BAB 2

TINJAUAN PUSTAKA

2.1. Kinerja Mahasiswa

Menurut Naomi & Nindyati (2008) dalam buku Indra et al., (2021) Kinerja akademik merupakan hasil akhir yang dicapai oleh siswa/mahasiswa sebagai keberhasilan selama mengikuti pendidikan dalam sebuah institusi pendidikan (I Made Indra P., S.KM., MPH., QRGP. et al., 2021).

2.2. *Data Mining*

Data Mining adalah suatu proses yang menggunakan teknik statistik, matematika, kecerdasan buatan, dan *machine learning* untuk memperoleh dan mengidentifikasi informasi yang berguna dan pengetahuan yang terkait dari berbagai *database* besar (Turban, dkk. 2005). Berikut merupakan beberapa teknik dari *data mining* antara lain (Osman, 2019):

a. Asosiasi

Asosiasi merupakan salah satu teknik *data mining* yang cukup dikenal dalam menemukan pola berdasarkan hubungan antar variabel dalam transaksi yang sama. Selain itu, asosiasi juga dikenal sebagai teknik relasi karena menggunakan relasi antara item dan menemukan seringnya kemunculan item berbeda yang muncul dengan frekuensi tertinggi dalam dataset.

b. Klasifikasi

Klasifikasi merupakan teknik yang digunakan untuk mengklasifikasikan kumpulan data ke dalam kelompok atau kelas untuk mendapatkan prediksi dan analisis yang akurat dalam kumpulan data yang besar.

c. Klastering

Klastering merupakan salah satu teknik pertama yang digunakan dalam *data mining*. Dalam prosesnya melibatkan satu atau lebih atribut untuk mengidentifikasi antara data yang mirip satu sama lain untuk mengetahui perbedaan dan persamaan antara kumpulan data.

d. Prediksi

Prediksi merupakan topik yang komprehensif, mulai dari memprediksi kegagalan komponen hingga memahami dan memprediksi keuntungan perusahaan. Hal ini digunakan dengan kombinasi teknik *data mining* lainnya, teknik prediksi ini termasuk analisis tren, klasifikasi, pencocokan pola, dan hubungan. Selain itu, prediksi juga dibuat dengan menganalisis peristiwa masa lalu.

2.2.1. Algoritma Klasifikasi

Klasifikasi adalah jenis analisis data yang dapat membantu orang memprediksi label kelas sampel harus diklasifikasikan. Berbagai macam teknik klasifikasi telah diusulkan dalam bidang-bidang seperti pembelajaran mesin, sistem pakar dan statistik (Yu et al., 2008). Menurut (Ardiyansyah et al., 2018), Algoritma Klasifikasi merupakan suatu teknik pembelajaran untuk memprediksi dari beberapa atribut yang menggambarkan dan membedakan kelas data atau konsep dengan tujuan memprediksi dari objek kelasnya yang tidak diketahui.

2.2.2. *Random Forest*

Random Forest adalah suatu teknik yang dikembangkan dari algoritma *Classification and Regression Tree* (CART), yang pada prosesnya menggunakan pendekatan *bootstrap aggregating* (*bagging*) dan *random feature selection* (Breiman, 2001). *Random Forest* adalah salah satu algoritma *machine learning* yang merupakan pengembangan dari algoritma *Decision Tree*, *RF* bisa dilihat sebagai gabungan beberapa buah *decision tree* (Primartha, 2021). Rumus dari *RF* yang terdiri dari *N trees* dinyatakan sebagai berikut (Liparas et al., 2014):

$$l(y) = argmax_c(\sum_{n=1}^{N} I_{hn}(y) = c) \quad \textbf{(2.1)}$$

Dimana variabel I adalah fungsi indikator dan *hn* merupakan *tree* ke-n dari *RF*.

Metode *CART* yang digunakan untuk membangun pohon pada algoritma *Random Forest Classifier* menggunakan aturan *Gini Impurity* untuk menentukan pecahan dari pohon keputusan (Daniya et al, 2020). Perhitungan dimulai dengan

penentuan nilai *Gini Index* untuk menentukan distribusi probabilitas atribut terhadap kelas target dan dilanjutkan pada perhitungan *Gini Impurity*. Berikut adalah rumus perhitungan ***Gini*** (Daniya et al, 2020):

$$Gini = \sum_{i=1}^{n} P_{i^2} \quad \textbf{(2.2)}$$

Dimana:

n = Merupakan jumlah kelas target

I = Merupakan kelas target

p = Merupakan rasio kelas target

Berikut ini merupakan rumus perhitungan *Gini Impurity*:

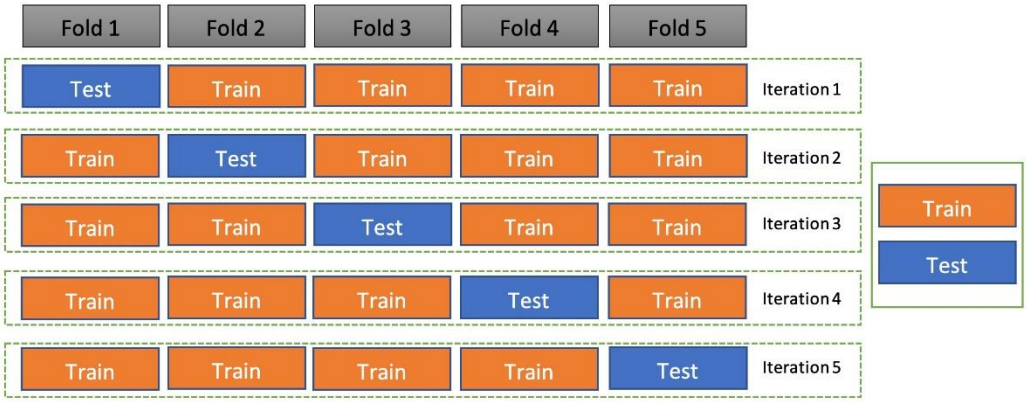
$$Gini\ impurity = 1 - \sum_{i=1}^{n} P_{i^2} \quad \textbf{(2.3)}$$

2.2.3. *Data Preprocessing*

Data Preprocessing merupakan proses mengubah data mentah menjadi data yang lebih mudah dipahami serta mudah diproses. Hal ini diperlukan untuk memperbaiki kesalahan yang ada pada data mentah seperti data tidak lengkap, kosong, tidak sesuai, dan formatnya salah. Tujuan dari *data preprocessing* untuk membuat kualitas data menjadi lebih baik, termasuk kelengkapan, konsistensi, ketepatan waktu, dan meningkatkan hasil dari akurasi (rifqi dharma, 2022). Tahapan-tahapan dari *data preprocessing* yaitu, *data selection*, *data cleaning*, *data transformation*, dan *data reduction* (rifqi dharma, 2022):

2.2.4. *K-Fold Cross Validation*

Cross Validation merupakan metode yang dapat digunakan untuk mengevaluasi kinerja dari suatu model atau algoritma yang dimana memisahkan data menjadi 2 subset yaitu data latih dan data uji (Daqiqil, 2021). Subset pembelajaran melatih model atau algoritma dan subset validasi memvalidasinya lalu pemilihan jenis CV berdasarkan ukuran dari datasetnya. *K-fold* adalah sebuah metode yang memecah dataset menjadi dua bagian yaitu data latih dan data uji sebanyak *k* kelompok yang dimana jumlah data latih dan data uji pada tiap kelompok sama.



Gambar 2. 1 K-Fold Cross Validation

Sumber: (Gopal Krishna Ranjan, 2021)

2.2.5. *Confusion Matrix*

Confusion Matrix merupakan matrik yang membandingkan hasil aktual dengan hasil prediksi, terutama digunakan dalam *supervised learning* untuk mengevaluasi akurasi prediksi dari suatu klasifikasi (Pu et al., 2021). Dalam *Confusion Matrix* terdapat istilah TP, TN, FP, dan FN yang antara lain (Gde Agung Brahmana Suryanegara et al., 2021):

Tabel 2. 1 Confusion Matrix

		Actual Value	
		TRUE	FALSE
Prediction value	TRUE	True Positive (TP)	False Positive (FP)
	FALSE	False Negative (FN)	True Negative (TN)

TP = *True Positive* artinya data positif yang diprediksi secara positif

FP = *False Positive* artinya data negatif yang diprediksi secara positif

FN = *False Negative* artinya data positif yang diprediksi secara negatif

TN = *True Negative* artinya data negatif yang diprediksi secara negatif

Dari tabel diatas, *confusion matrix* dapat diukur dan dievaluasi tingkat performa model klasifikasinya dengan menghitung akurasi. Akurasi adalah persentase dari data uji yang diklasifikasikan dengan benar oleh model klasifikasi yang dibangun. Akurasi ini dapat dihitung dengan rumus berikut (Gde Agung Brahmana Suryanegara et al., 2021):

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} X\ 100\%$$

(2.4)

2.2.6. CRISP-DM

Cross-Industry Standard Process for Data Mining atau yang disingkat dengan CRISP-DM merupakan suatu model atau tahapan dalam mengelola dan menyempurnakan *data mining* (Huber et al., 2019). Ada beberapa tahapan yang ada di dalam CRISP-DM diantaranya:

Gambar 2. 2 CRISP-DM

Sumber: (Hotz, 2018)

1. *Business Understanding*

   Tahap awal dari CRISP-DM adalah menentukan tujuan bisnis lalu menilai situasi saat ini dan setelahnya menetapkan tujuan dilakukannya *data mining*.

2. *Data Understanding*

   Tahap ini merupakan persiapan, mengevaluasi persyaratan data, dan mengumpulkan data. Data yang dikumpulkan kemudian dideskripsikan untuk mengetahui mana yang menjadi atribut, kelas, dan jenis tipe datanya.

3. *Data Preparation*

Selanjutnya pada tahap ini data-data perlu diidentifikasi, dipilih, dibersihkan dan diubah ke dalam format yang diinginkan. Pada tahapan ini juga bisa disebut sebagai data *preprocessing.*

4. *Modeling*

Pada tahap ini dilakukan penerapan algoritma dalam mencari, mengidentifikasi, dan menemukan pola. Pemilihan algoritma yang digunakan menyesuaikan dengan jenis tipe data karena dari tipe datanya, dapat diketahui apakah data tersebut akan diprediksi, diklasifikasi, diklaster atau dengan melihat hubungan asosiasi.

5. *Evaluation*

Di tahap ini evaluasi dilakukan pada model atau algoritma yang dilakukan sebelumnya, Pada algoritma klasifikasi, evaluasi yang sering digunakan adalah akurasi, *sensitivity*, *F-Measure*, dan lain-lain.

6. *Deployment*

Tahap *Deployment* merupakan tahap akhir dalam CRISP-DM, yang bertujuan untuk melakukan otomatisasi model atau pengembangan aplikasi yang terintegrasi dengan sistem informasi manajemen atau operasional yang ada.

2.2.7. *Analysis of Variance (ANOVA)*

Analysis of Variance merupakan teknik standar untuk mengukur signifikansi statistik dari suatu set variabel independen dalam memprediksi variabel dependen (Wenda, 2022). *ANOVA* digunakan untuk memangkas fitur tanpa memengaruhi keakuratan prediktor (Chen et al., 2022). Dalam penyelesaian menggunakan *ANOVA*, berikut langkah-langkahnya:

a. Semua fitur dipilih dari dataset.

b. Fungsi fitur target *dari scikit-learn* dihitung menggunakan *ANOVA F-Score* untuk setiap fitur. Dibawah ini merupakan rumus untuk menghitung *ANOVA*.

$$F = \frac{variance\ between\ groups}{variance\ within\ groups}$$

$$Variance\ between\ groups = \frac{\sum_{i}^{n} n_i(\overline{Y}_{i-} \overline{Y})^2}{(k-1)}$$

$$Variance\ within\ groups = \frac{\sum_{i=1}^{k} \sum_{j=1}^{ni} (Y_{ij-} \overline{Y}_i)^2}{(n-k)}$$

(2.5)

c. Hasil dari pengujian digunakan untuk melakukan pemilihan fitur yang memungkinkan pembuangan fitur yang tidak terkait dengan variabel target. Fitur yang memiliki pengaruh paling tinggi dengan varian terendah dipilih dalam eksperimen ini dan diuji dengan *SelectKBest(); K* mewakili jumlah fitur yang ada untuk dataset akhir.

d. Jumlah fitur dengan peringkat tertinggi digunakan untuk membuat berbagai subset fitur.

2.3. Penelitian Terdahulu

Tabel 2. 2 Penelitian Terdahulu

No.	Nama Peneliti	Judul Penelitian	Hasil Penelitian
1	(Hendrawan et al., 2021)	Implementasi Pemilihan Fitur Metode *Wrapper* dan *Embedded* dalam Prediksi Ketepatan Kelulusan Mahasiswa	Hasil dari pengujian menggunakan algoritma *Random Forest* tanpa pemilihan fitur memiliki nilai akurasi sebesar 73% dengan nilai *AUC* 0.5. Sedangkan hasil pengujian menggunakan *algoritma Random Forest* dengan pemilihan fitur metode *wrapper* dan metode *embedded* memiliki nilai akurasi sebesar 100% dengan nilai *AUC* 1.
2	(Oon Wira Yuda et al., 2022)	Penerapan *Data Mining* Untuk Klasifikasi Kelulusan Mahasiswa Tepat Waktu Menggunakan Metode *Random Forest*	Berdasarkan dari hasil pengujian menggunakan algoritma *Random Forest*, akurasi yang diperoleh sebesar 0.98 atau 98%. Dengan diperolehnya tingkat akurasi tersebut, dapat disimpulkan bahwa pengujian dengan algoritma *Random Forest* memiliki hasil yang akurat

			dalam klasifikasi kelulusan mahasiswa.
3	(Rachmatika & Bisri, 2020)	Perbandingan Model Klasifikasi untuk Evaluasi Kinerja Akademik Mahasiswa	Dalam hasil uji Friedman, algoritma *Random Forest* memiliki peringkat tertinggi dengan nilai rata-rata 8,38. Berdasarkan hasil tersebut, dapat disimpulkan bahwa perbandingan pada Sembilan algoritma klasifikasi yang di evaluasi pada empat dataset, algoritma *Random Forest* lebih unggul dan dapat diandalkan dalam mengevaluasi kinerja akademik mahasiswa
4	(Linawati et al., 2020)	Prediksi Prestasi Akademik Mahasiswa menggunakan Algoritma *Random Forest* dan C4.5	Algoritma *Decision Tree* C4.5 menghasilkan akurasi sebesar 87.1%, presisi sebesar 85.4%, dan *recall* sebesar 87.1%. Sedangkan *Random Forest* menghasilkan akurasi sebesar 92.4%, presisi sebesar 91.4%, dan *recall* sebesar 92.4%. Dalam hal ini *Random Forest* memiliki akurasi, presisi, dan *recall* yang lebih baik dibandingkan algoritma *Decision Tree* C4.5 yang menjadikannya sebagai solusi untuk memprediksi prestasi akademik mahasiswa.
5	(Hasan et al., 2022)	Perbandingan K-*Nearest Neighbor* dan *Random Forest* dengan Seleksi Fitur *Information Gain* untuk Klasifikasi Lama Studi Mahasiswa	Hasil menunjukkan bahwa 15 atribut dan salah satu diantaranya merupakan atribut target, didapatkan 4 atribut terbaik berdasarkan seleksi fitur *Information Gain*. Algoritma *Random Forest* memiliki akurasi sebesar 100% dibandingkan algoritma K-NN yang memiliki akurasi sebesar 86.67% yang menjadikan algoritma *Random Forest* bekerja lebih baik dalam mengklasifikasi lama studi mahasiswa.

6	(Alhassan et al., 2020)	*Predict Students' Academic Performance based on their Assessment Grades and Online Activity Data*	Dalam model prediksi, hasil menunjukkan bahwa model dasar dan sub model bekerja dengan baik dalam memprediksi prestasi akademik siswa. *Random Forest* mengungguli algoritma lain dalam memprediksi kinerja siswa dengan akurasi tertinggi untuk model dasar dan sub model, diikuti oleh *decision tree*.

Berdasarkan dari penelitian-penelitian terdahulu di atas tentang klasifikasi nilai mahasiswa yang telah dilakukan, ada banyak metode atau teknik *data mining* yang digunakan dalam melakukan klasifikasi dan prediksi. Dalam penelitiannya (Rachmatika & Bisri, 2020), (Linawati et al., 2020), dan (Alhassan et al., 2020) memiliki pembahasan yang berkaitan dengan nilai mahasiswa menggunakan algoritma *Random Forest*. Namun, yang membedakan penelitian ini dengan penelitian sebelumnya adalah indikator, dan seleksi fitur yang digunakan dalam klasifikasi nilai mahasiswa. Penelitian ini mengangkat studi kasus pada Universitas Muhammadiyah Kalimantan Timur dengan indikator atribut data yang diperoleh dari MKDU dan BAA UMKT dan penerapan seleksi fitur *Analysis of Variance* (*ANOVA*) pada algoritma *Random Forest Classifier* dalam mengklasifikasi nilai mahasiswa.