GO INFOTECH: JURNAL ILMIAH STMIK AUB
Vol.30, No.1, June 2024
ISSN (p): 1693-590x, ISSN (e): 2686-4711
DOI: 10.36309/goi.v30i1.266

Implementasi CRISP-DM pada Data Mining untuk Melakukan Prediksi Pendapatan dengan Algoritma C.45

Diyah Ruswanti*[1], Dahlan Susilo[2], Riani[3]
[1,2,3]Universitas Sahid Surakarta; Jalan Adi Sucipto No 154 Jajar-Laweyan, Indonesia
[3]Program Studi Informatika, Universitas Sahid Surakarta, Surakarta, Indonesia
e-mail: [1]dyahruswanti@usahidsolo.ac.id, [2]dahlan.susilo@usahidsolo.ac.id,
[3]rianiadvance@gmail.com

Abstrak

Penelitian ini membahas penerapan metodologi CRISP-DM (Cross Industry Standard Process for Data Mining) untuk melakukan analisis prediksi pendapatan menggunakan algoritma C4.5. Tujuan utama penelitian ini adalah menprediksi jenis kendaraan yang menyumbang pendapatan terbesar pada uji kir kendaraan di Dinas Pehubungan. Pada tahap pertama, Tahapan Business Understanding, penelitian ini mendefinisikan tujuan bisnisnya sebagai pengembangan model prediksi pendapatan untuk mendukung keputusan strategis. Selanjutnya, pada Tahapan Data Understanding, data pendapatan individu dikumpulkan dan dieksplorasi untuk pemahaman awal sebelum dilakukan pemrosesan data lebih lanjut. Pada tahap Data Preparation, data yang diperoleh dibersihkan, diubah, dan dipersiapkan untuk analisis. Faktor-faktor yang dianggap penting untuk prediksi pendapatan, seperti pendidikan, pekerjaan, dan status perkawinan, dipilih dan diolah dengan cermat. Tahapan Modeling melibatkan penerapan algoritma C4.5 untuk membangun model prediksi. Evaluasi dilakukan terhadap kualitas model, termasuk akurasi dan interpretasi aturan yang dihasilkan oleh algoritma. Hasil analisis menunjukkan bahwa model prediksi pendapatan menggunakan algorithma C4.5 mampu memberikan hasil yang memuaskan dengan tingkat akurasi sebesar 75% dengan jenis kendaraan yang diprediksi pendapatan uji kir naik adalah jenis Light Truk.

Kata kunci: *crisp-dm, prediksi, pendapatan, c.45, data mining*

Abstract

This research discusses the application of the CRISP-DM (Cross Industry Standard Process for Data Mining) methodology in conducting revenue prediction analysis using the C4.5 algorithm. The main objective of this research is to predict the type of vehicle that contributes the largest income to vehicle driver's license tests at the Transportation Department. In the first stage, the Business Understanding phase, this research defines its business objectives as developing a revenue prediction model to support strategic decisions. Next, at the Data Understanding stage, individual income data is collected and explored for initial understanding before further data processing is carried out. In the Data Preparation stage, the data obtained is cleaned, transformed, and prepared for analysis. Factors considered important for income prediction, such as education, employment, and marital status, are carefully selected and processed. The Modeling phase involves applying the C4.5 algorithm to build a prediction model. Evaluation is carried out on the quality of the model, including accuracy and interpretation of the rules generated by the algorithm. The results of the analysis show that the revenue prediction model using the C4.5 algorithm is able to provide satisfactory results with an accuracy level of 75% with the type of vehicle predicted by the KIR test increasing revenue being the Light Truck type.

Keywords: *crisp-dm, prediksi, pendapatan, c.45, data mining*

Received June 1st, 2024; Revised June 9th, 2024; Accepted June 19th, 2024

1. PENDAHULUAN

Data merupakan aset yang sangat berharga. Dalam era perkembangan teknologi yang pesat, dimana data dikumpulkan dan disimpan mampu menghasilkan basis data yang sangat besar. Namun, seringkali data yang tersebut tidak dilihat karena terlalu besar, kurang menarik, dan membutuhkan waktu yang lama untuk diolah. Akibatnya, pengambilan keputusan yang seharusnya didasarkan pada data sering dibuat berdasarkan pemikiran atau perasaan pembuat keputusan. Inilah yang mendorong lahirnya data mining. Data mining adalah proses otomatisasi pencarian data dalam database yang mempunyai skala besar untuk menghasilkan informasi yang berguna. Perkembangan data mining mengikuti dari kemajuan teknologi informasi yang memungkinkan akumulasi data dalam jumlah besar. Keterkaitan ini membuat data mining berkembang pesat di berbagai sektor bisnis.

Dalam kegiatan bisnis biasanya ada persaingan dan perkembangan yang dinamis, para pelaku bisnis harus dapat memikirkan strategi untuk bertahan sekaligus mengembangkan skala bisnisnya. Untuk mencapai hal tersebut, analisis data perusahaan menjadi salah satu cara yang efektif. Dengan melakukan analisis data, perusahaan dapat meningkatkan kapasitas produk, efisiensi biaya operasional, dan efektivitas kegiatan pemasaran.

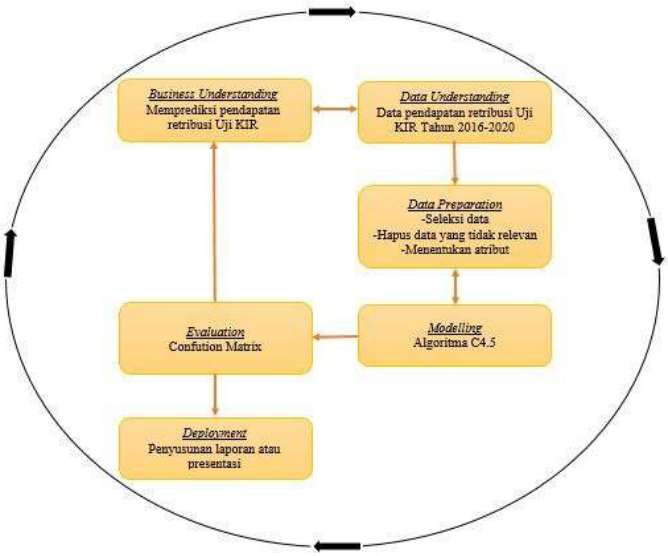
Dinas Perhubungan merupakan bagian dari Perangkat Daerah dengan tugas urusan pemerintahan pad abidang perhubungan. Kepala Dinas Perhubungan memimpin Dinas Perhubungan dan mempunyai tanggung jawab kepada Bupati melalui Sekda. Salah satu tugasnya adalah melaksanakan Uji Kendaraan apakah layak digunakan di jalan raya atau tidak (Uji KIR). Retribusi adalah salah satu sumber PAD (Pendapatan asli Daerah) yang untuk membiayai kegiatan Pemerintah Daerah. Retribusi Pengujian Kendaraan Bermotor atau disebut juga Retribusi Uji KIR, termasuk dalam kategori retribusi Jasa Umum. Menurut peraturan yang berlaku, biaya ini merupakan kompensasi atas layanan pengujian kendaraan bermotor. Tarif untuk uji KIR diatur dalam Peraturan Daerah Kabupaten Karanganyar Nomor 20 Tahun 2019, yang merupakan Perubahan Ketiga Atas Peraturan Daerah Kabupaten Karanganyar Nomor 4 Tahun 2012 Tentang Retribusi Jasa Umum.

Menurut data yang dikumpulkan oleh Dinas Perhubungan Kabupaten Karanganyar dari tahun 2016 hingga 2020, pendapatan retribusi Uji KIR telah melebihi target setiap tahun. Realisasi pada tahun 2016 mencapai 116,37%; pada tahun 2017, naik menjadi 101,99%; dan pada tahun 2020, naik lagi menjadi 105,10%. Namun, pada tahun 2018, dan 2019 terjadi penurunan, dengan realisasi penerimaan yang tidak mencapai target masing-masing 97,47% dan 93,74%.

Dengan data mining, kita dapat melakukan prediksi terhadap pendapatan uji KIR, melalui metode crisp-dm diharapkan proses yang berjalan dimulai dari *business understanding* sampai dengan *evaluation* dapat dilaksanakan dengan baik agar dapat memprediksi pendapatan dari hasil Uji KIR ini. Tujuan dari penelitian ini adalah mengembangkan pemodalam data mining denagn crips-dm untuk menemukan informasi pada data pendapatan uji KIR dari tahun-tahun sebelumnya menggunakan algorithma c.45 dan dievaluasi dengan *confussion matrix*.

2. METODE PENELITIAN

Metodologi crisp-dm digunakan pada penelitian ini yang merupakan kepanjangan dari *Cross Industry Standard Process for Data Mining.* Methodologi ini meliputi enam Tahapan seperti yang ada pada Gambar 1.



Gambar 1. Metode CRISP-DM

1. Tahap Pemahaman Bisnis (*Business Understanding*)
   Tujuan penelitian adalah menemukan pola kenaikan dan penurunan pendapatan terhadap retribusi Uji KIR yang digunakan untuk melakukan prediksi pendapatan yang berasal dari Uji KIR pada tahun berikutnya. Tujuan dari data mining ini adalah untuk mengetahui bagaimana pola pendapatan retribusi Uji KIR berdasarkan jenis kendaraan terkait dengan prediksi pendapatan retribusi Uji KIR pada tahun berikutnya. Situasi bisnis ini menunjukkan bahwa pendapatan dari retribusi pengujian KIR setiap bulan berbeda, dipengaruhi oleh jenis kendaraan dan jumlah kendaraan yang melakukan pengujian.
2. Tahapan Pemahaman Data (*Data Understanding*)
   Untuk mengenal data yang akan digunakan, Tahapan ini melibatkan pengumpulan data awal dan pemahaman tentang data tersebut. Pengumpulan data dilakukan melalui observasi dan wawancara. Data ini dikumpulkan langsung dari bendahara pendapatan Kabupaten Karanganyar oleh Dinas Perhubungan. Data yang diminta berasal dari Uji KIR pendapatan retribusi dari tahun 2016 hingga 2020.
3. Tahapan Persiapan Data (*Data Preparation*)
   Tahapan ini melibatkan pengolahan data yang telah dikumpulkan dengan menyiapkan data, melakukan seleksi dengan menghapus data yang tidak relevan, dan menentukan karakteristik. Bulan, Minibus, Microbus, Bus, Pick Up, Light Truck, Truck, Mobil Baru, dan Status (naik/turun) adalah atribut yang digunakan dalam data pendapatan retribusi Uji KIR..
4. Tahapan Pemodelan (*Modelling*)
   Tahapan ini mencakup menentukan teknik data mining yang akan digunakan; algoritma yang digunakan adalah C4.5, dan alat yang digunakan adalah aplikasi *RapidMiner 5.3*.
5. Tahapan Evaluasi (*Evaluation*)
   Pada Tahapan ini dilakukan pengukuran tingkat akurasi dari model data yang dihasilkan pada Tahapan Pemodel
6. Tahapan Penyebaran (*Deployment*)

2.1 Tinjauan Pustaka

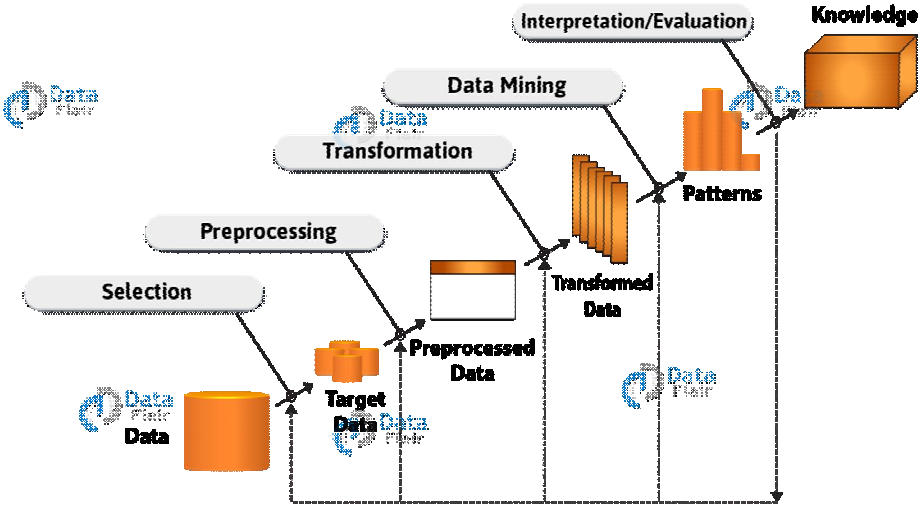
Dalam kelemahan sistem, penulis menggunakan analisis *PIECES* yang terdiri dari *performance, information, control, efficiency dan service* [5]. Sebagai alat ukur untuk menganalisa kelemahan sistem diperlukan langkah untuk mengidentifikasi dan memberikan solusi terhadap kelemahan sistem yang ada, seperti Tabel 1 berikut:

Penelitian mengenai CRISP-DM menunjukkan bahwa metode ini dapat diterapkan dalam penelitian data mining, dukungan data yang baik akan memberikan hasil penambagan data yang dapat memberikan informasi yang baik dan benar dalam pengambilan keputusan [1].

2.1.1 Data Mining

Data mining adalah suatu teknik yang digunakan untuk mengekstrak informasi bermanfaat dari gudang basis data yang sangat besar. Selain itu, proses ini juga dapat disebut sebagai pengekstrakan informasi baru dari kumpulan data yang sangat besar, yang membantu pengambilan keputusan. Data mining melibatkan penggunaan berbagai alat dan teknik analisis data untuk menemukan pola dan hubungan yang tersembunyi dalam data. Dalam proses pemecahan masalah dan pencarian pengetahuan baru, data mining dapat melakukan banyak hal, seperti klastering, klasifikasi, asosiasi, estimasi, dan prediksi [2].

Data mining, yang juga disebut sebagai penemuan pengetahuan dalam *database* (KDD), mencakup pengumpulan dan penggunaan data historis untuk menemukan pola, keteraturan, atau hubungan dalam kumpulan data yang sangat besar [3]. Data mining dapat membantu pengambilan keputusan di masa depan [4]. Gambar 2 memberikan penjelasan lebih lanjut tentang proses data mining.



Gambar 2. Proses pada Data Mining

Proses pemrosesan data terdiri dari beberapa tahapan, termasuk pembersihan data, integrasi data, seleksi data, transformasi data, proses pemrosesan, evaluasi pola, dan presentasi pengetahuan. Pembersihan data adalah proses untuk menghilangkan data yang tidak konsisten atau tidak relevan. Seleksi data dilakukan untuk memilih data yang sesuai untuk dianalisis atau cocok dengan data uji yang akan diambil dari *database*; transformasi data dilakukan untuk mengubah atau menggabungkan data ke dalam format yang sesuai untuk diproses dalam pengolahan data [5]. Menemukan pengetahuan penting dan tersembunyi dari data terdiri dari proses penggalian. Pola berbasis pengetahuan ditemukan melalui evaluasi pola. Pada tahap ini, hipotesis saat ini diuji dengan mengevaluasi model prediksi dan pola khas teknik data mining. Terakhir, Presentasi Pengetahuan menampilkan dan menyampaikan informasi tentang cara mendapatkan informasi untuk pengguna [6].

2.1.2 Prediksi

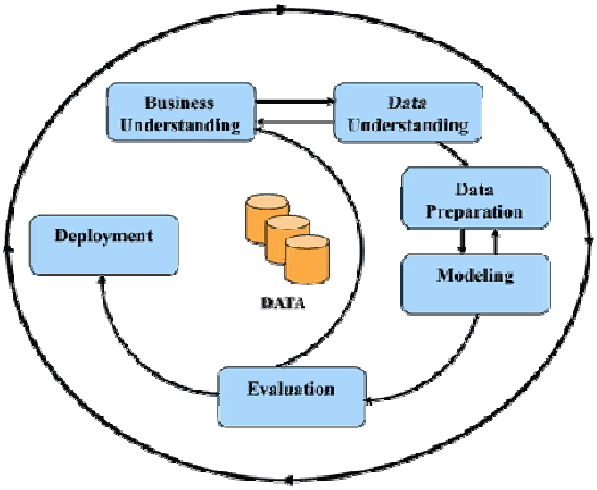
Pada dasarnya, peramalan atau prediksi adalah cara untuk memprediksi apa yang akan terjadi di masa depan. Menurut Kamus Besar Bahasa Indonesia (KBBI), prediksi adalah hasil dari kegiatan memprediksi, meramal, atau memperkirakan nilai pada masa yang akan datang dengan menggunakan data dari masa lalu. Prediksi adalah bagian dari proses perencanaan dan pengambilan keputusan dan menunjukkan apa yang akan terjadi dalam situasi tertentu. Prediksi

adalah proses sistematis untuk memperkirakan hasil yang paling mungkin berdasarkan informasi saat ini dan sebelumnya dengan tujuan mengurangi kesalahan dalam perkiraan [7]. Prediksi tidak selalu memberikan jawaban yang pasti tentang apa yang akan terjadi, tetapi mereka berusaha untuk mendapatkan jawaban yang paling dekat. Jika hasil peramalan terlalu tinggi atau rendah dibandingkan dengan kenyataan yang sebenarnya, hasil tersebut dikatakan biasa [8].

Selama kesalahan peramalan relatif kecil, hasil peramalan dianggap konsisten. Peramalan dapat dilakukan dalam dua cara: pertama, secara kualitatif—menggunakan pendapat seseorang; ini penting karena hasil peramalan ditentukan oleh pemikiran, pendapat, dan pengetahuan orang yang membuatnya; dan kedua, secara kuantitatif—menggunakan perhitungan angka menggunakan berbagai teknik statistik. Hasil peramalan sangat bergantung pada teknik yang digunakan [9].

2.1.3 CRISP-DM

Integral Solutions Ltd (ISL), Teradata, Daimler AG, NCR Corporation, dan OHRA adalah lima perusahaan yang membangun model proses data mining standar, atau CRISP-DM [10]. Model ini kemudian dikembangkan oleh ratusan organisasi dan perusahaan di Eropa untuk dijadikan metode standar *non-proprietary* untuk data mining. Versi pertama dari metode ini dipresentasikan pada 4th CRCR pada tahun 1996 [1]. Siklus hidup proyek data mining digambarkan oleh model proses CRISP-DM, yang terdiri dari enam tahapan: pemahaman bisnis, pemahaman data, persiapan data, modeling, evaluasi, dan penerapan [11] seperti yang terlihat pada gambar 3.



Gambar 3. Proses CRIPS-DM [12]

2.1.4 Algorithma C.45

Algoritma C4.5, yang merupakan evolusi dari ID3, digunakan untuk membuat pohon keputusan berdasarkan data latihan yang telah disediakan. Ini memiliki kemampuan untuk mengatasi data kontinu dan nilai yang hilang. Pohon keputusan adalah hasil dari perhitungan entropy dan pendapatan informasi yang dilakukan berulang kali sampai setiap atribut pohon memiliki kelas yang tidak dapat diperhitungkan lagi [13].

Algoritma C4.5 dipilih karena kemampuan untuk membuat subsistem model dasar yang dapat digunakan dalam sistem pendukung keputusan. Metode pohon keputusan sangat efektif dan terkenal dalam prediksi dan klasifikasi [14]. Setelah mengubah data yang rumit menjadi aturan yang dapat dipahami dengan mudah dalam bahasa alami, aturan tersebut juga dapat dikomunikasikan dalam bentuk bahasa basis data seperti Bahasa Pertanyaan Struktural untuk mencari *record* dalam kategori tertentu [15]. Ada beberapa tahap algoritma C4.5 yaitu:

1. Menyiapkan data. Data diambil dari data sebelumnya dan sudah dikelompokkan ke dalam kelas- kelas tertentu.

2. Menghitung nilai *entropy* dan gain.
   Berikut adalah persamaan untuk menghitung nilai *entropy*

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

   Keterangan Penamaan:
   S = himpunan kasus
   n = jumlah partisi S
   pi = proporsi dari S_i terhadap S

   Sedangkan untuk menghitung nilai *gain* itu sendiri dengan formula sebagai berikut:

$$Gain\ (S, A) = Entropy(S) - \sum_{i=1}^{n} \frac{|Si|}{|S|} x\ Entropy(Si) \quad (2)$$

   Keterangan Penamaan:
   S = himpunan kasus A A = atribut
   n = jumlah partisi atribut A
   $|S_i|$ = jumlah kasus pada partisi ke-i
   $|S|$ = jumlah kasus dalam S
3. Penentuan akar pohon. Akar diambil dari atribut yang terpilih, dengan nilai *gain* yang paling tinggi.
4. Mengulangi proses di atas sampai semua kasus memiliki kelas yang sama. Proses partisi pohon keputusan akan berhenti saat : Semua *record* dalam simpul berada didalam satu kelas yang sama dan atau Tidak ada atribut di dalam *record* yang dipartisi lagi.

3. HASIL DAN PEMBAHASAN

Metode yang digunakan dalam penelitian ini adalah metode CRIPS-DM yang meliputi enam Tahapan:

3.1 Tahapan Pemahaman Bisnis (Business Understanding)

Data pendapatan retribusi Uji KIR dari tahun 2016 hingga 2020 menunjukkan perbedaan antara hasil dan tujuan. Penghitungan target masih dilakukan secara manual saat ini. Bisnis seperti ini membutuhkan teknik yang dapat membantu memprediksi pendapatan agar tujuan dan hasil dapat sesuai. Tujuan perusahaan ini adalah untuk mengetahui pola pendapatan untuk memperkirakan pendapatan tahun berikutnya. Studi ini berfokus pada pola pendapatan dari jenis kendaraan yang dipengaruhi oleh jenis dan jumlah kendaraan yang melakukan uji KIR. Tujuan data mining ini adalah untuk mengetahui bagaimana pola pendapatan dari jenis kendaraan ini berhubungan dengan prediksi pendapatan tahun berikutnya. Untuk memulai, penelitian akan diajukan ke Dinas Perhubungan Kabupaten Karanganyar. Selanjutnya, data akan dikumpulkan dari Dinas tersebut, diproses, dan dianalisis hasilnya. Studi ini direncanakan dimulai dari Maret hingga Juni 2021.

3.2 Tahapan Pemahaman Data (Data Understanding)

Bapak Sri Suboko, S.Sos., M.Si., Kepala Dinas Perhubungan Kabupaten Karanganyar, telah diwawancarai. Tujuan wawancara ini adalah untuk mengajukan pertanyaan lisan yang berkaitan dengan penelitian yang akan dilakukan untuk mendukung penelitian dan menjawab masalah yang ada. Data pendapatan retribusi Uji KIR selama lima tahun periode 2016–2020 dikumpulkan melalui wawancara dan observasi. Data ini dikumpulkan langsung dari Dinas

Perhubungan Kabupaten Karanganyar melalui Ibu Eni Hastuti, bendahara pendapatan. Data yang diminta berasal dari *dataset* "Pendapatan Retribusi Uji KIR.xlsx", yang berisi data pendapatan retribusi Uji KIR tahun 2016–2020.

3.3 Tahapan Persiapan Data (Data Preparation)

Pada tahap ini, proses seleksi data dilakukan dengan menghapus data yang tidak diperlukan, seperti atribut Total. Hal ini dilakukan karena atribut tersebut tidak berdampak pada pengolahan data yang akan datang. Bulan, Minibus, Mocrobus, Bus, Pick Up, Light Truck, Truck, Mobil Baru, dan Status adalah attribut yang digunakan. Status naik dan turun diperoleh dari perbandingan data bulan sebelumnya.

3.4 Tahapan Pemodelan (Modelling)

Untuk tujuan penelitian ini, data Pendapatan Retribusi Uji KIR telah dipilih, seperti yang ditunjukkan dalam Lampiran 2. Data yang sudah dikategorikan dapat dilihat pada Tabel 3.1. Setiap atribut yang memiliki nilai numerik disusun dalam bentuk kategori sesuai dengan peraturan yang ditetapkan oleh Badan Keuangan Daerah Kabupaten Karanganyar, dengan nilai berikut:

Sangat Kecil = Pendapatan kurang dari Rp. 10.000.000,-

Kecil = Pendapatan antara Rp. 10.000.000 s/d Rp. 20.000.000

Sedang = Pendapatan antara Rp. 20.000.001 s/d Rp. 30.000.000

Besar = Pendapatan antara Rp. 30.000.001 s/d Rp. 40.000.000

Sangat Besar = Pendapatan lebih dari Rp. 40.000.000

Tabel 1. Kategori Data

No	Bulan	Minibus	Microbus	Bus	Pick Up	Light Truck	Truck	Mobil Baru	Status
1	Januari 2016	Sangat Kecil	Sangat Kecil	Sangat Kecil	Sangat Besar	Sedang	Sangat Kecil	Sangat Kecil	Turun
.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.
60	Desember 2020	Sangat Kecil	Sangat Kecil	Sangat Kecil	Besar	Kecil	Sangat Kecil	Sangat Kecil	Naik

Menghitung *Nilai Entropy dan Gain*

Dari data yang ada di Tabel 1, diketahui jumlah kasus (n) sebanyak 60, dengan status “Naik” 2sebanyak 9 data dan “Turun” sebanyak 32 data

Sehingga didapat nilai *entropy* total:

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

$$= \left(\frac{-\text{Jml Naik Total}}{\text{Jml Kasus Total}} * \log_2\left(\frac{\text{Jml Naik Total}}{\text{Jml Kasus Total}}\right)\right) + \left(\frac{-\text{Jml Turun}}{\text{Jml Kasus Total}} * \log_2\left(\frac{\text{Jml Turun Total}}{\text{Jml Kasus Total}}\right)\right)$$

$$= \left(\frac{-29}{60} * \log_2\left(\frac{29}{60}\right)\right) + \left(\frac{-31}{60} * \log_2\left(\frac{31}{60}\right)\right)$$

$$= 0{,}50697 + 0{,}49222$$

= 0,9992

Langkah selanjutnya, menghitung nilai *entropy* dari setiap atribut:

Entropy Minibus SK (S) $= \sum_{i=1}^{n} -$ pi * $\log_2 pi$

$= \left(\frac{-\text{Jml Naik Minibus SK}}{\text{Jml Kasus Minibus SK}} * \log_2\left(\frac{\text{Jml Naik Minibus SK}}{\text{Jml Kasus Minibus SK}}\right)\right)$
$+ \left(\frac{-\text{Jml Turun Minibus SK}}{\text{Jml Kasus Minibus SK}} * \log_2\left(\frac{\text{Jml Turun Minibus SK}}{\text{Jml Kasus Minibus SK}}\right)\right)$
$= \left(\frac{-29}{60} * \log_2\left(\frac{29}{60}\right)\right) + \left(\frac{-31}{60} * \log_2\left(\frac{31}{60}\right)\right)$
= 0,50697 + 0,49222
= 0,9992

Dan seterusnya setiap atribut dihitung nilai entrpoynya

Setelah mendapatkan nilai *entropy* dari setiap atribut di atas, berikutnya menghitung nilai *gain*.

Gain Minibus (S,A) = *Entropy*(S) - $\sum_{i=1}^{n}$ x *Entropy*(Si)

= *Entropy* Total - $\left(\left(\frac{|Si|}{\text{Jml Kasus Total}}\right)\right.$ *Entropy* Minibus SK)

$= (0{,}9992) - \left(\left(\frac{60}{60}\right) * 0{,}9992\right)$
= 0,9992 – 0,9992
= 0

Hasil perhitungan nilai *entropy* dan *gain* pada Tabel 2 berikut ini:

Tabel 2 Perhitungan Node

Data Jenis Kendaraan/Kontribusi Retribusi		Naik (S_1)	Turun (S_2)	Jml Kasus (S)	Entropy	Gain
Minibus	Sangat Kecil	29	60	31	0,99	0
Microbus	Sangat Kecil	29	60	31	0,99	0
Bus	Sangat Kecil	29	60	31	0,99	0
Pick Up	Sedang	1	6	5	0,65	0,08
	Besar	2	9	7	0,76	
	Sangat Besar	26	45	19	0,98	
Light Truck	Kecil	2	16	14	0,54	0,16
	Sedang	26	43	17	0,96	
	Besar	1	1	0	0	
Truck	Sangat Kecil	29	60	31	0,99	0
Mobil Baru	Sangat Kecil	28	59	31	0,99	0,018
	Kecil	1	1	0	0	

Lakukan pengulangan sampai kondisi kasus memiliki kelas yang sama atau semua *record* dalam simpul berada didalam satu kelas yang sama dan atau tidak ada atribut di dalam *record* yang dipartisi lagi.

3.5 Tahapan Evaluasi (Evaluation)

Pada tahap dalam penelitian evaluasi ini dilakukan dengan mengukur tingkat akurasi dari algoritma C4.5 saat memodelkan data untuk prediksi pendapatan. Pengukuran data dilakukan dengan Confusion Matrix,

3.6 Tahapan Penyebaran (Deployment)

Pada tahap ini, pengetahuan atau informasi yang telah diperoleh akan disusun dalam bentuk laporan. Hasil dari penelitian ini berupa prediksi naik atau turunnya pendapatan retribusi Uji KIR pada Dinas Perhubungan Kabupaten Karanganyar yang diharapkan dapat digunakan oleh instansi sebagai bahan pertimbangan dalam menentukan target pendapatan pada tahun yang akan datang.

Hasil dari proses perhitungan algoritma C4.5 menunjukkan bahwa atribut Light Truck terpilih sebagai akar pohon dan prediksi yang dihasilkan adalah naik. Hasil pengujian yang diperoleh dari perhitungan RapidMiner, dengan melakukan pengukuran terhadap 60 data menggunakan *split validation*, dimana sebanyak 48 sebagai data latih dan 12 sebagai data uji dengan rasio partisi 80% dan 20% serta pengambilan sampel secara acak, menghasilkan nilai *Accuracy*, *Precision* dan *Recall* dengan persamaan *confusion matrix* seperti pada Tabel 3 sebagai berikut:

Tabel 4.1 Hasil *Confusion Matrix*

		Aktual	
Prediksi		Turun	Naik
	Turun	3	1
	Naik	2	6

Pada tabel di atas dari 12 kendaraan akan diprediksi apakah pendapatannya akan turun atau naik, dengan 5 kendaraan turun pendapatannya dan 7 kendaraan naik pendapatannya.

a. Ada 3 kendaraan yang diprediksi turun pendapatannya dan memang benar bahwa kendaraan tersebut turun pendapatannya.
b. Ada 2 kendaraan yang diprediksi naik pendapatannya tetapi kenyataannya kendaraan tersebut turun pendapatannya.
c. Ada 1 kendaraan yang diprediksi turun pendapatannya tetapi kenyataannya kendaraan tersebut naik pendapatannya.
d. Ada 6 kendaraan yang diprediksi naik pendapatannya dan memang benar bahwa kendaraan tersebut naik pendapatannya.

Nilai *Accuracy*, *Precision* dan *Recall* dari Tabel 4.1 diperoleh dengan persamaan sebagai berikut:

a. *Accuracy* menggambarkan seberapa akurat model dapat mengklasifikasikan dengan benar. Nilai *accuracy* diperoleh dengan persamaan:

$$Accuracy = \frac{\text{Jml kendaraan diprediksi benar (turun+naik)}}{\text{Jml kendaraan keseluruhan}} *100\%$$

$$= \frac{3+6}{3+2+1+6} *100\%$$

$$= \frac{9}{12} *100\%$$

$= \frac{}{12}$ *100% = 75%

b. *Precision* dapat menggambarkan seberapa akurat hubungan antara data yang diminta dengan hasil prediksi yang diberikan oleh model. Nilai *precicion* diperoleh dengan persamaan:

$$Precision = \frac{\text{Jml kendaraan naik diprediksibenar}}{\text{Jml kendaraan diprediksi naik}} *100\%$$

$$= \frac{6}{6+2} *100\%$$

$$= \frac{6}{8} *100\% = 75\%$$

c. *Recall* menggambarkan keberhasilan model dalam menemukan kembali sebuah informasi. Nilai *recall* diperoleh dengan persamaan:

$$Recall = \frac{\text{Jml kendaraan naik diprediksi benar}}{\text{Jml kendaraan naik}} *100\%$$

$$= \frac{6}{6+1} *100\%$$

$$= \frac{6}{7} *100\% = 85{,}71\%$$

4. KESIMPULAN

Kesimpulan penelitian menggunakan metodologi CRISP-DM pada data mining dapat dicapai melalui rangkuman temuan, evaluasi proses, dan implikasi praktis dari hasil analisis. Crisp-dm dengan 6 tahapannya dapat memingkatkan kemampuan untuk melakukan prediksi pendapatan dari pengujian kendaraan. Hasil prediksi yang diperoleh dari perhitungan *algoritma C4.5*, bahwa pendapatan retribusi Uji KIR untuk tahun 2021 adalah naik. Jenis kendaraan yang memiliki pengaruh besar dalam memprediksi kenaikan pendapatan retribusi Uji KIR adalah Light Truck, Pick Up, Mobil Baru dan Minibus. Nilai akurasi menggunakan confussion matrix mendapatkan tingkat akurasi sebesar 75%. Algoritma C4.5 dapat membantu dalam memprediksi pendapatan Uji KIR dengan cukup baik.

5. SARAN

Untuk saran penelitian selanjutnya, implementasi CRISP-DM dapat diterapkan pada algoritm data mining lainnya seperti Fuzzy, K-Means dan lainnya. Selain itu implementasi Crisp-dm juga diterapkan pada kegiatan data mining lainnya tidak hanya untuk melakukan prediksi, seperti melakukan klasifikasi, klastering dan lainnya.

DAFTAR PUSTAKA

[1] C. Schröer, F. Kruse, and J. M. Gómez, “A systematic literature review on applying CRISP-DM process model,” *Procedia Comput. Sci.*, vol. 181, no. 2019, pp. 526–534, 2021, doi: 10.1016/j.procs.2021.01.199.

[2] M. S. Brown, “(For Dummies) Meta S. Brown-Data Mining For Dummies-Wiley Publishing Inc. (2014).pdf.” 2014. [Online]. Available: www.wiley.com

[3] D. Zhu *et al.*, “A Cluster Separation Measure,” *Procedia Comput. Sci.*, vol. 2, no. 1, pp. 1–6, 2016, doi: 10.1016/j.procs.2016.09.180.

[4] M. A. A. Riyadi and K. Fithriasari, “Data Mining Peramalan Konsumsi Listrik dengan Pendekatan Cluster Time Series sebagai Preprocessing,” *J. Pengemb. Teknol. Inf. dan*

ilmu Komput., vol. 2, no. 4, pp. 1–6, 2016.

[5] K. Englmeier, "The role of text mining in mitigating the threats from fake news and misinformation in times of corona," *Procedia Comput. Sci.*, vol. 181, no. 2019, pp. 149–156, 2021, doi: 10.1016/j.procs.2021.01.115.

[6] A. A. Az-zahra, A. F. Marsaoly, I. P. Lestyani, R. Salsabila, and W. O. Z. Madjida, "Penerapan Algoritma K-Modes Clustering Dengan Validasi Davies Bouldin Index Pada Pengelompokkan Tingkat Minat Belanja Online Di Provinsi Daerah Istimewa Yogyakarta," *J. MSA (Mat. dan Stat. serta Apl.)*, vol. 9, no. 1, p. 24, 2021, doi: 10.24252/msa.v9i1.18555.

[7] J. M. Raimundo and P. Cabrita, "Artificial intelligence at assisted reproductive technology," *Procedia Comput. Sci.*, vol. 181, pp. 442–447, 2021, doi: 10.1016/j.procs.2021.01.189.

[8] L. M. A. da Costa, F. A. Bernardi, T. L. M. Sanches, A. Kritski, R. M. Galliez, and D. Alves, "Operational modeling for testing diagnostic tools impact on tuberculosis diagnostic cascade: A model design," *Procedia Comput. Sci.*, vol. 181, no. 2019, pp. 650–657, 2021, doi: 10.1016/j.procs.2021.01.214.

[9] C. Bergmeir, R. J. Hyndman, and B. Koo, "A note on the validity of cross-validation for evaluating autoregressive time series prediction," *Comput. Stat. Data Anal.*, vol. 120, pp. 70–83, 2018, doi: 10.1016/j.csda.2017.11.003.

[10] D. Feblian and D. U. Daihani, "Implementasi Model Crisp-Dm Untuk Menentukan Sales Pipeline Pada Pt X," *J. Tek. Ind.*, vol. 6, no. 1, 2017, doi: 10.25105/jti.v6i1.1526.

[11] B. E. Adiana, I. Soesanti, A. E. Permanasari, J. G. No, J. G. No, and J. G. No, "Analisis Segmentasi Pelanggan Menggunakan Kombinasi RFM Model dan Teknik Clustering," no. 2, pp. 23–32, 2018, doi: 10.21460/jutei.2017.21.76.

[12] Y. Suhanda, I. Kurniati, and S. Norma, "Penerapan Metode Crisp-DM Dengan Algoritma K-Means Clustering Untuk Segmentasi Mahasiswa Berdasarkan Kualitas Akademik," *J. Teknol. Inform. dan Komput.*, vol. 6, no. 2, pp. 12–20, 2020, doi: 10.37012/jtik.v6i2.299.

[13] T. W. Liao, *Recent Advances in Data Mining of Enterprise Data: Algorithma Aplications*. 2007.

[14] T. B. Santoso and D. Sekardiana, "Penerapan Algoritma C4.5 untuk Penentuan Kelayakan Pemberian Kredit," *J. Algoritm. Log. dan Komputasi*, vol. II, no. 1, pp. 130–137, 2019, [Online]. Available: https://journal.ubm.ac.id/index.php/alu

[15] J. D. Rodr??guez, A. P??rez, and J. A. Lozano, "A general framework for the statistical analysis of the sources of variance for classification error estimators," *Pattern Recognit.*, vol. 46, no. 3, pp. 855–864, 2013, doi: 10.1016/j.patcog.2012.09.007.