Jurnal Riset Mahasiswa Matematika
Volume 2(2) (2022), Pages 40-46
p-ISSN: 2808-1552; e-ISSN: 2808-4926

Implementasi Data mining Menggunakan Algoritma C4.5 pada Klasifikasi Penjualan Hijab

Faridatul Husna*, Hairur Rahman, Juhari

Program Studi Matematika, Fakultas Sains dan Teknologi, Universitas Islam Negeri Maulana Malik Ibrahim Malang, Indonesia

frdtlna02@gmail.com*, hairur@mat.uin-malang.ac.id, juhari@uin-malang.ac.id

Abstrak

Indonesia dikenal sebagai negara dengan mayoritas penduduk beragama Islam, ini membuat kebutuhan sandang di Indonesia juga harus memperhatikan bagaimana kriteria pakaian umat Islam, salah satunya yaitu hijab. Perkembangan bisnis dalam dunia *fashion* khususnya hijab telah menjadi *trend setter* saat ini sehingga banyaknya data dalam dunia bisnis *fashion* menciptakan kondisi di mana terdapat pelaku bisnis memiliki banyak data namun minim informasi dari data tersebut. Untuk menghadapi kondisi tersebut, perlu dilakukan klasifikasi pada data. Klasifikasi merupakan suatu proses untuk menemukan properti-properti yang sama dalam suatu himpunan data untuk diklasifikasikan ke dalam kelas-kelas yang berbeda. Salah satu metode klasifikasi adalah *Decision tree* menggunakan Algoritma C4.5. penelitian ini bertujuan untuk mengetahui model dan keakuratan algoritma dalam melakukan klasifikasi terhadap penjualan hijab dari beberapa *brand* hijab. Diperoleh model *Decision tree* menggunakan Algoritma C4.5 dengan akar pertama adalah atribut Harga di mana yang menjadi akar pertama adalah atribut yang paling mempengaruhi penjualan hijab. Hasil perhitungan nilai akurasi adalah 87% sehingga model *Decision tree* dan proses klasifikasi menggunakan Algoritma C4.5 tergolong baik. Penelitian ini diharapkan dapat membantu pelaku bisnis dalam bidang *fashion* terutama hijab untuk mengetahui faktor-faktor yang mempengaruhi daya minat konsumen terhadap suatu produk hijab.

Kata kunci: akurasi; algoritma c4.5; *decision tree*; klasifikasi.

Abstract

Indonesia is known as a country with a majority Muslim population, this makes the need for clothing in Indonesia must also pay attention to the criteria for Muslim clothing, one of which is the hijab. Business developments in the fashion world, especially hijab, have become a trend setter at this time so that the large amount of data in the fashion business world creates conditions where there are businesspeople who have a lot of data but lack of information from that data. To deal with these conditions, it is necessary to classify the data. A classification is a process to find the same properties in a data set to be classified into different classes. One of the classification methods is the Decision tree using the C4.5 Algorithm. This research aims to determine the model and the accuracy of the C4.5 algorithm in classifying hijab sales from several hijab brands. The Decision tree model is obtained using the C4.5 algorithm with the first root being the price attribute, where the first root is the attribute that most affected the sale of the hijab. The result of calculating the accuracy value is 87% so that the Decision tree model and the classification process using the C4.5 Algorithm are classified as good. This research is expected to help businesspeople in the fashion sector, especially hijab, to find out the factors that influence consumer interest in a hijab product.

Keywords: accuracy; c4.5 algorithm; classification; decision tree.

Submitted: 30 Desember 2021 Reviewed: 26 September 2022 Accepted: 28 Desember 2022
Doi: http://dx.doi.org/10.18860/jrmm.v2i2.14891

PENDAHULUAN

Indonesia mempunyai 16 subsektor kreatif yang menjadi fokus pengembangan, salah satunya adalah *fashion*. Pada tahun 2020, *fashion* menempati urutan pertama dalam memberikan kontribusi untuk ekonomi kreatif di Indonesia[1]. Sebagai negara yang dikenal dengan mayoritas penduduk beragama Islam, tentu membuat kebutuhan sandang di Indonesia harus memperhatikan kriteria pakaian bagi umat muslim salah satunya yaitu hijab. Saat ini terdapat puluhan juta penduduk Indonesia yang telah memakai hijab, hal ini selaras dengan semakin berkembangnya industri pakaian muslim khususnya hijab.

Tingginya angka pengguna hijab, mengakibatkan terciptanya kondisi di mana sekarang ini terdapat banyak data namun minim informasi, maka diperlukan proses klasifikasi penjualan hijab untuk mengolah data tersebut agar didapatkan informasinya. Informasi dari data dapat ditemukan *data mining*. Salah satu teknik *data mining* yaitu klasifikasi sesuai untuk diterapkan pada kasus yang akan digunakan pada penelitian. Klasifikasi dilakukan dengan menggunakan data *set* yang dibandingkan untuk mengembangkan model yang mampu mengklasifikasikan seluruh data yang ada [2]. Terdapat banyak model klasifikasi salah satunya adalah *decision tree* menggunakan algoritma C4.5. Penelitian ini dilakukan menggunakan algoritma C4.5 untuk mengklasifikasikan data penjualan hijab dari beberapa *brand* hijab dan menggunakan beberapa atribut.

Decision tree

Decision tree digunakan untuk mengeksplorasi data, mencari keterkaitan tersembunyi antara sejumlah variabel input dengan variabel target. *Decisison tree* merupakan struktur yang dapat digunakan untuk membagi fakta yang besar menjadi pohon keputusan yang mempresentasikan aturan, kemudian aturan tersebut dapat dengan mudah untuk diinterpretasi oleh manusia[3].

Banyak variasi *model decision tree* dengan tingkat kemampuan dan syarat yang berbeda, Namun pada umumnya ciri kasus yang dapat diterapkan pada decision tree adalah [4]:

1. *Data/example* dinyatakan dengan pasangan atribut dan nilainya.
2. *Label/output* data bernilai diskrit.
3. Data mempunyai *missing value*.

Klasifikasi

Klasifikasi didefinisikan sebagai *supervise learning* yang membutuhkan label dalam prosesnya untuk mengekstrak model yang digunakan untuk memprediksi suatu label [5]. Proses dalam klasifikasi yaitu untuk menemukan properti-properti yang sama dalam suatu himpunan obyek pada suatu database kemudian diklasifikasikan ke dalam kelas-kelas yang berbeda sesuai dengan model klasifikasi yang dipilih. Proses klasifikasi bertujuan untuk mencari model dari training set yang memisahkan atribut ke dalam kategori atau kelas yang sesuai, kemudian model tersebut digunakan untuk klasifikasi atribut yang kelasnya belum diketahui sebelumnya [6].

Algoritma C4.5

Salah satu metode klasifikasi yang dipakai dalam penelitian ini adalah klasifikasi *decision tree* menggunakan Algoritma C4.5. *Decision tree* dibangun untuk mengklasifikasikan data yang dirancang sebagai input pada algoritma yang terdiri dari beberapa objek dan atribut. Proses pembuatan *decision tree* yang pertama adalah mengentropi masing-masing kelas dan atribut, kemudian penghitungan perolehan informasi[7]. Informasi yang didapatkan dari pohon keputusan akan lebih mudah dipahami karena atribut dengan nilai gain tertinggi akan menjadi akar pohon yang kemudian diikuti cabang-cabangnya.

Proses untuk membuat sebuah pohon keputusan harus melalui beberapa tahapan berikut ini:

1. Menyiapkan *data training*
2. Menentukan akar dari pohon dengan menghitung nilai *Entropy* dan *gain* terlebih dahulu. *Entropy* adalah nilai informasi yang menyatakan ukuran ketidakpastian *(impurity)* atribut

dari sekumpulan obyek data [8]. Untuk menghitung nilai *Entropy*, digunakan persamaan berikut [9]:

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

Keterangan:
s = himpunan kasus
n = jumlah partisi
p_i = proporsi kelas i dalam data yang diproses pada node s. Nilai p_i didapat dengan membagikan jumlah baris data sebagai label kelas dengan jumlah baris semua data.

Setelah menghitung nilai *Entropy,* kemudian menghitung nilai *gain* dari setiap atribut. *Gain* merupakan tingkat pengaruh suatu atribut terhadap keputusan atau ukuran efektifitas suatu variabel dalam mengklasifikasikan data [8].Untuk menghitung nilai *Entropy*, digunakan persamaan berikut [9]:

$$Gain(s, A) = Entropy(s) - \sum_{i=1}^{n} \frac{|s_i|}{|s|} \cdot Entropy(s_i) \quad (2)$$

Keterangan:
s = himpunan kasus
A = nilai yang mungkin dari atribut A
n = jumlah partisi atribut A
$|s_i|$ = subset dari s di mana A mempunyai nilai i
$|s|$ = himpunan kasus
Setelah menemukan nilai *gain* tertinggi dari suatu atribut, maka atribt tersebut yang digunakan sebagai akar pertama. Dari atribut yang tertinggi, kemudian dihitung nilai *Entropy* dan nilai *gain* dari kasus dengan nilai pada atribut yang digunakan sebagai akar pertama untuk dijadikan akar simpul pertama.

3. Pemberhentian proses partisi akar pohon keputusan jika:
   a. Seluruh *record* pada simpul *N* mendapatkan kelas yang sama.
   b. Tidak ada atribut yang dipartisi lagi.
   c. Tidak ada *record* di dalam cabang yang kosong.

Confusion matrix

Confusion matrix adalah metode yang digunakan untuk melakukan perhitungan akurasi pada konsep *data mining* [10]. *Confusion matrix* digambarkan dengan tabel yang menyatakan jumlah data uji yang benar dan jumlah data uji yang salah. *Confussion matrix* terdiri atas data yang memiliki 2 kelas bersifat positif atau negatif. Tabel *confusion matrix* tersusun atas 4 sel yaitu *true positif, false positif, true negative,* dan *false negative* [10].

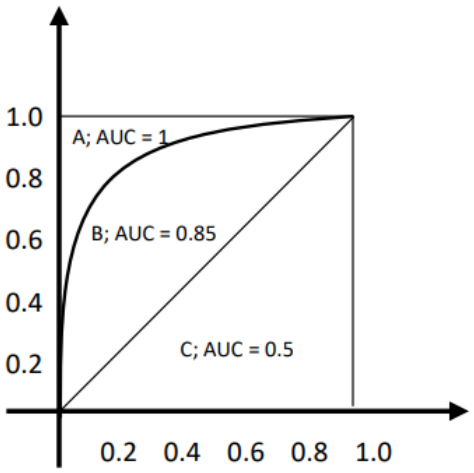
Nilai yang dihasilkan melalui metode *confusion matrix* adalah berupa evaluasi akurasi sebagai berikut [11].

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} 100\% \quad (3)$$

Keterangan:
- *True Positive* (TP) adalah jumlah *record* data positif yang diklasifikasikan sebagai nilai positif.
- *False Positive* (FP) adalah jumlah *record* data negatif yang diklasifikasikan sebagai nilai positif.
- *False Negative* (FN) adalah jumlah *record* data positif yang diklasifikasikan sebagai nilai negatif.
- *True Negative* (TN) adalah jumlah *record* data negatif yang diklasifikasikan sebagai nilai positif.

Setelah mendapatkan nilai *accuracy*, maka dapat diketahui klasifikasi tersebut berada pada kelompok klasifikasi baik atau buruk. Pada akurasi klasifikasi data mining, nilai *Area Under Curve (AUC)* dapat dibagi menjadi beberapa kelompok yaitu [12]:

1. 90% - 100% = Klasifikasi sangat baik
2. 80% - 90% = Klasifikasi baik
3. 70% - 80% = Klasifikasi cukup
4. 60% - 70% = Klasifikasi buruk
5. 50% - 60% = Klasifikasi salah



Gambar 1. Kurva AUC

K-fold cross validation

Rata-rata keberhasilan suatu sistem dapat diketahui dengan melakukan *cross validation,* yaitu dengan melakukan iterasi dengan mengacak atribut *input* hingga sistem tersebut dapat diujikan untuk atribut *input* yang lain [13]. Pada proses *cross validation*, data akan dikelompokkan dalam *k* buah partisi dengan ukuran yang sama, selanjutnya proses *testing* dan *training* dilakukan sebanyak *k* kali. Pada perulangan ke-*i* partisi akan menjadi data *testing* dan sisanya akan menjadi data *training*.

METODE

Tahapan-tahapan penelitian dilakukan sebagai berikut:

1. Langkah pertama pada penelitian ini adalah menghitung nilai *Entropy* total dari seluruh data menggunakan persamaan (1).
2. Kemudian menghitung nilai *Entropy* dari setiap nilai pada atibut menggunakan persamaan (1).
3. Setelah nilai *Entropy* dihitung, maka hitung nilai *gain* dari setiap atribut menggunakan persamaan (2).
4. Membuat akar *decision tree* menggunakan atribut yang memiliki nilai *gain* tertinggi.
5. Atribut yang memiliki nilai *gain* tertinggi kemudian dihitung nilai *Entropy* totalnya. Misalnya pada atribut harga, terdapat dua nilai yaitu harga murah dan harga mahal. Maka dihitung nilai *Entropy* total untuk kasus harga murah dan harga mahal menggunakan persamaan (2). Setelah itu dihitung nilai *gain* dari setiap atribut dengan kasus harga murah dan harga mahal menggunakan persamaan (2). Maka akan didapatkan nilai *gain* tertinggi dari kasus harga murah dan harga mahal dari perhitungan nilai *gain* tersebut yang kemudian dijadikan sebagai akar simpul pertama pada *decision tree*.
6. Melakukan proses pada poin ke-5 sampai semua atribut masuk pada *decision tree*.
7. Setelah *decision tree* terbentuk, maka dilakukan evaluasi akurasi untuk mengetahui keakuratan algoritma C4.5 menggunakan *confusion matrix* dan rumus *accuracy* pada persamaan (3).

8. Menguji *k-fold cross validation* untuk mengetahui akurasi dari keseluruhan proses *data mining*. Pengujian *k-fold cross validation* dilakukan dengan membagi *data testing* menjadi beberapa bagian.
9. Analisis hasil dari *decision tree* yang sudah terbentuk dan uji akurasi pada *confusion matrix.*

HASIL DAN PEMBAHASAN

Analisis Data

Data yang digunakan berjumlah 104 data. Kemudian dilakukan pembersihan data dan transformasi data untuk diproses pada pemrograman Google Colab menggunakan bahasa Python.

Tabel 1. Transformasi Data

Brand	Jenis	Bahan	Harga	Bintang	Terjual
0	0	0	93.390	4,9	Rendah
1	0	3	104.800	4,9	Rendah
2	3	0	159.000	5	Rendah
…	…	…	…	…	…
12	1	1	299.000	4,9	Rendah
12	1	1	195.000	5	Rendah

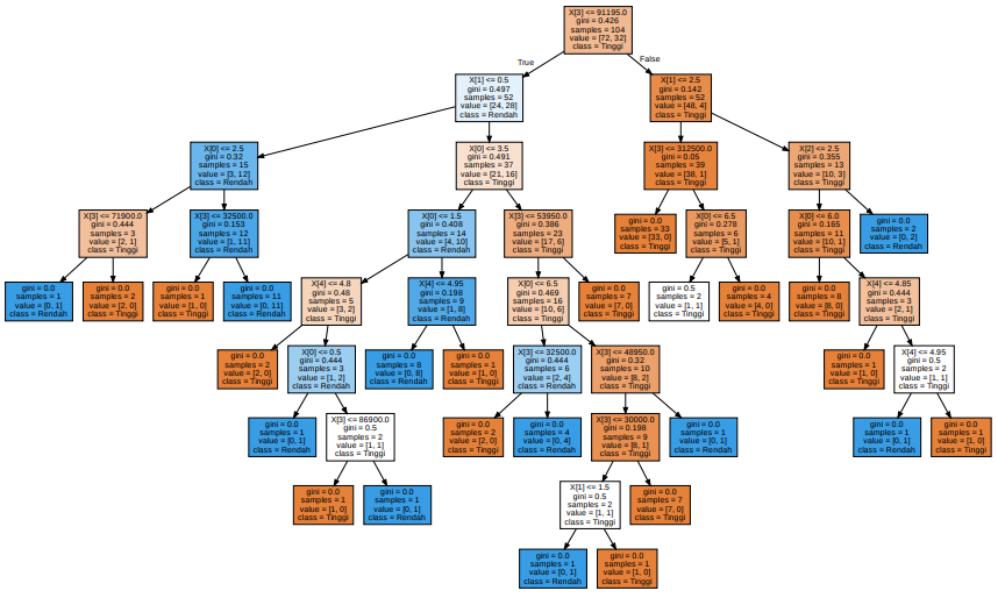
Setelah melakukan transformasi data, dihitung nilai *Entropy* dan nilai *gain* dari setiap atribut menggunakan persamaan (1) dan (2).

Tabel 2. Hasil Penghitungan Nilai *Entropy* dan *Gain*

Atribut	Nilai	Jumlah Kasus	Kelas Tinggi	Kelas Rendah	*Entropy*	*Gain*
Total					0,878927	
Brand						0,238
	Rabbani	12	2	10	0,650022	
	Zoya	8	1	7	0,543564	
	Nafisa	8	4	4	1	
	Elzatta	11	8	3	0,845351	
	Nadiraa	6	3	4	0,985228	
	Zealcofa	9	5	4	0,991076	
	Kami	8	1	7	0,543564	
	Shafira	6	0	6	0	
	Umama	7	2	6	0,811278	
	WMD	7	4	3	0,985228	
	Alsyahra	5	2	3	0,721928	
	Meccanism	7	0	7	0	
	RM	8	0	8	0	
Jenis						0,004
	Instan	25	12	13	0,987693	
	Segi Empat	37	10	27	0,841852	
	Pashmina	23	4	19	0,666587	
	Instan Syari	18	5	13	0,852405	
Bahan						0,025
	Polyester	23	7	16	0,886541	
	Cerutty	18	4	14	0,764205	
	Crepe	20	10	10	1	
	Voal	43	11	32	0,82036	
Harga						0,221

	≤ 91.195	52	28	24	0,99572	
	> 91.195	52	3	49	0,31821	
Bintang						0,568
	4,6	3	1	2	0,91829	
	4,7	8	3	5	0,95443	
	4,8	22	12	10	0,99403	
	4,9	47	14	33	0,87867	
	5	24	1	23	0,24988	

Berikut adalah model *decision tree* yang telah dibangun dengan memproses data pada pemrograman Google Colab:



Gambar 2. *Decision tree*

Uji Akurasi

Pengujian dilakukan untuk mengetahui kinerja metode algoritma C4.5 dalam melakukan klasifikasi terhadap kelas yang telah ditentukan. Pada pengujian ini, data yang berjumlah 104 data dibagi menjadi 70% data *training* dan 30% data *testing*.

Tabel 3. *Confusion matrix*

Kelas Observasi	Kelas Prediksi	
Kelas Sebenarnya	Kelas = Tinggi	Kelas = Rendah
Kelas = Tinggi	TP = 20	FN = 0
Kelas = Rendah	FP = 4	TN = 8

Setelah melakukan pengujian *confusion matrix*, data *testing* dilakukan uji akurasi, *precision*, dan *recall* pada Google Colab.

Tabel 4. Hasil Uji Akurasi

Uji	*Accuracy*	*Precision*	*Recall*
Tingkat Akurasi	87%	83%	100%

KESIMPULAN

Dari *decision tree* yang terbentuk, informasi yang didapatkan bahwa faktor utama yang mempengaruhi penjualan hijab adalah harga. Hasil pengujian tingkat akurasi sebesar 87% berdasarkan nilai *Area Under Curve* (AUC), klasifikasi algoritma C4.5 tergolong baik.

DAFTAR PUSTAKA

[1] Yasyi, Dini N. "Tahun 2020, Sektor Ekonomi Kreatif Akan Sumbang Rp.1.100 Triliun ke PDB Indonesia", https://www.goodnewsfromindonesia.id/2020/08/31/tahun2020-sektor-ekonomi-kreatif-akan-sumbang-rp1-100-triliun-ke-pdb-indonesia, diakses pada 2 Juni 2021 pukul 14.02.

[2] Ramageri, Bharati M. 2010.Data mining Techniques and Aplications. Indian Journal of Computer Science and Engineering. 1(4).

[3] Micahel J. A. Berry, G. L. 2004. Data mining Techniques For Marketing, Sales, and Customer Relationship Management. Wiley Publishing..

[4] Santosa, B. 2007. Data mining: Teknik Pemanfaatan Data untuk Keperluan Bisnis (1st ed.). Graha Ilmu

[5] Han, J, Kamber, M, & Pei, J. 2012. Data mining: Concept and Techniques, Third Edition. Waltham: Morgan Kaufmann Publishers.

[6] Azwanti, N. 2018. Analisa Algoritma C4.5 untuk Memprediksi Penjualan Motor pada PT. Capella Dinamik Nusantara Cabang Muka Kuning. Informatika Mulawarman : Jurnal Ilmiah Ilmu Komputer, 13(1), 6. http://ejournals.unmul.ac.id/index.php/JIM/article/view/629

[7] Mienye, I. D., Sun, Y., & Wang, Z. 2019. Prediction performance of improved decision tree-based algorithms: A review. Procedia Manufacturing, 35, 698–703. https://doi.org/10.1016/j.promfg.2019.06.011

[8] Lakshmi, B. N., Indumathi, T. S., & Ravi, N. (2016). A Study on C.5 Decision tree Classification Algorithm for Risk Predictions During Pregnancy. Procedia Technology, 24, 1542–1549. https://doi.org/10.1016/j.protcy.2016.05.128

[9] Kusrini & Luthfi, Emha T. 2009. Algoritma Data mining. ANDI OFFSET.

[10] Rogers, Simon & Girolami, Mark. 2012. A First Course in Machine Learning. CRC Press Taylor & Francis Group.

[11] Rahman, M. F., Alamsah, D., Darmawidjadja, M. I., & Nurma, I. 2017. Klasifikasi Untuk Diagnosa Diabetes Menggunakan Metode Bayesian Regularization Neural Network (RBNN). Jurnal Informatika, 11(1), 36. https://doi.org/10.26555/jifo.v11i1.a5452

[12] Gorunescu, Florin. 2011. Data mining: Concepts, Models, and Techniques. Verlag Berlin Heidelberg: Springer .

[13] Hartama, A. A. K. 2017. Klasifikasi Penyakit Hipertensi Menggunakan Algoritma C4.5 Studi Kasus RSU Provinsi NTB. 1–154.