BAB II
TINJAUAN PUSTAKA

2.1 Penelitian Terdahulu

Penelitian yang berkaitan dengan data mining sudah banyak dilakukan di perusahaan, instansi pemerintahan dan pendidikan. Tinjauan pustaka bertujuan sebagai referensi dan rujukan terhadap hasil penelitian sebelumnya yang berkaitan dengan penelitian yang akan dilakukan.

Menurut penelitian yang dilakukan oleh [8]yang berjudul "Penerapan Metode Klasifikasi Data Mining Untuk Prediksi Kelulusan Tepat Waktu". Metode yang digunakan dalam penelitian ini adalah algoritma C4.5, Naive Bayes, dan Neural Network untuk memperkirakan kelulusan tepat waktu mahasiswa dengan melihat pengaruh dari IMK dan IPK. Dari penelitian ini menunjukan bahwa algoritma terbaik adalah algoritma yang paling tinggi tingkat accuracy pada model klasifikasi yaitu C4.5 dan Neural Network dengan tingkat accuracy 100% sedangkan Naive Bayes 99.8878%. Hasil data mining dari algoritma terpilih dalam penelitian ini menggunakan C4.5, *interface* dirancang menggunakan java engine yang dapat menampilkan prediksi kelulusan tepat waktu beserta jumlah kelulusan tepat waktu setiap Program Studi.

Menurut penelitian yang dilakukan oleh [9] yang berjudul "Penerapan Algoritma C4.5 Untuk Klasifikasi Predikat Kelulusan Mahasiswa Fakultas Komunikasi Dan Informatika Universitas Muhammadiyah Surakarta". Metode yang digunakan dalam penelitian ini adalah algoritma C4.5 untuk klasifikasi mahasiswa berdasarkan predikat kelulusannya. Berdasarkan hasil penelitian yang dilakukan maka dapat disimpulkan bahwa telah diperoleh klasfikasi predikat kelulusan mahasiswa Fakultas Komunikasi dan Informatika UMS. Variabel yang paling tinggi pengaruhnya terhadap predikat kelulusan adalah partisipasi mahasiswa menjadi asisten. Interprestasi hasil penelitian ini adalah mengindikasikan bahwa

variabel yang perlu digunakan sebagai pertimbangan bagi 7 Fakultas Komunikasi dan Informatika UMS untuk memperoleh tingkat predikat kelulusan yang maksimal adalah peran serta mahasiswa untuk menjadi asisten. Secara umum probabilitas predikat "*Cumlaude*" pada kelompok mahasiswa yang pernah menjadi asisten lebih tinggi dibandingkan dengan yang tidak pernah menjadi asisten jika berasal dari jurusan IPA semasa sekolah menengah atas memiliki probabilitas predikat kelulusan "*Cumlaude*" yang lebih tinggi dibandingkan dengan mahasiswa dari jurusan lainnya

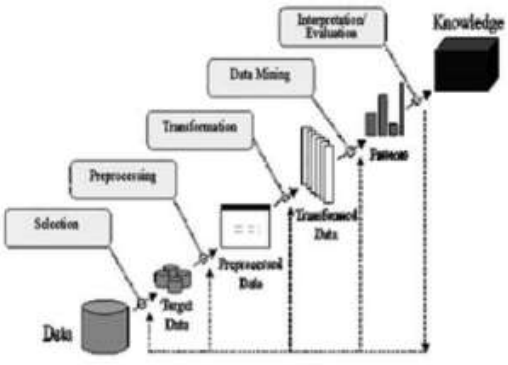
2.2 Data Mining

Data mining merupakan proses penggalian dan pertambangan pengetahuan dari sejumlah data yang besar, database atau *repository database* lainnya. Tujuan utama dari penambangan data ini untuk menemukan pengetahuan baru yang tersembunyi dari database tersebut[10].

Data mining adalah suatu rangkaian dari proses kemudian dapat dipisah-pisah menjadi beberapa tahapan. Tahapan-tahapan yang ada dalam data mining bersifat interaktif terhadap pengguna yang terlibat langsung dengan perantara *knowledge base*. Tahap-tahap dalam data mining antara lain:

1. Pembersihan Data Tahap pembersihan data dilakukan untuk membuang data yang tidak konsisten dan noise. Selain itu, terdapat atribut data yang tidak sesuai dengan hipotesis data mining yang ada. Pembersihan data dapat mempengaruhi kinerja dari sistem data mining karena data yang diolah akan berkurang jumlah dan kompleksitasnya.
2. Integrasi Data Integrasi data digunakan untuk menggabungkan data dari beberapa sumber karena dapat terjadi data yang dibutuhkan dapat berasal dari beberapa database atau file task. Tahap ini dilakukan pada atribut-atribut yang unik seperti nama, jenis produk, dan nomor pelanggan. Untuk menghasilkan data yang tepat dan tidak menyimpang maka harus dilakukan dengan cermat pada tahap ini.

3. Transformasi Data Transformasi data dilakukan dengan mengubah data menjadi bentuk atau format yang sesuai. Sebagai contoh beberapa teknik dasar seperti analisis asosiasi dan klastering hanya dapat menerima input data kategorikal. Karena data yang berupa angka numerik perlu dipecah menjadi beberapa interval. Proses tersebut yang dinamakan binning. Transformasi dan pemilihan data ini menentukan ketepatan hasil dari data mining karena ada beberapa karakteristik dari teknik-teknik yang ada pada data mining tertentu bergantung dengan tahap ini
4. Aplikasi Teknik Data Mining Tahapan aplikasi teknik data mining adalah bagian dari salah satu proses data mining. Sehingga harus diperhatikan bahwa teknik-teknik yang ada tidak selamanya dapat mencukupi untuk melaksanakan data mining tertentu.
5. Evaluasi Pola yang Ditemukan Tahap evaluasi pola yang ditemukan digunakan untuk menemukan pola-pola yang dengan ciri khas maupun prediksi yang bernilai. Apabila hasil yang ada tidak cocok dengan hipotesis yang ada maka terdapat cara lain yang dapat dilakukan.
6. Presentasi Pola yang Ditemukan Selanjutnya tahap presentasi pola yang ditemukan digunakan untuk menghasilkan tindakan atau langkah yang harus dilakukan dari analisis yang diperoleh dengan bentuk pengetahuan yang dapat dipahami semua orang. Dalam presentasi ini visualisasi membantu menampilkan hasil data mining. Gambar 2.1 merupakan proses data mining.



Gambar 2. 1 Proses Data Mining

Dalam data mining terdapat beberapa metode pengolahan. Berikut adalah pengelompokan metode pengolahan data mining antara lain:

a. Classification

Classification adalah suatu teknik dengan melihat atribut dari kelompok data yang telah didefinisikan. Teknik ini dilakukan pada data dengan memanipulasi data yang ada, kemudian diklasifikasi sehingga dapat memperoleh hasil berupa sejumlah aturan. Salah satu contoh yang mudah dan populer adalah decision tree. Decision tree merupakan model prediksi menggunakan struktur pohon atau struktur berhirarki. Perbedaan antara metode clustering dan classification terletak pada data karena metode clustering tidak ada variabel target dalam pengklusteran, sedangkan classification harus ada target variabel kategori.

b. Association

Association sebuah metode yang digunakan untuk mengetahui beberapa kejadian-kejadian khusus atau proses yang muncul pada setiap kejadian yang berhubungan dengan asosiasi. Salah satu contoh adalah Market Basket Analysis, yaitu salah satu metode asosiasi yang digunakan untuk menganalisis kemungkinan para pelanggan untuk membeli sejumlah barang secara bersamaan.

c. Clustering

Clustering sebuah metode yang digunakan untuk menganalisis pengelompokkan pada data yang berbeda, hampir sama dengan klasifikasi tetapi dalam proses pengelompokkannya belum diketahui saat dijalankan pada tool data mining. Metode yang sering digunakan adalah metode statistik atau neural network.

d. Predictive

Modelling Predictive modelling sebuah metode berupa metode pengolahan data mining dengan melakukan dengan cara prediksi atau peramalan. Dan tujuan dari metode ini yaitu untuk membentuk sebuah model prediksi suatu nilai yang mempunyai ciri-ciri khusus.

2.3 9 (*Decision Tree*)

Pohon keputusan adalah sebuah struktur yang digunakan untuk mengubah data menjadi pohon keputusan sehingga akan menghasilkan aturan-aturan keputusan besar menjadi himpunan-himpunan record yang lebih kecil dengan menerapkan serangkaian aturan keputusan. Pohon keputusan yang dihasilkan oleh algoritma C4.5 dapat digunakan untuk klasifikasi [12].

2.4 Software RapidMiner

Software RapidMiner adalah sebuah software yang digunakan untuk mengolah data mining. Biasanya berkaitan tentang analisis teks, mengekstrak polapola dari dataset yang besar dan digabungkan dengan metode statistika, database, dan kecerdasan buatan. Analisis teks ini bertujuan untuk mendapatkan informasi bermutu tinggi dari teks yang diolah [9].

Di dalam software RapidMiner menyediakan prosedur data mining dan machine lerning termasuk ETL (extraction, transformation, loading), data prepocessing, visualisasi, modelling dan evaluasi. Prosesnya tersusun dari operatoroperator, dideskripsikan dengan XML, dan dibuat dengan GUI. Disajikan dengan tulisan bahasa pemrograman Java.

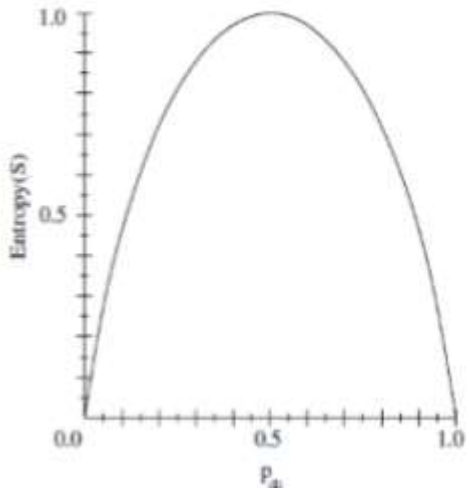
2.5 Algoritma C4.5

Algoritma C4.5 merupakan sebuah algoritma data yang sering digunakan dan diterapkan untuk sebuah proses klasifikasi data dengan atribut numerik dan kategorial. Hasil dari proses klasifikasi dapat berupa aturan-aturan yang digunakan untuk memprediksi nilai atribut bertipe diskret atau tidak saling berhubungan dari record yang baru. Algoritma C4.5 berasal dari pengembangan algoritma ID3, antara lain dapat mengatasi missing data, dapat mengatasi data kontinu, dan pruning [6][11]

2.6 Entropy

Entropy (S) adalah jumlah bit yang dibutuhkan untuk dapat mengekstrak suatu kelas (+ atau -) dari beberapa data acak pada ruang sampel S. Sehingga entropy dapat dikatakan sebagai kebutuhan bit untuk menyatakan suatu kelas. Semakin

kecil nilai entropy maka akan semakin baik entropy yang digunakan dalam mengekstrak suatu kelas. Entropy digunakan untuk mengukur ketidakaslian S. Gambar 2.2 merupakan tampilan dari grafik entropy.



Gambar 2. 2 Tampilan Grafik Entropy

Berikut definisi nilai Entropy: Entropy (S) = $\sum -pi * log2(pi)\ n\ i$=0
.....(1)

Rumus (1) merupakan rumus yang digunakan dalam menghitung entropy untuk menentukan seberapa informatif atribut tersebut. Berikut keterangannya:
S : Himpunan kasus
n : Jumlah partisi
pi : Jumlah kasus pada partisi ke-i

A. Information Gain

Information Gain adalah informasi yang didapatkan dari perubahan entropy di suatu kumpulan data, baik melalui observasi maupun disimpulkan dengan cara melakukan partisipasi terhadap suatu set data.

Gain (S, A) = Entropy (S) - ∑ − |Si| |S| n i=1 ∗ $Entropy$ (S) .....(2)

Rumus (2) merupakan rumus dalam perhitungan information gain setelah menemukan nilai entropy. Berikut keterangannya:

S: Himpunan kasus

n: Jumlah partisi atribut A

|Si|: Jumlah kasus pada partisi ke-i

|S|: Jumlah kasus dalam S

2.7 Split Info

Split Info merupakan rumus yang menyatakan informasi potensial atau entropy. dapat dilihat dalam rumus (3). Dan keterangannya:

Split Info (S, A) = -∑ Si S $log2$ Si S n i=1 .....(3)

S: Himpunan kasus

A: Atribut

Si: Jumlah kasus pada partisi ke- i

2.8 Gain Ratio

Gain Ratio adalah modifikasi dari information gain yang digunakan untuk mengurangi bias atribut yang memiliki banyak cabang. Gain ratio memiliki sifat:

Bernilai besar jika data menyebar rata

➢ Bernilai kecil jika semua data masuk ke dalam satu cabang GainRatio

(S, A) = Gai (S, A) $SpiltInfo$ (S, A) .....(4) dan Keterangannya sebagai berikut:

S: Himpunan kasus

A: Atribut Gain

(S, A): Information gain pada atribut A

SpiltInfo (S, A): SplitInfo pada atribut A

Berikut langkah-langkah dalam membuat pohon keputusan pada algoritma C4.5, yaitu:

- Pertama adalah memilih atribut sebagai akar, dan yang akan dipilih sebagai akar adalah atribut yang memiliki nilai gain ratio tertinggi dari semua atribut yang ada.
- Membuat cabang pada masing-masing nilai, artinya membuat cabang sesuai dengan jumlah nilai variabel gain ratio tertinggi.
- Membagi setiap kasus dalam cabang, berdasarkan perhitungan nilai gain ratio tertinggi dan perhitungan dilakukan setelah perhitungan nilai gain ratio tertinggi awal dan kemudian dilakukan proses perhitungan gain ratio tertinggi kembali tanpa menyertakan nilai variabel gain ratio awal.
- Terakhir adalah mengulangi proses pada setiap cabang sehingga semua kasus pada cabang memiliki kelas yang sama, mengulangi semua proses perhitungan gain ratio tertinggi untuk masing-masing cabang kasus sampai tidak bisa dilakukan proses perhitungan.

AUC (Area Under the Curve) adalah metrik evaluasi yang digunakan untuk mengevaluasi kinerja model klasifikasi. AUC mengukur seberapa baik model mampu membedakan antara dua kelas dengan melihat seberapa besar area di bawah kurva ROC (Receiver Operating Characteristic) yang dihasilkan oleh model. Kurva ROC adalah grafik yang menunjukkan tingkat sensitivitas (True Positive Rate) versus tingkat spesifisitas (False Positive Rate) dari model pada berbagai titik pemotongan (threshold) yang berbeda. AUC memiliki rentang nilai antara 0 hingga 1, dimana semakin besar nilai AUC, semakin baik kinerja model dalam membedakan kedua kelas yang berbeda. AUC adalah salah satu metrik evaluasi yang umum digunakan dalam pembelajaran mesin (machine learning) dan statistik.

Beberapa hal yang perlu diperhatikan terkait AUC antara lain:

1. AUC dapat digunakan untuk membandingkan performa model klasifikasi yang berbeda. Model dengan nilai AUC yang lebih besar dianggap lebih baik dalam membedakan kedua kelas yang berbeda.
2. AUC dapat digunakan pada klasifikasi biner (dua kelas) maupun multi-kelas. Namun, pada klasifikasi multi-kelas, AUC dihitung sebagai rata-rata dari AUC pada setiap kelas.
3. AUC tidak sensitif terhadap distribusi kelas yang tidak seimbang (imbalanced class), sehingga tetap dapat memberikan hasil yang baik meskipun jumlah sampel pada kelas positif dan negatif berbeda jauh.
4. AUC dapat menjadi metrik evaluasi yang berguna dalam pemilihan threshold pada model klasifikasi. Threshold yang digunakan akan mempengaruhi trade-off antara True Positive Rate dan False Positive Rate yang dihasilkan oleh model.
5. Meskipun AUC memiliki kelebihan sebagai metrik evaluasi, namun AUC tidak memberikan informasi detail tentang kinerja model pada setiap titik threshold tertentu. Oleh karena itu, perlu diperhatikan juga metrik evaluasi lain seperti Precision, Recall, F1-score, dll. untuk melihat kinerja model secara lebih detail.
6. Selain itu, AUC juga dapat terpengaruh oleh data yang memiliki noise atau outlier. Oleh karena itu, sebaiknya data dipreproses terlebih dahulu sebelum menghitung AUC pada model klasifikasi.
7. AUC dapat dihitung dengan menggunakan beberapa teknik, seperti trapezoidal rule, Simpson's rule, dan Mann-Whitney U test. Namun, teknik yang paling umum digunakan adalah trapezoidal rule.
8. AUC dapat dihitung menggunakan berbagai jenis model klasifikasi, seperti logistic regression, decision tree, random forest, dan neural network.
9. AUC dapat dihitung menggunakan berbagai software atau library pemrograman, seperti Python dengan library Scikit-learn atau R dengan library pROC.
10. Penting untuk diingat bahwa AUC tidak memberikan informasi tentang akurasi absolut dari model klasifikasi, melainkan hanya memberikan informasi tentang seberapa baik model dapat membedakan kedua kelas yang berbeda. Oleh karena itu, AUC perlu dilihat bersama dengan metrik evaluasi lainnya untuk mengevaluasi kinerja model secara komprehensif.

Tabel nilai AUC dapat digunakan sebagai acuan untuk mengevaluasi kinerja model klasifikasi. Berikut ini adalah tabel yang menunjukkan interpretasi nilai AUC:

Tabel 2.1 Nilai AUC

Nilai AUC	Interpretasi
0.5	Tidak lebih baik dari model yang acak
0.5 - 0.6	Kinerja yang buruk
0.6 - 0.7	Kinerja yang cukup
0.7 - 0.8	Kinerja yang baik
0.8 - 0.9	Kinerja yang sangat baik
0.9 - 1.0	Kinerja yang luar biasa

Dalam interpretasi nilai AUC, semakin tinggi nilai AUC, semakin baik kinerja model dalam membedakan antara kedua kelas yang berbeda. Namun, perlu diingat bahwa interpretasi ini bersifat relatif dan dapat bervariasi tergantung pada konteks masalah dan data yang digunakan. Oleh karena itu, penting untuk selalu mempertimbangkan metrik evaluasi lainnya dan konteks masalah secara keseluruhan dalam mengevaluasi kinerja model klasifikasi.

Penelitian terkait terdapat pada Tabel 3.1.

2.1 Tabel Penelitian Sebelunnya

No	Judul	Peneliti (Tahun)	Metode	Hasil Penelitian
	Penerapan mengunakan Metode Klasifikasi *Data Mining* Untuk Prediksi Kelulusan Tepat Waktu	Saefulloh dan Moedjiono (2013)	algoritma C4.5, Naive Bayes	Klasifikasi performance keakurasian AUC maka diperoleh hasil penelitian yaitu, dapat diketahui metode yang terbaik adalah Neural Network dengan nilai akurasi 87,32%, yang kedua adalah algoritma Decision Tree dengan nilai akurasi 85.37%, dan yang terakhir adalah K-Nearest Neigboard dengan nilai 83,66%.
	Penerapan Algoritma C4.5 Klasifikasi predikat	Nugroho (2014)	algoritma C4.5	Penerapan algoritma klasifikasi C4.5 akan dapat diimplementasikan pada rekomendasi penerimaan mitra penjualan di PT. Atria Artha Persada, dilihat dari tingkat *accuracy* yang mencapai 96.26 % dan *recall* 71.43%., yang menyatakan bahwa perhitungan yang dilakukan akan mampu memprediksi dan merekomendasikan penerimaan mitra

				penjualan dengan baik.
	Keberhasilan Mahasiswa Di Amik Tunas Bangsa Penerapan Algoritma C4.5	Luvia, Windarto, Solihun, & Hartama(2017)	C4.5	Kombinasi 3-itemset antara lain sebanyak 5,2% lulusan yang berasal dari jurusan matematika dan masuk melalui jalur masuk kemitraan maka menempuh studi selama kurang dari 4 tahun dengan nilai confidence 96,9%.
	Classification Techniques in Data Mining-Case Study	AgarwalBabu, & Reddy (2016)	ID3, *decision tree* (DT), C4.5, Bayesian *classification*	Hasil percobaan kami menunjukkan bahwa pengklasifikasi yang agak sederhana memberikan hasil yang berguna dengan akurasi antara 75 dan 80% yang sulit dikalahkan dengan model lain yang lebih canggih. daripada negatif palsu.
	Penerapan Data Mining Klasifikasi Pola Nasabah Menggunakan Algoritma C4.5 Pada Bank BRI Batang	Ayu Rizqi Oktaviana, 2016)	*Decision Tree*	Aplikasi yang diimplementasikan akan dibandingkan dengan hasil menggunakan sofrware rapidminer. Sehingga diperoleh akurasi dengan decision tree sebesar 89,5%.

Berdasarkan Tabel 3.1 dapat dijelaskan bahwa hasil penelitian-penelitian terdahulu dapat disimpulkan bahwa penerapan algoritma klasifikasi C4.5 dapat diimplementasikan pada berbagai bentuk keperluan klasifikasi, selain itu tingkat *accuracy* penerapan algoritma ini dapat mencapai lebih dari 80%. Dengan demikian, hasil penelitian ini bahwa penerapan algoritma C4.5 dan tingkat

accuracy C4.5 ini diharapkan dapat digunakan secara optimal dalam memprediksi Prestasi Peserta Didik Berdasarkan Sosial Ekonomi, Motivasi, Kedisiplinan dan Prestasi Masa Lalu di SMKN 1 Penawartama Tulang Bawang