BAB II

LANDASAN TEORI

1.1 Tinjauan Pustaka

1) Oleh Mochammad Faid Program Studi Teknik Informatika dan Teknik Elektronika, STT Nurul Jadid dengan judul Klasifikasi Mutu Padi Organik Menggunakan C4.5 di Dinas Pertanian Bondowoso. Dimana penelitian ini mengangkat masalah dari data mutu padi organik yang masih berbentuk manual sehingga sulit untuk melihat pola mutu padi organik secara menyeluruh, untuk kepentingan pengembangan yang akan dijadikan sebuah *software* klasfikasi yang akan sangat bermanfaat bagi dinas pertanian tersebut. Proses klasfikasi menggunakan algoritma c4.5 yang menghasilkan akurasi sebesar 83.0372% dimana nilai akurasi tersebut merupakan nilai tertinggi dari beberapa algoritma lain yakni *naïve bayes, decision stump, LADTree* dan *FTTree*.
2) Oleh Dito Putro Utomo, Pahala Sirait, Roni Yunis Program Studi Teknik Informatika, Universitas Budi Darma dengan judul Reduksi Atribut Pada Dataset Penyakit Jantung dan Klasifikasi Menggunakan Algoritma C5.0 dimana penelitian ini menggunakan dataset yang berasal dari UCI *Repository Machine Learning*, dimana terdapat 12 atribut dan 2 kelas target pada dataset penyakit jantung dan memiliki jumlah *record* sebanyak 303. Penelitian ini menggunakan metode *principal component analysis* (PCA) lalu melakukan klasifikasi padas setiap proses reduksi atributnya hingga menemukan kombinasi atribut optimal dari dataset. Kombinasi atribut

optimal dilihat dari nilai *eigenvalue* dari proses reduksi dengan PCA dan tingkat akurasi yang didapatkan sebesar 87,46%.

3) Oleh Carlis Hutabarat Program Studi Teknik Informatika STMIK Budi Darma dengan judul Penerapan *Data Mining* Untuk Memprediksi Permintaan Produk Kartu Perdana Internet Menggunakan Algoritma C5.0 dalam penelitian ini data yang dianalisa berhubungan dengan data penjualan kartu perdana internet. Dimana dalam data penjualan tersebut atribut yang diprediksi tentang penjualan jenis kartu Tri, Axis, XL, Telkomsel, dan Indosat. Data penjualan yang digunakan merupakan data periode januari hingga maret 2017. Implementasi dari data tersebut menggunakan *tools* See5 dan menghasilkan bahwa factor yang mempengaruhi besarnya permintaan adalah jenis jaringan yang dipakai di daerah tersebut.

4) Oleh Ridawati Manik, Pristiwanto, Keenedi Tampubolon Program Studi Teknik Informatika, STMIK Budi Darma dengan judul Prediksi Kolektibilitas Kredit Anggota Dengan Algoritma C5.0. Penelitian ini membahas status kolektibilitas yang dijadikan sebagai acuan bagi pihak bank unutk menindak lanjuti proses analisa pemberian kredit untuk mengurangi kredit macet yang bisa berdampak pada meningkatnya tingkat profitabilitas lembaga. Data yang digunakan pada prediksi ini berjumlah 100 data anggota yang memiliki track record pada CU Damai Sejahtera. Implementasi dari data tersebut menggunakan algoritma C5.0 menghasilkan sebuah *rule* dan pohon keputusan. *Tools* yang digunakan untuk proses implementasi tersebut adalah Rapid Miner. Dalam klasifikasi ini

menggunakan data 3 tahun terakhir dari 2014-2016, dengan jumlah transaksi yang diperoleh adalah 515 data transaksi dengan 11 atribut.

5) Oleh M. Erfan Rianto, Hendy Aang Hiriana, Imam Agus Hidayat dari Universitas Nurul Jadid, dengan judul Komparasi Metode KNN dan C4.5 dalam Klasifikasi Data Mutu Padi Organik. Penelitian ini membahas perbandingan hasil klasifikasi dengan *Grade* antar objek berdasarkan data yang jaraknya paling dekat dengan objek tersebut. Algoritma KNN didasarkan perbadingan contoh tes yang diberikan contoh pelatihan yang mirip dengan ide metode ini adalah untuk mengindentifikasi k sampel dalam training set yang independen variable x mirip dengan y dan menggunakan sample k ini unutk mengklasifikasi sampel baru ini ke dalam kelas. Pengujian menggunakan program *python* dengan *memanfaatkan library numpy, matplotlib, pandas, classification report* dan *accuracy score*.

Berdasarkan penelitian sebelumnya pada (Faid, 2017) yang membahas mutu padi organik terdapat beberapa perbedaan yaitu :

1. Algoritma yang digunakan yaitu Algoritma C5.0.
2. Metode pengujian evaluasi model yang digunakan yaitu *k-fold cross-validation.*
3. Bahasa yang digunakan yaitu bahasa pemrograman R.
4. Tools yang digunakan yaitu RStudio.

1.2 *Data Mining*

Tidak dapat disangkal lagi kebutuhan manusia akan data dan informasi. Hal ini tidak telepas dari dunia teknologi, arus informasi dapat mengalir dengan cepat dan mudah. Perlu adanya penataan dan pengelolaan data dan informasi yang ada supaya tidak membingungkan hingga sulit untuk dipahami, apalagi jika data

tersebut tidak valid karena data akan diubah menjadi sebuah informasi dimana banyak persepsi yang akan mengganggunya. Di sinilah peran data mining dibutuhkan. *Data mining* adalah teknologi penggalian pengetahuan atau yang bisa disebut informasi dari kumpulan data sehingga hasilnya dapat digunakan untuk pengambilan sebuah keputusan.

Data mining secara umum adalah kegiatan pencarian (*discovery*) secara berulang (*iterative*) dan intensif yang bertujuan untuk mengekstrak pengetahuan dari sekumpulan data yang awalnya tidak/belum memiliki arti yang penting. Pengetahuan yang dimaksud dapat berupa *pattern*/pola, hubungan, perubahan, *anomaly*, ataupun model yang muncul dari data. Hasil yang didapatkan harus valid, berguna dan mudah dimengerti (Setia *dkk.*, 2018).

1.3 *Imbalance Class*

Imbalancing Class atau yang biasa disebut sebagai kelas tidak seimbang adalah suatu keadaan dimana terdapat kelas yang memiliki jumlah instance lebih banyak daripada kelas lainnya. Dampak dari kelas yang tidak seimbang ini adalah kinerja tiap algoritme cenderung menghasilkan nilai yang sama, namun ketika memprediksi kelas yang minoritas cenderung menghasilkan nilai *False Positive* yang besar. Terdapat metode yang dapat digunakan untuk melakukan penyeimbangan kelas yaitu pendekatan berbasis sampel. ROSE merupakan teknik penyeimbangan distribusi kelas dengan melakukan replikasi pada kelas minoritas secara acak sehingga jumlah data pada kelas minoritas sama dengan jumlah data kelas mayoritas (Nasution dan Basuki, 2021).

1.4 *Preprocessing* Data

Preprocessing data adalah hal yang harus dilakukan dalam proses *data mining*, karena tidak semua data atau atribut data digunakan dalam proses *data mining*. Proses ini dilakukan agar data yang akan digunakan sesuai dengan kebutuhan.

Preprocessing data dapat dilakukan dengan langkah-langkah sebagai berikut:

1. *Integration* adalah suatu langkah untuk menggabungkan data dari beberapa sumber. Data *integration* hanya dilakukan jika data berasal dari tempat yang berbeda-beda.
2. *Cleaning* adalah proses menghilangkan *noise* dan menghilangkan data yang tidak relevan atau inkonsisten disebut pembersihan data. Dalam hal ini, transaksi yang memiliki jumlah item kurang dari dua (item tunggal) akan dihilangkan.
3. *Transformation* adalah menormalisasikan data dan aggresi data. Data *transformation* biasanya digunakan untuk mengubah data dalam bentuk yang sesuai dalam proses *data mining*.

1.5 Klasifikasi

Pengklasifikasi *data mining* melakukan proses dengan cara belajar dengan data yang ada kemudian mengklasifikasikan data baru, hasil dari metode klasifikasi adalah klasifikasi (nominal atau ordinal). Untuk melihat apakah perkiraan akurasi yang diberikan oleh model klasifikasi sudah benar, ada sesuatu yang disebut matriks konfusi. Dari matriks ini, penambang dapat memperkirakan keakuratan proses yang telah dijalankan.

Klasifikasi melakukan proses pembangunan model didasarkan dengan data latih yang tersedia, lalu memanfaatkan model tersebut guna mengklasifikasikan terhadap data yang baru. Klasifikasi merupakan pekerjaan yang melakukan pembelajaran dari fungsi target yang mengelompokan setiap *set* atribut ke *label* kelas yang tersedia. Sistem yang melakukan proses klasifikasi diharapkan mampu melakukan klasifikasi *dataset* dengan benar, tetapi kinerja sistem tidak bisa 100% benar sehinga sebuah sistem klasifikasi juga harus diukur kinerjanya (Utomo *dkk.*, 2020).

1.6 *Decision Tree*

Decision Tree adalah metode penambangan data yang umum digunakan untuk membangun sistem klasifikasi berdasarkan beberapa kovariat atau untuk mengembangkan algoritma prediksi untuk variabel target. Metode ini mengklasifikasikan populasi ke dalam segmen seperti cabang yang membangun pohon terbalik dengan simpul akar, simpul internal, dan simpul daun. Algoritma non-parametrik dan dapat secara efisien menangani kumpulan data yang besar dan rumit tanpa memaksakan struktur parametrik yang rumit. Ketika ukuran sampel cukup besar, data studi dapat dibagi menjadi set data pelatihan dan validasi. Menggunakan dataset pelatihan untuk membangun model pohon keputusan dan dataset validasi untuk memutuskan ukuran pohon yang sesuai yang dibutuhkan untuk mencapai model akhir yang optimal. Algoritma yang sering digunakan untuk mengembangkan pohon keputusan (termasuk CART, C4.5, CHAID, dan QUEST) dan menjelaskan program SPSS dan SAS yang dapat digunakan untuk memvisualisasikan struktur pohon (Song dan Lu, 2015).

Menurut (Darmawan *dkk.*, 2017), Karakteristik dari *decision tree* dibentuk dari sejumlah elemen sebagai berikut :

1) *Node*, yang menyatakan variabel. *Node* bisa berupa variabel akar, variabel cabang, dan kelas.
2) *Arm*, setiap cabang menyatakan nilai hasil pengujian di *node* bukan daun.
3) *Node* akar, tidak mempunyai input arm yaitu lengan masukan dan mempunyai nol atau lebih *output arm* yaitu lengan keluar.
4) *Node* internal, setiap *node* yang bukan daun (*non terminal*) yang mempunyai tepat satu *input arm* dan dua atau lebih *output arm*, *node* ini menyatakan pengujian yang didasarkan pada nilai fitur.
5) *Node* daun (*terminal*) adalah *node* yang mempunyai tepat satu *input arm* dan tidak mempunyai *output arm*. *Node* ini menyatakan label kelas (keputusan).

1.7 Algoritma C5.0

C5.0 adalah algortima klasifikasi yang dapat dilakukan dalam kumpulan data besar. Algoritma C5.0 ini lebih baik daripada C4.5 pada kecepatan proses, memori dan efisiensi. Algoritma C5.0 bekerja dengan membedakan sampel berdasarkan pada atribut yang menyediakan informasi. C5.0 dapat membagi atribut berdasarkan dari nilai informasi gain yang paling besar. Proses akan berlanjut hingga bagian sampel tidak dapat dibagi. Algoritma C5.0 dapat menangani atribut kontinyu dan diskrit. Pemilihan atribut dalam algoritma ini akan diproses menggunakan *information gain*. Atribut dengan nilai *Gain* tertinggi akan dipilih sebagai akar bagi *node* selanjutnya (Utomo *dkk.*, 2020).

Strategi pengembangan decision tree dengan menggunakan alogritma C5.0 adalah sebagai berikut:

1) Pada tahap awal, *tree* digambarkan sebagai node tunggal yang merepresentasikan *training set*.
2) Jika sampel seluruhnya berisi kelas yang sama, maka *node* tersebut menjadi *leaf* dan dilabeli dengan kelas tersebut.
3) Jika tidak, algoritma dengan menggunakan ukuran berbasis entropi *(information gain*) akan memilih variabel prediktor yang akan memisahkan *record* ke dalam kelas-kelas individual. Variabel tersebut menjadi variabel tes atau keputusan pada *node* tersebut.
4) Cabang dikembangkan untuk tiap nilai yang diketahui dari variabel tes, dan sampel dipartisi berdasarkan cabang tersebut.
5) Algoritma menggunakan proses yang sama secara rekursif membentuk *decision tree*.
6) Partisi rekursif berakhir hanya ketika satu dari kondisi-kondisi berikut terpenuhi :
    a. Seluruh *record* pada *node* tertentu memiliki kelas yang sama.
    b. Tidak ada atribut yang tersisa pada *record* yang dapat dipartisi lebih lanjut. Dalam kasus ini suara mayoritas digunakan. *Node* tersebut menjadi *leaf node* dan dilabeli dengan kelas yang menjadi mayoritas dalam *record* yang ada.
    c. Tidak ada *record* untuk cabang variabel tes. Dalam kasus ini, *leaf* terbentuk dengan mayoritas kelas sebagai *label record* tersebut.

Langkah-langkah membuat pohon keputusan di algoritma C5.0 mirip dengan membuat pohon keputusan di algoritma C4.5. Analoginya meliputi perhitungan entropi dan *gain*. Jika algoritma C4.5 berhenti sampai *gain* dihitung,

maka pada algoritma C5.0 akan terus menskalakan gain menjadi menggunakan *gain* dan entropi yang ada.

Adapun persamaan rumus untuk mencari nilai *entropy* adalah sebagai berikut :

$$Entropy(S) = -\sum_{i=1}^{m} p_i log_2^{pi} \quad (2.1)$$

Dimana S adalah kumpulan data yang terdiri dari data sampel, adalah proporsi yang dapat dihitung dengan $pi = \frac{ni}{|s|}$, ni adalah jumlah data yang termasuk dalam kelas i, dan $|s|$ adalah jumlah data dalam himpunan S. Untuk menghitung kondisional entropi untuk atribut A, Persamaan 2.2 digunakan.

$$E(S|A) = -\sum_{j=1}^{v} p_i^{'} \sum_{i=1}^{m} p_{ij}\, log_2(p_{ij}) \quad (2.2)$$

Dimana adalah proporsi yang dapat dihitung dengan $p'_j = \frac{|S_j|}{S} = \frac{\sum_i^m n_{ij}}{n}$, p_{ij} adalah peluang bersyarat dimana dapat dihitung dengan $p_{ij=\frac{n_{ij}}{|S_j|}}$, dan $|S_j|$ adalah jumlah data dengan atribut A. Kemudian, nilai *gain* dari atribut A dapat dihitung dengan Persamaan 2.3.

$$Gain\ (A) = E(A) - E(S|A) \quad (2.3)$$

Setelah mendapat nilai *entropy* dan *gain*, maka langkah berikutnya adalah menghitung nilai *gain ratio*. Nilai dari *gain ratio* pada atribut A dapat dihitung pada persamaan 2.4.

$$Gain\ Ratio\ (A) = \frac{Gain\ (A)}{Split\ (A)} \tag{2.4}$$

Dimana,

$$Split(A) = -\sum_{j=1}^{v} p'_j log_2\,(p'_j)$$

Evaluasi model dilakukan dengan menghitung akurasi, yang akan menunjukkan tingkat yang benar dari memprediksi data terhadap data yang sebenarnya. Semakin tinggi nilai akurasi berarti semakin rendah prediksi data uji *error* sehingga model memiliki performansi yang baik. Metode evaluasi penelitian ini adalah *K-fold cross-validation*, dimana membagi himpunan sampel secara acak menjadi k himpunan bagian. Dalam metode ini, diulang k kali untuk pelatihan dan pengujian data. Satu subset digunakan untuk pengujian di setiap iterasi, sedangkan subset yang tersisa digunakan untuk pelatihan. Akurasi diperoleh berdasarkan data uji terhadap model klasifikasi menggunakan Persamaan 2.5.

$$Accuracy(\%) = \frac{\sum Test\ data\ is\ correctly\ classified}{\sum Test\ data} \times 100 \tag{2.5}$$

1.8 Bahasa Pemrograman R

R Merupakan bahasa yang digunakan dalam komputasi statistik yang pertama kali dikembangkan oleh Ross Ihaka dan Robert Gentlement di University of Auckland New Zealand yang merupakan akronim dari nama depan kedua pembuatnya. Sebelum R dikenal ada S yang dikembangkan oleh John Chambers dan rekan-rekan dari Bell Laboratories yang memiliki fungsi yang sama untuk komputasi statistik. Saat ini, R secara luas diakui sebagai salah satu perangkat lunak

paling kuat untuk analisis data dan ilmu data. R dibuat dengan tujuan awal untuk menghitung statistik dan grafik. Awalnya digunakan oleh para ilmuwan dalam penelitian dan akademisi mereka. Pada penelitian ini untuk mendukung permodelan adalah versi terbarunya yakni versi R-4.2.1

Selain karena R dapat digunakan secara gratis terdapat kelebihan lain yang ditawarkan, antara lain:

1) *Protability*.

   Penggunaan *software* dapat digunakan kapanpun tanpa terikat oleh masa berakhirnya lisensi.

2) *Multiplatform*. R bersifat *Multiplatform Operating Systems*

   Dimana software R lebih kompatibel dibanding *software* statistika lainnya.

3) *General and Cutting-edge.*

   Berbagai metode statistik dapat digunakan untuk analisis statistika dengan pendekatan klasik dan pendekatan modern.

4) *Programable*

   Pengguna dapat memprogram metode baru atau mengembangakan modifikasi dari analisis statistika yang telah ada pada sistem R.

5) Berbasis analisis matriks

   Bahasa R sangat baik digunakan untuk *programming* dengan basis matriks.

6) Fasiltas grafik yang lengkap.

Adapun kekurangan dari R antara lain:

1) *Point and Click GUI*. Interaksi utama dengan R bersifat CLI

R-Commander sendiri merupakan GUI yang diciptakan dengan tujuan untuk keperluan pengajaran sehingga analisis statistik yang disediakan adalah yang klasik.

2) *Missing statistical function*.

R merupakan *lingua franca* untuk keperluan komputasi statistika modern saat ini, dapat dikatakan ketersediaan fungsi tambahan dalam bentuk Paket hanya masalah waktu saja.

Berikut ini beberapa library atau package yang digunakan dalam pengolahan data dengan algoritma C5.0 pada bahasa pemrograman R :

1) Dplyr

dplyr (menggugah "data tang") mendefinisikan transformasi data melalui lima fungsi sederhana, sering disebut "kata kerja:", dbplyr (backend database dplyr) untuk mengakses data dalam database. sintaks akrab dplyr, ditambah dengan linter SQL seperti sqlfluff, memungkinkan untuk menulis kueri SQL yang dapat dibaca dan bebas kesalahan.

2) Tidyverse

Menurut penulisnya, “the tidyverse adalah kumpulan paket ilmu data R yang berpendirian,” dianggap sebagai cerminan praktik terbaik karena tata bahasa, filosofi desain, dan struktur data yang intuitif dan terpadu.

3) Tidyrules

Paket tidyrules dimaksudkan untuk mengekstrak aturan yang dapat diuraikan dari objek model ke format tibble/data.frame. Paket mendukung model berikut - C5.0, rpart dan cubist. Aturan keluaran dapat diuraikan oleh R, python (kueri panda) dan SQL (dengan klausa WHERE).

4) Caret

Library/package caret yang disebut sebagai *Classification And Regression Training* adalah serangkaian fungsi yang membantu merampingkan setiap proses yang membuat model prediktif. *Library/package* ini menyusun alat yang dapat digunakan untuk pra-pemrosesan data *tuning* / penyetelan model dengan bantuan resampling estimasi kepentingan variabel dan pemisahan data.

5) C50

C50 adalah implementasi R dari algoritma pembelajaran mesin terawasi C5. 0 yang dapat menghasilkan pohon keputusan. Algoritma asli dikembangkan oleh Ross Quinlan. Ini adalah versi perbaikan dari C4.5..

1.9 R Studio

R Studio adalah *Integrated Development Environment* (IDE) untuk R yang banyak digunakan saat ini. Bahasa pemrograman untuk perhitungan statistik dan grafik. R Studio didirikan oleh JJ Allaire, pencipta bahasa pemrograman ColdFusion. Hadley Wickham adalah Kepala Ilmuwan di R Studio. R Studio tersedia dalam dua edisi: R Studio Desktop, di mana program berjalan secara lokal seperti aplikasi desktop biasa dan R Studio Server, yang memungkinkan R Studio diakses menggunakan *browser web* saat dijalankan di server Linux jarak jauh. Pada penelitian ini untuk mendukung kinerja permodelan. RStudio yang akan digunakan yaitu versi 9.1.191.26 yang rilis terbaru di bulan juli 2022.