PENERAPAN TEKNIK DATA MINING UNTUK MENENTUKAN HASIL SELEKSI MASUK SMAN 1 GIBEBER UNTUK SISWA BARU MENGGUNAKAN DECISION TREE

Castaka Agus Sugianto
Program Studi Teknik Informatika Politeknik TEDC Bandung
E-mail: castaka@poltektedc.ac.id

Abstrak
Pada proses penentuan hasil seleksi siswa di SMA Negeri 1 Cibeber dilakukan dengan memperhatikan aturan-aturan tertentu sehingga siswa dapat dengan mudah memperoleh informasi hasil dari seleksi mereka. Penulis mencoba menggali pola dari sebuah data penerimaan siswa baru dengan metode klasifikasi untuk membantu pengambilan keputusan dari seorang pimpinan supaya kebijakan lebih proaktif dan benar. Objek penelitian merupakan daftar siswa yang mendaftar ke SMA Negeri 1 Cibeber tahun 2011 dan tahun 2012. Uji yang dilakukan menggunakan 3 (tiga) jenis Algoritma, yaitu Alogaritma C4.5 (Decsion Tree), Naive Bayes dan Naural Network, yang menunjukan hasil sebagai berikut : Pengujian dengan menggunakan Algoritma C4.5 (Decisian Tree), adalah Accuracy 99.05%, Precisian 96.33%, pengujian dengan Algoritma Naive Bayes, Accuracy 90.62%, Precision unknawn, sedangkan pengujiandengan Al goritma Neural Network, Accuracy 95.02%, Precision 7 3.10%. Memperhatikan penentukan prediksi tingkat akurasi hasil seleksi siswa yang masuk di SMA Negeri 1 Cibeber yang dilakukan bahwa Algoritma C4.5 memperoleh tingkat akurasi yang cukup tinggi yaitu mencapai sebesar 99.05%, Sehingga tingkat akurasi seleksi masuk SMA Negeri 1 Cibeber dari penyeleksian nilai yang disajikan dalam pengelompokan matia pelajaran dalam Ujian Nasional merupakan metode yang tepat untuk digunakan dalam memprediksi seleksi masuk siswa ke SMA Negeri 1 Cibeber.

Kata kunci : Data Mining, Decision Tree, C4.5, Naive Bayes, Neural Network.

Pendahuluan

SMA Negeri 1 Cibeber – Cianjur tiap tahunya melakukan penerimaan siswa baru, sehingga Penentuan Hasil Seleksi Masuk pada SMA Negeri 1 Cibeber dilakukan disetiap pergantian Tahun Ajaran. Penentuan Hasil Seleksi Masuk ini cukup menyita banyak waktu, membutuhkan tenaga ekstra juga ketelitian yang lebih untuk membuatnya, karena penentuan keputusan disini harus memperhatikan aturan-aturan tertentu dari sekolah SMA Negeri 1 Cibeber dan juga waktu pengumuman hasil seleksi agar siswa/siswi dapat dengan mudah dan cepat memperoleh informasi hasil dari seleksi mereka.

Beberapa metode yang digunakan untuk mengolah data yang sifatnya besar untuk menemukan pola yang terdapat didalamnya diantaranya adalah : teorema bayes, *decision tree, artificial neural networks, support vector machine,* regresi linear[1].

Dari metode tersebut memiliki kelebihan dan kekurangan masing – masing. Namun pada penelitan kali ini penulis mengangkat mengenai penggunaan algoritma C4.5 yang merupakan algoritma dari metode *decision tree.* Berdasarakan masalah yang telah diuraikan maka penulis mengangkat Untuk dapat mengatasi permasalahan tersebut diperlukan

pola atau metode komputasi untuk memprediksi atau perkiraan dari siswa yang masuk ke SMA Negeri 1 Cibeber.

Pengambilan keputusan berdasarkan *decision tree analysis* (analisis pohon keputusan) merupakan salah satu alat pengambilan keputusan prediksi dari berbagai altematif yang tersedia. Analisis pohon keputusan biasa digambarkan dengan simbol standar. Prediksi ini bisa dilakukan dengan metode klasifikasi dalam data mining.

Tujuan dari penelitian ini Mengetahui tingkat akurasi prediksi masuk SMAN 1 Cibeber. Membangun pola untuk prediksi seleksi masuk ke SMAN 1 Cibeber. dengan menggunakan algoritma C4.5, kemudian membandingkan hasil dan akurasi algoritma C4.5 dengan Naïve bayes dan Neural Network.

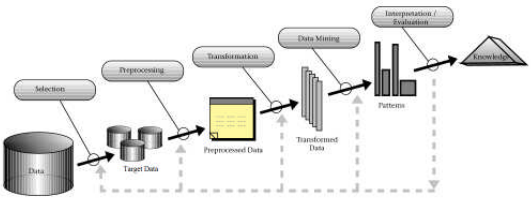
Landasan Teori

1 . Data Mining

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 retieval, dan komputasi tingkat tinggi. Selain itu, data mining didukung oleh ilmu lain seperti neural network, pengenalan pola, spatial data *analysis, image database, signal processing*[2]. *Data mining* didefinisikan sebagai proses menemukan pola-pola dalam data. Proses ini otomatis atau seringnya semiotomatis. Pola yang ditemukan harus penuh arti dan pola tersebut memberikan keuntungan, biasanya keuntungan secara ekonomi. Data yang dibutuhkan dalam jumlah besar [3].

2. Proses Data Mining

Menurut Hermawati Fajar Astuti (2009), Tahapan proses dalam penggunaan *Data Mining* yang merupakan proses K*nowledge Discovery in Databases* (KDD) adalah sebagai berikut [4] :



Gambar 1. proses K*nowledge Discovery in Databases*

1. Menentukan target data yang meliputi pemilihan data dan lebih fokus pada sub set data.
2. *Cleaning and integration*
   Untuk data c*leaning*, yaitu menghilangkan noise dan data yang tidak konsisten, sedangkan *Integration* yaitu menggabungkan data dari berbagai sumber yang berbeda.
3. *Selectiion and transformation*
   Dalam data *selection* yaitu mengambil data yang sesuai dengan tugas analisis dari database, sedangkan Data *transformation*, yaitu menggabungkan data ke dalam sebuah bentuk atau model yang sesuai untuk penggalian melalui operasi *summary* atau *aggregation*.
4. *Data mining* merupkaan proses yang penting dan utama untuk mengekstrak pola dari data dengan metode yang labih mutakhir..
5. *Pattern evaluation*, yaitu mengidentifikasikan pola yang menarik serta merepresentasikan pengetahuan berdasarkan *interestingness measures.*
6. *Knowledge presentation*, berupa penyajian pengetahuan yang digali dan disajikan kepada pengguna dengan menggunakan visualisasi dan teknik representasi pengetahuan.

3. Klasifikasi

Dalam Data Mining akan ditemukan dengan namanya Klasifikasi. Menurut Fajar Astuti Hermawat klasifikasi yaitu menentukan sebuah record data baru ke salah satu dari beberapa kategori (kelas) yang telah didefinisikan sebelumnya[4].

Komponen-komponen utama dari proses klasifikasi antara lain [5] :

1. Kelas, merupakan variable tidak bebas yang merupakan label dari hasil klasifiikasi.
2. Prodiktor, merupakan variable bebas suatu model berdasarkan dari karakteristik atribut data yang diklasifikasi.
3. Set Data Pelatihan, merupkan sekumpuln data lengkap yang berisi kelas dan predictor untuk dilatih agar model dapat mengelompokkan ke dalam kelas yang tepat.
4. Set Data uji, yang berisi data-data baru yang akan dikelompokkan oleh model, guna mengetahui akurasi dari model yang telah dibuat.

Terdapat beberapa elemen yang menjadi kunci utama dalam klasifikasi, yang berperan untuk menentukan suatu model itu termasuk baik atau tidak. Elemen-elemen itu diantaranya :

1. Akurasi Prediksi, menentukan tingkat akurasi suatu model dalam memprediksi keluaran.
2. Kecepatan, ini menunjukkan cepat atau lambatnya dalam memproses data masukan
3. *Robustness*, disini menggambarkan kemampuan suatu model dalam melakukan prediksi yang akurat, walaupun dalam kondisinya seringkali banyak terjadi gangguan.
4. Skalabilitas, merupakan kemampuan suatu model dalam memproses data yang berukuran lebih besar maupun data yang berbeda yang didapat dari bidang-bidang lainnya.
5. *Interpretability*, ini menggambarkan kemudahan suatu model untuk dapat dipahami dan diinterpretasikan.
6. Kesederhanaan, ini merupakan sifat yang cenderung dipilih untuk menyelesaikan suatu permasalahan.

4. Decision Tree

Decision tree adalah kelasifikasi yang melakukan partisi rekursif pada ruang sample [6]. *Decision tree* merupakan salah satu teknik yang dapat digunakan untuk melakukan klasifikasi terhadap sekumpulan objek atau record. Teknik ini terdiri dari kumpulan *decision node*, dihubungkan oleh cabang, bergerak ke bawah dari *root node* sampai berakhir di leaf node..

4.1. Kelebihan Metode Decision Tree

Dalam penggunaannya, Metode Decision Tree ini memiliki kelebihan sebagai berikut :

1. Daerah pengambilan keputusan yang sbelumnya kompleks dan sangat global, dapat diubah menjadi lebih simpel dan spesifik.
2. Eliminasi perhitungan-perhitungan yang tidak diperlukan, karena ketika menggunakan metode decision tree maka *sample* diuji hanya berdasarkan criteria atau kelas tertentu.
3. Fleksibel untuk memilih fasilitas atau fitur dari internal nodes yang berbeda, *fitur* yang terpilih akan membedakan kriteria satu dengan kriteria yang lain dalam *node* yang sama. Karena metode *decision tree* ini bersifat fleksibel, sehingga akan mampu untuk meningkatkan kualitas keputusan yang dihasilkan jika dibandingkan ketika menggunakan metode penghitungan satu tahap yang masih berisifat konvensional.
4. Dalam analisis multivariat, dengan kriteria dan kelas yang jumlahnya sangat banyak, seorang penguji biasanya perlu untuk mengestimasikan baik itu distribusi dimensi tinggi ataupun parameter tertentu dari distribusi kelas tersebut. Metode *Decision Tree* dapat menghindari munculnya permasalahan ini, yaitu dengan menggunakan kriteria yang jumlahnya lebih sedikit pada setiap node internal tanpa harus banyak mengurangi kualitas keputusan yang dihasilkan.

4.2 Algoritma C4.5

Algoritma C4.5 merupakan struktur pohon yang terdapat simpul yang akan mendeskripsikan atribut-atribut, dimana setiap cabang menggambarkan hasil dari atribut yang diuji, dan setiap daun menggambarkan kelas. Algoritma C4.5 secara rekursif mengunjungi setiap simpul keputusan, memilih pembagian yang optimal, hingga akhirnya tidak bisa dibagi lagi. Algoritma C4.5 menggunakan konsep lnformation Gain atau Entrcpy Reduction untuk memilih pembagian yang optimal[2].

Adapun langkah-langkah dari algoritma c4.5 sebagai berikut:

4.3 *Confusion Matrix*

Confusion Matrix merupakan visualisasi untuk mengevaluasi dari kinerja model klasifikasi [7]. Untuk melakukan klasifikasi evaluasi komparatif, maka dalam penelitian ini menggunakan *Confusion Matrix*. *Confusion Matrix* ini meliputi informasi tentang kelas yang sebenarnya dan kelas prediksi. Hal ini akan ditemukan pada kolom matriks yang mewakili kelas yang diprediksi, sedangkan setiap baris mewakili kejadian pada kelas tersebut.

Confusion Matrix adalah salah satu alat ukut berbentuk matrik 2x2 yang digunakan untuk mendapatkan jumlah ketepatan algoritma yang dipakai.

Confusion Matrix disajikan pada tabel 1. di bawah ini [5] :

Tabel 1. *Confusion Matrix*

		Actual Class	
		Class=Yes	Class=No
Predicted Class	Class=Yes	TP (True Positive)	FP (False Positive)
	Class=No	FN (False Negative)	TN (True Negative)

Keterangan:

True Positive (TP) : Jika data yang diprediksi bernilai positif dan sesuai dengan nilai aktual (positif).

False Positive (FP) : Jika yang di tidak sesuai dengan nilai aktual.

False Negative (FN) : Jika yang diprediksi bernilai negatif dan aktualnya positif.

True Negative (TN). Jika benar antara prediksi negatif dan kenyataannya negative.

Untuk mengukur performance dari hasil data mining salah satunya menggunakan akurasi, adapun rumusnya bisa dilihat di bawah ini:

$$Accuracy = \frac{Number\ of\ correct}{total\ number\ of\ prediction}\ x\ 100\%$$

Selain akurasi, metrik evaluasi lainnya yaitu *precision, recall*, [8]. Metrik yang sering digunakan untuk alat penilaian diantaranya, yaitu :

$$Recall = \frac{TP}{TP + FN}\ x\ 100\%$$

$$Precision = \frac{TP}{TP + FP}\ x\ 100\%$$

Pembahasan

1. Pengumpulan Data

Data yang digunakan dalam penelitian ini data siswa diambil dari data siswa yang mendaftar ke SMAN 1 Cibeber, yang berasal dari berbagai jenis sekolah SLTP/Sederajat yang ada di Kecamatan Cibeber bahkan ada yang dari luar Kecamatan Cibeber. Data yang digunakan data-data yang diambil dari bagian Kesiswaan khususnya. Data yang diambil yaitu data siswa yang masuk atau pendaftar ke SMAN 1 Cibeber tahun 2011 dan 2012. Aplikasi perangkat lunak klasifikasi yang digunakan yaitu RapidMiner.

Tabel 2. Kategori kriteria kelulusan

No	Jml Nilai UN	Hasil
1	>24.050	Diterima
2	<24.050	Tidak Diterima

Adapun Sample Data Siswa yang mendaftar bisa di lihat pada tabel 3 di bawah ini.

Tabel 3. Data siswa yang mendaftar di SMAN 1 Cibeber

no	nm_calon	Bind	Bing	Mat	IPA	Jml	Status	Prestasi
1	YULYANTI	6.60	9.20	9.75	9.25	34.80	Diterima	Tidak Ada
2	SEPTIAN AGOSTINE	8.20	8.20	8.00	9.25	33.65	Diterima	Tidak Ada
3	MUHAMAD ABDUL ROIS	7.40	7.80	8.50	8.00	31.70	Diterima	Tidak Ada
4	IKBAL	7.80	5.40	3.50	6.25	22.95	tdk_diterima	Tidak Ada
5	ANGGIA AKTIFA	8.20	7.40	8.00	7.75	31.35	Diterima	Tidak Ada
6	RIDHA HUSNI QONIAH	7.40	7.00	7.75	7.25	29.40	Diterima	Tidak Ada
7	LILIS SUPRIATI	6.20	8.00	6.50	7.25	27.95	Diterima	Tidak Ada
8	SULASTRI	6.80	6.40	7.75	7.50	28.45	Diterima	Tidak Ada
9	MUHAMMAD MUSTHOFA	9.00	8.80	9.75	8.00	35.55	Diterima	Tidak Ada
10	LISDA NISA N	8.20	7.00	8.75	8.75	32.70	Diterima	Tidak Ada
11	NAZAR ABDUL AZIZ	7.20	7.20	7.50	7.75	29.65	Diterima	Tidak Ada
12	SITI AISYAH	9.00	7.00	8.00	7.75	31.75	Diterima	Tidak Ada
13	NINING BAIZURAH	7.40	8.60	8.00	8.50	32.50	Diterima	Tidak Ada
14	SITI FATIMAH	7.00	6.60	7.25	7.25	28.10	Diterima	Tidak Ada
15	WULAN MAULIDHA K	8.80	8.40	9.00	7.50	33.70	Diterima	Tidak Ada
...	...	...	...	...	...	...	...	...
522	ENENG RESTI	7.00	4.60	5.00	4.75	21.35	tdk_diterima	Tidak Ada

2. Hasil Pengujian

Setelah dilakukan pengujian pada RapidMiner terhadap 3 (tiga) jenis Algoritma, yaitu Algoritma C4.5, naïve Bayes. dan Neural Network, menunjukkan tingkat perbedaan hasil yang begitu beragam. pada pengujian saat ini Algoritma C4.S mencapai tingkat akurasi yang cukup tinggi yaitu 99.05%. Bahkan untuk kriteria lain seperti Precision dan Recall, masih berada diatas kedua algoritma lainnya. Serta jika dilihat dari sisi waktu proses juga tidak kalah dengan algoritma lain Berikut Tabel perbandingan dari ketiga Algoritma yang digunakan sebagai pengujian :

Tabel 4. Perbandingan Algoritma C4.5, Naive Bayes, dan Neural Network

Algoritma	Accuracy	Precision	Recall
Decission Tree	99.05%	96.33%	93.83%
Naive Bayes	90.62%	-	0.00%
Neural Network	95.02%	73.10%	87.83%

Hasil Perhitungan manual Akurasi Algoritma Decission Tree.

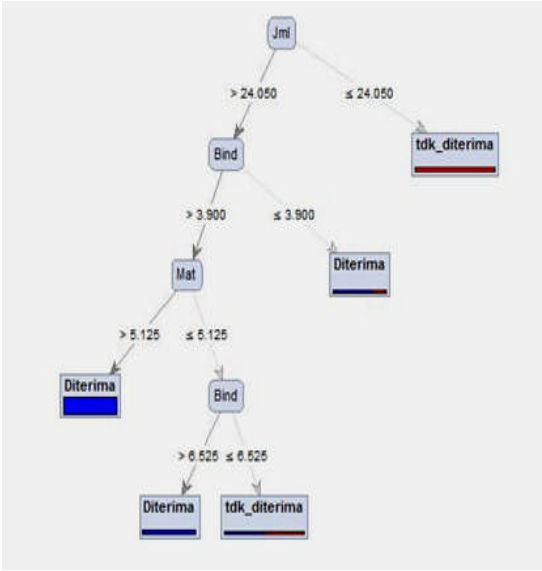
$$Akurasi = \frac{471 + 46}{522}\ x\ 100\% = 90.04\%$$

Dari hasil perhitungan manual dengan hasil dari software simulasi menggunakan rapidminer tidak menunjukan hasil perbedaan yang signifikan.

Secara keseluruhan bahwa Algoritma C4.5 atau Decision Tree ini merupakan salah satu jenis Algoritma yang sesuai dengan data-data yang disajikan, sehingga dapat memperoleh tingkat akurasi yang cukup tinggi

dibanding dengan penggunaan algoritma pembanding lainnya, yaitu algoritma Naive Bayes dan Neural Network.

Berikut *Graph View Tree*, hasil generate RapidMiner, dengan menggunakan jenis Algoritma C4.5:



Gambar 2. Pohon Keputusan untuk Seleksi Masuk

Dalam pengujian menggunakan Algoritma C4.5 menghasilkan pohon keputusan (Decisian Tree) seperti diatas, yang menggambarkan bahwa jika Jumlah Nilai UN siswa <24.050, maka siswa tersebut dinyatakan tidak diterima. Akan tetapi jika jumlah nilai siswa pendaftar itu >24,050, maka siswa dinyatakan diterima, walaupun pada salah satu mata pelajaran siswa tersebut nilainya kurang, tetapi jika nilai mata pelajaran lain tinggi, sehingga jumlah nilai keseluruhan mencapai nilai lebih dari 24.050, maka siswa tersebut diterima, dan jika nilai matematika <= 5.125 dan nilai bahasa Indonesia <= 6.525 maka tidak diterima.

Kesimpulan dan Saran

1. Kesimpulan

Untuk melakukan penentukan prediksi tingkat akurasi siswa baru di SMA Negeri 1 Cibeber yang diterima dilakukan dengan menggunakan Algoritma C4.5 mencapai tingkat akurasi yang cukup tinggi yaitu mencapai sebesar 99.05%. Hal ini menunjukan tingkat akurasi yang cukup tinggi dari data penyeleksian nilai yang disajikan dalam pengelompokan mata pelajaran dalam Ujian Nasional, sehingga dapat dikatakan proses seleksi siswa baru menggunakan metode ini merupakan metode yang tepat untuk digunakan dalam metode komputasi untuk seleksi masuk siswa ke SMA Negeri 1 Cibeber.

2. Saran

Di masa yang akan dating penelitian ini bisa ditambah lagi atribut prediktornya yang mempengaruhi hasil prediksi penerimaan siswa baru, misalnya prestasi akademik dan non akademik. jadi tidak hanya nilai UN. Dan jumlah nilai UN.

Ketika preparasi data menggunakan algoritma *feature selection* mungkin akan memberikan *performance* lebih bagus.

Daftar Pustaka

[1] M. S. Suhartinah and Ernastuti, "GRADUATION PREDICTION OF GUNADARMA UNIVERSITY STUDENTS USING ALGORITHM AND NAIVE BAYES C4 . 5 ALGORITHM," pp. 1–15, 2010.

[2] H. Jiawei, *Data Mining Consepts and Technique*, Second Edi. Elsevier Inc., 2006, p. 5.

[3] I. H. Witten and E. Frank, *Data Mining*, Second edi. Elsevier, 2005, p. 9.

[4] F. A. Hermawati, *Data Mining*. Yogyakarta: Andi Offset, 2009, p. 9.

[5] P. P. Widodo, R. T. Handayanto, and Herlawati, *Penerapan Data Mining dengan Matlab*. 2013.

[6] O. Maimon and L. Rokach, *Data Mining and Kenoledge Discovery Handbook*, Second Edi. Springer, 2010, p. 9.

[7] S. S. Imas and H. ismail Mohd, "Hotspot Occurrences Classification using Decision Tree Method," in *ICT and Knowledge Engineering*, 2010, pp. 46–50.

[8] C. . KrishnaVeni and T. S. Rani, "On the Classification of Imbalanced Datasets," *Computer Sciences and Technology*, vol. 2, pp. 145–148, 2011.