

BAB 2

TINJAUAN PUSTAKA

2.1. Kinerja Mahasiswa

Menurut Naomi dan Nindyati tahun 2008, kinerja akademik merupakan hasil akhir yang dicapai oleh peserta didik sebagai keberhasilan selama mengikuti pendidikan dalam sebuah institusi pendidikan (I Made Indra P., S.KM., MPH., QRG. et al., 2021).

2.2. Pembelajaran *Hybrid*

Pembelajaran *Hybrid* adalah metode pembelajaran campuran, antara pembelajaran tatap muka dengan pembelajaran dalam jaringan (Inspektorat Jenderal Kementerian Pendidikan dan Kebudayaan, 2020). Konsep pembelajaran *hybrid* memungkinkan sebagian mahasiswa yang berada di satu kelas yang sama diminta untuk melakukan perkuliahan secara daring dari rumah, sebagian mahasiswa lainnya melakukan perkuliahan di kampus secara luring di saat yang bersamaan. Di UMKT sendiri, pembelajaran *hybrid* menggunakan sebuah platform *Learning Management System* (LMS) yang bernama OpenLearning guna memajemen sumber daya pembelajaran.

2.3. *Data Mining*

Data Mining adalah proses penggalian dan analisis menggunakan perangkat otomatis atau semi-otomatis untuk mengolah sejumlah besar data dengan tujuan mengidentifikasi pola yang signifikan (Muflikhah et al., 2018). Algoritma *Data Mining* dibagi menjadi beberapa kelompok berdasarkan fungsi dan tujuannya masing-masing, di antaranya (Buuololo, Efori, S.kom., 2020):

a. Deskripsi

Algoritma deskripsi digunakan untuk mengidentifikasi pola yang sering terjadi dan berulang. Pola tersebut kemudian diubah menjadi sebuah aturan yang dapat diterapkan

b. Klasifikasi

Algoritma klasifikasi melakukan pengelompokan data dari hubungan antara variabel kriteria dan variabel target

c. Prediksi

Prinsip kerja dari algoritma prediksi hampir serupa seperti algoritma klasifikasi, yang menjadi pembeda adalah algoritma prediksi tidak hanya mengidentifikasi target dari data yang ada, namun juga menghasilkan *output* yang bisa digunakan untuk memprediksi masa depan

d. Estimasi

Algoritma estimasi didefinisikan sebagai perkiraan atau prediksi. Perbedaannya terletak pada fokus pengelompokannya, di mana pengelompokan pada estimasi lebih berfokus pada nilai numerik dan daripada kategori

e. Pengklasteran

Pengklasteran melakukan pengelompokan data dari kemiripan nilai (homogen). Data yang dapat dikelompokkan merupakan hasil pengamatan *record* data, kelas dan objek-objek yang memiliki kesamaan. Jika dibandingkan dengan klasifikasi, pengklasteran tidak melibatkan penggunaan variabel keputusan atau target

f. Asosiasi

Asosiasi adalah suatu kelompok, kumpulan, atau persekutuan. Proses asosiasi melakukan pencarian atribut yang muncul atau selalu muncul secara bersamaan.

2.3.1. Algoritma Klasifikasi

Klasifikasi merupakan salah satu peran dalam data *mining* yang menerapkan metode pendekatan prediktif, di mana sebuah kumpulan data uji digunakan untuk mengevaluasi keakuratan suatu model, yang pada umumnya *dataset* dibagi menjadi dua bagian, yaitu data latih dan data uji. Data latih digunakan untuk membentuk model, sedangkan data uji digunakan untuk menguji model tersebut (Muflikhah et al., 2018). Beberapa algoritma klasifikasi termasuk

Cart, ID3, C4.5, J48, C5.0, *nearest neighbor*, *naive bayes*, dan masih banyak lagi (Buulolo, Efori, S.kom., 2020).

2.3.2. *k-Nearest Neighbor*

Sesuai dengan namanya, "*nearest neighbor*", KNN mengklasifikasikan data berdasarkan prinsip "kedekatan" dengan tetangga terdekatnya (Primartha, 2021). Dalam konteks ini, jumlah data/tetangga terdekat dalam algoritma KNN ditentukan oleh pengguna yang disebut dengan "K". Secara sederhana, KNN dapat dijelaskan sebagai berikut (Primartha, 2021):

1. Tentukan k buah tetangga terdekat (*nearest neighbor*) dengan cara acak
2. Ubahlah himpunan data ke dalam bentuk ruang vektor
3. Bagi himpunan data menjadi data pelatihan dan data pengujian
4. Lakukan perhitungan jarak (d) antara data pengujian dengan data pelatihan
5. Urutkan nilai d dari yang terkecil hingga yang terbesar
6. Ambil dan pisahkan data dari hasil penyortiran (*sorting*) sejumlah k
7. Perhatikan kelas (*class*) mayoritas
8. Lakukan klasifikasi data pengujian berdasarkan mayoritas.

Untuk menemukan kedekatan data uji dengan tetangga terdekat, dapat dicari menggunakan rumus jarak *Euclidean distance* (Primartha, 2021).

2.3.3. *Euclidean Distance*

Menurut Banjarsari dkk (2015), fungsi dari *Euclidean distance* yaitu untuk mengukur kedekatan antara dua buah objek dengan menggambarkannya sebagai garis lurus yang teratur dan langsung (Habibi & Santika, 2020). Rumus persamaan *Euclidean distance* dapat dituliskan sebagai berikut (Indriyanto, 2021):

$$Euclidean\ Distance = \sqrt{\sum_{i=1}^p (a_k - b_k)^2} \quad (2.1)$$

- a_k = Sampel data
 b_k = Data *testing*
 p = Dimensi data

i = Variabel data

2.3.4. Data Preprocessing

Data preprocessing melibatkan tahapan dalam proses *Data Mining* yang bertujuan untuk mendapatkan hasil analisis yang lebih akurat, serta bermanfaat untuk mengurangi waktu komputasi dan mengompresi nilai data tanpa mengubah informasinya (Santosa & Umam, 2018). *Data preprocessing* terdiri dari empat tahap utama, yaitu seleksi data, pemrosesan data atau pembersihan data, transformasi data dan reduksi data (Nofriansyah & Widi Nurcahyo, 2015).

2.3.5. K-Fold Cross-Validation

Cross-Validation adalah metode statistik yang digunakan untuk mengevaluasi kinerja suatu model algoritma dengan melibatkan pembagian data ke dalam dua subset, yaitu data pelatihan (*training*) dan data pengujian (*testing*) (Daqiqil, 2021). Model dilatih menggunakan *subset* pelatihan dan divalidasi oleh *subset* pengujian, kemudian pemilihan jenis *Cross-Validation* bergantung pada ukuran kumpulan data yang digunakan (Daqiqil, 2021). Metode *K-fold* sendiri merupakan salah satu dari beberapa metode dalam *Cross-Validation*, di mana metode ini *dataset* dibagi menjadi k kelompok, dan setiap kelompok terdiri dari data *training* dan data *testing* dengan jumlah yang sama (Daqiqil, 2021).

2.3.6. Confusion Matrix

Matriks kebingungan, juga dikenal sebagai *confusion matrix*, digunakan untuk menggambarkan hasil perbandingan klasifikasi yang dilakukan oleh suatu model (Mustika et al., 2021). Biasanya, matriks kebingungan divisualisasikan dalam bentuk tabel yang menunjukkan kinerja model klasifikasi pada rangkaian data pengujian dengan nilai sebenarnya yang diketahui. Gambar 2.1 menunjukkan contoh tabel tersebut untuk mempermudah pemahaman.

		Actual Values	
		Positive (1)	Negative (0)
Predicted Values	Positive (1)	TP (True Positive)	FP (False Positive) <i>Type I Error</i>
	Negative (0)	FN (False Negative) <i>Type II Error</i>	TN (True Negative)

Gambar 2. 1 Confusion Matrix
Sumber: (Mustika et al., 2021)

TP = jumlah data *point* berlabel *yes* dengan nilai yang diidentifikasi benar.

TN = jumlah data *point* berlabel *no* dengan nilai yang diidentifikasi salah.

FP = jumlah data *point* berlabel *yes* dengan nilai sebenarnya yang diidentifikasi salah.

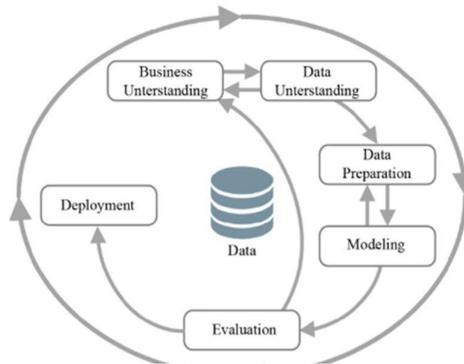
FN = jumlah data *point* berlabel *no* dengan nilai sebenarnya yang diidentifikasi benar.

Nilai akurasi pada model dapat dihitung dengan menggunakan rumus di bawah (Pratiwi et al., 2021):

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \times 100\% \quad (2.2)$$

2.3.7. CRISP-DM

CRISP-DM, singkatan dari *Cross-Industry Standard Process for Data Mining*, adalah suatu model atau rangkaian langkah-langkah yang digunakan untuk mengelola dan menyempurnakan proses data mining (Huber et al., 2019).



Gambar 2. 2 Alur CRISP-DM

Sumber: (Huber et al., 2019)

Adapun poin-poin pada *Gambar 2.1* dapat dijelaskan sebagai berikut (Huber et al., 2019):

1. *Business Understanding*

Langkah awal dalam CRISP-DM adalah menentukan tujuan bisnis dengan melakukan evaluasi terhadap kondisi saat ini dan menetapkan tujuan yang ingin dicapai melalui penerapan data mining.

2. *Data Understanding*

Persiapan, evaluasi persyaratan data, dan pengumpulan data menjadi tahapan kedua. Data yang telah terkumpul dideskripsikan untuk diketahui mana yang menjadi atribut, kelas dan jenis tipe datanya.

3. *Data Preparation*

Setelah melalui tahap satu dan dua, data-data tersebut perlu diidentifikasi, dipilih, dibersihkan, dan diubah ke dalam format yang diinginkan. Hal ini disebut data preparation atau data preprocessing.

4. *Modelling*

Pengaplikasian algoritma untuk mencari, mengidentifikasi, dan menemukan pola. Pemilihan algoritma disesuaikan dari jenis tipe data, karena dari tipe data tersebut dapat diketahui apakah data tersebut akan diestimasi, diprediksi, diklasifikasi, diklaster atau ingin dilakukan penglihatan hubungan asosiatif.

5. *Evaluation*

Melakukan evaluasi pada model yang telah digunakan. Pada model algoritma klasifikasi, evaluasi yang banyak digunakan adalah akurasi, *G-Mean*, *F-Measure*, dan sebagainya.

6. *Deployment*

Setelah pemilihan model algoritma dan evaluasi dilakukan, tahap terakhir adalah *deployment*. Pada tahap ini, tujuannya adalah mengotomatisasi model atau mengembangkan aplikasi yang terintegrasi dengan sistem informasi manajemen atau operasional yang ada.

2.3.8. *Correlation Pearson*

Analisis korelasi digunakan untuk mengukur kekuatan hubungan antara dua variabel, di mana variabel lainnya dianggap dikendalikan atau tetap (sebagai variabel kontrol), tujuan dari analisis korelasi adalah untuk mengetahui sejauh mana hubungan tersebut berkorelasi satu sama lain (Romadloni & Hilman F Pardede, 2019). *Koefisien Pearson* memiliki jangkauan dari angka -1 hingga +1, korelasi 1 menunjukkan setiap kenaikan positif dalam sebuah variabel maka terjadi peningkatan positif dalam variabel lain, -1 berarti jika ada peningkatan positif dalam satu variabel maka terjadi penurunan negatif pada variabel lainnya, dan nol untuk setiap variabel yang tidak berhubungan (Glen, 2023). Rumus *Correlation Pearson* dapat ditulis sebagai berikut (Glen, 2023):

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[n \sum x^2 - (\sum x)^2] [n \sum y^2 - (\sum y)^2]}} \quad (2.3)$$

- n = banyak data (*record*)
- x = sum dari nilai atribut x
- y = sum dari nilai atribut y
- xy = sum dari nilai nilai atribut x dan y
- x^2 = sum dari nilai atribut x^2
- y^2 = sum dari nilai atribut y^2

Menurut Jonathan Sarwono, berikut adalah tabel interval kekuatan hubungan korelasi dari *Correlation Pearson* (Safitri, W, 2014):

Tabel 2. 1 Tabel Nilai Korelasi

No.	Nilai r	Interpretasi
1	0,00	Tidak ada hubungan
2	0,01 – 0,09	Hubungan kurang berarti
3	0,10 – 0,29	Hubungan moderat
4	0,30 – 0,49	Hubungan kuat
5	0,50 – 0,69	Hubungan sangat kuat
6	0,70 – 0,89	Hubungan mendekati sempurna
7	>0,90	

Sumber: (Safitri, W, 2014)

2.4. Penelitian Terdahulu

Tabel 2. 2 Penelitian Terdahulu

No.	Nama Peneliti	Judul Penelitian	Hasil Penelitian
1	(manullang & sianturi, 2021)	Penerapan Algoritma K-Nearest Neighbor Untuk Memprediksi Kelulusan Mahasiswa	Dari hasil penelitian yang dilakukan dapat disimpulkan bahwa metode KNN dapat digunakan untuk memprediksi kelulusan mahasiswa di STMIK Pelita Nusantara
2	(Imaning Dyah Larasati, Ahmad Afif Supianto, 2019)	Prediksi Kelulusan Mahasiswa Berdasarkan Kinerja Akademik Menggunakan Metode Modified K-Nearest Neighbor (MK-NN)	Untuk setiap pengujian yang dilakukan, akurasi dari MK-NN berada di atas akurasi K-NN. Sehingga dapat disimpulkan bahwa MK-NN menghasilkan akurasi lebih baik dibanding K-NN
3	(Rahman, 2023)	Optimalisasi dalam Mengidentifikasi Mahasiswa Jalur Cepat (<i>Fast-track</i>) menggunakan Metode K-Nearest Neighbor	Hasil yang diuji untuk memperoleh ketepatan metode K-Nearest Neighbor pada saat penyeleksian penerimaan mahasiswa fast-track Fakultas MIPA di Universitas Andalas berdasarkan IPK semester 1

No.	Nama Peneliti	Judul Penelitian	Hasil Penelitian
			sampai dengan 6 dan jumlah sks sampai dengan semester 6 menggunakan k = 3 adalah 95,79%
4	(Gumanti et al., 2022)	Penerapan Algoritma K-Nearest Neighbor Untuk Klasifikasi Topik Skripsi Mahasiswa di Fakultas Ilmu Komputer	Optimasi nilai k menggunakan k-fold cross validation menghasilkan tingkat akurasi yaitu 56,67% dengan nilai k-fold cross validation = 2 dan nilai K=5
5	(Wiyono & Abidin, 2018)	<i>Implementation of K-Nearest Neighbour (Knn) Algorithm to Predict Student'S Performance</i>	Nilai K terbaik adalah 3, 6, dan 9 untuk mendapatkan prediksi terbaik. Hasil ini diperoleh dengan mencoba nilai K adalah 3 sampai 60. Nilai prediksi tersebut kemudian dibandingkan, hasil prediksi salah yang persentase terkecil adalah yang terbaik
6	(Nugraheni et al., 2022)	Perbandingan Performa Antara Algoritma Naive Bayes Dan K-Nearest Neighbour Pada Klasifikasi Kanker Payudara	Pengujian model K-NN menghasilkan akurasi, presisi sehat, recall sakit dan AUC lebih tinggi dibandingkan Naive Bayes. Sedangkan untuk hasil presisi sakit dan recall sehat Naive Bayes. Sehingga dapat ditarik kesimpulan berdasarkan penelitian bahwa K-NN lebih baik dibandingkan Naive Bayes
7	(Rahmatullah, 2019)	Prediksi Tingkat Kelulusan Tepat Waktu dengan Metode Naive Bayes dan k-Nearest Neighbor	hasil akurasi tertinggi diraih oleh algoritma kNN yaitu 98.7%
8	(Taslim et al., 2021)	Optimasi Nilai k pada Algoritma k-Nearest	Optimasi nilai menggunakan 5-fold cross-

No.	Nama Peneliti	Judul Penelitian	Hasil Penelitian
		<i>Neighbor</i> untuk Prediksi Akademik Mahasiswa yang Bekerja	<i>validation</i> menghasilkan tingkat akurasi sebesar 0,55 dan nilai $k=3$. Kemudian, hasil uji <i>performance</i> dengan <i>confusion matrix</i> menunjukkan nilai sebesar 85,71%
9	(Nikmatun & Waspada, 2019)	Implementasi Data Mining untuk Klasifikasi Masa Studi Mahasiswa menggunakan Algoritma <i>k-Nearest Neighbor</i>	Hasil akurasi 75,95%

Perbedaan semua penelitian pada tabel di atas adalah pada penelitian ini melakukan optimasi atau seleksi fitur pada algoritma KNN untuk meningkatkan akurasi dari algoritma tersebut yang memiliki masalah dalam pemilihan nilai K . Kemudian melakukan perbandingan nilai akurasi sebelum dan sesudah menerapkan seleksi fitur menggunakan *Correlation Pearson*. Lalu, indikator yang digunakan pada penelitian ini adalah data riwayat perkuliahan yang didapat dari MKDU UMKT.