BAB III
METODOLOGI PENELITIAN

3.1 Metode Penelitian

Metode yang diterapkan dalam penelitian ini melibatkan penggunaan tiga algoritma, yaitu CART (Classification And Regression Trees), C4.5, dan C5.0, untuk melakukan klasifikasi terhadap email spam dan ham (non-spam). Dan nantinya akan dilakukan perbandingan nilai akurasi diantara ketiga algoritma ini, sehingga dapat menghasilkan pengetahuan algoritma mana yang efektif untuk proses klasifikasi email spam dan ham (non-spam). Dataset text email spam dan ham (non-spam) yang sudah terlabel digunakan untuk menjadi bahan pengolahan data dalam penerapan ketiga algoritma ini dalam klasifikasi email spam dan ham(non-spam).

3.2 Teknik Pengumpulan Data

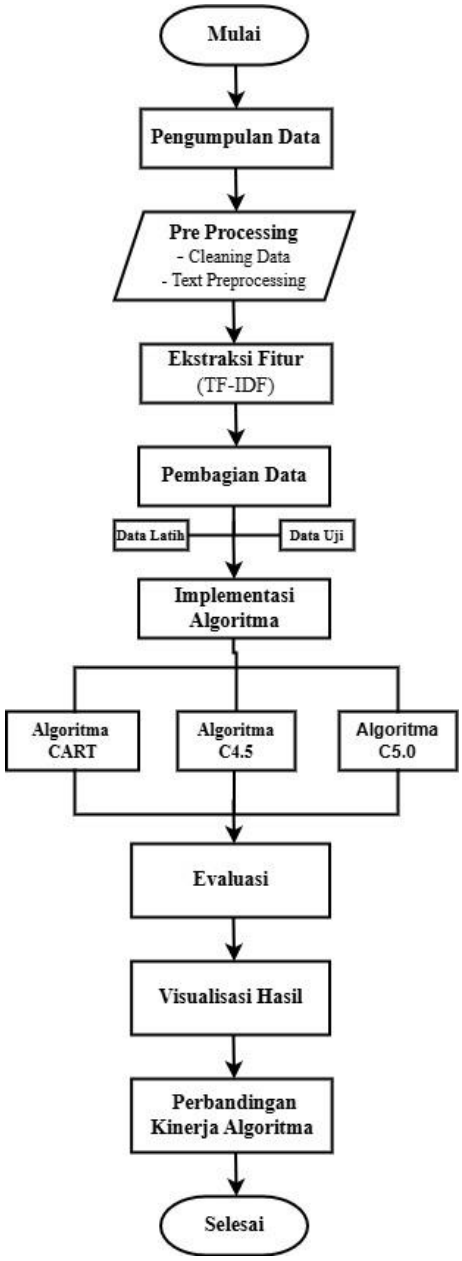
Penelitian ini menggunakan data berupa teks email spam dan ham (non-spam) terlabel yang diperoleh dari repository *Kaggle*, sebuah platform yang menyediakan dataset publik untuk berbagai keperluan penelitian dan pengembangan model. Dataset ini dipilih karena memuat informasi yang relevan dengan topik perbandingan algoritma data mining dalam klasifikasi email spam dan ham (non-spam).

Dataset email spam dan ham (non-spam) yang sudah terlabel diambil dari Kaggle melalui langkah-langkah berikut :

1. Dataset email spam dan ham (non-spam) diunduh melalui platform Kaggle di *kaggle.com*.
2. Dataset yang digunakan untuk penelitian ini ditemukan melalui pencarian dengan kata kunci "*Indonesian Email Spam*". Dataset yang terpilih merupakan kumpulan text email yang terlabel spam dan ham.
3. Dataset ini diperoleh dari *kaggle.com*, dengan rincian sebagai berikut:
    - Nama Dataset : Indonesian Email Spam
    - Jumlah Data : Terdapat 2.638 data
    - Atribut Data : Terdapat dua atribut data yang diberikan.

3.3 Tahapan Penelitian

Tahapan penelitian yang akan dilakukan oleh peneliti dalam proses penelitian terdiri dari beberapa langkah. Langkah-langkah penelitian merupakan prosedur yang dilakukan oleh peneliti untuk memperoleh hasil yang diinginkan. Dalam metode ini memiliki beberapa tahapan atau alur penelitian adalah sebagai berikut:



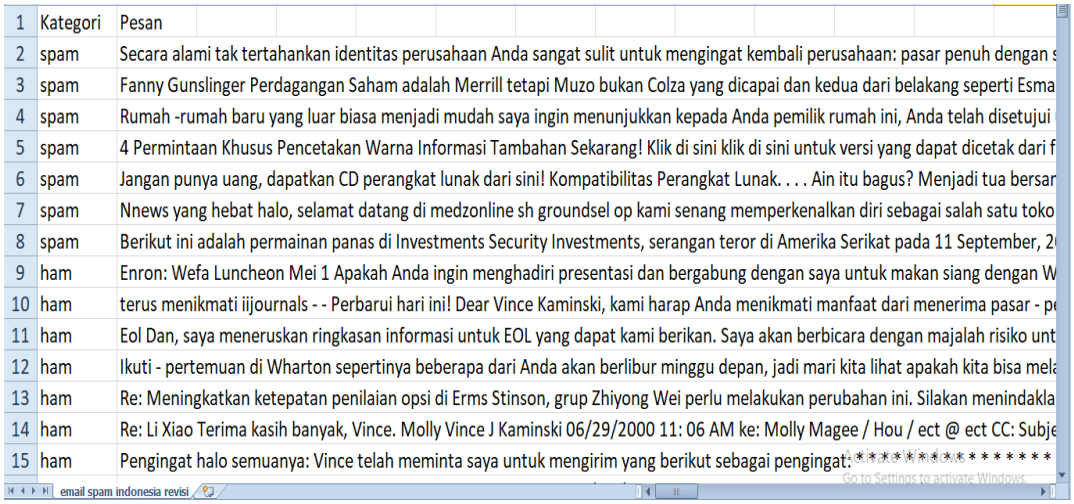
Gambar 3. 1 Tahapan Penelitian

3.3.1 Studi Literatur

Untuk tahapan yang pertama peneliti melakukan proses studi literatur, yaitu mencari, membahas dan memahami penelitian sebelum yang berkaitan dengan objek yang akan dibahas.

3.3.2 Pengumpulan Data

Untuk membandingkan kinerja algoritma CART (Classification And Regression Trees), C4.5, dan C5.0 dalam klasifikasi email spam dan ham (non-spam), dibutuhkan data yang cukup banyak dan relevan. Dalam penelitian ini, peneliti menggunakan dataset yang berisi 2.638 data teks email spam dan ham (non-spam) yang sudah terlabel dari repository Kaggle.



	Kategori	Pesan
2	spam	Secara alami tak tertahankan identitas perusahaan Anda sangat sulit untuk mengingat kembali perusahaan: pasar penuh dengan s
3	spam	Fanny Gunslinger Perdagangan Saham adalah Merrill tetapi Muzo bukan Colza yang dicapai dan kedua dari belakang seperti Esma
4	spam	Rumah -rumah baru yang luar biasa menjadi mudah saya ingin menunjukkan kepada Anda pemilik rumah ini, Anda telah disetujui
5	spam	4 Permintaan Khusus Pencetakan Warna Informasi Tambahan Sekarang! Klik di sini klik di sini untuk versi yang dapat dicetak dari f
6	spam	Jangan punya uang, dapatkan CD perangkat lunak dari sini! Kompatibilitas Perangkat Lunak. . . . Ain itu bagus? Menjadi tua bersar
7	spam	Nnews yang hebat halo, selamat datang di medzonline sh groundsel op kami senang memperkenalkan diri sebagai salah satu toko
8	spam	Berikut ini adalah permainan panas di Investments Security Investments, serangan teror di Amerika Serikat pada 11 September, 2
9	ham	Enron: Wefa Luncheon Mei 1 Apakah Anda ingin menghadiri presentasi dan bergabung dengan saya untuk makan siang dengan W
10	ham	terus menikmati iijournals - - Perbarui hari ini! Dear Vince Kaminski, kami harap Anda menikmati manfaat dari menerima pasar - p
11	ham	Eol Dan, saya meneruskan ringkasan informasi untuk EOL yang dapat kami berikan. Saya akan berbicara dengan majalah risiko unt
12	ham	Ikuti - pertemuan di Wharton sepertinya beberapa dari Anda akan berlibur minggu depan, jadi mari kita lihat apakah kita bisa mel
13	ham	Re: Meningkatkan ketepatan penilaian opsi di Erms Stinson, grup Zhiyong Wei perlu melakukan perubahan ini. Silakan menindakla
14	ham	Re: Li Xiao Terima kasih banyak, Vince. Molly Vince J Kaminski 06/29/2000 11: 06 AM ke: Molly Magee / Hou / ect @ ect CC: Subje
15	ham	Pengingat halo semuanya: Vince telah meminta saya untuk mengirim yang berikut sebagai pengingat: **********

email spam indonesia revisi

Gambar 3. 2 Data Mentah

Deskripsi Dataset :

Dataset ini diambil dari *https://www.kaggle.com/ datasets/ mfaisalqureshi/spam-email/* dan diterjemahkan ke dalam Bahasa Indonesia oleh penerbit. Dataset ini terdiri dari 2638 data yang terdiri dari 1372 pesan spam dan 1266 pesan non-spam (ham) 52%:48%.

3.3.3 Pre Processing

Tahapan yang selanjutnya adalah preprocessing data, dimana tahapan ini adalah tahapan terpenting dalam pengolahan data untuk memastikan bahwa informasi tersebut siap untuk diproses dan dianalisis oleh model pembelajaran mesin,

termasuk algoritma CART yang merupakan Pohon Klasifikasi dan Regresi, C4.5, serta C5.0. Preprocessing data bertujuan untuk membersihkan, mempersiapkan, dan memformat data sedemikian rupa agar model bisa memanfaatkan informasi yang relevan dari data tersebut secara efektif.

Tahapan Preprocessing :

1. Cleaning Data (Pembersihan data)

   Dilakukan tahap pembersihan data mentah untuk memastikan kualitas data yang diterapkan dalam analisis. Tujuan dari data *cleaning* adalah menghilangkan atau memperbaiki kesalahan, data yang kurang lengkap, dan data yang kurang relevan agar hasil analisis lebih akurat.

2. Text Preprocessing (Pembersihan Teks Email)

   Untuk mempersiapkan teks ulasan menjadi format yang sesuai untuk analisis atau pemodelan dengan menghapus unsur-unsur yang tidak diperlukan.

   Berikut tahapan Text Preprocessing :

   - Lowercasing / Transform Cases

     Lowercasing / Transform Cases adalah Proses konversi huruf teks menjadi huruf kecil untuk memastikan konsistensi dan memudahkan pencocokan string, terutamanya ketika perbedaan antara huruf kapital dan huruf kecil tidak memiliki arti yang penting.

Tabel 3. 1 Proses Lowercasing / Transform Cases

SEBELUM	SESUDAH
Pengingat halo semuanya: Vince telah meminta saya untuk mengirim yang berikut sebagai pengingat: * Ini adalah pengingat lain tentang tanggung jawab yang kami miliki sebagai kelompok untuk membersihkan ruang konferensi setelahnya rapat. Bagi banyak pengunjung untuk Enron ruang konferensi, mungkin paparan pertama	pengingat halo semuanya: vince telah meminta saya untuk mengirim yang berikut sebagai pengingat: * ini adalah pengingat lain tentang tanggung jawab yang kami miliki sebagai kelompok untuk membersihkan ruang konferensi setelahnya rapat. bagi banyak pengunjung untuk enron ruang konferensi, mungkin paparan pertama

dan terakhir ke perusahaan. Kamis lalu, beberapa anggota kelompok pergi, meninggalkan kotak makan siang mereka di atas meja. Tolong, jangan membuat saya malu dengan harus menunjukkan kepada Anda bahwa ibu Anda tidak bekerja di sini. Vince	dan terakhir ke perusahaan. kamis lalu, beberapa anggota kelompok pergi, meninggalkan kotak makan siang mereka di atas meja. tolong, jangan membuat saya malu dengan harus menunjukkan kepada anda bahwa ibu anda tidak bekerja di sini. vince

- Tokenization (Tokenisasi)

  Tokenisasi merupakan langkah yang memisahkan tulisan menjadi bagian-bagian kecil, misalnya kata - kata atau tanda baca. Metode ini mendukung pemrosesan bahasa alami dengan memecah teks menjadi bagian-bagian yang lebih mudah untuk diolah.

Tabel 3. 2 Proses Tokanize

SEBELUM	SESUDAH
Pengingat halo semuanya: Vince telah meminta saya untuk mengirim yang berikut sebagai pengingat: * * * * * * * * * * * * * * * * Ini adalah pengingat lain tentang tanggung jawab yang kami miliki sebagai kelompok untuk membersihkan ruang konferensi setelahnya rapat. Bagi banyak pengunjung untuk Enron ruang konferensi, mungkin paparan pertama dan terakhir ke perusahaan. Kamis lalu, beberapa anggota kelompok pergi, meninggalkan kotak makan siang mereka di atas meja. Tolong, jangan	“pengingat”, "halo", "semuanya", "vince", "telah", "meminta", "saya", "untuk", "mengirim", "yang", "berikut", "sebagai", "pengingat", "ini", "adalah", "pengingat", "lain", "tentang", "tanggung", "tanggung", "jawab", "yang", "kami", "miliki", "sebagai", "kelompok", "untuk", "membersihkan", "ruang", "konferensi", "setelahnya", "rapat", ".", "bagi", "banyak", "pengunjung", "untuk", "enron", "ruang", "konferensi", "mungkin", "paparan", "pertama", "dan", "terakhir", "ke", "perusahaan", ".", "kamis", "lalu",

membuat saya malu dengan harus menunjukkan kepada Anda bahwa ibu Anda tidak bekerja di sini. Vince	",", "beberapa", "anggota", "kelompok", "pergi", ",", "meninggalkan", "kotak", "makan", "siang", "mereka", "di", "atas", "meja", ".", "tolong", ",", "jangan", "membuat", "saya", "malu", "dengan", "harus", "menunjukkan", "kepada", "anda", "bahwa", "ibu", "anda", "tidak", "bekerja", "di", "sini", ".", "vince"

- Stopword Removal

  Stopwords removal merupakan langkah untuk menghapus istilah yang tidak terlalu penting atau tidak berhubungan dari kumpulan kata-kata yang tidak memiliki arti.

Tabel 3. 3 Proses Stopword Removal

SEBELUM	SESUDAH
Pengingat halo semuanya: Vince telah meminta saya untuk mengirim yang berikut sebagai pengingat: * Ini adalah pengingat lain tentang tanggung jawab yang kami miliki sebagai kelompok untuk membersihkan ruang konferensi setelahnya rapat. Bagi banyak pengunjung untuk Enron ruang konferensi, mungkin paparan pertama dan terakhir ke perusahaan. Kamis lalu, beberapa anggota kelompok pergi, meninggalkan kotak makan siang mereka di atas meja. Tolong, jangan membuat saya malu dengan harus menunjukkan kepada Anda bahwa ibu Anda tidak bekerja di sini. Vince	"pengingat", "vince", "meminta", "kirim", "pengingat", "pengingat", "tanggung", "tanggung", "jawab", "kelompok", "membersihkan", "ruang", "konferensi", "rapat", "pengunjung", "enron", "ruang", "konferensi", "paparan", "pertama", "terakhir", "perusahaan", "kamis", "anggota", "kelompok", "pergi", "meninggalkan", "kotak", "makan", "siang", "meja", "tolong", "malu", "menunjukkan", "ibu", "bekerja", "vince"

- Filter tokens

  Filter tokens bertujuan untuk penyaringan token (kata atau unit teks) berdasarkan panjangnya untuk menghapus kata-kata yang sangat pendek atau sangat panjang yang dianggap tidak berhubungan atau tidak relevan.

Tabel 3. 4 Proses Filter tokens

SEBELUM	SESUDAH
Pengingat halo semuanya: Vince telah meminta saya untuk mengirim yang berikut sebagai pengingat: * Ini adalah pengingat lain tentang tanggung jawab yang kami miliki sebagai kelompok untuk membersihkan ruang konferensi setelahnya rapat. Bagi banyak pengunjung untuk Enron ruang konferensi, mungkin paparan pertama dan terakhir ke perusahaan. Kamis lalu, beberapa anggota kelompok pergi, meninggalkan kotak makan siang mereka di atas meja. Tolong, jangan membuat saya malu dengan harus menunjukkan kepada Anda bahwa ibu Anda tidak bekerja di sini. Vince	"Pengingat", "Vince", "meminta", "kirim", "pengingat", "tanggung", "tanggung", "jawab", "kelompok", "membersihkan", "ruang", "konferensi", "rapat", "pengunjung", "Enron", "ruang", "konferensi", "paparan", "pertama", "terakhir", "perusahaan", "Kamis", "anggota", "kelompok", "pergi", "meninggalkan", "kotak", "makan", "siang", "meja", "tolong", "malu", "menunjukkan", "ibu", "bekerja", "Vince".

3.3.4 Ekstraksi Fitur

Setelah data dibersihkan dan di-tokenisasi, setiap pesan email spam dan ham (non-spam) harus diubah menjadi bentuk yang bisa dipahami oleh algoritma CART (Classification And Regression Trees). Biasanya, dilakukan Ekstraksi Fitur menggunakan teknik TF-IDF atau Bag of Words.

Pada penelitian ini diterapkan pembobotan fitur TF-IDF. TF-IDF, yang merupakan kepanjangan dari Term Frequency-Inverse Document Frequency,

adalah metode dalam analisis teks dan informasi yang berfungsi untuk mengevaluasi sejauh mana pentingnya sebuah kata dalam suatu dokumen dibandingkan dengan kumpulan dokumen yang ada. Term Frequency (TF) mengukur frekuensi kemunculan suatu kata dalam dokumen, memberikan bobot lebih tinggi pada kata-kata yang muncul dengan frekuensi tinggi. Sedangkan, Inverse Document Frequency (IDF) menghitung seberapa jarang suatu kata muncul di antara semua dokumen dalam korpus, dengan memberikan bobot yang lebih besar pada kata-kata yang tidak sering muncul. Penggabungan kedua elemen ini, TF-IDF, menghasilkan skor yang mewakili relevansi suatu kata dalam dokumen tertentu sambil mengurangi dampak dari kata-kata umum seperti "dan" atau "adalah".

Misalnya, dalam dataset email, kata "diskon" mungkin sering dijumpai dalam email spam, tapi jarang ditemui di email non-spam. Jadi, TF-IDF akan memberi bobot tinggi pada kata "diskon" dalam email spam dan membantu model mengenali bahwa email dengan kata-kata seperti ini kemungkinan besar adalah spam.

3.3.5 Pembagian Data

Setelah tahap preprocessing, langkah selanjutnya adalah pembagian data. Dalam penelitian ini, metode yang digunakan untuk membagi data adalah *k-Fold Cross Validation*. Teknik ini digunakan untuk mengevaluasi model secara lebih akurat dengan memisahkan dataset menjadi beberapa bagian (fold) yang digunakan secara bergantian untuk pelatihan dan pengujian model. Proses ini membantu dalam meminimalkan bias yang bisa terjadi akibat pembagian data yang kurang merata atau tidak representatif.

Pada penelitian ini dilakukan *5-Fold Cross Validation* pada dataset yang berjumlah 2.638 data. Dengan menggunakan metode ini, Dataset ini dibagi menjadi lima bagian, di mana setiap bagian secara bergantian berfungsi sebagai data pengujian, sedangkan empat bagian lainnya digunakan sebagai data pelatihan.

Proses Cross Validation :

- Dataset yang digunakan berjumlah 2.638 data. Data ini dibagi menjadi 5 fold yang masing-masing berisi sekitar 528 data (karena 2.638 dibagi 5 adalah sekitar 528).
- Pembagian data dilakukan secara acak dengan memastikan bahwa setiap fold memiliki proporsi data yang representatif dari keseluruhan dataset.

Proses Iterasi (Iterasi 1, 2, 3, 4, 5) :

- Data latih : 2.110 data
- Data uji : 528 data

Setiap iterasi menghasilkan nilai akurasi berdasarkan data uji. Rata-rata akurasi dari kelima iterasi dihitung untuk memberikan gambaran umum mengenai kinerja model. Pembagian data menggunakan *5-Fold Cross Validation* ini memastikan bahwa setiap data dipakai dalam tahap pelatihan dan pengujian, yang mengurangi bias dan memberikan evaluasi yang lebih stabil terhadap performa model.

3.3.6 Implementasi Algoritma CART, C4.5, C5.0

Implementasi algoritma CART, C4.5, C5.0 adalah proses penerapan algoritma untuk membangun model klasifikasi. Untuk tahapannya adalah pelatihan model**,** pengujian model atau klasifikasi, dan evaluasi.

1. Pelatihan Model , pada proses ini algoritma menggunakan data latih untuk mempelajari pola. Dan model akan menemukan pola dan hubungan antara fitur dan label yang ada. Sehingga akan menghasilkan kemampuan untuk membuat prediksi yang digunakan dalam tahap klasifikasi.
2. Pengujian Model/Klasifikasi, Setelah model dilatih pengujian model dilakukan menggunakan data uji atau biasa disebut dengan data baru, Ini berarti model akan memprediksi label dari data yang belum pernah dilihat sebelumnya berdasarkan pola yang telah dipelajari. Misalnya, dalam klasifikasi email, model dapat menentukan apakah email baru adalah spam atau ham (non-spam).

3. Evaluasi, Evaluasi adalah proses di mana menganalisis kinerja model setelah pelatihan untuk memahami seberapa baik model tersebut bekerja pada data baru/dalam konteks klasifikasi. Dengan perhitungan seperti akurasi, akurasi adalah metrik yang digunakan untuk menilai kinerja model klasifikasi. Rumus untuk menghitung akurasi adalah sebagai berikut:

$$Akurasi = \frac{Jumlah\ Prediksi\ Benar}{Total\ Prediksi} = \frac{TP + TN}{TP + TN + FP + FN}$$

Keterangan :

- TP (True Positives): Jumlah prediksi positif yang benar (misalnya, jumlah email spam yang benar-benar diklasifikasikan sebagai spam).
- TN (True Negatives): Jumlah prediksi negatif yang benar (jumlah email ham yang benar-benar diklasifikasikan sebagai ham).
- FP (False Positives): Jumlah prediksi positif yang salah (jumlah email ham yang salah diklasifikasikan sebagai spam).
- FN (False Negatives): Jumlah prediksi negatif yang salah (jumlah email spam yang salah diklasifikasikan sebagai ham).

3.3.7 Visualisasi Hasil

Visualisasi adalah proses menyajikan data atau informasi dalam bentuk grafis seperti grafik, diagram, atau peta untuk memudahkan pemahaman dan analisis. Tujuannya adalah membantu pengguna melihat pola, tren, atau perbandingan yang mungkin sulit dikenali dari data mentah. Dengan visualisasi, data kompleks bisa ditampilkan dengan lebih ringkas, jelas, dan mudah diinterpretasi.

Contoh visualisasi meliputi grafik batang, diagram garis, pie chart, atau heatmap**.** Dalam konteks perbandingan kinerja algoritma, visualisasi seperti Grouped Bar Chart (grafik batang berkelompok) digunakan untuk memudahkan pemahaman mengenai perbedaan kinerja setiap algoritma berdasarkan beberapa metrik seperti akurasi. Dengan demikian, pengguna bisa cepat menarik kesimpulan dan mengambil keputusan berdasarkan data yang tersedia.

3.3.8 Perbandingan Kinerja Algoritma

Setelah melakukan evaluasi kinerja algoritma, seperti CART, C4.5, dan C5.0 untuk klasifikasi email spam dan ham, langkah selanjutnya adalah membandingkan hasil kinerja setiap algoritma**.** Nilai yang dibandingkan adalah nilai akurasi disetiap algoritma, untuk melihat nilai akurasi algoritma mana yang paling tinggi.

3.3.9 Perhitungan Manual Algoritma (CART, C4.5 dan C5.0)

Perhitungan secara manual dalam penggunaan algoritma pembelajaran machine learning merujuk pada cara menghitung setiap langkah algoritma secara langsung, tanpa bergantung pada aplikasi atau library tertentu. Tujuan utamanya adalah untuk mendapatkan pemahaman yang lebih mendalam tentang cara kerja algoritma dan bagaimana model mengambil keputusan berdasarkan data yang diberikan. Dengan melakukan perhitungan manual, kita dapat lebih jelas melihat proses pengolahan data, pemilihan fitur, serta cara model belajar dari data untuk membuat prediksi. Selain itu, cara ini juga membantu memperjelas pemahaman konsep-konsep dasar sebelum menerapkan algoritma secara otomatis dengan menggunakan alat atau library yang lebih canggih. Berikut perhitungan manual setiap algoritma, dengan menggunakan data sampel dan tools Microsoft Excel 2010 :

	A	B	C	D
1	vince	Kategori		
2	0.0	spam	Total Data	1842
3	0.0	spam	Spam	1095
4	0.0	spam	Ham	747
5	0.0	spam		
6	0.0	spam		
7	0.0	spam		
8	0.0	spam		
9	0.0	spam		
10	0.0	spam		
11	0.0	spam		
12	0.0	spam		
13	0.2043100	ham		
14	0.0645785	ham		
15	0.0498582	ham		
16	0.0224185	ham		
17	0.1941571	ham		
18	0.1092432	ham		
19	0.0386876	ham		
20	0.0825336	ham		
21	0.0	ham		
22	0.1504637	ham		
23	0.0	spam		
24	0.0	spam		
25	0.0	spam		

Gambar 3. 3 Data Sampel Perhitungan Manual

1. Algortima CART (*Classification and Regression Tree*)

CART (Classification and Regression Tree) merupakan algoritma yang menciptakan pohon keputusan dengan memanfaatkan Gini Index untuk memilih split terbaik pada setiap node. Gini Index berfungsi untuk menilai impurity data dan dipakai untuk membagi data menjadi subset yang lebih seragam. Tahapan ini terus berlangsung shingga pohon terbentuk, yang selanjutnya digunakan untuk tujuan klasifikasi. Hasil akhir dari klasifikasi berupa kategori (seperti spam atau bukan spam).

Berikut adalah rumus dan langkah-langkah untuk mengimplementasikan algoritma CART (Classification and Regression Trees), secara manual menggunakan data sampel dengan tools excel 2010 :

- Gini Impurity (Ketidakmurnian Gini),

  Gini impurity adalah ukuran ketidakpastian dalam dataset. Rumusnya adalah:

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

  Langkah pertama mencari nilai Proposi Spam dan Ham, dengan rumus :

  = Jumlah Spam / Total data

  = Jumlah Ham / Total Data

  Untuk memperjelas hasil dari pehitungan proporsi spam dan ham bisa dilihat pada Gambar 3.4.

F	G
Proporsi Spam dan Ham	
Spam	Ham
0.594462541	0.405537459

Gambar 3. 4 Proporsi Spam dan ham (CART)

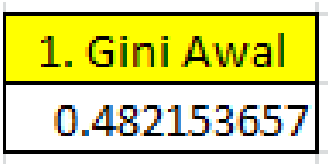
Selanjutnya mencari nilai Gini atau Gini awal.

Rumus :

= - 1 (Proporsi Spam Log 2+ Proporsi Ham Log2)

= - 1 (0.594462541^2 + 0.405537459^2)

Gini = 0.482153657



Gambar 3. 5 Nilai Gini

- Penghitungan Gini Impurity untuk Subset

  Dataset dibagi menjadi subset berdasarkan atribut, Gini impurity untuk setap subset, subset 1 dan subset 2 dapat dihitung dengan cara yang sama.

  Rumus :

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

  Tahapan pertama adalah menetapkan nilai batas untuk memisahkan data menjadi dua bagian. Dalam proses pemisahan data pehitungan manual CART, nilai 0.1 dijadikan sebagai batasan. Jika nilai Vince kurang dari 0.1, maka data tersebut akan ditempatkan dalam subset 1, sebaliknya, jika nilai Vince lebih dari 0.1, data akan dimasukkan ke dalam subset 2.

  Rumus :

  Subset 1

  = IF (value (0.0) < 0.1 ; ” Subset 1” ; “”)

  Jika nilai 0.0 kurang dari 0.1, maka masukkan ke dalam subset 1. Jika nilai 0.0 lebih besar dari 0.1, maka tidak ada subset yang relevan atau kosong.

  Subset 2

  = IF (value (0.0) > 0.1 ; ” Subset 2” ; “”)

Jika nilai 0.0 lebih besar dari 0.1, maka masukkan ke dalam subset 2. Jika nilai 0.0 lebih kecil dari 0.1, maka tidak ada subset yang relevan atau kosong.

vince	Kategori	Subset 1 (<0.1)	Subset 1 (>0.1)
0.0	spam	Subset 1	
0.0	spam	Subset 1	
0.0	spam	Subset 1	
0.0	spam	Subset 1	
0.0	spam	Subset 1	
0.0	spam	Subset 1	
0.0	spam	Subset 1	
0.0	spam	Subset 1	
0.0	spam	Subset 1	
0.0	spam	Subset 1	
0.199556494	ham		Subset 2
0.051520289	ham	Subset 1	
0.028028008	ham	Subset 1	
0.0	ham	Subset 1	
0.140979895	ham		Subset 2
0.0	ham	Subset 1	
0.056281491	ham	Subset 1	
0.101765894	ham		Subset 2
0.153155790	ham		Subset 2
0.209408621	ham		Subset 2
0.061666363	ham	Subset 1	
0.103576251	ham		Subset 2
0.114835209	ham		Subset 2

Gambar 3. 6 Pembagian Data Gini Impurity 0.1 (CART)

Pada Gambar 3.6 ini adalah hasil dari pemisahan data dengan batas nilai 0.1.

Selanjutnya menentukan nilai Gini Index untuk setiap subset.

Subset 1

vince	Kategori	Jumlah Spam	Jumlah Ham
0.0	spam	1	0
0.0	spam	1	0
0.0	spam	1	0
0.0	spam	1	0
0.0	spam	1	0
0.0	spam	1	0
0.0	spam	1	0
0.0	spam	1	0
0.0	spam	1	0
0.0	spam	1	0
0.0	ham	0	1
0.0637747	ham	0	1
0.0	ham	0	1
0.0936587	ham	0	1
0.0401918	ham	0	1
0.0504007	ham	0	1
0.0	ham	0	1
0.0459733	ham	0	1
0.0560010	ham	0	1
0.0507370	ham	0	1
0.0	ham	0	1
0.0	ham	0	1
0.0194176	ham	0	1
0.0975691	ham	0	1
0.1351475	ham	0	0
Jumlah		1094	624

Gambar 3. 7 Hasil Pembagian Data Subset 1 (CART)

Gambar 3.7 memperlihatkan hasil pembagian data pada subset 1, di mana data subset 1 yang termasuk dalam kategori spam diberikan nilai 1 di kolom jumlah spam, sementara data yang bukan spam mendapat nilai 0. Hal yang sama berlaku untuk kategori Ham.

Langkah selanjutnya menghitung Proporsi Spam dan Ham.

Rumus :

Spam dan ham = Jumlah Spam / Total Data Spam dan Ham subset 1.

Spam = 1094 / 1718 = 0.636786962

Ham = 624 / 1718 = 0.363213038

Untuk memperjelas hasil proporsi spam dan ham subset 1 bisa dilhat pada Gambar 3.8.

A. Proporsi Spam dan Ham

Subset 1	
Spam	0.63678696
Ham	0.36321304

Gambar 3. 8 Proporsi Subset 1

Selanjutnya menghitung nilai gini pada subset 1 dengan Rumus :

Gini Subset 1 = 1 - Jumlah spam pangkat 2 + Jumlah Ham pangkat 2

Gini Subset 1 = 1 - (0.636786962^2+0.33876221^2)

= 1 - 0.462579

= 0.537421

0.537421	Gini Subset 1

Gambar 3. 9 Hasil Gini Subset 1

Subset 2

vince	Kategori	Jumlah Spam	Jumlah Ham
0.0	spam	0	0
0.0	spam	0	0
0.0	spam	0	0
0.0	spam	0	0
0.0	spam	0	0
0.0	spam	0	0
0.0	spam	0	0
0.0	spam	0	0
0.0	spam	0	0
0.0	spam	0	0
0.0	spam	0	0
0.1581609	ham	0	1
0.2198307	ham	0	1
0.0686006	ham	0	0
0.0	ham	0	0
0.0430797	ham	0	0
0.0265389	ham	0	0
0.0	ham	0	0
0.1218433	ham	0	1
0.0586649	ham	0	0
0.1092927	ham	0	1
0.0783729	ham	0	0
0.0	ham	0	0
Jumlah		0	127

Gambar 3. 10 Hasil Pembagian Data Subset 2 (CART)

Gambar 3.10 memperlihatkan hasil pembagian data pada subset 2, di mana data subset 2 yang termasuk dalam kategori spam diberikan nilai 1 di kolom jumlah spam, sementara data yang bukan spam mendapat nilai 0. Hal yang sama berlaku untuk kategori Ham.

Selanjutnya sama seperti subset 1, menghitung proporsi spam dan ham. Dengan Rumus :

Spam dan ham = Jumlah Spam / Total Data Spam dan Ham subset 2.

Spam = 0.1 / 127 = 0.00078678

Ham = 127 / 127 = 0.99921322

Subset 2	
Spam	0.00078678
Ham	0.99921322

Gambar 3. 11 Proporsi Subset 1

Selanjutnya menghitung nilai gini pada subset 2 dengan Rumus :

Gini Subset 2 = 1 - Jumlah spam pangkat 2 + Jumlah Ham pangkat 2

Gini Subset 2 = 1 - (0^2+1^2)

= 1 - 0

= 1

1	Gini Subset 2

Gambar 3. 12 Hasil Gini Subset 2

- Gini Index Setelah Pemisahan (Weighted Gini Index)

  Rumus :

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

4. Gini Index Setelah Pemisahan (Weighted Gini Index)	
Rumus :	
GiniAfter Split= (Total_Subset1 / Total_Data) * Gini_Subset1 + (Total_Subset2 / Total_Data) * Gini_Subset2	
Hasil =	
0.570189619	

Gambar 3. 13 Weighted Gini Index

Pada Gambar 3.13 menggunakan Rumus :

GiniAfter Split = (Total_Subset1 / Total_Data) * Gini_Subset1 + (Total_Subset2 / Total_Data) * Gini_Subset2

= (1718 / 1842) * 0.537421 + (127 / 1842) * 1

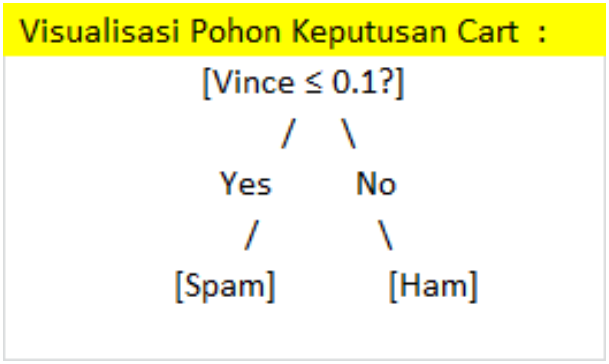
Weighted Gini Index = 0.570189619

- Hitung Gain (Reduction in Impurity)

5. Hitung Gain (Reduction in Impurity)	
Rumus :	
Gain = = Gini_Awal - Gini_After_Split	
Hasil =	
-0.088035962	

Gambar 3. 14 Hasil Reduction in Impurity

- Visualisasi Pohon Keputusan CART



Gambar 3. 15 Pohon Keputusan Algoritma CART

Gambar 3.15 menunjukkan pohon keputusan yang dihasilkan oleh algoritma CART berdasarkan perhitungan secara manual. Dalam pohon tersebut, vince muncul sebagai node akar pertama karena memiliki nilai Gini Index yang paling tinggi. Jika nilai vince berada di bawah 0,1, maka data akan diklasifikasikan sebagai spam. Sebaliknya, jika nilai vince melebihi 0,1, maka data tersebut akan diklasifikasikan sebagai ham.

2. Algoritma C4.5

Algoritma C4.5 merupakan metode pembuatan pohon keputusan yang diciptakan oleh Ross Quinlan sebagai perbaikan dari ID3. Algoritma C4.5 membangun pohon keputusan dengan memilih atribut yang memiliki Gain Ratio tertinggi untuk menjadi root, kemudian membagi data ke dalam cabang sesuai dengan nilai dari atribut tersebut. Proses ini dilakukan secara rekursif sampai semua data dalam cabang memiliki kelas yang sama atau tidak ada atribut yang tersisa. Apabila terdapat data yang hilang, C4.5 mengatasi hal ini dengan memangkas untuk mencegah overfitting.

Berikut adalah rumus dan langkah-langkah untuk mengimplementasikan algoritma C4.5, secara manual menggunakan data sampel dengan tools Microsoft Excel 2010 :

- Entropy

  Entropy adalah ukuran ketidakpastian dalam data.

  Rumus :

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

  Langkah pertama mencari nilai Proposi Spam dan Ham, dengan rumus :

  = Jumlah Spam / Total data

  = Jumlah Ham / Total Data

  Untuk memperjelas hasil dari pehitungan proporsi spam dan ham bisa dilihat pada Gambar 3.16.

Hitung Probabilitas (p_i)	
Spam	0.594463
Ham	0.405537

Gambar 3. 16 Proporsi Spam dan ham (C4.5)

  Selanjutnya menghitung Log 2 pada masing – masing proporsi spam dan ham.

  Rumus :

  = Log (nilai proporsi spam atau ham ; 2)

  Untuk hasilnya ada pada Gambar 3.17 dibawah ini.

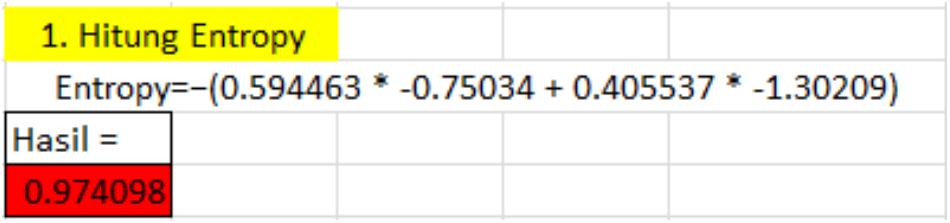
Hitung Log^2 :	
Spam	-0.75034
Ham	-1.30209

Gambar 3. 17 Nilai Log2 Spam dan Ham (C4.5)

Tahap selanjutnya menghitung nilai Entropy, dengan rumus :

Entropy = - (Jumlah Proporsi Spam * Nilai Log2 spam + Jumlah Proporsi Ham * Nilai Log2 Ham

Untuk memperjelas pehitungan entropy ada pada Gambar 3.17



Gambar 3. 18 Entropy C4.5

- Gain Informasi

Gain informasi mengukur seberapa banyak informasi yang diperoleh dari atribut A saat membagi dataset S.

Rumus :

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

Langkah pertama adalah membagi data berdasarkan fitur vince, dengan menetapkan nilai batas untuk memisahkan data menjadi dua bagian. Dalam proses pemisahan data pehitungan manual C4.5, nilai 0.1 dijadikan sebagai batasan. Jika nilai Vince kurang dari 0.1, maka data tersebut akan ditempatkan dalam subset 1, sebaliknya, jika nilai Vince lebih dari 0.1, data akan dimasukkan ke dalam subset 2.

Rumus :

Subset 1

= IF (value (0.0) < 0.1 ; ” Subset 1” ; “”)

Jika nilai 0.0 kurang dari 0.1, maka masukkan ke dalam subset 1. Jika nilai 0.0 lebih besar dari 0.1, maka tidak ada subset yang relevan atau kosong.

Subset 2

= IF (value (0.0) > 0.1 ; ” Subset 2” ; “”)

Jika nilai 0.0 lebih besar dari 0.1, maka masukkan ke dalam subset 2. Jika nilai 0.0 lebih kecil dari 0.1, maka tidak ada subset yang relevan atau kosong.

vince	Kategori	Subset 1 (<0.1)	Subset 1 (>0.1)
0.0	spam	Subset 1	
0.0	spam	Subset 1	
0.0	spam	Subset 1	
0.0	spam	Subset 1	
0.0	spam	Subset 1	
0.0	spam	Subset 1	
0.0	spam	Subset 1	
0.0	spam	Subset 1	
0.0	spam	Subset 1	
0.0	spam	Subset 1	
0.199556494	ham		Subset 2
0.051520289	ham	Subset 1	
0.028028008	ham	Subset 1	
0.0	ham	Subset 1	
0.140979895	ham		Subset 2
0.0	ham	Subset 1	
0.056281491	ham	Subset 1	
0.101765894	ham		Subset 2
0.153155790	ham		Subset 2
0.209408621	ham		Subset 2
0.061666363	ham	Subset 1	
0.103576251	ham		Subset 2
0.114835209	ham		Subset 2

Gambar 3. 19 Pembagian Data Gain Informasi 0.1 (C4.5)

Gambar 3.19 adalah hasil dari pemisahan data dengan batas nilai 0.1.

Selanjutnya menentukan nilai Gain Informasi untuk setiap subset.

Subset 1

vince	Kategori	Jumlah Spam	Jumlah Ham
0.0	spam	1	0
0.0	spam	1	0
0.0	spam	1	0
0.0	spam	1	0
0.0	spam	1	0
0.0	spam	1	0
0.0	spam	1	0
0.0	spam	1	0
0.0	spam	1	0
0.0	spam	1	0
0.0	ham	0	1
0.0637747	ham	0	1
0.0	ham	0	1
0.0936587	ham	0	1
0.0401918	ham	0	1
0.0504007	ham	0	1
0.0	ham	0	1
0.0459733	ham	0	1
0.0560010	ham	0	1
0.0507370	ham	0	1
0.0	ham	0	1
0.0	ham	0	1
0.0194176	ham	0	1
0.0975691	ham	0	1
0.1351475	ham	0	0
Jumlah		1094	624

Gambar 3. 20 Hasil Pembagian Data Subset 1 (C4.5)

Pada Gambar 3.20 memperlihatkan hasil pembagian data pada subset 1, di mana data subset 1 yang termasuk dalam kategori spam diberikan nilai 1 di kolom jumlah spam, sementara data yang bukan spam mendapat nilai 0. Hal yang sama berlaku untuk kategori Ham.

Langkah selanjutnya menghitung Proporsi Spam dan Ham.

Rumus :

Spam dan ham = Jumlah Spam / Total Data Spam dan Ham subset 1.

Spam = 1094 / 1718 = 0.636786962

Ham = 624 / 1718 = 0.363213038

A. Proporsi Spam dan Ham

Subset 1	
Spam	0.63678696
Ham	0.36321304

Gambar 3. 21 Proporsi Subset 1 (C4.5)

Selanjutnya menghitung nilai Log2 pada subset 1.

Rumus :

= LOG (Proporsi Spam/Ham ; 2)

Spam = LOG (0.63678696 ; 2)

Ham = LOG (0.36321304 ; 2)

Untuk hasilnya bisa dilihat pada Gambar 3.22.

B. Hitung Log^2 :

Subset 1	
Spam	-0.6511173
Ham	-1.4611121

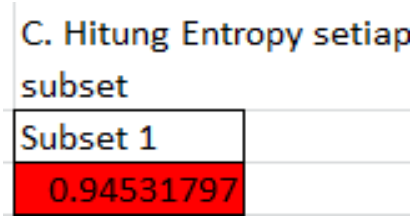
Gambar 3. 22 Nilai Log2 Subset 1 (C4.5

Hitung nilai entropy pada subset 1, dengan rumus :

Entropy = - (Proporsi Spam * Nilai Log2 Spam + Proporsi Ham * Nilai Log2 Ham)

= - (0.63678696 * - 0.6511173 + 0.36321304 * - 1.4611121)

Hasilnya bisa dilihat pada Gambar 3.23.



Gambar 3. 23 Nilai Entropy Subset 1 (C4.5)

Subset 2

vince	Kategori	Jumlah Spam	Jumlah Ham
0.0	spam	0	0
0.0	spam	0	0
0.0	spam	0	0
0.0	spam	0	0
0.0	spam	0	0
0.0	spam	0	0
0.0	spam	0	0
0.0	spam	0	0
0.0	spam	0	0
0.0	spam	0	0
0.0	spam	0	0
0.1581609	ham	0	1
0.2198307	ham	0	1
0.0686006	ham	0	0
0.0	ham	0	0
0.0430797	ham	0	0
0.0265389	ham	0	0
0.0	ham	0	0
0.1218433	ham	0	1
0.0586649	ham	0	0
0.1092927	ham	0	1
0.0783729	ham	0	0
0.0	ham	0	0
Jumlah		0	127

Gambar 3. 24 Hasil Pembagian Data Subset 2 (C4.5)

Gambar 3.24 memperlihatkan hasil pembagian data pada subset 2, di mana data subset 2 yang termasuk dalam kategori spam diberikan nilai 1 di kolom jumlah spam, sementara data yang bukan spam mendapat nilai 0. Hal yang sama berlaku untuk kategori Ham.

Selanjutnya sama seperti subset 1, menghitung proporsi spam dan ham. Dengan Rumus :

Spam dan ham = Jumlah Spam / Total Data Spam dan Ham subset 2.

Spam = 0.1 / 127 = 0.00078678

Ham = 127 / 127 = 0.99921322

Subset 2	
Spam	0.00078678
Ham	0.99921322

Gambar 3. 25 Proporsi Subset 2 (C4.5)

Selanjutnya menghitung nilai Log2 pada subset 2.

Rumus :

= LOG (Proporsi Spam/Ham ; 2)

Spam = LOG (0.00078678 ; 2)

Ham = LOG (0.99921322 ; 2)

Untuk hasilnya bisa dilihat pada Gambar 3.26

Subset 2	
Spam	-10.311748
Ham	-0.0011355

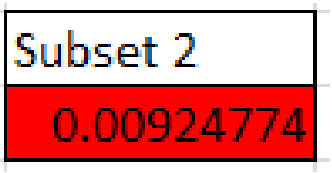
Gambar 3. 26 Nilai Log2 Subset 1 (C4.5

Hitung nilai entropy pada subset 2, dengan rumus :

Entropy = - (Proporsi Spam * Nilai Log2 Spam + Proporsi Ham * Nilai Log2 Ham)

= - (0.00078678 * - 10.311748 + 0.99921322 * - 0.0011355)

Hasilnya bisa dilihat pada Gambar 3.27.



Gambar 3. 27 Nilai Entropy Subset 2 (C4.5)

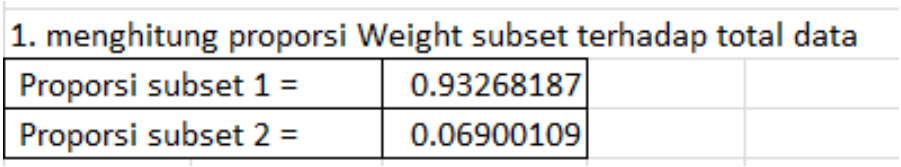
Selanjutnya Hitung weight entropy subset untuk masing- masing subset.

Langkah pertama hitung proporsi subset 1 dan subset 2.

Rumus :

Proporsi Subset = Jumlah spam dan ham subset 1 dan subset 2/ Total data

Untuk hasilnya ada pada Gambar 3.28.

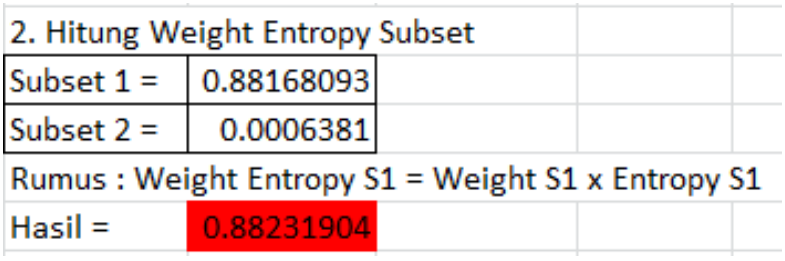


Gambar 3. 28 Proporsi Subset 1.2 (C4.5)

Selanjutnya hitung Weight Entropy setiap Subset, dengan rumus :

= Weight Entropy S1 = Weight S1 x Entropy S1

Untuk lebih memperjelas hasil Weight Entropy setiap Subset bisa dilihat pada Gambar 3.29.



Gambar 3. 29 Nilai Weight Entropy Subset 1,2 (C4.5)

Hitung Information Gain (IG) / Gain Informasi, dengan rumus :

= Information Gain=Entropy Total−Total Weight Entropy

3. Hitung Information Gain (IG):	
Information Gain=Entropy Total−Total Weight Entropy	
Information Gain =	0.09177874

Gambar 3. 30 Nilai IG (C4.5)

Gambar 3.30 adala nilai Information Gain yang berjumlah 0.09177874.

- Split Information

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

Langkah pertama menghitung Log2 untuk setiap proporsi, dengan rumus :

= LOG (Proporsi subset 1 atau subset 2 ; 2)

Log^2 Proporsi	
Subset 1 =	-0.100543025
Subset 2 =	-3.857237126

Gambar 3. 31 Nilai Log2 Subset 1,2 (C4.5)

Selanjutnya menghitung nilai Kontribusi Split Information untuk setiap subset.

Rumus :

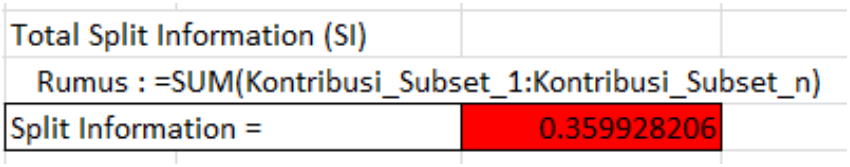
= Nilai proporsi subset 1 atau 2 * Nilai Log2 subset 1 atau subset 2

Untuk lebih memperjelas hasil Kontribusi Split Information , bisa dilihat pada Gambar 3.32

Kontribusi Split Information	
Subset 1 =	-0.093774656
Subset 2 =	-0.26615355

Gambar 3. 32 Nilai Kontribusi (C4.5)

Tahap selanjutnya menghitung total Split Information (SI), dengan rumus : =SUM(Kontribusi_Subset_1:Kontribusi_Subset_n)



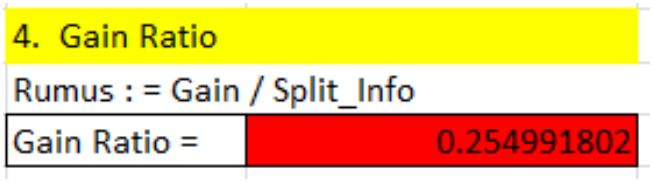
Gambar 3. 33 Nilai Split Information (C4.5)

Gambar 3.33 adalah nilai dari Split Information (SI), dengan jumlah 0.359928206.

- Gain Ratio

  Rumus :

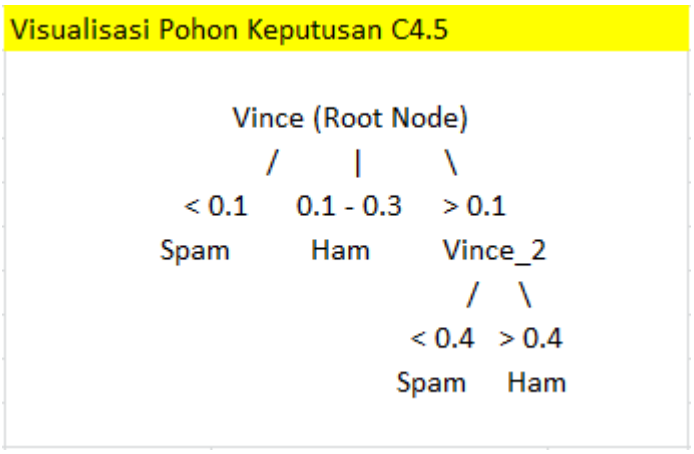
$$Gain\ Ratio\ (S,A) = \frac{IG\ (S,A)}{SplitInfo\ (S,A)}$$



Gambar 3. 34 Nilai Gain ratio (C4.5)

Pada Gambar 3.34, ditampilkan nilai Gain Ratio dari algoritma C4.5. Nilai ini digunakan untuk menentukan root (akar) pertama dalam pembentukan pohon keputusan.

- Visualisasi pohon keputusan algoritma C4.5



Gambar 3. 35 Pohon Keputusan Algoritma C4.5

Gambar 3.35 menunjukkan pohon keputusan yang dihasilkan oleh algoritma C4.5 berdasarkan perhitungan secara manual. Dalam pohon tersebut, vince muncul sebagai node akar pertama karena memiliki nilai Gain Ratio tertinggi Jika nilai vince berada di bawah 0,1, maka data akan diklasifikasikan sebagai spam. Sebaliknya, Jika nilai tersebut berada dalam kisaran 0.1 hingga 0.3, maka akan diklasifikasikan sebagai Ham. Namun, jika nilainya lebih besar dari 0.1, data akan dialihkan untuk diproses lebih lanjut di node Vince_2. Di node ini, terdapat dua pembagian lainnya: jika nilainya kurang dari 0.4, maka akan dikategorikan sebagai Spam, sedangkan jika lebih dari 0.4, akan diklasifikasikan sebagai Ham.

3. Algoritma C5.0

Algoritma C5.0 menentukan root (akar) pertama dalam pohon keputusan dengan memilih atribut yang memiliki Gain Ratio paling tinggi. Langkah ini diawali dengan penghitungan entropi untuk menilai ketidakpastian dalam data, diikuti dengan perhitungan keuntungan informasi dari setiap atribut. Selanjutnya, nilai Gain Ratio dihitung dengan memperhatikan pembagian data yang lebih seimbang. Atribut yang memiliki Gain Ratio tertinggi ditetapkan sebagai akar, karena memberikan pemisahan terbaik untuk proses klasifikasi. Proses ini memastikan bahwa pemilihan akar memaksimalkan informasi yang diperoleh dan meningkatkan efisiensi dalam pembuatan pohon keputusan. Cara perhitungan algoritma C5.0 untuk menentukan Gain Ratio sama seperti algoritma C4.5.

Berikut adalah rumus dan langkah-langkah untuk mengimplementasikan algoritma C4.5, secara manual menggunakan data sampel dengan tools Microsoft Excel 2010 :

- Entropy

  Entropy adalah ukuran ketidakpastian dalam data, bisa dihitung dengan Rumus :

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

  Langkah pertama mencari nilai Proposi Spam dan Ham, dengan rumus :

  = Jumlah Spam / Total data

  = Jumlah Ham / Total Data

  Untuk memperjelas hasil dari pehitungan proporsi spam dan ham bisa dilihat pada Gambar 3.16.

Hitung Probabilitas	
Spam	0.594462541
Ham	0.405537459

Gambar 3. 36 Proporsi Spam dan ham (C5.0)

Selanjutnya menghitung Log 2 pada masing – masing proporsi spam dan ham.

Rumus :

= Log (nilai proporsi spam atau ham ; 2)

Hasil Log2 ada pada Gambar 3.37 dibawah ini.

Hitung Log^2 :	
Spam	-0.750342192
Ham	-1.302092913

Gambar 3. 37 Nilai Log2 Spam dan Ham (C5.0)

Tahapan selanjutnya menghitung nilai Entropy, dengan rumus :

Entropy = - (Jumlah Proporsi Spam * Nilai Log2 spam + Jumlah Proporsi Ham * Nilai Log2 Ham

Untuk memperjelas pehitungan entropy ada pada Gambar 3.17

1. Hitung Entropy					
Entropy=−(0.594463 * -0.75034 + 0.405537 * -1.30209)					
Hasil =					
0.974098					

Gambar 3. 38 Nilai Entropy C5.0

- Gain Informasi

Gain informasi mengukur seberapa banyak informasi yang diperoleh dari atribut A saat membagi dataset S.

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

Langkah pertama adalah membagi data berdasarkan fitur vince, dengan menetapkan nilai batas untuk memisahkan data menjadi dua bagian. Dalam proses pemisahan data pehitungan manual C5.0, nilai 0.1 dijadikan sebagai batasan. Jika nilai Vince kurang dari 0.1, maka data tersebut akan ditempatkan dalam subset 1, sebaliknya, jika nilai Vince lebih dari 0.1, data

akan dimasukkan ke dalam subset 2.

Rumus :

Subset 1

= IF (value (0.0) < 0.1 ; ” Subset 1” ; “”)

Jika nilai 0.0 kurang dari 0.1, maka masukkan ke dalam subset 1. Jika nilai 0.0 lebih besar dari 0.1, maka tidak ada subset yang relevan atau kosong.

Subset 2

= IF (value (0.0) > 0.1 ; ” Subset 2” ; “”)

Jika nilai 0.0 lebih besar dari 0.1, maka masukkan ke dalam subset 2. Jika nilai 0.0 lebih kecil dari 0.1, maka tidak ada subset yang relevan atau kosong.

vince	Kategori	Subset 1 (<0.1)	Subset 1 (>0.1)
0.0	spam	Subset 1	
0.0	spam	Subset 1	
0.0	spam	Subset 1	
0.0	spam	Subset 1	
0.0	spam	Subset 1	
0.0	spam	Subset 1	
0.0	spam	Subset 1	
0.0	spam	Subset 1	
0.0	spam	Subset 1	
0.0	spam	Subset 1	
0.199556494	ham		Subset 2
0.051520289	ham	Subset 1	
0.028028008	ham	Subset 1	
0.0	ham	Subset 1	
0.140979895	ham		Subset 2
0.0	ham	Subset 1	
0.056281491	ham	Subset 1	
0.101765894	ham		Subset 2
0.153155790	ham		Subset 2
0.209408621	ham		Subset 2
0.061666363	ham	Subset 1	
0.103576251	ham		Subset 2
0.114835209	ham		Subset 2

Gambar 3. 39 Pembagian Data Gain Informasi 0.1

Gambar 3.19 adalah hasil dari pemisahan data dengan batas nilai 0.1.

Selanjutnya menentukan nilai Gain Informasi untuk setiap subset.

Subset 1

vince	Kategori	Jumlah Spam	Jumlah Ham
0.0	spam	1	0
0.0	spam	1	0
0.0	spam	1	0
0.0	spam	1	0
0.0	spam	1	0
0.0	spam	1	0
0.0	spam	1	0
0.0	spam	1	0
0.0	spam	1	0
0.0	spam	1	0
0.0	ham	0	1
0.0637747	ham	0	1
0.0	ham	0	1
0.0936587	ham	0	1
0.0401918	ham	0	1
0.0504007	ham	0	1
0.0	ham	0	1
0.0459733	ham	0	1
0.0560010	ham	0	1
0.0507370	ham	0	1
0.0	ham	0	1
0.0	ham	0	1
0.0194176	ham	0	1
0.0975691	ham	0	1
0.1351475	ham	0	0

Gambar 3. 40 Hasil Pembagan Data Subset 1 (C5.0)

Pada Gambar 3.40 memperlihatkan hasil pembagian data pada subset 1, di mana data subset 1 yang termasuk dalam kategori spam diberikan nilai 1 di kolom jumlah spam, sementara data yang bukan spam mendapat nilai 0. Hal yang sama berlaku untuk kategori Ham.

Langkah selanjutnya menghitung Proporsi Spam dan Ham.

Rumus :

Spam dan ham = Jumlah Spam / Total Data Spam dan Ham subset 1.

Spam = 1094 / 1718 = 0.636786962

Ham = 624 / 1718 = 0.363213038

A. Proporsi Spam dan Ham

Subset 1	
Spam	0.63678696
Ham	0.36321304

Gambar 3. 41 Proporsi Subset 1 (C5.0)

Selanjutnya menghitung nilai Log2 pada subset 1.

Rumus :

= LOG (Proporsi Spam/Ham ; 2)

Spam = LOG (0.63678696 ; 2)

Ham = LOG (0.36321304 ; 2)

Untuk hasilnya bisa dilihat pada Gambar 3.42.

B. Hitung Log^2 :

Subset 1	
Spam	-0.6511173
Ham	-1.4611121

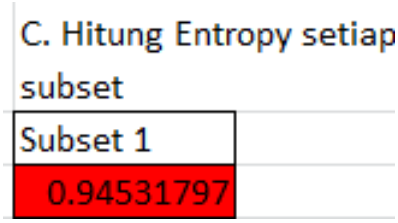
Gambar 3. 42 Nilai Log2 Subset 1 (C5.0)

Tahapan selanjutnya Hitung nilai entropy pada subset 1, dengan rumus :

Entropy = - (Proporsi Spam * Nilai Log2 Spam + Proporsi Ham * Nilai Log2 Ham)

= - (0.63678696 * - 0.6511173 + 0.36321304 * - 1.4611121)

Hasilnya bisa dilihat pada Gambar 3.43.



Gambar 3. 43 Nilai Entropy Subset 1 (C4.5)

Subset 2

vince	Kategori	Jumlah Spam	Jumlah Ham
0.0	spam	0	0
0.0	spam	0	0
0.0	spam	0	0
0.0	spam	0	0
0.0	spam	0	0
0.0	spam	0	0
0.0	spam	0	0
0.0	spam	0	0
0.0	spam	0	0
0.0	spam	0	0
0.0	spam	0	0
0.1581609	ham	0	1
0.2198307	ham	0	1
0.0686006	ham	0	0
0.0	ham	0	0
0.0430797	ham	0	0
0.0265389	ham	0	0
0.0	ham	0	0
0.1218433	ham	0	1
0.0586649	ham	0	0
0.1092927	ham	0	1
0.0783729	ham	0	0
0.0	ham	0	0
Jumlah		0	127

Gambar 3. 44 Hasil Pembagan Data Subset 2 (C5.0)

Pada Gambar 3.44 memperlihatkan hasil pembagian data pada subset 2, di

mana data subset 2 yang termasuk dalam kategori spam diberikan nilai 1 di kolom jumlah spam, sementara data yang bukan spam mendapat nilai 0. Hal yang sama berlaku untuk kategori Ham.

Selanjutnya sama seperti subset 1, menghitung proporsi spam dan ham. Dengan Rumus :

Spam dan ham = Jumlah Spam / Total Data Spam dan Ham subset 2.

Spam = 0.1 / 127 = 0.00078678

Ham = 127 / 127 = 0.99921322

Subset 2	
Spam	0.00078678
Ham	0.99921322

Gambar 3. 45 Proporsi Subset 2 (C5.0)

Selanjutnya menghitung nilai Log2 pada subset 2.

Rumus :

= LOG (Proporsi Spam/Ham ; 2)

Spam = LOG (0.00078678 ; 2)

Ham = LOG (0.99921322 ; 2)

Untuk hasilnya bisa dilihat pada Gambar 3.26

Subset 2	
Spam	-10.311748
Ham	-0.0011355

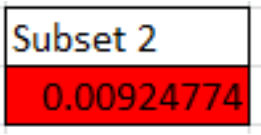
Gambar 3. 46 Nilai Log2 Subset 1 (C4.5

Hitung nilai entropy pada subset 2, dengan rumus :

Entropy = - (Proporsi Spam * Nilai Log2 Spam + Proporsi Ham * Nilai Log2 Ham)

= - (0.00078678 * - 10.311748 + 0.99921322 * - 0.0011355)

Hasilnya bisa dilihat pada Gambar 3.27.



Subset 2
0.00924774

Gambar 3. 47 Nilai Entropy Subset 2 (C4.5)

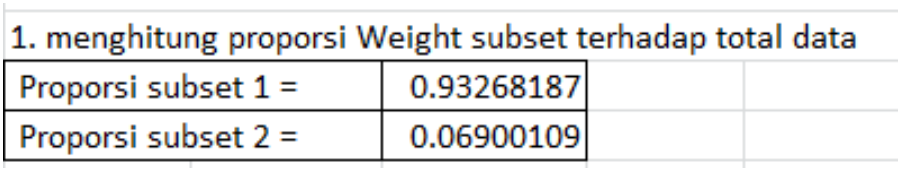
Selanjutnya Hitung weight entropy subset untuk masing- masing subset.

Langkah pertama hitung proporsi subset 1 dan subset 2.

Rumus :

Proporsi Subset = Jumlah spam dan ham subset 1 dan subset 2/ Total data

Untuk hasilnya ada pada Gambar 3.48.



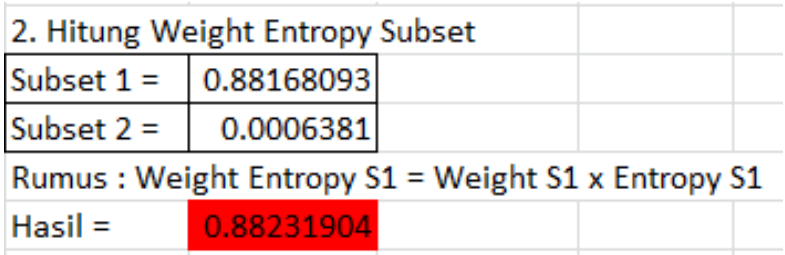
1. menghitung proporsi Weight subset terhadap total data	
Proporsi subset 1 =	0.93268187
Proporsi subset 2 =	0.06900109

Gambar 3. 48 Proporsi Subset 1.2 (C4.5)

Selanjutnya hitung Weight Entropy setiap Subset, dengan rumus :

= Weight Entropy S1 = Weight S1 x Entropy S1

Untuk lebih memperjelas hasil Weight Entropy setiap Subset bisa dilihat pada Gambar 3.49.



2. Hitung Weight Entropy Subset	
Subset 1 =	0.88168093
Subset 2 =	0.0006381
Rumus : Weight Entropy S1 = Weight S1 x Entropy S1	
Hasil =	0.88231904

Gambar 3. 49 Nilai Weight Entropy Subset 1,2 (C4.5)

Hitung Information Gain (IG) / Gain Informasi, dengan rumus :

= Information Gain=Entropy Total−Total Weight Entropy

3. Hitung Information Gain (IG):	
Information Gain=Entropy Total−Total Weight Entropy	
Information Gain =	0.09177874

Gambar 3. 50 Nilai IG (C4.5)

Gambar 3.50 adala nilai Information Gain yang berjumlah 0.09177874.

- Split Information

Rumus :

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

Langkah pertama menghitung Log2 untuk setiap proporsi, dengan rumus :

= LOG (Proporsi subset 1 atau subset 2 ; 2)

Log^2 Proporsi	
Subset 1 =	-0.100543025
Subset 2 =	-3.857237126

Gambar 3. 51 Nilai Log2 Subset 1,2 (C5.0)

Selanjutnya menghitung nilai Kontribusi Split Information untuk setiap subset.

Rumus :

= Nilai proporsi subset 1 atau 2 * Nilai Log2 subset 1 atau subset 2

Untuk lebih memperjelas hasil Kontribusi Split Information , bisa dilihat pada Gambar 3.52.

Kontribusi Split Information	
Subset 1 =	-0.093774656
Subset 2 =	-0.26615355

Gambar 3. 52 Nilai Kontribusi (C4.5)

Tahapan selanjutnya menghitung total Split Information (SI), dengan rumus : =SUM(Kontribusi_Subset_1:Kontribusi_Subset_n)

Total Split Information (SI)	
Rumus : =SUM(Kontribusi_Subset_1:Kontribusi_Subset_n)	
Split Information =	0.359928206

Gambar 3. 53 Nilai Split Information (C5.0)

Gambar 3.53 adalah nilai dari Split Information (SI), dengan jumlah 0.359928206.

- Gain Ratio

  Rumus :

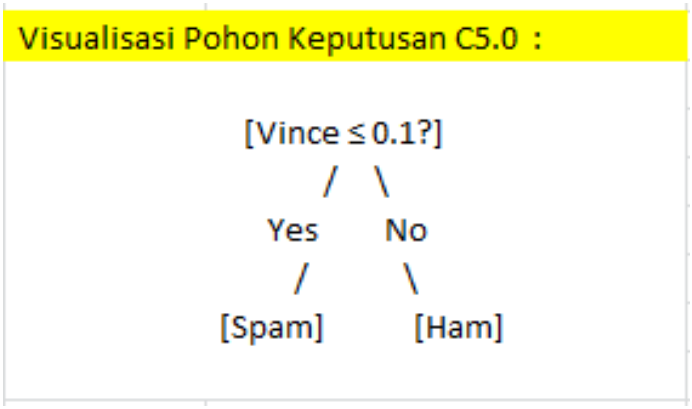
$$Gain\ Ratio\ (S,A) = \frac{IG\ (S,A)}{SplitInfo\ (S,A)}$$

4. Gain Ratio	
Rumus : = Gain / Split_Info	
Gain Ratio =	0.254991802

Gambar 3. 54 Nilai Gain ratio (C4.5)

Pada Gambar 3.54, ditampilkan nilai Gain Ratio dari algoritma C5.0. Nilai ini digunakan untuk menentukan root (akar) pertama dalam pembentukan pohon keputusan.

- Visualisasi pohon keputusan C5.0



Gambar 3. 55 Pohon Keputusan Algoritma C5.0

Gambar 3.55 menunjukkan pohon keputusan yang dihasilkan oleh algoritma C5.0 berdasarkan perhitungan secara manual. Dalam pohon tersebut, vince muncul sebagai node akar pertama karena memiliki nilai Gini Index yang paling tinggi. Jika nilai vince berada di bawah 0,1, maka data akan diklasifikasikan sebagai spam. Sebaliknya, jika nilai vince melebihi 0,1, maka data tersebut akan diklasifikasikan sebagai ham.

Perhitungan yang dilakukan secara manual oleh algoritma CART, C4.5, dan C5.0 pada perhitungan dengan menggunakan data sampel. Dataset yang digunakan dalam perhitungan manual ini merupakan fitur akar (root) pertama yang dihasilkan dari perhitungan menggunakan tools *RapidMiner*. Perhitungan manual ini bertujuan untuk memberikan contoh konkret mengenai bagaimana setiap algoritma menentukan root pertama serta melakukan pembagian data. Dengan kata lain, meskipun perhitungan dilakukan dengan menggunakan tools, setiap atribut atau fitur yang dihitung melalui tools tersebut pada dasarnya telah melewati tahapan-tahapan yang sama seperti yang diterapkan dalam perhitungan manual. Pendekatan ini diterapkan untuk memperoleh pengetahuan yang lebih rinci dan jelas tentang bagaimana algoritma CART, C4.5, dan C5.0 berfungsi dalam proses klasifikasi.