SIGMA
Jurnal Teknologi Pelita Bangsa

Vol. 11 No. 3 September 2020 ISSN: 2407-3903

Diterima, 16 Juli 2020 | Direvisi, 28 Agustus 2020 | Dipublikasikan, 28 September 2020

MENENTUKAN PREDIKSI KELULUSAN SISWA DENGAN MEMBANDINGKAN ALGORITMA C4.5 DAN NAIVE BAYES STUDI KASUS SMKN. 1 CIKARANG SELATAN

Muhammad Makmun Effendi[1], Arie Setiawan[2]

Program Studi Teknik Informatika Universitas Pelita Bangsa
[1]effendiyan@pelitabanngsa.ac.id

Abstraksi

Proses mengidentifikasi informasi dengan menggunakan teknik statistik, dan machine learning merupakan pengertian dari data mining. Data mining dapat diterapkan diberbagai bidang kehidupan seperti bidang kesehatan, bisnis, dan pendidikan. Salah satu penerapan data mining pada bidang pendidikan seperti untuk memprediksi kelulusan siswa sekolah. Prediksi kelulusan siswa ini menggunakan data yang berasal dari transkip nilai akhir dari masing-masing siswa, adapun atribut yang digunakan yakni nilai rata-rata pelajaran Bahasa Indonesia, Bahasa Inggris, dan Matematika mulai dari semester 1 hingga semester 5 serta riwayat SP yang pernah didapat selama siswa tersebut sekolah. Pada penelitian ini menggunakan dua metode data mining, yaitu algoritma C4.5 dan Naïve bayes . Penggunaan dua metode pada penelitian ini bertujuan untuk membandingkan kinerja dari kedua algoritma dalam memprediksi kelulusan siswa berdasarkan tingkat accuracy, precision, dan recall yang didapatkan. Dari hasil pengujian dengan menggunakan data testing sebanyak 222 data yang menyatakan algoritma C4.5 memiliki tingkat nilai accuracy 98,64%, precision 100% dan recall 100% sedangkan naïve bayes memiliki tingkat accuracy 97,75%, precision 95,52% dan recall 95,52%. Dan jika pengujian menggunakan data training sebanyak 890 data maka akan menyatakan algoritma C4.5 memiliki tingkat nilai accuracy 98,99%, precision 98,68% dan recall 98,68% sedangkan naïve bayes memiliki tingkat accuracy 97,42%, precision 99,39% dan recall 99,39%. Dari perbandingan diatas algoritma C4.5 memiliki nilai tingkat accuracy yang cenderung lebih tinggi dibandingkan dengan algoritma naïve bayes , maka di putuskan dalam memprediksi kelulusan siswa, algoritmaC4,5 lebih baik dibanding algoritma naïve bayes dalam memprediksi data kelulusan siswa.

Kata Kunci: *Data mining, Clacification, C4.5 Algoritma, Naive Bayes*

Abstract

The process of identifying information using statistical techniques, and machine learning is the meaning of data mining. Data mining can be applied in various fields of life such as health, business, and education. One of the applications of data mining in the field of education is to predict the graduation of school students. Prediction of student graduation using data derived from transcripts of the final grades of each student, while the attributes used are the average value of Indonesian, English, and Mathematics lessons from semester 1 to semester 5 as well as the history of SP that has been obtained during the student is in school. In this study, two data mining methods are used, namely the C4.5 algorithm and Naïve Bayes algorithm. The use of the two methods in this study aims to compare the performance of the two algorithms in predicting student graduation based on the level of accuracy, precision, and recall obtained. From the test results using data testing as much as 222 data which states that the C4.5 algorithm has an accuracy value of 98.64%, 100% precision and 100% recall, while Nave Bayes has an accuracy level of 97.75%, precision 95.52% and recall. 95.52%. And if the test uses 890 training data, it will state that the C4.5 algorithm has an accuracy level of 98.99%, precision 98.68% and recall 98.68% while nave Bayes has an accuracy level of 97.42%, precision 99, 39% and recalls 99.39%. From the above comparison, the C4.5 algorithm has an accuracy level that tends to

be higher than the nave Bayes algorithm, so it was decided that in predicting student graduation, the C4.5 algorithm is better than the nave Bayes algorithm in predicting student graduation data.

Keywords**:** *Data mining, Clacification, C4.5 Algoritma, Naive Bayes*

1. Pendahuluan

Seiring berkembangnya teknologi bidang sosial, bidang ekonomi, bidang informasi, kebutuhan informasi yang akurat transportasi, bidang pemerintahan, dan juga sudah seperti kebutuhan pokok bagi manusia bidang pendidikan. tanpa sadar setiap dalam kehidupan sehari-hari. hampir segala saatnya manusia menghasilkan banyak data dari kegiatan mereka sehari-hari. namun dibalik banyaknya data tersebut tidak diimbangi dengan pengetahuan yang memadai, alhasil data tersebut tidak memiliki nilai manfaat bagi manusia. bidang pendidikan di indonesia masih merupakan aspek penting dalam kehidupan masyarakat. di indonesia memberlakukan wajib belajar 9 tahun atau diartikan dari jenjang sekolah dasar (sd) hingga sekolah menengah pertama (smp), setelah lulus dari jenjang smp biasanya para siswa melanjutkan ke jenjang berikutnya yakni sekolah menengah atas (sma). dalam sebuah instansi sekolah, terdapat banyak data dari para siswa, baik data administrasi, akademik, biodata siswa, dan data lainnya, namun data tersebut hanya menjadi arsip berkas milik sekolah. alhasil data tersebut tidak bisa menjadi informasi yang dapat dimanfaatkan baik untuk guru, maupun untuk para siswa. proses mengidentifikasi informasi dengan menggunakan teknik statistik, dan machine learning merupakan pengertian dari data mining. data mining dapat diterapkan kedalam bidang pendidikan, salah satunya untuk memprediksi kelulusan siswa. prediksi kelulusan siswa ini menggunakan data yang berasal dari transkip nilai akhir dari masing-masing siswa. ada beberapa metode data mining yang biasa digunakan untuk menentukan prediksi kelulusan siswa sekolah, antara lain algoritma c4.5, metode k-nearest neighbor (k-nn), dan metode naïve bayes.

2. Landasan Pemikiran

Data Mining adalah kegiatan menemukan pola yang menarik dari data dalam jumlah besar, data dapat disimpan dalam *database*, *data warehouse*, atau penyimpanan informasi lainnya. *Data Mining* berkaitan dengan bidang ilmu-ilmu lain, seperti *database system*, *data warehousing*, statistik, *machine learning*, *information retrieval*, dan komputasi tingkat tinggi. Selain itu, *Data Mining* didukung oleh ilmu lain seperti *neural network*, pengenalan pola, *spatial data analysis*, *image database*, *signal processing* [1].

Naive bayes merupakan pengklasifikasian dengan probabilitas dan statistik yang dikemukakan oleh ilmuwan Inggris Thomas Bayes yaitu memprediksi peluang dimasa depan berdasarkan pengalaman dimasa sebelumnya [1] pengertian lain dari *naive bayes* yaitu sebuah pengklasifikasian probabilistik sederhana yang menghitung sekumpulan probabilitas dengan menjumlahkan frekuensi dan kombinasi nilai dari dataset yang diberikan, algoritma menggunakan *teorema bayes* dan mengasumsikan semua atribut independen atau tidak saling ketergantungan yang diberikan oleh nilai pada variabel kelas [6].

Algoritma C4.5 adalah algoritma yang sudah banyak dikenal dan digunakan untuk klasifikasi data yang memiliki atribut-atribut numerik dan kategorial. Hasil dari proses klasifikasi yang berupa aturan-aturan dapat digunakan untuk memprediksi nilai atribut bertipe diskret dari record yang baru. Algortima C4.5 sendiri merupakan pengembangan dari algortima ID3, dimana pengembangan dilakukan dalam hal, bisa mengatasi missing data, bisa mengatasi data kontinu dan *pruning* [4].

RapidMiner adalah salah satu software untuk pengolahan *Data Mining*. Pekerjaan yang dilakukan oleh *RapidMiner text mining* adalah berkisar dengan analisis teks, mengekstrak pola-pola dari data set yang besar dan mengkombinasikannya dengan metode statistika, kecerdasan buatan, dan *database*. Tujuan dari analisis teks ini adalah untuk mendapatkan informasi bermutu tertinggi dari teks yang diolah [8].

Salah satu klasifikasi tugas yang dapat dilakukan dengan *Data Mining* adalah pengklasifikasian. Klasifikasi pertama kali diterapkan pada bidang tanaman yang mengklasifikasi suatu spesies tertentu, seperti yang dilakukan oleh Carolus von Linne (atau dikenal dengan nama Carolus Linnaeus) yang pertama kali mengklasifikasi spesies berdasarkan karakteristik fisik. Selanjutnya dia dikenal sebagai bapak klasifikasi [3]

Dalam klasifikasi terdapat target variabel kategori. Metode-metode / model-model yag telah dikembangkan oleh periset untuk menyelesaikan kasus klasifikasi antara lain [3] :

- Pohon keputusan
- Pengklasifikasi bayes / *naive bayes*
- Jaringan saraf tiruan
- Analisis statistik
- Algoritma genetik
- *Rough sets*
- Pengklasifikasi *k-nearest neighbour*
- Metode berbasis aturan
- *Memory based learning*
- *Support vector machine*

Diantara beberapa metode yang dapat digunakan untuk klasifikasi adalah metode pohon keputusan atau *decision tree*. Metode pohon keputusan merupakan sebuah metode yang dapat mengubah fakta yang sangat besar menjadi sebuah pohon keputusan yang merepresentasikan aturan. Aturan dapat dengan mudah dipahami dengan bahasa alami [3].

3. Metode Penelitian

3.1 Jenis Data

Adapun jenis data yang digunakan pada penelitian ini adalah data yang bertipe atribut nominal. Atribut

nominal adalah atribut yang nilainya tidak memiliki urutan, nilai yang dikandung berupa simbol-simbol atau nama kelas yang menunjukkan nama sampel [1].

3.2 Data Yang Di Gunakan

Penelitian ini menggunakan data internal sekolah yang diperoleh penulis melalui bagian tata usaha pada SMK Negeri 1 Cikarang Selatan. Data yang digunakan pada penelitian ini merupakan data nilai rata-rata rapot dari mata pelajaran yang ditentukan mulai dari semester 1 hingga semester 5. Untuk data transkip nilai siswa diambil dari angkatan kelulusan 3 tahun terakhir yakni angkatan tahun kelulusan 2016, 2017, dan 2018, setelah data dikumpulkan diperoleh data set sebanyak 1.112 data. Adapun data yang dijadikan data *training* adalah sebanyak 890 *record* data, sedangkan data *testing* sebanyak 222 data. Adapun atribut yang digunakan pada penelitian ini dijelaskan pada tabel berikut.

Tabel 3.1 Atribut yang di gunakan

No	Atribut	Keterangan
1	Nilai Rata-rata Bahasa Indonesia	Nilai Rata-rata mata pelajaran Bahasa Indonesia mulai dari semester 1 hingga 5
2	Nilai Rata-rata Bahasa Inggris	Nilai Rata-rata mata pelajaran Bahasa Inggris mulai dari semester 1 hingga 5
3	Nilai Rata-rata Matematika	Nilai Rata-rata mata pelajaran Matematika mulai dari semester 1 hingga 5
4	Riwayat SP	Data riwayat Sp (Surat peringatan) Yang pernah di dapat oleh siswa berdasarkan pelanggaran yang di lakukan oleh siswa

3.3 Metode Yang Di Gunakan

1. Algoritma C4.5

Algoritma C4.5 adalah algoritma yang sudah banyak dikenal dan digunakan untuk klasifikasi data yang memiliki atribut-atribut numerik dan kategorial. Hasil dari proses klasifikasi yang berupa aturan-aturan dapat digunakan untuk memprediksi nilai atribut bertipe diskret dari record yang baru. Algortima C4.5 sendiri merupakan pengembangan dari algortima ID3, dimana pengembangan dilakukan dalam hal, bisa mengatasi missing data, bisa mengatasi data kontinu dan pruning [2].

2. Naive Bayes

Naïve bayes Clasifier merupakan salah satu algoritma pemecahan yang termasuk kedalam metode klasifikasi pada data mining. *Naïve bayes Clasifier* mengadopsi ilmu statistika yaitu dengan menggunakan teori kemungkinan (Probabilitas) untuk menyelesaikan sebuah kasus *Supervised Learning*, artinya dalam himpunan data terdapat label, Class atau Target sebagai acuan atau gurunya [4].

3. Pengujian

Pengujian dilakukan untuk mengetahui hasil perhitungan dan juga mengetahui apakah fungsi bekerja dengan baik atau tidak. Pada penelitian ini proses pengujian dilakukan melalui 2 langkah, yakni dengan melakukan perhitungan secara manual kemudian data diuji menggunakan *software tools RapidMiner* agar memastikan bahwa hasil perhitungan secara manual sesuai atau tidak dengan hasil yang diperoleh dengan *tools RapidMiner*.

4. Evaluasi Dan Validasi Hasil

Evaluasi dilakukan dengan cara menganalisa hasil dari masing-masing algoritma yang digunakan, untuk memastikan bahwa hasil dari perhitungan dan pengujian sesuai dengan tujuan. Sedangkan validasi dilakukan dengan mengukur hasil prediksi dari masing-masing algoritma untuk mengetahui tingkat accuracy, precision, dan recall, dengan itu dapat dilihat algoritma mana yang memiliki tingkat akurasi lebih baik pada proses penelitian ini.

4. Pembahasan

4.1 Hasil

Dari hasil yang ada kemudian data dikategorikan dengan variable, atribut kemudian dijadikan data training sebanyak 17 data dan data testing sebanyak 220 data. Dari proses tersebut kemudian membandingkan tingkat accuracy dan precision dari algoritma naïve bayes dengan C4.5 dalam memprediksi kelulusan siswa. Adapun pada Tabel 4.1 menjelaskan mengenai sampel dari isi data training, dan pada Tabel 4.2 menjelaskan mengenai sampel dari isi data testing.

Tabel 4.1 Data

No	Rata-Rata B.Indonesia	Rata-Rata B.Inggris	Rata-Rata Matematika	Riwayat SP	Keterangan Lulus
1	80	80	85	Tidak Ada	Tepat Waktu
2	80	80	80	Tidak Ada	Tepat Waktu
3	80	80	80	Tidak	Tepat

				Ada	Waktu
4	80	80	80	Tidak Ada	Tepat Waktu
5	80	80	80	Tidak Ada	Tepat Waktu
6	80	80	80	Tidak Ada	Tepat Waktu
7	80	80	80	Tidak Ada	Tepat Waktu
8	80	80	80	Tidak Ada	Tepat Waktu
9	80	80	80	Tidak Ada	Tepat Waktu
10	80	80	80	Tidak Ada	Tepat Waktu
11	80	80	80	Tidak Ada	Tepat Waktu
12	85	80	80	Tidak Ada	Tepat Waktu
13	80	80	85	Tidak Ada	Tepat Waktu
14	80	80	80	Tidak Ada	Tepat Waktu
15	80	85	85	Tidak Ada	Tepat Waktu
16	80	80	80	Tidak Ada	Tepat Waktu
17	80	80	80	Tidak Ada	Tepat Waktu

Tabel 4.2 Data Testing

No	Rata-Rata B.Indonesia	Rata-Rata B.Inggris	Rata-Rata Matematika	Riwayat SP	Keterangan Lulus
1	80	80	75	SP1	Tidak Tepat Waktu
2	80	75	80	Tidak Ada	Tidak Tepat Waktu
3	75	80	75	Tidak Ada	Tidak Tepat Waktu
4	80	80	80	Tidak Ada	Tepat Waktu
5	80	80	80	Tidak Ada	Tepat Waktu
6	75	75	75	Tidak Ada	Tidak Tepat Waktu
7	80	80	80	Tidak Ada	Tepat Waktu
8	75	80	80	Tidak Ada	Tidak Tepat Waktu
9	75	75	80	Tidak Ada	Tidak Tepat Waktu
10	75	80	80	Tidak Ada	Tidak Tepat Waktu
11	80	80	80	Tidak Ada	Tepat Waktu
12	80	80	80	Tidak Ada	Tepat Waktu
13	80	80	80	Tidak Ada	Tepat Waktu
14	80	80	80	Tidak Ada	Tepat Waktu
15	80	80	75	Tidak Ada	Tidak Tepat Waktu
16	80	80	80	Tidak Ada	Tepat Waktu
17	80	80	80	Tidak Ada	Tepat Waktu

Data tersebut disimpan dalam format excel workbook yang selanjutnya diubah menjadi data.frame dengan perintah read.excel. Sebelum melakukan uji coba dengan menggunakan tools Rapid miner, penulis

melakukan perhitungan manual terhadap data uji atau data testing dengan menggunakan rumus dari algoritma C4.5 dan algoritma Naïve bayes
.

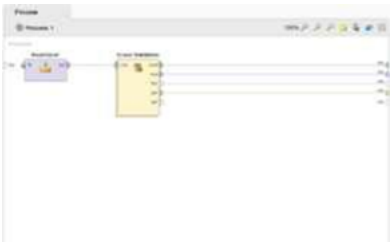
Tabel 4.3 Perhitungan data uji coba dengan algoritma c4.5

C45						
Tahap 1						
Langkah		Jumlah Kasus	Tepat Waktu	Tidak Tepat Waktu	Entropi	Informasi Gain
	Total	222	156	66	0,877962001	
Rata B.Indonesia	90	0	0	0	0	0,223178983
	85	1	0	1	0	
	80	197	156	41	0,73787731	
	75	24	0	24	0	
	70	0	0	0	0	
Rata B.Inggris	90	0	0	0	0	0,128745185
	85	1	1	0	0	
	80	206	155	51	0,807408414	
	75	15	0	15	0	
	70	0	0	0	0	
Rata Matematika	90	0	0	0	0	0,320574803
	85	0	0	0	0	
	80	188	156	32	0,658191266	
	75	34	0	34	0	
	70	0	0	0	0	
Riwayat SP	Tidak Ada	215	155	60	0,854180205	0,038306003
	SP1	3	1	2	0,918295834	
	SP2	3	0	3	0	
	SP3	1	0	1	0	

4.2 Pembahasan

Langkah kedua, dilakukan implementasi algoritma naïve bayes dan C4.5 dengan menggunakan tools rapidminer, Berikut adalah tahapan dalam penerapan algoritma naïve bayes dan C.45:

1. Menentukan pohon keputusan
2. Menentukan nilai accuracy, recall dan precision
3. Menentukan performance



Gambar 4.1 Pengajuan tools

Dari implementasi pengujian pada tools Rapid miner maka terbentuk simpul-simpul diperoleh decision tree untuk klasifikasi prediksi kelulusan siswa seperti yang diperlihatkan pada Gambar berikut :

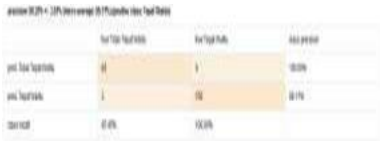
Dataset yang digunakan berupa dataset berformat excel workbook yang kemudian dibaca dengan menggunakan fungsi Read Excel pada Rapid miner seperti yang diperlihatkan pada Gambar berikut.

Gambar 4.2 Hasil Pohon Keputusan

Pada Gambar di bawah menjelaskan tentang hasil perhitungan accuracy dari algoritma C4.5 yang dilakukan pada tools Rapid miner. Perhitungan accuracy dari algoritma C4.5, dilakukan dengan cara jumlah TP + TN dibagi jumlah total data testing yang diuji.

Gmabar 4.3 Hasil Accuracy Algoritma C4.5 pada rapid miner

Pada Gambar 4.4 menjelaskan tentang hasil perhitungan precision dari algoritma C4.5 yang dilakukan pada tools Rapid miner. Nilai precision dihitung dengan cara membagi jumlah data benar yang bernilai positif (True Positive) dibagi dengan jumlah data benar yang bernilai positif (True Positive) dan data salah yang bernilai positif (False Positive).



Gmabar 4.3 Hasil Accuracy Algoritma C4.5 pada rapid miner

Berdasarkan data yang dijelaskan pada Tabel 4.5 terlihat bahwa terdapat bahwa dari beberapa perbandingan diatas algoritma C4.5 memiliki nilai tingkat *accuracy* yang cenderung lebih tinggi dibandingkan dengan algoritma *naïve bayes*, maka di putuskan dalam memprediksi kelulusan siswa, algoritma C4.5 lebih baik dalam membanding algoritma *naïve bayes* untuk memprediksi data kelulusan siswa.

5. Penutup

C4.5 memiliki nilai tingkat accuracy yang cenderung lebih tinggi dibandingkan dengan algoritma naïve bayes, maka di putuskan dalam memprediksi kelulusan siswa, algoritma C4.5 lebih baik dibanding algoritma naïve bayes dalam memprediksi data kelulusan siswa.

Daftar Pustaka

[1] B. D. Meilani and N. Susanti, "Aplikasi Data Mining Untuk Menghasilkan Pola Kelulusan Siswa Dengan Metode Naive Bayes," Jurnal Ilmiah NERO, p. 184, 2015.

[2] E. Elisa, "Analisa dan Penerapan Algoritma C4.5 Dalam Data Mining Untuk Mengidentifikasi Faktor- Faktor Penyebab Kecelakaan Kerja Kontruksi PT. Arupadhatu Adisesanti," JOIN, pp. 36-41, 2017.

[3] S. Adinugroho and Y. A. Sari, Implementasi Data Mining Menggunakan WEKA, Malang: UB Press, 2018.

[4] R.T.Vulandari, Data Mining, Teori Dan Aplikasi Rapidminer, Yogyakarta: Penerbit Gava Media, 2017.

[5] Y. Mardi, "Data Mining : Klasifikasi menggunakan Algoritma C4.5," Jurnal Edik Informatika, pp. 213-219, 2016.

[6] N. Azwanti, "Analisa Algoritma C4.5 Untuk Memprediksi Penjualan Motor Pada PT. Capella Dinamik Nusantara Cabang Muka Kuning," Jurnal Ilmiah Ilmu Komputer, pp. 33-38, 2018.

[7] M. R. Faisal and D. T. Nugrahadi, Belajar Data Science: Klasifikasi dengan Bahasa Pemrograman R, Banjarbaru, Kalimantan Selatan: Scripta Cendekia, 2019.

[8] D. Nofriansyah and G. W. Nurcahyo, Algoritma Data Mining dan Pengujian, Yogyakarta: Deepublish, 2015.

[9] Nendrabertus, "Eksplorasi Data Mining Menggunakan RapidMiner," 24 April 2015. [Online]. Available:https://www.softovator.com/eksplorasi-data-mining-menggunakan-rapidminer/.

[10] D. H. Kamagi and S. Hansun, "Implementasi Data Mining dengan Algoritma C4.5 Untuk Memprediksi Tingkat kelulusan Mahasiswa," ULTIMATICS, vol. VI, pp. 15-20, 2014.

[11] R. P. Sari Putri and I. Waspada, "Penerapan Algoritma C4.5 pada Aplikasi Prediksi Kelulusan Mahasiswa Prodi Informatika," Khazanah Informatika, vol. IV, pp. 1-7, 2018.