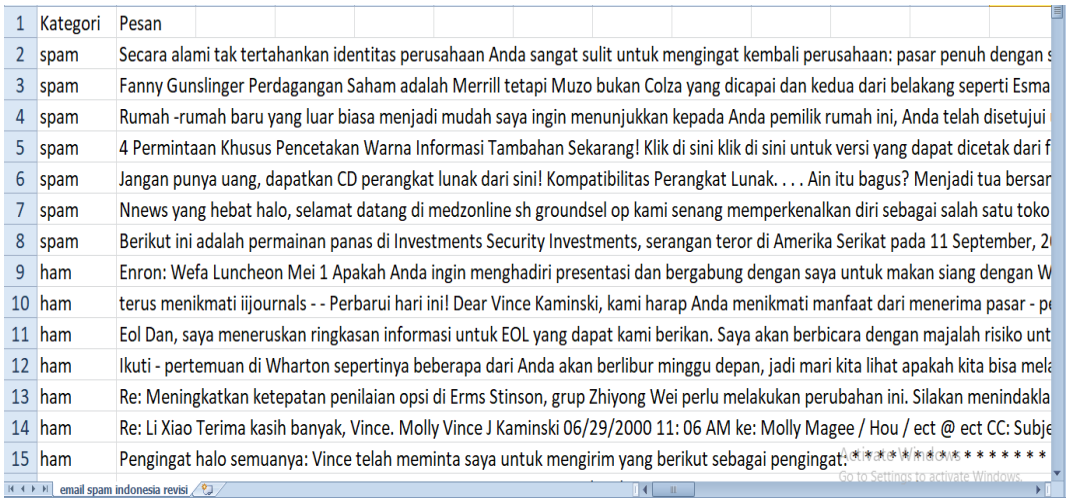
BAB IV
HASIL DAN PEMBAHASAN

4.1 Pengumpulan Data

Penelitian ini menggunakan dataset publik yaitu dataset yang tersedia di repository kaggle dengan jumlah data sebanyak 2.638 dataset email spam dan ham yang belum di cleaning dengan 2 Atribut yaitu Kategori dan Pesan, dan untuk Class/Label yaitu atribut Kategori. Dataset dalam penelitian ini adalah dataset teks bernama Indonesian Email Spam yang sudah diterjemahkan kedalam bahasa indonesia. Dataset teks ini nantinya akan di preprocessing menjadi data yang bisa dibaca oleh model atau dataset yang siap digunakan untuk proses klasifikasi email spam dan ham menggunakan Software *Rapidminer* dan visualisasinya menggunakan Tools Tablue.



	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 4. 1 Data Mentah

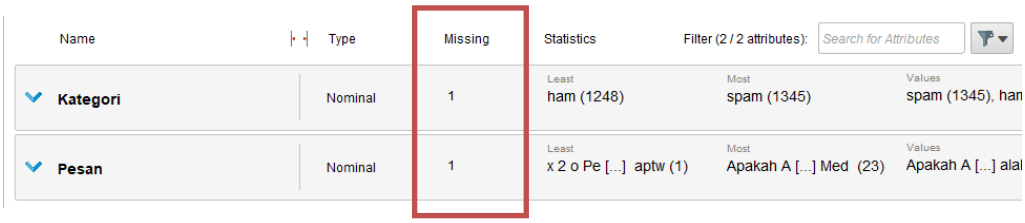
4.2 Pre-processing

Pada tahapan ini data akan dilakukan cleaning atau proses pembersihan data, menggunakan tools *RapidMiner*.

4.2.1 Cleaning Data (Missing Values dan Duplikat Data)

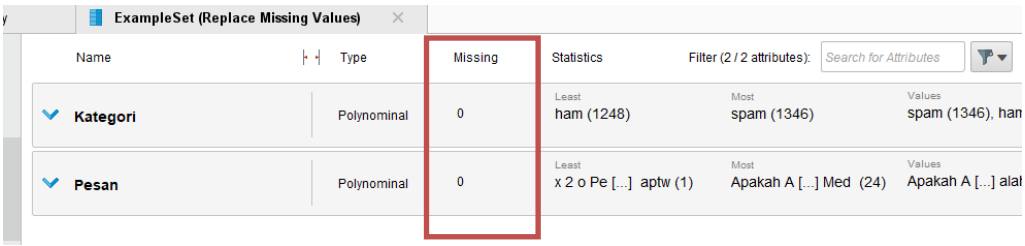
1. Missing Values

   Missing Values adalah nilai yang hilang atau kosong dalam dataset.



Gambar 4. 2 Dataset Sebelum Missing Values

Dataset dalam penelitian ini, pada Gambar 4.2 terdapat 1 Missing Values atau nilai yang kosong pada atribut kategori dengan type data nominal dan atribut pesan dengan type data nominal Untuk menghapus Missing Values pada setiap atribut dilakukan "*Replace Missing Values"* sehingga Missing Values pada masing – masing atibut menjadi kosong dengan type data Polynominal. Untuk melihat data yang sudah di *Replace Missing Values* bisa dilihat pada Gambar 4.3.



Gambar 4. 3 Dataset Sesudah Missing Values

2. Duplikat Data

   Duplikat data adalah data yang muncul lebih dari sekali dalam dataset atau data yang sama.

Row No.	Kategori	Pesan
1	spam	Secara alami tak tertahankan ide...
2	spam	Fanny Gunslinger Perdagangan ...
3	spam	Rumah -rumah baru yang luar bi...
4	spam	4 Permintaan Khusus Pencetaka...
5	spam	Jangan punya uang, dapatkan C...
6	spam	Nnews yang hebat halo, selamat ...
7	spam	Berikut ini adalah permainan pan...
8	spam	Simpan uang Anda beli untuk me...
9	spam	Tidak Terkirim: Bisnis Berbasis ...
10	spam	Simpan uang Anda beli untuk me...
11	spam	Las Vegas High Rise Boom Las ...
12	spam	Simpan uang Anda beli untuk me...
13	spam	mencerahkan gigi itu membuat g...
14	spam	Fenomena Wall Street mendapat...
15	spam	PEMBERITAHUAN FPA: Ebay Mis...

ExampleSet (2,594 examples,0 special attributes,2 regular attributes)

40 duplicate values found and removed; 2598 unique values remain.

Gambar 4. 4 Dataset Sebelum Remove Duplicates

Pada dataset ini ditemukan data duplikat sebanyak 40 duplikat data, untuk menghapus duplikat data menggunakan "*Remove Duplicates"*. Sehingga dataset awal berjumlah 2.594 dataset menjadi 2.548 dataset. Untuk lebih jelas melihat jumlah data yang sudah di *Remove Duplicates* bisa dilihat pada Gambar 4.5.

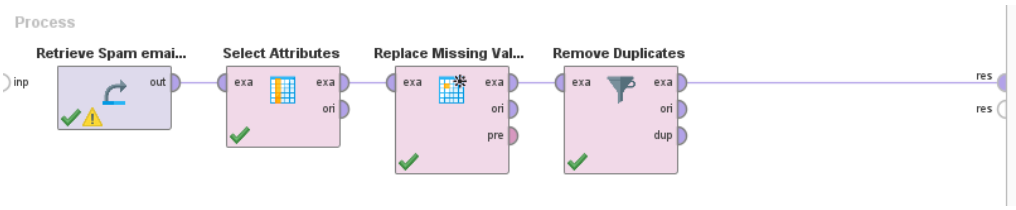
Row No.	Kategori	Pesan
1	spam	Secara alami tak tertahankan identitas ...
2	spam	Fanny Gunslinger Perdagangan Saha...
3	spam	Rumah -rumah baru yang luar biasa m...
4	spam	4 Permintaan Khusus Pencetakan War...
5	spam	Jangan punya uang, dapatkan CD pera...
6	spam	Nnews yang hebat halo, selamat datan...
7	spam	Berikut ini adalah permainan panas di l...
8	spam	Simpan uang Anda beli untuk mendapa...
9	spam	Tidak Terkirim: Bisnis Berbasis Rumah...
10	spam	Simpan uang Anda beli untuk mendapa...
11	spam	Las Vegas High Rise Boom Las Vegas...
12	spam	Simpan uang Anda beli untuk mendapa...
13	spam	mencerahkan gigi itu membuat gigi An...
14	spam	Fenomena Wall Street mendapatkan h...
15	spam	PEMBERITAHUAN FPA: Ebay Misrepres...

ExampleSet (2,548 examples,0 special attributes,2 regular attributes)

Gambar 4. 5 Dataset Sesudah Remove Duplicates

3. Proses Cleaning Data

   Proses cleaning data menggunakan Replace Missing Values, Remove Duplicates di *Rapidminer*, untuk memperjelas proses penghapusan missing values dan duplikat data di *rapidminer* bisa dilihat pada Gambar 4.6.



Gambar 4. 6 Tahapan Cleaning

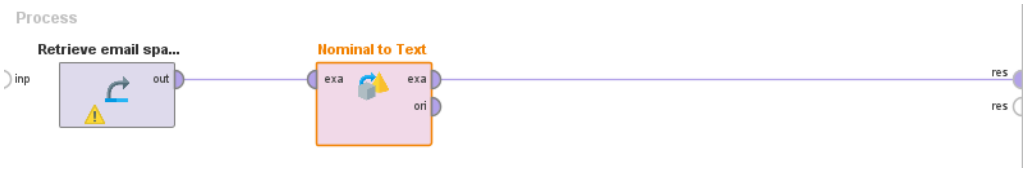
Pada Gambar 4.6 adalah proses untuk menghapus atau operator yang ada di Rapiminer yang digunakan untuk menghapus data missing values dan data duplikat menggunakan operator Replace Missing Values dan Remove Duplicate.

4.2.2 Text Preprocessing

Tahapan teks preprocessing (pembersihan teks) bertujuan untuk menghapus unsur-unsur tidak penting dari teks agar data email spam dan ham lebih siap untuk diproses oleh model, sehingga dapat digunakan dalam pengklasifikasian email spam dan ham.

1. Nominal to text

   Tahapan pertama adalah menambahkan Nominal to text tujuannya untuk mengelompokkan email berdasarkan kategori (spam atau ham) dan menjadi target output dalam klasifikasi.



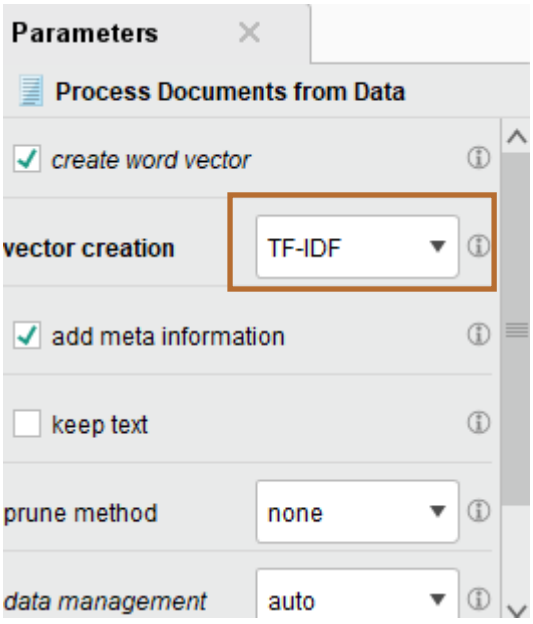
Gambar 4. 7 Proses Nominal To Text

Gambar 4.7 adalah proses pertama untuk text preprocessing, yaitu menambahkan operator Nominal to Text di *RapidMiner* digunakan untuk

mengubah atribut bertipe Nominal menjadi Text, sehingga dapat diproses dalam analisis teks sama seperti TF-IDF.

2. Process Documents from Data

   Selanjutnya menambahkan Process Documents from Data , proses ini digunakan untuk mempersiapkan data teks sebelum diterapkan ke model pembelajaran mesin. Dan membantu mengubah teks mentah menjadi format yang dapat dipahami dan diolah oleh algoritma klasifikasi. Proses ini melibatkan pembobotan fitur TF-IDF (*Term Frequency-Inverse Document Frequency*), TF-IDF ini membantu model untuk fokus pada kata-kata yang paling relevan dalam menentukan apakah sebuah email termasuk spam atau bukan. Pada Process Documents from Data, TF-IDF ini ada di parameters Process Documents from Data di *Rapidminer* dan tidak terpisah prosesnya dengan tahapan text preprocessing di Tools *Rapidminer*. Untuk melihat lebih jelas parameters Process Documents from Data bisa dilihat pada Gambar 4.8. Dan di dalam Process Documents from Data ini terdiri dari proses Tokenize, Transfrom Cases, Filter Stopwords (dictionary) dan Filter Tokens (by Leght).



Gambar 4. 8 Parameters Proses Documents from Data

3. Tokenize

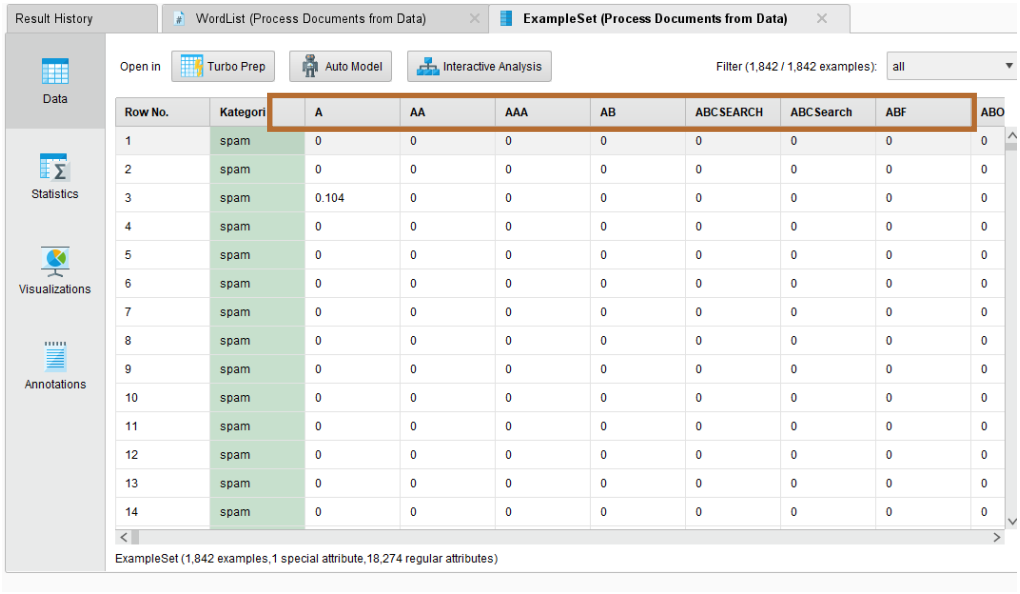
   Proses Tokenize berfungsi untuk memecah teks menjadi bagian-bagian kecil yang dikenal sebagai token.

Gambar 4. 9 Parameters Tokenize

Pada Gambar 4.9 Parameters Tokenize, dibagian "*regular expression*" masukan simbol – simbol di expression untuk menghapus simbol – simbol yang tidak diperlukan atau untuk membersihkan data, proses ini bertujuan untuk menghapus simbol – simbol atau tanda baca yang ada pada dataset supaya mendapatkan hasil token yang relevan dan menghindari token seperti "adalah@!".

Untuk memperjelas hasil tokenize bisa dilihat pada Gambar 4.10.

Gambar 4.10 menunjukkan data hasil tokenisasi, di mana setiap fitur telah diubah menjadi token. Setiap token berbentuk teks tanpa adanya simbol-simbol.

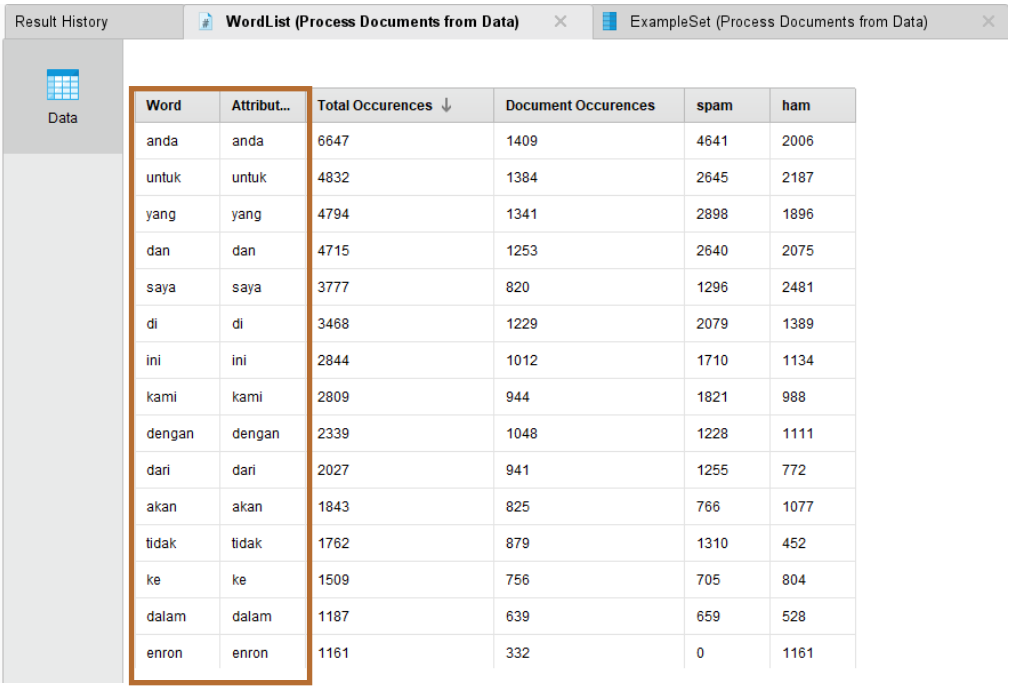


Row No.	Kategori	A	AA	AAA	AB	ABCSEARCH	ABCSearch	ABF	ABO
1	spam	0	0	0	0	0	0	0	0
2	spam	0	0	0	0	0	0	0	0
3	spam	0.104	0	0	0	0	0	0	0
4	spam	0	0	0	0	0	0	0	0
5	spam	0	0	0	0	0	0	0	0
6	spam	0	0	0	0	0	0	0	0
7	spam	0	0	0	0	0	0	0	0
8	spam	0	0	0	0	0	0	0	0
9	spam	0	0	0	0	0	0	0	0
10	spam	0	0	0	0	0	0	0	0
11	spam	0	0	0	0	0	0	0	0
12	spam	0	0	0	0	0	0	0	0
13	spam	0	0	0	0	0	0	0	0
14	spam	0	0	0	0	0	0	0	0

Gambar 4. 10 Data Hasil Tokenize

4. Transform Cases / Lowercasing

Transform Cases atau Lowercasing digunakan untuk mengubah teks menjadi format huruf kecil semua.



Word	Attribut...	Total Occurences	Document Occurences	spam	ham
anda	anda	6647	1409	4641	2006
untuk	untuk	4832	1384	2645	2187
yang	yang	4794	1341	2898	1896
dan	dan	4715	1253	2640	2075
saya	saya	3777	820	1296	2481
di	di	3468	1229	2079	1389
ini	ini	2844	1012	1710	1134
kami	kami	2809	944	1821	988
dengan	dengan	2339	1048	1228	1111
dari	dari	2027	941	1255	772
akan	akan	1843	825	766	1077
tidak	tidak	1762	879	1310	452
ke	ke	1509	756	705	804
dalam	dalam	1187	639	659	528
enron	enron	1161	332	0	1161

Gambar 4. 11 Hasil Transform Cases / Lowercasing

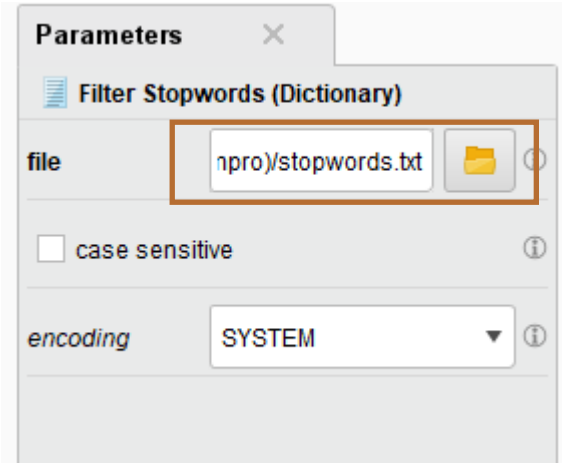
Pada Gambar 4.11 Hasil Transform Cases atau Lowercasing, bisa dilihat pada word dan atribut dataset yang sudah di tokenize berubah menjadi token behuruf kecil semua, proses ini bertujuan untuk membuat teks lebih konsisten, menghindari perbedaan antara kata yang sama yang ditulis dengan huruf besar.

5. Stopwords

Digunakan untuk menghilangkan kata-kata umum yang tidak membawa informasi penting dalam teks. Di *Rapidminer* Menggunakan "Filter Stopwords (Dictionary)". Dalam proses ini harus menggunakan contoh stopwords, stopwords yang digunakan pada proses ini diunduh repository kanggle, seperti Gambar 4.12.

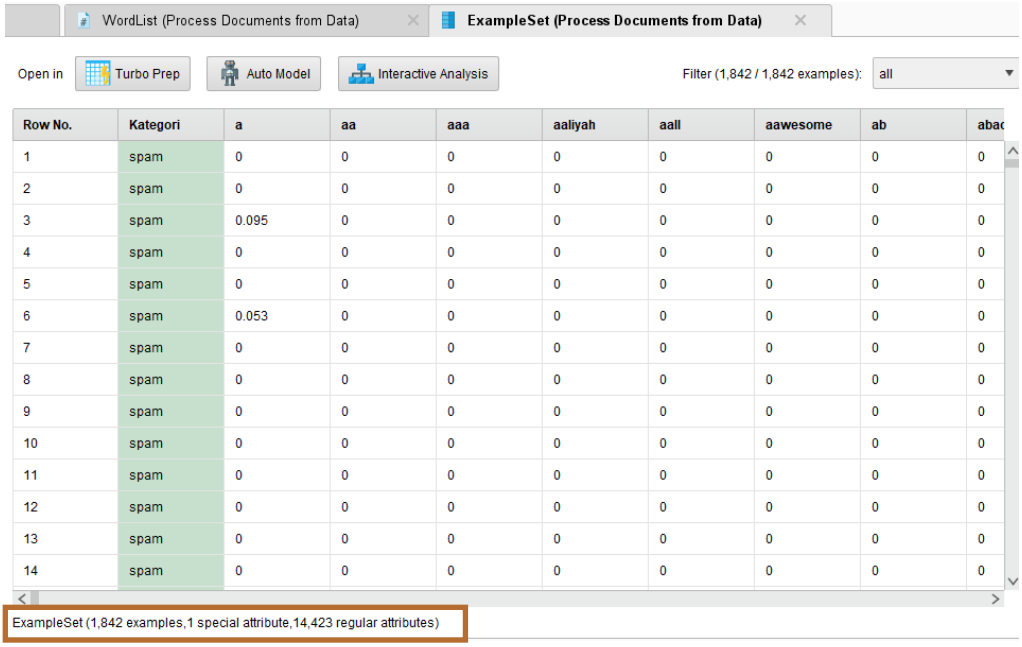
Gambar 4. 12 Stopwords

Pada Gambar 4.12, adalah contoh tabel stopwords yang diunduh di repository Kaggle dan digunakan sebagai contoh untuk penghapusan kata-kata umum yang tidak membawa informasi penting. Dan dimasukkan di parameters Filter Stopwords (Dictionary) seperti pada Gambar 4.13.



Gambar 4. 13 Parameters Filter Stopwords (Dictionary)

Hasil dari penghapusan stopwords menghasilkan pengurangan jumlah fitur, dari 18.274 atribut reguler menjadi 14.423 atribut reguler, karena kata-kata umum yang tidak memberikan informasi relevan telah dihilangkan. Untuk memahami hasil tersebut dengan lebih baik, lihat bagian bawah Gambar 4.14, yang diwarnai merah, memperlihatkan ExampleSet (1.842 examples, 1 special attribute, 14.423 regular attributes).

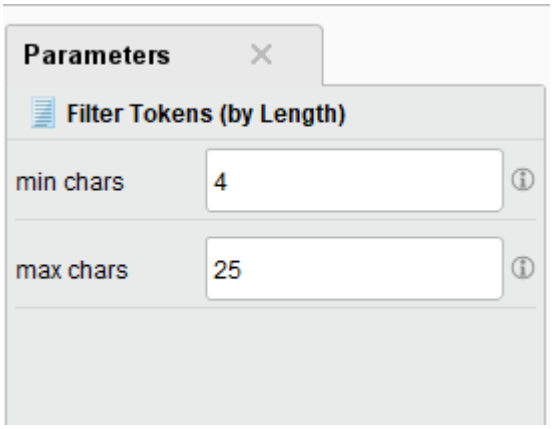


Row No.	Kategori	a	aa	aaa	aaliyah	aall	aawesome	ab	abac
1	spam	0	0	0	0	0	0	0	0
2	spam	0	0	0	0	0	0	0	0
3	spam	0.095	0	0	0	0	0	0	0
4	spam	0	0	0	0	0	0	0	0
5	spam	0	0	0	0	0	0	0	0
6	spam	0.053	0	0	0	0	0	0	0
7	spam	0	0	0	0	0	0	0	0
8	spam	0	0	0	0	0	0	0	0
9	spam	0	0	0	0	0	0	0	0
10	spam	0	0	0	0	0	0	0	0
11	spam	0	0	0	0	0	0	0	0
12	spam	0	0	0	0	0	0	0	0
13	spam	0	0	0	0	0	0	0	0
14	spam	0	0	0	0	0	0	0	0

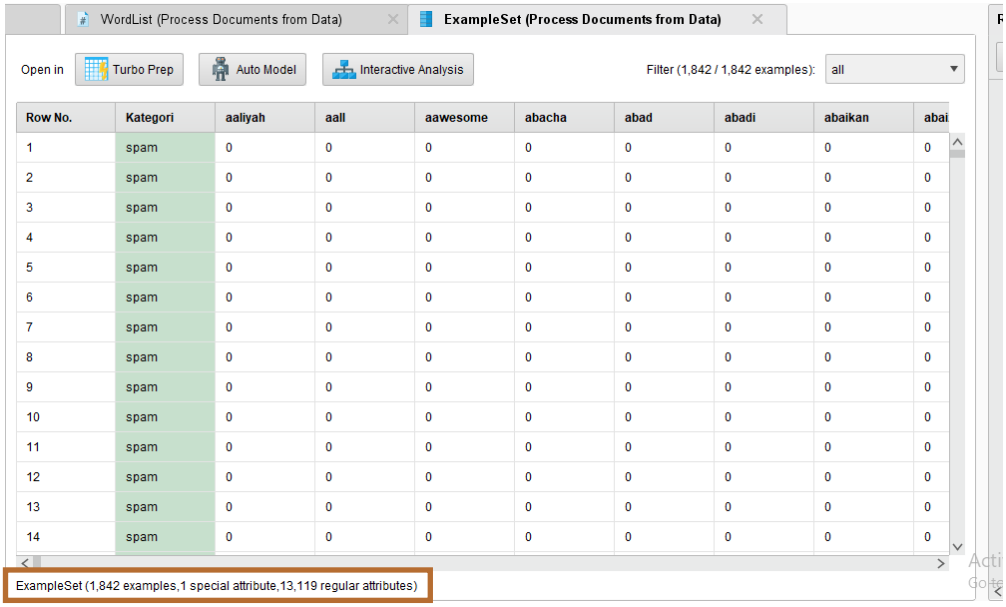
Gambar 4. 14 Hasil Filter Stopwords (Dictionary)

6. Filter Tokens

Proses selanjutnya adalah Filter Tokens, yang bertujuan untuk menyaring token (kata atau unit teks) berdasarkan panjangnya. Penyaringan ini dilakukan untuk menghapus kata-kata yang sangat pendek atau sangat panjang sehingga dianggap tidak relevan. Untuk memfilter panjang dan pendeknya sebuah kata, digunakan operator "Filter Tokens (by Length)" dengan parameter min chars sebesar 4 dan max chars sebesar 25. Parameter yang digunakan dalam Filter Tokens (by Length) dapat dilihat pada Gambar 4.15.



Gambar 4. 15 Parameters Filter Tokens (by Length)

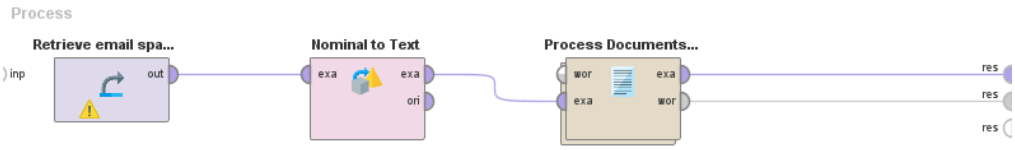


Row No.	Kategori	aaliyah	aall	aawesome	abacha	abad	abadi	abaikan	abai
1	spam	0	0	0	0	0	0	0	0
2	spam	0	0	0	0	0	0	0	0
3	spam	0	0	0	0	0	0	0	0
4	spam	0	0	0	0	0	0	0	0
5	spam	0	0	0	0	0	0	0	0
6	spam	0	0	0	0	0	0	0	0
7	spam	0	0	0	0	0	0	0	0
8	spam	0	0	0	0	0	0	0	0
9	spam	0	0	0	0	0	0	0	0
10	spam	0	0	0	0	0	0	0	0
11	spam	0	0	0	0	0	0	0	0
12	spam	0	0	0	0	0	0	0	0
13	spam	0	0	0	0	0	0	0	0
14	spam	0	0	0	0	0	0	0	0

Gambar 4. 16 Hasil Filter Tokens (by Length)

Pada Gambar 4.16, ditampilkan hasil dari proses Filter Tokens (by Length). Dengan membatasi panjang dan pendek token, jumlah fitur dalam dataset mengalami pengurangan, yaitu dari 14.423 regular atribut menjadi 13.119 regular atribut.

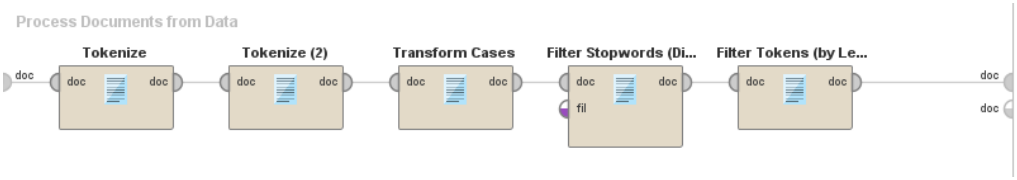
7. Operator Cleaning Teks



Gambar 4. 17 Tahapan Text Preprocessing

Pada Gambar 4.17, terlihat operator yang digunakan dalam proses preprocessing text di *RapidMiner*. Dataset yang digunakan mencakup email spam dan ham, sedangkan Nominal to Text berperan dalam mengkonversi atribut dengan tipe Nominal menjadi format Teks. Selain itu, Process Documents from Data digunakan untuk menyiapkan data teks serta mendukung

pengubahan teks mentah menjadi bentuk yang mudah dimengerti dan diproses oleh algoritma klasifikasi.

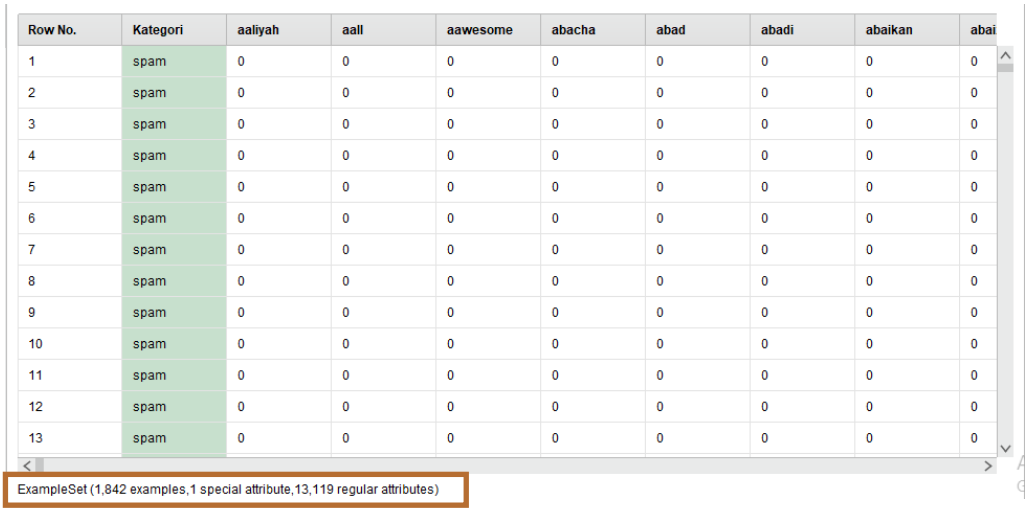


Gambar 4. 18 Tahapan Process Documents from Data

Gambar 4.18 menampilkan operator-operator yang terdapat dalam Process Documents from Data, yang digunakan dalam proses preprocessing text. Operator ini terdiri dari Tokenize, yang berfungsi membagi teks menjadi potongan-potongan kecil (token), Transform Case, yang mengubah semua huruf menjadi kecil untuk memastikan keseragaman data, penyaringan kata umum (Filter Stopwords), yang menghilangkan kata-kata biasa yang tidak mengandung informasi berarti,serta Filter Tokens (by Length), yang menyaring token berdasarkan panjangnya untuk menghilangkan kata-kata yang terlalu pendek atau terlalu panjang agar lebih relevan. Operator-operator ini bekerja secara berurutan untuk memastikan data teks menjadi lebih rapi dan siap untuk analisis selanjutnya.

8. Hasil Preprocessing

Dataset yang sudah di cleaning dan siap digunakan untuk proses klasifikasi email spam dan ham.



Row No.	Kategori	aaliyah	aall	aawesome	abacha	abad	abadi	abaikan	abai
1	spam	0	0	0	0	0	0	0	0
2	spam	0	0	0	0	0	0	0	0
3	spam	0	0	0	0	0	0	0	0
4	spam	0	0	0	0	0	0	0	0
5	spam	0	0	0	0	0	0	0	0
6	spam	0	0	0	0	0	0	0	0
7	spam	0	0	0	0	0	0	0	0
8	spam	0	0	0	0	0	0	0	0
9	spam	0	0	0	0	0	0	0	0
10	spam	0	0	0	0	0	0	0	0
11	spam	0	0	0	0	0	0	0	0
12	spam	0	0	0	0	0	0	0	0
13	spam	0	0	0	0	0	0	0	0

ExampleSet (1,842 examples,1 special attribute,13,119 regular attributes)

Gambar 4. 19 Data Final Cleaning

Gambar 4.19 menampilkan Data Final Cleaning yang telah melalui proses preprocessing, di mana dataset teks diubah menjadi Feature Matrix menggunakan teknik TF-IDF. Matriks ini terdiri dari 1.841 baris (rows), yang ditunjukkan oleh angka pada kolom pertama seperti Row No 1, 2, 3, dan seterusnya, di mana setiap baris mewakili satu contoh atau instance data, dalam hal ini email yang diklasifikasikan sebagai spam atau ham. Selain itu, matriks ini memiliki 13.119 kolom, yang merupakan jumlah fitur kata penting yang dihasilkan dari proses preprocessing text. Matriks ini digunakan sebagai input untuk model pembelajaran mesin pada tahapan berikutnya. Untuk mempermudah membaca dataset, dapat dilihat pada Gambar 4.20 yang menampilkan final dataset.

Word	Attribute Name	Total Occurences ↓	Document Occurences	spam	ham
enron	enron	1161	332	0	1161
vince	vince	968	426	0	968
situs	situs	717	333	664	53
memiliki	memiliki	647	500	337	310
perusah...	perusahaan	571	260	399	172
kaminski	kaminski	531	261	0	531
email	email	530	279	408	122
terima	terima	528	382	87	441
kasih	kasih	516	375	74	442
orang	orang	513	320	295	218
informasi	informasi	461	310	311	150
adobe	adobe	446	46	443	3
http	http	444	322	350	94
bisnis	bisnis	436	254	317	119
perangkat	perangkat	410	202	358	52

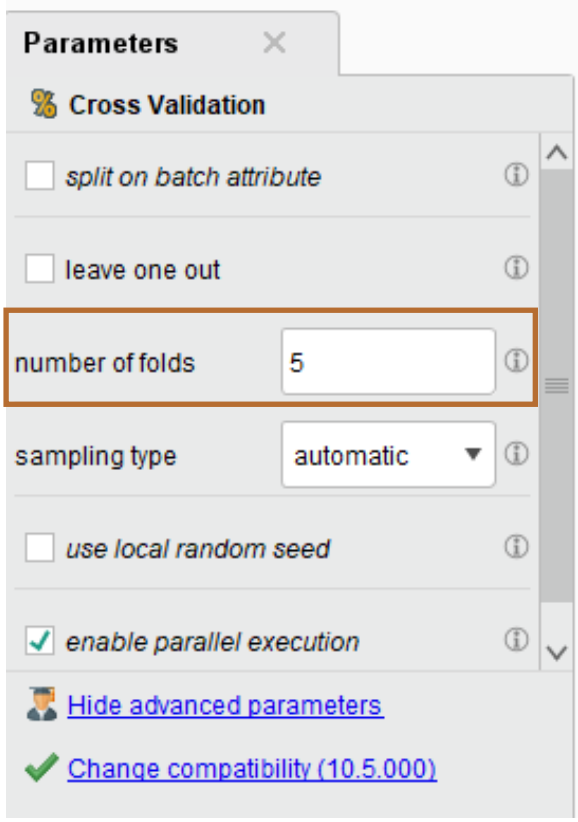
Gambar 4. 20 Final Dataset

Final dataset pada Gambar 4.20 ini sama seperti Data Final Cleaning yang ada pada Gambar 4.19. Perbedaannya terletak pada cara penyajian data, di mana pada Data Final di Gambar 4.20, data telah dikelompokkan dan disusun secara berurutan dan teratur, sehingga lebih mudah untuk dibaca dan dipahami.

4.3 Pembagian Data

Proses selanjutnya adalah pembagian data, Pada penelitian ini pembagian data dilakukan dengan menggunakan teknik 5-fold Cross Validation. 1.841 dataset dibagi menjadi lima bagian yang sama besar. Pada setiap iterasi, empat bagian digunakan sebagai data latih yang berjumlah 1.473 data untuk 4 *folds* dan untuk *folds* terakhir berjumlah 1.472 data, sementara untuk data uji pada 4 folds masing-masing berisi 368 data dan *folds* terakhir berisi 369 data (sisa 1 data karena 1.841 − (368×4) = 369. Proses ini diulang sebanyak lima kali, sehingga setiap bagian dataset akan digunakan sekali sebagai data uji dan empat kali sebagai data latih.

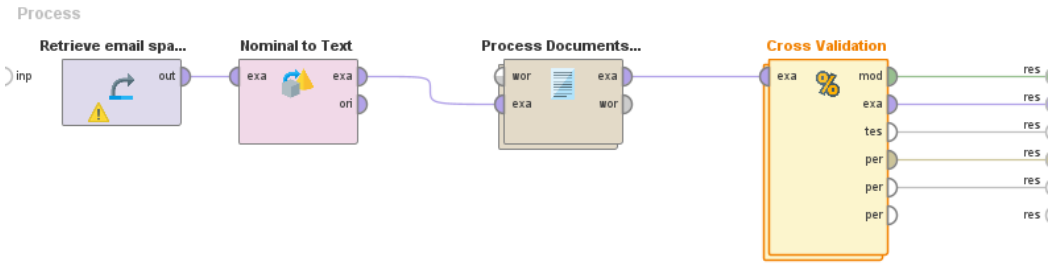
Untuk memperjelas pembagian data menggunakan metode cross validation bisa dilihat pada Gambar 4.21 dan 4.22.



Gambar 4. 21 Parameters Cross Validation

Pada Gambar 4.21, parameter cross-validation menunjukkan bagian number of folds, di mana kita bisa memasukkan jumlah folds yang diinginkan. Sebagai

contoh, jika memasukkan angka "5", proses cross-validation akan menggunakan 5 folds, yang berarti data dibagi menjadi lima bagian untuk validasi silang.



Gambar 4. 22 Proses Cross Validation

Gambar 4.22 menunjukkan penambahan operator Cross Validation, yang digunakan untuk membagi data menjadi beberapa folds. Penjelasan lebih lanjut mengenai parameter-parameter yang digunakan oleh Cross Validation dapat dilihat pada Gambar 4.21.

4.4 Implementasi Algoritma Klasifikasi

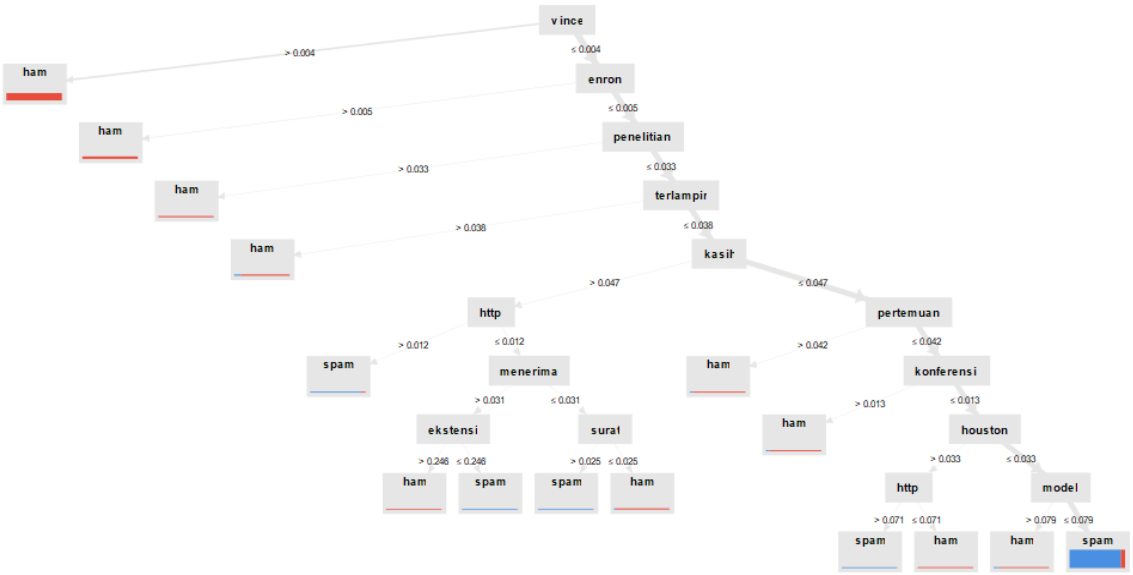
Tahapan selanjutnya dilakukan penerapan Algoritma klasifikasi terhadap data teks email spam dan ham yang sudah di preprocessing menjadi dataset berbentuk *Feature Matrix*. Algoritma yang digunakan yaitu Algoritma Decision Tree versi Algoritma CART (*Classification and Regression Tree*), C4.5 dan C5.0.

4.4.1 Algoritma CART (Classification and Regression Tree)

1. Proses Pelatihan (Training)

   Pada proses training, model CART dibangun dengan membagi data berdasarkan nilai atribut yang meminimalkan nilai Gini Index. Pembagian ini dilakukan secara rekursif pada setiap node hingga mencapai kondisi berhenti yang telah ditentukan, seperti kedalaman pohon atau jumlah data dalam satu leaf yang terlalu sedikit. Model ini menghasilkan pohon keputusan dengan sejumlah node dan leaf yang masing-masing merepresentasikan aturan klasifikasi.

Berikut adalah pohon keputusan yang dihasilkan oleh model CART :



Gambar 4. 23 Pohon Keputusan Algoritma CART

Pohon keputusan yang dihasilkan oleh algoritma CART pada Gambar 4.23 memiliki struktur berikut :

- Akar Pohon (Root Node), keputusan dimulai dari fitur vince dengan batas nilai (threshold) > 0.004. Berdasarkan nilai ini, data akan diarahkan ke cabang berikutnya.
- Cabang (Branches), Setiap cabang mewakili pengambilan keputusan berdasarkan nilai atribut atau fitur. Contohnya, jika atribut vince ≤ 0.004, maka data akan diarahkan ke jalur kiri, sedangkan jika > 0.004, ke jalur kanan.
- Node Daun (Leaf Nodes), Pada ujung cabang, terdapat node daun yang memberikan hasil klasifikasi (spam dan ham). Pada pohon ini, node daun menampilkan klasifikasi dengan probabilitas tertentu (ditunjukkan oleh grafik batang kecil di bawah label).

Kesimpulannya, dalam proses klasifikasi menggunakan algoritma CART dengan kumpulan data teks ini, hasil dari pohon keputusan menunjukkan bahwa mayoritas data dikategorikan sebagai spam dibandingkan ham.

2. Proses Pengujian (Testing)

Setelah model CART dilatih, langkah selanjutnya adalah menguji kinerjanya dengan menggunakan metode cross-validation.

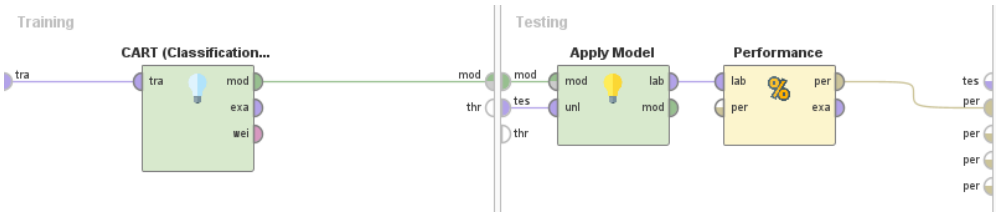
Berikut metrik evaluasi kinerja Algoritma CART :

accuracy: 92.29% +/- 0.82% (micro average: 92.29%)

	true spam	true ham	class precision
pred. spam	1063	110	90.62%
pred. ham	32	637	95.22%
class recall	97.08%	85.27%	

Gambar 4. 24 Hasil Klasifikasi CART

Diperoleh hasil akurasi untuk algoritma CART sebesar 92.29%. Artinya model dapat mengklasifikasikan email spam dan ham sejauh 92.29%. Untuk memperjelas *Confusion Matrix* akurasinya ada pada Gambar 4.24.



Gambar 4. 25 Model Rapidminer CART

Pada Gambar 4.25, terlihat berbagai operator yang ada dalam operator Cross Validation, termasuk operator CART yang berfungsi untuk mengklasifikasikan email menjadi spam dan ham, Apply Model yang berfungsi untuk menerapkan model yang telah dilatih ke dataset baru guna menghasilkan prediksi, serta Performance yang digunakan untuk mengukur dan mengevaluasi kualitas hasil prediksi model, seperti akurasi, presisi, recall, dan F1-score.

4.4.2 Algoritma C4.5

1. Proses Pelatihan (Training)

   Pada proses pelatihan, algoritma C4.5 membuat model dengan cara membangun pohon keputusan dari data latih. Langkah pertama adalah menghitung entropi untuk mengukur tingkat keragaman data. Semakin rendah entropi, semakin seragam datanya. Kemudian, algoritma menghitung Gain Ratio untuk memilih atribut terbaik sebagai simpul utama (*root node*). Atribut yang memiliki Gain Ratio tertinggi akan dipilih. Setelah simpul utama dipilih, data dibagi berdasarkan nilai atribut tersebut untuk membentuk cabang-cabang pada pohon keputusan. Proses ini terus dilakukan hingga semua data pada setiap cabang termasuk ke dalam satu kelas yaitu spam dan ham atau berhenti karena jumlah data sudah terlalu sedikit untuk dilanjutkan.

Berikut adalah pohon keputusan yang dihasilkan oleh Algoritma C4.5 :

```
vince <= 0.0
|   enron <= 0.0
|   |   penelitian <= 0.03125081831341921
|   |   |   terlampir <= 0.036826563774414484
|   |   |   |   terima <= 0.03969590573982193
|   |   |   |   |   pagi <= 0.03714908223623516
|   |   |   |   |   |   keuangan <= 0.05233845330097178
|   |   |   |   |   |   |   lunak <= 0.0
|   |   |   |   |   |   |   |   klik <= 0.010342191308728531
|   |   |   |   |   |   |   |   |   online <= 0.02210289619158463
|   |   |   |   |   |   |   |   |   |   negara <= 0.0
|   |   |   |   |   |   |   |   |   |   |   dijamin <= 0.0
|   |   |   |   |   |   |   |   |   |   |   |   pemasaran <= 0.016434333432990778
|   |   |   |   |   |   |   |   |   |   |   |   |   seks <= 0.0
|   |   |   |   |   |   |   |   |   |   |   |   |   |   menyenangkan <= 0.0
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   tulis <= 0.03189214509832559
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   gratis <= 0.05044818067421766
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   obat <= 0.0
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   menghemat <= 0.015182618971929395
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   qmail <= 0.0
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   disetujui <= 0.0
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   pengiriman <= 0.024998020823970644
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   penuh <= 0.15455659846700065
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   http <= 0.06645381133654625
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   pribadi <= 0.02676040282592395
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   asli <= 0.09546401007461774
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   memiliki <= 0.0534210502600842
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   hubungi <= 0.0
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   daftar <= 0.022238220006316627
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   berhasil <= 0.0
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   orang <= 0.006119462185737039
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   salam <= 0.0
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   kartu <= 0.033964757667643994
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   telepon <= 0.021482828620544615
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   http <= 0.03955987212946771
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   harga <= 0.009939579786123377
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   selamat <= 0.07708841202667607: spam (130.0/47.0)
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   selamat > 0.07708841202667607
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   dunia <= 0.03311868970842497: ham (5.0)
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   dunia > 0.03311868970842497: spam (2.0)
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   harga > 0.009939579786123377
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   berkualitas <= 0.033528781372951726: ham (7.0)
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   berkualitas > 0.033528781372951726: spam (2.0)
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   http > 0.03955987212946771: spam (12.0)
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   telepon > 0.021482828620544615
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   balas <= 0.028612841453390558: ham (7.0/1.0)
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   balas > 0.028612841453390558: spam (2.0)
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   kartu > 0.033964757667643994: spam (9.0/1.0)
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   salam > 0.0
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   bank <= 0.01848264247335408: ham (8.0/1.0)
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   bank > 0.01848264247335408: spam (2.0)
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   orang > 0.006119462185737039: spam (17.0/1.0)
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   berhasil > 0.0: spam (11.0)
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   daftar > 0.022238220006316627
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   subjek <= 0.010079910347979669: ham (9.0/2.0)
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   subjek > 0.010079910347979669: spam (3.0)
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   hubungi > 0.0
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   ditolak <= 0.0: ham (11.0/2.0)
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   ditolak > 0.0: spam (2.0)
|   |   |   |   |   |   |   |   |   |   |   |   |   |   memiliki > 0.0534210502600842: ham (13.0/1.0)
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   asli > 0.09546401007461774: spam (15.0)
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   pribadi > 0.02676040282592395: spam (15.0)
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   http > 0.06645381133654625: spam (28.0)
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   penuh > 0.15455659846700065: spam (17.0)
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   pengiriman > 0.024998020823970644: spam (21.0)
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   disetujui > 0.0: spam (23.0)
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   qmail > 0.0: spam (21.0)
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   menghemat > 0.015182618971929395: spam (21.0)
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   obat > 0.0: spam (22.0)
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   gratis > 0.05044818067421766: spam (44.0)
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |   tulis > 0.03189214509832559: spam (25.0)
|   |   |   |   |   |   |   |   |   |   |   |   |   |   menyenangkan > 0.0: spam (26.0)
|   |   |   |   |   |   |   |   |   |   |   |   |   seks > 0.0: spam (25.0)
|   |   |   |   |   |   |   |   |   |   |   |   pemasaran > 0.016434333432990778: spam (25.0)
|   |   |   |   |   |   |   |   |   |   |   dijamin > 0.0: spam (25.0)
|   |   |   |   |   |   |   |   |   |   negara > 0.0: spam (106.0/1.0)
|   |   |   |   |   |   |   |   |   online > 0.02210289619158463: spam (92.0)
|   |   |   |   |   |   |   |   klik > 0.010342191308728531: spam (174.0/1.0)
|   |   |   |   |   |   |   lunak > 0.0: spam (160.0)
|   |   |   |   |   |   keuangan > 0.05233845330097178
|   |   |   |   |   |   |   matematika <= 0.04340442489024796
|   |   |   |   |   |   |   |   deskripsi <= 0.04528464188065518
|   |   |   |   |   |   |   |   |   faks <= 0.0
|   |   |   |   |   |   |   |   |   |   keuangan <= 0.1365777478677908: spam (14.0)
|   |   |   |   |   |   |   |   |   |   keuangan > 0.1365777478677908: ham (3.0/1.0)
|   |   |   |   |   |   |   |   |   faks > 0.0: ham (2.0)
|   |   |   |   |   |   |   |   deskripsi > 0.04528464188065518: ham (2.0)
|   |   |   |   |   |   |   matematika > 0.04340442489024796: ham (4.0)
|   |   |   |   |   pagi > 0.03714908223623516
|   |   |   |   |   |   premier <= 0.0
```

```
|   |   |   |   |   |   |   |   |   keuangan > 0.1365777478677908: ham (3.0/1.0)
|   |   |   |   |   |   |   |   |   faks > 0.0: ham (2.0)
|   |   |   |   |   |   |   |   deskripsi > 0.04528464188065518: ham (2.0)
|   |   |   |   |   |   |   |   matematika > 0.04340442489024796: ham (4.0)
|   |   |   |   |   pagi > 0.03714908223623516
|   |   |   |   |   |   premier <= 0.0
|   |   |   |   |   |   |   dihapus <= 0.038427230845943613: ham (13.0)
|   |   |   |   |   |   |   dihapus > 0.038427230845943613: spam (2.0)
|   |   |   |   |   |   premier > 0.0: spam (10.0)
|   |   |   |   terima > 0.03969590573982193
|   |   |   |   |   klik <= 0.010342191308728531
|   |   |   |   |   |   http <= 0.05901052388518529
|   |   |   |   |   |   |   pelanggan <= 0.03306567728236988
|   |   |   |   |   |   |   |   email <= 0.0910620182114889
|   |   |   |   |   |   |   |   |   situs <= 0.024336172762630108: ham (35.0/1.0)
|   |   |   |   |   |   |   |   |   situs > 0.024336172762630108: spam (3.0/1.0)
|   |   |   |   |   |   |   |   email > 0.0910620182114889: spam (3.0)
|   |   |   |   |   |   |   pelanggan > 0.03306567728236988: spam (3.0)
|   |   |   |   |   |   http > 0.05901052388518529: spam (9.0)
|   |   |   |   |   klik > 0.010342191308728531: spam (13.0)
|   |   |   terlampir > 0.036826563774414484
|   |   |   |   program <= 0.0989891518258461: ham (29.0)
|   |   |   |   program > 0.0989891518258461: spam (4.0)
|   |   penelitian > 0.03125081831341921: ham (27.0)
|   enron > 0.0: ham (103.0)
vince > 0.0: ham (426.0)

Number of Leaves  :     56

Size of the tree :      111
```

Gambar 4. 26 Pohon Keputusan Algoritma C4.5

Pohon keputusan yang dihasilkan oleh algoritma C4.5 pada Gambar 4.26 sedikit berbeda dengan pohon keputusan decision tree lainnya, karena algoritma ini menggunakan J.48 untuk implementasinya di *rapidminer*.

2. Proses Pengujian (Testing)

   Setelah model C4.5 dilatih, langkah berikutnya adalah menguji performa model menggunakan dataset uji. Pada Proses Testing C4.5 ini, pembagian data sama seperti Algoritma CART yaitu dilakukan dengan metode Cross-Validation sebanyak 5-fold.

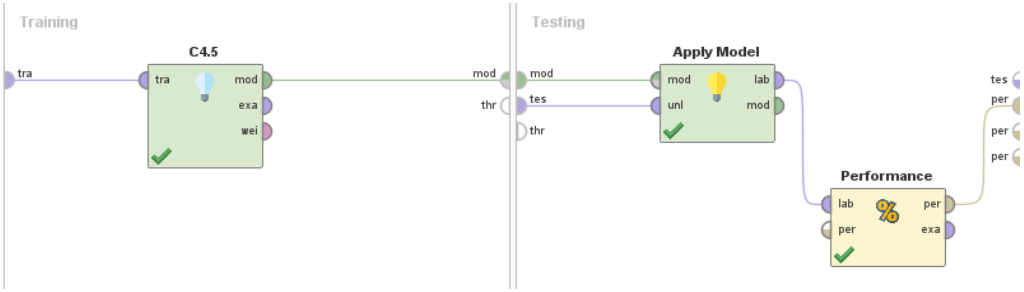
   Berikut Confusion Matrix yang dihasilkan oleh Algoritma C4.5 :

accuracy: 91.53% +/- 1.39% (micro average: 91.53%)

	true spam	true ham	class precision
pred. spam	1051	112	90.37%
pred. ham	44	635	93.52%
class recall	95.98%	85.01%	

Gambar 4. 27 Hasil Klasifikasi C4.5

Setelah pengujian dilakukan, model C4.5 menghasilkan rata-rata akurasi sebesar 91.53%, Confusion Matrix pada Gambar 4.27 menunjukkan bahwa model ini cukup efektif dalam membedakan email spam dan ham, meskipun masih terdapat kesalahan kecil pada beberapa prediksi.



Gambar 4. 28 Model Rapidminer C4.5

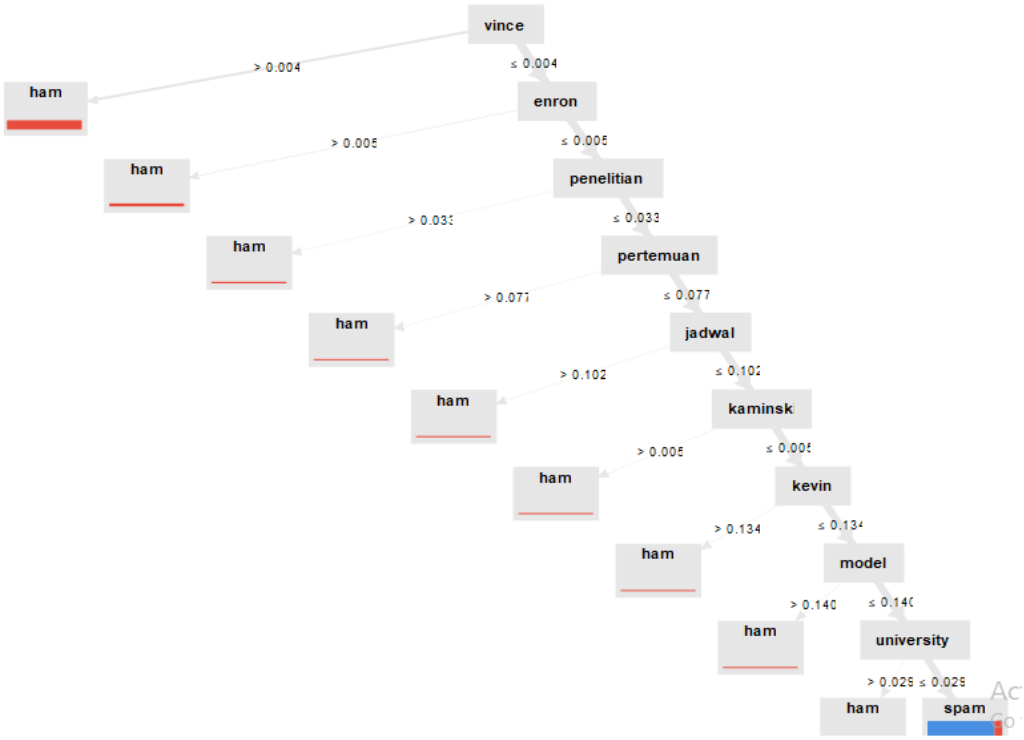
Pada Gambar 4.28, terlihat berbagai operator yang ada dalam operator Cross Validation, termasuk operator C4.5 yang berfungsi untuk mengklasifikasikan email menjadi spam dan ham, Apply Model yang berfungsi untuk menerapkan model yang telah dilatih ke dataset baru guna menghasilkan prediksi, serta Performance yang digunakan untuk mengukur dan mengevaluasi kualitas hasil prediksi model, seperti akurasi, presisi, recall, dan F1-score.

4.4.3 Algoritma C5.0

1. Proses Pelatihan (Testing)

   Pada proses pelatihan, algoritma C5.0 membangun model pohon keputusan berdasarkan data latih. Proses ini dimulai dengan memilih atribut terbaik untuk dijadikan simpul utama (*root node*). Pemilihan atribut dilakukan dengan menghitung Gain Ratio, mirip dengan algoritma C4.5. Setelah atribut utama dipilih, data dipisahkan berdasarkan nilai atribut tersebut untuk membentuk cabang-cabang pohon keputusan. Proses ini dilakukan secara rekursif hingga semua data pada setiap cabang termasuk dalam kelas spam dan ham.

Berikut adalah pohon keputusan yang dihasikan oleh Algoritma C5.0 :



Gambar 4. 29 Pohon Keputusan Algoritma C5.0

Pohon keputusan yang dihasilkan oleh Algoritma C5.0 pada Gambar 4.29 memiliki struktur berikut :

- Akar Pohon (Root Node), Pohon dimulai dari kata kunci "vince". Jika nilai atribut "vince" lebih kecil atau sama dengan 0.004, maka klasifikasi akan mengarah ke cabang berikutnya (ke kiri). Jika lebih besar dari 0.004, maka langsung diputuskan sebagai ham.
- Cabang Pertama, jika nilai "vince" ≤ 0.004, maka dipertimbangkan kata kunci "enron" dan jika "enron" ≤ 0.006, klasifikasi dilanjutkan ke cabang berikutnya. Jika lebih besar, maka diputuskan sebagai ham.
- Proses Berulang, setiap cabang pohon menampilkan kata kunci penting lainnya seperti "penelitian", "pertemuan", "jadwal", "kaminski", "kevin", "model", dan "university". Atribut-atribut ini memiliki nilai ambang batas tertentu yang memengaruhi hasil klasifikasi. Misalnya, pada atribut "penelitian" : (Jika nilainya ≤ 0.033, maka klasifikasi dilanjutkan ke cabang

"pertemuan" dan Jika nilainya > 0.033, maka langsung diputuskan sebagai ham.

- Daun Pohon (Leaf Nodes), Node paling akhir (daun) menunjukkan hasil klasifikasi, yaitu ham atau spam. Misalnya, jika cabang pohon mencapai atribut "university" dengan nilai > 0.140, hasilnya adalah ham. Jika nilainya ≤ 0.140, maka dilihat lebih lanjut untuk memastikan apakah itu spam.

Pohon ini menggambarkan proses pengambilan keputusan berbasis aturan, di mana setiap atribut menjadi faktor penentu hingga hasil akhir (spam atau ham) ditemukan. Dan proses klasifikasi menggunakan algoritma C5.0 dengan kumpulan data teks ini, hasil dari pohon keputusan menunjukkan bahwa mayoritas data dikategorikan sebagai spam dibandingkan ham.

2. Proses Pengujian (Testing)

Pengujian Algoritma C5.0 dilakukan dengan metode 5-fold Cross-Validation, sama seperti proses pembagian data yang dilakukan oleh Algoritma CART dan C4.5.

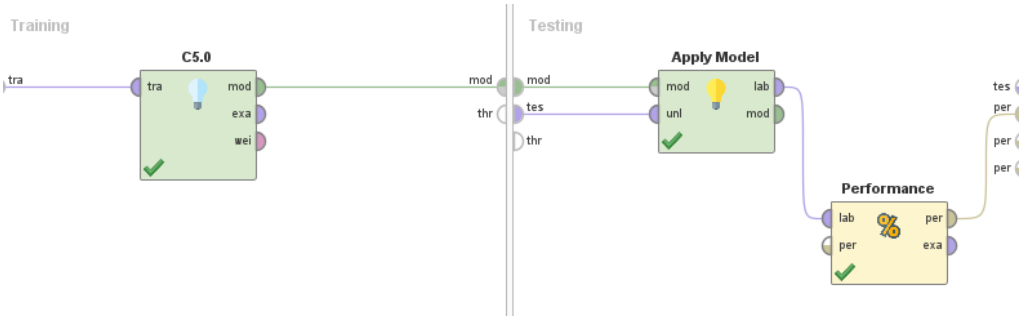
Berikut Confusion Matrix yang dihasilkan oleh Algoritma C5.0 :

accuracy: 91.97% +/- 0.24% (micro average: 91.97%)

	true spam	true ham	class precision
pred. spam	1080	133	89.04%
pred. ham	15	614	97.62%
class recall	98.63%	82.20%	

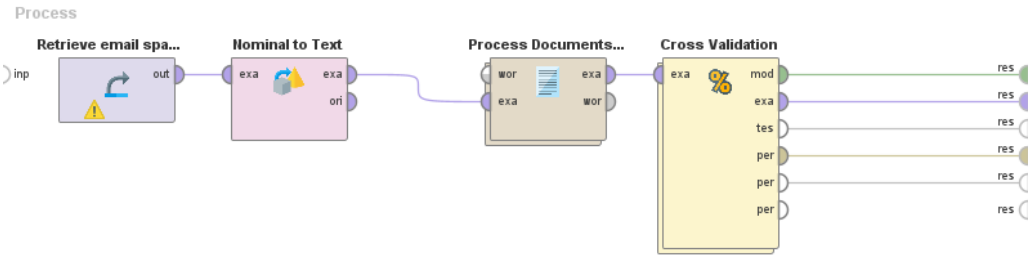
Gambar 4. 30 Hasil Klasfikasi C5.0

Diperoleh hasil akurasi untuk algoritma C5.0 sebesar 91.97%. Artinya model dapat mengklasifikasikan email spam dan ham sejauh 91.97%. Untuk melihat lebih jelas Confusion Matrix akurasi algoritma C5.0 ada pada Gambar 4.30.



Gambar 4. 31 Model Rapidminer C5.0

Pada Gambar 4.31, terlihat berbagai operator yang ada dalam operator Cross Validation, termasuk operator C5.0 yang berfungsi untuk mengklasifikasikan email menjadi spam dan ham, Apply Model yang berfungsi untuk menerapkan model yang telah dilatih ke dataset baru guna menghasilkan prediksi, serta Performance yang digunakan untuk mengukur dan mengevaluasi kualitas hasil prediksi model, seperti akurasi, presisi, recall, dan F1-score.



Gambar 4. 32 Repository Model *Rapidminer*

Repository Model di *RapidMiner* dapat dilihat pada Gambar 4.32, yang menunjukkan proses untuk mengklasifikasikan email ke dalam kategori spam atau ham, serta menguji kinerja algoritma CART, C4.5, dan C5.0.

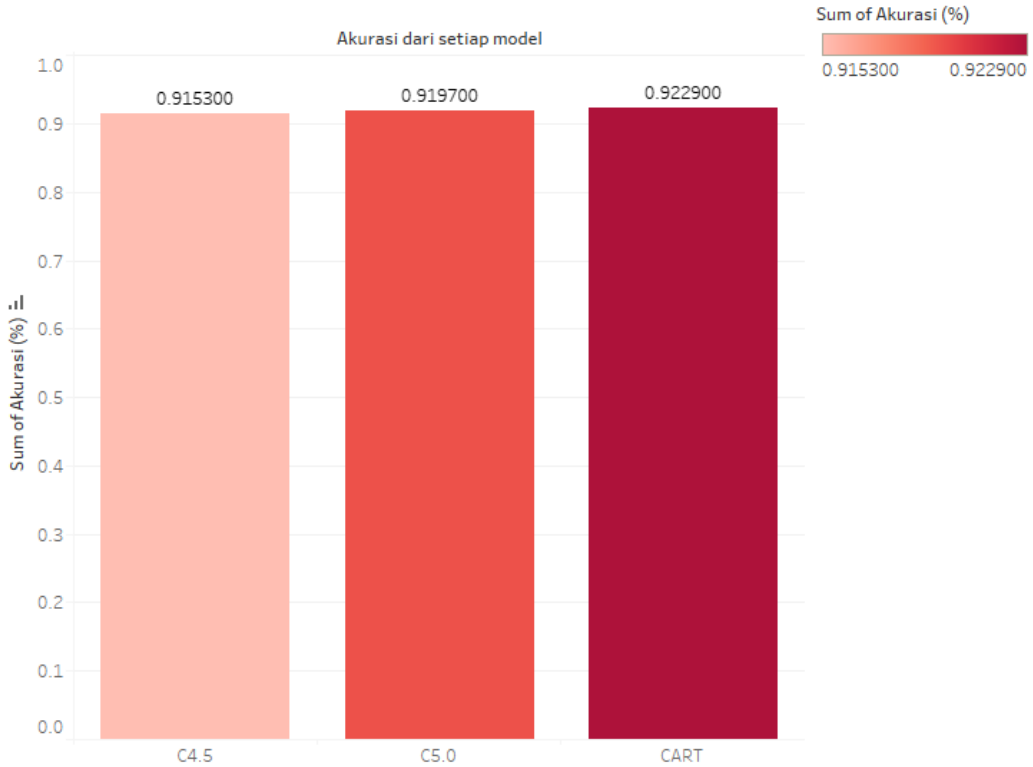
4.5 Visualisasi Hasil

Visualisasi hasil dalam penelitian ini ditampilkan dalam bentuk bar chart untuk mempermudah perbandingan kinerja algoritma CART, C4.5, dan C5.0 dalam klasifikasi email spam dan ham. Grafik ini menampilkan matrik evaluasi seperti akurasi, precision, recall, dan waktu eksekusi, sehingga dapat memberikan gambaran yang lebih jelas mengenai performa masing – masing algoritma.

Dengan adanya visualisasi ini, pola perbedaan kinerja antar algoritma dapat lebih mudah dianalisis dan memungkinkan penarikan kesimpulan mengenai metode yang paling optimal berdasakan hasil klasifikasi yang diperoleh.

4.5.1 Akurasi

Akurasi yang didapatkan oleh algortima CART, C4.5 dan C5.0 cenderung besar karena berada diatas 90%. Algoritma CART memiliki nilai akurasi sebesar 0.9229 dan algoritma C4.5 nilai akurasinya sebesar 0.9153 sedangkan algoritma C5.0 akurasinya sebesar 0.9197. Dari masing – masing nilai akurasi yang dihasilkan dari setiap model akurasi CART memiliki nilai yang paling tinggi dengan nilai akurasi sebesar 0.9229. Dari hasil performance setiap model maka bisa diartikan ke-3 algoritma ini sangat baik dalam proses klasifikasi.

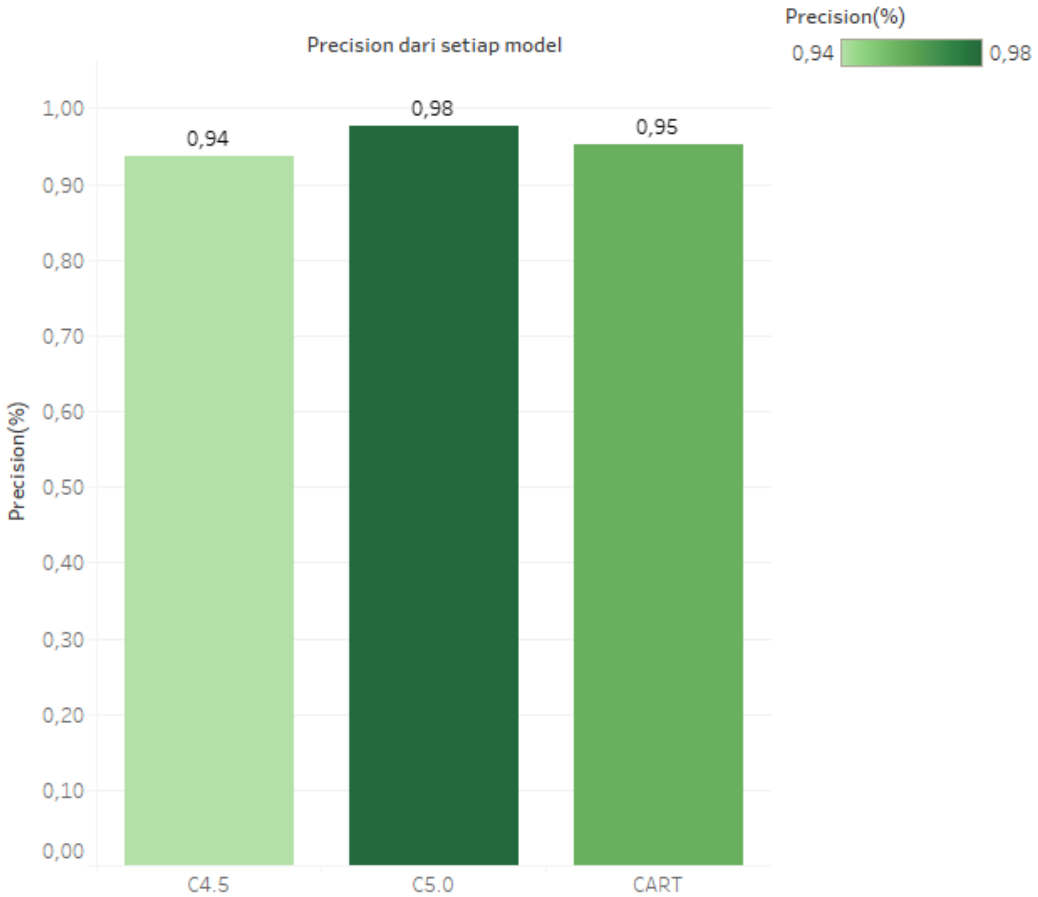


Gambar 4. 33 Perbandingan Akurasi Dari Setiap Algoritma

Visualisasi hasil perbandingan Akurasi yang terdapat pada Gambar 4.33 disajikan dalam bentuk bar chart, di mana bar dengan warna paling terang menunjukkan nilai akurasi tertinggi, sementara bar dengan warna paling gelap menunjukkan nilai akurasi terendah.

4.5.2 Precision

Nilai Precision yang dihasilkan oleh ke-3 model cenderung tinggi karena diatas 90% dan bisa dikatakan bagus kerana nilai tertingginya memiliki nilai 0.98 dari algoritma C5.0 sedangkan precision terendahnya di algoritma C4.5 dengan nilai 0.94. Dari hasil precesion yang dihasilkan, dapat disimpulakan bahwa setiap model dapat mengklasifikasikan variabel terget *true positive* dengan sangat baik. Dan model yang paling baik adalah algoritma C5.0, karena keakuratannya dalam memprediksi variabel target spam. Untuk memperjelas hasil visualisasinya ada pada Gambar 4.34.

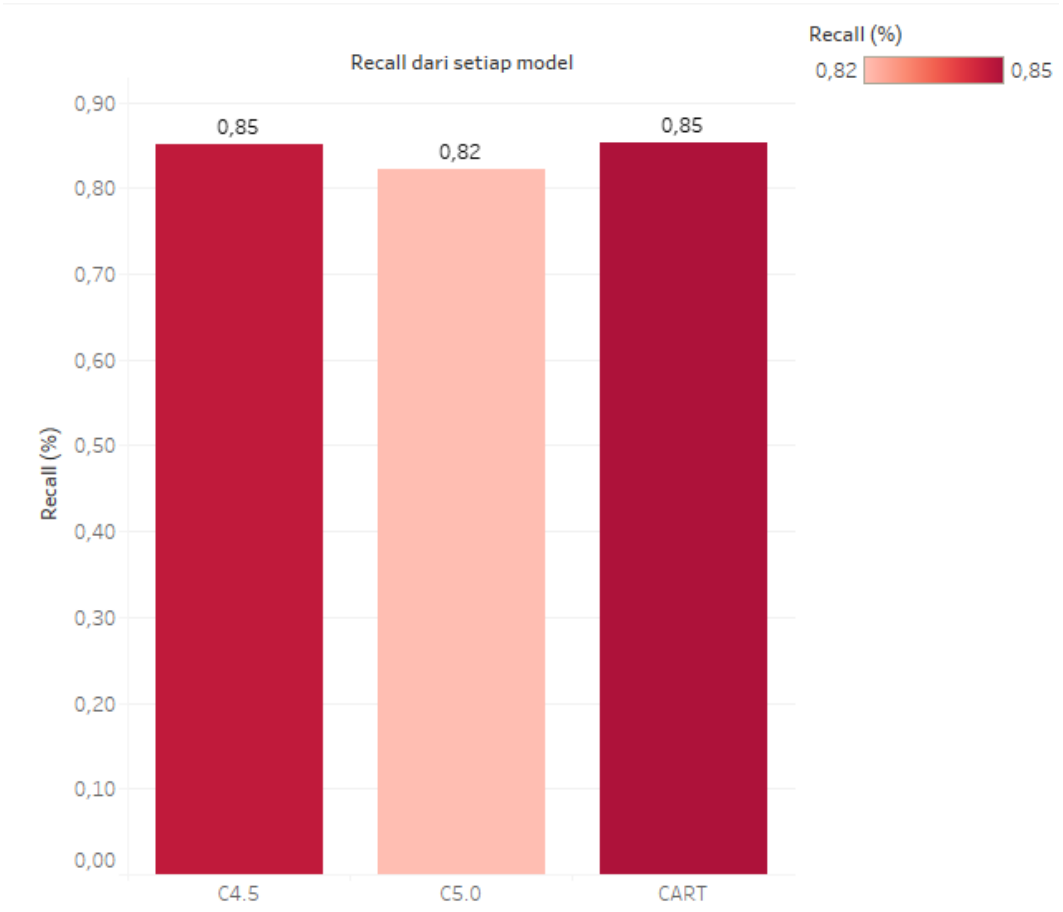


Gambar 4. 34 Perbandingan Precision Dari Setiap Algoritma

Visualisasi hasil perbandingan Precision yang terdapat pada Gambar 4.34 disajikan dalam bentuk bar chart, di mana bar dengan warna paling terang menunjukkan nilai akurasi tertinggi, sementara bar dengan warna paling gelap menunjukkan nilai akurasi terendah.

4.5.3 Recall

Nilai Recall juga tergolong tinggi meskipun tidak sebesar nilai Akurasi dan Precision. Nilai Recall algoritma CART dan C4.5 sama atau rata yaitu 0.85 sedangkan nilai recall C5.0 berada diangka terendah dengan nilai 0.82. Dari performance ke-3 model ini menunjukan performance yang cukup baik dalam mengidentifikasi kasus positif atau spam.

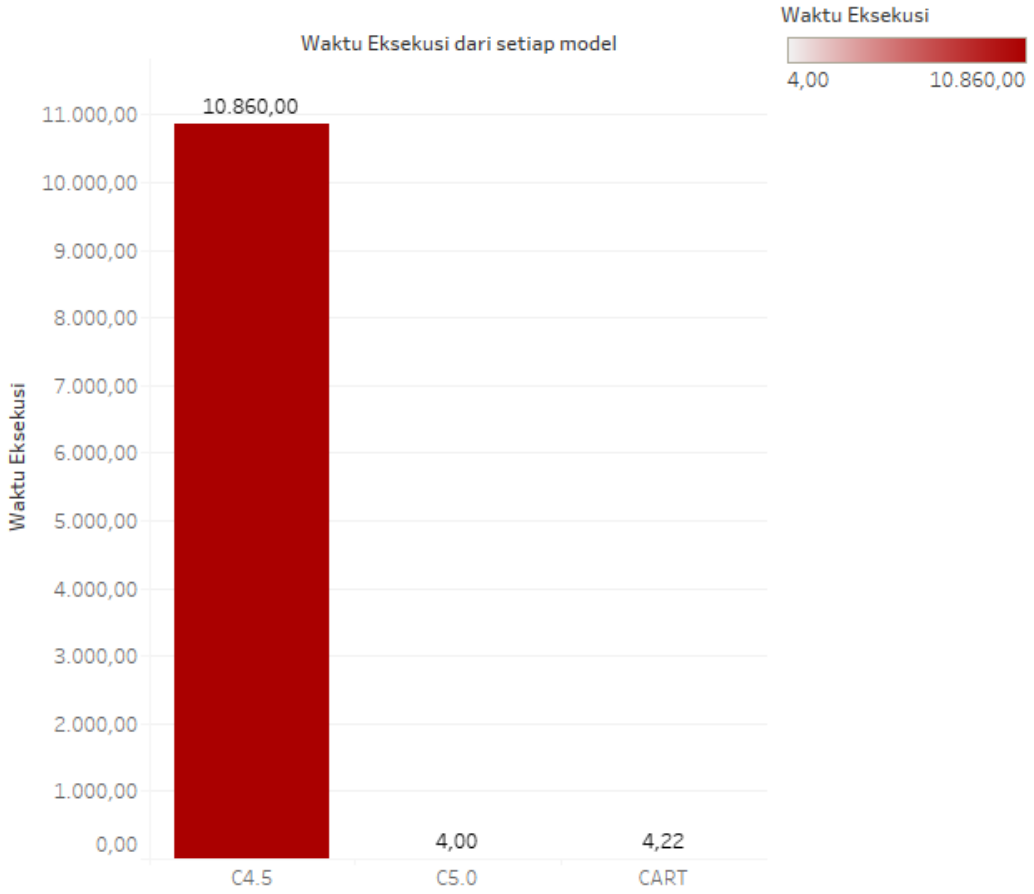


Gambar 4. 35 Perbandingan Recall Dari Setiap Algoritma

Visualisasi hasil pada Gambar 4.35 menunjukkan bahwa nilai Recall untuk CART dan C4.5 berada pada tingkat yang sama, sehingga warna bar keduanya sama, yaitu lebih terang dibandingkan bar untuk C5.0. Algoritma C5.0 memiliki nilai Recall terendah, sehingga warna barnya lebih gelap.

4.5.4 Waktu Eksekusi Algoritma

Waktu eksekusi algoritma atau berapa lama proses algortima dalam mengklasfikasikan email spam dan ham. Untuk algoritma C4.5 cukup lama yaitu 10.860 detik atau 3.01 jam sedangkan untuk algoritma C4.5 dan C5.0 hanya membutuhkan waktu 4.00 dan 4.22 detik untuk proses klasifikasi email spam dan ham. Waktu eksekusi ke-3 algoritma ini sudah termasuk proses cleaning data. Algoritma C4.5 meskipun performance Akurasi, Precision dan Recall cukup baik tapi algoritma C4.5 ini lemah dalam waktu proses klasifikasinya.



Gambar 4. 36 Waktu Eksekusi Dari Setiap Algoritma

Pada Gambar 4.36, terlihat visualisasi bar yang mewakili algoritma C4.5, yang menunjukkan waktu eksekusi lebih lama dibandingkan dengan algoritma lainnya. Hal ini terlihat dari tingginya bar C4.5, yang menggambarkan durasi waktu eksekusi yang lebih panjang.

4.6 Perbandingan Kinerja Algoritma

Setelah melakukan pengujian menggunakan 5-fold cross-validation, model CART menunjukkan akurasi tertinggi dengan nilai akurasi 92.29%, diikuti oleh C5.0 yang mencapai 91.97% dan C4.5 yang memperoleh akurasi 91.53%. Meskipun algoritma CART yang paling sederhana dibandingkan algortima C4.5 dan C5.0, CART terbukti lebih efektif dalam mengklasifikasikan data, terutama dalam hal mengidentifikasi email spam dan ham. Hasil ini menunjukkan bahwa CART lebih unggul dibandingkan kedua algoritma lainnya dalam konteks penelitian ini. Berikut Tabel perbandingan algoritma CART, C4.5 dan C5.0 :

Tabel 4. 1 Perbandingan Kinerja Algoritma

Algoritma	Akurasi (%)	Presisi (%)	Recall (%)	Waktu Eksekusi
CART	92.29%	95.30%	85.28%	4.22 detik
C4.5	91.53%	93.59%	85.00%	3.01 jam
C5.0	91.97%	97.65%	82.20%	4.0 detik

Ket :

- : Nilai Akurasi Tertinggi
- : Nilai Presisi Tertinggi
- : Nilai Recall Tertinggi
- : Waktu Eksekusi yang paling lama

Kesimpulan :

- CART unggul dalam hal akurasi dan recall.
- C4.5 memiliki performa yang cukup baik tetapi lambat dalam waktu eksekusi.
- C5.0 unggul dalam hal presisi dan efisiensi waktu dibandingkan dengan C4.5 dan CART.

4.7 Pembahasan

Berdasarkan hasil penelitian, ketiga algoritma klasifikasi, yaitu CART, C4.5, dan C5.0, menunjukkan performa yang baik dengan akurasi di atas 90%, yang menandakan kemampuan tinggi dalam klasifikasi email spam dan ham. Akurasi dalam konteks ini dihitung sebagai persentase jumlah prediksi yang benar dibandingkan dengan keseluruhan data yang diuji, sehingga mencerminkan seberapa sering model menghasilkan klasifikasi yang tepat secara umum. CART memiliki akurasi tertinggi sebesar 92.29%, diikuti oleh C5.0 sebesar 91.97%, dan C4.5 sebesar 91.53%. Selain unggul dalam akurasi, CART juga menunjukkan keunggulan dalam recall sebesar 85.28%, lebih tinggi dibandingkan dengan C5.0 yang memiliki recall terendah (82.20%). Recall sendiri mengukur seberapa baik model dalam mengidentifikasi semua email spam yang sebenarnya spam. Di sisi lain, C5.0 menunjukkan keunggulan dalam precision tertinggi, yaitu 97.65%, yang berarti model ini lebih sedikit melakukan kesalahan dalam mengklasifikasikan email sebagai spam. Precision dihitung dengan rumus Precision = TP / (TP + FP), di mana FP adalah jumlah email ham yang salah diklasifikasikan sebagai spam. Precision yang tinggi menunjukkan bahwa ketika model memprediksi sebuah email sebagai spam, kemungkinan besar prediksi tersebut benar. Selain itu, dari segi efisiensi waktu eksekusi, C5.0 menjadi yang tercepat dengan waktu hanya 4 detik, sedangkan C4.5 meskipun memiliki akurasi dan recall yang kompetitif, memiliki kelemahan dalam waktu eksekusi yang sangat lama, yaitu 3.01 jam, sehingga kurang efisien untuk aplikasi real-time.

Hasil ini menunjukkan bahwa CART menjadi pilihan terbaik secara keseluruhan karena keseimbangan antara akurasi, recall, dan efisiensi waktu. Sementara itu, C5.0 lebih cocok untuk kasus yang memerlukan precision tinggi dan waktu eksekusi cepat, sedangkan C4.5 dapat digunakan dalam situasi di mana waktu bukan menjadi kendala utama.