

## PERBANDINGAN ALGORITMA NAÏVE BAYES DAN C4.5 DALAM MENENTUKAN TINGKAT PENJUALAN MOTOR HONDA

Donni Prabowo<sup>1</sup>, Firman Hidayat<sup>2</sup>, Gagah Gumelar<sup>3</sup>, Dewa Qintoro<sup>4</sup>, Aji Setiawan<sup>5</sup>

<sup>1,2,3,4,5</sup>Program Studi Informatika, Universitas Amikom Yogyakarta

E-mail : <sup>1</sup>[donniprabowo@amikom.ac.id](mailto:donniprabowo@amikom.ac.id), <sup>2</sup>[firman.hidayat@students.amikom.ac.id](mailto:firman.hidayat@students.amikom.ac.id),  
<sup>3</sup>[gagah.gumelar@students.amikom.ac.id](mailto:gagah.gumelar@students.amikom.ac.id), <sup>4</sup>[dewaqintoro@students.amikom.ac.id](mailto:dewaqintoro@students.amikom.ac.id),  
<sup>5</sup>[aji.setiawan@students.amikom.ac.id](mailto:aji.setiawan@students.amikom.ac.id)

### Abstract

*At this time, the mobility of the community activities will be very high. The mobility of the high society will have an impact to the transportation needs are increasing. Coupled with the opportunities of the community towards taxi online, will have an impact on the number of sales of motorcycles in Indonesia. In this research we will classify the sales data of the motor with rapidminer software by applying two algorithm: Algorithm C45, and Naive Bayes. The second algorithm is a method of classification based on statistics and probability. Two of the above algorithms compared the rate of speed in the best-selling motorcycle classification of data, to get the decision tree the best selling motorcycle was purchased by the society, and should be in more production. Research using Algorithm C45 motor with the lowest prices is the best selling, than Algorithm Naive Bayes where the best selling honda motor is a type of Scooter, medium of CC, and low price.*

**Keywords:** honda best-selling motorcycle, naive bayes algorithm, classification, algorithm C45.

### PENDAHULUAN

Honda Motor merupakan salah satu perusahaan motor terbesar di dunia. Di Indonesia, Honda memiliki perusahaan yang bernama PT. Astra Honda Motor (AHM) yang bergerak di bidang manufaktur, perakitan, dan distributor sepeda motor merk Honda. Perusahaan ini merupakan satu-satunya perusahaan yang memiliki hak sebagai Agen Tunggal Pemegang Merek (ATPM) sepeda motor Honda. Penjualan motor di Indonesia khususnya untuk motor merk Honda sendiri mencapai ratusan ribu unit per tahunnya, tetapi tidak jarang Honda merilis jenis sepeda motor baru dengan spesifikasi tertentu tetapi kurang diminati di pasar Indonesia. Penelitian kali ini dilakukan untuk mengetahui atribut atau parameter yang menentukan tingkat kelarisan motor dengan menggunakan data mining.

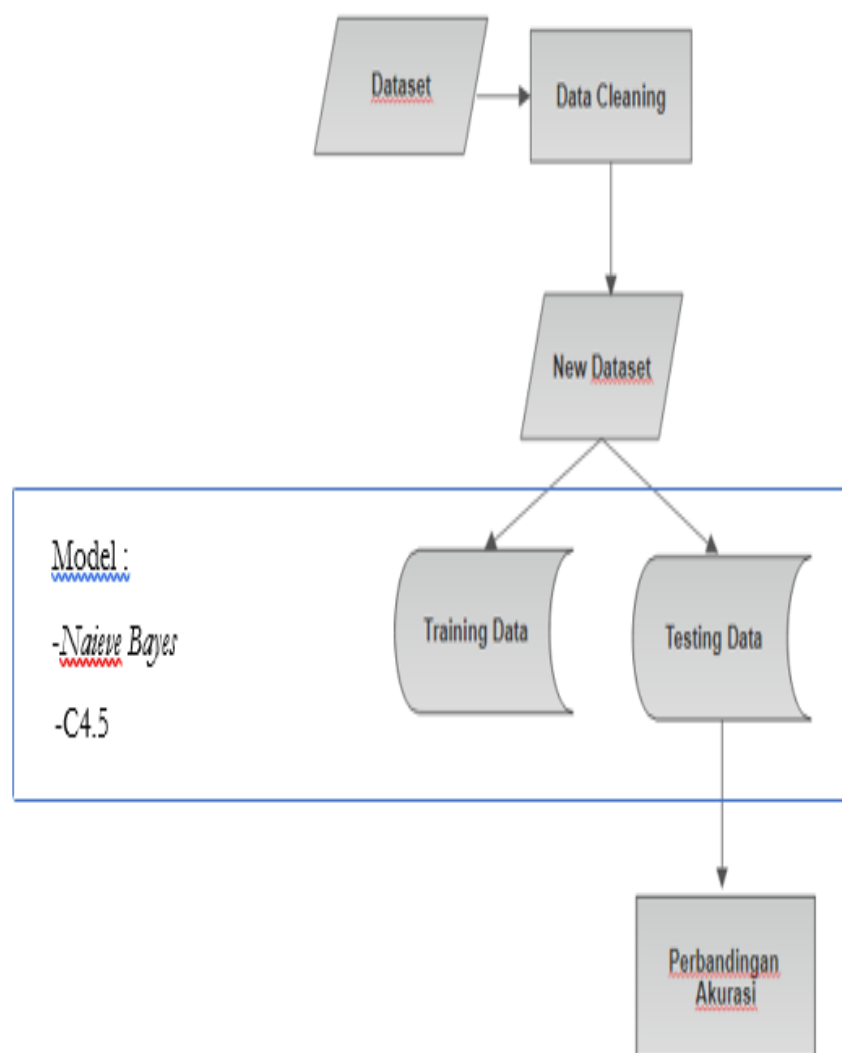
Data mining telah banyak diaplikasikan dalam berbagai bidang, seperti bidang bisnis, kedokteran, pendidikan, dan lain-lain. Dalam bidang bisnis, Teknik data mining biasanya digunakan untuk beberapa kasus seperti menargetkan penjualan, manajemen pekerjaan, menilai loyalitas customer, informasi summary, dan banyak kasus lainnya. Data mining sendiri merupakan sebuah teknik untuk menggali nilai tambah berupa informasi dengan melakukan penggalian pola pola dari data dengan tujuan untuk memanipulasi data menjadi informasi yang lebih berharga yang selama ini tidak diketahui secara manual dari suatu basis data atau sekumpulan data. Beberapa fungsi dari data mining yaitu prediksi, deskripsi, klasifikasi, dan asosiasi.

Data Mining memiliki banyak sekali metode/algoritma yang dapat digunakan untuk penentuan berbagai macam pola, prediksi, klasifikasi dll, sesuai dengan kebutuhan peneliti. Dalam kasus ini digunakan dua metode yang akan dibandingkan untuk

memprediksi tinggi rendahnya penjualan dari sebuah motor honda. Penelitian sebelumnya yang membahas tentang kinerja algoritma C4.5 menggunakan perangkat lunak *WEKA* mengatakan pengujian dengan *WEKA* lebih efektif dan fleksibel karena pohon keputusan yang terbentuk hampir sama dengan perhitungan yang dilakukan secara manual[1]. Dalam penelitian ini peneliti mencoba membandingkan 2 algoritma dengan menggunakan Perangkat lunak lain yang tentunya dikhususkan untuk mining data.

#### METODE PENELITIAN

Penelitian ini termasuk jenis penelitian eksperimen dengan menguji dua buah algoritma yaitu *Naïve Bayes* dan C4.5 dalam menentukan tingkat penjualan motor honda menggunakan dataset penjualan motor honda tahun 2017. Pengujian dilakukan dengan menggunakan perangkat lunak *RapidMiner 9.0* yang dikhususkan untuk melakukan olah data. Adapun alur penelitian yang digunakan secara umum adalah sebagai berikut :



Gambar 1. Alur Penelitian

Dimulai dari menentukan dataset, cleaning dataset hingga uji menggunakan 2 algoritma untuk membandingkan tingkat akurasi keduanya.

#### A. Penentuan Dataset

Data yang digunakan pada penelitian ini adalah data penjualan motor Honda pada tahun 2017 yang selanjutnya di proses dan diklasifikasikan berdasarkan jenis, cc, dan harga menjadi dataset baru yang siap uji. Dataset baru disimpan dalam format excel (\*.xls) sesuai dengan format yang akan dijadikan inputan pada aplikasi *rapid miner*, sebagai berikut :

Table 1. Training Data

JENIS	*CC	**Harga	***Laris
Skutik	Sedang	Murah	ya
Sport	Sedang	Sedang	tidak
Skutik	Sedang	Sedang	ya
Bebek	Kecil	Murah	ya
Sport	Sedang	Mahal	ya
Sport	Besar	Mahal	tidak
Skutik	Sedang	Sedang	ya
Sport	Besar	Mahal	tidak
Bebek	Sedang	Sedang	ya
Sport	Besar	Mahal	tidak
Bebek	Sedang	Mahal	tidak
Bebek	Sedang	Mahal	tidak
Sport	Sedang	Mahal	tidak
Skutik	Sedang	Mahal	tidak

\*CC : Kecil < 125; Sedang < 150; Besar > 150.

\*\*Harga : Murah < 15jt; Sedang < 22jt; Mahal > 22jt.

Tabel 2. Testing Data

JENIS	CC	Harga	Laris
Skutik	Sedang	Murah	ya
Skutik	Sedang	Sedang	ya
Skutik	Sedang	Mahal	tidak
Bebek	Kecil	Murah	ya
Bebek	Sedang	Sedang	ya
Bebek	Sedang	Mahal	tidak
Sport	Sedang	Mahal	ya
Sport	Sedang	Mahal	tidak
Sport	Besar	Mahal	tidak

## B. Algoritma yang Digunakan

Pengujian dilakukan dengan menggunakan 2 metode yaitu naïve bayes dan C4.5, kedua metode/ algoritma ini merupakan metode data mining dengan fungsi klasifikasi.

### 1. Algoritma *Naïve Bayes*

Naïve Bayes merupakan algoritma machine learning yang menggunakan perhitungan probabilistik yang berdasar pada teorema Bayes[2]. Pada penelitian ini algoritma *Naïve Bayes* akan di komparasikan dengan algoritma C4.5. Dengan memanfaatkan perhitungan probabilitas dan statistik, algoritma ini digunakan untuk memprediksi probabilitas dimasa depan berdasarkan pengalaman di masa sebelumnya. Keuntungan metode ini yaitu membutuhkan data penelitian yang kecil untuk menentukan estimasi parameter dalam proses pengklasifikasiannya.

Tahap dari proses algoritma Naïve Bayes adalah :

1. Menghitung jumlah kelas / label.
2. Menghitung jumlah kasus per kelas.
3. Mengalikan semua variable kelas.
4. Membandingkan hasil per kelas.

Berikut Persamaan teorema *Naïve Bayes* yang digunakan :

$$P(Y|X) = \frac{P(Y) \prod_{i=1}^q P(X_i|Y)}{P(X)} \quad (2)$$

Dimana :

$X$  = Data dengan kelas yang belum diketahui.

$Y$  = Hipotesis data yang merupakan suatu class spesifik

$P(Y|X)$  = Probabilitas hipotesis berdasarkan kondisi (posteriori probability)

$P(Y)$  = Probabilitas hipotesis (prior probability)

$P(X_i|Y)$  = Probabilitas Berdasarkan kondisi pada hipotesis

$P(X)$  = Probabilitas X

### 2. Algoritma C4.5

C4.5 adalah algoritma *statistical classifier* yang digunakan untuk membuat *decision tree* atau pohon keputusan. Entropy digunakan untuk memilih atribut yang paling optimal yang bisa membagi dataset secara efektif dalam beberapa kelas yang berbeda. Metode ini menghitung nilai entropy setiap atribut dan memilih atribut dengan nilai entropi tertinggi [3]. *Decision tree* dengan algoritma C4.5 merupakan salah satu metode klasifikasi yang menggunakan representasi struktur pohon (tree) dimana setiap node merepresentasikan nilai dari atribut, dan merepresentasikan kelas [3]. Algoritma ini merupakan salah satu algoritma induksi pohon keputusan yaitu ID3 yang dikembangkan oleh J. Ross Quinlan. Kelebihan dari metode ini yaitu mampu mengeliminasi perhitungan atau data yang sekiranya tidak diperlukan, daerah keputusan yang sebelumnya luas dapat diubah menjadi lebih simple dan spesifik, lebih fleksibel. Berikut langkah-langkah metode *decision tree* dengan algoritma C4.5 dalam membangun sebuah pohon keputusan adalah sebagai berikut :

1. Pilih Atribut sebagai akar.
2. Hitung jumlah data kolom berdasarkan anggota atribut dengan syarat tertentu.
3. Pilih atribut sebagai *Node*.

4. Buatlah cabang dari setiap nilai anggota dari *Node*.
5. Menentukan *entropy* dari tiap anggota *Node*. apabila ada *entropy Node* yang bernilai nol maka proses berhenti.
6. Perhitungan diulang apabila ada anggota *Node* yang memiliki nilai *entropy* lebih dari nol.

*Node* adalah atribut yang memiliki nilai *gain* tertinggi, sebelum menghitung *gain* kita perlu menghitung nilai *entropy* masing masing atribut. Berikut rumus yang digunakan :

$$\mathbf{Entropy(S)} = \sum_{i=1}^n - p_i * \log_2 p_i \quad (4)$$

Keterangan :

**S** = Himpunan Kasus

**n** = Jumlah partisi S

***p<sub>i</sub>*** = Proporsi S<sub>i</sub> terhadap S

Gain dihitung dengan rumus sebagai berikut :

$$\mathbf{Gain(S,A)} = \mathbf{Entropy(s)} - \sum_{i=1}^n \frac{|S_i|}{|S|} * \mathbf{Entropy(S_i)} \quad (4)$$

Keterangan :

**S** = Himpunan Kasus

**A** = Atribut

**n** = Jumlah partisi atribut A

**|S<sub>i</sub>|** = Proporsi S<sub>i</sub> terhadap S

**|S|** = Jumlah kasus dalam S

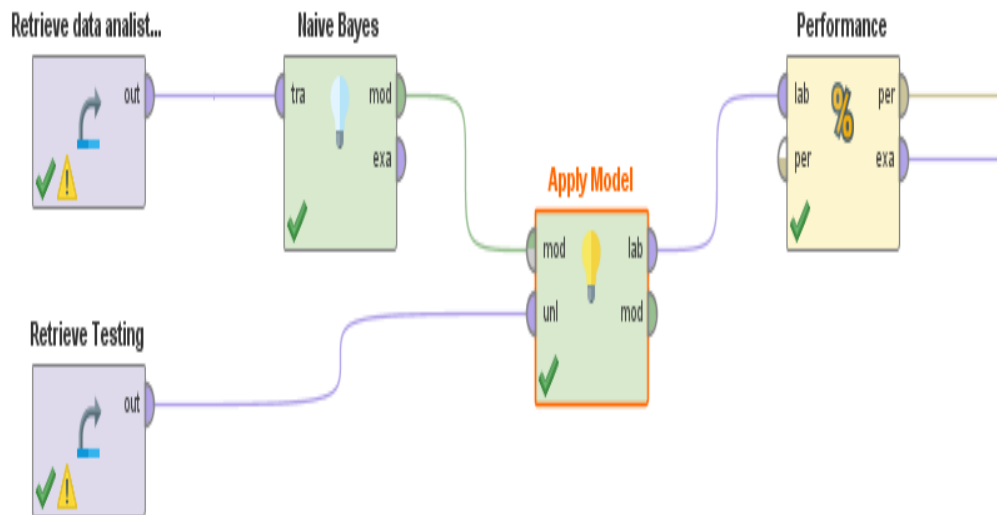
Setelah keduanya di dapatkan, maka pohon keputusan dibentuk dengan variable yang memiliki nilai *gain* tertinggi akan menjadi root, begitu seterusnya.

### C. Implementasi

#### 1. *RapidMiner*

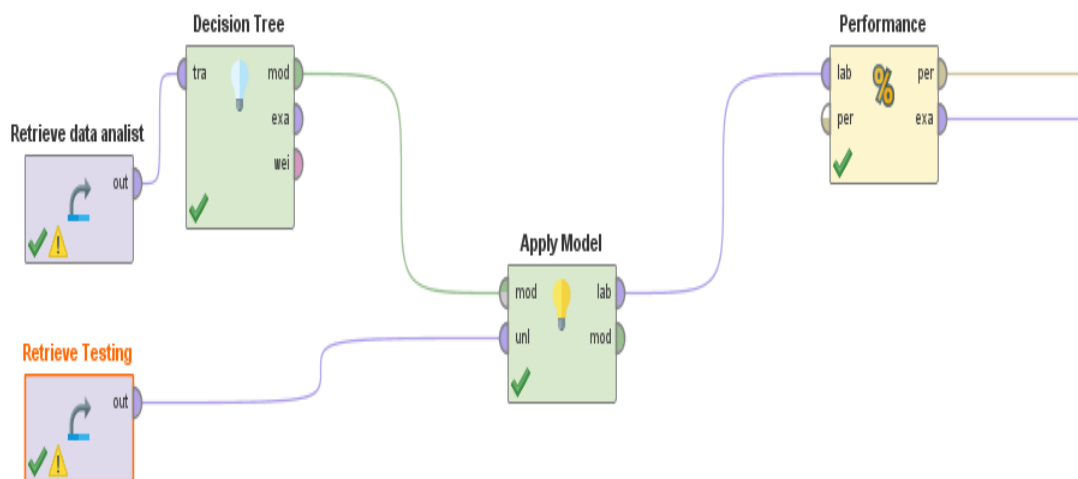
Untuk melakukan pengujian kedua metode, peneliti menggunakan sebuah perangkat lunak yaitu *RapidMiner 9.0*. *RapidMiner* merupakan Perangkat lunak yang digunakan untuk melakukan analisis text mining, data mining, dan analisis prediksi. Teknik deskriptif maupun prediksi diterapkan dalam perangkat lunak ini guna memudahkan pengguna dalam membuat keputusan yang terbaik dengan 500 operator data mining yang dimilikinya, meliputi operator *input*, *output*, *visualisasi*, dan *datapreprocessing*. Dibuat dengan Bahasa pemrograman java, *software* ini sangat mudah digunakan karena selain memiliki UI/UX yang bagus, perangkat lunak ini juga ringan sehingga bisa bekerja hamper di semua sistem.

Pengujian Algoritma *Naïve Bayes* dan C4.5 pada *RapidMiner* menggunakan data training dan data testing, kedua data tersebut adalah data penjualan motor Honda tahun 2017 yang telah di bersihkan sesuai dengan kebutuhan mining. Tahap implementasi dilakukan 2 kali sesuai dengan Metode yang kita gunakan dan ditambah dengan segmentasi *Performance* guna menentukan tingkat akurasi pada masing-masing metode. Design implementasi pengujian algoritma *Naïve Bayes* dengan menggunakan *tools RapidMiner* ditunjukkan.



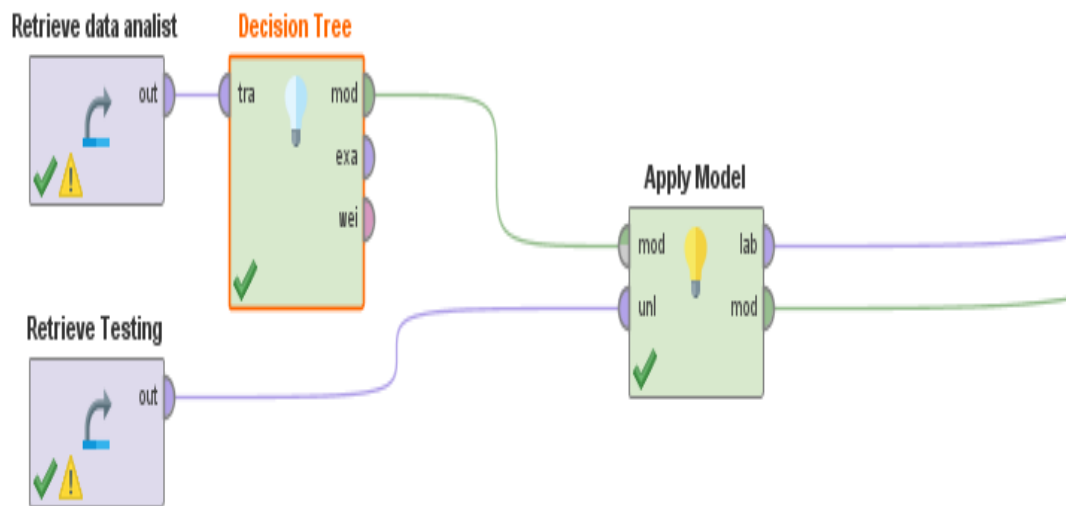
Gambar 2. Rapid Miner Dengan Metode Naive Bayes

Retrieve Testing adalah Testing data, Operator Apply Model digunakan untuk menerapkan model yang sudah dipelajari pada data pelatihan.dari algoritma *naive bayes*, Operator Performance digunakan untuk evaluasi kinerja statistik dari proses klasifikasi.



Gambar 3. RapidMiner Dengan Metode C4.5

Pada Gambar 3. Operator Retrieve data analyst adalah Training data, Operator Retrieve Testing adalah Testing data, Operator Apply Model digubakan untuk menerapkan model yang sudah dipelajari pada data pelatihan.dari algoritma *C4.5*, Operator Performance digunakan untuk evaluasi kinerja statistik dari proses klasifikasi.



Gambar 4. adalah design implementasi untuk menentukan desain C4.5

Gambar 4 adalah design implementasi untuk membuat pohon keputusan menggunakan algoritma *decision tree* atau C4.5. *Decision Tree* berguna untuk menentukan parameter apa yang mempengaruhi tingkat penjualan produk motor *Honda*.

## HASIL DAN PEMBAHASAN

### A. Uji Algoritma

Percobaan ini dilakukan untuk mengetahui tingkat akurasi dari kedua metode dan atribut yang mempengaruhi tingkat penjualan sepeda motor. Percobaan ini menggunakan dataset sebanyak 14 sampel yang diambil dari data penjualan tahun 2017 dengan hasil sebagai berikut:

accuracy: 85.71%

	true ya	true tidak	class precision
pred. ya	5	1	83.33%
pred. tidak	1	7	87.50%
class recall	83.33%	87.50%	

Gambar 5. Screenshoot Hasil Uji dengan Algoritma Naïve Bayes

accuracy: 85.71%

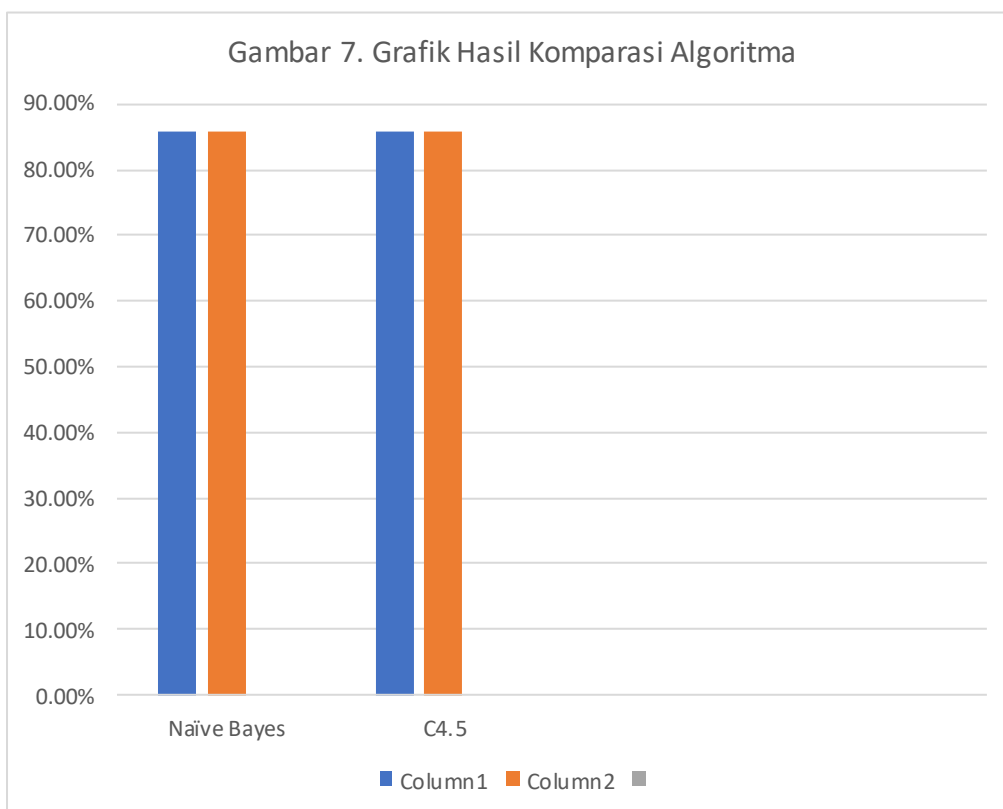
	true ya	true tidak	class precision
pred. ya	5	1	83.33%
pred. tidak	1	7	87.50%
class recall	83.33%	87.50%	

Gambar 6. Screenshoot Hasil Uji dengan Algoritma C4.5

Hasil dari pengujian algoritma *Naïve Bayes* dan C4.5 menggunakan *RapidMiner* menunjukkan keduanya memiliki tingkat akurasi yang sama, yaitu 85,71% seperti yang ditunjukkan pada Gambar 5 dan 6.

Tabel 3. Hasil Perbandingan Algoritma Naïve Bayes dan C4.5

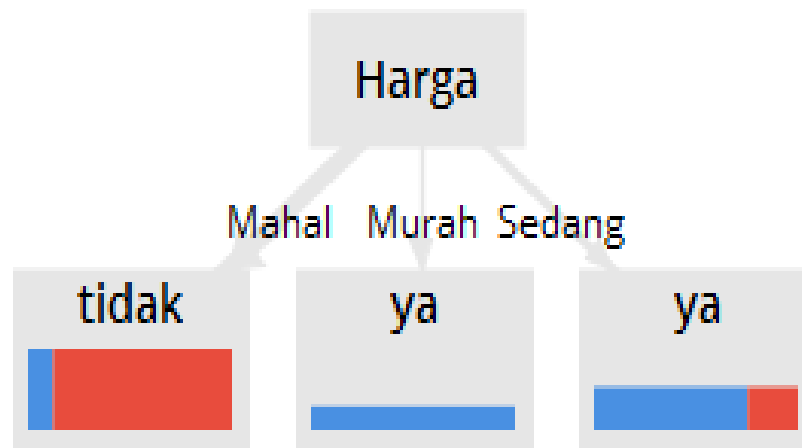
Algoritma	Akurasi
Naïve Bayes	85,71%
C4.5	85,71%





## B. Penentuan Pohon Keputusan

Setelah pengujian kedua algoritma, peneliti mencoba mencari atribut yang menjadi penentu tingkat kelarisan penjualan sepeda motor Honda di Indonesia. Dengan implementasi design pada Gambar 4 hasilnya sebagai berikut :



Gambar 8. Decision Tree dari Algoritma C4.5

Gambar 8 adalah pohon keputusan yang terbentuk dari penerapan algoritma C4.5 pada design yang ditunjukkan pada gambar 4. Dari hasil tersebut menunjukkan atribut Harga adalah yang menentukan tingkat kelarisan sepeda motor honda berdasarkan dataset yang ada.

Dari hasil analisa pohon keputusan yang terbentuk maka Motor Honda dengan harga murah dan sedang adalah yang paling laris, sehingga dealer dan atpm bisa memperluas pasar pada dua kategori tersebut. Sedangkan Motor Honda dengan harga mahal tidak diminati di pasar Indonesia.

## KESIMPULAN

Berdasarkan hasil dari implementasi algoritma *Naïve Bayes* dan C4.5 pada penentuan tingkat penjualan Sepeda motor honda dapat disimpulkan sebagai berikut:

- Penggunaan *software RapidMiner 9.0* sangat mempermudah proses penelitian dalam membandingkan kedua algoritma.
- Perbandingan Algoritma *Naïve Bayes* dan C4.5 dalam penentuan tingkat penjualan motor honda menunjukkan tingkat akurasi yang sama yaitu sebesar 85,71%.
- Decision tree* yang dihasilkan dari implementasi metode C4.5 menunjukkan atribut harga sangat mempengaruhi tingkat penjualan (laris atau tidaknya) motor honda. Motor dengan harga murah (< 15 jt) akan sangat laris dipasaran, dan motor dengan harga sedang (< 22 jt) cukup diminati, sedangkan motor dengan harga mahal (> 22 jt) sedikit sekali peminatnya.

## DAFTAR PUSTAKA

- [1] Azwanti, Nurul, 2018, Analisa Algoritma C4.5 Untuk Memprediksi Penjualan Motor pada PT. Capela Dinamik Nusantara Cabang Muka Kuning, *Informatika Mulawarman: Jurnal Ilmiah Ilmu Komputer*, Vol.13 No. 1 Februari 2018 Hal. 33-37.

- [2] Supriyanti, W., Kusriani., dan Amborowati, A., 2016, Perbandingan Kinerja Algoritma C4.5 dan Naïve Bayes Untuk Ketepatan Pemilihan Konsentrasi Mahasiswa, *Jurnal INFORMA Politeknik Indonusa Surakarta*, ISSN : 2442-7942, Vol 1 No. 3 Tahun 2016.
- [3] Georgina, Obuandike N., Alhasan, John, B., Abdullahi M., 2017, Clasification of Crime Data for Crime Control Using C4.5 and Naïve Bayes Techniques, *International Journal of Mathematiccal Analysis And Optimization: Theory And Applications*, Vol. 2017,PP. 139-153.
- [4] Harahap, F., 2015, Penerapan Data Mining Dalam Memprediksi Pembelian Cat, *Konferensi Nasional Sistem & Informatika STMIK STIKOM Bali*, Hal 856-862.
- [5] Ardiansyah, M. S., Fitri, Mustika L. S., Sapta, Andy, 2018, Analisa Kinerja Algoritma C4.5 Dalam Memprediksi Hasil Belajar, *Journal of Science and Social Research*, ISSN 2615-3262 (Online), February 2018 I(1): 73-79.
- [6] Edyson, Thomas T., Nova, Febri L., 2018, Analisa Pengaruh Padding Citra Bitmap 24 bit Terhadap Kebutuhan Memori Penyimpanan, *Jurnal Fahma*, Vol 16 No. 2 Mei 2018