BAB II

TINJAUAN PUSTAKA

2.1 *Knowledge Discovery in Database* (KDD)

Knowledge Discovery in Databases adalah singkatan dari KDD dan KDD adalah cara memperoleh pengetahuan dengan menggunakan data dari database atau data yang telah dikumpulkan dan telah disimpan. Setelah pengetahuan ini ditemukan, pada akhirnya digunakan sebagai basis pengetahuan dam diproses untuk mengambil keputusan dalam menentukan hasil akhir. Berikut tahapan KDD yaitu :

1. *Seleksi Data*

   Langkah ini dilakukan pada awal proses KDD yang meliputi pengumpulan informasi yang juga melalui proses pemilihan data yang nantinya akan dijadikan sumber data akhir untuk diolah dalam Data Mining.

2. *Preprocessing atau pembersihan*

   Langkah ini dilakukan dengan tujuan untuk menghilangkan beberapa data duplikat pada data, memeriksa data yang tidak memenuhi kebutuhan karena proses penambahan data harus sesuai dengan kebutuhan permintaan penelitian.

3. *Konversi*

   Berikutnya ada proses memodifikasi dan juga mengadaptasi model penyimpanan database.

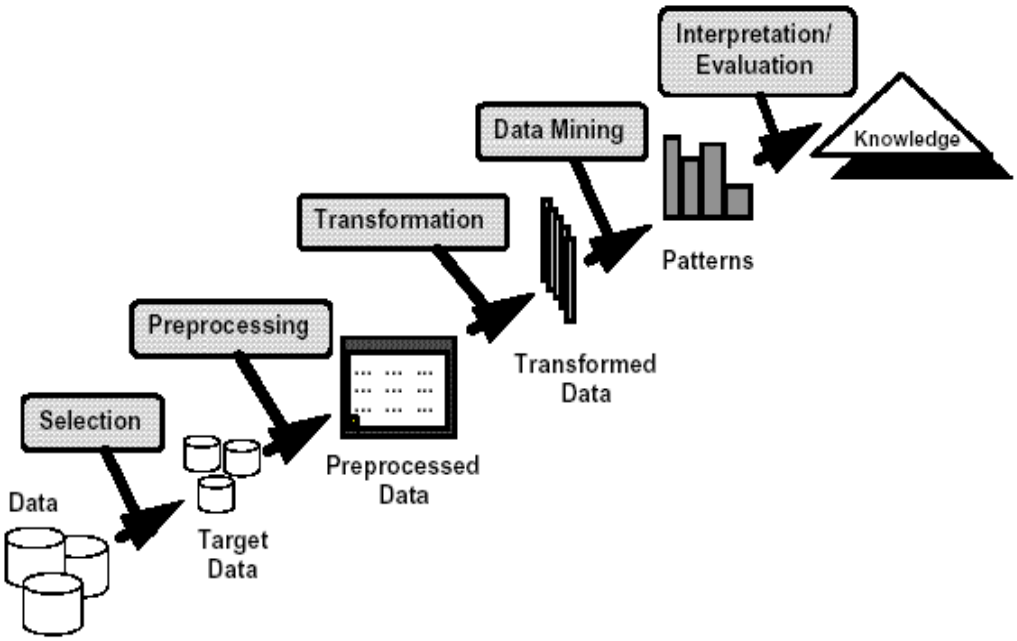
4. Data Mining

   Pada tahap data mining yaitu cara menemukan dan untuk melatih model aturan agar menghasilkan informasi dalam bentuk keputusan berdasarkan tujuan penelitian.

5. Interpretasi(interpretation)

   Menampilkan hasil rule model dari data mining agar dapat dipahami terutama informasi yang bertentangan dengan hipotesis penelitian.

Dari sudut pandang lain, data mining di anggap sebagai salah satu langkah penting dalam proses KDD. Knowledge discovery in databases (KDD) adalah keseluruhan proses untuk mencari dan mengidentifikasi pola (*pattern*) dalam data, dimana pola yang ditemukan dapat bermanfaat dan dapat dimengerti. KDD berhubungan dengan teknik integrasi, interpretasi dan visualisasi dari pola-pola sejumlah kumpulan data.



Gambar 2. 1 Tahapan KDD

2.2 *Data Mining*

Data mining adalah suatu istilah yang digunakan untuk menguraikan penemuan pengetahuan didalam database.

Data mining merupakan proses pertama yang akan dilakukan yaitu pencarian informasi baru (Elisa et al., 2019).

Data mining digunakan untuk mengekstrak esensi pengetahuan dari sekumpulan data yang mudah dipahami manusia, meliputi database dan manajemen data, preprocessing data, pertimbangan dan inferensi model, pengukuran kepentingan, pertimbangan kompleksitas, pasca-pemrosesan struktur yang terdeteksi dan juga divisualisasikan dan diperbarui secara online (Laia et al., 2018).

Gartner Group menyatakan bahwa data mining adalah proses menemukan hubungan, pola, dan kebiasaan baru yang bermakna dengan mengorganisasikan sejumlah besar data yang disimpan pada media penyimpanan dengan menggunakan teknologi pemrosesan data, pengenalan pola sebagai teknik statistik dan matematika.

Data mining merupakan gabungan beberapa disiplin ilmu yang menggabungkan pembelajaran mesin, pengenalan pola, statistik, database, dan teknik visualisasi untuk memecahkan masalah pengambilan informasi dari database besar . (Idris, 2019).

Data mining dikelompokan berdasarkan tugas yang dapat dilakukan, yaitu:

1. *Description* (Deskripsi)

   Peneliti mencari cara untuk mendeskripsikan pola dan tren pada dalam sample data. Misalnya, pelaksana pemilu mustahil jika tidak ditemukannya bukti atau data yang menunjukkan bahwa orang yang tidak profesional hanya akan mendapat sedikit dukungan dalam pemilihan presiden.

2. *Estimation* (Estimasi)

   Estimasi kurang lebih sama dengan kategorikal, kecuali bahwa variabel target diestimasi secara numerik dan bukan kategoris. Model dibangun dari record lengkap yang memberikan nilai variabel target sebagai nilai prediksi. Selain itu, pada scan berikutnya nilai estimasi variabel target didasarkan pada nilai variabel prediktor. Contohnya

adalah memperkirakan IPK kumulatif mahasiswa pascasarjana dengan memeriksa IPK siswa tersebut saat mendaftar di program sarjana.

3. Prediction (Prediksi)

   Prediksi seperti klasifikasi dan estimasi, kecuali memprediksi nilai hasil di masa depan. Berikut adalah beberapa contoh prediksi dan penelitian bisnis:

   a. Prediksi harga gula dalam lima bulan yang akan datang.
   b. Prediksi tingkat kemiskinan tiga tahun akan datang.
   c. Predisksi persentase kenaikan harga saham 10 tahun mendatang.

Beberapa metode dan teknik yang digunakan dalam klasifikasi dan estimasi juga dapat digunakan (bila sesuai) untuk prediksi.

2.3 Metode *Data Mining*

Data mining dapat digunakan sebagai prediksi untuk memperkirakan nilai masa mendatang. Dengan menerapkan teknik ini akan dibangun pohon keputusan (*decision tree*) untuk kemungkinan siswa yang dapat menyelesaikan studi dengan baik. Salah satu metode data mining decision tree yang terkenal dan dapat digunakan sebagai prediksi adalah algoritma C4.5. Dimana algoritma C4.5 merupakan algoritma klasifikasi data dengan teknik pohon keputusan yang dapat mengolah data numerik (kontinyu) dan diskrit, dapat menangani nilai atribut yang hilang, menghasilkan aturan-aturan yang mudah diinterpretasikan dan tercepat diantara algoritma-algoritma lain.

2.3.1 Algoritma C4.5

Algoritma C4.5 adalah suatu deretan algoritma untuk permasalahan klasifikasi didalam sebuah mesin dan himpunan data. Dengan nilai data yang bervariasi, dimana kejadian diuraikan oleh koleksi atribut dan mempunyai salah satu dari satu set kelas yang eksklusif.

Lalu hasil atau data yang telah dikumpulkan selanjutnya akan digunakan untuk mengolah data baru yang disebut test atau dataset. Menggunakan algoritma C4.5 dalam melakukan proses klasifikasi,sehingga hasilnya adalah pengelompokan test atau dataset ke dalam beberapa kelas. Secara umum, memakai algoritma C4.5 adalah langkah yang digunakan untuk menentukan pohon keputusan adalah.

a. Pilih atribut sebagai root.
b. Buat cabang untuk setiap nilai.
c. Bagi tiap cabang kedalam kelas.
d. Ulangi proses untuk setiap cabang sampai semua kasus pada tiap cabang memiliki kelas yang sama.

Pertama menentukan atribut yang akan dijadikan root, menentukan berdasarkan nilai gai yang paling tinggi dari semua atribut yang ada. Dan berikutnya untuk mendapat nilai paling tinggi sebelum nya melakukan perhitungan nilai *entropy* dari semua data yang ada untuk mengetahui ukuran dari beberapa varian data . ketika telah ditentukan nilai *entropy* makan akan dijadikan atribut yang paling penting untuk menentukan pengelompokan data , dan ini disebut *information gain*. Saat penelitian, untuk memilih atribut sebagai akar didasarkan pada nilai

information gain tertinggi dari atribut yang ada. Berikut rumus perhitungan yang digunakan.

1. Entropy, merupakan langkah awal dalam perhitungan algortima C4.5.

$$Entropy(S) = \sum_{i=1}^{n} - Pi * log_2 \ Pi$$

Rumus 2. 1 Entropy

Keterangan :

S: himpunan kasus

n: jumlah partisi S

Pi: proporsi dari Si terhadap S

2. Information gain, merupakan yaitu kriteria yang digunaan untuk memilih suatu atribut yang popular, dapat dihitung dengan cara pengelompokan berdasarkan masing-masing atribut dalam satu data. Notasi information gain adalah Gain (S.A) yang berarti dalam data atribut A relative terhadap output.

Rumus 2. 2 Gain

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

Keterangan :

S : himpunan kasus

A : atribut

N : jumlah partisi atribut A

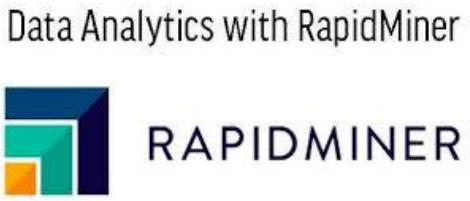
|Si| : jumlah kasus pada partisi ke-i

|S| : jumlah kasus dalam S

2.4 Software Pendukung

Dengan menggunakan prinsip dan algoritma data mining RapidMiner digunakan sebagai software pendukung pada penelitian ini, RapidMiner mengekstrak pola-pola dari data set yang besar dengan mengkombinasikan metode statistika, kecerdasan buatan dan database.

2.4.1 RapidMiner



Gambar 2. 2 Aplikasi RapidMiner

RapidMiner (YALE) adalah perangkat lunak open source untuk knowledge discovery dan data mining. Rapidminer memiliki kurang lebih 400 prosedur (operator) data mining termasuk operator untuk masukan, output, data preprocessing dan visualisasi (Sulianta, dkk 2010:101).

Beberapa fitur RapidMiner antara lain:

1. Lisensi gratis (sumber terbuka).
2. Cross-platform karena diprogram dalam bahasa Java.
3. Data internal berbasis XML untuk memfasilitasi pertukaran data pengalaman.
4. Dilengkapi dengan bahasa scripting untuk mengotomatisasi percobaan.

5. Memiliki GUI (antarmuka pengguna grafis), mode baris perintah (mode batch) dan API Java yang dapat dipanggil dari program lain.
6. Dapat diperluas dengan menambahkan plugin dan ekstensi.

2.5 Penelitian Terdahulu

1. *Implementation of Data Mining to Analyze Drug Cases Using C4.5 Decision Tree* (Sri Wahyuni, 2018). Data mining was the process of finding useful information from a large set of databases. One of the existing techniques in data mining was classification. The purpose of this study is to analyze the data of prisoners in Labuhan Deli prison extracted with data mining decision tree algorithm C4.5, then it will generate new knowledge to learn about the factors of detainees doing drug crimes. The method used is the decision tree method and the algorithm used is the C4.5 algorithm. From the solution of the decision tree, a number of rules will arise for a case. In this case, the researcher classified prisoner data at Labuhan Deli prison to know the factors that lead prisoners to commit drug-related crimes. By applying this C4.5 algorithm, knowledge is obtained in the form of information aimed at minimizing drug-related criminal acts.
2. *Implementation Data Mining using Decision Tree Method- Algorithm C4.5 for Postpartum Depression Diagnosis* (Aris Supriyanto, 2018). t. Postpartum depression is a serious problem that needs to be addressed because it negatively affects the family, the child's health, cognitive function, and mother-child interactions. Diagnosis is performed based

on data on psychological status, blood pressure, respiration, body temperature, and classification extracted using the C4.5 decision tree algorithm method. This study aims to apply data mining using the C4.5 algorithm to determine the level of postpartum depression by categories, factors related to postpartum and appropriate solutions using an information system. This study aims to improve services for postpartum patients using an online information system. The use of information systems can improve organizational performance because processes can be performed automatically, thereby increasing economic benefits. The results show the largest increase for psychological variables 0.57 node 1, blood pressure 0.54 node 2, body temperature 0.54 node 3, meaning these three variables have more influence on health. The patient is depressed and needs priority treatment.

3. *Implementasi Data Mining Dengan Algoritma Decision Tree C4.5 Untuk Prediksi Kelulusan Mahasiswa Di Universitas Pandanaran* (Abdul Rohman1, Anief Rufiyanto2, 2019) Sangat penting bagi penyelenggara pendidikan memprediksi bagaimana prestasi akademik semua mahasiswa karena program strategis tersebut dapat direncanakan dalam mempertahankan meningkatkan atau kinerja mahasiswa selama masa studi di perguruan tinggi. sangat penting untuk mengambil suatu keputusan untuk data mahasiswa, menggunakan data mining semua data tersebut dianalisa lebih lanjut. Penggunaan data mining dengang menggunakan algoritma C4.5 Decision Tree untuk melihat pola

kelulusan siswa dapat digunakan untuk mengembangkan sistem yang dapat. Untuk memprediksi kelulusan siswa dengan metode algoritma pohon keputusan sudah banyak penelitian yang mengimplementasikan data mining, dengan data siswa normal dan sebagian besar siswa tidak memenuhi syarat. Sedangkan pada penelitian ini mahasiswa Universitas Pandanaran memiliki data mahasiswa kelas reguler dan mahasiswa kelas pekerja dan sebagian besar berstatus pekerja. Langkah-langkah yang dilakukan dalam penelitian ini adalah: (1) mendapatkan semua data dari mahasiswa Universitas Pandanaran, (2) mengolah dan menganalisa semua data dari mahasiswa dengan memakai algoritma klasifikasi data mining decision tree(3) lalu lakukan percobaan eksperimen dan percobaan terakhir algoritma (4) mengevaluasi dan mencocokkan hasil akhir (5) untuk mengambil keputusan penerimaan universitas dan kemudian menghasilkan pola/model kelulusan mahasiswa yang dapat digunakan. Hasil penelitian ini menghasilkan 10 rule dengan nilai AUC sebesar 0.874 dengan nilai akurasi sebesar 65.98 dan dapat tergolong klasifikasi data yang baik. Oleh karena itu, hasil ini penting untuk pengambilan keputusan dalam Organisasi.

4. *Data Mining Menggunakan Algoritma C4.5 Untuk Memprediksi Kepuasan Mahasiswa Terhadap Kinerja Dosen Di Kota Batam* (Yulia, Anggia Dasa Putri, 2019) This study aims to determine the influence of student satisfaction on teacher's performance in lessons using data mining techniques using the C4.5 algorithm in which the variables are

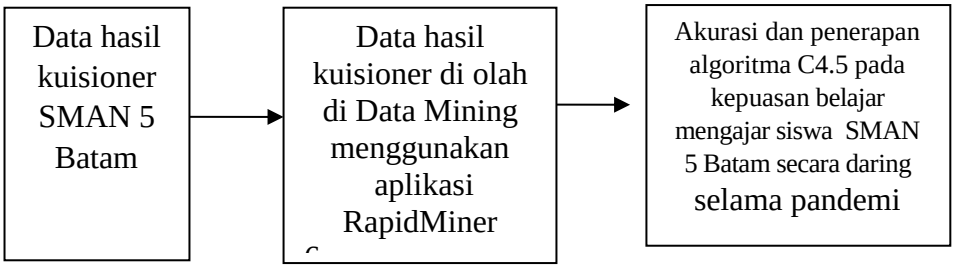
Use includes satisfaction. reliability, responsiveness, appearance, empathy and trust. The objective of the study is to predict the level of student satisfaction with the teaching and learning activities of university professors in Batam city. The concern of this study is to know the capacity of a teacher to create a quality teaching staff and a quality education. Data mining using C4.5 algorithms, data classification and pattern finding processes to provide data insights for decision making. Based on the research results performed, the decision tree calculated manually using the C4.5 algorithm is similar to the decision tree created by Weka software with an accuracy of 94.12%. We can then conclude that the prediction of student satisfaction with teacher performance was met.

5. *Implementasi Teknik Data Mining untuk Prediksi Peminatan Jurusan Siswa Menggunakan Algoritma C4.5* (Novitaria Manullang 1), Rahmat Widia Sembiring 2), Indra Gunawan 3), Iin Parlina 4), Irawan, 2021). SMK Persiapan Swasta merupakan salah satu SMK yang ada di kota Pematangsiantar yang terletak di Jalan Pane No. 66 Desa Tomuan Kecamatan Siantar Timur Kota Pematangsiantar. SMK Persiapan Swasta Pematangsiantar menerima siswa baru pada tahun 1968 dengan menawarkan jurusan khususnya teknik mesin, dan beberapa tahun kemudian seiring dengan minat yang terus meningkat, sekolah tersebut membuka keterampilan 'sekarang mencakup 7 (tujuh) keterampilan.

Penelitian ini bertujuan untuk melakukan prediksi terhadap peminatan jurusan siswa dengan menggunakan teknik data mining dengan algoritma C4.5.Metode yang digunakan adalah algoritma C4.5 yang menentukan jurusan yang dipelajari siswa berdasarkan latar belakang, minat, dan keterampilannya. Variabel yang digunakan adalah nilai ujian utama siswa, minat, dan bakat. Penelitian ini diharapkan dapat mempercepat proses pengambilan keputusan dalam memprediksi jurusan mahasiswa pada proses penerimaan baru. Penerapan data mining dengan teknik algoritma C4.5 dilakukan pada saat pengujian terhadap 100 record yang diperiksa. Ditemukan bahwa algoritma C4.5 dapat mencapai akurasi 100,00%.

2.6 Kerangka Pemikiran

Dalam penulisan kali ini, penulis membuat gambaran singkat sebagai alur penyusunan mengumpulan Data melalui pengisian kuisioner yang berisi pertanyaan seputar pembelajaran online selama masa pandemi COVID kepada siswa-siswi SMAN 5 Batam. Tahap ini memproses data hasil kuisioner yang di isi oleh siswa-siswi SMAN 5 Batam sebelumnya di Data Mining dengan menggunakan aplikasi RapidMiner. Pada tahap ini dilakukan pengecekan hasil data yang diolah menggunakan software Rapid Miner kemudian melakukan prediksi untuk melihat keakuratan data menggunakan algoritma C4.5 dan terakhir dilakukan evaluasi akhir untuk mengetahui apakah pengolahan data tersebut sesuai dengan hasil pengujian yang diharapkan.



Gambar 2. 3 Kerangka Pemikiran