BAB II

LANDASAN TEORI

1.1. Tinjauan Literatur

Di dalam penerapan data mining terdapat beberapa penelitian yang telah dilakukan menggunakan algoritma C5.0 sehingga dalam penelitian ini diperlukan tinjauan Pustaka sebagai alat dalam penerapan algoritma ini, agar dapat menghindari pembuatan ulang, mengindentifikasi kesenjangan, mengetahui algoritma yang sudah diterapkan, mengetahui penelitan yang sama dibidang ini, serta melanjutkan untuk penelitian sebelumnya.

Tabel 2. 1. Daftar Literatur

No	Penulis	Judul	Metode	Hasil
1	(Fatiya Nur Umma, Budi Warsito, Di Asih 1 Maruddani, 2021)	Klasifikasi Status Kemiskinan Rumah Tangga Dengan Algoritma C5.0 Di Kabupaten Pemalang	Algoritma C5.0	Tujuan utama dalam penulisan judul Klasifikasi Status Kemiskinan Rumah Tangga Dengan Algoritma C5.0 Di Kabupaten Pemalang ialah mengkaji klasifikasi status kemiskinan rumah tangga di kabupaten Pemalang Hasil dari menggunakan Algoritma C5.0 dalam penelitian ini menghasilkan tingkat akurasi yang sangat baik dengan relatif dari segi presisi dan efiensi, dengan tingkat akurasi 91,16%.
2	(Edi Wijaya, Feriani	Aplikasi Prediksi Penentuan	Algoritma C5.0	Tujuan utama dalam penulisan judul Aplikasi Prediksi Penentuan Kelancaran Pembayaran Koperasi

	Astuti Tarigan,	Kelancaran Pembayaran		Dengan Algoritma C5.0 ialah mengembangkan Aplikasi Prediksi
No	**Penulis**	**Judul**	**Metode**	**Hasil**
	Michael, Juni 2019)	Koperasi Dengan Algoritma C5.0		untuk penentuan kelancaran pembayaran pada koperasi menerapkan Algoritma C5.0. Hasil dari menggunakan metode Algoritma C5.0 ialah mengembangkan Aplikasi Prediksi untuk penentuan kelancaran pembayaran pada koperasi menerapkan Algoritma C5.0. Hasil dari menggunakan metode Algoritma C5.0 dalam penelitian ini mengembangkan aplikasi prediksi dengan tingkat nilai entropy >45 tahun 0.918 dan nilai gain 0.049.
3	(Fajar Fathur Rachman, Setia Pramana, Desember 2020)	Analisis Sentimen Pro dan Kontra Masyarakat Indonesia tentang Vaksin COVID-19 pada Media Sosial *Twitter*	*Web Scraping*	Melakukan riset bagaimana respon & opini masyarakat Indonesia terhadap vaksin COVID-19 dengan menggunakan data yang bersumber dari media sosial *twitter*, dengan Hasil kesimpulan bahwa masyarakat lebih banyak memberikan respon yang bersentimen positif terhadap vaksin COVID-19 dibandingkan dengan respon yang bersentimen

				negatif hasil presentase positif 1461 (29,6), Netral 2313 (46,8), Negatif 1167 (23,6).
No	**Penulis**	**Judul**	**Metode**	**Hasil**
4	(Ilham Kurniawan, Rizal Amegia Saputra, September 2017)	Penerapan Algoritma C5.0 Pada Sistem Pendukung Keputusan Kelayakan Penerimaan Beras Masyarakat Miskin	Algoritma C5.0	Tujuan utama dalam penelitian ini ialah melakukan riset menggunakan metode sistem pendukung keputusan untuk kelayakan penerimaan beras masyarakat miskin karena pemanfaatan sistem pendukung keputusan dapat dimanfaatkan untuk membantu manusia mengambil keputusan dengan cepat, tepat, dan konsisten, dengan hasil aplikasi sistem pendukung keputusan ini dapat menjadi alternatif pemecahan masalah, diantaranya: Sistem pendukung keputusan dengan algoritma C5.0 dibuat agar membantu para pengguna khususnya para petugas kelurahan yang bersangkutan dalam menentukan keputusan mengenai siapa yang benar-benar layak menerima bantuan beras untuk masyarakat miskin.

5	(Aswan S. Sunge, Faradilla Laksmita Devi, Juni 2020)	Analisis Pemilihan Jurusan Siswa Dengan Metode Klasifikasi Algoritma	Algoritma C5.0	Tujuan utama dalam penelitian ini adalah menganalisa pemilihan jurusan siswa dengan metode klasifikasi menggunakan algortima C5.0 dengan hasil *confusion matrix true* positif (tp) sebanyak 6 *record, false posifitve* (fp) sebanyak 0
No	**Penulis**	**Judul**	**Metode**	**Hasil**
		C5.0(Studi Kasus : Smk Ma'Arif Nu Al-Mawardi Bekasi		*record, true negative* (tn) sebanyak 16 *record, false negative* (fn). sebanyak 1 record dengan akurasi 95.65%
6	(Tedi Permana, Amril Mutoi Siregar, Anis Fitri Nur Masruriyah, Ayu Ratna Juwita, Desember 2020)	Perbandingan Hasil Prediksi Kredit Macet Pada Koperasi Menggunakan Algoritma KKN dan C5.0	Algoritma C5.0 *dan* KKN	Kesimpulan berdasarkan hasil perhitungan dengan algoritma KNN dan C5.0 dengan dataset yang digunakan sebanyak 30 data transaksi simpan pinjam, dan lima atribut yang digunakan, nilai akurasi yang diperoleh algoritma C5.0 lebih baik dari KNN yaitu 86.67%. Sedangkan untuk algoritma KNN mendapatkan nilai akurasi 83.33%, sehingga algoritma C5.0 lebih cocok digunakan sebagai patokan untuk prediksi kredit macet di koperasi.

2.1.1 Literatur 1

Pemalang merupakan kabupaten dengan dengan persentase kemiskinan sebesar 16,04%. Garis kemiskinan di Pemalang berada pada angka Rp 351.183 per bulan. Berbagai program telah dijalankan oleh pemerintah yang tujuannya untuk dapat mengurangi jumlah dan proporsi penduduk miskin. Sejumlah permasalahan masih ditemui meskipun pemerintah telah mengalokasikan dana yang sangat besar untuk penerapan kebijakan. Menurut Bappenas pemenuhan kebutuhan dasar atas masyarakat miskin atas pendidikan dan kesehatan menemui kendala pendataan dan akurasinya. Tujuan utama dalam penulisan judul Klasifikasi Status Kemiskinan Rumah Tangga Dengan Algoritma C5.0 Di Kabupaten Pemalang ialah mengkaji klasifikasi status kemiskinan rumah tangga di kabupaten Pemalang hasil dari menggunakan Algoritma C5.0 dalam penelitian ini menghasilkan tingkat akurasi yang sangat baik dengan relatif dari segi presisi dan efiensi, dengan tingkat akurasi 91,16%.

2.1.2 Literatur 2

Pada praktiknya, biasanya pihak koperasi akan mengevaluasi kelayakan kredit dari nasabahnya secara konvensional. Proses tersebut tentunya sangat tidak efisien dan efektif dikarenakan apabila jumlah nasabah sangat banyak, maka tentunya memerlukan waktu yang sangat lama dalam melakukan proses evaluasi. Selain itu, proses penilaian kelancaran pembayaran kredit dari nasabah secara konvensional tentunya kurang akurat dikarenakan penilaian yang berbeda-beda dari pihak nasabah membuat persentase terjadinya penunggakan pembayaran kredit menjadi sangat besar. Tujuan utama dalam penulisan judul Aplikasi Prediksi Penentuan Kelancaran

Pembayaran Koperasi Dengan Algoritma C5.0 ialah mengembangkan Aplikasi Prediksi untuk penentuan kelancaran pembayaran pada koperasi menerapkan algoritma C5.0. Pada hasil menggunakan Algoritma C5.0 dalam penelitian ini mengembangkan aplikasi prediksi dengan tingkat nilai entropy >45 tahun 0.918 dan nilai gain 0.049.

2.1.3 Literatur 3

Wabah penyakit baru yang disebabkan oleh virus korona (2019-nCoV) atau yang biasa disebut dengan COVID-19 ditetapkan secara resmi sebagai pandemi global oleh *World Health Organization* (WHO) pada tanggal 11 Maret 2020 lalu. Melihat pesatnya penyebaran COVID-19 dan bahaya yang akan muncul jika tidak segera ditangani, salah satu cara yang sangat mungkin untuk mencegah penyebaran virus ini adalah dengan mengembangkan vaksin. Vaksin tidak hanya melindungi mereka yang divaksinasi tetapi juga masyarakat luas dengan mengurangi penyebaran penyakit dalam populasipenelitian ini ingin melihat bagaimana respon & opini masyarakat Indonesia terhadap vaksin COVID-19 dengan menggunakan data yang bersumber dari media sosial twitter. Untuk menjawab permasalahan tersebut, maka penelitian ini akan melakukan analisis sentimen dengan mengklasifikasikan respon masyarakat tersebut ke dalam sentimen positif & negatif, dan mengelompokkan opini masyarakat terhadap vaksin COVID-19 dengan menggunakan metode *Latent Dirichlet Allocation* (LDA).

2.1.4 Literatur 4

Pada penelitian ini sistem pendukung keputusan digunakan untuk penerimaan beras masyarakat miskin. Algoritma yang digunakan pada sistem pendukung

keputusan yaitu algoritma C5.0 dengan model klasifikasi tree. Penerapan algoritma C5.0 pada dataset kelurahan Caringin Wetan dan kelurahan Gunung parang tahun 2015. Model pohon keputusan yang dihasilkan dari penerapan algoritma C5.0 dengan pengolahan data menggunakan SPSS itu dapat digunakan pada aplikasi penentuan penerimaan beras raskin yang akan di gunakan oleh pihak tertentu. Sistem pendukung keputusan dengan algoritma C5.0 dibuat agar membantu para pengguna khususnya para petugas kelurahan yang bersangkutan dalam menentukan keputusan mengenai siapa yang benar-benar layak menerima bantuan beras untuk masyarakat miskin. Sistem pendukung keputusan ini dirancang dalam bentuk aplikasi android, sehingga memudahkan para pengguna khususnya para petugas kelurahan yang bersangkutan dalam penggunaannya. Selain itu, karena dibuat dalam aplikasi android, maka informasi yang didapat akan lebih real-time atau bisa didapatkan pada saat itu juga, dan bisa langsung digunakan dimanapun.

2.1.5 Literatur 5

Penentuan jurusan di SMK biasanya berdasarkan 2 program keahlian yang telah dipilih ketika mendaftar. Contoh-contoh pilihan jurusan yang ada di SMK seperti Teknik Kendaraan Ringan, Teknik Sepeda Motor, Rekayasa Perangkat Lunak, Administrasi Perkantoran, Teknik Komputer dan Jaringan, dan lain sebagainya. Kemudian dari hasil tes dan nilai rapor diolah lagi untuk mendapatkan nilai yang memenuhi syarat untuk pilihan jurusan pertama. Jika standar nilai pada jurusan pertama tidak terpenuhi, maka dicocokan untuk pilihan jurusan yang kedua. Jika sesuai dengan standar nilai jurusan yang ke 2 maka di masukan ke pilihan jurusan yang kedua

dan jika tidak sesuai dengan kedua pilihan jurusan maka siswa di nyatakan tidak diterima oleh SMK tersebut. Tujuan Utama dalam penelitian ini adalah menganalisa pemilihan jurusan siswa dengan metode klasifikasi menggunakan algortima C5.0 dengan hasil confusion matrix true positif (tp) sebanyak 6 *record, false posifitve* (fp) sebanyak 0 *record, true negative* (tn) sebanyak 16 *record*, *false negative* (fn) sebanyak 1 *record* dengan akurasi 95.65%.

2.1.6 Literatur 6

Koperasi adalah organisasi yang dibuat untuk membantu masyarakat dan para anggotanya dalam urusan keuangan, salah satunya yaitu simpan pinjam dana. Masyarakat Indonesia masih tergolong cukup besar dalam hal kredit macet pada peminjaman dana. Permasalahan kredit macet masih menjadi masalah utama di perusahaan pembiayaan, karena akan membuat kondisi keuangan perusahaan tersebut terganggu. Pada umumnya kredit macet di koperasi disebabkan karena pengurus koperasi masih terlalu sederhana dalam melakukan analisa data. Antisipasi yang biasa dilakukan oleh pihak koperasi hanya sebatas pendekatan personal kepada nasabah. Namun antisipasi tersebut masih belum efektif, ketika jumlah nasabah sangat banyak. Salah satu cara untuk menangani masalah tersebut adalah dengan memprediksi kredit macet menggunakan teknik komputasi. Kemudian penelitian tentang prediksi penyakit jantung yang dilakukan oleh Rohman, Suhartono, dan Supriyanto. Algoritma yang digunakan yaitu, C4.5 dan C4.5 berbasis Adabost. Penelitian ini menggunakan datasets pasien penyakit jantung yang didapatkan dari Universitas California, Invene (UCI), dan menggunakan 14 atribut. Hasil yang diperoleh dari penelitian tersebut yaitu algoritma

C4.5 berbasis adabost mendapatkan nilai akurasi yang lebih tinggi yaitu 92,24%, sehingga praktisi kesehatan dapat menggunakan hasil dari penelitian tersebut sebagai masukan dalam prediksi penyakit jantung. Selanjutnya penelitian oleh Praningki dan Budi yaitu, *prediksi* kanker serviks menggunakan algoritma Classification and Regression Tree (CART), Naïve Bayes, dan K-Neaerest Neighbor (KNN). Datasets dikumpulkan dari Rumah Sakit Daerah (RSUD) Kediri dan Yayasan Kanker Indonesia Cabang Kediri dan terdapat 12 atribut yang digunakan. Hasil dari penelitian tersebut menyatakan bahwa algoritma Naïve Bayes mampu melakukan *klasifikasi* dengan baik dengan nilai akurasi 94,44%, sehingga hasil dari penelitian tersebut dapat memberikan keputusan klinis bagi tenaga medis. Berdasarkan dari teknik – teknik yang telah dilakukan dari penelitian ini, terbukti mampu menangani permasalahan yang terjadi. Maka pada penelitian ini akan mengambil judul "Perbandingan Hasil Prediksi Kredit Macet pada Koperasi Menggunakan Algoritma KNN dan C5.0".

2.2. Analisis Sentimen

Analisis Sentimen atau Opinion Mining merupakan proses memahami, mengekstrak dan mengolah data tekstual secara otomatis untuk mendapatkan informasi sentimen yang terkandung dalam suatu kalimat opini. Analisis sentimen dilakukan untuk melihat pendapat atau kecendrungan opini terhadap sebuah 13 masalah atau objek oleh seseorang, apakah cenderung berpandangan atau beropini negatif atau positif (Rachman and Pramana, 2020).

2.3. Twitter Crawling

Twitter menyediakan *Application Programming Interface Streaming* (APIS) untuk memfasilitasi *data crawling*. API memudahkan pengguna untuk mengambil data *tweet* secara *real time*. Tujuan awal dibentuknya *Twitter* API ini adalah untuk mengetahui relasi dan interaksi antara pengguna, namun sebaliknya *Twitter* API banyak digunakan untuk menggali informasi komunitas tertentu atas pandangannya terhadap topik yang sedang trend (Eka Sembodo, Budi Setiawan and Abdurahman Baizal, 2016).

Crawling adalah proses pengambilan sejumlah besar halaman web dengan cepat ke dalam suatu tempat penyimpanan lokal dan mengindeksnya berdasar sejumlah kata kunci (Eka Sembodo, Budi Setiawan and Abdurahman Baizal, 2016). Mesin pencari web bekerja dengan cara menyimpan informasi tentang banyak halaman web, yang diambil langsung dari situs dan untuk penelitian ini akan mengambil opini dari akun twitter tertentu. Halaman-halaman ini diambil dengan *twitter crawler* otomatis yang mengikuti setiap pranala/*link* yang dilihatnya. Isi setiap halaman lalu dianalisis untuk menentukan cara indeks-nya (misalnya, kata-kata diambil dari judul, subjudul, atau *field* khusus yang disebut *meta tag*).

Data tentang halaman web disimpan dalam sebuah *database* indeks untuk digunakan dalam pencarian selanjutnya. Sebagian mesin pencari, seperti *Google*, menyimpan seluruh atau sebagian halaman sumber (yang disebut *cache*) maupun informasi tentang halaman web itu sendiri. Penelitian ini juga melakukan cara yang sama dengan memanfaatkan *twitter* API, kemudian membuat aplikasi berbasis PHP

untuk menangkap kata kunci yang perusahaan telekomunikasi beserta produk-produknya.

2.4. Term Frequency-Inverse Document Frequency

Term Frequency-Inverse Document Frequency (TF-IDF) adalah metode yang digunakan untuk menghitung bobotan setiap kata yang telah diekstrak. Penggunaan metode ini umumnya dilakukan untuk mengihitung kata umum yang ada pada information retrieval. Model pembobotan TF-IDF merupakan metode yang mengintegrasikan model *term frequency* (tf) dan *inverse document frequency* (idf). *Term frequency* (tf) merupakan proses untuk menghitung jumlah kemunculan term dalam satu dokumen dan *inverse document frequency* (idf) digunakan untuk menghitung term yang muncul di berbagai dokumen (komentar) yang dianggap sebagai term umum, yang dinilai tidak penting (Kurniawan and Saputra, 2017).

Tahapan pembobotan dengan TF-IDF adalah :

1. Term Frequency (tf)

Term frequency atau tf merupakan jumlah kemunculan atau frekuensi kata pada suatu dokumen. Sementara Wtf adalah jumlah bobot dari tf yang telah dihitung dengan logaritma. Perhitungan Term Frequency dilakukan dengan persamaan 1.1:

$$\mathrm{Wtf}_{t,d} = \begin{cases} 0, \\ 1 + \log_{10} \mathrm{tft,d}, \text{ if } \mathrm{tft,d} > 0 \end{cases} \quad (1)$$

2. Document Frequency (df) Document Frequency (df) merupakan frekuensi atau jumlah dokumen yang mengandung suatu kata.

3. Inverse Document Frequency (idf) Inverse Document Frequency (idf) adalah bobot kebalikan dari bobot document frequency. Kata yang jarang muncul di banyak

$$idf_t = log_{10}(N/df_t)$$

dokumen mempunyai bobot Inverse Document Frequency yang tinggi. Perhitungan dari Inverse Document Frequency (idf) dilakukan dengan persamaan 1.2:

(2)

Keterangan:

N : jumlah dokumen teks.

dft : jumlah dokumen yang mengandung suatu kata t.

4. Term Frequency-Inverse Document Frequency (tf-idf) Pembobotan ini adalah hasil perkalian dari pembobotan term frequency dan inverse document frequency dari suatu term.

Perhitungannya dapat dilihat melalui persamaan

2.3: $$\mathrm{W}_{t,d} = \mathrm{W}tf_{t,d} \times idf_{\mathrm{t}} \quad (3)$$

Keterangan :

$wtft,d$: Term Frequency.

$idft$: Inverse Document Frequency.

2.5. Preprocessing

Saat melakukan analisis sentimen, teks dokumen harus disiapkan terlebih dahulu agar dapat digunakan dalam proses utama. Proses menyiapkan teks atau data mentah disebut preprocessing teks. Fungsi pra-pemrosesan teks mengubah data tidak terstruktur menjadi data terstruktur(Rachman and Pramana, 2020). Proses yang dilakukan dalam preprocessing adalah sebagai berikut

2.5.1. Cleansing

Tahapan ini bertujuan untuk menghilangkan karakter atau *symbol*, tweet yang

terdapat pada twitter memiliki berbagai komponen atau karakteristik tweet yang khas seperti "@" yang diidentifikasi sebagai komponen username, URL yang dikenal melalui operasi regular, hashtag yang menandakan kata sebagai topik yang sedang dibicarakan, dan "RT" yang diidentifikasi sebagai mengulang kembali tweet yang telah diposting. Komponen-komponen tersebut tidak memiliki pengaruh apapun terhadap sentimen, maka akan dibuang (Athira, Gholissodin and Perdana, 2018).

2.5.2. Case Folding

Case Folding merupakan proses penyeragaman bentuk huruf serta penghapusan angka dan tanda baca. Dalam kasus ini yang akan dipakai hanyalah huruf latin dari a hingga z (Athira, Gholissodin and Perdana, 2018).

2.5.3. Tokenizing

Pada proses tokenizing dokumen yang masih berupa kalimat dipecah per kata menjadi beberapa bagian dan secara bersamaan hilangkan semua karakter maupun tanda baca yang ada pada kalimat tersebut, hasil dari proses inilah yang disebut token (Athira, Gholissodin and Perdana, 2018). Untuk lebih jelasnya dapat dilihat pada tabel dibawah ini:

Tabel 2. 2. Proses Tokenizing

Sebelum Preprocessing	Sesudah Preprocessing
Vaksin merupakan solusi terbaik untuk saat ini	Vaksin Merupakan Solusi Terbaik Untuk Saat ini

2.5.4. Stopword Removal (Filtering)

Stopword Removal merupakan proses menghilangkan kata yang tidak mendeskripsikan sesuatu dalam bahasa Indonesia seperti “di”, ”ke”, “dari”, “yang”, “sedang”, “ini”, dan lain sebagainya. Di dalam *text classification*, kata seperti “tidak”, “bukan”, “tanpa” biasanya tidak termasuk kedalam kata yang akan dihilangkan. Dalam penerapannya kalimat yang mengandung teks tersebut perlu diubah atau disesuaikan pada proses preprocessing (Athira, Gholissodin and Perdana, 2018).

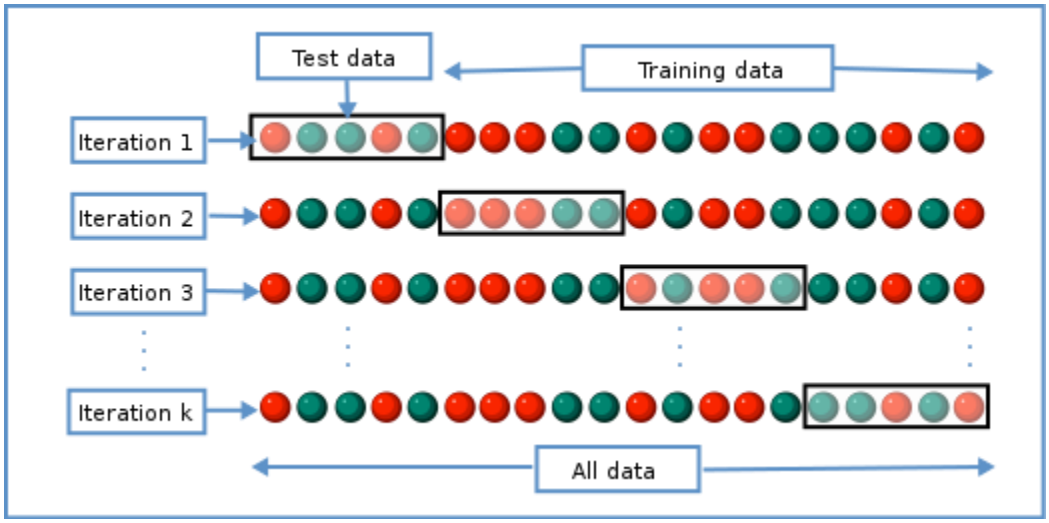
2.5.5. Stemming

Proses Stemming merupakan proses penghilangan imbuhan yang masih melekat sehingga diperoleh sebuah kata dasar, contohnya: “membaca”, “dibaca”, “dibacakan” akan dikonversi menjadi kata dasar (stem) “baca” (Athira, Gholissodin and Perdana, 2018).

2.6. K-Fold Cross Validation

Cross validation adalah metode statistik yang digunakan untuk mengevaluasi dan membandingkan algoritma pembelajaran dengan cara membagi data menjadi dua bagian: satu digunakan untuk belajar atau melatih model, satu untuk menguji model tersebut (Athira, Gholissodin and Perdana, 2018). Dalam penelitian ini, metode *cross validation* digunakan untuk mencari akurasi dari setiap model klasifikasi. Salah satu bentuk dari *cross validation* adalah *k-fold cross validation*. Dalam metode *k-fold cross validation*, data dibagi ke dalam beberapa partisi yang disebut dengan *fold*. Masing-masing *fold* memiliki jumlah data dengan ukuran yang sama atau mendekati sama. Selama k iterasi, dipilih salah satu *fold* sebagai data uji, sedangkan sisa k-1 *fold*

dijadikan data latih (Athira, Gholissodin and Perdana, 2018). Gambar berikut merupakan ilustrasi pembagian data dalam *4-fold cross validation*.



Gambar 2. 1. Ilustrasi 4-*fold cross validation* (Kurniawan and Saputra, 2017).

Pada gambar diatas, seluruh data dibagi menjadi 4 fold dengan setiap *fold* berisi 5 data. Dalam setiap iterasi, dipilih salah satu *fold* sebagai data uji dan sisanya menjadi data latih. Setiap data hanya boleh sekali menjadi data uji. Perhitungan akurasi penilaian terhadap data uji dilakukan di setiap iterasi.

2.7. Algoritma C5.0

Algoritma C5.0 sebuah sistem dapat dibangun untuk menciptakan sebuah pohon keputusan yang dapat digunakan untuk mendiagnosis sebuah penyakit, salah satunya penyakit diabetes melitus. Pada algoritma C5.0 dapat dilakukan tahapan pruning untuk menjadikan pohon (Kastawan, Wiharta and Sudarma, 2018). Keputusan yang lebih baik. Pruning menjadi salah satu cara untuk merampingkan struktur dari *decision tree* sehingga proses generalisasi datanya menjadi lebih baik, dan memudahkan interpretasinya (Riadi, Azhar and Wicaksono, 2020).

Algoritma C5.0 merupakan penyempurnaan dari algoritma terdahulu yang dibentuk oleh Ross Quinlan pada tahun 1987, yaitu algoritma ID3 dan C4.5. Strategi pengembangan *decision tree* dengan menggunakan alogritma C5.0 adalah sebagai berikut:

1. Menghitung information gain setiap atribut.
2. Membuat node berdasarkan atribut yang memiliki information gain tertinggi.
3. Cabang dikembangkan untuk tiap nilai yang diketahui dari atribut node.
4. Jika sampel seluruhnya berisi kelas yang sama, maka node tersebut memiliki leaf dan dilabeli dengan kelas tersebut.
5. Jika tidak, hitung information gain setiap atribut yang tersisa yang dapat memisahkan record ke dalam kelas-kelas individual.
6. Algoritma menggunakan proses yang sama secara rekursif atau perulangan membentuk pohon keputusan.
7. Partisi rekursif berakhir hanya ketika satu dari kondisi-kondisi berikut terpenuhi:
8. Seluruh record pada node tertentu memiliki kelas yang sama.

(Yusuf, 2007).

Untuk menentukan akar dari pohon keputusan ditentukan oleh gain yang tertinggi, sebelum menemukan gain terdebih dahulu menghitung nilai entrophy keseluruhan pada rumus persamaan dibawah:

$$Entropy(S) = \sum = 0 - * 2 \qquad (4)$$

Keterangan : S adalah himpunan (dataset) kasus n: banyaknya data pi: probabilitas yang di dapat dari kelas dibagi total kasus setelah kita mendapatkan nilai

entropy dari masing-masing atribut, langkah selanjutnya yaitu menghitung nilai dari Information Gain. Information gain adalah kriteria yang sering digunakan untuk menetukan akar utama dalam suatu pohon keputusan. Cara menghitung information gain adalah menghitung nilai dari *output* tiap atribut. Adapun rumus yang digunakan adalah sebagai berikut :

$$Gain(\mathrm{S},\mathrm{A}) = Entropy(\mathrm{S}) - \sum_{\mathrm{i}-1}^{\mathrm{n}} \left|\frac{S_i}{\mathrm{S}}\right| * Entropy(S_i) \qquad (5)$$

Keterangan :

S : Himpunan kasus A : Atribut

n : Jumlah partisi atribut A

|Sì| : Jumlah kasus pada partisi ke i

|S| : Jumlah kasus dalam S

(Umma, Warsito and Maruddani, 2021).