BAB 2

TINJAUAN PUSTAKA

2.1 Kinerja Mahasiswa

Menurut Naomi & Nindyati (2008) dalam buku Indra dkk (2021) Kinerja akademik merupakan hasil akhir yang dicapai oleh peserta didik sebagai keberhasilan selama mengikuti pendidikan dalam sebuah institusi pendidikan.

2.2 *Data Mining*

Menurut Pramudiono (2006) dalam buku (Nofriansyah & Nurcahyo, 2015) *data mining* adalah analisis otomatis dari data yang berjumlah besar atau kompleks dengan tujuan untuk menemukan pola atau kecenderungan yang penting yang biasanya tidak disadari keberadaannya.

Menurut Santoso & Umam (2018) *Data mining* adalah aktivitas mengekstraksi informasi atau pengetahuan penting dari kumpulan data besar menggunakan teknik tertentu. Informasi atau pengetahuan yang dihasilkan oleh *data mining* dapat digunakan untuk meningkatkan proses pengambilan keputusan.

Adapun jenis-jenis algoritma *data mining* sebagai berikut (Santoso & Umam, 2018) :

1. Klastering

   Mengelompokkan objek ke dalam beberapa kelompok berdasarkan kemiripan antar objek, dimana dalam satu klaster harus berisi objek yang saling mirip dan antar klaster objek salin tidak mirip. Klastering ini tidak memerlukan data pelatihan yang sudah diberi label.

2. Klasifikasi

   Melakukan pengelompokan objek berdasarkan kelompok yang sudah ada. Berbeda dengan klastering, klasifikasi ini memerlukan data pelatihan yang sudah diberi label kelompok atau kelas.

3. Regresi / Estimasi

   Regresi pada dasarnya mirip dengan klasifikasi, yakni memerlukan data

pelatihan yang sudah diberi label. Bedanya, output klasifikasi adalah nilai diskrit, sedangkan output dari regresi adalah nilai kontinyu.

4. Asosiasi

   Melakukan asosiasi antar objek dalam suatu set data, biasanya data transaksional.

2.3 Metode Klasifikasi

Klasifikasi adalah suatu teknik dengan melihat pada kelakuan dan atribut dari kelompok yang telah didefinisikan. Teknik ini dapat memberikan klasifikasi pada data baru dengan memanipulasi data yang ada yang telah diklasifikasi dengan menggunakan hasilnya untuk memberikan sejumlah aturan. Salah satu contoh yang mudah dan popular dengan *decision tree* yaitu salah satu metode klasifikasi yang paling popular karena mudah untuk interpretasi seperti algoritma C4.5, ID3 (Nofriansyah & Nurcahyo, 2015).

2.4 Algoritma C4.5

Menurut Berry & Linoff (2018) dalam buku (Nofriansyah & Nurcahyo, 2015) Algoritma C4.5 merupakan salah satu solusi pemecahan kasus yang sering digunakan dalam pemecahan masalah pada teknik klasifikasi. Keluaran dari algoritma C4.5 berupa sebuah *decision tree* layaknya teknik klasifikasi lain.

Menurut Santoso (2007) dalam buku (Nofriansyah & Nurcahyo, 2015) algoritma C4.5 yaitu salah satu algoritma C4.5 induksi pohon keputusan yaitu ID3 (*Iterative Dichotomiser* 3). Input berupa sampel *training*, label *training* dan atribut. Algoritma C4.5 merupakan pengembangan dari ID3. Jika suatu set data mempunyai beberapa pengamatan dengan *missing value* dapat diganti dengan nilai rata-rata dari variabel yang bersangkutan.

Menurut Nofriansyah & Nurcahyo (2015) Untuk penyelesaian kasus di dalam algoritma C4.5 terdapat 2 elemen beberapa elemen yaitu *entropy* dan *gain*. *Entropy*(S) merupakan jumlah bit yang diperkirakan dibutuhkan untuk dapat mengekstrak suatu kelas (+ atau -) dari sejumlah data acak pada ruang sampel S. *Entropy* dapat dikatakan sebagai kebutuhan bit untuk menyatakan

suatu kelas untuk digunakan dalam mengekstrak suatu kelas. *Entropy* digunakan untuk mengukur ketidakaslian S . Adapun rumus untuk mencari nilai *entropy*.

$$Entropy(S) \equiv \sum_{i=1}^{n} -pi * log_2\, p \quad (2.1)$$

Keterangan :

S : himpunan kasus

A : atribut

n : jumlah partisi S

p_i : proporsi dari S_i terhadap S

Gain (S,A) merupakan perolehan informasi dari atribut A relatif terhadap output data S. Perolehan informasi didapat dari *output* data atau variabel dependen S yang dikelompokkan berdasarkan atribut, dinotasikan dengan *gain* (S,A). Adapun rumus untuk mencari nilai *gain* yaitu :

$$Gain(S, A) \equiv Entropy(S) - \sum_{i=1}^{n} \frac{|Si|}{|S|} * Entropy(Si) \quad (2.2)$$

Dimana :

A : atribut

S : sampel

n : jumlah partisi himpunan atribut a

|Si| : jumlah sampel pada partisi ke-i

|S| : jumlah sampel dalam S

2.5 *Data Preprocessing*

Data preprocessing adalah tahapan awal dari *data mining* untuk menghasilkan analisis yang lebih akurat dalam pemakaian teknik-teknik *machine learning* (Santoso & Umam, 2018).

Langkah-langkah penerapan data preprocessing (Nofriansyah & Nurcahyo, 2015) :

1. *Data Selection*

   Merupakan data kasus dalam proses operasional *data mining*. Pada data yang ada, kolom yang diambil adalah hasil yang disebut dengan atribut keputusan, sedangkan kolom yang diambil dalam pembentukan pohon keputusan adalah atribut penentuan.

2. *Data Cleaning*

   Data cleaning diterapkan untuk menambahkan konten atribut yang hilang atau kosong dan merubah data yang tidak konsisten.

3. *Data Transformation*

   Pada proses ini, data ditransferkan ke dalam bentuk yang sesuai untuk proses *data mining*.

4. *Data Reduction*

   Proses reduksi data dilakukan dengan menghilangkan atribut yang tidak diperlukan sehingga ukuran dari *database* menjadi kecil dan hanya menyertakan atribut yang diperlukan pada proses *data mining*.

2.6 *Confusion Matrix*

Confusion matrix adalah matrik yang berukuran N x N dimana N adalah jumlah kelas yang diprediksi. Jadi matrik ini cocok digunakan untuk permasalahan klasifikasi. *Confusion matrix* menyajikan ringkasan semua hasil prediksi yang dihasilkan dengan membandingkan antara hasil prediksi dan hasil yang diharapkan (Daqiqil Id, 2021).

Tabel 2. 1 *Confusion Matrix*

	Positif	Negatif
Positif	TP	FP
Negatif	FN	TN

Pada *confusion matrix* tersebut ada 4 kolom dengan beberapa istilah yaitu :

1. TP (*True* Positif) berisi jumlah data points diberi label *Yes* yang sebenarnya bernilai *Yes*.

2. TN (*True* Negatif) berisi jumlah data points diberi label No yang sebenarnya bernilai No.
3. FP (*False* Positif) berisi jumlah data points diberi label No yang sebenarnya bernilai No.
4. FN (*False* Negatif) berisi jumlah data points diberi label No yang sebenarnya bernilai *Yes*.

Menghitung *confusion matrix* dapat menggunakan rumus sebagai berikut :

1. *Accuracy* (Akurasi)

   Mengukur akurasi dengan menggunakan rumus dengan jumlah prediksi yang benar dibagi dengan total seluruh populasi.

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{2.3}$$

2. *Precision* (ketetapan)

   Precision untuk mengukur jumlah data yang sukses diprediksi sebagai positif, dibandingkan dengan seluruh data yang diprediksi positif, baik yang kenyataannya benar maupun tidak benar.

$$precision = \frac{TP}{TP + FP} \tag{2.4}$$

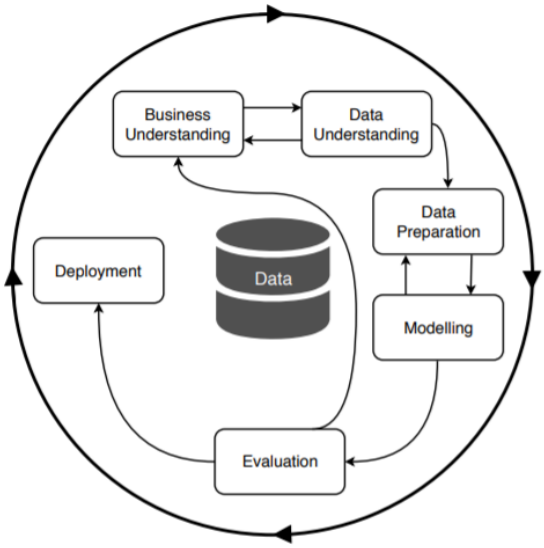
3. *Sensitivity* (*recall*)

   Mengukur banyaknya data yang sukses diprediksi sebagai positif dibandingkan dengan seluruh data yang pada kenyataannya positif.

$$sensitivity = \frac{TP}{TP + FN} \tag{2.5}$$

2.7 CRISP-DM (*Cross Industry Standard Process for Data Mining*)

Data mining menjadi salah satu cabang bidang keilmuan baru dan populer dalam dunia komputer sehingga beberapa perusahaan-perusahaan besar terus mengembangkan dan menyempurnakan metodologi *dalam mining* (Suntoro, 2019)



Gambar 2.1 Tahapan CRISP-DM

Sumber : Martinez-Plumed et al (2019)

1. *Business Understanding*

   Tahap awal dari metodologi CRISP-DM adalah tahapan *business understanding* berisi tentang menentukan tujuan bisnis, menilai situasi saat ini dan menetapkan tujuan dilakukan *data mining*. Tahapan ini sangat penting, namun sering diabaikan ketika seseorang terjun ke dunia *data mining*.

2. *Data Understanding*

   Tahap kedua adalah kegiatan persiapan, mengevaluasi persyaratan data, dan termasuk pengumpulan data. Pada tahapan ini, data yang berhasil dikumpulkan kemudian dideskripsikan bagian mana yang atribut, kelas, dan tipe data.

3. *Data Preparation*

   Tahap ketiga setelah data dikumpulkan, data-data tersebut perlu di identifikasi, dipilih, dibersihkan, kemudian dibangun ke dalam bentuk atau format yang diinginkan. *Data preparation* disebut juga dengan *data pre-processing*.

4. *Modeling*

   Modeling adalah aplikasi dari algoritma untuk mencari, mengidenifikasi,

dan menampilkan pola. Pemilihan algoritma berdasarkan tipe data karena dari tipe data kita bisa mengetahui apakah data tersebut akan diestimasi, prediksi, klasifikasi, *clustering*, atau melihat hubungan asosiatif.

5. *Evaluation*

   Tahap kelima yang digunakan untuk membantu pengukuran evaluasi pada model adalah kita bisa mengukur model mana yang paling baik digunakan untuk proses *data mining*. Pada penerapan klasifikasi, pengukuran evaluasi yang banyak digunakan adalah akurasi, *sensitivity*, *G-Mean*, *F-Measure*, dan lain sebagainya.

6. *Deployment*

   Deployment adalah tahap akhir dalam CRISP-DM. Setelah model dievaluasi dan dipilih algoritma dengan hasil pengukuran terbaik, dilanjutkan ke tahapan *deployment*. Tahapan *deployment* digunakan untuk melakukan otomatisasi model atau pengembangan aplikasi, terintegrasi dengan sistem informasi manajemen atau operasional yang ada.

2.8 Penelitian Terdahulu

Tabel 2. 2 Penelitian Terdahulu

No	Nama Peneliti	Judul Penelitian	Hasil Penelitian
1	Fitriani dkk (2021)	Prediksi hasil belajar siswa secara daring pada masa pandemi covid-19 menggunakan metode C4.5	Penelitian ini menghasilkan klasifikasi metode C4.5 terhadap nilai hasil belajar siswa menghasilkan rule-rule dan klowledge yang dapat digunakan untuk memprediksi hasil belajar secara daring mendapatkan akurasi sebesar 83,33%
2	Irnanda dkk (2021)	Faktor penyebab menurunnya prestasi belajar mahasiswa pada masa pandemi dengan menggunakan klasifikasi C4.5	Hasil penelitian dilakukan dengan menggunakan bantuan *software rapidminer* dan diperoleh akurasi 97,5%

Tabel 2. 3 Penelitian Terdahulu (Lanjutan)

No	Nama Peneliti	Judul Penelitian	Hasil Penelitian
3	Satria dkk (2020)	Prediksi ketetapan waktu lulus mahasiswa menggunakan algoritma C4.5 pada UIN Raden Intan Lampung	Penelitian ini menggunakan perhitungan F-Measure yang mendapatkan nilai 71% bahwa algoritma C4.5 dinilai baik dalam mengklasifikasi dan melakukan prediksi terhadap mahasiswa yang lulus tepat waktu
4	Gaol dkk (2021)	Prediksi kelulusan mahasiswa STIKOM dengan menggunakan algoritma C4.5	Diperoleh hasil akurasi sebesar 90,00% dengan precision sebesar 91,38% dan recall sebesar 98,15%
5	Budiyantara dkk (2020)	Komparasi algoritma *decision tree*, *naïve bayes* dan *k-nearest neighbor* untuk memprediksi mahasiswa lulus tepat waktu	Hasil evaluasi dan validasi yang telah dilakukan menggunakan *tools rapidminer* diperoleh hasil akurasi dari algoritma C4.5 sebesar 98,04%, metode *naïve bayes* sebesar 96,00% dan akurasi pada metode K-NN sebesar 90,00%. Dari hasil komparasi menunjukkan algoritma C4.5 memiliki tingkat akurasi yang paling tinggi untuk menentukan prediksi mahasiswa lulus tepat waktu.

Berdasarkan uraian penelitian yang telah dilakukan sebelumnya, algoritma C4.5 banyak digunakan pada berbagai studi kasus. Terdapat dua penelitian yang dilakukan pada dunia pendidikan yaitu pada Fitriani dkk (2021) dan Irnanda dkk (2021). Penulis pada penelitian ini akan mengangkat permasalahan tentang prediksi performa mahasiswa dalam perkuliahan *online* di Universitas Muhammadiyah Kalimantan Timur dengan metode yang sama pada dua penelitian tersebut. Namun yang membedakan penelitian ini adalah indikator yang diperoleh dari *platform OpenLearning.* Sehingga penelitian ini memunculkan sebuah penelitian baru yang belum pernah dilakukan sebelumnya.