BAB II

LANDASAN TEORI

2.1 Tinjauan Pustaka

2.1.1 Klasifikasi

Klasifikasi merupakan salah satu metode dalam data mining yang digunakan untuk mengelompokkan data kedalam kelas atau kelompok tertentu, Model ini berfungsi untuk memperkirakan nilai dari suatu data berdasarkan pola yang telah dipelajari dari data lain yang sudah diketahui hasilnya. Tujuan utama dari model klasifikasi adalah untuk memprediksi nilai suatu variabel target yang belum diketahui dengan memanfaatkan informasi dari variabel-variabel lain yang telah tersedia (Nugraha, Sunandar, & Julian, 2022).

2.1.2 Penilangan

Tilang adalah sanksi berupa denda yang diberikan oleh polisi kepada pengguna jalan yang melanggar aturan lalu lintas, seperti tidak mematuhi rambu, tidak memakai helm, atau tidak membawa surat kendaraan. Meskipun aturan telah diatur dalam Undang-Undang Nomor 22 Tahun 2009, masih banyak pelanggaran yang terjadi. Oleh karena itu, tilang digunakan untuk menegakkan hukum dan meningkatkan kesadaran masyarakat demi keselamatan di jalan (Yusri, 2020).

2.1.3 Algoritma C4.5

Algoritma C4.5 merupakan salah satu metode pohon keputusan yang banyak digunakan karena berbagai keunggulannya. Salah satu kelebihan utama algoritma ini adalah kemampuannya dalam mengolah data numerik maupun diskrit, serta tetap dapat bekerja meskipun terdapat nilai atribut yang hilang dalam data. C4.5 menghasilkan aturan-aturan keputusan yang mudah dipahami, dan memiliki kecepatan pemrosesan yang relatif tinggi dibandingkan dengan algoritma sejenis lainnya (Matondang, Ramadhan Nasution, Armansyah, & Furqan, 2024)**.**

Algoritma C4.5 menggunakan dua jenis data utama sebagai input, yaitu training sample dan samples. *Training sample* adalah data contoh yang sudah diketahui hasil atau labelnya, dan digunakan untuk membangun model pohon keputusan. Data ini berfungsi sebagai dasar pembelajaran algoritma agar dapat mengenali pola. Sementara itu, *samples* adalah data yang berisi nilai-nilai atribut yang akan digunakan sebagai parameter dalam proses klasifikasi, yaitu untuk menentukan ke kategori mana data tersebut termasuk (Setio, Saputro, & Bowo Winarno, 2020).

Berikut ini merupakan prinsip dasar kerja algoritma C4.5 dalam membentuk pohon keputusan (decision tree).

Tahapan dari proses algoritme C4.5 diuraikan sebagai berikut:

1) Proses dimulai dengan mempersiapkan data *training*.
2) Menghitung nilai *entropy*. Entropy merupakan ukuran seberapa beragam atau tidak pastinya data. Jika semua data dalam satu kelompok memiliki hasil yang sama, maka entropy-nya rendah. Sebaliknya, jika hasilnya beragam, entropy-nya tinggi. Nilai Entopy dihitung dengan rumus yang ditulis sebagai

   $$\text{Entropy} = -\sum_{i=1}^{k} p_i * log_2 p_i$$

   dengan S adalah himpunan kasus, pi adalah probabilitas yang diperoleh dari sum (ya) dibagi dengan total kasus.
3) Menghitung nilai gain. Gain digunakan untuk menentukan atribut terbaik dalam membagi data saat membangun pohon keputusan atau dengan kata lain gain merupakan tingkat pengaruh suatu atribut terhadap keputusan atau ukuran efektifitas suatu variabel dalam mengklasifikasikan data. Gain dihitung dengan rumus yang ditulis sebagai

   $$\text{Gain(S,A)} = \text{Entropy (S)} - \sum_{i=1}^{k} \left|\frac{S_i}{S}\right| * Entropy\ (S_i)$$

   dengan S adalah himpunan kasus, A adalah atribut, $|Si|$ adalah jumah kasus pada partisi ke i, dan $|S|$ adalah jumlah kasus dalam S. Pada algoritma C4.5,

nilai gain digunakan untuk menentukan variable mana yang menjadi node dari suatu pohon keputusan. Suatu variabel yang memiliki gain tertinggi akan dijadikan node di pohon keputusan.

4) Menghitung nilai split info dengan rumus

$$SplitInfo(S,A) = -\sum_{j=1}^{k} \frac{S_j}{S} X \log_2 \frac{S_j}{S}$$

dengan S adalah ruang sample, A adalah atribut, dan Sj adalah jumlah sample untuk atribut ke j.

5) Menentukan nilai gain ratio dengan rumus yang ditulis

$$\text{GainRatio (S,A)} = \frac{Gain(S,A)}{Split(S,A)}$$

dengan Gain(S,A) adalah information gain pada atribut (S,A), A adalah atribut, dan Split(S,A) adalah split information pada atribut (S,A).

Nilai gain ratio tertinggi akan digunakan sebagai atribut akar. Dengan demikian akan terbentuk pohon keputusan sebagai node 1.

6) Mengulangi proses ke-2 hingga semua cabang memiliki kelas yang sama. Proses percabangan akan berhenti apabila
    1) semua kasus dalam simpul n mendapat kelas yang sama;
    2) tidak ada variabel independen di dalam kasus yang dipartisi lagi;
    3) tidak ada kasus di dalam cabang yang kosong.

2.2 Penelitian Terdahulu

Dalam penelitian ini peneliti mengambil beberapa penelitian terdahulu untuk menjadi referensi dalam penelitian ini. Berikut beberapa hasil penelitian yaitu:

Penerapan Data Mining Algoritma C4.5 Untuk Klasifikasi Hasil Pengujian Kendaraan Bermotor (Andarista & Jananto, 2022). Pada penelitian ini menggunakan software Rapidminer versi 9.10.000 dengan menggunakan 424 record. Hasil dari penelitian ini dengan menggunakan perhitungan manual menggunakan excel dan tools Rapidminer dengan komposisi pembagian data 80% data training dan 20% data testing menghasilkan pohon keputusan atribut Kedalaman Alur Ban sebagai root node dengan nilai gain sebesar 0,24 dan tingkat

keakurasian sebesar 94,12% menghasilkan 15 rule/aturan. Dari perhitungan manual menggunakan excel dan menggunakan tools Rapidminer menghasilkan nilai akurasi yang sama dan pohon keputusan yang sama. Untuk mengelompokan faktor yang mempengaruhi kendaraan bermotor tidak lulus uji, dengan menggunakan metode klasifikasi Algoritma C4.5 yang diharapkan dapat membantu mengetahui prediksi hasil pengujian kendaraan bermotor yang dilihat dari faktor yang mempengaruhinya.

Analisis Klasifikasi Kasus Tindak Pidana Pencurian Dengan Pohon Keputusan Menggunakan Algoritma C4.5 (Studi Kasus Polsek Telanaipura) (Puspitorini, Simorangkir, Dwi Larasati, Yanuar, & Yandri, 2023). Pada penelitian ini, pengelompokan terhadap kasus tindak pidana serta menetukan pelanggaran pasal KUHP (kitab Undang-undang Hukum Pidana) dilakukan dengan data mining klasifikasi. Representasi hasil klasifikasi akan disajikan dalam bentuk pohon keputusan (decision tree) yang dibangun dengan algoritma C4.5. aturan (rule) yang terbentuk dapat digunakan untuk menentukan kelas data tindak pidana yang baru.

Perbandingan Kinerja Algoritma C4.5 Dan Naive Bayes Untuk Klasifikasi Penerima Beasiswa (Anam & Santoso, 2018)**.** Hasil penelitian menunjukkan tingkat akurasi model algoritma C4.5 sebesar 96.40% lebih baik dari tingkat akurasi model algoritma Naive Bayes sebesar 95.11%, sedangkan waktu proses dari kedua model algoritma yang diteliti menunjukkan hasil 0 s. Namun jika di telusuri lebih lanjut ternyata masih belum bisa dinyatakan sebagai algoritma yang lebih unggul.

Penerapan Algoritma C4.5 Untuk Klasifikasi Tren Pelanggaran Kendaraan Angkutan Barang Dengan Metode Crisp-Dm (Purnomo et al., 2023). Hasil dari penelitan ini dapat mengetahui pola klasifikasi tren pelanggaran kendaraan angkutan barang berdasarkan hasil pohon keputusan algoritma C.45, sehingga hasil penelitian dapat menjadikan acuan dalam pengembilan keputusan dan membuat kebijakan. Hasil dari penelitian ini menunjukkan bahwa performa akurasi pada pengujian data mining klasifikasi tren pelanggaran kendaraan angkutan barang dengan 10 fold cross validation linear sampling menghasilkan akurasi 86.31% +/- 1.23% (micro average: 86.31%), shuffled sampling menghasilkan akurasi 86.34% +/- 0.67% (micro average: 86.34%) dan stratified

sampling menghasilkan akurasi 86.34% +/- 0.67% (micro average: 86.34%).(Purnomo, Pamungkas, & Juliane, et al., 2023).

Analisis Perbandingan Algoritma C4.5 Dan Cart Untuk Klasifikasi Penyakit Stroke (Suryani et al., 2022). Berdasarkan hasil pembahasan dan percobaan yang telah dilakukan, percobaan pertama dengan holdout data 60:40 menunjukkan hasil algoritma C4.5 memiliki tingkat akurasi yang baik mencapai 96% dari pada algoritma CART dengan akurasi hanya 95,76%. Kemudian, dilakukan percobaan kedua dengan hold-out data 70:30. Algoritma C4.5 menunjukkan akurasi sebesar 95,76% sedangkan algoritma CART hanya sebesar 95,11%. Secara keseluruhan, dapat disimpulkan bahwa algortima C4.5 merupakan metode klasifikasi terbaik ketika dibandingkan dengan algoritma CART.

Pengelompokan Penerima Vaksinasi Covid-19 Di Kota Bengkulu Menggunakan Algoritma K-Means Clustering (Muntahanah, Wendanado, & Toyib, 2022). System yang dibuat menggunakan bahasa pemprogram Php MySql, dalam pembuatan system ini digunakan versi offline atau localhost. Dengan menerapkan algoritma k-means Clustering setiap warga yang akan melakukan vaksin dapat melihat ketersediaan di tiap puskesmas dari stok vaksin yang tersedia. Dengan menerapkan algoritma k-means Clustering dapat mengklasifikasikan ketersedian stok vaksin untuk disalurkan kepada warga yang akan melakukan vaksin.