BAB II
LANDASAN TEORI

2.1 Tinjauan Pustaka

Dalam melakukan penelitian penting adanya melakukan tinjauan pustaka *(literatur review)* yang merupakan sebuah aktivitas untuk meninjau atau mengkaji kembali literature yang telah dipublikasikan oleh akademisi atau peneliti lain sebelumnya terkait topik yang akan diteliti. Dan agar menghindari melakukan penelitian yang sama dengan penelitian yang sudah dilakukan sebelumnya . berikut ini merupakan beberapa sumber penelitian terdahulu :

Tabel 2. 1 Tinjauan Pustaka

1.	Judul	Implementasi Algoritma *Decision Tree* untuk Klasifikasi Produk Laris
	Penulis	Asmaul Husnah Nasrullah
	Tahun	2021
	Tujuan Penelitian	Menguji akurasi dari algoritma C4.5 dalam melakukan klasifikasi produk laris (data privat)
	Permasalahan	Bagaimana menentukan produk laris di perusahaan dagang Cipta Karya Gorontalo ?
	Subjek Penelitian	PT Cipta Karya Gorontalo
	Metode Penelitian	Algoritma *Decision Tree*
	Hasil Penelitian	Hasil akurasi model klasifikasi produk laris menggunakan *Decision Tree* C4.5 yang diperoleh dari penelitian ini adalah sebesar 90% dan nilai AUC 0.709 dimana nilai ini termasuk dalam *Good Classification* . Sehingga dapat disimpulkan bahwa model klasifikasi data mining Algoritma *Decision Tree* C4.5 akurat dalam menentukan klasifikasi untuk produk laris.

2.	Judul	Implementasi Algoritma C5.0 untuk Menentukan Pelanggan Potensial di Kantor Pos Cimahi
	Penulis	Nisa Hanum Harani, Fanny Shafira Damayanti
	Tahun	2021
	Tujuan penelitian	Mengetahui cara identifikasi pelanggan
	Permasalahan	Bagaimana cara mengetahui pelanggan mana yang memiliki potensial dan mana yang tidak memiliki potensial?
	Subjek Penelitian	Kantor Pos Cimahi
	Metode Penelitian	Algoritma *Decision Tree*
	Hasil Penelitian	Berdasarkan hasil yang didapat, Implementasi Algoritma C5.0 untuk menentukan Pelanggan Potensial di Kantor Pos Cimahi yaitu : a. Sistem yang telah dibangun dapat membantu bagian penjualan di kantor Pos Cimahi untuk menentukan pelanggan korporat yang potensial dan tidak sehingga pelanggan yang telah ditentukan sebagai pelanggan potensial dapat diperlakukan secara khusus agar pelanggan tersebut tetap menggunakan jasa dari Kantor Pos Cimahi. b. Hasil akurasi yang diperoleh dari data transaksi periode bulan Januari – Oktober 2020 yaitu sebesar 96%
3	Judul	Implementasi Algoritma C5.0 Pada Analisa Data Potensi Pertanian dan Peternakan
	Penulis	Nurnia Zamasi
	Tahun	2021
	Tujuan Penelitian	Mengetahui cara melakukan pendataan potensi pertanian dan peternakan dengan baik.

	Permasalahan	Bagaimana cara meningkatkan data potensi pertanian dan peternakan ?
	Subjek Penelitian	Unit pembinaan pemeliharaan tanaman (UPPT)
	Metode Penelitian	Algoritma *Decision Tree*
	Hasil Penelitian	Dapat mengetahui potensi pertanian dan peternakan disetiap wilayah mana yang lebih berpotensi pada setiap tahunnya. Algoritma C5.0 dapat memberikan informasi berupa rule Analisa untuk menggambarkan proses yang terkait dengan data potensi pertanian dan peternakan. Aplikasi *RapidMiner Classification Decision Tree* digunakan sebagai aplikasi pendukung keputusan dan pengujian atas hasil yang didapatkan secara manual, yang menghasilkan sebuah pohon keputusan. Dari pohon keputusan inilah akan menghasilkan sebuah aturan-aturan yang dapat membantu pihak UPPT Biru-biru dalam menganalisa potensi pertanian dan peternakan di setiap wilayahnya serta mudah dipahami oleh pengguna aplikasi.
4	Judul	Algoritma Klasifikasi *Decision Tree* Untuk Rekomendasi Buku Berdasarkan Kategori Buku
	Penulis	Mawadatul Maulidah, Windu Gata, Rizki Aulianita, Cucu Ika Agustyaningrum (STMIK Nusa Mandiri)
	Tahun	2020
	Tujuan Penelitian	Mendapatkan model algoritma terbaik untuk rekomendasi buku berdasarkan prediksi kategori buku pada dataset *goodreads book*

	Permasalahan	Bagaimana cara mendapatkan rekomendasi buku yang relevan ?
	Subjek Penelitian	Dataset *goodreads book*
	Metode Penelitian	Metode dasar Algoritma *Decision Tree ,* dan algoritma pembanding *K-Nearest Neighbor (K-NN), Naïve Bayes, Random Forest ,* dan *Support Vector Classifier (SVC).*
	Hasil Penelitian	*Decision Tree* memiliki akurasi paling tinggi diantara metode yang dikomparasikan yaitu sebesar 99,95% *precision* sebesar 100%, recall sebesar 96%, fl-score sebesar 98% dengan rata-rata kesalahan sebesar 0.05 dan AUC sebesar 99,96%, diikuti oleh algoritma *Random Forest, K-Nearest Neighbor (K-NN), Support Vector Classifier (SVC).* Dan *Naïve Bayes* yang memiliki akurasi paling rendah. Dengan demikian hasil evaluasi menggunakan curva ROC/AUC yaitu algoritma klasifikasi *Decision Tree* bernilai 99,96% dengan tingkat diagnose *excellent classification.* Hal tersebut dapat membuktikan bahwa Algoritma Decision Tree dapat digunakan sebagai rekomendasi buku untuk memprediksi kategori buku pada *goodreads book.*
5.	Judul	Implementasi Algoritma Decision Tree untuk Klasifikasi Data Peserta Didik
	Penulis	Imam Sutoyo
	Tahun	2018
	Tujuan Penelitian	Menghasilkan model klasifikasi untuk keperluan pengelompokan peserta didik

	Permasalahan	Bagaimana cara mengelompokkan peserta program Pendidikan agar kegiatan pembelajaran dapat disesuaikan?
	Subjek Penelitian	Dataset peserta didik
	Metode penelitian	Algoritma *Decision Tree*
	Hasil Penelitian	Berdasarkan percobaan didapatkan bahwasannya algoritma *Decision Tree* yang diuji coba menunjukkan hasil yang memuaskan , baik C4.5 maupun *Random Forest* telah menunjukkan kinerja yang tinggi dalam ukuran akurasi, yakni 97,63% untuk C4.5 dan 95,13% untuk *Random Forest* . Berdasarkan ukuran akurasi ini, C4.5 menggungguli *Random Forest* sebesar 2,5% oleh karena itu model rule yang dihasilkan oleh C4.5 digunakan sebagai dasar pengembangan prototipe aplikasi klasifikasi.
6.	Judul	Implementasi Algoritma C5.0 pada Penilaian Kinerja Pegawai Negeri Sipil
	Penulis	Putu Wirya Kastawan, Dewa Made Wiharta, I Made Sudarma
	Tahun	2018
	Tujuan penelitian	Mempermudah pengelompokan data yang besar pada penilaian kepegawaian
	Permasalahan	Pengolahan penilaian dan evaluasi kinerja aparatur yang masih dilakukan secara manual
	Subjek Penelitian	Badan Kepegawaian Daerah Provinsi Bali
	Metode Penelitian	Algoritma *Decision Tree*
	Hasil penelitian	Berdasarkan hasil evaluasi yang telah dilakukan dapat diambil beberapa kesimpulan sebagai berikut :

		a. Algoritma C5.0 dapat memproses data kinerja pegawai menjadi sebuah pohon keputusan serta aturan-aturan yang berguna sebagai sebuah masukan. Hasil yang diperoleh bisa dikembangkan menjadi sebuah system penentu keputusan sehingga dapat dijadikan untuk membantu dalam menentukan keputusan kinerja pegawai. b. Secara umum berdasarkan hasil evaluasi untuk data staf dengan data *training* sebanyak 184 data didapat tingkat akurasi sebesar 96,08% dimana salah satu faktoryang mempengaruhi akurasi adalah data kinerja tidak memenuhi syarat yang masih kecil, hal ini bisa ditingkatkan dengan menambah jumlah data *training* yang terkait dengan data tersebut.

2.2 *Data Mining*

Data Mining merupakan proses menemukan hubungan, pola dan kecenderungan dalam sekumpulan besar data yang tersimpan di dalam penyimpanan menggunakan Teknik pengenalan pola serta Teknik statistic dan matematika . *Data Mining* juga adalah suatu proses dalam menemukan pola yang menarik dan pengetahuan dari dalam data dalam jumlah besar, sumber data dapat berupa database, web, data Gudang, serta data dari penyimpanan informasi lainnya atau data yang dialirkan ke system secara dinamis .

Ada pula yang berpendapat bahwa, *data mining* adalah kegiatan mengekstrak informasi atau pengetahuan (*knowledge)* penting dari suatu set data berukuran besar dengan menggunakan Teknik tertentu, dimana informasi atau *knowledge* yang

dihasilkan dari *data mining* bisa dipakai untuk memperbaiki pengambilan keputusan (Apriori, 2021).

Proses pengolahan data dalam *data mining* dibutuhkan algoritma-algoritma untuk melakukan ekstraksi data menjadi informasi atau pengetahuan. Penggunaan algoritma pada *data mining* diklasifikasi berdasarkan masing-masing peranan *data mining* . Pada peranan estimasi dan prediksi, algoritma yang banyak digunakan seperti *Linear Regression, Support Vector Machine, Neural Network,* dan sebagainya, sedangkan algoritma yang banyak digunakan pada peranan klasifikasi adalah *K-Nearest Neighbos (k-NN), Naïve Bayes, ID3, C4.5, CART,* dan lainnya. Pada peranan clustering algoritma yang biasa digunakan seperti *K-Means, Fuzzy C-Means, K-Medoid, Self-Organization Map (SOM),* dan lainnya, sedangkan pada peranan asosiasi beberapa algoritma yang banyak digunakan seperti *Eclat, FP-Growth, Apriori, Chi Square, Coeffcient of Correlation,* dan masih banyak lagi algoritma yang dapat digunakan untuk peranan asosiasi .

2.2.1 Tahapan *Data Mining*

Tahapan *data mining* merupakan proses penemuan pengetahuan dari data menurut (Apriori, 2021) sebagai berikut:

1. **Pembersihan Data *(Data Cleaning)***

   Pembersihan data yaitu tahapan untuk menghilangkan noise dan data yang tidak koefisien. Pembersihan data seperti penghapusan atribut data yang tidak dibutuhkan.

2. **Integrasi Data *(Data Integration)***

   Integrasi data merupakan proses kombinasi bebrapa sumber data. Tahapan ini melakukan penggabungan data dari berbagai sumber data yang berbeda.

3. **Seleksi Data *(Data Selection)***

   Seleksi data adalah proses pengambilan data yang berkaitan dengan tugas analisis dari basis data. Tahapan ini melakukan Teknik pengurangan representasi dari data dan meminimalkan hilangnya informasi data, mulai dari pengurangan atribut dan kompresi data.

4. **Transformasi Data *(Data Transformation)***

   Transformasi data merupakan tahapan dimana data diubah dan dikonsolidasikan kedalam bentuk yang sesuai untuk penambangan dengan melakukan ringkasan atau penggabungan operasi.

5. **Penambangan Data *(Data Mining)***

   Penambangan data merupakan tahapan penting dalam proses penemuan pengetahuan dari dalam data dengan menggunakan Teknik dan algoritma tertentu.

6. **Evaluasi Pola *(Pattern Evaluation)***

   Tahapan ini merupakan proses untuk mengidentifikasi pola menarik yang mewakili pengetahuan berdasarkan Langkah-langkah yang diberikan.

7. **Representasi Pengetahuan *(Knowledge Presentation)***

   Knowledge Presentation merupakan tahapan dalam penemuan pengetahuan yang direpresentasikan secara visual kepada pengguna untuk membantu dalam memahami hasil data mining.

2.3 Klasifikasi

Klasifikasi dapat didefinisikan sebagai *supervise learning* yang membutuhkan label dalam prosesnya untuk mengekstrak model yang digunakan untuk memprediksi suatu label. Proses dalam klasifikasi adalah untuk menemukan property-properti yang sama dalam suatu himpunan obyek pada suatu *database* kemudian diklasifikasikan ke dalam kelas-kelas yang berbeda sesuai dengan model klasifikasi yang dipilih. Proses klasifikasi bertujuan untuk menacari model dari *training set* yang memisahkan atribut kedalam kategori atau kelas yang sesuai, kemudian model tersebut digunakan untuk klasifikasi atribut yang kelasnya belum diketahui sebelumnya (Hermawati, n.d.)

2.4 *Decision Tree*

Decision Tree atau pohon keputusan merupakan salah satu teknik klasifikasi. *Decision Tree* adalah *top-down* pohon rekursif dari algoritma induksi, yang menggunakan ukuran seleksi atribut untuk memilih atribut yang diuji. Dengan pohon keputusan, manusia dapat dengan mudah melihat hubungan antara fakto-faktor yang mempengaruhi suatu masalah. Pohon keputusan ini juga dapat menganalisa nilai suatu informasi yang terdapat dalam suatu alternatif pemecahan

masalah. Pohon keputusan adalah salah satu metode klasifikasi yang mudah untuk diaplikasikan oleh manusia. *Decision Tree* digunakan untuk mempelajari klasifikasi dan prediksi pola dari data dan menggambarkan relasi dari variable atibut x dan variable target y dalam bentuk pohon . *Decision Tree* adalah struktur menyerupai *flowchart* dimana setiap internal node (node yang bukan *leaf* atau bukan node terluar) merupakan pengujian terhadap variable attribute, tiap cabangnya merupakan hasil dari pengujian tersebut, sedangkan node terluar yakni *leaf* menjadi labelnya (Kastawan et al., 2018)

2.5 Algoritma C5.0

Algoritma C5.0 merupakan penyempurnaan dari algoritma ID3 dan C4.5. Dalam proses pembentukan pohon keputusan nilai informasi *gain* tertinggi akan terpilih sebagai *root* bagi *node* selanjutnya. Algoritma ini dimulai dengan semua data yang dijadikan akar dari pohon keputusan sedangkan atribut yang dipilih akan menjadi pembagi bagi sampel tersebut (Kastawan et al., 2018). Adapun rumus untuk mengetahui *entropy* keseluruhan dan *entropy* setiap atribut :

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

Keterangan :

S = Himpunan Kasus

n = Jumlah partisi S

pi = probabilitas yang didapat dari jumlah kelas dibagi total kasus

Setelah menghitung nilai entropy, pemilihan atribut dilakukan dengan menggunakan information Gain. Untuk menghitung gain, yang bisa dihitung dengan formula sebagai berikut :

$$Gain(S, A) = Entropy\ (S) - \sum_{i=1}^{n} \frac{|Si|}{|S|} * Entropy\ (Si) \quad (2)$$

Keterangan :

S = Himpunan kasus

A = Atribut

n = Jumlah atribut

|Si| = Jumlah partisi ke-i

|S| = Jumlah kasus dalam S

Kemudian lakukan perhitungan pada rumus ke tiga :

$$Gain\ Ratio = \frac{Gain\ (S,A)}{\sum_{i=1}^{n} Entropy\ (Si)} \quad (3)$$

Gain (S,A) = Nilai gain dari variable

$\sum_{i=1}^{n} Entropy\ (Si)$ = Banyaknya nilai *entropy* dalam suatu variable

2.5.1 Tahapan Algoritma C5.0

C5.0 adalah algoritma pembelajaran mesin berbasis pohon keputusan yang dikembangkan oleh Ross Quinlan. Algoritma ini dapat digunakan untuk klasifikasi dan regresi pada data yang terstruktur (Kastawan et al., 2018). Berikut adalah tahapan-tahapan algoritma C5.0 :

1. **Pengumpulan Data**

   Tahapan ini melibatkan pengumpulan data yang akan digunakan untuk melatih model.

2. **Preprocessing Data**

   Data yang dikumpulkan kemudian diproses untuk menghapus nilai yang hilang dan memperbaiki data yang salah atau tidak konsisten.

3. **Pemilihan Fitur**

   Tahap ini melibatkan pemilihan fitur yang paling penting dalam data untuk melatih model. Hal ini dapat dilakukan menggunakan berbagai metode seperti *Entropy, Information Gain*, dan *Gain Ratio*

4. **Pembuatan Model**

   Setelah fitur yang paling penting telah dipilih, model pohon keputusan dibangun berdasarkan data pelatihan dan fitur yang telah dipilih. Model pohon keputusan ini digunakan untuk melakukan klasifikasi atau regresi pada data uji.

5. Pruning Model

Setelah model dibangun, model tersebut dapat dipangkas untuk menghindari overfitting pada data pelatihan. Untuk menghitung nilai *pruning* digunakan rumus :

$$e = \frac{r + \frac{z^2}{2n} + z\sqrt{\frac{r}{n} - \frac{r^2}{n} + \frac{z^2}{4n^2}}}{1 + \frac{z^2}{n}} \quad (4)$$

Dimana :

r = nilai perbandingan *error rate*

n = total sample

z = Φ – 1 (c)

c = confidence level

6. Evaluasi Model

Tahap ini melibatkan evaluasi model untuk mengetahui seberapa baik model tersebut melakukan klasifikasi atau regresi pada data uji. Evaluasi dapat dilakukan dengan menggunakan *confusion matrix* seperti akurasi, presisi, recall, dan F1-score.

2.6 Cross Validation

Cross validation merupakaan teknik validasi, yang dimana teknik ini digunakan untuk melihat nilai akurasi algoritma yang digunakan dari berbagai kasus atau dengan berbagai model dataset yang digunakan. Pada dasarnya *cross validation* adalah teknik validasi silang yang membagi sebuah dataset menjadi dua bagian yang mana dinamakan *training data* dan *test data.* Hal ini dilakukan dengan membagi data dalam berbagai partisi. Itulah mengapa *cross validation* juga sering disebut dengan *k-fold cross validation* karena dimana percobaan dilakukan sebanyak nilai *k* (Abdul Muiz Khalimi, 2022). Berikut adalah contoh table dari cara kerja *cross validation*

Tabel 2. 2 Cross Validation

Percobaan 1	Test	Train	Train	Train	Train
Percobaan 2	Train	Test	Train	Train	Train
Percobaan 3	Train	Train	Test	Train	Train
Percobaan 4	Train	Train	Train	Test	Train
Percobaan 5	Train	Train	Train	Train	Test

Percobaan diatas adalah contoh ilustrasi dari 5-*fold cross validation* yang artinya adalah melakukan percobaan sebanyak 5 kali tahapan.

Percobaan 1, yaitu menjadikan bagian partisi pertama menjadi data testing dan partisi lainnya menjadi data training.

Percobaan 2, yaitu menjadikan bagian partisi kedua menjadi data testing dan partisi lainnya menjadi data training.

Percobaan 3, yaitu menjadikan bagian partisi ketiga menjadi data testing dan partisi lainnya menjadi data training dan begitu seterusnya.

Dari 5 hasil percobaan ini, kita akan catat nilai evaluasi performa dari model tersebut dengan menggunakan confussion matrix. Kemudian tentukan nilai rata-

rata dari setiap percobaan. Maka disitu akan ditemukan percobaan mana yang dapat dijadikan acuan dari penggunaan suatu model algoritma yang telah terpilih.

2.7 Confusion Matrix

Confusion matrix juga sering disebut *error matrix* .Pada dasarnya *confusion matrix* memberikan informasi perbandingan hasil klasifikasi yang dilakukan oleh system (model) dengan hasil klasifikasi sebenarnya. *Confusion matrix* berbentuk table matriks yang menggambarkan kinerja model klasifikasi pada serangkaian data uji yang nilai sebenarnya diketahui (Kuncahyo Setyo Nugroho, 2019). Adapun table dibawah ini merupakan *confusion matrix* dengan 4 kombinasi nilai prediksi dan nilai actual yang berbeda.

Tabel 2. 3 Confusion Matrix

		Actual Values	
		1 (Positive)	0 (Negative)
Predicted Values	1 (Positive)	TP (True Positive)	FP (False Positive Type I Error
	0 (Negative)	FN (False Negative) Type II Error	TN (True Negative)

Dimana *TP* adalah *True Positive* yaitu jumlah data positif yang terklasifikasi dengan benar oleh system, *TN* adalah *True Negative* yaitu jumlah data negative yang terklasifikasi dengan benar oleh system, *FN* adalah *False Negative yaitu* jumlah data negative namun terklasifikasi salah oleh system dan *FP* adalah *False Positive,* yaitu jumlah data positif namun terklasifikasi salah oleh system.