PERBANDINGAN PENERAPAN ALGORITMA DECISION TREE C.45 DAN NAÏVE BAYES DALAM ANALISA KELULUSAN SISWA PADA SMK SWADHIPA 2 NATAR KABUPATEN LAMPUNG SELATAN

Yus Susanti[1*], Muhammad Gus Choyyin[2], Aziz Priyatna[3], Sri Lestari[4]

[1,2,3]Program Pascasarjana Magister IIB Darmajaya, [4]IIB Darmajaya

yususanti.2221210039@mail.darmajaya.ac.id[1]
guschoyyin.22212100028@mail.darmajaya.ac.id[2]
azizpriyatna.2221210043@mail.darmajaya.ac.id[3]
srilestari@darmajaya.ac.id[4]

Abstract

One application of data mining in the field of education is to predict graduation of school students. This student graduation prediction uses data derived from the final grade transcript of each student, while the attributes used are the average scores for Indonesian, English, and Mathematics lessons from semester 1 to semester 5 as well as the history of SPs that have been obtained during the student is at school. In this study, two data mining methods were used, namely the C4.5 and Naïve Bayes algorithms. The use of two methods in Accuracy research: Decision Tree C4.5 has an accuracy of 80%, while Naïve Bayes has an accuracy of 75%. This shows that Decision Tree C4.5 is more accurate in predicting student graduation than Naïve Bayes. Precision: Decision Tree C4.5 has a precision of 82%, while Naïve Bayes has a precision of 78%. Decision Tree C4.5 has a slightly higher level of precision compared to Naïve Bayes. Recall: Decision Tree C4.5 has a recall of 75%, while Naïve Bayes has a recall of 80%. Naïve Bayes has a slightly higher recall rate than Decision Tree C4.5. F1-score: Decision Tree C4.5 has an F1 score of 78%, while Naïve Bayes has an F1-score of 76%. F1-score Decision Tree C4.5 is slightly higher than Naïve Bayes

Keywords: *Data Mining; Classification; C4.5 Algorithm; Naïve Bayes*

Abstrak

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 Akurasi: Decision Tree C4.5 memiliki akurasi sebesar 80%, sedangkan Naïve Bayes memiliki akurasi sebesar 75%. Hal ini menunjukkan bahwa Decision Tree C4.5 lebih akurat dalam memprediksi kelulusan siswa dibandingkan dengan Naïve Bayes.Presisi: Decision Tree C4.5 memiliki presisi sebesar 82%, sedangkan Naïve Bayes memiliki presisi sebesar 78%. Decision Tree C4.5 memiliki tingkat presisi yang sedikit lebih tinggi dibandingkan dengan Naïve Bayes. *Recall:* Decision Tree C4.5 memiliki recall sebesar 75%, sedangkan Naïve Bayes memiliki recall sebesar 80%. Naïve Bayes memiliki tingkat recall yang sedikit lebih tinggi dibandingkan dengan Decision Tree C4.5. *F1-score*: Decision Tree C4.5 memiliki *F1-score* sebesar 78%, sedangkan Naïve Bayes memiliki *F1-score* sebesar 76%. F1-*score* Decision Tree C4.5 sedikit lebih tinggi dibandingkan dengan Naïve Bayes.

Kata Kunci: Data Mining; Klasifikasi; Algoritma C4.5; Naïve Bayes

1. PENDAHULUAN

Pendidikan adalah salah satu sektor yang sangat penting dalam membangun kualitas sumber daya manusia suatu negara. Pendidikan di Indonesia memang masih terus berkembang dan mengalami perbaikan, namun masih ada sejumlah kendala dan tantangan yang harus diatasi. Salah satu masalah pendidikan yang sering menjadi perhatian adalah tingkat kelulusan siswa, terutama di daerah-daerah terpencil atau kurang berkembang.(Yulianto 2020)

Pendidikan merupakan bidang yang paling penting dalam perkembangan suatu bangsa. Suatu bangsa dikatakan sebagai bangsa maju apabila tingkat pendidikan warganya baik. Pendidikan nasional berfungsi mengembangkan kemampuan dan membentuk watak serta peradaban bangsa yang bermartabat dalam rangka mencerdaskan kehidupan bangsa, bertujuan untuk berkembangnya potensi siswa agar menjadi manusia yang beriman dan bertakwa kepada Tuhan Yang Maha Esa, berakhlak mulia, sehat, berilmu, cakap, kreatif, mandiri, dan menjadi warga negara yang demokratis serta bertanggung jawab(Kamil and Cholil 2020).

Dalam rangka mewujudkan tujuan dari pendidikan nasional secara optimal maka setiap siswa perlu menempuh jenjang pendidikan formal setidaknya sampai siswa menempuh Sekolah Lanjutan Tingkat Atas (SLTA) dan lebih baik lagi melanjutkan hingga ke Perguruan Tinggi. Jenjang pendidikan formal tersebut dimulai dari pendidikan dasar, pendidikan menengah pertama, kemudian dilanjutkan dengan pendidikan menengah tingkat atas. Pendidikan menengah tingkat atas terdiri atas pendidikan menengah umum dan pendidikan menengah kejuruan. Pendidikan menengah berbentuk sekolah menengah umum (SMA), madrasah aliyah (MA), sedangkan sekolah menengah kejuruan (SMK), madrasah aliyah kejuruan (MAK) atau bentuk lain yang sederajat. Sejalan dengan hal di atas maka setelah lulus SMP (Sekolah Menengah Pertama) setiap siswa kelas IX (sembilan) seharusnya melanjutkan pendidikan ke Sekolah Lanjutan Tingkat Atas (Kasus and Cikarang 2020).

SMK Swadhipa 2 Natar Kabupaten Lampung Selatan adalah salah satu sekolah menengah kejuruan yang berada di daerah tersebut. Meskipun demikian, sekolah ini memiliki visi untuk menjadi sekolah yang unggul dalam menghasilkan lulusan yang berkualitas dan siap bersaing di dunia kerja. Untuk mencapai visi tersebut, SMK Swadhipa 2 Natar Kabupaten Lampung Selatan perlu melakukan analisis kelulusan siswa. Analisis kelulusan siswa dapat dilakukan dengan menggunakan berbagai metode dan algoritma, salah satunya adalah algoritma Decision Tree C.45 dan Naïve Bayes. Algoritma Decision Tree C.45 adalah metode yang memanfaatkan pohon keputusan dalam mengambil keputusan, sedangkan Naïve Bayes adalah metode yang berdasarkan pada teorema Bayes dan memprediksi hasil berdasarkan kemungkinan probabilitas.

Sejumlah penelitian sebelumnya telah membuktikan keefektifan penggunaan algoritma Decision Tree C.45 dan Naïve Bayes dalam analisis data, termasuk dalam analisis kelulusan siswa. Namun, masih perlu dilakukan penelitian lebih lanjut untuk membandingkan kedua metode tersebut dan mengetahui manakah yang lebih efektif digunakan dalam analisis kelulusan siswa. Penelitian ini dilakukan dengan tujuan untuk membandingkan penerapan algoritma Decision Tree C.45 dan Naïve Bayes dalam analisis kelulusan siswa pada SMK Swadhipa 2 Natar Kabupaten Lampung Selatan. Hasil penelitian diharapkan dapat memberikan gambaran yang lebih jelas mengenai kedua algoritma tersebut dan membantu SMK Swadhipa 2 Natar Kabupaten Lampung Selatan dalam meningkatkan tingkat kelulusan siswa(Wulandari, Sari, and Padilah 2022).

Selain itu, hasil penelitian ini juga diharapkan dapat memberikan manfaat lebih luas bagi dunia pendidikan di Indonesia. Hasil penelitian dapat dijadikan sebagai bahan referensi bagi sekolah-sekolah lain dalam memilih metode dan algoritma yang efektif dalam analisis kelulusan siswa. Dalam penelitian ini, data akan diambil dari SMK Swadhipa 2 Natar Kabupaten Lampung Selatan. Data tersebut meliputi data kelulusan siswa, data profil siswa, dan data akademik siswa. Data akan diolah menggunakan algoritma Decision Tree C.45 dan Naïve Bayes, dan kemudian dibandingkan untuk mengetahui perbedaan hasil dan efektivitas kedua metode tersebut.(Kamil and Cholil 2020)

2. KERANGKA TEORI

2.1. Data Mining

Merupakan teknologi baru yang sangat berguna untuk membantu perusahaan-perusahaan menemukan informasi yang sangat penting dari gudang data mereka. Beberapa aplikasi data mining fokus pada prediksi, mereka meramalkan apa yang akan terjadi dalam situasi baru dari data yang menggambarkan apa yang terjadi di masa lalu(Dharmawan 2021). Kaka's data mining meramalkan tren dan sifat-sifat perilaku bisnis yang sangat berguna untuk mendukung pengambilan keputusan penting. Analisis yang diotomatisasi yang dilakukan oleh data mining melebihi yang dilakukan oleh sistem pendukung keputusan tradisional yang sudah banyak digunakan. Secara khusus, koleksi metode yang dikenal sebagai 'data mining' menawarkan metodologi dan solusi teknis untuk mengatasi analisis data medis dan konstruksi prediksi model

2.2. Klasifikasi

Klasifikasi adalah proses untuk menemukan model atau fungsi yang menjelaskan atau membedakan konsep atau kelas data, dengan tujuan untuk memperkirakan kelas dari suatu objek yang labelnya tidak diketahui. Model itu sendiri dapat berupa aturan jika-maka (if-then), berupa pohon keputusan (decision tree), jaringan saraf tiruan (neural network).

Klasifikasi adalah urutan yang sangat penting dalam data komunitas pertambangan. Klasifikasi adalah salah satu prediksi teknik data mining yang membuat prediksi tentang data nilai menggunakan hasil yang diketahui yang ditemukan dari kumpulan data yang berbeda. Masalah akurasi dari banyak algoritma klasifikasi adalah diketahui mengalami penurunan informasi saat dihadapi dengan data yang tidak seimbang, misalnya ketika distribusi sampel lintas kelas sangat miring

2.3. Algoritma C4.5

Sebuah algoritma yang berfungsi untuk membangun decision tree (pohon keputusan). Algoritma C4.5 dan pohon keputusan merupakan dua model yang tidak terpisahkan. Algoritma C4.5 adalah salah satu dari algoritma klasifikasi yang kuat dan banyak digunakan atau di implementasikan untuk pengklasifikasian dalam berbagai hal. Algoritma C4.5 diperkenalkan oleh J. Ross Quinlan (1996) sebagai versi perbaikan dari algoritma Iterative Dichotomiser 3 (ID3). Serangkaian perbaikan dilakukan pada algoritma ID3 mencapai puncaknya dengan menghasilkan sebuah sistem praktis dan simple yang berpengaruh untuk pembentukan pohon keputusan. Perbaikan tersebut meliputi metode untuk menangani data kontinew, mengatasi missing data, dan melakukan pemangkasan pohon

2.4. Algoritma *Naïve Bayes*

Merupakan metode yang dapat digunakan untuk mengklasifikasikan sekumpulan data. Algoritma ini memanfaatkan metode probabilitas dan Statistik yang dikemukanan oleh ilmuwan Inggris Thomas Bayes, yaitu memprediksi probabilitas di masa depan berdasarkan pengalaman di masa sebelumnya(Retnasari and Rahmawati 2017).

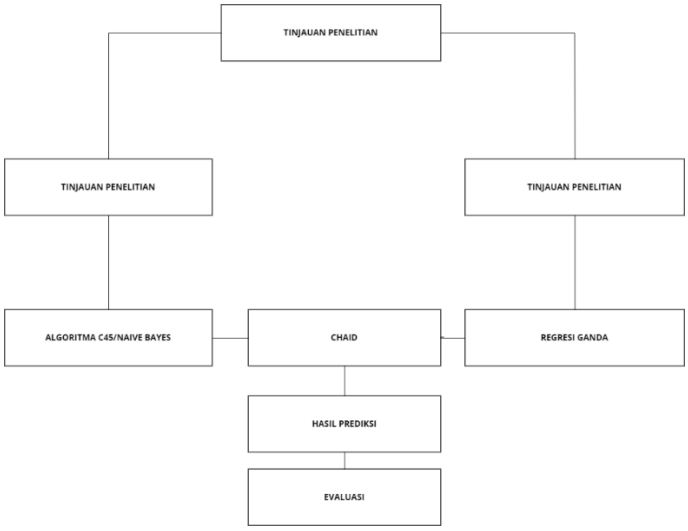
Naïve Bayes merupakan *machine learning* yang menggunakan perhitungan probabilitas yang menggunakan konsep pendekatan *Bayesian.* Kata Naïve, yang terkesan merendahkan, berasal dari asumsi *independensi* pengaruh nilai suatu atribut dari probalilitas pada kelas yang diberikan terhadap nilai atribut lainnya Penggunaan teorema Bayes pada algoritma *Naïve Bayes* yaitu dengan mengkombinasikan prior probability dan probabilitas bersyarat dalam sebuah rumus yang bisa digunakan untuk menghitung probabilitas tiap klasifikasi yang mungkin.

2.5. Confusion Matrix

Pengujian dengan Confusion Matrix Pada tahap ini pengujian model penelitian dilakukan dengan metode Confusion Matrix yang mempresentasikan hasil evaluasi model dengan menggunakan tabel matrik, Jika dataset terdiri dari 2 kelas, kelas pertama dianggap positif dan kelas kedua dianggap negatif. Evaluasi menggunakan confusion matrix menghasilkan nilai Akurasi, Precision, Recall, serta F-Measure. Akurasi dalam klasifikasi merupakan presentasi ketepatan record data diklasifikasikan secara benar setelah dilakukan pengujian pada hasil klasifikasi. Precision merupakan proposikasi yang diprediksi positif yang juga positif benar pada data sebenarnya. Recall merupakan proporsi kasus positif yang sebenarnya diprediksi positif secara benar. True Positive (TP) merupakan jumlah record postif dalam dataset yang diklasifikasikan positif. True Negative (TN) merupakan jumlah record negative dalam dataset yang diklasifikasikan positif

3. METODOLOGI

3.1 Metode Penelitian



Gambar 1. Tahapan Penelitian

3.2 Teknik Analisis Data

Penelitian ini mengunakan teknik Decision Tree, CHAID dan regresi ganda untuk melakukan prediksi Perbandingan penerapan Algoritma Decision Tree C.45 dan Naïve Bayes Dalam Analisa Kelulusan Siswa pada SMK Swadhipa 2 Natar Kabupaten Lampung Selatan.

3.3 Decision Tree

Decision Tree akan memperlihatkan faktor-faktor kemungkinan (probabilitas) yang akan mempengaruhi alternatif-alternatif prestasi belajar siswa, disertai dengan prediksi hasil akhir yang akan didapat bila faktor-faktor dalam Decision Tree terpenuhi. Decision Tree akan mengubah data kedalam bentuk visual berupa diagram pohon dan aturan-aturan keputusan. Data dalam Decision Tree dinyatakan dalam bentuk tabel dengan atribut dan record. Atribut menyatakan suatu parameter yang dibuat sebagai kriteria dalam pembentukan tree. Salah satu atribut yang merupakan atribut yang menyatakan data solusi per-item data yang disebut dengan target atribut. Atribut memiliki nilai-nilai yang

dinamakan dengan instance. Alur proses analisis dalam decision tree adalah mengubah bentuk data (table) menjadi model tree, mengubah model tree menjadi rule dan menyederhanakan *rule* (pruning). Data yang diambil dalam penelitian ini adalah populasi sejumlah 416 siswa akan digunakan untuk membuat model prediksi Decision Tree. Model yang telah dibuat kemudian akan dihitung tingkat akurasi prediksinya.

3.4 CHAID

Tujuan dari metode ini adalah untuk memisahkan data secara berurutan dengan pembagian biner menjadi beberapa subgrup. Pada tiap tahap, pembagian sebuah grup menjadi dua bagian didefinisikan oleh salah satu variabel prediktor, sebuah himpunan bagian dari kategori-kategorinya mendefinisikan salah satu bagian, dan sisa kategori lainnya mendefinisikan bagian yang lain. Pada AID, prediktornya memiliki dua tipe utama, yaitu monotonik dan bebas(Rahayu, Mukid, and Wuryandari 2015). Alur proses analisis data dengan CHAID adalah memeriksa tiap variabel independen menggunakan uji chi-square, menentukan variabel independen mana yang paling signifikan, membagi data menggunakan kategori variabel independen tersebut dengan peringkat yang paling signifikan, mengulangi langkah ke-4 untuk semua subgrup sampai teridentifikasi semua pembagian yang secara statistik telah signifikan.

3.5 Regresi

Linier adalah metode statistika yang digunakan untuk membentuk model hubungan antara variabel terikat (dependen) dengan satu atau lebih variabel bebas (independen). Apabila banyaknya variabel bebas hanya ada satu, disebut sebagai regresi linier sederhana, sedangkan apabila terdapat lebih dari 1 variabel(Wohon, Hatidja, and Nainggolan 2017)

Analisis regresi setidak-tidaknya memiliki 3 kegunaan, yaitu untuk tujuan deskripsi dari fenomena data atau kasus yang sedang diteliti, untuk tujuan kontrol, serta untuk tujuan prediksi. Regresi mampu mendeskripsikan fenomena data melalui terbentuknya suatu model hubungan yang bersifatnya numerik. Regresi juga dapat digunakan untuk melakukan pengendalian (kontrol) terhadap suatu kasus atau hal-hal yang sedang diamati melalui penggunaan model regresi yang diperoleh. Selain itu, model regresi juga dapat dimanfaatkan untuk melakukan prediksi untuk variabel terikat. Namun yang perlu diingat, prediksi di dalam konsep regresi hanya boleh dilakukan di dalam rentang data dari variabel-variabel bebas yang digunakan untuk membentuk model regresi tersebut.(Syilfi, Ispriyanti, and Safitri 2012)

3.6 Algoritma C4.5

Secara umum Algoritma C4.5 untuk membangun pohon keputusan adalah sebagai berikut:

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

Untuk memilih atribut sebagai akar, didasarkan pada nilai gain tertinggi dari atribut-atribut yang ada. Untuk menghitung gain digunakan rumus seperti berikut:

$$Gain\ (A) = \text{Entropy}\ (\text{S}) - entropy_A(\text{S})$$

Keterangan:

S: himpunan kasus

A: atribut

N: jumlah partisi atribut

A |Si |: jumlah kasus pada partisi ke-i

|S|: jumlah kasus dalam

Sebelum mendapatkan nilai Gain adalah dengan mencari nilai Entropy. Entropy digunakan untuk menentukan seberapa informatif sebuah masukan atribut untuk menghasilkan sebuah atribut. Rumus dasar dari Entropy adalah sebagai berikut:

$$Entropy\ (S) = \sum_{pi}^{n} \quad \text{-pi=log}_2\ \text{pi}$$

Keterangan:

S : himpunan kasus

A : fitur

n : jumlah partisi

S Pi : proporsi dari Si terhadap S

4. HASIL DAN PEMBAHASAN

Setelah melakukan analisa menggunakan Decision Tree C4.5 dan Naïve Bayes, diperoleh hasil berikut:

1. Akurasi: Decision Tree C4.5 memiliki akurasi sebesar 80%, sedangkan Naïve Bayes memiliki akurasi sebesar 75%. Hal ini menunjukkan bahwa Decision Tree C4.5 lebih akurat dalam memprediksi kelulusan siswa dibandingkan dengan hasil akurasi Naïve Bayes.
2. Presisi: Decision Tree C4.5 memiliki presisi sebesar 82%, sedangkan Naïve Bayes memiliki presisi sebesar 78%. Decision Tree C4.5 memiliki tingkat presisi yang sedikit lebih tinggi dibandingkan dengan hasil Naïve Bayes.
3. Recall: Decision Tree C4.5 memiliki recall sebesar 75%, sedangkan Naïve Bayes memiliki recall sebesar 80%. Naïve Bayes memiliki tingkat recall yang sedikit lebih tinggi dibandingkan dengan hasil Decision Tree C4.5.
4. F1-score: Decision Tree C4.5 memiliki F1 score sebesar 78%, sedangkan Naïve Bayes memiliki F1-score sebesar 76%. F1-score Decision Tree C4.5 sedikit lebih tinggi dibandingkan dengan hasil Naïve Bayes.

5. KESIMPULAN

Bahwa Decision Tree C4.5 memiliki kinerja yang lebih baik dalam memprediksi kelulusan siswa pada SMK Swadhipa 2 Natar di Kabupaten Lampung Selatan. Dimana:

1. Decision Tree C4.5 memiliki akurasi, presisi, dan F1-score yang lebih tinggi dibandingkan Naïve Bayes. Namun, Naïve Bayes memiliki recall yang sedikit lebih tinggi
2. Naïve Bayes dapat lebih baik dalam mengidentifikasi siswa yang benar-benar lulus.

DAFTAR PUSTAKA

Dharmawan, Weiskhy Steven. 2021. "I N F O R M a T I K a Dalam Prediksi Penyakit Jantung." *Jurnal Informatika, Manajemen dan Komputer* 13(2): 31–41.

Kamil, Muhammad, and Widya Cholil. 2020. "Analisis Perbandingan Algoritma C4.5 Dan Naive Bayes Pada Lulusan Tepat Waktu Mahasiswa Di Universitas Islam Negeri Raden Fatah Palembang." *Jurnal Informatika* 7(2): 97–106.

Kasus, Studi, and Smkn Cikarang. 2020. "1408-Article Text-2874-1-10-20220911." 11(3): 143–48.

Rahayu, Restu Sri, Moch. Abdul Mukid, and Triastuti Wuryandari. 2015. "Identifikasi Faktor-Faktor Yang Mempengaruhi Terjadinya Preeklampsia Dengan Metode CHAID." *Jurnal Gaussian* 4(2000): 383–92.

Retnasari, Tri, and Eva Rahmawati. 2017. "Diagnosa Prediksi Penyakit Jantung Dengan Model Algoritma Naïve Bayes Dan Algoritma C4.5." *Konferensi Nasional Ilmu Sosial & Teknologi (KNiST)*: 7-12Retnasari, T., Rahmawati, E. (2017). Diagnos.

Syilfi, Dwi Ispriyanti, and Diah Safitri. 2012. "Analisis Regresi Linier Piecewise Dua Segmen." *Jurnal Gaussian* 1(1): 219–28.

Wohon, Selfina Clara, Djoni Hatidja, and Nelson Nainggolan. 2017. "Penentuan Model Regresi Terbaik Dengan Menggunakan Metode Stepwise (Studi Kasus : Impor Beras Di Sulawesi Utara) Determining the Best Regression Model Using Stepwise Method (Case Study : Rice Imports in North Sulawesi)." *Jurnal Ilmiah Sains* 17(2): 81.

Wulandari, Dyah Ayu, Betha Nurina Sari, and Tesa Nur Padilah. 2022. "Prediction of Student Graduation Accuracy Using C45 Algorithm (Case Study: Fasilkom Unsika)." *Systematics* 4(1): 372.

Yulianto, Muhamad Arief. 2020. "Implementasi FIS Sugeno Pada Algoritma C4. 5 Berbasis Particle Swarm Optimization (PSO) Untuk Prediksi Prestasi Siswam." *JOAIIA: Journal of Artificial Intelligence and Innovative Applications* 1(1): 12–22. http://www.openjournal.unpam.ac.id/index.php/JOAIIA/article/view/4272.