BAB II
TINJAUN PUSTAKA

2.1 Email dan Spam

2.1.1 Email / Surat Elektronik

Email, atau yang dikenal dalam bahasa Indonesia sebagai surat elektronik, merupakan sebuah alat yang memungkinkan para pengguna internet untuk bertukar pesan menggunakan alamat elektronik di dunia maya. Setiap pengguna email memiliki kotak pesan yang tersimpan di server email. Kotak surat memiliki alamat untuk mengidentifikasi dan berkomunikasi dengan kotak surat lainnya, baik untuk menerima maupun mengirim pesan. *Domain Name System* (DNS) memastikan bahwa semua pengguna memiliki alamat yang unik, tanpa duplikat di antara jutaan pengguna internet.

Layanan email biasanya dikategorikan menjadi dua jenis utama yaitu email yang diakses melalui perangkat lunak dan email berbasis web. Untuk pengguna yang memanfaatkan email melalui perangkat lunak, proses pengiriman pesan dilakukan dengan aplikasi klien email, contohnya Eudora atau Outlook Express. Program ini memungkinkan pengguna untuk mengedit dan membaca email tanpa memerlukan koneksi internet. Koneksi hanya diperlukan untuk mengirim atau menerima email dari kotak surat. [7]

2.1.2 Spam dan Ham

Email spam adalah pesan email yang tidak diminta atau tidak diharapkan yang disebarkan secara luas kepada ribuan atau bahkan jutaan alamat email. Email yang termasuk dalam kategori spam biasanya mencakup iklan atau promosi barang, upaya penipuan (phishing), virus atau malware, dan konten yang tidak relevan. penipuan lainnya. Email spam dapat sangat mengganggu dan berbahaya karena dapat menguras waktu dan sumber daya pengguna komputer. [1]

Ham atau email non-spam adalah pesan yang diinginkan atau relevan bagi penerima, biasanya dari pengirim tepercaya. Tidak seperti spam, ham biasanya berisi komunikasi penting seperti email pribadi, pesan bisnis, pemberitahuan

layanan, atau informasi terkait transaksi dari bank dan platform resmi. Email tersebut tidak berisi konten berbahaya dan tidak dikirim secara massal tanpa izin penerima.[1]

Dampak dari Spam email ini dapat mengganggu produktivitas dengan memenuhi kotak masuk pengguna dengan pesan yang tidak diinginkan dan tidak relevan. Hal ini dapat mempersulit pencarian email penting dan membuang waktu untuk menyortir email spam. Spam email dapat berisi tautan atau lampiran berbahaya yang dapat memasang malware di komputer atau mencuri informasi pribadi pengguna. Spam email dapat merusak reputasi pengguna dengan membuat mereka tampak tidak profesional atau mengabaikan keamanan email. Selain itu, email spam dapat mengambil data pribadi, termasuk nama, lokasi, nomor ponsel, dan informasi keuangan. Informasi ini nantinya dapat digunakan untuk pencurian identitas atau kejahatan lainnya. [8]

2.2 Data Mining

2.2.1 Pengertian Data Mining

Data Mining merupakan bidang yang menggabungkan metode dari pembelajaran mesin, pengenalan pola, statistik, basis data, dan visualisasi untuk menarik informasi dari kumpulan data yang besar. Dengan cara lain, data mining dapat diartikan sebagai proses menemukan pola dari data yang sangat banyak, data yang disimpan dalam repositori menggunakan teknologi pengenalan pola, teknik statistik, dan matematika. Data mining juga dikenal sebagai *knowledge discovery in database (KDD)*, adalah proses pengumpulan dan pemanfaatan data historis untuk mengungkap pola, keteraturan, atau hubungan dalam kumpulan data besar. Keluaran dari data mining dapat digunakan untuk meningkatkan pengambilan keputusan di masa mendatang. Istilah *pattern recognition* sekarang jarang digunakan karena merupakan bagian dari data mining. [9]

2.2.2 Fungsi Data Mining

1. Prediction

   Fungsi pertama dari data mining adalah prediksi. Ini adalah proses mengidentifikasi pola dalam data dan menggunakan beberapa variabel untuk memprediksi nilai atau jenis variabel lain yang nilainya masih belum diketahui.

2. Description

   Berikutnya adalah fungsi deskripsi, ini merupakan langkah untuk mengidentifikasi ciri-ciri utama dari suatu informasi dalam database.

3. Classification (Klasifikasi)

   Proses ini melibatkan pengenalan contoh atau fungsi yang digunakan untuk menggambarkan kelompok atau ide dari sekumpulan data tertentu. Metode yang digunakan untuk menggambarkan data itu sangat penting dan juga dapat memprediksi pola data di masa depan.

4. Association (Asosiasi)

   Yang terakhir adalah metode yang diterapkan untuk mengidentifikasi keterkaitan antara nilai-nilai atribut dalam satu kumpulan data.

2.2.3 Pengelompokan Data Mining

Pengelompokkan data mining dibagi menjadi beberapa kelompok, menurut Kusrini dan Luthfi (2009), yaitu :

1. Deskripsi

   Deskripsi merupakan metode untuk menjelaskan pola serta kecenderungan yang terdapat dalam data yang ada.

2. Estimasi

   Estimasi sangat mirip dengan proses klasifikasi, tetapi dalam estimasi, variabel yang dijadikan target cenderung lebih dekat kepada angka daripada kepada kategori. Model dikembangkan dengan memanfaatkan data lengkap yang memberikan angka sebagai nilai prediksi variabel target.

3. Prediksi

   Prediksi adalah tindakan memperkirakan nilai yang tidak diketahui dan memperkirakan nilai untuk masa mendatang.

4. Klasifikasi

   Dalam pengklasifikasian, ada variabel sasaran yang termasuk dalam kategori. Sebagai contoh, klasifikasi penghasilan bisa dibedakan menjadi tiga kelompok yang berbeda, yaitu tinggi, menengah, dan rendah.

5. Pengklasteran

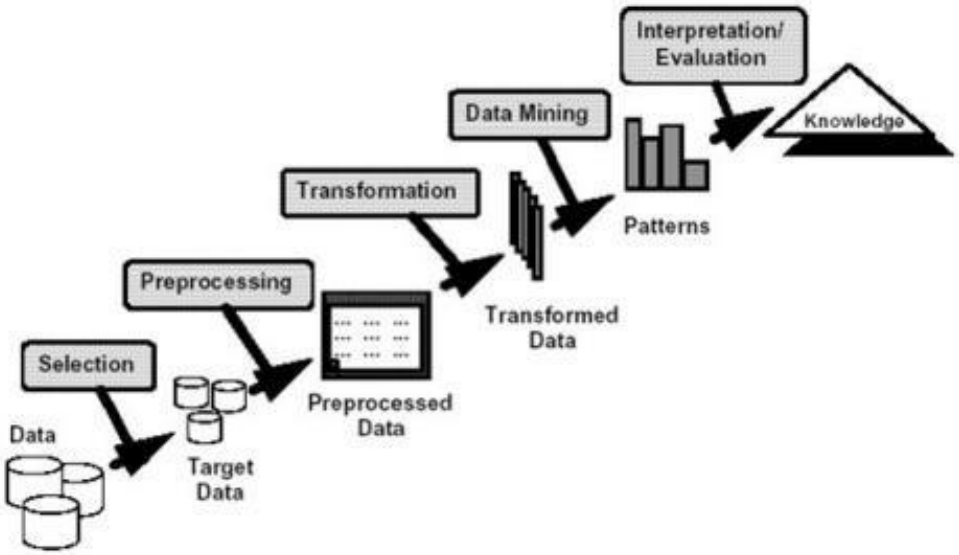
   Adalah klasifikasi record, analisis, atau penilaian objek-objek ke dalam kategori menurut persamaan yang dimilikinya.

6. Asosiasi

   Ini bertanggung jawab untuk menentukan ciri-ciri yang muncul bersamaan. Dalam dunia bisnis, hubungan ini lebih umum disebut sebagai analisis keranjang belanja.

2.2.4 Tahapan Data Mining

Salah satu kebutuhan dalam data mining, terutama dalam implementasi data berukuran besar, adalah adanya pendekatan yang terstruktur dan sistematis tidak hanya selama analisis tetapi juga dalam penyiapan data dan penafsiran hasil untuk memperoleh tindakan atau keputusan yang bermanfaat. Sebagai suatu rangkaian proses, data mining dapat dipecah menjadi beberapa tahap proses yang digambarkan dalam gambar 2.1. Tahap-tahapan tersebut bersifat interaktif, di mana pengguna berpartisipasi secara langsung atau melalui basis pengetahuan (knowledge base).



Gambar 2. 1 Tahapan Data Mining (Sri Diantika et al., 2021)

Tahap-tahap data mining adalah sebagai berikut :

1. Pembersihan data (data cleaning)

   Pembersihan data adalah langkah untuk menghilangkan noise serta informasi yang tidak konsisten atau yang tidak relevan.

2. Integrasi data (data integration)

   Integrasi data adalah penggabungan data dari beberapa data base ke dalam data base baru.

3. Seleksi data (data selection)

   Data yang tersimpan dalam database sering kali tidak sepenuhnya dimanfaatkan, oleh sebab itu hanya informasi yang relevan untuk analisis yang akan diambil dari database.

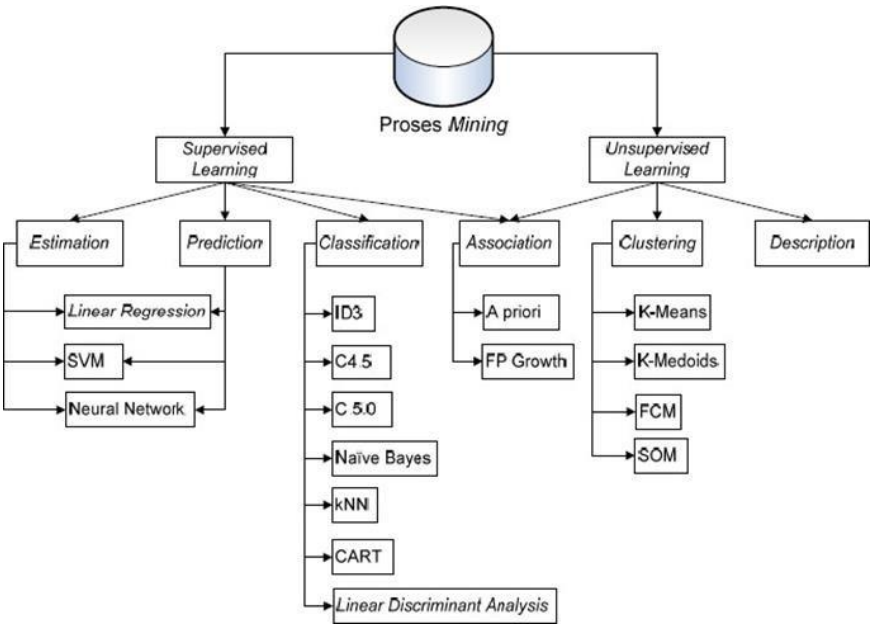
4. Transformasi data (data transformation)

   Data diubah atau digabungkan ke dalam format yang sesuai untuk diproses dalam data mining.

5. Proses Mining

   Merupakan suatu proses utama ketika metode diterapkan untuk menemukan pengetahuan berharga dan tersembunyi dari data.

Beberapa metode yang dapat digunakan berdasarkan pengelompokan data mining dapat dilihat pada Gambar



Gambar 2. 2 Beberapa Metode Data Mining

6. Evaluasi pola (pattern evaluation)

   Untuk mengidentifikasi pola-pola menarik ke dalam pengetahuan berbasis (knowledge based) yang ditemukan.

7. Presentasi pengetahuan (knowledge presentation)

   Merupakan pemetaan atau visualisasi dan penyampaian informasi mengenai teknik yang diterapkan untuk mendapatkan wawasan yang diperoleh oleh pengguna.[10]

2.3 Klasifikasi Text

Klasifikasi merupakan model data mining yang bertujuan untuk menemukan Cara mengidentifikasi perbedaan antara berbagai kategori data dengan maksud untuk memprediksi kategori yang belum diketahui.[6]

Klasifikasi teks adalah proses untuk mengelompokkan teks ke dalam kategori tertentu. Dalam dunia text mining, klasifikasi diartikan sebagai proses untuk meneliti atau mengevaluasi sekumpulan dokumen teks yang telah

dikategorikan sebelumnya. Tujuan dari proses ini adalah untuk menciptakan suatu model atau algoritma yang dapat digunakan untuk mengategorikan dokumen teks yang masih tanpa kategori ke dalam satu atau lebih kategori yang telah ditetapkan sebelumnya.[11]

2.3.1 Proses Klasifikasi

1. Pengumpulan Data: Mengumpulkan informasi yang berkaitan untuk analisis. Informasi ini harus memiliki karakteristik yang dapat digunakan untuk membedakan antara kategori..
2. Text Preprocessing : Merujuk pada serangkaian metode yang umum diterapkan dalam analisis text. Secara teknis, langkah-langkah yang dilakukan meliputi penghapusan kata-kata yang tidak memiliki makna penting (stop words), penyederhanaan huruf (case folding), pemecahan teks menjadi unit-unit kata (tokenizing), dan pengembalian kata ke bentuk dasar (stemming). Proses-proses ini telah teruji dan menjadi standar dalam berbagai penelitian analisis text.[12]
3. Ekstraksi Fitur : Data teks diubah menjadi format yang bisa dipahami oleh algoritma klasifikasi. Beberapa metode ekstraksi fitur yang umum digunakan adalah (*Bag of Words (BoW, TF-IDF (Term Frequency-Inverse Document Frequency, Word Embeddings*).
4. Pembagian Data: Memisahkan dataset menjadi dua bagian, data pelatihan (training set) dan data pengujian (test set). Data pelatihan dimanfaatkan untuk mengembangkan model, sementara data pengujian digunakan untuk menilai performa model.
5. Pelatihan Model: Menggunakan algoritma klasifikasi untuk membangun model berdasarkan data pelatihan. Model ini belajar dari pola dan hubungan antara fitur dan kelas.
6. Pengujian Model: Menggunakan data pengujian untuk menilai kemampuan model dalam memprediksi kategori dari data yang belum pernah dilihat sebelumnya. Hasil evaluasi ini sering kali dinyatakan dalam bentuk metrik kinerja seperti akurasi, precision, recall, dan F1 score.

7. Evaluasi Model : Hasil klasifikasi dievaluasi menggunakan beberapa metrik, seperti (Akurasi, Presisi, Recall, F1-Score.

2.4 Algoritma Klasifikasi

Algoritma klasifikasi merupakan teknik dalam machine learning dan data mining yang berfungsi untuk memprediksi kategori atau label dari informasi berdasarkan pola atau ciri tertentu. Metode ini umumnya diterapkan dalam tugas pengelompokan data ke dalam dua kelas atau lebih, seperti email yang dianggap spam atau bukan, diagnosis kesehatan sehat atau sakit, atau prediksi apakah seorang pelanggan akan setia atau tidak.

2.4.1 CART (Classification and Regression Trees)

CART (Classification And Regression Trees), adalah algoritma yang diperkenalkan oleh Leo Breiman, Jerome H. Friedman, Richard S. Olshen, dan Charles J pada tahun 1984. CART termasuk dalam kategori metode atau teknik algoritma yang digunakan untuk analisis data, yang juga dikenal sebagai pohon keputusan. Metode klasifikasi yang digunakan dalam CART adalah salah satu pendekatan dalam klasifikasi. data mining non-parametrik yang berguna untuk memperoleh sekelompok data akurat sebagai deskriptor/penciri suatu klasifikasi. [4]

Metode CART mencakup dua pendekatan, yaitu Metode pengklasifikasian pohon dan Metode regresi pohon. CART akan membentuk pohon klasifikasi ketika variabel dependen adalah kategoris. Sebaliknya, CART akan menghasilkan pohon regresi jika variabel dependen berupa kontinu atau numerik. Dalam pengembangan algoritma CART, terdapat tiga tahap, yaitu pengembangan pohon klasifikasi, pemangkasan pohon, dan penentuan pohon klasifikasi yang optimal.[4]

CART memiliki kelebihan yang signifikan, seperti kemudahan interpretasi dan kemampuan untuk menangani data yang hilang, ada juga kelemahan yang perlu diperhatikan, seperti risiko overfitting dan ketidakstabilan.

Berikut adalah rumus dan langkah-langkah untuk mengimplementasikan algoritma CART (Classification and Regression Trees) dalam klasifikasi email spam dan ham (non-spam) :

1. Gini Impurity (Ketidakmurnian Gini)

   Gini impurity adalah ukuran ketidakpastian dalam dataset, mirip dengan entropi, tetapi lebih sederhana untuk perhitungan. Rumusnya adalah:

$$Gini\ (S) = 1 - \sum_{i=1}^{c} P(i)^2$$

   Keterangan :

   - S adalah dataset.
   - C adalah jumlah kelas (misalnya, spam dan ham).
   - P (*i*) adalah proporsi dari kelas iii dalam dataset S

2. Penghitungan Gini Impurity untuk Subset

   Ketika dataset dibagi menjadi subset berdasarkan atribut, Gini impurity untuk subset S_v dapat dihitung dengan cara yang sama:

$$Gini\ (S_v) = 1 - \sum_{i=1}^{c} P(i)^2$$

3. Gain Informasi Gini

   Untuk menentukan seberapa baik atribut A membagi dataset S, kita menghitung penurunan Gini impurity yang dihasilkan:

$$Gain\ (S,A) = Gini(S) - \sum_{v \in Values\,(A)} \frac{|S_v|}{|S|}\ Gini\ (S_v)$$

   Keterangan :

   - S_v adalah subset dari dataset S di mana atribut A memiliki nilai *v*.

- |S| adalah jumlah total contoh dalam dataset S.
- $|S_v|$ adalah jumlah contoh dalam subset S_v

2.4.2 C4.5

C4.5 adalah sekumpulan algoritma yang digunakan untuk klasifikasi dalam pembelajaran mesin serta penambangan data. C4.5 bertujuan untuk memahami bagaimana nilai atribut berhubungan dengan kategori, yang dapat diterapkan untuk mengelompokkan item yang tidak dikenal ke dalam kategori yang baru. Algoritma C4.5 dibangun di atas ID3, dengan perbaikan yang mencakup atribut kontinu, nilai atribut, dan pengolahan informasi, yang menghasilkan pohon untuk membuat keputusan melalui proses pemangkasan yang dilakukan dalam dua tahap. Dengan,perhitungan algoritma C4.5, kita dapat menentukan rasio perolehan informasi Gain Ratio untuk setiap atribut.[5] Keunggulannya dalam menangani data berkelanjutan, nilai yang hilang, dan proses pemangkasan membuatnya ideal untuk berbagai aplikasi klasifikasi. Akan tetapi, keterbatasan yang terkait dengan kompleksitas komputasi dan sensitivitas terhadap perubahan data memerlukan perhatian dalam implementasinya, terutama untuk kumpulan data besar atau berdimensi tinggi.

Berikut adalah rumus dan langkah-langkah untuk mengimplementasikan algoritma C4.5 dalam klasifikasi email spam dan ham (non-spam) :

1. Entropy

   Entropy adalah ukuran ketidakpastian dalam data. Rumus untuk menghitung entropy dari dataset S adalah:

$$Entropy(S) = -\sum_{i=1}^{C} P(i) LOG_2 P(i)$$

   Dimana :

   - C adalah jumlah kelas (spam dan ham).
   - P (*i*) adalah proporsi dari kelas i dalam dataset S

2. Gain Informasi

Gain informasi mengukur seberapa banyak informasi yang diperoleh dari atribut A saat membagi dataset S. Rumus untuk menghitung gain informasi adalah: [13]

$$Gain\ (S,A) = Entropy(S) - \sum_{v \in Values\,(A)} \frac{|S_v|}{|S|}\ Entropy\,(S_v)$$

Dimana :

- S : Dataset awal.
- S_v adalah subset dari *S* di mana atribut *A* memiliki nilai *v*.
- |S| adalah jumlah total contoh dalam dataset.
- $|S_v|$ adalah jumlah contoh dalam subset S_v

3. Split Information

$$SplitInfo(S,A) = - \sum_{v \in Values\,(A)} \frac{|S_v|}{|S|}\ log_2 \left(\frac{|S_v|}{|S|}\right)$$

Keterangan :

- S : Dataset utama yang ingin dibagi.
- A : Atribut yang digunakan untuk membagi dataset.
- v ∈ Values (A) : Setiap nilai unik dari atribut A.
- S_v : Subset data di mana atribut A memiliki nilai v.
- $|S|$: Jumlah total data di dataset S.
- S_v.: Jumlah data dalam subset S_v .

4. Rasio gain (gain ratio)

   C4.5 memanfaatkan rasio gain untuk mengatasi bias yang terkait dengan atribut yang memiliki banyak nilai. Rumusnya adalah:

$$Gain\ Ratio\ (S,A) = \frac{IG\ (S,A)}{SplitInfo\ (S,A)}$$

   Keterangan :

   - IG(S, A): Information Gain dari atribut A terhadap dataset S.
   - SplitInfo(S, A): Split Information untuk atribut A.

2.4.3 C5.0

Algoritma C5.0 adalah metode klasifikasi dalam data mining yang diterapkan khususnya pada teknik pohon keputusan dan model berbasis aturan. Algoritma C5.0 adalah pengembangan dan peningkatan dari algoritma C4.5 dan ID3. Algoritma C5.0 memiliki berbagai keunggulan seperti klasifikasi yang serba guna dan berkinerja baik dalam menyelesaikan berbagai jenis permasalahan, kemampuan dalam mengolah fitur numerik dan nominal, kesesuaian untuk dataset kecil maupun besar, serta efisiensi yang lebih tinggi. C5.0 rentan terhadap overfitting jika tidak diatur dengan baik, terutama saat menggunakan boosting dengan jumlah pohon yang berlebihan. Algoritma ini juga mengalami ketidakstabilan, di mana perubahan kecil pada data pelatihan dapat menghasilkan struktur pohon yang berbeda. [6]

Berikut adalah rumus-rumus algoritma C5.0, yang menggabungkan konsep dasar dari algoritma sebelumnya (C4.5) dan memperkenalkan beberapa peningkatan :

1. Entropy

$$Entropy(S) = -\sum_{i=1}^{c} P(i) LOG_2 P\ (i)$$

Keterangan :

- C adalah jumlah kelas (spam dan ham).
- P (*i*) adalah proporsi dari kelas i dalam dataset S

2. Gain Informasi

$$Gain\ (S,A) = Entropy\ (S) - \sum_{v \in Values\,(A)} \frac{|S_v|}{|S|}\ Entropy\ (S_v)$$

Keterangan :

- S_v adalah subset dari *S* di mana atribut *A* memiliki nilai *v*.
- |S| adalah jumlah total contoh dalam dataset.
- $|S_{v|}$ adalah jumlah contoh dalam subset S_v

3. Gain Ratio

$$Gain\ Ratio\ (S,A) = \frac{Gain\ Ratio\ (S,A)}{\text{Entropy (A)}}$$

Keterangan :

- Entropy(A) : Entropy dari atribut A yang dihitung dengan rumus yang sama

4. Gini Impurity (Alternatif)

Sebagai alternatif untuk entropi, C5.0 juga dapat menggunakan Gini impurity:

$$Gini\ (S) = 1 - \sum_{i=1}^{c} P\ (i)^2$$

Keterangan :

- Gini (S): Ketidakmurnian Gini dari dataset S.
- P(*i*)P(: Proporsi dari kelas iii dalam dataset S.

5. Penurunan Gini Impurity

   Untuk menghitung seberapa baik atribut A membagi dataset S menggunakan Gini impurity:

$$Gain\ (S,A) = Gini\ (S) - \sum_{v \in Values\,(A)} \frac{|S_v|}{|S|}\ Gini\ (S_v)$$

2.5 Software Pendukung (*RapidMiner*)

RapidMiner merupakan sistem perangkat lunak yang canggih untuk analisis data dan machine learning. Sistem ini menyediakan sejumlah alat untuk pengolahan, pembentukan model, penilaian, dan penerapan data. *RapidMiner* dibuat dengan antarmuka yang intuitif, memungkinkan pengguna untuk merancang dan menguji berbagai model dengan mudah, bahkan bagi mereka yang tidak memiliki latar belakang pemrograman.

RapidMiner menyediakan sebuah antarmuka yang memungkinkan pengguna menyusun alur kerja dengan cara drag-and-drop untuk menangani serta menganalisis data. *RapidMiner* dapat terhubung dengan berbagai sumber data, termasuk file datar, basis data, serta platform big data seperti Hadoop dan Spark. Perangkat lunak ini juga memiliki sejumlah operator yang terintegrasi, yang berfungsi sebagai komponen utama alur kerja, mencakup setiap langkah dalam proses data mining, seperti pembersihan data, pemilihan fitur, dan pemodelan. *RapidMiner* dikembangkan menggunakan bahasa Java agar dapat berjalan di berbagai sistem operasi.

1. Sejumlah algoritma untuk data mining seperti pohon keputusan dan peta yang bersifat otonom.
2. Visualisasi yang canggih meliputi berbagai grafik seperti histogram tumpang tindih, diagram pohon, serta grafik sebar 3D.
3. Ada beragam pilihan plugin, termasuk plugin teks yang dapat digunakan untuk menganalisis penambangan teks.

4. Menyediakan teknologi terkait penambangan data dan pembelajaran mesin yang mencakup ETL (Extract, Transform, Load), persiapan data awal, visualisasi data, pemodelan, dan evaluasi hasil.
5. Proses penambangan data didefinisikan dalam format XML dan terdiri dari operator yang dapat saling terhubung, diuraikan dengan XML, dan dibangun dalam antarmuka grafis.
6. Menggabungkan proyek penambangan data Weka dengan analisis statistik R.[14]

2.6 Tableu

Tableau adalah sistem visual yang memungkinkan eksplorasi dan analisis data pelanggan secara mendalam. Dikembangkan sebagai kelanjutan komersial dari proyek penelitian Polaris, Tableau kini telah menjadi bagian penting dari tumpukan Business Intelligence (BI) di berbagai organisasi dalam beberapa tahun terakhir (Association for Computing Machinery, ACM Digital, 2011).

Tableau memiliki berbagai kemampuan, termasuk pembuatan peta tematik seperti dalam GIS, manipulasi data layaknya di Excel, plot seri waktu mirip dengan FRED, dan penyusunan infografis yang setara dengan alat infografis khusus. Namun, keunggulan utama Tableau terletak pada kemampuannya mengelola kumpulan data besar untuk menciptakan visualisasi interaktif pada dasbor, memungkinkan pengguna menyelidiki data secara mendetail guna menemukan pola tersembunyi. Selain itu, antarmuka Tableau yang berbasis drag-and-drop sangat ramah bagi pengguna. Bagi lulusan yang memasuki dunia kerja, keterampilan dalam visualisasi data sangat dicari di bidang analitik data.[15]

2.7 Penelitian Terdahulu

Dalam penelitian ini, terdapat penelitian yang telah dilaksanakan sebelumnya dan dapat dijadikan sebagai acuan bagi peneliti. Berikut adalah penelitian sebelumnya yang digunakan sebagai referensi, penelitian terdahulu dapat dilihat pada Tabel 2.1.

Penelitian berjudul "Analisa perbandingan kinerja algoritma C4.5 dan algoritma k-nearest neighbors untuk klasifikasi penerima beasiswa." ini dilakukan oleh Agung Purwanto1), Handoyo Widi Nugroho2), pada tahun 2023. Penelitian ini memperkenalkan pendekatan insentif yang terinspirasi dari pemrograman dinamis sebagai pengganti metode pengambilan keputusan konvensional dalam penempatan beasiswa. Tujuannya adalah untuk mengidentifikasi skema penempatan beasiswa yang terbaik dengan tingkat ekuitas yang paling tinggi, sambil mempertimbangkan batasan praktis dan kebutuhan ekuitas. Penelitian ini menggunakan beberapa data sampel calon penerima dari jurusan, yang mendapatkan hasil perbandingan algoritma, Algoritma K-Nearest Neighbors menunjukkan kinerja yang lebih unggul dengan tingkat presisi 98,08%, akurasi 98,30%, dan nilai recall 98,00%. Hasil AUC yang diperoleh adalah 1,000. Sementara itu, algoritma C4.5 mencatatkan presisi 97,23% dengan nilai precision 94,43%, nilai recall 100,00%, dan AUC sebesar 0,956.

Penelitian yang dilakukan oleh Chindu Lintang Bhuana1 , Handoyo Widi Nugroho2, pada tahun 2023, dengan judul "Perbandingan decision tree C4.5 dan Support Vector Machine (svm) dalam klasifikasi penderita stroke berbasis pso". Fokus penelitian ini adalah peningkatan akurasi klasifikasi penderita stroke dengan mengoptimalkan pemilihan fitur menggunakan teknik feature selection atau feature engineering. Selain itu, penelitian juga berfokus pada perbandingan antara algoritma Decision Tree C4.5 dan Support Vector Machine (SVM) untuk mengevaluasi kinerja keduanya dalam menangani dataset yang lebih besar atau beragam, serta menjajaki penerapan teknik pembelajaran mesin lainnya seperti ensemble methods untuk meningkatkan hasil klasifikasi.

Penelitian yang dilakukan oleh Halimah 1 , Deppi Linda 2 , Felisita Klaralia 3, pada tahun 2020, dengan judul "Penerapan Algoritma Naïve Bayes untuk Memprediksi Penyakit Malaria pada Puskesmas Hanura". Penelitian ini memiliki tujuan untuk menciptakan sebuah sistem untuk memprediksi penyakit malaria di Puskesmas Hanura dengan menggunakan pendekatan Naïve Bayes serta memanfaatkan kumpulan data rekam medis dari pasien. Fokusnya adalah pada analisis gejala yang paling relevan untuk meningkatkan akurasi prediksi, serta penerapan model Naïve Bayes untuk memprediksi kemungkinan penderita malaria berdasarkan gejala yang ada. Sistem ini membantu pihak medis memberikan dugaan sementara, mempercepat penanganan atau pencegahan sebelum pemeriksaan medis lebih lanjut, serta meningkatkan efisiensi layanan kesehatan, mengurangi ketergantungan pada diagnosa manual, dan mendukung pengambilan keputusan medis yang lebih tepat di wilayah dengan kasus malaria tinggi.

Penelitian yang dilakukan oleh Putra Kurniawan 1 , Wasilah2 , Sutedi 3 , Handoyo Widi Nugroho 4, pada tahun 2024, dengan judul "Implementasi Data Mining Dalam Klasifikasi Tingkat Kesenjangan Kompetensi PNS Menggunakan Metode Naive Bayes". Penelitian ini berfokus pada evaluasi kinerja algoritma Naïve Bayes dalam mengklasifikasikan level ketidakmerataan kompetensi Pegawai Negeri Sipil di Pemerintah Provinsi Lampung. Data yang digunakan berasal dari penilaian mandiri yang dikumpulkan melalui kuesioner yang didasarkan pada kamus kompetensi. Tujuan penelitian ini adalah untuk mengevaluasi keefektifan metode Naïve Bayes untuk mengukur level kekurangan keterampilan PNS sebagai landasan dalam strategi peningkatan keterampilan. Temuan penelitian menampilkan bahwa algoritma Naïve Bayes dapat menghasilkan klasifikasi yang akurat dengan tingkat akurasi sebesar 98,02%. Manfaat dari hasil klasifikasi ini adalah dapat digunakan oleh Pemerintah Provinsi Lampung dalam merencanakan pengembangan kompetensi PNS, dengan tujuan meningkatkan kualitas dan efisiensi pelayanan publik.

Penelitian yang dilakukan oleh Robby Anggriawan1 , Handoyo Widi Nugroho 2,

pada tahun 2023, dengan judul "Komparasi algoritma C4.5 dan Naive Bayes dalam prediksi penderita penyakit gagal jantung". Fokus penelitian ini adalah untuk analisis dan pencegahan penyakit kardiovaskuler, yang merupakan penyebab kematian utama akibat Penyakit Tidak Menular (PTM), dengan mempertimbangkan faktor risiko seperti hipertensi, penyakit jantung koroner, dan stroke. Penelitian ini berfokus pada pengembangan sistem prediksi atau deteksi dini penyakit kardiovaskuler menggunakan data kesehatan, seperti riwayat medis, gaya hidup, dan faktor lingkungan, yang dapat membantu dalam mengidentifikasi individu memiliki risiko yang signifikanMelalui metode ini, diharapkan angka kematian awal akibat penyakit jantung dan pembuluh darah dapat diminimalisir, terutama di negara-negara dengan pendapatan rendah dan menengah.

Penelitian yang dilakukan oleh Fida Maisa Hanaa 1, Widya Cholid Wahyudin 2, Saiful Ulyac, Deka Setia Negarad 3, pada tahun 2023, dengan judul "Implementasi algoritma cart dalam klasifikasi penyakit diabetes". Penelitian ini berfokus pada penerapan algoritma CART (Classification and Regression Trees) untuk klasifikasi penyakit diabetes dengan tujuan untuk memanfaatkan teknik data mining dalam pencatatan dan pencegahan penyakit tersebut. Hasil penelitian menunjukkan bahwa algoritma CART yang tidak melalui proses pruning dan prepruning memperoleh akurasi 100%, sementara penerapan pruning dan prepruning memberikan akurasi sebesar 96,15%. Keuntungan dari penelitian ini adalah kemampuannya untuk mengidentifikasi pasien yang berisiko diabetes dengan lebih cepat dan tepat, yang mendukung upaya pencegahan serta penanganan yang lebih awal. Kesimpulannya, metode CART efektif dalam klasifikasi penyakit diabetes dan dapat diterapkan dalam sistem kesehatan untuk meningkatkan manajemen penyakit ini.

Penelitian yang dilakukan oleh Rahmanita Widiyanti 1, Cucu Suhery 2, Rahmi Hidayati 3, pada tahun 2022, dengan judul "Implementasi Algoritma C5.0 Untuk Klasifikasi Kepuasan Masyarakat Terhadap Pelayanan Kantor Kecamatan". Penelitian ini bertujuan untuk mengklasifikasikan kepuasan masyarakat terhadap

pelayanan publik di Kantor Kecamatan Pontianak Kota menggunakan algoritma C5.0. Dengan menggunakan data sebanyak 250 sampel, penelitian ini mengidentifikasi unsur penilaian yang mempengaruhi kepuasan, aspek kesesuaian pelayanan, pemahaman prosedur, kecepatan pelayanan, biaya, standar pelayanan, keterampilan petugas, perilaku staf, fasilitas dan infrastruktur, serta pengaduan layanan. Sebagai hasilnya, algoritma C5.0 membentuk pohon keputusan dan aturan yang menunjukkan bahwa keterampilan petugas adalah faktor utama yang memengaruhi kepuasan. Model ini menghasilkan tingkat akurasi 90,67%, presisi 68,085%, recall 55,89%, dan tingkat kesalahan 9,33%. Penelitian ini menawarkan solusi untuk meningkatkan kualitas layanan dengan menekankan pada elemen-elemen yang berpengaruh terhadap kepuasan publik.

Penelitian yang dilakukan oleh Sadimin 1 , Handoyo Widi Nugroho 2, pada tahun 2023 , dengan judul "Perbandingan kinerja algoritma datamining untuk Prediksi kelulusan mahasiswa". Penelitian ini fokus pada perbandingan kinerja algoritma dalam prediksi terkait data mahasiswa di organisasi pendidikan, khususnya dalam prediksi kinerja mahasiswa, penerima beasiswa, dan kelulusan mahasiswa. Tujuan dari penelitian ini adalah untuk mengevaluasi dan mengkomperasi algoritma C4.5 serta Naïve Bayes dengan memperhatikan akurasi, precision, recall, dan AUC guna menemukan pendekatan mana yang lebih efisien dalam mengklasifikasikan data mahasiswa. Hasil dari penelitian ini menunjukkan bahwa C4.5 mencatat akurasi sebesar 79,91%, precision 89,06%, dan recall 81,38%. Di sisi lain, Naïve Bayes memiliki akurasi 76,95%, precision 75,95%, recall 98,38%, serta AUC 0,838. Dari sini, dapat disimpulkan bahwa algoritma C4.5 unggul dalam akurasi, sedangkan Naïve Bayes memiliki tingkat recall yang lebih tinggi dan AUC yang sedikit lebih baik, yang mencerminkan perbedaan dalam kemampuan masing-masing algoritma dalam mengolah data. Penelitian ini memberikan pemahaman yang berguna bagi institusi pendidikan dalam menentukan algoritma yang paling sesuai untuk prediksi berdasarkan data.

Penelitian yang dilakukan oleh Lintang Gladyza Adinda Putri 1, Satrio A.

Wicaksono 2, Bayu Rahayudi 3, pada tahun 2025, dengan judul "Perbandingan kinerja algoritma datamining untuk Prediksi kelulusan mahasiswa". Penelitian ini fokus pada deteksi spam email dengan memanfaatkan teknik Extreme Gradient Boosting (XGBoost) untuk mengategorikan email sebagai spam atau bukan spam. Tujuan utama dari penelitian ini adalah untuk mengembangkan suatu sistem yang dapat mengidentifikasi spam dengan tingkat ketepatan yang tinggi, serta menilai kinerja model melalui Stratified K-Fold Cross Validation dan Confusion Matrix. Temuan dari penelitian ini menunjukkan bahwa model XGBoost mencapai akurasi sebesar 95,3%, precision 95,1%, recall 95,6%, dan F1-score 95,2%, dengan tingkat kesalahan klasifikasi yang cukup rendah. Kesimpulannya, dapat diartikan bahwa metode Extreme Gradient Boosting sangat berhasil dalam mendeteksi email yang dikategorikan sebagai spam, dengan proporsi yang memadai antara deteksi email spam dan menghindari kesalahan dalam klasifikasi email yang non-spam. Penelitian ini mengarah pada penerapan XGBoost sebagai solusi yang tepat dalam menangani masalah spam email secara efektif.

Penelitian yang dilakukan oleh Yulia Verawati 1, Muhammad Said Hasibuan 2. pada tahun 2021, dengan judul "Perbandingan Data Set IRIS Dengan Aplikasi RapidMiner dan Orange menggunakan Algoritma Klasifikasi". Penelitian ini fokus pada perbandingan performa algoritma decision tree dalam klasifikasi data menggunakan dua tools yang berbeda, yaitu RapidMiner dan Orange, dengan dataset Iris yang memiliki atribut numerik dan class nominal. Tujuan penelitian ini adalah untuk membandingkan hasil klasifikasi dan performa algoritma decision tree pada kedua tools tersebut. Hasil penelitian menunjukkan bahwa RapidMiner menghasilkan accuracy 86,67%, classification error 13,33%, kappa 0,800, serta weighted mean recall 86,67% dan weighted mean precision 90,48%, sementara Orange menghasilkan nilai klasifikasi untuk atribut-atribut seperti petal length dan petal width yang bervariasi berdasarkan spesies Iris. Kesimpulannya, kedua tools, RapidMiner dan Orange, dapat digunakan untuk melakukan klasifikasi data Iris dengan hasil yang berbeda, menunjukkan kelebihan dan kekurangan masing-masing alat, dan memberikan wawasan dalam pemilihan tool

yang tepat untuk aplikasi klasifikasi menggunakan algoritma decision tree.

Tabel 2. 1 Penelitian Terdahulu

No	Judul	Metode	Tujuan	Penulis/Tahun	Hasil
1.	Analisa perbandingan kinerja algoritma c4.5 dan algoritma k-nearest neighbors untuk klasifikasi penerima beasiswa.[5]	Penelitian ini mengusulkan metode insentif yang terinspirasi oleh pemrograman dinamis untuk menggantikan proses pengambilan keputusan tradisional dalam penugasan beasiswa.	Tujuan dari pelaksanaan penelitian ini Tujuannya adalah untuk mengidentifikasi skema pembagian beasiswa yang paling efisien dengan tingkat keadilan yang paling tinggi. sambil memperhitungkan kendala praktis dan persyaratan ekuitas	Agung Purwanto1), Handoyo Widi Nugroho2) / 2023	Berdasarkan hasil penelitian, dapat disimpulkan bahwa algoritma K-Nearest Neighbors menunjukkan kinerja yang lebih unggul, yakni dengan presisi 98,08%, akurasi 98,30%, serta nilai recall 98,00%, dengan hasil AUC mencapai 1,000. Sementara itu, algoritma C4.5 memperoleh presisi sebesar 94,43%, dengan nilai recall 100,00% dan AUC yang tercatat sebesar 0,956.
2.	Perbandingan decision tree C4.5 dan	Algoritma C4.5 dan support	Tujuan penelitian ini untuk menganalisis dan	Chindu Lintang Bhuana1 , Handoyo Widi	Dari riset yang telah dilakukan, pengembangan

	support vector machine (svm) dalam klasifikasi penderita stroke berbasis pso.[10]	vector machine (svm)	membandingkan Pohon Keputusan C4.5 dan Support Vector Machine (SVM) Berbasis PSO untuk menguji perbandingan tingkat akurasi dan presisi dalam klasifikasi penyakit stroke .	Nugroho2 / 2023	menggunakan algoritma C4.5 dan Support Vector Machine yang berbasis pada PSO serta memanfaatkan data dari pasien yang mengalami stroke. Model yang dihasilkan dibandingkan untuk menentukan metode yang paling efektif dengan mengukur performa ketiga metode tersebut melalui confusion matrix dan kurva ROC. Algoritma C4.5 yang didasarkan pada PSO memperoleh tingkat akurasi sebesar 96,11 persen dan nilai AUC sebesar 0,827. Sementara itu, metode Support Vector Machine

					yang berbasis PSO mendapatkan akurasi sebesar 95,33 persen serta nilai AUC 0,867. Tingkat akurasi tertinggi dicapai oleh algoritma C4.5, sedangkan nilai AUC tertinggi dimiliki oleh algoritma Support Vector Machine yang berbasis PSO. Jadi, dapat disimpulkan bahwa metode Support Vector Machine dan algoritma C4.5 yang dikembangkan dengan Particle Swarm Optimization merupakan metode yang paling unggul.
3.	Penerapan Algoritma Naïve Bayes untuk Memprediksi Penyakit	Algoritma Naive Bayes	Penelitian ini bertujuan Untuk mencegah malaria sejak dini. dilakukan prediksi penyakit	Halimah 1 , Deppi Linda 2 , Felisita Klaralia 3 / 2020	Hasil dari penelitian ini adalah Metode Naïve Bayes dapat digunakan

	Malaria pada Puskesmas Hanura.[16]		malaria dengan metode Naïve Bayes pada Puskesmas Hanura.		untuk memprediksi kemungkinan terjadinya penyakit malaria di Puskesmas Hanura. Dengan penerapan teknik data mining, maka pihak Puskesmas dapat mengolah data pasien dan mendapatkan informasi secara mudah.
4.	Implementasi Data Mining Dalam Klasifikasi Tingkat Kesenjangan Kompetensi PNS Menggunakan Metode Naive Bayes.[17]	Algoritma Naive Bayes	Penelitian ini bertujuan untuk menguji penerapan metode Naïve Bayes dalam klasifikasi tingkat kesenjangan kompetensi PNS di Pemerintah Provinsi Lampung. Melalui pendekatan ini, diharapkan dapat diperoleh tingkat akurasi yang optimal dalam mengidentifikasi kesenjangan	Putra Kurniawan 1 , Wasilah2 , Sutedi 3 , Handoyo Widi Nugroho 4 / 2024	Hasil dari penelitian yaitu Model klasifikasi kesenjangan kompetensi PNS di Pemerintah Provinsi Lampung menggunakan Naïve Bayes dalam RapidMiner mencapai akurasi 98,02%, menunjukkan kinerjanya yang

			kompetensi PNS.		baik.
5.	Komparasi algoritma C45 dan naive bayes dalam prediksi penderita penyakit gagal jantung.[13]	Algoritma Decision Tree C4.5 dan Naive Bayes	Penelitian ini bertujuan untuk menyelesaikan permasalahan prediksi penyakit jantung menggunakan data mining klasifikasi. Diperlukan metode yang dapat mengolah data yang tersedia, salah satunya dengan teknik klasifikasi. Penelitian ini juga bertujuan untuk menentukan algoritma yang paling akurat dalam prediksi penyakit jantung.	Robby Anggriawan1 , Handoyo Widi Nugroho 2 / 2023	Berdasarkan pengukuran kinerja dengan membandingkan dua algoritma berdasarkan jumlah data, dapat disimpulkan bahwa algoritma Decision Tree memiliki kemampuan dalam pengambilan keputusan untuk menentukan penderita gagal jantung.
6.	Implementasi algoritma CART dalam klasifikasi penyakit diabetes.[4]	Algoritma CART	Tujuan penelitian untuk mengimplementasi Algoritma CART (Classification and Regression Trees) digunakan dalam klasifikasi penyakit diabetes untuk membangun model keputusan yang akurat dalam mengidentifikasi pola dan	Fida Maisa Hanaa 1, Widya Cholid Wahyudin 2, Saiful Ulyac, Deka Setia Negarad 3 / 2023	Hasil dari penelitian ini yaitu menghasilka nilai Akurasi tertinggi dalam klasifikasi penyakit diabetes diperoleh saat menggunakan algoritma CART dengan pruning dan

			menentukan hasil diagnosis.		prepruning, mencapai 100%.
7.	Implementasi Algoritma C5.0 Untuk Klasifikasi Kepuasan Masyarakat Terhadap Pelayanan Kantor Kecamatan.[6]	Algoritma C5.0	Tujuan penelitian ini membangun sistem yang diimplementasikan menggunakan website.	Rahmanita Widiyanti 1, Cucu Suhery 2,, Rahmi Hidayati 3 /2022	Hasil dari penelitian ini adalah Unsur penilaian yang paling berpengaruh terhadap kepuasan masyarakat dalam pelayanan menggunakan algoritma C5.0 adalah kemampuan petugas, yang diperoleh dari node akar pohon keputusan algoritma C5.0.
8.	Perbandingan kinerja algoritma datamining untuk Prediksi kelulusan mahasiswa.[18]	Algoritma C4.5 dan Naïve Bayes	Tujuan penelitian ini adalah untuk menguji dan membandingkan algoritma C4.5 dan Naïve Bayes dalam hal akurasi, precision, recall, dan AUC, guna menentukan metode yang lebih efektif dalam mengklasifikasikan	Sadimin1) , Handoyo Widi Nugroho2)	Hasil penelitian menunjukkan bahwa C4.5 memiliki akurasi 79,91%, precision 89,06%, dan recall 81,38%, sementara Naïve Bayes memiliki akurasi 76,95%, precision

			data mahasiswa.		75,95%, recall 98,38%, dan AUC 0,838. Kesimpulannya, algoritma C4.5 unggul dalam akurasi, sedangkan Naïve Bayes memiliki recall yang lebih tinggi dan AUC yang sedikit lebih baik, menunjukkan perbedaan kekuatan kedua algoritma dalam memproses data.
9.	Analisis Klasifikasi Spam Email Menggunakan Metode Extreme Gradient Boosting (XGBoost). [19]	Metode Extreme Gradient Boosting	Tujuan penelitian ini adalah untuk mengembangkan sistem deteksi spam yang dapat mengklasifikasikan email dengan akurat, serta untuk mengevaluasi Performa model diuji menggunakan Stratified K-Fold Cross Validation dan Confusion Matrix untuk mengevaluasi akurasi, precision,	Lintang Gladyza Adinda Putri 1 , Satrio A. Wicaksono 2 , Bayu Rahayudi 3, pada tahun 2025	Hasil penelitian menunjukkan bahwa model XGBoost memiliki akurasi 95,3%, precision 95,1%, recall 95,6%, dan F1-score 95,2%, dengan kesalahan klasifikasi yang relatif kecil. Kesimpulannya, metode Extreme

			recall, dan metrik lainnya dalam klasifikasi data.		Gradient Boosting terbukti efektif dalam mengklasifikasikan spam email dengan keseimbangan yang baik antara mengenali email spam dan menghindari kesalahan klasifikasi pada email non-spam.
10.	Perbandingan Data Set IRIS Dengan Aplikasi Rapid Miner dan Orange menggunakan Algoritma Klasifikasi. [20]	Algoritma Decision Tree	ujuan penelitian ini adalah untuk membandingkan hasil klasifikasi dan performa algoritma decision tree pada kedua tools yaitu *RapidMiner* dan *Orange*.	Yulia Verawati 1, Muhammad Said Hasibuan 2. pada tahun 2021	Hasil penelitian menunjukkan bahwa RapidMiner menghasilkan accuracy 86,67%, classification error 13,33%, kappa 0,800, serta weighted mean recall 86,67% dan weighted mean precision 90,48%, sementara Orange menghasilkan nilai klasifikasi

					untuk atribut-atribut seperti petal length dan petal width yang bervariasi berdasarkan spesies Iris.

Bersadarkan penelitian terdahulu tersebut maka penulis melakukan penelitian yang berjudul "Perbandingan Kinerja Algoritma CART, C4.5 dan C5.0 Dalam Klasfikasi Email Spam Ham (Non-Spam)" dengan menggunakan tools *Rapidminer* , Tableau untuk visualisasi dan Microsoft Excel 2010 untuk pehitungan manual.