

Analisis Penerapan Data Mining dalam Klasifikasi Penjualan Pakaian Pada Toko Online Shopee Menggunakan Algoritma C4.5

M. Redinal Muktar¹, Muhammad Rifqi Faidhil Syam², Asri Kunda, Muhammad Syahlan Natsir

^{1,2} *Jurusan Sistem Informasi Universitas Dipa Makassar*
Jln. Perintis Kemerdekaan KM. 9 Makassar

¹redinalmuktar@gmail.com ²rifqifaidhil.al@gmail.com ³kundaasri@yahoo.com ⁴sahlan@undipa.ac.id

Abstrak

Shopee adalah salah satu contoh platform penjualan online yang dapat diakses oleh publik sehingga shopee dapat menjadi wadah untuk penjual melakukan transaksi jual beli. Perkembangan trend fashion dikemas melalui kecanggihan teknologi informasi yang tiada henti, antara lain iklan yang menarik, penjualan yang mudah diakses, sehingga pengaruh ini berimbas kepada kebutuhan gaya hidup anak milenial zaman sekarang. Beberapa dari mereka menghabiskan uang dengan terus-menerus membeli produk-produk yang sesuai dengan trend tetapi tidak menyesuaikan dengan diri sendiri. Semakin pesatnya pertumbuhan telah menciptakan kondisi banyaknya data jenis produk namun minim informasi mengenai minat pembeli, hal ini tentu saja dibutuhkan oleh penjual khususnya penjual online. Maka penelitian ini akan melakukan analisis data mining dalam klasifikasi penjualan pakaian menggunakan algoritma C4.5 dengan model decision tree dengan tingkat akurasi berdasarkan klasifikasi nilai AUC (Area Under Curve). yang terbagi dengan 2 kelas berdasarkan atribut toko/brand, jenis, bahan, rating dan harga. Hasil urutan klasifikasi minat konsumen yang dapat mempengaruhi penjualan pakaian dilihat dari hasil decision tree yakni atribut rating, akan tetapi berdasarkan nilai gain tertinggi yakni atribut toko/brand akan tetapi dataset dengan nilai weight/bobot tertinggi dari setiap atribut ialah atribut harga dengan weight atau bobot 0,667., dengan urutan klasifikasi selanjutnya ialah toko/brand, rating, bahan dan jenis pakaian dengan nilai weight atau bobot masing-masing adalah 0,125., 0,110., 0,098. Hasil perhitungan confusion matrix menghasilkan nilai accuracy mencapai 80,95% precision rendah sebesar 57,45% dan nilai precision tinggi sebesar 87,32% serta nilai recall rendah 55,10% dan nilai recall tinggi 88,34%. Hasil nilai AUC (Area Under Curve) pada aplikasi rapidminer ialah 0,721 dan dikategorikan ke dalam Klasifikasi Cukup.

Kata kunci: *Rapidminer, Decision Tree, Algoritma C4.5, Confussion Matrix, AUC.*

I. PENDAHULUAN

Perkembangan trend fashion dikemas melalui kecanggihan teknologi informasi yang tiada henti, antara lain iklan yang menarik, penjualan yang mudah diakses, sehingga pengaruh ini berimbas kepada kebutuhan gaya hidup anak milenial zaman sekarang. Shopee adalah salah satu contoh platform penjualan online yang dapat diakses oleh publik sehingga shopee dapat menjadi wadah untuk penjual melakukan transaksi jual beli [8].

Beberapa dari mereka menghabiskan uang dengan terus-menerus membeli produk-produk yang sesuai dengan trend tetapi tidak menyesuaikan dengan diri sendiri. Semakin pesatnya pertumbuhan telah menciptakan kondisi banyaknya data jenis produk namun minim informasi mengenai minat pembeli, hal ini tentu saja dibutuhkan oleh penjual khususnya penjual online.

Data mining adalah suatu proses kumpulan data yang jumlahnya besar untuk mengetahui nilai tambahnya [2]. Beberapa pengelompokan data mining berdasarkan tujuan yang akan dicapai, yaitu ada *description* (deskripsi), *estimation* (estimasi), *prediction* (prediksi), *classification*

(klasifikasi), *clustering* (pengklusteran) dan *association* (asosiasi) [7].

Pembangunan pohon keputusan dilakukan untuk mengklasifikasikan data yang dibentuk menjadi data input pada algoritma yang terdiri dari beberapa objek dan atribut [6].

Algoritma C4.5 ini adalah salah satu metode klasifikasi dalam bentuk pohon keputusan. Algoritma pohon keputusan adalah berbasis logika, dan mudah diartikan dan populer serta dapat digunakan secara luas pada berbagai jenis kasus dalam data mining dan dapat menghasilkan kinerja yang baik dan cukup mudah dipahami [3].

klasifikasi adalah agar dapat menemukan model dari *training set* agar dapat melakukan pemisahan atribut berdasarkan kategori atau kelas yang sama/sesuai, selanjutnya penggunaan model ini digunakan untuk pengklasifikasian kelas suatu atribut yang belum diketahui sebelumnya [1].

Confussion matrix adalah penggunaan metode untuk menemukan hasil nilai akurasi data mining. Pembentukan ini digunakan untuk mengevaluasi suatu model klasifikasi berdasarkan proses hitung dari data testing yang akan memprediksi suatu nilai benar atau [10].

Melakukan analisis data mining dalam klasifikasi penjualan pakaian menggunakan algoritma C4.5 dengan model decision tree dengan tingkat akurasi berdasarkan klasifikasi nilai AUC (Area Under Curve).

Dengan adanya hasil klasifikasi penjualan pakaian beberapa toko/brand dengan menggunakan metode algoritma C4.5 diharapkan Shopee dan toko/brand mendapatkan feedback yang baik dan dapat meningkatkan angka penjualannya.

II. METODOLOGI PENELITIAN

Pada penelitian ini akan dilakukan analisis terhadap hasil berupa decision tree yang diperoleh dari subjek data yaitu data penjualan pakaian pada 8 toko online shopee dengan range waktu 1 Tahun yaitu dari Bulan Oktober 2021 sampai Bulan Oktober 2022.

Jenis penelitian yang akan dilakukan adalah jenis penelitian kuantitatif. Bahan Penelitian yang akan digunakan ialah 8 toko online di Shopee yaitu Erigo, starcross, Russ&Co, Roughneck 1991, Shinning Bright, RSCH, Epidemic dan Reclays. Variabel bebas dan variabel terikat pada penelitian ini ialah, variabel bebas dalam penelitian ini kelima atribut yaitu Toko/Brand, Jenis, Bahan, Harga dan Rating, sedangkan variabel terikat dalam penelitian ini yaitu daya minat konsumen berdasarkan angka terjual pakaian dengan label rendah dan tinggi.

A. Menghitung nilai rata – rata dari total penjualan dari beberapa brand pakaian. Dengan rumus sebagai berikut [4]:

$$\text{Rata - rata} = \frac{(\text{Total Terjual})}{(\text{Banyak Data})} \tag{1}$$

B. Membentuk decision tree menggunakan algoritma C4.5 dengan menghitung nilai entropy total seluruh data menggunakan rumus [4]:

$$\text{Entropy (s)} = \sum_{i=1}^n - p_i \log_2 p_i \tag{2}$$

C. Menghitung nilai entropy dari nilai setiap atribut dengan rumus [4]:

$$\text{Entropy Atribut} = \sum_{i=1}^n - \frac{p_i}{p} \log_2 \left(\frac{p_i}{p} \right) - \frac{p_i}{p} \log_2 \left(\frac{p_i}{p} \right) \tag{3}$$

D. Menghitung nilai gain dari setiap atribut dengan rumus [4]:

$$\text{Gain[s, A]} = \text{Entropy (s)} - \left(\left(\frac{p_i}{p} \text{ Atribut} \right) + \left(\frac{p_i}{p} \text{ Atribut} \right) + \dots + \left(\frac{p_i}{p} \text{ Atribut} \right) \right) \tag{4}$$

E. Membuat akar decision tree menggunakan atribut yang memiliki nilai gain tertinggi

F. Melakukan kembali proses pada point 6 sampai semua atribut masuk pada decision tree

G. Melihat atribut yang memiliki nilai gain tertinggi

H. Melakukan evaluasi akurasi untuk mengetahui keakuratan algoritma C4.5 menggunakan confusion matrix dengan rumus [9]:

$$\text{Accuracy} = \frac{TT + TR}{TT + TR + FT + FR} \tag{5}$$

$$\text{Precision} = \frac{TT}{TT + FT}; \frac{\sum \text{PrecisionR,T}}{\text{JumlahKelas}} \tag{6}$$

$$\text{Recall} = \frac{TT}{TT + FR}; \frac{\sum \text{RecallR,T}}{\text{JumlahKelas}} \tag{7}$$

Penelitian ini menggunakan metode pengujian confusion matrix. Confusion matrix digambarkan dengan tabel yang menyatakan jumlah data uji yang benar dan jumlah data uji yang salah. Konsep ini dilakukan untuk mengevaluasi model klasifikasi berdasarkan proses hitung data testing yang memprediksi nilai benar atau salah.

Pada akurasi klasifikasi data mining, nilai AUC (Area Under Curve) dapat dibagi menjadi beberapa kelompok yaitu [4]:

- 1) 0,9 – 1,0 = Klasifikasi sangat baik
- 2) 0,8 - 0,9 = Klasifikasi baik
- 3) 0,7 – 0,8 = Klasifikasi cukup
- 4) 0,6 – 0,7 = Klasifikasi buruk
- 5) 0,5 – 0,6 = Klasifikasi salah

Berikut perhitungan manual untuk mendapatkan hasil nilai AUC (Area Under Curve) [4]:

$$\text{Specificit y} = \frac{TR}{TR + FT}; \frac{\sum \text{SpecificityR,T}}{\text{JumlahKelas}} \tag{8}$$

$$\text{AUC} = \frac{\text{Recall} + \text{Specificit y}}{\text{JumlahKelas}} \tag{9}$$

III. HASIL DAN PEMBAHASAN

Proses klasifikasi pada penelitian ini menggunakan Algoritma C4.5 pada penjualan pakaian pada 8 toko online di Shopee yaitu Erigo, starcross, Russ&Co, Roughneck 1991, Shinning Bright, RSCH, Epidemic dan Reclays. Dataset yang digunakan berupa atribut toko/brand, jenis, bahan, harga dan rating penjualan dengan metode confusion matrix menggunakan aplikasi Rapidminer. Data yang digunakan berjumlah 1471 data.

Tabel 1 Data Penjualan Pakaian

NO	TOKO	JENIS	HARGA	BAHAN	BINTANG	TERJUAL
1	ERIGO	KAOS PENDEK	300000	Katun	4,9	531
2	ERIGO	KAOS PENDEK	250000	Katun	4,9	829
3	ERIGO	KAOS PENDEK	300000	Katun	4,9	464
4	ERIGO	KAOS PENDEK	300000	Katun	4,9	185
5	ERIGO	KAOS PENDEK	300000	Katun	4,8	105
6	ERIGO	KAOS PENDEK	300000	Katun	4,9	463
7	ERIGO	KAOS PENDEK	300000	Katun	4,9	179
8	ERIGO	KAOS PENDEK	300000	Katun	4,8	96
9	ERIGO	KAOS PENDEK	250000	Katun	4,9	442
10	ERIGO	KAOS PENDEK	250000	Katun	4,9	672
11	ERIGO	KAOS PENDEK	250000	Katun	4,9	475
12	ERIGO	KAOS PENDEK	250000	Katun	4,9	438
13	ERIGO	KAOS PENDEK	250000	Katun	4,9	625
14	ERIGO	KAOS PENDEK	250000	Katun	4,9	773
...						
1457	RECLAYS	CELANA PANJANG	450000	Denim	4,9	14
1458	RECLAYS	CELANA PANJANG	350000	Denim	4,6	1500
1459	RECLAYS	SWEATER	25500	Fleece	4,8	609
1460	RECLAYS	SWEATER	255000	Fleece	4,8	593
1461	RECLAYS	SWEATER	255000	rench Terr	4,8	786
1462	RECLAYS	SWEATER	265000	rench Terr	4,9	14
1463	RECLAYS	SWEATER	350000	rench Terr	4,8	5
1464	RECLAYS	SWEATER	315000	rench Terr	4,8	126
1465	RECLAYS	SWEATER	315000	rench Terr	4,6	127
1466	RECLAYS	SWEATER	315000	rench Terr	4,8	648
1467	RECLAYS	SWEATER	315000	Katun	4,8	228
1468	RECLAYS	SWEATER	400000	FLEECE	4,5	6
1469	RECLAYS	SWEATER	400000	FLEECE	4	1
1470	RECLAYS	SWEATER	400000	FLEECE	5	9
1471	RECLAYS	SWEATER	400000	FLEECE	5	5

Selain atribut, kelas/label output yang digunakan juga dilakukan pengubahan nilai numerik ke dalam bentuk kata dengan perhitungan rata-rata. Rumus dan perhitungan rata-rata ialah sebagai berikut:

$$\text{Rata - rata} = \frac{414707}{1471} = 282$$

Setelah didapatkan nilai rata-rata dari data tersebut, ketentuan untuk pengubahan nilai numerik ke nilai label kata adalah sebagai berikut: *terjual* ≤ 282 pcs : Rendah *terjual* > 282 pcs : Tinggi

Berikut adalah data yang telah dilakukan pengubahan nilai numerik menjadi nilai label kelas kata.

Tabel 2 Data Perubahan Nilai Numerik Menjadi Nilai Label

NO	TOKO	JENIS	HARGA	BAHAN	BINTANG	TERJUAL
1	ERIGO	KAOS PENDEK	300000	Katun	4,9	RENDAH
2	ERIGO	KAOS PENDEK	250000	Katun	4,9	RENDAH
3	ERIGO	KAOS PENDEK	300000	Katun	4,9	RENDAH
4	ERIGO	KAOS PENDEK	300000	Katun	4,9	RENDAH
5	ERIGO	KAOS PENDEK	300000	Katun	4,8	RENDAH
6	ERIGO	KAOS PENDEK	300000	Katun	4,9	RENDAH
7	ERIGO	KAOS PENDEK	300000	Katun	4,9	RENDAH
8	ERIGO	KAOS PENDEK	300000	Katun	4,8	RENDAH
9	ERIGO	KAOS PENDEK	250000	Katun	4,9	RENDAH
10	ERIGO	KAOS PENDEK	250000	Katun	4,9	RENDAH
11	ERIGO	KAOS PENDEK	250000	Katun	4,9	RENDAH
12	ERIGO	KAOS PENDEK	250000	Katun	4,9	RENDAH
13	ERIGO	KAOS PENDEK	250000	Katun	4,9	RENDAH
14	ERIGO	KAOS PENDEK	250000	Katun	4,9	RENDAH
...						
1457	RECLAYS	CELANA PANJANG	450000	Denim	4,9	RENDAH
1458	RECLAYS	CELANA PANJANG	350000	Denim	4,6	RENDAH
1459	RECLAYS	SWEATER	25500	Fleece	4,8	RENDAH
1460	RECLAYS	SWEATER	255000	Fleece	4,8	RENDAH
1461	RECLAYS	SWEATER	255000	rench Terr	4,8	RENDAH
1462	RECLAYS	SWEATER	265000	rench Terr	4,9	RENDAH
1463	RECLAYS	SWEATER	350000	rench Terr	4,8	RENDAH
1464	RECLAYS	SWEATER	315000	rench Terr	4,8	RENDAH
1465	RECLAYS	SWEATER	315000	rench Terr	4,6	RENDAH
1466	RECLAYS	SWEATER	315000	rench Terr	4,8	RENDAH
1467	RECLAYS	SWEATER	315000	Katun	4,8	RENDAH
1468	RECLAYS	SWEATER	400000	FLEECE	4,5	RENDAH
1469	RECLAYS	SWEATER	400000	FLEECE	4	RENDAH
1470	RECLAYS	SWEATER	400000	FLEECE	5	RENDAH
1471	RECLAYS	SWEATER	400000	FLEECE	5	RENDAH

A. Proses Pembentukan Decision Tree Menggunakan Algoritma C4.5

1) Perhitungan Nilai Entropy Total

Diketahui jumlah kasus pada seluruh data adalah 1471 kasus dengan kelas terjual tinggi sebanyak 326 dan kelas terjual rendah sebanyak 1145. Maka nilai entropy totalnya yang didapatkan sebagai berikut:

$$\text{Entropy Total [326,1145]} = - \frac{326}{1471} \log_2 \left(\frac{326}{1471} \right) - \frac{1145}{1471} \log_2 \left(\frac{1145}{1471} \right) = 0,7631$$

2) Perhitungan Nilai Entropy dan Gain Setiap Atribut

Setelah hasil entropy total telah didapatkan maka selanjutnya menghitung nilai entropy dan nilai gain pada setiap atribut sebagai berikut:

Atribut Toko/Brand

Brand Erigo:

Jumlah Kasus : 291

Kelas Tinggi : 103

Kelas Rendah : 188

$$\text{Entropy Total [188,103]} = - \frac{188}{291} \log_2 \left(\frac{188}{291} \right) - \frac{103}{291} \log_2 \left(\frac{103}{291} \right) = 0,9375$$

$$\text{Gain}[s, A] = 0,7631 - \left(\left(\frac{291}{1471} \cdot 0,9375 \right) + \left(\frac{163}{1471} \cdot 0,7268 \right) + \left(\frac{168}{1471} \cdot 0,5435 \right) + \left(\frac{179}{1471} \cdot 0,6521 \right) + \left(\frac{140}{1471} \cdot 0,6769 \right) + \left(\frac{194}{1471} \cdot 0,0828 \right) + \left(\frac{198}{1471} \cdot 0,8230 \right) + \left(\frac{138}{1471} \cdot 0,9902 \right) \right)$$

....
Atribut Jenis

Kaos Pendek

Jumlah Kasus : 528
 Kelas Tinggi : 173
 Kelas Rendah : 355

$$\text{Entropy Total [355,173]} = -\frac{355}{528} \log_2\left(\frac{355}{528}\right) - \frac{173}{528} \log_2\left(\frac{173}{528}\right) = 0,9125$$

$$\text{Gain}[s, A] = 0,7631 - \left(\left(\frac{528}{1471} \cdot 0,9125\right) + \left(\frac{86}{1471} \cdot 0,6127\right) + \left(\frac{79}{1471} \cdot 0,5480\right) + \left(\frac{162}{1471} \cdot 0,5032\right) + \left(\frac{143}{1471} \cdot 0,6019\right) + \left(\frac{137}{1471} \cdot 0,8309\right) + \left(\frac{28}{1471} \cdot 0,7496\right) + \left(\frac{106}{1471} \cdot 0,5882\right) + \left(\frac{66}{1471} \cdot 0,6500\right) + \left(\frac{136}{1471} \cdot 0,6385\right) \right)$$

.....

Atribut Bahan
 Katun

Jumlah Kasus : 834
 Kelas Tinggi : 203
 Kelas Rendah : 631

$$\text{Entropy Total [631,203]} = -\frac{631}{834} \log_2\left(\frac{631}{834}\right) - \frac{203}{834} \log_2\left(\frac{203}{834}\right) = 0,2517$$

$$\text{Gain}[s, A] = 0,7631 - \left(\left(\frac{834}{1471} \cdot 0,8007\right) + \left(\frac{100}{1471} \cdot 0,7219\right) + \left(\frac{83}{1471} \cdot 0,7315\right) + \left(\frac{22}{1471} \cdot 0,9024\right) + \left(\frac{52}{1471} \cdot 0,3912\right) + \left(\frac{179}{1471} \cdot 0,8222\right) + \left(\frac{50}{1471} \cdot 0,6808\right) + \left(\frac{14}{1471} \cdot 0,7496\right) + \left(\frac{6}{1471} \cdot 0\right) + \left(\frac{8}{1471} \cdot 0\right) + \left(\frac{5}{1471} \cdot 0\right) + \left(\frac{14}{1471} \cdot 0,5917\right) + \left(\frac{40}{1471} \cdot 0,2864\right) + \left(\frac{38}{1471} \cdot 0,3984\right) + \left(\frac{18}{1471} \cdot 0,9183\right) + \left(\frac{1}{1471} \cdot 0\right) + \left(\frac{4}{1471} \cdot 0,8113\right) + \left(\frac{2}{1471} \cdot 1\right) + \left(\frac{1}{1471} \cdot 0\right) \right)$$

Berikut merupakan tabel hasil dari perhitungan entropy total, entropy dan gain setiap atribut.

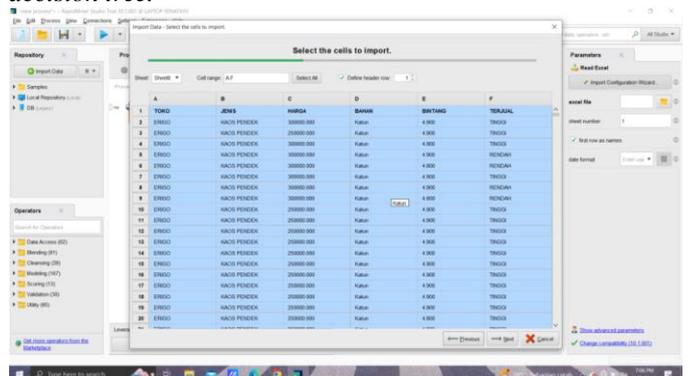
Tabel 3 Hasil Perhitungan Entropy dan Gain

KETERANGAN	JUMLAH KASUS	RENDAH	TINGGI	ENTROPY	GAIN
TOTAL	1471	1145	326	0,7631	
BRAND					
ERIGO	291	188	103	0,9375	
STARCROSS	163	130	33	0,7268	
RUSS&CO	168	147	21	0,5436	
ROUGHNECK 1991	179	149	30	0,6522	
SHINING BRIGHT	140	115	25	0,6769	
RSCH	194	192	2	0,0828	
EPIDEMIC	198	147	51	0,8231	
RECLAYS	138	77	61	0,9903	
JENIS					
KAOS PENDEK	528	355	173	0,9125	0,0766
KAOS PANJANG	86	73	13	0,6127	
KEMEJA PENDEK	79	69	10	0,5480	
KEMEJA PANJANG	162	144	18	0,5033	
JAKET	143	122	21	0,6019	
HOODIE	137	101	36	0,8309	
BOMBER/varsity	28	22	6	0,7496	
CELANA PENDEK	106	91	15	0,5882	
CELANA PANJANG	66	55	11	0,6500	
SWEATER	136	114	22	0,6385	
BAHAN					
Katun	834	631	203	0,8007	0,0227
FLANEL	100	80	20	0,7219	
POLYESTER	83	66	17	0,7315	
Nilon	22	15	7	0,9024	
PARASUT	52	48	4	0,3912	
Fleece	179	133	46	0,8222	
Denim	50	41	9	0,6801	
CANVAS	14	11	3	0,7496	
MARSOTO	6	6	0	0	
CORDUROY	8	8	0	0	
WRINKLE FABRIC	5	5	0	0	
CHINO	14	12	2	0,5917	
LINEN	40	38	2	0,2864	
BABY TERRY	38	35	3	0,3985	
FRENCH TERRY	18	12	6	0,9183	
LAKEN	1	1	0	0	
POPLIN	4	1	3	0,8113	
AMERIKAN DRILL	2	1	1	1	
WOOL	1	1	0	0	

3) Decision Tree

Hasil decision tree dari setiap toko ialah atribut rating. Berdasarkan nilai gain yang didapatkan dari semua atribut, nilai gain tertinggi adalah atribut toko/brand akan tetapi dataset dengan nilai weight/bobot tertinggi dari setiap atribut ialah atribut harga maka pada pemodelan decision tree akar pertama pada setiap toko ialah atribut rating yang kemudian minat konsumen pada setiap toko/brand yaitu berdasarkan atribut harga.

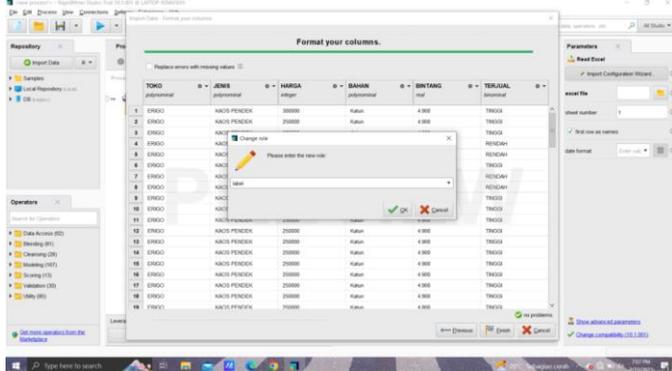
Berikut adalah model *decision tree* yang telah dibangun dengan memproses data menggunakan aplikasi *Rapidminer*. Pertama memasukkan data mentah untuk pengolahan *decision tree*.



Gambar 1 Data Penjualan Pakaian

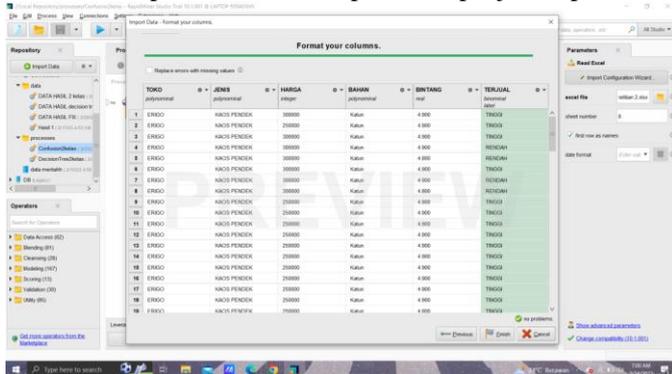
kemudian menentukan label kelas yang akan digunakan.

Berikut menentukan label kelas yang akan digunakan.



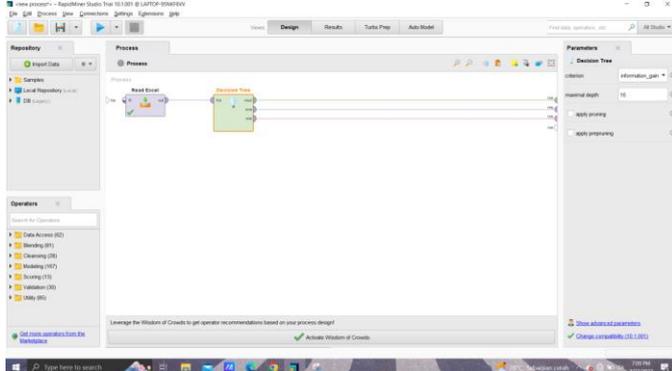
Gambar 2 Proses Penambahan Label Pada Data Penjualan Pakaian

Berikut hasil penambahan label pada data penjualan pakaian.



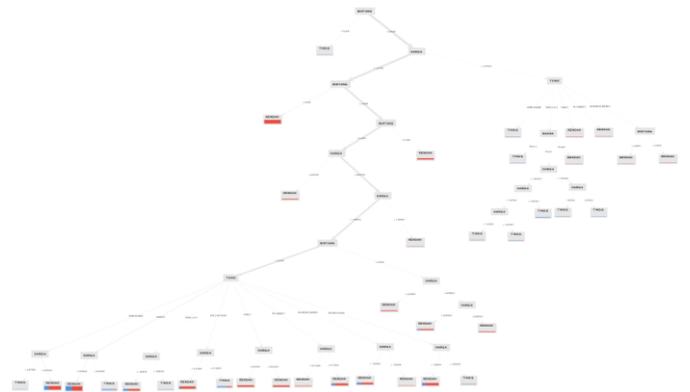
Gambar 3 Penambahan Label Pada Data Penjualan Pakaian

Selanjutnya memasukkan Operator *Decision tree* untuk mengolah dataset menjadi model *decision tree*.



Gambar 4 Proses Pengolahan Decision Tree menggunakan Aplikasi RapidMiner

Berikut adalah model *decision tree* yang telah dibangun dengan memproses data menggunakan aplikasi Rapidminer.



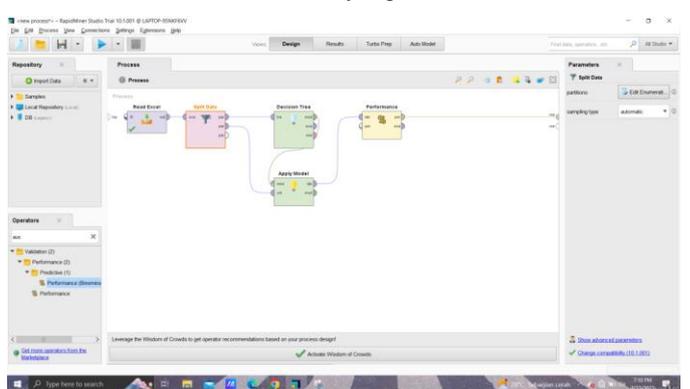
Gambar 5 Model Decision Tree

B. Uji Akurasi Algoritma C4.5 Menggunakan Confusion Matrix

Hasil perhitungan confusion matrix dari olahan aplikasi rapidminer menunjukkan jumlah True Tinggi (TT) pada kelas rendah sebanyak 303, True Tinggi (TT) pada kelas tinggi sebanyak 54, True Rendah (TR) pada kelas rendah sebanyak 54, True Rendah (TR) pada kelas tinggi sebanyak 303, False Tinggi (FT) pada kelas rendah sebanyak 44, False Tinggi (FT) pada kelas tinggi sebanyak 40, dan False Rendah (FR) pada kelas rendah sebanyak 44, False Rendah (FR) pada kelas tinggi sebanyak 40.

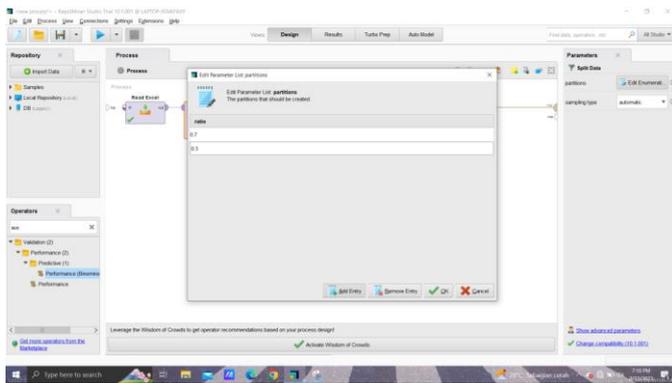
1) Pengolahan confusion matrix dan AUC (Area Under Curve) menggunakan aplikasi rapidminer

Memasukkan dataset dan Operator *Decision tree* selanjutnya memasukkan *performance* untuk perhitungan *confusion matrix* dengan menggunakan operator split data dan apply model agar dapat menjalankan perhitungan *confusion matrix* berdasarkan hasil *decision tree* yang telah diolah.



Gambar 6 Proses Pengolahan Confusion Matrix

Selanjutnya memasukkan nilai partitions dengan perbandingan 70% dan 30%.



Gambar 7 Proses Input Nilai Pembagian Dataset

2) Pengolahan accuracy, recall dan precision menggunakan aplikasi Rapidminer

accuracy: 80.95%

	true TINGGI	true RENDAH	class precision
pred. TINGGI	54	40	57.45%
pred. RENDAH	44	303	87.32%
class recall	55.10%	88.34%	

Gambar 8 Hasil Nilai Accuracy Confusion Matrix

recall: 88.34% (positive class: RENDAH)

	true TINGGI	true RENDAH	class precision
pred. TINGGI	54	40	57.45%
pred. RENDAH	44	303	87.32%
class recall	55.10%	88.34%	

Gambar 9 Hasil Nilai Recall Confusion Matrix

precision: 87.32% (positive class: RENDAH)

	true TINGGI	true RENDAH	class precision
pred. TINGGI	54	40	57.45%
pred. RENDAH	44	303	87.32%
class recall	55.10%	88.34%	

Gambar 10 Hasil Nilai Precision Confusion Matrix

3) Perhitungan accuracy, recall, precision, spesificity dan AUC

$$Accuracy = \frac{TT + TR}{TT + TR + FT + FR}$$

$$Accuracy = \frac{54 + 303}{54 + 44 + 40 + 303} = 0,8095$$

$$Precision = \frac{TT}{TT + FT}; \frac{\sum PrecisionR,T}{JumlahKelas}$$

$$Precision Rendah = \frac{303}{303 + 44} = 0,8732$$

$$Precision Tinggi = \frac{54}{54 + 40} = 0,5745$$

$$Precision = \frac{0,8732 + 0,5745}{2} = 0,7238$$

$$Recall = \frac{TT}{TT + FR}; \frac{\sum RecallR,T}{JumlahKelas}$$

$$Recall Rendah = \frac{303}{303 + 40} = 0,8834$$

$$Recall Tinggi = \frac{54}{54 + 44} = 0,5510$$

$$Recall = \frac{0,8834 + 0,5510}{2} = 0,7172$$

$$Specificity = \frac{TR}{TR + FT}; \frac{\sum SpecificityR,T}{JumlahKelas}$$

$$Specificity Rendah = \frac{54}{54 + 40} = 0,5745$$

$$Specificity Tinggi = \frac{303}{303 + 347} = 0,8732$$

$$Specificity = \frac{0,5745 + 0,8732}{2} = 0,7238$$

$$AUC = \frac{Recall + Specificity}{JumlahKelas} = \frac{0,7172 + 0,7238}{2} = 0,7210$$

Selanjutnya menampilkan hasil nilai AUC (Area Under Curve) dengan model algoritma decision tree aplikasi Rapidminer.



Gambar 11 AUC (Optimistic)



Gambar 12 AUC (Pessimistic)



Gambar 13 AUC (Area Under Curve)

C. Hasil Analisis Decision Tree Menggunakan Algoritma C4.5

Berdasarkan hasil perhitungan nilai entropy total, entropy dan gain pada setiap atribut ditemukan urutan klasifikasi minat konsumen yang dapat mempengaruhi penjualan pakaian dilihat dari hasil decision tree yakni atribut rating. Berdasarkan nilai gain yang didapatkan dari semua atribut, nilai gain tertinggi adalah atribut toko/brand akan tetapi dataset dengan nilai weight/bobot tertinggi dari setiap atribut ialah atribut harga dengan weight atau bobot 0,667., dengan urutan klasifikasi berdasarkan nilai weight atau bobot minat konsumen ialah toko/brand, rating, bahan dan jenis pakaian dengan nilai weight atau bobot masing-masing adalah 0,125., 0,110., 0,098. Maka pada pemodelan decision tree akar pertama pada setiap toko ialah atribut rating yang kemudian minat konsumen pada setiap toko/brand yaitu berdasarkan atribut harga.

Tingkat daya minat konsumen dalam membeli suatu pakaian pada setiap toko dilihat dari rating dan harga pakaian. Jika minat konsumen dilihat dari rating dan harga pakaian maka toko/brand tentu harus lebih memperhatikan mengenai harga dari suatu pakaian yang dijual, karena hal itu mempengaruhi minat konsumen dalam membeli suatu pakaian. Apabila minat konsumen dalam membeli suatu pakaian meningkat maka rating penjualan juga akan semakin meningkat sehingga toko/brand bisa mendapatkan banyak feedback baik dari konsumen dan hal ini tentu juga akan meningkatkan larisnya penjualan pakaian.

Urutan klasifikasi atribut untuk melihat minat konsumen dalam membeli suatu pakaian pada setiap toko/brand ialah berdasarkan rating penjualan, harga pakaian, toko/brand disusul dengan bahan dan jenis pakaian. Dari hasil ini toko/brand akan terbantu dalam mengetahui minat konsumen membeli suatu pakaian dengan melihat hasil decision tree dengan menggunakan algoritma C4.5. Hasil tersebut juga sesuai dengan penelitian [5] yaitu penggunaan decision tree dapat membantu penjual dalam pemenuhan kebutuhan dan kepuasan dari konsumen, hasil decision tree ini dapat mengetahui faktor yang paling mempengaruhi daya tarik konsumen terhadap suatu produk. Oleh karena itu, proses jual beli dapat dilakukan tanpa kecurangan, karena pelaku usaha

dapat melakukan penjualan dengan angka yang tinggi dengan mengetahui apa yang menjadi daya tarik dari konsumen.

D. Hasil Analisis Akurasi Algoritma C4.5 Menggunakan Confusion Matrix

Pengujian akurasi dilakukan untuk mengetahui kinerja metode algoritma C4.5 dalam melakukan klasifikasi terhadap 2 kelas yang telah ditentukan. Pada pengujian ini data yang berjumlah 1471 data dibagi menjadi 70% dan 30%. Dari Gambar 4.16 hasil keseluruhan nilai akurasi data nilainya mencapai 80,95%, hasil capaian nilai precision rendah sebesar 57,45% dan nilai precision tinggi sebesar 87,32% serta nilai recall rendah 55,10% dan nilai recall tinggi 88,34%.

Hasil nilai AUC (Area Under Curve) pada aplikasi rapidminer ialah 0,721 dan dikategorikan ke dalam Klasifikasi Cukup. Kurva Receiver Operating Characteristic (ROC) digunakan untuk mengekspresikan data confusion matrix. Garis horizontal mewakili nilai False Tinggi (FT) dan garis vertikal mewakili True Tinggi (TP). Nilai AUC (Area Under Curve) model algoritma decision tree adalah 0,721 hal ini menunjukkan bahwa model algoritma decision tree berdasarkan [4] mencapai klasifikasi Cukup. Hasil tersebut menunjukkan bahwa Algoritma C4.5 dengan model decision tree menggunakan aplikasi rapidminer dapat digunakan untuk mengetahui klasifikasi dan urutan atribut suatu penjualan berdasarkan minat pembelian konsumen

IV. KESIMPULAN

Dari decision tree tingkat daya minat konsumen dalam membeli suatu pakaian dilihat dari rating dan harga pakaian, jika minat konsumen dilihat dari rating harga pakaian maka toko/brand tentu harus lebih memperhatikan mengenai harga dari suatu pakaian yang dijual, karena hal itu mempengaruhi minat konsumen dalam membeli suatu pakaian dan rating penjualan juga akan semakin meningkat sehingga toko/brand bisa mendapatkan banyak feedback baik dari konsumen dan hal ini tentu juga akan meningkatkan larisnya penjualan pakaian.

Hasil pengujian tingkat akurasi Algoritma C4.5 menggunakan aplikasi Rapidminer dalam model decision tree diperoleh hasil accuracy sebesar 80,95% menggunakan perhitungan confusion matrix. Sedangkan klasifikasi pengujian berdasarkan nilai AUC (Area Under Curve) masuk kedalam kategori klasifikasi Cukup. Hasil tersebut menunjukkan bahwa Algoritma C4.5 dengan model decision tree menggunakan aplikasi rapidminer dapat digunakan untuk mengetahui klasifikasi dan urutan atribut suatu penjualan berdasarkan minat pembelian konsumen

V. SARAN

Penelitian ini dapat dikembangkan dengan menggabungkan metode lain sehingga didapatkan hasil prediksi dari data. Jumlah data dalam hal ini atribut toko/brand ditambah, sehingga dapat diperoleh hasil akurasi fungsi algoritma yang lebih baik. Dapat mengembangkan penelitian ini dengan menambahkan jumlah kelas pada label

REFERENSI

- [1] Azwanti N. (2018). “Analisa Algoritma C4.5 Untuk Memprediksi Penjualan Motor Pada Pt. Capella Dinamik Nusantara Cabang Muka Kuning” *Jurnal Ilmiah Ilmu Komputer*. 13(1), 33-38.
<http://E-Journals.Unmul.Ac.Id/Index.Php/JIM/Article/View/629>
- [2] Bramer M. (2007). “*Principles of Data Mining*” Springer. London, UK.
https://www.researchgate.net/publication/220688376_Principles_of_Data_Mining
- [3] Lakshmi B. N, Dr. Indumathi. T. S Dan Dr. Ravi N. (2016). “A Study On C.5 Decision Tree Classification Algorithm For Risk Predictions During Pregnancy” *Procedia Technology* 24.
<https://www.sciencedirect.com/science/article/pii/S22121017316302171>
- [4] Gorunescu F. (2011). “*Data Mining Concept, Models and Techniques*” Springer. Verlag Berlin Heidelberg.
https://tktung.files.wordpress.com/2015/12/florin_gorunescu_data_mining_concepts_models.pdf
- [5] Husna F. (2021). “*Implementasi Data Mining Menggunakan Algoritma C4.5 Pada Klasifikasi Penjualan Hijab*” Program Studi Matematika Fakultas Sains Dan Teknologi Universitas Islam Negeri Maulana Malik Ibrahim Malang.
<http://etheses.uin-malang.ac.id/32882/1/17610069.pdf>
- [6] Mienye I. D, Sun Y Dan Wang Z. (2019). “Prediction Performance Of Improved Decision Tree-Based Algorithms: A Review” *Procedia Manufacturing* 35 .
https://www.researchgate.net/publication/335186630_Prediction_performance_of_improved_decision_tree-based_algorithms_a_review
- [7] Pramadhani A. E & Setiadi T. (2014). “Penerapan Data Mining Untuk Klasifikasi Prediksi Penyakit Ispa (Infeksi Saluran Pernapasan Akut) Dengan Algoritma Decision Tree (Id3)” *Jurnal Sarjana Teknik Informatika*, 2(1), 831-839.
<https://Docplayer.Info/30460448-Penerapan-Data-Mining-Untuk-Klasifikasi-Prediksi-Penyakit-Ispa-Infeksi-Saluran-Pernapasan-Akut-Dengan-Algoritma-Decision-Tree-Id3.Html>
- [8] Rahayu S. (2020). “*Pengaruh Trend Fashion Dan Pergaulan Terhadap Gaya Hidup Konsumtif Pada Mahasiswi Fakultas Ekonomi Dan Bisnis Islam UIN Sulthan Thaha Saifuddin Jambi*” Program Studi Ekonomi Syariah Fakultas Ekonomi Dan Bisnis Islam Universitas Islam Negeri Sulthan Thaha Saifuddin Jambi.
<http://repository.uinjambi.ac.id/4827/>
- [9] Rahman, M. F., Darmawidjadja, D., & Alamsah, M. I. (2017). “*Klasifikasi Untuk Diagnosa Diabetes Menggunakan Metode Bayesian Regularization Neural Network (Rbnn)*” *Jurnal Informatika*, 11(1).
http://journal.uad.ac.id/index.php/JIFO/article/view/5452/pdf_2
- [10] Rogers, S & Girolami, M. (2012). “*A First Course In Machine Learning*. CRC Press Taylor & Francis Group” London, New York.
<http://repo.darmajaya.ac.id/3793/1/A%20First%20Course%20in%20Artificial%20Intelligence%20%28%20PDF%20Drive%20%29.pdf>