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 :

a. 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*".
b. Regresi, adalah memprediksi nilai dari suatu variabel kontinyu yang diberikan berdasarkan nilai dari variabel yang lain, dengan mengasumsikan sebuah model ketergantungan linier atau nonlinier.
c. Klasifikasi. Adalah menemukan sebuah *record* data baru ke salah satu dari beberapa kategori (kelas) yang telah didefinisikan sebelumnya dan disebut dengan "*supervised learning*".
d. 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. (Hermawati, 2013).

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 (Saputra & Primadasa, 2018). Menurut (Hermawati, 2013), 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 0.1**

Gambar 0.1. 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 akurasi klasifikasi evaluasi.

2.3. ALGORITME *K-NEAREST NEIGHBOR*

Menurut (Prasetyo, Purbaningtyas, & Adityo, 2019), 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. 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 (Putra, Pardede, & Syahputra, 2022). *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 :

$$d(x,y) = \sqrt{\sum_{i=1}^{n} (X_{training} - Y_{testing})^2} \qquad (0.1)$$

Keterangan :

d (x,y)	: jarak *Euclidean*
$X_{training}$	: data training ke-*i*
$Y_{testing}$	: 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. ALGORITME *NAÏVE BAYES*

Algoritme *Naïve Bayes* merupakan salah satu algoritme yang masuk dalam *supervised learning*. Algoritme *Naïve Bayes* merupakan metode pengklasifikasian berdasarkan nilai probabilitas setiap atribut yang terdapat di dalam data set, klasifikasi *naïve bayes* didasarkan pada teorema bayes (*bayes theorem*). Algoritma *Naïve Bayes* diperkenalkan oleh ilmuwan inggris Thomas Bayes. Algoritma *Naïve Bayes* mengadopsi ilmu statistika yaitu menggunakan teori kemungkinan (probabilitas) untuk menyelesaikan sebuah kasus supervised learning, artinya dalam himpunan data terdapat label, class atau target sebagai acuan (Tias Mugi Rahayu, 2021). Keunggulan dari algoritme ini adalah tingkat pemrosesan yang cepat dan tingkat akurasi yang tinggi mesikipun dalam jumlah data set yang besar.

2.5. AKURASI

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

$$\text{Akurasi} = \frac{TP+TN}{TP+TN+FP+FN}\ x\ 100\% \qquad (0.2)$$

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

2.6. 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 :

1. Aninda Zulaifa Abidin dan Yogiek Indra Kurniawan (2019) dengan judul penelitian "Aplikasi Klasifikasi Penerima Kartu Indonesia Sehat Menggunakan Algoritme *K-Nearest Neighbor*" mendapatkan kesimpulan, bahwa penerapan Algoritme *K-Nearest Neighbor* didalam sistem dirasa sesuai, dapat dilihat dari nilai pengujian data testing sebanyak 12 kali percobaan menghasilkan rata-rata nilai accuracy 97,66% precison 98,5% dan recall 96,5%.
2. Eka Wahyu Sholeha, Selviana Yunita, Rifqi Hammad, Veny Cahya Hardita, dan Kaharuddin (2022) dengan judul penelitian "Analisis Sentimen Pada Agen Perjalanan Online Menggunakan *Naïve Bayes* dan *K-Nearest Neighbor*" mendapatkan kesimpulan, bahwa didapatkan akurasi tertinggi ketika seluruh data menggunakan huruf kecil untuk kedua Algoritme dengan akurasi 52,35%, namun berdasarkan hasil penelitian ditemukan bahwa Algoritme *K-Nearest Neighbor* memiliki nilai akurasi yang lebih baik pada nilai rata-rata dibandingkan *Naïve Bayes*.
3. Aria Pratama, Farid Ali Ma'ruf, Iin, Ade Rizki Rinaldi, dan Faturrhohman (2021) dengan judul penelitian "Klasifikasi Penerimaan Beasiswa Menggunakan Algoritme *K-Nearest Neighbor*" mendapat kesimpulan, bahwa dalam penerapan Algoritme *K-Nearest Neighbor* dalam melakukan klasifikasi penerima beasiswa mendapat nilai akurasi sebesar 78,45%, nilai *precision* sebesar 25%, dan nilai *recal* sebesar 9,52%.

4. Yulia Rizki Amalia (2018) dengan judul penelitian "Penerapan Data Mining Untuk Prediksi Penjualan Terlaris Menggunakan Metode *K-Nearest Neighbor*" mendapatkan kesimpulan, bahwa dengan menerapkan metode *K-Nearest Neighbor* didapatkan hasil prediksi penjualan produk terlaris ebanyak 6 jenis produk dari 22 jenis produk, dengan produk terlaris dengan nilai tertinggi yaitu Mesin Cuci dan LCD dengan nilai akurasi sebesar 92,51%.
5. Lalu Abd Rahman Hakim, Ahmad Ashril Rizal, dan Dwi Ratnasari (2019) dengan judul penelitian "Aplikasi Prediksi Kelulusan Mahasiswa Berbasis *K-Nearest Neighbor*" mendapatkan kesimpulan, hasil perhitungan dari metode *K-Fold Cross Validation* mendapatkan hasil tertinggi pada model ketiga 80% ketika K = 4 dan 61% ketika K = 1, hasil yang didapatkan adalah 143 mahasiswa lulus tepat waktu dan 104 mahasiswa lulus tidak tepat waktu, sedangkan hasil perhitungan dari metode *Confusion Matrix* menunjukkan hasil akurasi tertinggi dengan nilai 98% pada K = 1 untuk "Tepat Waktu", dan nilai 98% pada K = 2 untuk "Tidak Tepat Waktu", hasil yang didapatkan 51 mahasiswa lulus tepat waktu dan 196 mahasiswa lulus tidak tepat waktu.
6. Pandu Yuli Santoso dan Dewi Kusumaningsih (2018) dengan judul penelitian "Algoritme *K-Nearest Neighbor* Untuk Memprediksi Kelulusan Ujian Nasional Berbasis Desktop Pada SMAN 12 Tangerang" mendapatkan kesimpulan, nilai akurasi rata-rata sistem yang didapat setelah dilakukan 5 kali testing dalam melakukan prediksi sebesar 88,5% dengan nilai tertinggi pada K = 7.
7. Ari Rudiyan, Akhmad Erik Dzulkifli, dan Khabib Munazar (2022) dengan judul penelitian "Klasifikasi Kebaran Hutan Menggunakan Algoritme *K-Nearest Neighbor*" mendapat kesimpulan, bahwa dalam penerapan Algoritme *K-Nearest Neighbor* dilakukan pengujian dengan data testing 30% dan data training 70% dari 14.201 data, didapatkan nilai akurasi tertinggi sebesar 92% dengan nilai K =18, dan diperoleh nilai yang sama pada nilai K = 28.
8. Inna Alvin Nikmatun dan Indra Waspada (2019) dengan judul penelitian "Implementasi *Data Mining* Untuk Klasifikasi Masa Studi Mahasiswa

Menggunakan Algoritme *K-Nearest Neighbor*" mendapatkan kesimpulan, bahwa dalam penerapan Algoritme *K-Nearest Neighbor* hasil dari enam skenario pengujian mendapatkan nilai akurasi tertinggi pada skenario yang menggunakan atribut data Mata Kuliah Pilihan (MKP) dengan nilai akurasi sebesar 75,95%.

9. Gilang Atala Panharesi (2022) dengan judul penelitian "Klasifikasi Penyelesaian Skripsi Mahasiswa Menggunakan Metode *Weighted Naïve Bayes*" mendapat kesimpulan, hasil dari klasifikasi waktu penyelesaian skripsi mahasiswa dengan waktu 1 semester, 2 semester, dan > 2 semester adalah 0,14286, 0,53968, dan 0,31746. Hasil pengujian performa menggunakan *Confusion Matrix* mendapat nilai akurasi sebesar 77,5%, *precision* kelas 1 sebesar 100%, *precision* kelas 2 sebesar 72,3%, dan *precision* kelas 3 82,4%, *recall* kelas 1 sebesar 11,1%, *recall* kelas 2 sebesar 100%, dan *recall* kelas 3 sebesar 93,3%, *specificity* kelas 1 sebesar 100%, *specificity* kelas 2 sebesar 75%, dan *specificity* kelas 3 sebesar 88%.
10. Tias Mugi Rahayu, Besse Arnawisuda Ningsi, Isnurani, dan Irvana Arofah (2021) dengan judul penelitian "Klasifikasi Ketepatan Waktu Kelulusan Mahasiswa Dengan Metode *Naïve Bayes*" mendapat kesimpulan, dari total data yang berjumlah 2478 data mahasiswa, diketahui 61,9% lulus tepat waktu berjumlah 697 mahasiswa dan 38,1% lulus tidak tepat waktu berjumlah 429 mahasiswa. Dari perhitungan menggunakan Algoritme *Naïve Bayes* bahwa 225 data testing yang diuji menunjukkan 156 data terklasifikasi secara benar. Hasil klasifikasi ketepatan kelulusan mahasiswa menggunakan Algoritme *Naïve Bayes* diperoleh tingkat akurasi sebesar 69,33%.
11. Sri Hartati dan Haris Anom SAN (2022) dengan judul penelitian "Algoritme *Naïve Bayes* untuk Prediksi Kelulusan Mahasiswa" mendapatkan kesimpulan dari total dataset kelulusan mahasiswa sejumlah 321 mendapatkan nilai akurasi hasil prediksi sebesar 80%, dengan nilai *precision* sebesar 88% dan nilai *recall* sebesar 88%.
12. Lila Setiyani, Mokhamad Wahidin, Dudi Awaludin, dan Sri Purwani (2020) dengan judul penelitian "Analisis Prediksi Kelulusan Mahasiswa Tepat

Waktu Menggunakan Metode *Data Mining Naïve Bayes* : *Systematic Review*" mendapat kesimpulan, bahwa metode *Naïve Bayes* dirasa sesuai dalam melakukan prediksi kelulusan mahasiswa dengan menghitung dengan mendapat nilai akurasi dari ketiga literatur diatas 90%.

13. Fajar Edi Prabowo dan Achmad Kodar (2019) dengan judul penelitian "Analisis Prediksi Masa Studi Mahasiswa Menggunakan Algoritme *Naïve Bayes*" mendapatkan kesimpulan, dari total 244 data testing dan 62 data testing nilai akurasi yang didapatkan algoritme *Naïve Bayes* dalam melakukan prediksi masa studi mahasiswa sebesar 82,26%.
14. Endang Etriyanti, Dedy Syamsuar, dan Yesi Novaria Kunang (2020) dengan judul penelitian "Implementasi Data Mining Menggunakan Algoritme *Naive Bayes Classfier* dan *C4.5* Untuk Memprediksi Kelulusan Mahasiswa" mendapat kesimpulan, dari total dataset yang digunakan algoritme *Naïve Bayes* mendapat nilai *accuracy* sebesar 78,46% sedangkan untuk algoritme *C4.5* mendapat nilai *accuracy* sebesar 79,08%. Dengan demikian algoritme *C4.5* dinilai lebih baik dalam melakukan prediksi kelulusan mahasiswa.
15. Jaka Tirta Samudra, B. Herawan Hayadi, dan Puji Sari Ramadhan (2022) dengan judul penelitian "Komparasi 3 Metode Algoritme Klasifikasi Data Mining Pada Prediksi Kenaikan Jabatan" mendapatkan kesimpulan, berdasarkan hasil evaluasi klasifikasi kenaikan jabatan menggunakan klasifikasi model Naïve Bayes, K-Nearest Neighbor, dan Neural Network ini, dapat disimpulkan bahwa dalam kasus menggunakan dataset dari kampus universitas quality untuk yang terbaik mengklasifikasikan dataset kenaikan jabatan terlihat dari nilai accuracy, F1, precision dan recall yang dihasilkannya dari naïve bayes stabil untuk hasil nilainya serta untuk 5-fold cross validation pada accuracy, recall dan F1 naïve bayes nilainya tinggi 76.6% untuk precision K-Nearest Neighbor lebih tinggi 61.5%. Setelah itu untuk 10-fold cross validation pada accuracy naïve bayes nilainya tinggi 76.6%, untuk nilai F1 neural network lebih tinggi 66.4%, untuk precision k-nearest neighbor lebih tinggi 63.9%, untuk nilai recall model naïve bayes dan neural network sama-sama tinggi 76.6%. Setelah itu untuk 20-fold cross

validation pada accuracy naïve bayes nilainya tinggi 76.6%, untuk nilai F1 k-nearest neighbor lebih tinggi 67.8%, untuk precision k-nearest neighbor lebih tinggi 65.9%, untuk nilai recall model naïve bayes lebih tinggi 76.6%.

16. Silvana Puspa Nabila, Nurissaidah Ulinnuha, dan Ahmad Yusuf (2021) dengan judul penelitian "Model Prediksi Kelulusan Tepat Waktu Dengan Metode *Fuzzy C-Means* dan *K-Nearest Neighbor* Menggunakan Data Registrasi Mahasiswa" mendapatkan hasil dengan menggunakan teknik pengujian 10-fold cross validation adalah k=1 dengan menggunakan 3 cluster sedangkan jika menggunakan pengujian training 60% dan testing 30% didapatkan hasil 3 cluster dengan K=7, dan performa dari metode FCM-KNN dengan pengujian 10-fold cross validation diperoleh tingkat rata rata akurasinya adalah 71%.
17. Anjelika Hutapea, M. Tanzil Furqon, Indriati (2018) dengan judul penelitian "Penerapan Algoritme *Modified K-Nearest Neighbor* (M-KNN) Pada Pengklasifikasian Penyakit Kejiwaan *Skizofrenia*" mendapat kesimpulan, bahwa hasil pengujian pengaruh nilai K, memperoleh nilai persentase optimum dengan akurasi 37,045% pada pengujian nilai K=7. Pada pengujian tersebut menggunakan nilai K=1 hingga K=10, hasil pengujian pengaruh nilai *K-Fold*, diperoleh nilai persentase optimum dengan akurasi 28,4462% dengan nilai *K-Fold*=7.
18. Ismail Habibi Herman, Didit Widiyanto, dan Iin Ernawati (2020) dengan judul penelitian "Penggunaan *K-Nearest Neighbor* Untuk Mengidentifikasi Citra Batik Pewarna Alami dan Pewarna Sintetis Berdasarkan Warna" mendapat kesimpulan, bahwa algoritma *K-Nearest Neighbor* (KNN) mampu mengidentifikasi jenis batik pewarna alami dan batik pewarna sintetis menggunakan citra yang diklasifikasi berdasakran nilai ciri warna, algoritma KNN dengan nilai neighbor K=1, nilai K=3, dan nilai K=5 dapat mengklasifikasi citra batik dengan tingkat akurasi 100%, nilai *neighbor* (K) yang lebih besar dari 5 akan terjadi penurunan akurasi, hal ini ditunjukan pada nilai K=7 dengan nilai akurasi sebesar 50% dan nilai K=9 dengan nilai akurasi sebesar 75%.

19. Huzain Azis, Purnawansyah, Farniwati Fattah, dan Inggrianti Pratiwi Putri (2020) dengan judul penelitian "Performa Klasifikasi *K-Nearest Neighbor* dan *Cross Validation* Pada Data Pasien Penyakit Jantung" mendapat kesimpulan, bahwa pada dataset1 (dataset 50:50) di peroleh nilai performa paling baik pada nilai akurasi sebesar 82%, presisi 82%, recall 82% dan f-measure 82%, pada K=13. Dataset2 (dataset 20:80) di peroleh nilai performa paling baik pada nilai akurasi sebesar 87%, presisi 87%, recall 97%, dan f-measure 92%, pada K=3. Dataset3 (dataset 80:20) di peroleh nilai performa paling baik pada nilai akurasi sebesar 91%, presisi 92%, recall 60% dan f-measure 72%, pada K=5.
20. Muhammad Sofi Yuniarto dan Eko Adi Sarwoko (2018) dengan judul penelitian "Implementasi Metode *K-Nearest Neighbor* untuk Diagnosis Kanker Kolorektal Dengan *Biomarker Micro-RNA*" mendapat kesimpulan, bahwa hasil performa terbaik dari pengujian beberapa nilai K dengan metode KNearest Neighbor menghasilkan accuracy 94,17%, specificity 94,43%, dan sensitivity 94,41% pada K=3, dan berdasarkan kurva ROC, pada pengujian K=3 didapatkan model dengan performa terbaik pada fold 9 dengan nilai accuracy 100%, specificity 100%, dan sensitivity 100%. Sedangkan model dengan performa terjelek pada fold 2 dengan nilai accuracy 91,667%, specificity 86,667%, dan sensitivity 96,667%.

Berdasarkan hasil analisis penelitian terdahulu, maka penelitian klasifikasi kelulusan mahasiswa dengan menggunakan metode *K-Nearest Neighbor* dapat dilakukan.