BAB II

TINJAUAN PUSTAKA DAN DASAR TEORI

2.1 Tinjauan Pustaka

Pada penelitian ini menggunakan beberapa referensi terkait dengan metode Algoritma C5.0. hal ini berfungsi sebagai pedoman sekaligus pembanding dengan penelitian terdahulu terhadap penelitian yang akan dilakukan. Referensi tersebut diantaranya sebagai berikut:

Nusari, Purbasari, dan Puspaningrum (2020) menggunakan algoritma *decision tree* C5.0 dalam memprediksi ketepatan waktu kelulusan mahasiswa. Dari evaluasi model yang dilakukan, diperoleh tingkat akurasi sebesar 83,78% dengan nilai *precision* sebesar 92,85%, *recall* sebesar 72,22%, dan nilai *error* sebesar 16,21%.

Sungkar dan Qurohman (2021) menerapkan algoritma C5.0 dalam memprediksi kelulusan pembelajaran mahasiswa pada mata kuliah Arsitektur Sistem Komputer. Proses prediksi dilakukan berdasarkan dengan klasifikasi algoritma C5.0 menggunakan atribut Nilai Kehadiran, Nilai Tugas, Nilai UTS dan Nilai UAS. Kinerja algoritma C5.0 mendapatkan tingkat akurasi yang tinggi sebesar 93.33%.

Giustin, Sari, dan Padilah (2022) memanfaatkan algoritma C5.0 dalam memprediksi hasil belajar pada mata kuliah Kalkulus. Penelitian ini menerapkan metodologi *Cross Industry Standard Process for Data Mining* (CRISP –DM) dengan algoritma C5.0 yang menggunakan atribut wali, jumlah anggota keluarga, status tempat tinggal, internet, aktivitas, keinginan melanjutkan sekolah,

pendidikan terakhir orang tua (ayah dan ibu), pekerjaan orang tua, nilai tugas, UAS, dan UTS. Memperoleh hasil akurasi sebesar 95%.

Dewi, Cholissodin, dan Sutrisno (2019) melakukan klasifikasi penyimpangan tumbuh kembang anak dengan menggunakan algoritma C5.0. Penelitian ini bertujuan untuk mengembangkan model klasifikasi dan mengklasifikasikannya menjadi tiga jenis yaitu *autisme*, *down syndrome*, dan ADHD (*Attention Deficit Hyperactivity Disorder*). Pada tahap pengujian dilakukan perbandingan hasil akurasi dua algoritma, pada algoritma C4.5 menghasilkan 87,61% dan algoritma C5.0 menghasilkan 93,33%.

Esananda, Nugroho, dan Anggraeny (2021) melakukan penelitian mengenai implementasi fase *boosting* pada algoritma C5.0 untuk menentukan prestasi akademik siswa. Penelitian ini menggunakan fase *boosting* yang dapat meningkatkan akurasi model pohon keputusan. Tetapi proses *boosting* tidak selalu menghasilkan kondisi yang lebih akurat, hal tersebut disebabkan adanya pembentukan data *training* baru secara random. Sehingga penelitian ini menghasilkan pohon keputusan dengan akurasi 93,15%, nilai *recall* sebesar 84,32%, *precision* sebesar 81,62%, dan nilai *accuracy* sebesar 84,03%.

Penelitian ini menggunakan algoritma C5.0 untuk memprediksi ketepatan lulus mahasiswa penerima KIP-Kuliah merdeka yang kemudian akan digunakan untuk memonitoring mahasiswa sehingga dapat lulus tepat waktu. Perbandingan antara beberapa penelitian terdahulu disajikan pada tabel 2.1.

Tabel 2. 1 Tinjauan Pustaka

No	Nama Penulis	Objek	Metode	Hasil
1.	Anita Nusari, Intan Yunia Purbasari, Eva Yulia Puspaningrum(2020)	Prediksi Ketepatan Waktu Kelulusan Mahasiswa	Algoritma C5.0	Hasil pengujian dengan menerapkan algoritma C5.0 dan menggunakan data mahasiswa dengan jumlah 125 data dimana sebanyak 88 data merupakan data latih dan 37 merupakan data uji. Memperoleh nilai *recall* sebesar 72,22%, nilai *precision* sebesar 92,85%, nilai *accuracy* sebesar 83,78%, dan nilai *error* sebesar 16,21%.
2.	Muchamad Sobri Sungkar dan M Taufik Qurohman (2021)	Prediksi Kelulusan Mahasiswa pada Matakuliah Arsitektur Sistem Komputer	Algoritma C5.0	Penggunaan algoritma C5.0 dalam memprediksi kelulusan pembelajaran mahasiswa pada matakuliah Arsitektur Sistem Komputer memberikan hasil yang cukup baik dengan tingkat akurasi mencapai 93,33%.
3.	Fida Nafisah Giustin, Betha Nurina Sari, dan Tesa Nur Padilah(2022)	Prediksi Hasil Belajar Mahasiswa pada Matakuliah Kalkulus	Algoritma C5.0	Hasil yang diperoleh dengan menggunakan algoritma C5.0 memberikan hasil yang cukup baik mencapai 95%, menunjukkan algoritma tersebut dapat digunakan sebagai alat bantu untuk memprediksi

No	Nama Penulis	Objek	Metode	Hasil
				hasil belajar mahasiswa dengan cukup akurat.
4.	Dyah Ayu Dewi, Imam Cholissodin dan Sutrisno(2019)	Penyimpangan Tumbuh Kembang Anak	Algoritma C5.0	Menunjukkan bahwa algoritma C5.0 dapat digunakan untuk membangun model klasifikasi yang akurat dalam memprediksi penyimpangan tumbuh kembang anak, model yang dihasilkan memiliki tingkat akurasi yang tinggi.
5.	Salsabila Citra Esananda, Budi Nugroho dan Fetty Tri Anggraeny(2021)	Menentukan Prestasi Akademik Siswa	Algoritma C5.0, fase *Boosting*	Hasil penelitian menunjukkan bahwa implementasi fase *boosting* pada algoritma C5.0 dapat meningkatkan performa atau hasil akurasi dari model dalam memprediksi prestasi akademik siswa, penerapan fase *boosting* ini memiliki tingkat akurasi yang lebih tinggi dibandingkan dengan model yang dibangun tanpa menggunakan fase *boosting*.
6.	Dewi Wulansari	Prediksi Ketepatan Kelulusan Mahasiswa Penerima KIP-Kuliah Merdeka	Algoritma C5.0	Algoritma C5.0 dapat digunakan untuk memprediksi Ketepatan Kelulusan mahasiswa pada penerima KIP-Kuliah Merdeka dengan akurasi sebesar 97%.

2.2 Dasar Teori

2.2.1 Data Mining

Data mining adalah suatu proses untuk menemukan pola atau pengetahuan yang bermanfaat dari kumpulan data besar. *Data mining* mencakup berbagai teknik dan algoritma yang dapat digunakan untuk mengidentifikasi pola dalam data seperti *clustering, association rule mining, decision tree*, dan lain seba*gain*ya. (Han, Kamber, & Pei, 2011).

Penggunaan *data mining* umumnya melibatkan proses pengolahan data untuk mengekstraksi informasi yang berharga dari data yang besar dan kompleks. Dengan menerapkan proses tersebut *data mining* akan menghasilkan sebuah informasi yang baru.

2.2.2 Metode Prediksi Data Mining

Prediksi dalam data mining adalah suatu teknik yang digunakan untuk memprediksi nilai atau kelas dari suatu data berdasarkan pola-pola yang terdapat pada data tersebut. Teknik prediksi dalam data mining ini dapat dilakukan dengan menggunakan berbagai metode seperti *decision tree*, neural network, dan regression. (Han, Kamber, & Pei, 2011).

Metode prediksi dalam data mining termasuk ke dalam kelompok *supervised learning*. *Supervised learning* adalah teknik pembelajaran mesin di mana model dipelajari dari data yang telah diberi label atau jawaban yang benar, sehingga dapat memprediksi label atau jawaban yang tepat untuk data baru. (Witten, Frank, & Hall 2016).

2.2.3 Algoritma C5.0

Menurut (Quinlan, 1993) Algoritma C5.0 adalah salah satu algoritma *decision tree* yang digunakan untuk klasifikasi dan prediksi. Algoritma ini merupakan pengembangan dari algoritma *CART* (*Classification and Regression Trees*) yang lebih cepat dan memiliki performa yang lebih baik (Amalda, 2019). Algoritma C5.0 menggunakan teknik pengklasifikasian dengan membangun pohon keputusan (*decision tree*) berdasarkan data latih, kemudian digunakan untuk memprediksi kelas dari data uji (Darmawan & Sari,2019).

Algoritma C5.0 bekerja dengan menghitung nilai *entropy* dan *gain ratio* setiap atribut yang ada dalam data, dan memilih atribut yang memiliki nilai *gain ratio* paling tinggi untuk dijadikan *node* dalam *decision tree*. Algoritma ini juga menggunakan teknik *pruning* untuk mengurangi *overfitting* yang bisa terjadi pada *decision tree* yang terlalu kompleks. *Pruning* dilakukan dengan menghilangkan beberapa aturan dari pohon keputusan yang dihasilkan oleh algoritma C5.0, sehingga menghasilkan pohon keputusan yang lebih sederhana dan mudah dipahami.

Salah satu hal menarik dari algoritma C5.0 adalah kemampuannya dalam membangun pohon keputusan yang sangat efisien dengan memilih atribut yang paling signifikan pada setiap level dari pohon tersebut. Selain itu, algoritma C5.0 juga mampu menangani data yang tidak lengkap.

Langkah-langkah algoritma C5.0 yaitu:

1. Persiapkan data

Langkah pertama dalam algoritma C5.0 adalah mempersiapkan data yang akan digunakan, data ini merupakan data awal yang masih berekstensi excel yang kemudian diubah sesuai dengan kebutuhan berupa file ekstensi csv. Serta memastikan dataset yang mencakup atribut yang ingin diprediksi. Dataset ini terdiri dari data yang terstruktur.

2. Memilih atribut untuk membangun *decision tree*

   Salah satu langkah penting dalam algoritma C5.0 adalah memilih atribut yang paling informatif untuk membangun *decision tree.* Dalam algoritma ini proses pemilihan atribut menggunakan *information gain* dimana nilai informasi *gain* tertinggi akan dipilih sebagai *parent* bagi *node* selanjutnya. Untuk memilih atribut ini dapat menggunakan rumus *entropy*, *gain* dan *gain ratio*. Berikut rumus *entropy*, *gain* dan *gain ratio*.

   Menghitung *Entropy* menggunakan rumus pada persamaan 2.1.

$$\boldsymbol{Entropy(S)} = -\sum_{i=1}^{n} \mathbf{P_i \log_2 P_i} \quad \text{.....(2.1)}$$

   Keterangan:

   S : Himpunan kasus atau jumlah kasus

   n : Jumlah dari partisi S

   p_i : Rasio S_i terhadap S

   Menghitung *Gain* menggunakan rumus pada persamaan 2.2.

$$\boldsymbol{Gain(S,A) = Entropy(S) - \sum_{i=1}^{n} \frac{|S_i|}{|S|} \; x \; Entropy(S_i)} \quad \text{.....(2.2)}$$

   Keterangan:

   S : Himpunan kasus atau jumlah kasus

A : Variabel atau atribut yang digunakan

n : Jumlah partisi pada variabel A

|Si| : Jumlah kasus pada partisi ke-i

|S| : Jumlah kasus dalam S

Si : Himpunan kasus pada partisi ke-i

Menghitung *Gain ratio* menggunakan rumus pada persamaan 2.3.

$$\textbf{\textit{Gain Ratio}} = \frac{\textbf{\textit{Gain}}(\boldsymbol{S},\boldsymbol{A})}{\sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} \textbf{Entropy}(\boldsymbol{S_i})} \quad \text{..... (2.3)}$$

Keterangan:

Gain(S, A) : Nilai *gain* dari variabel

$\sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} \textbf{Entropy}(\boldsymbol{S_i})$: Banyaknya nilai *entropy* dalam suatu variabel

3. Membuat objek *DecisionTreeClassifier*

Setelah memilih atribut, langkah selanjutnya adalah membuat objek *DecisionTreeClassifier* yang akan digunakan untuk membangun pohon keputusan. Objek ini diimplementasikan menggunakan *library* atau *framework* seperti *scikit-learn*.

4. Melatih model

Setelah objek *DecisionTreeClassifier* dibuat, model akan dilatih menggunakan dataset yang dipersiapkan sebelumnya dan memasukkan data ke dalam model, mengoptimalkan parameter pohon keputusan berdasarkan fungsi objektif yang digunakan untuk menghasilkan model pohon keputusan dan meningkatkan akurasi dan kemampuan prediksi model.

5. Melakukan prediksi pada data uji

   Setelah model dilatih, langkah selanjutnya melakukan prediksi pada data uji. Data uji dimasukkan ke dalam model, dan model akan menghasilkan prediksi berdasarkan aturan yang ada pada pohon keputusan.

6. Evaluasi model

   Tahap terakhir adalah evaluasi model, untuk mengukur kinerja model dalam melakukan prediksi. Salah satu metrik yang digunakan adalah *confusion matrix* dimana akan memberikan informasi tentang prediksi yang benar dan salah.

2.2.4 Confusion Matrix

Menurut (Han, Kamber, & Pei, 2012). *Confusion matrix* adalah alat evaluasi yang digunakan dalam metode prediksi dan klasifikasi untuk mengukur kinerja model dalam memprediksi kelas target. Ada beberapa istilah yang digunakan dalam *confusion matrix* yaitu *TP*(*True Positive*) dan *TN*(*True Negative*) memberikan informasi jika klasifikasi benar serta *FP*(*False positive*) dan *FN*(*False Negative*) memberikan informasi Ketika klasifikasi salah.

Confusion matrix digunakan dalam pemodelan prediksi dan klasifikasi. Memungkinkan kita untuk memvisualisasikan dan mengukur sejauh mana model klasifikasi kita memprediksi dengan benar dan salah. *Confusion matrix* juga digunakan untuk mengevaluasi performa model klasifikasi atau prediksi. Ini memberikan gambaran tentang seberapa baik model kita dalam memprediksi kelas target. Berikut adalah representasi umum dari *confusion matrix*.

	Predict class		
Actual class		Positive	Negative
	Positive	TP	FN
	Negative	FP	TN

Gambar 2. 1 Confusion matrix dua kelas

Berdasarkan gambar 2.1 pada *confusion matrix* terdapat *predict class* dan *actual class*. Di mana *predict class* adalah kolom pada *confusion matrix* yang menunjukkan prediksi kelas oleh model sedangkan *actual class* adalah baris pada *confusion matrix* yang menunjukkan kelas sebenarnya dari data yang memiliki nilai true dan false. Berikut penjelasan dari *predict class* dan *actual class* pada *confusion matrix*.

1. True Positive (TP) : Model memprediksi positif dan secara aktual benar.
2. True Negative (TN) : Model memprediksi negatif dan secara aktual benar.
3. False Positive (FP) : Model memprediksi positif namun secara aktual salah.
4. False Negative (FN) : Model memprediksi negatif dan secara aktual salah.

Dari *confusion matrix* maka dapat menghitung nilai akurasi yang menampilkan seberapa akurat model klasifikasi yang telah dibuat. Berikut persamaan untuk melakukan klasifikasi secara benar (Anggreany, 2022).

Akurasi (2 kelas)= $\frac{TP+TN}{TP+TN+FP+FN}$. 100%…….(2.4)