

## **IMPLEMENTASI DATAMINING DENGAN ALGORITMA C4.5 UNTUK MEMPREDIKSI PEMBELIAN TIPE SEPEDA MOTOR**

**Siti Sundari**

Universitas Harapan Medan, Jl. H.M. Joni No 70 C Medan, [sundaristh@gmail.com](mailto:sundaristh@gmail.com)

### **Abstract**

Sales are one of the most important transactions in the business world. In managing a company, especially in the motorcycle showroom of CV Berlian Bintang Mas, purchasing the type of motorbike that consumers are interested in is very important and becomes a benchmark for a manager to make decisions in the future. There are several factors that determine consumer taste, including quality, price, design and technology. From these factors, a decision tree can be found using the Decision Tree. The C4.5 algorithm is a datamining method that is often used to make decisions. Rapid Miner 5, this application is very often used for decision trees. In this study, the design factor is very important and is the main factor affecting the purchase of a motorcycle. Furthermore, in the position of the two factors obtained from the decision tree is technology.

### **Keywords:**

*Data Mining, Decision Tree, C4.5 Algorithm, Rapid Miner.*

### **Abstrak**

Penjualan adalah salah satu transaksi yang sangat penting didunia bisnis. Dalam pengelolaan suatu perusahaan khususnya di showroom sepeda motor CV Berlian Bintang Mas, pembelian jenis sepeda motor yang diminati oleh konsumen sangatlah menjadi hal penting dan menjadi patokan seorang manager untuk mengambil keputusan kedepannya. Ada beberapa yang menjadi faktor penentu selera konsumen, diantaranya kualitas, harga, desain dan teknologi. Dari beberapa faktor inilah dapat dicari pohon keputusan menggunakan Decision Tree. Algoritma C4.5 adalah metode datamining yang sering digunakan untuk mengambil sebuah keputusan. Rapid Miner 5, aplikasi ini sangat sering digunakan untuk decision tree. Pada penelitian ini faktor desain sangatlah penting dan menjadi faktor utama yang mempengaruhi pembelian sepeda motor. Selanjutnya diposisi kedua faktor yang didapat dari pohon keputusan adalah teknologi.

### **Kata Kunci:**

*Data Mining, Decision Tree, Algoritma C4.5, Rapid Miner.*

## **1. PENDAHULUAN**

Didalam perkembangan dunia bisnis, terdapat beberapa hambatan dalam proses penjualan. Mulai dari pengolahan data yang masih manual menggunakan kertas hingga menggunakan *Microsoft excel*. Di CV Berlian Bintang Mas khususnya, pengambilan keputusan seorang manager dalam mengambil keputusan apakah produknya diterima oleh konsumen atau tidak, dan tipe sepeda motor yang paling diminati semua kalangan. Dengan melihat dari beberapa factor, jenis sepeda motor apakah yang diminati oleh konsumen. Diharapkan dengan menggunakan datamining, dapat mempermudah melakukan penjualan dikarenakan sudah mengetahui perkiraan jenis sepeda motor yang laris dipasar. Algoritma C4.5 merupakan algoritma klasifikasi pohon keputusan yang banyak digunakan karena memiliki kelebihan utama dari algoritma yang lainnya. Kelebihan algoritma C4.5 dapat menghasilkan pohon keputusan yang mudah diinterpretasikan, memiliki tingkat akurasi yang dapat diterima, efisien dalam menangani atribut bertipe diskret dan dapat menangani atribut bertipe diskret dan numerik. *Data mining* adalah proses yang menggunakan teknik statistik, matematika, kecerdasan buatan, dan *machine learning* untuk mengekstraksi dan mengidentifikasi informasi yang bermanfaat dan pengetahuan yang terkait dari berbagai *database* besar. Salah satu teknik yang ada pada *data mining* adalah klasifikasi [1].

Dalam penelitian sebelumnya dengan judul "Data Mining untuk Menganalisa Prediksi Mahasiswa berpotensi Non-Aktif Menggunakan Metode Decision Tree C4.5". Penelitian tersebut dilakukan untuk mengidentifikasi mahasiswa berpotensi non aktif menggunakan pohon keputusan C4.5, dari sisi keakuratan hasil prediksi. Hasil akhir dari tingkat akurasi penelitian ini diatas 90%. [2]. Yi Jiang et al melakukan penilaian terhadap kredit debitur. Penelitiannya menyatakan bahwa C4.5 adalah algoritma pembelajaran yang mengadopsi strategi pencarian lokal, dan dapat memperoleh aturan keputusan terbaik [3].

Algoritma C4.5 merupakan algoritma klasifikasi pohon keputusan yang banyak digunakan karena memiliki kelebihan utama dari algoritma yang lainnya. Kelebihan algoritma C4.5 dapat menghasilkan pohon keputusan yang mudah diinterpretasikan, memiliki tingkat akurasi yang dapat diterima, efisien dalam menangani atribut bertipe diskret dan dapat menangani atribut bertipe diskret dan numeric [4]. Pohon keputusan merupakan metode klasifikasi dan prediksi yang sangat kuat dan terkenal. Metode pohon keputusan mengubah fakta yang sangat besar menjadi pohon keputusan yang memprediksikan aturan. Aturan dapat dengan mudah dipahami dengan alami. Dan mereka juga dapat diekspresikan dalam bentuk bahasa basis data seperti *Structured Query Language* untuk mencari *record* pada kategori tertentu [5].

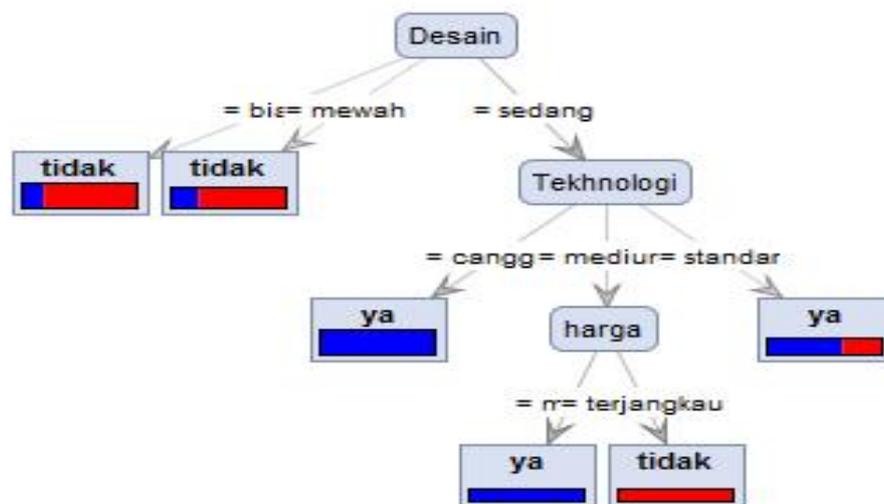
Proses pada pohon keputusan adalah mengubah bentuk data (tabel) menjadi model pohon, mengubah model pohon menjadi *rule* dan menyederhanakan *rule*. Secara umum algoritma C4.5 untuk membangun pohon keputusan adalah sebagai berikut [6] :

1. Pilih atribut sebagai akar
2. Buat cabang untuk tiap nilai
3. Bagi kasus dalam cabang
4. Ulangi proses untuk setiap cabang sampai semua kasus pada cabang memiliki kelas yang sama

## 2. HASIL DAN PEMBAHASAN

### 2.1. Implementasi

Tabel praproses dibuat menggunakan *Microsoft Excel* dengan format .xls, selanjutnya buka aplikasi Rapid Miner 5 pilih *new process*, pilih *importing* dan *drag Read excel* ke lembar kerja. Selanjutnya pilih *import configuration wizard* dan pilih data yang telah disimpan di *excel*. Kemudian pilih *modeling decision tree drag* ke lembar kerja dan hubungkan antara *read excel* dan *decision tree*. Setelah itu, klik *run* maka akan muncul pohon dibawah ini.



Gambar 1. Pohon Keputusan

### 2.2. Analisa Data

Data dalam pohon keputusan biasanya dinyatakan dalam bentuk tabel dengan atribut dan *record*. Atribut menyatakan suatu parameter yang dibuat sebagai kriteria dalam pembentukan pohon, salah satu atribut merupakan atribut yang menyatakan data solusi per item data yang disebut target atribut. Atribut memiliki nilai-nilai yang dinamakan dengan *instance* [7]

Adapun variable yang akan diproses *data mining* menggunakan algoritma C4.5 adalah sebagai berikut:

1. Kualitas

Variable ini berisi super, medium dan standar. Untuk kategori ini dilihat dari kapasitas mesin (CC) kategori super dilihat dari besarnya kapasitas sepeda motor yaitu diatas 150 CC, untuk medium antara 125 CC hingga 150 CC sedangkan untuk kategori standar 110 CC

2. Harga  
Variable harga terdiri atas tiga kategori, yaitu mahal, normal dan terjangkau. Adapun klasifikasi dari harga mahal adalah harga sepeda motor lebih dari 30 juta, untuk harga normal antara 20 jt sampai dengan 29 juta dan untuk harga terjangkau di bawah 20 jt.
3. Desain  
Adapun untuk variable desain terdapat tiga kategori antara lain mewah, sedang dan biasa. Desain dapat dikategorikan mewah jika *body* sport dan besar dan dilengkapi dengan fitur-fitur yang modern. Untuk kategori sedang meliputi fitur yang digunakan relatif modern, *body* juga tidak terlalu besar. Untuk kategori biasa desain *body* sepeda motor kecil.
4. Teknologi  
Variable teknologi meliputi canggih, medium dan standar.

### 2.3. Pemilihan Variabel

Pemilihan variabel didasarkan dari data-data yang diperoleh dilapangan. Ada beberapa factor yang menjadi penentu dalam pembelian sepeda motor di CV Berlian Bintang Mas antara lain adalah Kualitas dari sepeda motor, hal ini dilihat dari cc kendaraan, semakin besar cc nya akan digolongkan sebagai kendaraan sepeda motor mewah. Hal lain yang mempengaruhi pembelian yaitu harga yang ditawarkan. Teknologi juga sebagai hal yang tak lepas dari factor pembelian sepeda motor, yang tak kalah penting adalah bentuk desain dari sepeda motor yang dipasarkan. Berdasarkan uraian diatas, maka dapat diperoleh 4 variabel yaitu :kualitas, harga, desain dan teknologi yang ada di sepeda motor.

Tabel 1. Data Hasil Pra-proses

No	Kualitas	Harga	Desain	Teknologi	Beli
1	Super	Mahal	Mewah	Canggih	Ya
2	Medium	Mahal	Mewah	Canggih	Ya
3	Standar	Terjangkau	Sedang	Medium	Ya
4	Standar	Mahal	Sedang	Standar	Tidak
5	Medium	Normal	Biasa	Standar	Tidak
6	Medium	Normal	Sedang	Medium	Ya
7	Super	Mahal	Biasa	Standar	Tidak
8	Super	Normal	Sedang	Medium	Ya
9	Standar	Terjangkau	Sedang	Canggih	Ya
10	Super	Mahal	Biasa	Standar	Tidak
11	Medium	Mahal	Biasa	Canggih	Tidak
12	Super	Normal	Mewah	Medium	Ya
13	Super	Terjangkau	Sedang	Medium	Ya
14	Standar	Mahal	Biasa	Standar	Tidak
15	Standar	Normal	Mewah	Canggih	Ya
16	Medium	Mahal	Mewah	Canggih	Ya
17	Super	Normal	Mewah	Canggih	Ya
18	Standar	Mahal	Mewah	Canggih	Ya
19	Super	Normal	Biasa	Standar	Tidak
20	Medium	Terjangkau	Biasa	Standar	Ya
21	Standar	Mahal	Biasa	Canggih	Tidak

Perhitungan Entropy dan Gain

$$\text{Entropy (Total)} = \left( -\frac{8}{21} * \text{Log}_2 \left( \frac{10}{21} \right) \right) + \left( -\frac{13}{21} * \text{Log}_2 \left( \frac{11}{21} \right) \right)$$

$$= 0.958712$$

$$\text{Entropy (Kualitas Super)} = \left( -\frac{2}{8} * \text{Log}_2 \left( \frac{2}{8} \right) \right) + \left( -\frac{5}{8} * \text{Log}_2 \left( \frac{5}{8} \right) \right)$$

$$= 0.954434$$

$$\text{Entropy (Kualitas Medium)} = \left( -\frac{2}{6} * \text{Log}_2 \left( \frac{2}{6} \right) \right) + \left( -\frac{4}{6} * \text{Log}_2 \left( \frac{4}{6} \right) \right)$$

$$= 0.918296$$

$$\text{Entropy (Kualitas Standar)} = \left( -\frac{2}{7} * \text{Log}_2 \left( \frac{2}{7} \right) \right) + \left( -\frac{4}{7} * \text{Log}_2 \left( \frac{4}{7} \right) \right)$$

$$= 0.985228$$

$$\text{Entropy (Harga Mahal)} = \left(-\frac{9}{7} * \text{Log}_2\left(\frac{9}{7}\right)\right) + \left(-\frac{7}{7} * \text{Log}_2\left(\frac{7}{7}\right)\right)$$

$$= 0.970951$$

$$\text{Entropy (Harga Normal)} = \left(-\frac{2}{7} * \text{Log}_2\left(\frac{2}{7}\right)\right) + \left(-\frac{5}{7} * \text{Log}_2\left(\frac{5}{7}\right)\right)$$

$$= 0.863121$$

$$\text{Entropy (Harga Terjangkau)} = \left(-\frac{0}{4} * \text{Log}_2\left(\frac{0}{4}\right)\right) + \left(-\frac{4}{4} * \text{Log}_2(4)\right)$$

$$= 0$$

$$\text{Entropy (Desain Mewah)} = \left(-\frac{0}{7} * \text{Log}_2\left(\frac{0}{7}\right)\right) + \left(-\frac{7}{7} * \text{Log}_2\left(\frac{7}{7}\right)\right)$$

$$= 0$$

$$\text{Entropy (Desain Sedang)} = \left(-\frac{1}{6} * \text{Log}_2\left(\frac{1}{6}\right)\right) + \left(-\frac{5}{6} * \text{Log}_2\left(\frac{5}{6}\right)\right)$$

$$= 0.650022$$

$$\text{Entropy (Desain Biasa)} = \left(-\frac{7}{8} * \text{Log}_2\left(\frac{7}{8}\right)\right) + \left(-\frac{1}{8} * \text{Log}_2\left(\frac{1}{8}\right)\right)$$

$$= 0.543564$$

$$\text{Entropy (Teknologi Canggih)} = \left(-\frac{2}{9} * \text{Log}_2\left(\frac{2}{9}\right)\right) + \left(-\frac{7}{9} * \text{Log}_2\left(\frac{7}{9}\right)\right)$$

$$= 0.764205$$

$$\text{Entropy (Teknologi Medium)} = \left(-\frac{6}{5} * \text{Log}_2\left(\frac{6}{5}\right)\right) + \left(-\frac{5}{5} * \text{Log}_2\left(\frac{5}{5}\right)\right)$$

$$= -0.31564$$

$$\text{Entropy (Teknologi Standar)} = \left(-\frac{0}{7} * \text{Log}_2\left(\frac{0}{7}\right)\right) + \left(-\frac{7}{7} * \text{Log}_2\left(\frac{7}{7}\right)\right)$$

$$= 0$$

Setelah menghitung nilai *entropy*, hal yang dilakukan selanjutnya adalah menghitung nilai *Gain*. Adapun perhitungan *Gain* dapat dilihat sebagai berikut:

$$\text{Gain(Total, Kualitas)} = 0.958712 \left(\frac{8}{21} * 0.954434\right) + \left(\frac{6}{21} * 0.918296\right) + \left(\frac{7}{21} * 0.985228\right)$$

$$= 0.004338$$

$$\text{Gain(Total, Harga)} = 0.958712 \left(\frac{8}{21} * 0.970951\right) + \left(\frac{7}{21} * 0.863121\right) + \left(\frac{4}{21} * 0\right)$$

$$= 0.208648$$

$$\text{Gain(Total, Desain)} = 0.958712 \left(\frac{7}{21} * 0\right) + \left(\frac{6}{21} * 0.650022\right) + \left(\frac{8}{21} * 0.543564\right)$$

$$= 0.958712$$

$$\text{Gain(Total, Teknologi)} = 0.958712 \left(\frac{9}{21} * 0.764205\right) + \left(\frac{5}{21} * -0.31564\right) + \left(\frac{7}{21} * 0\right)$$

$$= 0.706348$$

**Tabel 2.** Perhitungan Node 1

Node			Jml Kasus	Tidak	Ya	Entropy	Gain
1	Total		21	8	13	0.958712	
	Kualitas						0.004338
		Super	8	3	5	0.954434	
		Medium	6	2	4	0.918296	
		Standar	7	3	4	0.985228	
	Harga						0.208648
		Mahal	10	6	4	0.970951	

		Normal	7	2	5	0.863121	
		Terjangkau	4	-	4	0	
	Desain						0.958712
		Mewah	7	-	7	0	
		Sedang	6	1	5	0.650022	
		Biasa	8	7	1	0.543564	
	Tekhnologi						0.706348
		Canggih	9	2	7	0.764205	
		Medium	5	6	5	-0.31564	
		Standar	7	-	1	0	

Dari perhitungan yang didapat di tabel 2, terlihat Gain Desain memiliki nilai yang paling tinggi. Dengan demikian, Desain dapat menjadi node akar dengan nilai 0,958712. Untuk atribut yang lain yaitu kualitas, harga dan teknologi perlu dilakukan perhitungan dengan cara yang sama.

### 3. KESIMPULAN

Dari hasil penelitian, dapat ditarik kesimpulan bahwa pembelian sepeda motor menggunakan C4.5 akan bermanfaat untuk pengambilan keputusan. Faktor desain sangatlah penting dan menjadi faktor utama yang mempengaruhi pembelian sepeda motor. Selanjutnya diposisi kedua faktor yang didapat dari pohon keputusan adalah teknologi. Saran penulis adalah agar dapat dibandingkan dengan metode datamining yang lain.

### DAFTAR PUSTAKA

- [1] Efraim Turban, Jay E. Aronson, Ting Peng Liang, 2005. *Decision Support System and Intelligent Systems Edisi 7 Jilid 1*, Andi Yogyakarta.
- [2] Dwi Untari, 2010. Data Mining untuk Menganalisa Prediksi Mahasiswa berpotensi Non-Aktif Menggunakan Metode Decision Tree C4.5, Universitas Dian Nuswantoro, Semarang.
- [3] Jiang, Yi. et al, (2007). "A Bank Customer Credit Evaluation Based on the Decision Tree and the Simulated Annealing Algorithm. *Journal of Department of Computer Science Xiamen University (IEEE International Co 8-11 July 2008)*".
- [4] Han, J., Kamber, M., 2001, "Data Mining Concepts and Techniques", Morgan Kaufman Pub., USA.
- [5] Kusriani, (2009). *Algoritma Data Mining*, Andi Yogyakarta
- [6] Muhammad Syahril, 2011. "Konversi Data Training Tentang Penyakit Hipertensi Menjadi Bentuk Pohon Keputusan dengan Teknik Klasifikasi Menggunakan Tools Rapid Miner 4.1, Vol 10, No.2, Mei.
- [7] Fitriana Harahap, (2015). *Penerapan Data Mining dalam Memprediksi Pembelian Cat*, STMIK Potensi Utama, Medan.