BAB 2

TINJAUAN PUSTAKA

2.1 Tinjauan Pustaka

Table 2.1 Penelitian terdahulu

Penulis	Tahun	Masalah	Metode	Hasil
Heri Susanto dan Sudiyanto	2014	Memprediksi prestasi siswa berdasarkan sosial ekonomi, motivasi, kedisiplinan dan prestasi masa lalu	Data mining, Decision tree algoritma J48, CHAID	Akurasi Decision tree algoritma J48 sebesar 95,7% sedangkan CHAID sebesar 82,1% dan analisis regresi ganda sebesar 90,6%
Siska Haryati, Aji Sudarsono, Eko Suryana	2015	Implementasi Data Mining Untuk Memprediksi Masa Studi Mahasiswa	Data mining, Algoritma C4.5	hasil evaluasi penelitian bahwa algoritma C4.5 mampu menganalisa tingkat ketepatan waktu mahasiswa
Senna Hendrian	2018	Memprediksi Siswa Dalam Memperoleh Bantuan Dana Pendidikan	data mining, algoritma klasifikasi, algoritma C4.5	nilai Accuracy Algoritma C4.5 adalah sebesar 98,80%, nilai untuk Precision sebesar 98,02%, dan nilai untuk Sensitivity atau Recall sebesar 99,00%. Dengan demikian Algoritma C4.5
Defri Kurniawan, Wibowo Wicaksono dan Yani Parti Astuti	2016	Memprediksi masa studi mahasiswa	Algoritma C4.5	Algoritma C4.5 memiliki akurasi baik (73,68%)
Indri Rahmayuni	2014	Klasifiksi data nilai mahasiswa	Algoritma C4.5 dan Cart	Algoritma C4.5 akurasi paling baik (85.61%) sedangkan cart sedikit dibawahnya (84.95%)
Andika Prajana	2016	Memprediksi tingkat kelulusan siswa dalam ujian nasional pada sma negeri 5 kota banda aceh	Roughset	Hasil nya siswa yang memiliki riwayat nilai yang bagus pada semua atau pada mata pelajaran tertentu yang di Ujian Nasional akan memiliki kecendrungan untuk mendapatkan nilai yang bagus juga dalam Ujian Nasional

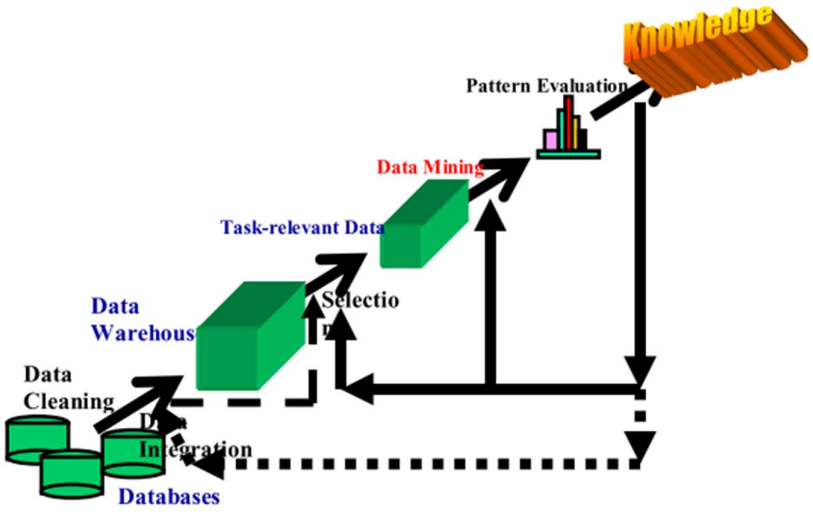
Muhamad Jamaris	2017	mplementasi Metode Rough Set Untuk Menentukan Kelayakan Bantuan Dana Hibah Fasilitas Rumah Ibadah	Rough set	analisis yang dilakukan dapat menghasilkan keputusan yang optimal sehingga dapat memberikan prediksi kepada pihak seleksi biro kesra provinsi Riau dalam penentuan kelayakan bantuan dana hibah fasilitas rumah ibadah, aturan atau pola rules yang terbentuk menjadi informasi yang bermanfaat dalam pengambilan keputusan. Pemohon yang keputusannya diterima layak diberikan bantuan sementara keputusan yang diproses dapat ditinjau kembali atau dipertimbangkan apakah layak diberikan bantuan atau tidak, sedangkan keputusan yang ditolak tidak layak untuk diberikan bantuan dana hibah
Edi Sutoyo	2018	Analisis data intelijen dan penambangan data yang mampu menangani pengetahuan kabur, belum pasti, mengandung fuzzy dan juga system informasi yang tidak lengkap	Rough Set	Hasil menujukan bahwa teknik ini mampu mencapai akurasi 96.04% dengan waktu eksekusi 3.1830 detik
Karmila Suryani	2016	Peluang Kelulusan Mahasiswa	Roughset	Predeksi tingkat kelulusan mahasiswa PTIK dalam menguasai ketiga materi, yaitu 4 orang yang pasti lulus 100%, 8 orang dengan peluang 67% serta 6 orang dengan peluang tidak lulus 33%

2.2 Data Mining

Data mining merupakan suatu metode menemukan suatu pengetahuan dalam suatu database yang cukup besar. Data mining adalah proses menggali dan menganalisa sejumlah data yang sangat besar untuk memperoleh sesuatu yang

benar, baru, sangat bermanfaat dan akhirnya dapat dimengerti suatu corak atau pola dalam data tersebut.

Data mining adalah bagian integral dari penemuan pengetahuan dalam database (KDD), yang merupakan proses keseluruhan mengubah data mentah menjadi informasi yang bermanfaat, seperti yang ditunjukkan pada Gambar 2.1. (Wahyudi, 2013).



Gambar 2.1 Proses *Data Mining*

Terdapat empat tugas utama dari *data mining* yakni:

1. Model prediktif

   Model prediktif digunakan untuk membangun model variabel target sebagai fungsi dari variabel penjelas. Variabel penjelas dalam hal ini adalah semua atribut yang digunakan untuk membuat prediksi, sedangkan variabel target adalah atribut yang nilainya diprediksi. Pemodelan prediktif dapat dibagi menjadi dua jenis. Artinya, klasifikasi digunakan untuk memprediksi nilai variabel target diskrit, dan regresi digunakan untuk memprediksi nilai variabel target kontinu.

2. Analisis terkait

   Analisis asosiasi digunakan untuk menemukan aturan asosiasi yang menunjukkan kondisi nilai atribut yang sering terjadi bersama-sama dalam sebuah dataset.

3. Analisis klaster

   Analisis klaster berbeda dengan klasifikasi, yang menganalisis kelas data objek yang berisi label. Clustering menganalisis objek data tanpa mencari nama kelas yang diketahui. Penunjukan kelas termasuk dalam data pelatihan. Karena sebelumnya tidak diketahui. Clustering adalah proses pengelompokan kelompok objek yang sangat mirip.

4. Deteksi anomali

   Deteksi anomali adalah metode pendeteksian data yang bertujuan untuk menemukan objek yang berbeda dari kebanyakan objek lainnya. Anomali dapat dideteksi dengan uji statistik yang menerapkan model distribusi atau model probabilistik pada data.

2.3 Education Data Mining

Educational Data Mining adalah disiplin yang muncul berkaitan dengan pengembangan metode untuk mendapatkan informasi unik dari dataset yang berasal dari pendidikan. (Hidayat et al., 2013). Data mining adalah proses yang mempekerjakan satu atau lebih teknik pembelajaran komputer (*machine learning*) untuk menganalisis dan mengekstraksi pengetahuan(*knowledge*) secara otomatis, Ada beberapa tahap dalam Data Mining yaitu :

1. Pembersihan Data (*Data Cleaning*)
2. Integrasi Data (*Data Integration*)
3. Seleksi Data (*Data Selection*)
4. Transformasi Data (*Data Transformation*)
5. Proses Mining
6. Evaluasi pola (*Pattern Evaluation*)
7. Presentasi Pengetahuan (*Knowledge Presentation*)

Pada penelitian sebelumnya, telah banyak penelitian membahas terkait dengan *educational data mining* seperti seperti (Rahmayuni, 2014) menggunakan algoritma C4.5 dan Cart Klasifiksi data nilai mahasiswa Algoritma C4.5 akurasi paling baik (85.61%) sedangkan cart sedikit dibawahnya (84.95%) validasi yang di gunakan *Cross-validation*. (Mayadewi & Rosely, 2015) menggunakan ID3, CHAID, *Naïve Bayes* bertujuan untuk membuat aturan yang dapat memprediksi nilai proyek akhir mahasiswa ID3 memiliki akurasi sebesar 62,66%, CHAID

63,66% dan *Naïve Bayes* 65,67% validasi yang digunakan *x-validation*. (Kurniawan et al., 2007) menggunakan Algoritma C4.5 Memprediksi masa studi mahasiswa. (Astuti et al., 2018) menggunakan Algoritma *Naive Bayes , forward selection.* Seleksi untuk mengetahui hubungan variabel nilai Algoritma C4.5 memiliki akurasi baik (73,68%). (Susanto & Sudiyatno, 2014) menggunakan *Data mining, Decision tree* algoritma J48, dan CHAID Memprediksi prestasi siswa berdasarkan sosial ekonomi, motivasi, kedisiplinan dan prestasi masa lalu Akurasi dari *naive bayes* 67,77% menjadi 78,08% stelah di optimalkan dengan *forward selection* validasi yang digunakan *Cross Validasi*. (Haryati et al., 2015) Menggunakan *Data mining*, Algoritma C4.5 Memprediksi Masa Studi Mahasiswa hasil evaluasi penelitian bahwa algoritma C4.5 mampu menganalisa tingkat ketepatan waktu mahasiswa. (Dahlan Abdullah, Cut Ita Erliana, 2015) menggunakan data mining, algoritma klasifikasi, algoritma C4.5 Memprediksi Siswa Dalam Memperoleh Bantuan Dana Pendidikan nilai Accuracy Algoritma C4.5 adalah sebesar 98,80%, nilai untuk Precision sebesar 98,02%, dan nilai untuk *Sensitivity* atau *Recall* sebesar 99,00%. Dengan demikian Algoritma C4.5.

2.4 Rough Set

Metode *Rough Set* dikembangkan oleh Zdizslaw Pawlak sebagai alat matematical untuk menangani ketidakjelasan dan ketidakpastian. Telah berhasil diterapkan dalam berbagai tugas, seperti fitur seleksi/ekstraksi, sintesis aturan dan klasifikasi, penemuan pengetahuan, dan lain-lain (Sembiring & Azhar, 2013).

Rough Set salah satu teknik data mining yang digunakan untuk menangani masalah *Uncertainty, Imprecision* dan *Vagueness* dalam aplikasi *Artificial Intelligence* (AI). Rough set merupakan teknik yang efisien untuk *Knowledge Discovery in Database* (KDD) dalam tahapan proses dan data mining. *Rough set* merupakan teknik yang efisien untuk KDD dalam tahapan proses dan Data Mining.

2.4.1 Indiscernibility Relation

Indiscernibility Relation adalah konsep utama yang digunakan dalam variabel *selection* pada *rough set*. Misal S = (U , A) sebagai sistem informasi, dimana U adalah himpunan objek yang tidak kosong dan A adalah himpunan atribut yang tidak kosong, jika α:U □ Vα, untuk setiap α ϵ A, maka Vα adalah himpunan nilai atribut α yang mungkin. Jika P ⊆ A dapat diasosiasikan dengan relasi ekivalen

IND(P); maka IND(P) = { (x,y) ϵ U2 | ∀α ϵ P, α(x) = α(y)} partisi himpunan U digenerate oleh IND(P) yang dinotasikan dengan U/IND(P).

2.4.2 Set Approximation

Untuk *decision system*, sangat penting menemukan seluruh subset menggunakan kelas yang ekivalen yaitu yang mempunyai nilai kelas yang sama. Tetapi, subset ini tidak selalu didefinisikan dengan tepat. Meskipun data tabel tidak dapat didefinisikan dengan tepat, hal ini juga dapat diatasi dengan melakukan perkiraan dengan menggunakan lower dan upper *approximations* yang didefinisikan sebagai :

vc$BX = x \in U: B\ x \subseteq X$ dan $B\ X = \{\ x \in U: B\ x \cap X \neq X'\}$ (2) di mana BX adalah *lower approximations* dari himpunan X, sedangkan $B\ X$ adalah *upper approximations* dari himpunan X, dimana X merupakan nilai dari atribut keputusan.B adalah boundary region dari himpunan X, jika $BNB\ X = B\ X - BX$ (3) Menurut Khaerunnisa (2016) Boundary Region B himpunan X terdiri dari objek-objek dari himpunan U yang tidak dapat diklasifikasikan ke dalam himpunan X maupun ke dalam himpunan UX yang diwakili oleh atribut himpunan B. Jika *boundary region* yang diberikan himpunan X adalah himpunan kosong, maka dapat dikatakan bahwa *X crisp* terhadap himpunan atribut (Sinaga et al., 2017)

2.4.3 Roughness

Pengukuran kekasaran dan kesamaan untuk dua bagian informasi yang berbeda dalam satu set universal hal yang sama juga penting untuk menjelaskan kekuatan dan kelengkapan informasi itu diberikan. Untuk *multiset neutrosopic* kasar, nilai aproksimasi bawah dan atas adalah properti dari bukti dalam menjelaskan kekasaran informasi yang dibutuhkan. Sedangkan modelnya informasi vektor, yaitu pengukuran kosinus dan pengukuran dadu merupakan hasil pengukuran kesamaan *multiset neutrosopic* kotor (Suriana et al., 2020).

2.4.4 Dependensi Atribut

8. Nilai *indiscernibility* yang pertama dicari adalah *indiscernibility* untuk kombinasi atribut yang terkecil yaitu 1.
9. Kemudian lakukan proses pencarian *dependency attributes*. Jika nilai *dependency attributes* yang didapat =1 maka *indiscernibility* untuk himpunan minimal variabel adalah variabel tersebut.

10. Jika pada proses pencarian kombinasi atribut tidak ditemukan *dependency attributes* =1, maka lakukan pencarian kombinasi yang lebih besar, di mana kombinasi variabel yang dicari adalah kombinasi dari variabel di tahap sebelumnya yang nilai *dependency attributes* paling besar.
11. Lakukan proses (3), sampai didapat nilai *dependency attributes* =1.

2.4.5 Reduct

Reduct merupakan proses penyeleksian atribut minimal (interesting attribute) dari sekumpulan kondisi dengan menggunakan *prime implicant* fungsi *Boolean*. (Raharjo & Windarto, 2021)

2.4.6 Generating Rules

Generating Rules adalah suatu metode *Rough Set* untuk menghasilkan *rules/knowledge* berdasarkan *equivalence class and reduct*. *Generating rules* dapat juga dikatakan sebagai suatu algoritma dari data mining, yang mana nantinya dari proses *generating rules* ini akan dihasilkan suatu *rules / knowledge* yang dapat digunakan dalam sebuah pengambilan keputusan. (Puspasari, 2015)

2.4.7 Pengukuran Kualitas Aturan

Pengukuran kualitas aturan (*quality measurement for rules*) ialah pengukuran yang dilakukan terhadap item *rule* yang dihasilkan (Soelaiman et al., 2008). Terdapat beberapa instrumen pengukuran yakni: yang meliputi *support, accuracy* dan *coverage*.

1) *Support*.

   Support dari *decision rules* ialah jumlah objek dari *decision rule* yang memiliki *antecedent* (f) *and conclusion* (g) yang sesuai.

$$Support(f \rightarrow g) = card(|f \cap g|)$$

2) *Accuracy.*

   Accuracy dari *decision rules* merupakan rasio perbandingan dari objek yang memenuhi *antecedent* dan juga memenuhi *conclusion* terhadap objek yang memenuhi *antecedent* saja.

$$accuracy(f \rightarrow g) = \frac{Support(f \rightarrow g)}{|f|}$$

3) *Coverage*

Coverage merupakan rasio perbandingan dari objek yang memenuhi *antecendent* dan juga memenuhi *conclusion* terhadap objek yang memenuhi *conclusion* saja.

$$coverage(f \rightarrow g) = \frac{Support(f \rightarrow g)}{|g|}$$

2.5 Validasi Model

Konfusi Matriks 1

	Prediksi Naik	Prediksi Tetap	Prediksi Turun
Naik	N	N	N
Tetap	T	T	T
Turun	R	R	R

Keputusan yang di notasikan N = Naik , T = Tetap dan R = Turun

NIM	Hasil (Real)	Prediksi	Keputusan
1	N	N	1
2	N	N	1
3	R	N	1
4	T	T	0
5	T	T	1
6	T	T	1
7	N	N	1
8	T	T	1
9	R	T	0
10	R	R	1
Hasil			8

Notasi	**Prediksi dan Keputusan**
NN	Naik Naik
NT	Naik Tetap
NR	Naik Turun
TN	Tetap Naik
TT	Tetap Tetap
TR	Tetap Turun
RN	Turun Naik
RT	Turun Tetap

RR	Turun Turun

NN	TN	RN
NT	TT	RT
NR	TR	RR

3	0	1
0	4	1
0	0	1

Akurasi

$$\frac{NN + TT + RR}{NN + NT + NR + TN + TT + TR + RN + RT + RR}$$

$$\frac{3 + 4 + 1}{3 + 0 + 0 + O + 4 + 0 + 1 + 1 + 1} = \frac{8}{10} = 0{,}8\ jadi\ akurasi\ 80\ \%$$

Precision adalah jumlah kelompok dokumen yang relavan dari total jumlah dokumen yang ditemukan oleh sistem, sedangkan *Recall* adalah perhitungan dokumen yang relevan dari seluruh dokumen yang berada didalam sistem.

Precision Naik

$$\frac{NN}{NN + TN + RN}$$

$$\frac{3}{3 + 0 + 1} = \frac{3}{4} = 0{,}75\ jadi\ Presisi\ Naik\ 75\%$$

Precision Tetap

$$\frac{TT}{TT + NT + RT}$$

$$\frac{4}{4 + 0 + 1} = \frac{4}{5} = 0{,}8\ jadi\ Presisi\ Tetap\ 80\%$$

Precision Turun

$$\frac{RR}{RR + NR + TR}$$

$$\frac{1}{1 + 0 + 0} = \frac{1}{1} = 1\ Jadi\ Presisi\ Turun\ 100\%$$

Recall Naik

$$\frac{NN}{NN + NT + NR}$$

$$\frac{3}{3+0+0} = \frac{3}{3} = 1\ jadi\ Recall\ Naik\ 100\%$$

Recall Tetap

$$\frac{TT}{TT+TN+TR}$$

$$\frac{4}{4+0+0} = \frac{4}{4} = 1\ jadi\ Recall\ Tetap\ 100\%$$

Recall Turun

$$\frac{RR}{RR+RN+RT}$$

$$\frac{1}{1+1+1} = \frac{1}{3} = 0{,}33\ Jadi\ Recall\ Turun\ 33{,}3\ \%$$