

IMPLEMENTASI DATA MINING ALGORITMA C4.5 UNTUK MEMPREDIKSI LALU LINTAS PENUMPANG BERANGKAT DAN DATANG MELALUI PELABUHAN TANJUNG PRIOK TAHUN 2021

Rukmania, Viona Cantika, Ahmad Yuda Pratama, Muhammad Rizky Erlangga

Sistem Informasi, Universitas Bina Sarana Informatika

Jalan Kayu Jati V Pemuda, Indonesia

19221168@bsi.ac.id.

ABSTRAK

Pelabuhan merupakan area yang mencakup Wilayah darat dan air di sekelilingnya, dengan batas yang jelas, yang berfungsi sebagai pusat perdagangan dan distribusi barang, Di dalam negeri maupun internasional. Saat ini, jumlah penumpang yang menggunakan transportasi laut tidak sebanyak pengguna transportasi darat maupun udara. Kondisi ini tentunya memberikan dampak signifikan bagi perusahaan yang bergerak di sektor transportasi laut, sehingga mendorong mereka untuk mencari solusi guna mengatasi permasalahan tersebut. Berdasarkan hal tersebut, Penulis mengaplikasikan metode C4.5 untuk memberikan jawaban dalam meramalkan tingkat lalu lintas penumpang di Pelabuhan Tanjung Priok. Prediksi ini didasarkan pada data jumlah kedatangan dan keberangkatan penumpang selama tahun 2021. Algoritma C4.5 diimplementasikan menggunakan perangkat lunak RapidMiner versi 7.1. Dari penerapan metode ini, diperoleh pohon keputusan beserta aturan yang menunjukkan bahwa atribut kedatangan memiliki pengaruh terbesar terhadap lalu lintas penumpang, diikuti oleh atribut keberangkatan. Implementasi algoritma ini juga telah diuji menggunakan RapidMiner, dan hasilnya konsisten dengan perhitungan manual. Model algoritma C4.5 yang digunakan menghasilkan tingkat akurasi mencapai 100%.

Kata kunci : Klasifikasi, Data Mining, Algoritma C4.5

1. PENDAHULUAN

Pelabuhan adalah suatu tempat di daratan dan lautan di sekitarnya yang mempunyai batas-batas tertentu, menurut Undang-Undang No. 21 Tahun 1992 tentang Pelayaran. Area ini digunakan untuk mendukung kegiatan pemerintahan dan ekonomi. Selain menerima dan menurunkan penumpang, kapal dapat berlabuh dan berlabuh di pelabuhan, serta memuat dan membongkar muatan. Lebih jauh, pelabuhan berfungsi sebagai titik transisi antara atau di dalam moda transportasi, dan dilengkapi dengan fasilitas keselamatan pelayaran dan fasilitas pendukung lainnya. Bangsa dengan populasi terbesar keempat secara global, yaitu Indonesia, adalah negara berkembang dengan status pendapatan menengah[1]. Efektivitas pertumbuhan ekonomi suatu negara dapat dilihat dari sektor perdagangannya [2].

Sarana transportasi utama untuk komoditas di Indonesia dan di seluruh dunia, Pelabuhan merupakan tempat yang meliputi daratan dan perairan dengan batas tertentu. yang ditentukan. Hampir 90% perdagangan Indonesia dilakukan melalui laut, dan angka ini meningkat menjadi 85% untuk seluruh dunia. Lima pelabuhan utama di Indonesia adalah Massar, Balikpapan, Tanjung Perak, Pelabuhan Belawan, dan Tanjung Priok. Di antara pelabuhan utama Indonesia untuk bongkar muat, Pelabuhan Tanjung Priok terletak di Jakarta.

Leonardo (2020) menemukan bahwa dibandingkan dengan transportasi darat dan udara, jumlah penumpang yang menggunakan transportasi laut kini lebih sedikit. Namun, rendahnya biaya perjalanan laut membuat banyak calon penumpang

masih memilih moda transportasi ini. Selain itu, transportasi laut memiliki kapasitas angkut yang lebih besar untuk barang dan menawarkan konektivitas yang lebih luas, mencakup hampir seluruh wilayah di Indonesia[3].

Tempo melaporkan pada 2 Februari 2020, secara keseluruhan, penumpang di semua pelabuhan utama menurun, dengan penurunan 76,51% di Pelabuhan Tanjung Priok dan penurunan 73,89% di Pelabuhan Belawan, Makassar 67,22%, Balikpapan 66,74%, dan Tanjung Perak 57,72%. Kondisi ini memberikan dampak signifikan bagi perusahaan yang bergerak di sektor transportasi laut, mendorong mereka untuk mencari solusi guna mengatasi berbagai tantangan yang dihadapi.

Berdasarkan data ini maka untuk mendukung sistem lalu lintas penumpang di Pelabuhan Tanjung Priok yang efektif, maka penulis mengimplementasikan metode algoritma C4.5 untuk memberi solusi dalam memprediksikan tingkat lalu lintas penumpang di Pelabuhan Tanjung Priok berdasarkan data penumpang datang dan berangkat pada tahun 2021.

2. TINJAUAN PUSTAKA

2.1. Pelabuhan

Pelabuhan ialah daerah perairan yang sudah terlindungi dari gelombang dan dilengkapi dengan sarana dan prasarana terminal laut seperti dermaga, dimana kapal ini bertambat untuk bisa melakukan pembongkaran dan pemuatan barang. Kran untuk pembongkaran dan muat, Gudang laut serta tempat untuk menyimpan kapal yang membongkar muatan.

Selain itu juga bisa jadi Gudang untuk barang yang disimpan dalam waktu yang lama selama menunggu untuk dikirimkan ke tempat tujuan[4].

2.2. Data Mining

Salah satu cara Untuk mengekstrak informasi yang berharga dari basis data adalah dengan menggunakan penambangan data. Beberapa metode, termasuk pembelajaran mesin dan kecerdasan buatan, digunakan dalam aplikasi praktis penambangan data, teknik statistik yang dipakai untuk memisahkan data serta melakukan identifikasi data untuk keperluan informasi dari suatu database yang lebih besar Data mining adalah teknik untuk mencari informasi dalam database. Dalam praktiknya, data mining menggunakan berbagai teknik seperti kecerdasan buatan, pembelajaran mesin, dan metode statistik untuk memisahkan dan mengidentifikasi data dari database besar untuk tujuan informasi [5].

2.3. Algoritma C4.5

Algoritma C4.5 yaitu merupakan sebuah pohon keputusan atau yang dikenal dengan decision tree[6].

Algoritma C4.5 dikembangkan oleh Ross Quinlan sebagai pengembangan dari algoritma ID3. Algoritma C4.5 adalah pembentukan pohon keputusan berdasarkan training data. Kelebihan algoritma ini yaitu dapat mengatasi *missing value*, dapat mengolah data numerik serta kategori *noisy* data. Dalam membuat pohon keputusan dengan algoritma C4.5 diperlukan langkah-langkah sebagai berikut

a. Siapkan Data Training

Data training adalah data histori yang telah disiapkan sebelumnya dan telah dikelompokkan ke dalam bentuk kelas-kelas tertentu.

b. Menentukan akar dari pohon

Akar akan ditentukan melalui atribut yang telah dipilih dengan cara menghitung *gain* dari setiap atribut, nilai *gain* yang paling tinggi nantinya akan menjadi akar pertama. Sebelum menghitung nilai *gain* dari atribut perlu menghitung *entropy* terlebih dahulu[7]. Untuk menghitung nilai *entropy* dapat digunakan rumus sebagai berikut

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

Dengan keterangan,

N : himpunan kasus

p : jumlah partisi N

Pi : proporsi dari Ni terhadap N

c. Setelah menghitung entropy, maka dapat dilanjutka dengan menghitung *gain* menggunakan rumus sebagai berikut

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

Dengan keterangan dimana :

N : himpunan kasus

A : atribut

p : jumlah partisi atribut A

|Ni| : jumlah kasus pada partisi ke i

|N| : jumlah kasus dalam N

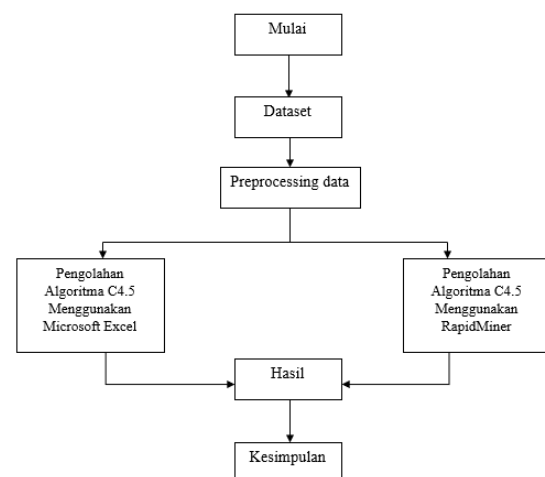
2.4. RapidMiner

RapidMiner adalah platform perangkat lunak data ilmu pengetahuan yang dikembangkan oleh perusahaan dengan nama yang sama, yang menyediakan lingkungan terpadu untuk pembelajaran mesin (machine learning), pembelajaran mendalam (deep learning), penambangan teks (text mining), dan analisis prediktif (predictive analytics). Aplikasi ini digunakan untuk aplikasi bisnis dan komersial serta untuk penelitian, pendidikan, pelatihan, pembuatan prototype dengan cepat, dan pengembangan aplikasi serta mendukung semua langkah proses pembelajaran mesin termasuk persiapan data, visualisasi hasil, validasi dan pengoptimalan[8].

3. METODE PENELITIAN

Metode penelitian memakai Algoritma C4.5 dengan bantuan Microsoft Excel dan software RapidMiner. Sumber data memakai data sekunder dari Badan Pusat Statistik Jakarta Utara[9].

Strategi pengumpulan data dalam penelitian ini mencakup mengambil data statistik lalu lintas datang dan berangkat di Pelabuhan Tanjung Priok Tahun 2021. Data yang diperoleh sebanyak 11 yakni dari bulan Januari sampai November. Teknik analisis data memakai bantuan software RapidMiner. Adapun alur penelitian ini bisa dilihat pada Gambar 3.



Gambar 1. Alur Penelitian

4. HASIL DAN PEMBAHASAN

4.1. Penerapan Algoritma C4.5

Didalam penelitian ini data yang dipakai ini berasal dari Badan Pusat Statistik 2021. Data ini selanjutnya akan diolah menurut Algoritma C4.5 di Excel untuk menghitung *entropy* dan *gain*. Selanjutnya data akan diolah di software RapidMiner versi 7.1 untuk menghasilkan klasifikasi dan

mengambil keputusan berdasarkan algoritma C4.5. Berikut adalah sampel data hasil rekapitulasi dari kuesioner yang telah diolah di Microsoft Excel, seperti yang disajikan pada Tabel 1:

Tabel 1. Data Lalu Lintas Penumpang yang Berangkat dan Datang melalui Pelabuhan Laut Tanjung Priok Menurut Bulan, 2021

Bulan	Datang	Berangkat	Kategori
Januari	3840	2846	Rendah
Februari	4721	4362	Rendah
Maret	4521	4071	Rendah
April	6606	6124	Tinggi
Mei	3066	2431	Rendah
Juni	9453	8179	Tinggi
Juli	3996	3590	Rendah
Agustus	3302	2768	Rendah

Bulan	Datang	Berangkat	Kategori
September	4335	4750	Rendah
Oktober	5318	7162	Tinggi
November	6635	8103	Tinggi

Perhitungan algoritma C4.5 ini dimulai dengan melakukan pemilihan pada atribut akar lebih dahulu yakni mencari total kasus secara menyeluruh, total kasus ini berkaitan dengan rendah atau tingginya tingkat keberangkatan dan kedatangan di Pelabuhan Tanjung Priok. Kemudian akan dihitung Ukuran ketidakpastian dari seluruh kasus yang dikelompokkan menurut atribut kedatangan dan keberangkatan. Sesudah itu dilaksanakan perhitungan gain untuk kedua atribut ini. Hasil perhitungan diperlihatkan oleh tabel berikut ini.

Tabel 2. Perhitungan Node 1

Attribute		Jumlah (s)	Kategori		Entropy	Gain
			Tinggi	Rendah		
Total		11	4	7	0.94566	
Kedatangan						0.94566
	< 5072	7	0	7	0.0000	
	> 5072	4	4	0	0.0000	
Keberangkatan						0.94566
	<4944	7	0	7	0.0000	
	>4944	4	4	0	0.0000	

Berikut perhitungan *entropy* dari table 2 sebagai berikut

Rumus Entropy:

$$Entropy(S) = \sum_{i=1}^n p_i \cdot \log_2 p_i$$

$$Entropy \text{ Total} = ((-4/11 \cdot \log_2 (4/11)) + ((-7/11) \cdot \log_2 (7/11))) = 0,945660305$$

Lalu dapat dilihat pada tabel 2 untuk hasil perhitungan atribut kedatangan dan keberangkatan dengan kategori rendah dan tinggi

$$Entropy \text{ Kedatangan, Kategori Rendah} = ((-7/7 \cdot \log_2 (7/7)) + ((-0/7) \cdot \log_2 (0/7))) = 0$$

$$Entropy \text{ Kedatangan, Kategori Tinggi} = ((-4/4 \cdot \log_2 (4/4)) + ((-0/4) \cdot \log_2 (0/4))) = 0$$

$$Entropy \text{ Keberangkatan, Kategori Rendah} = ((-7/7 \cdot \log_2 (7/7)) + ((-0/7) \cdot \log_2 (0/7))) = 0$$

$$Entropy \text{ Keberangkatan, Kategori Tinggi} = ((-4/4 \cdot \log_2 (4/4)) + ((-0/4) \cdot \log_2 (0/4))) = 0$$

Lalu. Menghitung *Gain* dengan rumus sebagai berikut

$$Gain(S, A) = entropy(S) - \sum |S_i|/|S| Entropy(S_i)$$

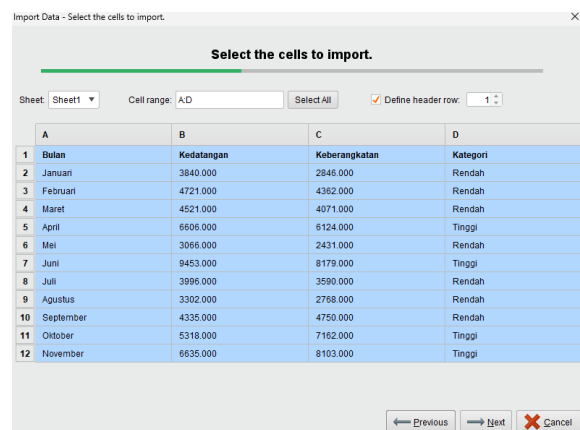
$$Gain \text{ Atribut Kedatangan} = (0,945660305) - ((7/11) \cdot 0) - ((4/11) \cdot 0) = 0,945660305$$

Sesudah dilaksanakan perhitungan dengan cara yang sama untuk atribut kedatangan dan keberangkatan maka didapatkan nilai *entropy* dan *gain* tiap atribut. Berdasarkan hasil perhitungan nilai *entropy* dan *gain* yang terdapat pada tabel 2, bahwa nilai *gain* kedua atribut sama yakni 0.94566 namun yang membedakan ialah banyaknya total penumpang. Sehingga dalam penelitian ini atribut kedatangan

dipakai sebagai pohon atau *Root*, dengan pembagian cabang berdasarkan jumlah partisi yang ada. Pada atribut kedatangan terdapat dua partisi yakni “Tinggi” dan “Rendah”, sehingga dari *Root* kedatangan akan dibentuk dua cabang yakni “Tinggi” dan “Rendah”.

4.2. Pengolahan Data Algoritma C4.5 Memakai RapidMiner

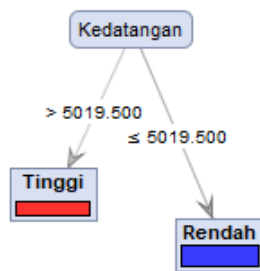
Berikutnya data diproses dengan memakai *software* RapidMiner serta untuk mengolahnya memakai Split Validation yang ada pada RapidMiner, berikutnya data bisa dilihat dari pohon keputusan, nilai akurasi, serta performancenya. Adapun proses penginputan data di RapidMiner bisa dilihat pada Gambar 2.



Gambar 2. Data yang di-import ke RapidMiner

Berikutnya dihitung memakai software Rapidminer versi 7.1. adapun untuk proses analisis datanya ini dengan memakai C4.5 di software RapidMiner yang dimulai dengan membuat pohon keputusan yang memberikan gambaran terkait hubungan antara atribut data dengan label yang akan diprediksi. Adapun algoritma yang dipakai di software RapidMiner ini ialah memakai *decision tree*. *Decision Tree* ialah salah satu cara data processing yang dipakai dalam memprediksikan masa depan dengan membangun klasifikasi dalam bentuk struktur pohon. Hal ini dilaksanakannya dengan memecah terus ke dalam himpunan bagian yang lebih kecil kemudian Ketika itu juga pohon keputusan akan bertahap berkembang. Hasil akhir dari proses ini ialah pohon dengan node keputusan serta node daun.

Atribut yang memiliki nilai gain terbesar akan berperan sebagai simpul (node) atau daun (leaf) berikutnya dalam struktur pohon keputusan. Setelah seluruh perhitungan selesai dilakukan, pohon keputusan yang dihasilkan menggambarkan hubungan antar atribut berdasarkan algoritma C4.5. Output akhir dari proses ini adalah pohon keputusan yang divisualisasikan, seperti yang ditunjukkan pada Gambar 3.



Gambar 3. Pohon Keputusan

Dari gambar 3 diperlihatkan bahwa untuk atribut kedatangan dibagi menjadi dua ada kategori tinggi dan rendah. Kategori tinggi untuk nilai > 5019.500 dan kategori rendah dengan nilai ≤ 5019.500 . Dalam hal ini daun pohon diperoleh karena nilai ini lebih besar dari yang lainnya khususnya untuk atribut keberangkatan meskipun nilai gainnya sama.

4.3. Tampilan deskripsi *decision tree* Tree

```

Kedatangan > 5019.500: Tinggi {Rendah=0, Tinggi=4}
Kedatangan ≤ 5019.500: Rendah {Rendah=7, Tinggi=0}
  
```

Gambar 4. *Decision Tree*

Gambar 4 diatas, memberikan representasi menyeluruh dari struktur pohon keputusan yang dihasilkan melalui penerapan algoritma C4.5. Pohon keputusan tersebut menggambarkan secara detail seluruh cabang dan node yang terbentuk dari proses klasifikasi data. Hasil ini memperlihatkan untuk

atribut kedatangan > 5019.500 dengan kategori tinggi ada 4 dan ≤ 5019.500 kategori rendah ada 7.

4.4. Tingkat Akurasi

Setelah proses pengolahan data selesai dilakukan, evaluasi tingkat akurasi dapat dilaksanakan untuk mengukur efektivitas metode C4.5 yang diterapkan. Salah satu tujuan utama penelitian ini adalah menganalisis tingkat ketepatan dalam mengklasifikasikan volume lalu lintas penumpang, khususnya dalam membandingkan frekuensi kedatangan dan keberangkatan. Untuk menilai tingkat akurasi ini, digunakan metode pengujian *confussion matrix* sebagai alat ukur yang dapat diandalkan.

accuracy: 100.00%

	true Rendah	true Tinggi	class precision
pred Rendah	7	0	100.00%
pred Tinggi	0	4	100.00%
class recall	100.00%	100.00%	

Gambar 5. Nilai Akurasi Algoritma C4.5

Dari hasil ini diperoleh tingkat akurasi 100% untuk kategori tinggi dan rendah pada atribut kedatangan.

PerformanceVector

```

PerformanceVector:
accuracy: 100.00%
ConfusionMatrix:
True:  Rendah  Tinggi
Rendah:  7      0
Tinggi:  0      4
  
```

Gambar 6. Nilai *Performance Vector* Algoritma C4.5

Pengujian yang dilakukan menggunakan perangkat lunak RapidMiner menghasilkan tingkat akurasi sistem yang mencapai 100%. Hal ini mengindikasikan bahwa aturan-aturan (rules) yang dihasilkan memiliki tingkat ketepatan yang sempurna. Validasi model telah dilakukan melalui metode split validation yang diimplementasikan pada platform RapidMiner versi 7.1, yang memastikan kehandalan dari hasil klasifikasi yang diperoleh.

5. KESIMPULAN DAN SARAN

Setelah menghitung entropi dan nilai perolehan untuk setiap atribut, penelitian ini menunjukkan bahwa penambahan data yang menggunakan algoritma C4.5 dapat digunakan untuk mengetahui lalu lintas penumpang yang masuk dan keluar dari Pelabuhan Tanjung Priok dengan Tingkat akurasi 100%. Menurut pohon keputusan dan aturan yang dihasilkan oleh Algoritma C4.5, atribut kedatangan memiliki dampak terbesar pada arus penumpang dengan gain, diikuti oleh atribut keberangkatan.

DAFTAR PUSTAKA

- [1] S. Sahara, V. K. Ladesi, W. Hadi, and K. Verawati, "Ramp check examination evaluation of public transport business," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 1098, no. 2, p. 022069, 2021, doi: 10.1088/1757-899x/1098/2/022069.
- [2] S. Sahara, "Analysis of Factors Causing Delay in Unloading Imported Goods," *Proc. Conf. Broad Expo. to Sci. Technol. 2021 (BEST 2021)*, vol. 210, no. Best 2021, pp. 345–351, 2022, doi: 10.2991/aer.k.220131.051.
- [3] F. Leonardo, "ANALISIS DIMENSI KUALITAS PELAYANAN DAN KEPUASAN PENUMPANG KAPAL LAUT (Studi Kasus Pada Terminal Penumpang Nusantara Pura PT. Pelabuhan Indonesia II (Persero) Cabang Tanjung Priok)," *J. Akunt. dan Keuang.*, vol. 3, pp. 1–9, 2020, [Online]. Available: <http://repository.stei.ac.id/1528/>
- [4] A. A. SHELEMO, "No Title," *Nucl. Phys.*, vol. 13, no. 1, pp. 104–116, 2023.
- [5] Dodi Nofri Yoliad, "Data mining Dalam Analisis Tingkat Penjualan Barang Elektronik Menggunakan Algoritma K-means," *Insearch (Information Syst. Res. J.)*, vol. 3, no. 1, 2023.
- [6] A. Gusti and S. Hadiani, "Application of the C4.5 Algorithm to Predict the Effectiveness of Google Adwords Ads," *Sistemasi*, vol. 12, no. 1, p. 21, 2023, doi: 10.32520/stmsi.v12i1.1978.
- [7] I. Iddrus and D. W. Sari, "Penerapan Data Mining Menggunakan Algoritma Decision Tree C4.5 Untuk Memprediksi Mahasiswa Drop Out Di Universitas Wiraraja," *J. Adv. Res. Inform.*, vol. 1, no. 02, pp. 1–7, 2023, doi: 10.24929/jars.v1i02.2684.
- [8] D. Nisrina, T. Informatika, and U. N. Waluyo, "Analisis Kepuasan Konsumen Menggunakan Metode Algoritma C4 . 5 Berbasis Rapidminer pada PT . Adeaksa," vol. 9, no. 1, pp. 26–33, 2024.
- [9] "Lalu Lintas Penumpang yang Berangkat dan Datang melalui Pelabuhan Laut Tanjung Priok Menurut Bulan, 2021-2022."