BAB 2

TINJAUAN PUSTAKA

2.1 Kinerja Mahasiswa

Mahasiswa yang mengikuti pendidikan di universitas atau institusi pendidikan tinggi memiliki performa akademis. Performa akademis dianggap sebagai salah satu indikator kualitas mahasiswa. Prestasi belajar mahasiswa didasarkan pada aktivitas perkuliahan yang mereka lakukan selama menempuh studi (Khalida Kumalasari, 2018).

Kinerja Mahasiswa menurut Naomi & Nindyati (2008) merupakan hasil akhir yang dicapai oleh perserta didik sebagai keberhasilan selama mengikuti pendidikan dalam sebuah institusi pendidikan (I Made Indra P., 2021).

2.2 Data Mining

Data mining adalah proses ekstraksi pengetahuan dari sejumlah besar data, namun seharusnya dinamai "penambangan pengetahuan dari data". Meskipun banyak istilah lain dengan arti serupa, data mining tetap populer karena menggambarkan proses menemukan informasi berharga dari bahan mentah. Beberapa orang juga menganggap data mining sebagai sinonim KDD (Jiawei Han & Micheline Kamber, 2006).

Data Mining merupakan sekumpulan data menjadi informasi yang memiliki potensi secara implisit (tidak nyata/ jelas) yang sebelumnya tidak diketahui, dengan menngunakan pranti otomatis atau semi otomatis, dari sejumlah besar data yang bertujuan untuk menemukan pola yang memiliki arti (Lailil Muflikhah et al., 2018).

Teknologi data mining dapat membantu mencari pengetahuan yang berharga dari sejumlah besar data dengan cara mempersiapkan dan menambang data, serta menganalisis hasilnya. Teknologi ini memanfaatkan teknologi basis data yang telah matang, yakni ilmu perangkat lunak yang mempelajari cara mengelola dan menerapkan basis data (Yang et al., 2020).

Data mining dapat digunakan oleh sebuah institusi untuk mengambil keputusan yang akurat dan juga untuk memprediksi hasil siswa (Rulin Swastika et al., 2023).

2.3 Klasifikasi

Klasifikasi merupakan adanya variabel target yang bertipe kategori. Model data mining mengujikan sejumlah record, dan setiap record nya berisikan variabel target dan sekumpulan variabel input atau pemrediksi. dalam data mining menggunakan metode prediktif untuk menentukan model atribut kelas dari sekumpulan record (training set) yang terdiri dari atribut lain dan satu atribut kelas. Tujuannya adalah untuk mengklasifikasikan rekord yang belum dilihat sebelumnya dengan seakurat mungkin. Data latih digunakan untuk membentuk model, dan data uji digunakan untuk menguji keakuratannya (Lailil Muflikhah et al., 2018).

Klasifikasi merupakan pengelompokan berdasarkan hubungan antara variabel target. Contohnya pengelompokn dampak gempa bumi yaitu rusak berat, rusak berat dan tsunami, atau tidak terdampak (Efori Buulolo, 2020).

2.4 Python

Python adalah bahasa pemerograman tingkat tinggi tujuan umum yang banyak di gunakan . ini dibuat oleh Guido van Rossum pada tahun 1991 dan di kembangkan lebih lanjut oleh oleh python Sofware Foundation. Itu dirancang dengan penekanan pada keterbacaan kode, dan sintaksnya memungkinkan programmer untuk mengekspresikan konsep mereka dalam baris kode yang lebih sedikit. Python adalah bahasa pemrograman yang memungkinkan Anda bekerja dengan cepat dan mengintegrasikan sistem dengan kebih efesien.(Yogi Aditya Saputra & Syafrial Fachri Pane, 2020)

2.5 Algoritma Decision Tree C 4.5

Decision Tree atau pohon keputusan adalah metode atau algoritma klasifikasi data mining dengan membentuk pola pohon keputusan yang digunakan untuk mendapatkan jawaban dari masalah yang dimasukan (Anief Rufiyanto et al., 2021).

Algoritma C4.5 menurut Rokach & Maimon (2012) dalam buku (Anjar Wanto et al., 2020) Data Mining Algoritma dan Imlementasi merupakan metode untuk membuat pohon keputusan berdasarkan data training yang diberikan. Algoritma C4.5 merupakan pengembangan dari ID3. Di antara proyek pengembangan yang dilakukan di Bagian C4.5 adalah, misalnya, menyilangkan nilai yang hilang, menyilangkan data kontinu, dan memangkas.

Dalam Algoritma C4.5, terdapat beberapa elemen yang dikenal untuk menyelesaikan kasus, yaitu Entropi dan Gain. Entropi (S) dapat diartikan sebagai jumlah bit yang dibutuhkan untuk mengekstraksi kelas (+ atau -) dari sejumlah data acak dalam ruang sampel S. Entropi dapat dianggap sebagai ukuran kebutuhan bit untuk menyatakan suatu kelas, sehingga semakin kecil nilai entropi, semakin sedikit bit yang diperlukan untuk mengekstraksi kelas tersebut. Oleh karena itu, nilai entropi yang rendah lebih diutamakan dalam penggunaannya dalam mengekstraksi suatu kelas (Dicky Nofriansyah & Gunadi Widi Nurcahyo, 2015) .

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

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 :

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

Keterangan :

S = Himpunan kasus

A = Atribut

n = Jumlah atribut

|Si| = Jumlah partisi ke -i

|S| = jumlah kasus dalam S

Beberapa peneliti terdahulu pernah melakukan teknik data mining diantaranya. Implementasi Data Mining untuk klasifikasi masa studi mahasiswa menggunakan algoritma Decision Tree dan analisa klasifikasi C4.5 terhadap faktor penyebab menurunnya prestasi belajar mahasiswa di masa pandemi. C3 (pemahaman materi) dipilih sebagai atribut paling berpengaruh dalam menurunnya prestasi belajar mahasiswa. Metode C4.5 memiliki akurasi 97.5% dalam pengujian menggunakan software Rapidminer (Irnanda et al., 2021). Penerapan Data Mining Klasifikasi Tingkat Pemahaman Siswa Pada Pelajaran Matematika menyimpulkan bahwa Algoritma C4.5 diterapkan untuk menentukan faktor yang berpengaruh pada tingkat pemahaman siswa pada pelajaran matematika. Hasilnya menunjukkan bahwa Motivasi Siswa adalah faktor utama yang mempengaruhi tingkat pemahaman siswa, diikuti oleh Cara Belajar Siswa, Sarana dan Prasarana, Minat Siswa, Cara Mengajar Guru, dan Media Pembelajaran. Hasil penerapan Algoritma C4.5 dapat diuji dengan software Rapidminer dan memperoleh akurasi 96% (Novika et al., 2021). Namun, algoritma C4.5 masih memiliki kelemahan dalam mengatasi data tidak seimbang, misalnya adanya kelas mayoritas dan kelas minoritas. Hal ini menyebabkan akurasi yang kurang optimal. Oleh karena itu, dilakukan pendekatan Split Feature Reduction Model (SFRM) untuk memberikan bobot pada setiap fitur dan menemukan fitur yang kuat. Hasilnya, akurasi klasifikasi C4.5 pada dataset mahasiswa dengan pendekatan SFRM meningkat menjadi 98% dalam mengatasi ketidakseimbangan kelas (Yusup et al., 2020).

2.6 Correlation-Based Feature Selection

Correlation-Based Feature Selection (CFS) adalah metode evaluasi subset fitur yang menggunakan prinsip bahwa fitur yang baik memiliki fitur keluaran yang sangat prediktif, tetapi tidak satu sama lain (Hall & Holmes, 2003).

Pendekatan *Correlation-Based Feature Selection* (CFS) didasarkan pada korelasi dan merupakan metode yang independen dari model klasifikasi akhir.

Metode ini mengevaluasi subset fitur hanya berdasarkan sifat-sifat data intrinsik, seperti korelasi, sesuai dengan namanya (Johannes S. Fischer, 2021).

$$M_S = \frac{k\, r_{cf}}{\sqrt{k + k\,(k-1)\, r_{ff}}} \tag{2.3}$$

Dimana Ms adalah "Merit" heuristik dari suatu subset fitur S yang berisi fitur k, rcf adalah korelasi fitur kelas rata-rata, dan rff adalah rata-rata antar fitur-fitur.

Secara umum, subset fitur yang baik harus terdiri dari fitur yang memiliki korelasi yang tinggi dengan kelas yang akan diprediksi, tetapi tidak saling berkorelasi satu sama lain. Untuk mengukur korelasi antara fitur nominal, digunakan teknik CFS yang akan memperhitungkan korelasi antara fitur-fitur tersebut setelah fitur numerik didiskritisasi. Oleh karena itu, teknik ini dapat digunakan untuk berbagai jenis masalah klasifikasi yang bersifat supervisi, bahkan untuk masalah di mana kelas yang akan diprediksi bersifat numerik.

Metode *Correlation-Based Feature Selection* adalah salah satu metode yang bisa digunakan untuk melakukan optimasi atau seleksi fitur pada algoritma *Decision Tree* Dengan menggunakan metode ini, bisa membantu meningkatkan akurasi dan performa dari algoritma *Decision Tree* dalam melakukan klasifikasi.

Penggunaan *Correlation Based Feature Selection* sudah dilakukan dalam penelitian terdahulu, diantaranya penelitian yang berjudul Klasifikasi Anomali *Intrusion Detection System* (IDS) Menggunakan Algoritma *Naïve Bayes Classifier* dan *Correlation-Based Feature Selection*, studi ini memanfaatkan data UNSW-NB15 untuk deteksi anomali. Empat atribut terpilih dengan menggunakan *Correlation-Based Feature Selection*. Hasilnya, akurasi meningkat dari 71,2% menjadi 74,8% setelah atribut dipilih dengan teknik korelasi (Anwar et al., 2019). Penelitian ini memfokuskan pada seleksi fitur *Decision Tree* C4.5 menggunakan metode *Correlation-Based Feature Selection* untuk klasifikasi kinerja mahasiswa.

2.7 Data Preprocessing

Data Preprocessing merupakan salah satu tahapan dalam melakukan mining data. Sebelum menuju ke tahap pemprosesan. Data mentah akan diolah terlebih dahulu. *Data Preprocessing* atau praproses data biasanya dilakukan melalui cara

eliminasi data yang tidak sesuai. Selain itu dalam proses ini data akan diubah dalam bentuk yang akan lebih dipahami oleh sistem (Suripto et al., 2022).

2.8 Cross Validation

Cross validation (CV) adalah metode statistik yang dapat digunakan untuk mengevaluasi kinerja model atau algoritma dimana data dipisahkan menjadi dua subset yaitu data latih dan data uji. Model atau algoritma dilatih oleh subset pembelajaran dan divalidasi oleh subset validasi. Selanjutnya pemilihan jenis CV dapat didasarkan pada ukuran dataset (Ibnu Daqiqil Id, 2021).

2.9 Confusion Matrix

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

Tabel 2. 1 Confusion Matrix

	Yes	NO
Yes	TP	FP
No	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 :

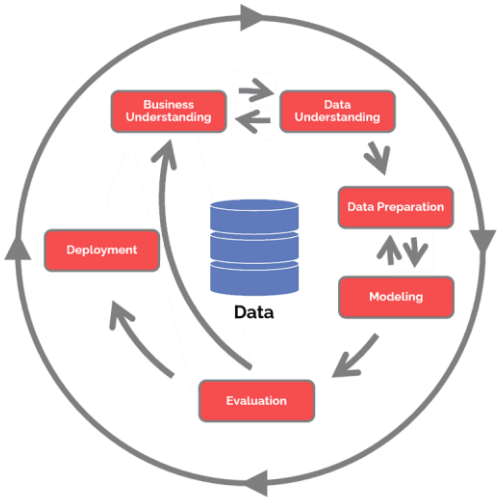
1. Accuracy (Akurasi)

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

$$accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (2.4)$$

2.10 CRISP-DM

Cross-Industry Standard Process for Data Mining atau CRISP-DM adalah model proses standar terbuka yang memberikan pendekatan untuk perencanaan proyek penambangan data (Cornellius Yudha Wijaya, 2021).



Gambar 2. 1 CRISP-DM

a. *Business Understanding* (Pemahaman Bisnis)

   Pada tahapan ini melakukan pentuan arah, tujuan dan strategi awal dalam lingkup bisnis atau penelitian secara keseluruhan.

b. *Data Understanding* (Pemahaman Data)

   Pada tahapan ini melakukan pengumpulan data awal kemudian melakukan identifikasi dan eksplorasi data.

c. *Data Preparation* (Pengolahan Data)

   Pada tahapan ini melakukan proses sleksi data dan pengubahan data, kemudian setelah data digabungkan akan diolah pada fase pengolahan data ini.

d. *Modelling* (Pemodelan)

Pada tahapan ini menerapkan beberapa permodelan degan data yang telah diklasifikasikan. Pemodelan ini dilakukan agar mendapatkan hasil yamg optimal.

e. *Evaluation* (Evaluasi)

Tahapan ini dilakukan untuk mengetahui model yang dirancang telah sesuai atau belum dengan tujuan pada fase awal. Adpun tujuan awal dirancangnya model ini untuk menghasilkan nilai akurasi yang tinggi, dengan begitu dapat membuktikan bahwa penelitian yang telah dilakukan berhasil.

f. *Deployment* (Penyebaran)

Deployment merupakan tahapan terakhir yang menghasilkan laporan penerapan proses data mining dan menentukan langkah apa yang harus diambil selanjutnya.

2.11 Penelitian Terdahulu

Tabel 2. 2 Penelitian Terdahulu

No	Nama Peneliti	Judul Penelitian	Hasil Penelelitian
1	(Khairunnissa Fanny Irnanda, 2021)	Analisa Klasifikasi C4.5 Terhadap Faktor Penyebab Menurunnya Prestasi Belajar Mahasiswa Pada Masa Pandemi	Penelitian ini menggunakan metode datamining C4.5 untuk mengklasifikasikan faktor penyebab menurunnya prestasi belajar mahasiswa pada masa pandemi. Hasil menunjukkan bahwa faktor pemahaman materi paling berpengaruh, dan pengujian software Rapidminer mencapai akurasi 97,5%.
2	(Tri Novika, 2021)	Penerapan Data Mining Klasifikasi Tingkat Pemahaman Siswa Pada Pelajaran Matematika	Pemodelan klasifikasi dengan Algoritma C4.5 pada RapidMiner diperoleh akurasi sebesar 96.00%. Algoritma C4.5 dapat diterapkan dan memberikan informasi baru tentang klasifikasi konsep pemahaman siswa pada pelajaran matematika.

3	(Senna, 2018)	Algoritma Klasifikasi Data Mining Untuk Memprediksi Siswa Dalam Memproleh Bantuan Dana Pendidikan	Hasil yang diperoleh untuk nilai Accuracy Algoritma C4.5 adalah sebesar 98,80%.
4	(Parawystia Prabasini Haryoto, 2021)	Algoritma C4.5 Dalam Data Mining Untuk Menentukan Klasifikasi Penerimaan Calon Mahasiswa Baru	Dari hasil Asal sekolah dan lama kuliah mempengaruhi prestasi mahasiswa. Hasil penelitian menunjukkan akurasi sebesar 81.32%.
5	(Mohammad Yusup, 2020)	Analisis Kinerja dalam Mendeteksi Student Loses Berdasarkan Nilai Gain dengan Splite Feature Reduction Model pada Algoritma C4,5	Hasil menunjukkan bahwa, kinerja akurasi klasifikasi C4.5 pada dataset mahasiswa dengan pendekatan SFRM sebelum proses Pengujian, 10 fold cross-validation, menunjukkan tingkat akurasi klasifikasi yang lebih baik yaitu akurasi 98% dalam penanganan ketidakseimbangan kelas.
6	(Saipul Anwar, 2019)	Klasifikasi Anomali Intrusion Detection System (IDS) Menggunakan Algoritma Naïve Bayes Classifier dan Correlation-Based Feature Selection	Dengan seleksi atribut menggunakan teknik korelasi, akurasi meningkat menjadi 74,8%. Seleksi atribut dengan teknik korelasi meningkatkan akurasi klasifikasi dengan algoritma naïve bayes.

Berdasarkan tabel 2.2 dari penelitian terdahulu, algoritma Decision Tree C4.5 dapat digunakan dalam penelitian di bidang pendidikan. Penelitian yang paling relevan dengan penelitian ini adalah "Klasifikasi *Anomali Intrusion Detection System* (IDS) Mengunakan Algoritma *Naïve Bayes Classifer* dan *Correlation-Based Feature Selection*", yang membedakannya adalah pada algoritma yang digunakan. Penelitian ini menggunakan Metode *Correlation-Based Feature Selection* pada algoritma *Decision Tree C4.5* untuk melakukan seleksi fitur dan mengevaluasi persentase peningkatan akurasi.