy yang baik dalam melakukan proses prediksi penyakit stroke.
Andi Maulida Argina (2020)	Penerapan Metode Klasifikasi *K-Nearest Neigbor* pada Dataset Penderita Penyakit Diabetes	Algoritma *K-Nearest Neighbor* (KNN)	Hasil Accuracy tertinggi yaitu 39% pada K=3, presisi tertinggi yaitu 65% pada K=3 dan K=5, recall tertinggi yaitu 36% pada K=3, dan F-Measure tertinggi yaitu 46% pada K=3. Nilai yang diperoleh tidak cukup baik dikarenakan jumlah data yang digunakan cukup kecil.
M. Syukri Mustafa dan I Wayan Simpen (2019)	Implementasi Algoritma *K-Nearest Neighbor* (KNN) Untuk Memprediksi Pasien Terkena Penyakit Diabetes Pada Puskesmas Manyampa Kabupaten Bulukumba	Algoritma *K-Nearest Neighbor* (KNN)	Hasil pengujian data dari sistem ini yang menggunakan 104 data pasien pada puskesmas Manyampa memperoleh hasil prediksi yang benasr sebanyak 71 dan salah atau ragu-ragu sebesar 33 dengan tingkat Accuracy sebesar 68.3%.

Permana Putra, Akim M H Pardede dan Siswan Syahputra (2022)	Analisis Metode *K-Nearest Neighbour* (Knn) Dalam Klasifikasi Data Iris Bunga	Algoritma *K-Nearest Neighbor* (KNN)	Hasil pengujian menunjukan metode *K-Nearest Neighbor* dalam klasifikasi data memiliki Accuracy persentase yang baik ketika menggunakan data random. Persentase variasi nilai K *K-Nearest Neighbor* 3,4,5,6,7,8,9 memiliki persentase 100 %.
Anida Zulaifa Abidin dan Yogiek Indra Kurniawan (2019)	Aplikasi Klasifikasi Penerima Kartu Indonesia Sehat Menggunakan Algoritma *K-Nearest Neighbhor*	Algoritma *K-Nearest Neighbor* (KNN)	Berdasarkan pengujian data testing sebanyak 12 kali percobaan menghasilkan rata-rata nilai *accuracy* 97,66% *precison* 98,5% dan *recall* 96,5%.
Lalu Abd Rahman Hakim, Ahmad Ashril Rizal dan Dwi Ratnasari (2019)	Aplikasi Prediksi Kelulusan Mahasiswa Berbasis *K-Nearest Neighbor* (K-NN)"	Algoritma *K-Nearest Neighbor* (KNN)	Hasil prediksi kelulusan mahasiswa S1 Teknik Informatika angkatan 2014 sebesar 51 mahasiswa diprediksi lulus tepat waktu dan 196 orang diprediksi tidak lulus tepat waktu.
Ari Rudiyan, Akhmad Erik Dzulkifli dan Khabib Munazar (2022)	Klasifikasi Kebakaran Hutan Menggunakan Metode *K-Nearest Neighbor :* Studi Kasus Hutan	Algoritma *K-Nearest Neighbor* (KNN)	hasil pengujian menggunakan data testing sejumlah 30% dari 14.201 data, dan data training yang digunakan sejumlah 70% dari 14.201 data, didapatkan tingkat Accuracy sebesar 92%

	Provinsi Kalimantan Barat		dengan nilai K = 18. Nilai yang sama diperoleh dengan jumlah K sampai dengan 28.
Surya Diansyah (2022)	Klasifikasi Tingkat Kepuasan Pengguna dengan Menggunakan *Metode K-Nearest Neighbour* (KNN)	Algoritma *K-Nearest Neighbor* (KNN)	Dimana k = 5 memiliki Accuracy yang tinggi yaitu sebesar 94.12% dengan *error* sebesar 5.88%. Sehingga penelitian ini dapat menjadi rujukan dalam mengklasifikasikan pengguna jasa.
Esty Purwaningsih dan Ela Nurelasari (2021)	Penerapan *K-Nearest Neighbor* Untuk Klasifikasi Tingkat Kelulusan Pada Siswa	Algoritma *K-Nearest Neighbor* (KNN)	Metode '*K-Nearest Neighbor* (KNN) yang diproses dengan tools rapidminer 9.0 didapatkan rata-rata Accuracy sebesar 96,49%.