Penerapan Komparasi Algoritma C4.5 dan Naïve Bayes untuk Menentukan Hasil Seleksi Masuk Perguruan Tinggi

Indah Sulihati
Teknik Informatika, Fakultas Teknik, Universitas Dian Nuswantoro
Email: adzamindah@gmail.com

Abstrak

Peningkatan pesat dalam pendataan telah menciptakan situasi di mana data berlimpah tetapi informasi langka, seperti informasi kepada orang tua tentang hasil penerimaan anak mereka ke perguruan tinggi. Oleh karena itu, penelitian ini dilakukan untuk membantu menyelesaikan permasalahan tersebut dengan menggunakan teknik data mining. Data mining itu sendiri adalah ekstraksi atau penemuan informasi baru dengan mencari pola atau aturan tertentu dari sejumlah besar data yang diharapkan dapat mengatasi kondisi tersebut. Metode data mining yang digunakan dalam white paper ini adalah klasifikasi, namun metode klasifikasi yang digunakan adalah pohon keputusan (decision tree) dan algoritma naive Bayes. Buku putih ini menggunakan teknik data mining untuk menemukan informasi berharga dalam data dan memungkinkan sekolah memberikan saran untuk memutuskan apakah seorang anak akan masuk perguruan tinggi setelah lulus dari SMA Kertosono. Dengan menggunakan algoritma Naive Bayes diketahui hasil yang didapat dari prediksi hasil masuk perguruan tinggi lebih akurat daripada algoritma Decision Tree.

Kata Kunci: *Data Mining, Decision Tree, Naive Bayes*

ABSTRACT

The rapid increase in data collection has created a situation where data is abundant but information is scarce, such as information to parents about their child's admission to college. Therefore, this research was conducted to help solve these problems by using

data mining techniques. Data mining itself is the extraction or discovery of new information by looking for certain patterns or rules from large amounts of data that are expected to overcome these conditions. The data mining method used in this white paper is classification, but the classification method used is a decision tree and a naive Bayes algorithm. This white paper uses data mining techniques to find valuable information in data and allows schools to provide advice for deciding whether a child will enter college after graduating from Kertosono High School. By using the Naive Bayes algorithm, it is known that the results obtained from predicting college entrance results are more accurate than the Decision Tree algorithm.

Keywords: *Data Mining, Decision Tree, Naive Bayes*

A. PENDAHULUAN

Mengandalkan data operasional saja tidak cukup untuk memanfaatkan data sistem informasi yang ada untuk mendukung pengambilan keputusan. Analisis data diperlukan untuk menyelidiki potensi informasi yang ada. Pengambil keputusan memanfaatkan gudang data yang ada untuk menemukan informasi yang dapat membantu mereka membuat keputusan. Hal ini memudahkan munculnya disiplin ilmu baru untuk mengatasi masalah penggalian informasi dan pola yang penting atau menarik dari sejumlah besar data, yang disebut data mining.

Dengan menggunakan teknologi data mining, kami bertujuan untuk memberikan pengetahuan yang sebelumnya tersembunyi di gudang data dan mengubahnya menjadi informasi yang berharga. Penambangan data adalah area yang relatif muda yang dikembangkan dalam beberapa tahun terakhir. Dengan perkembangan teknologi informasi dan komunikasi, teknologi data mining telah digunakan untuk menganalisis data dalam jumlah besar dan menjadi semakin populer saat ini. Data mining adalah disiplin ilmu interdisipliner yang mencakup sistem database, statistik, pembelajaran mesin, visualisasi, dan ilmu informasi. Selain itu, sistem data mining berbasis data mengintegrasikan teknologi

lain seperti analisis data spasial, temu kembali informasi, pengenalan pola, pemrosesan sinyal, grafik komputer, teknologi web, ekonomi, manajemen bisnis, bioinformatika, dan psikologi .

Klasifikasi adalah teknik data mining untuk memproses menempatkan suatu objek atau konsep ke dalam satu set kategori berdasarkan objek atau konsep yang bersangkutan dan untuk menemukan suatu model atau fungsi yang menggambarkan dan membedakan kelas data atau konsep dengan tujuan dapat menggunakan model untuk membuat prediksi kelas objek dimana kelas labelnya tidak diketahui.

Data mining adalah salah satu metode yang paling umum untuk menganalisis keberhasilan sebuah penelitian [1]. Data mining banyak digunakan dalam dunia pendidikan. Data mining dalam pendidikan adalah proses yang digunakan untuk mengekstrak informasi dan pola yang berguna dari database pendidikan yang sangat besar [2]. Akibatnya, akan membantu pendidik memberikan pendekatan pendidikan yang efektif [1]. Selain itu, pendidik dapat memantau kinerja siswa. Banyak penelitian telah dilakukan dengan menggunakan metode klasifikasi seperti pohon keputusan [3] untuk memprediksi kinerja siswa.

Dari sekian banyak metode klasifikasi, pohon keputusan banyak digunakan karena secara teoritis sederhana dan hasilnya mudah dibaca [4]. Pohon keputusan adalah algoritma yang kuat, populer, mudah ditafsirkan, dan banyak digunakan untuk berbagai masalah penambangan data. Algoritma ini memberikan kinerja yang sangat baik dan mudah dimengerti.

Pengklasifikasi Bayesian merupakan pengklasifikasi statistik untuk dapat memprediksi probabilitas keanggotaan kelas tertentu, menghitung peluang untuk suatu hipotesis, menghitung peluang dari suatu

kelas dari masing-masing kelompok atribut yang ada, dan menentukan kelas mana yang paling optimal [5].

Penelitian Terdahulu

Dewi Rahma Ente (2020) Penelitian tersebut didasari karena berkaitan dengan penyakit DM, status DM penderita penting untuk diketahui sebelum penderita DM mengalami komplikasi serius. Algoritma C4.5 telah popular digunakan untuk memprediksi status penyakit. Oleh karena itu dalam tulisan ini akan 6 digunakan algoritma C4.5 sebagai salah satu implementasikan data mining untuk mengklasifikasi penyakit DM. Faktor-faktor yang mempengaruhi status DM secara substansial adalah glukosa darah puasa (GDP), kolesterol LDL, usia dan berat badan. Dengan mengetahui faktor-faktor yang mempengaruhi status DM penderita maka komplikasi serius akibat DM ini dapat dicegah sedini mungkin Keunggulan dalam penelitian yang akan dikerjakan adalah pengukuran akurasi data latih dan data uji dari algoritma C4.5 dengan validasi silang lipat 10 setelah proses seleksi atribut dapat dilihat pada. Nilai akurasi memiliki rentang antara 50% sampai dengan 100% dengan tingkat akurasi rata-rata prediksi yaitu 98,5%. Ini berarti model yang didapatkan sangat baik dengan tingkat akurasi sangat tinggi[6].

Penelitian yang dilakukan oleh Amornsinlaphachai (2016) dengan tujuan Memilih model Data Mining untuk memprediksi prestasi akademik terhadap peserta didik program computer untuk membandingkan model efisiensi Data Mining dengan teknik klasifikasi dan membagun model pembelajaran berbasis web terhadap partisipasi peserta didik[7].

Untuk itu dalam penelitian ini akan dilakukan perbandingan metode klasifikasi data mining yaitu Algoritma C4.5 dan Naïve Bayes. Kemudian akan dilakukan komparasi terhadap kedua metode tersebut

sehingga didapatkan algoritma terbaik untuk prediksi penerimaan Perguruan Tinggi.

B. METODE

Penelitian ini menggunakan 2 metode algoritma klasifikasi yaitu algoritma decision tree dan algoritma naïve bayes. Yang selanjutnya akan dikomparasi kedua metode ini untuk menghasilkan metode mana yang terbaik yang bisa digunakan untuk memprediksi hasil seleksi masuk Perguruan Tinggi.

Decision Tree

Pohon keputusan adalah algoritma pembelajaran terawasi yang membutuhkan data dengan atribut kelas. Jika data hilang atau proses klasifikasi tidak dapat dilakukan, data harus beragam. Salah satu aspek yang paling menarik dari pohon keputusan adalah bagaimana aturan dibentuk.

Pohon keputusan adalah aturan keputusan khusus yang diatur dalam struktur pohon. Aturan keputusan dapat dibangun dari pohon keputusan hanya dengan melintasi jalur apa pun dari simpul akar ke daun apa pun. Satu set lengkap aturan keputusan dihasilkan. Pohon keputusan membagi ruang dokumen di daun menjadi area yang tidak tumpang tindih, dan prediksi dibuat pada setiap daun [8].

Algoritma pohon keputusan dibangun di atas dataset. Divide-and-conquer digunakan untuk membangun model pohon keputusan dengan menggunakan IG untuk memilih atribut dari kumpulan data gaya pohon. Pada setiap langkah dalam membangun pohon keputusan, satu dari setiap atribut dipilih untuk mengisolasi data.

Nilai atribut digunakan untuk menentukan nilai pembatas berdasarkan atribut yang dipilih. GI dan entropi biasanya digunakan untuk pohon klasifikasi. Untuk menghitung entropi didefinisikan (Shannon, 1948) sebagai berikut:

$$H_e(S) = 1 - \sum_{y \in C} p(y)^2$$

Di mana S adalah dataset, C adalah kelas dan p(y) adalah perbandingan jumlah data terhadap kelas C. entropi akan bernilai 0 apabila hanya ada 1 kelas dan mencapai nilai maksimum ketika semua kelas memiliki kemungkinan yang sama.

Naive Bayes

Naive Bayes adalah jaringan Bayesian sederhana. Naive Bayes sering digunakan untuk masalah klasifikasi. Klasifikasi merupakan aspek penting dari data mining. Dalam klasifikasi, pengklasifikasi dibuat dari contoh yang diberikan oleh spesifikasi kelas.

Setiap sampel E diwakili oleh sebuah vektor (a1, a2, ..., an). Dimana ai adalah nilai atribut Ai dan A1, A2, ..., An mewakili n atribut. Pengklasifikasi memprediksi kemungkinan label untuk data baru yang tidak berlabel. Naive Bayes memiliki struktur yang sangat sederhana dan mudah untuk dirancang..

Proses pembelajaran Naive Bayes hanya menghitung probabilitas, khususnya probabilitas bersyarat untuk setiap atribut, dari data latih. Dengan kata lain, nilai probabilitas p (ai | c) harus ditentukan dari sampel data pelatihan. Untuk setiap nilai ai, atribut Ai mempertimbangkan nilai dari variabel kelas C c. Dalam Naive Bayes, atribut diasumsikan sebagai berikut: Mereka adalah kelas independen satu sama lain [10].

Setiap atribut hanya memiliki variabel class sebagai induknya[10], P (E | c) dihitung oleh:

$$p(E|c) = p(a_1, a_2, \dots, a_n|c) = \prod_{i=1}^{n} p(a_i|c)$$

Di mana:

p (ai | c) mengacu pada probabilitas Ai, dan contohnya E = (a1, a2, an).

C. HASIL DAN PEMBAHASAN

Penelitian ini menggunakan 2 metode algoritma clasifikasi yaitu algoritma decision tree dan algoritma naïve bayes. Yang selanjutnya akan dikomparasi kedua metode ini untuk menghasilkan metode mana yang terbaik yang bisa digunakan untuk memprediksi hasil seleksi masuk kampus di sma Kertosono.

Pada perhitungan dengan Metode C4.5 terlebih dahulu harus mencari nilai Entrophy dan nilai Gain. Menghitung Jumlah yang "Diterima" dan "Ditolak" dari data latih dan mengetahui nilai entrophy. Jumlah kasus yang diterima sebanyak 20 dan ditolak sebanyak 16. Total keseluruhan menjadi 36 orang. Sehingga dapat dihitung nilai entrophy keseluruhan.

Hasil dari prediksi menggunakan metode algoritma decision tree. Dapat dilihat pada gambar 1.

accuracy: 82.34% +/- 8.22% (micro average: 82.33%)

	true Placed	true Not Placed	class precision
pred. Placed	136	26	83.95%
pred. Not Placed	12	41	77.36%
class recall	91.89%	61.19%	

Gambar 1. Hasil perhitungan akurasi di metode decision tree

Akurasi yang dihasilkan algoritma decision tree menghasilkan akurasi sebesar 82,34%. Hasil dari perhitungan menggunakan metode naïve bayes. Dapat dilihat pada gambar 2.

accuracy: 91.08% +/- 7.56% (micro average: 91.16%)

	true Placed	true Not Placed	class precision
pred. Placed	148	19	88.62%
pred. Not Placed	0	48	100.00%
class recall	100.00%	71.64%	

Gambar 2. Hasil perhitungan akurasi di metode Naive Bayes

Akurasi yang dihasilkan algoritma naïve bayes menghasilkan tingkat akurasi sebesar 91,08%.

Tabel 1 merupakan komparasi metode Decision Tree dan Naïve Bayes pada pengkasifikasian dataset campus recruitment. Pada tabel 1 memperlihatkan dengan menggunakan metode DT didapatkan akurasi 82,34%dan dengan menggunakan metode NB didapatkan akurasi 91,08.

Tabel 1. Hasil komparasi metode algoritma

Metode	Akurasi
Decision Tree	**82,34%**
Naive Bayes	**91,08%**

Dari hasil komparasi tersebut menunjukkan metode algoritma naïve bayes memiliki tingkat akurasi yang paling tinggi. Hal tersebut menunjukkan bahwa kinerja metode algoritma naïve bayes lebih baik dibanding dengan Decision Tree. Hal ini membuktikan penelitan penelitian sebelumnya metode naïve bayes lebih baik dari metode algortima decision tree.

D. PENUTUP

Simpulan dan Saran

Dalam pendidikan, kinerja siswa merupakan bagian yang penting. Pada penelitian ini dilakukan komparasi metode Decision Tree dan Naive Bayes untuk mengklasifikasikan kinerja siswa yang bisa diterima oleh Perguruan Tinggi dengan menggunakan dataset *campus recruitment*. Hasil penelitian pada dataset campus recruitment dengan menggunakan metode Decision Tree didapatkan akurasi 82,34%, dengan menggunakan metode Naive Bayes didapatkan akurasi 91,08%. Dari hasil tersebut dapat disimpulkan bahwa kinerja metode algoritma Naïve Bayes lebih baik dibanding metode Decision Tree.

Dari penelitian ini tentunya masih terdapat banyak kekurangan sehingga Analisa perbandingan algoritma klasifikasi diatas perlu dilakukan lagi penelitian lanjutan agar klasifikasi data seleksi perguruan tinggi dapat dilakukan secara lebih valid agar akurasi yang dihasilkan lebih tinggi.

DAFTAR PUSTAKA

Amornsinlaphachai P. 2016. *Efficiency of data mining models to predict academic performance and a cooperative learning model*. 8th International Conference on Knowledge and Smart Technology.

Angeline, D.M.D. 2013. *Association Rule Generation for Student Performance Analysis using Apriori Algorithm. The SIJ Transactions on Computer Science Engineering & Its Applications (CSEA)*. Vol.1 No.1: 12–16.

Ente, D. R., Thamrin, S. A., Arifin, S., Kuswanto, H., & Andreza, A. 2020. *Klasifikasi Faktor-Faktor Penyebab Penyakit Diabetes Melitus Di Rumah Sakit Unhas Menggunakan Algoritma* C4.5. Indonesian Journal Of Statistics And Its Applications. Vol.4 No.1: 80–88.

Gries, D., & Schneider, F. B. 2010. *Texts in Computer Science*. Media Vol. 42.

Shahiri, A. M., Husain, W., & Rashid, N. A. 2015. *A Review on Predicting Student's Performance Using Data Mining Techniques. Procedia Computer Science*. 414–422.

Hadjaratie, L. 2011. *Jaringan Saraf Tiruan Untuk Prediksi Tingkat Kelulusan Mahasiswa Diploma Program Studi Manajemen Informatika Universitas Negeri Gorontalo*. Thesis Postgraduated IPB, Bogor.

Lolli, F., Ishizaka, A., Gamberini, R., Balugani, E., & Rimini, B. 2017. *Decision Trees for Supervised Multi-criteria Inventory Classification. Procedia Manufacturing*. 1871–1881.

Lopez Guarin, C. E., Guzman, E. L., & Gonzalez, F. A. 2015. *A Model to Predict Low Academic Performance at a Specific Enrollment Using Data Mining. Revista Iberoamericana de Tecnologias Del Aprendizaje*. Vol.10 No.3: 119–125.

Wu, J., & Cai, Z. 2011. *Attribute Weighting via Differential Evolution Algorithm for Attribute Weighted Naive Bayes (WNB)*. Journal of Computational Information Systems. Vol.5 No.5: 1672–1679.