**MANAJEMEN TEKNOLOGI INFORMASI**

**ANALISIS JURNAL IMPLEMENTASI KNOWLEDGE IN DATABASES (KDD) MENGENAI BIDANG KELULUSAN MAHASISWA DI BERBAGAI UNIVERSITAS**

****

**Oleh :**

**IKAMARTHA DWAZAR**

**09031281621042**

**PROGRAM STUDI SISTEM INFORMASI**

**FAKULTAS ILMU KOMPUTER**

**UNIVERSITAS SRIWIJAYA**

**2018**

**Knowledge Discovery in Databases (KDD)**

***Data Mining*** adalah Serangkaian proses untuk menggali nilai tambah berupa informasi yang selama ini tidak diketahui secara manual dari suatu basisdata dengan melakukan penggalian pola-pola dari data dengan tujuan untuk memanipulasi data menjadi informasi yang lebih berharga yang diperoleh dengan cara mengekstraksi dan mengenali pola yang penting atau menarik dari data yang terdapat dalam basisdata.

***Data mining*** biasa juga dikenal nama lain seperti : Knowledge discovery (mining) in databases (KDD), ekstraksi pengetahuan (knowledge extraction) Analisa data/pola dan kecerdasan bisnis (business intelligence) dan merupakan alat yang penting untuk memanipulasi data untuk penyajian informasi sesuai kebutuhan user dengan tujuan untuk membantu dalam analisis koleksi pengamatan perilaku, secara umum definisi data-mining dapat diartikan sebagai berikut

* Proses penemuan pola yang menarik dari data yang tersimpan dalam jumlah besar.
* Ekstraksi dari suatu informasi yang berguna atau menarik (non-trivial, implisit, sebefumnya belum diketahui potensial kegunaannya) pola atau pengetahuan dari data yang disimpan dalam jumfah besar.
* Ekplorasi dari analisa secara otomatis atau semiotomatis terhadap data-data dalam jumlah besar untuk mencari pola dan aturan yang berarti.

**Konsep Data Mining**

***Data mining*** sangat perlu  dilakukan terutama dalam mengelola Data yang sangat besar untuk memudahkan aktifitas recording suatu transaksi dan untuk proses data warehousing agar dapat memberikan informasi yang akurat bagi penggunanya

Alasan utama mengapa data mining sangat menarik perhatian industri informasi dalam beberapa tahun belakangan ini adalah karena tersedianya data dalam jumlah yang besar dan semakin besarnya kebutuhan untuk mengubah data tersebut menjadi informasi dan pengetahuan yang berguna karena sesuai fokus bidang ilmu ini yaitu melakukan kegiatan mengekstraksi atau menambang pengetahuan dari data yang berukuran/berjumlah besar, informasi inilah yang nantinya sangat berguna untuk pengembangan. berikut langkah-langkahnya :

1. Data cleaning (untuk menghilangkan noise data yang tidak konsisten) Data integration (di mana sumber data yang terpecah dapat disatukan)
2. *Data selection* (di mana data yang relevan dengan tugas analisis dikembalikan ke dalam database)
3. Data transformation (di mana data berubah atau bersatu menjadi bentuk yang tepat untuk menambang dengan ringkasan performa atau operasi agresi)
4. Knowledge Discovery (proses esensial di mana metode yang intelejen digunakan untuk mengekstrak pola data)
5. Pattern evolution (untuk mengidentifikasi pola yang benar-benar menarik yang mewakili pengetahuan berdasarkan atas beberapa tindakan yang menarik)
6. Knowledge presentation (di mana gambaran teknik visualisasi dan pengetahuan digunakan untuk memberikan pengetahuan yang telah ditambang kepada user).

**ANALISIS JURNAL IMPLEMENTASI KNOWLEDGE IN DATABASES (KDD) MENGENAI BIDANG KELULUSAN MAHASISWA DI BERBAGAI UNIVERSITAS**

1. **Penerapan Data Mining Untuk Evaluasi Kinerja Akademik Mahasiswa Menggunakan Algoritma Naive Bayes Classifier**

Penelitian ini difokuskan untuk mengevaluasi kinerja akademik mahasiswa pada tahun ke-2 dan diklasifikasikan dalam kategori mahasiswa yang dapat lulus tepat waktu atau tidak. Kemudian dari klasifikasi tersebut, sistem akan memberikan rekomendasi solusi untuk memandu mahasiswa lulus dalam waktu yang paling tepat dengan nilai optimal berdasarkan histori nilai yang telah ditempuh mahasiswa. Input dari sistem ini adalah data induk mahasiswa dan data akademik mahasiswa. Sampel mahasiswa angkatan 2005-2009 yang sudah dinyatakan lulus akan digunakan sebagai data *training* dan *testing*. Sedangkan data mahasiswa angkatan 2010-2011 dan belum lulus akan digunakan sebagai data target. Data input akan diproses menggunakan teknik *data mining* algoritma *Naive Bayes Classifier* (NBC) untuk membentuk tabel probabilitas sebagai dasar proses klasifikasi kelulusan mahasiswa. Output dari sistem ini berupa klasifikasi kinerja akademik mahasiswa yang diprediksi kelulusannya dan memberikan rekomendasi untuk proses kelulusan tepat waktu atau lulus dalam waktu yang paling tepat dengan nilai optimal. *Naive Bayes Classifier* (NBC) NBC merupakan salah satu algoritma dalam teknik data mining yang menerapkan teori Bayes dalam klasifikas. Teorema keputusan Bayes adalah adalah pendekatan statistik yang fundamental dalam pengenalan pola *(pattern recoginition)*. Naive bayes didasarkan pada asumsi penyederhanaan bahwa nilai atribut secara konditional saling bebas jika diberikan. nilai *output*. Dengan kata lain, diberikan nilai *output*, probabilitas mengamati secara bersama adalah produk dari probabilitas individu.

Pada penelitian ini digunakan data sebagai berikut:

* 1. Data *training* dan data *testing*

Data ini akan digunakan sebagai proses *mining* dan pengujian, berupa sampel data induk dandata akademik mahasiswa angkatan 2005-2009yang sudah dinyatakan lulus. Data ini memiliki atribut NIM, jenis kelamin, asal sekolah, jalurmasuk, nilai ujian nasional, gaji orangtua, IPsemester 1-4, IPK semester 1-4, dan keteranganlulus.

* 1. Data target

Data ini berupa sampel data induk data akademik mahasiswa angkatan 2010-2011 yang diasumsikan belum lulus. Data ini memiliki atribut NIM, jenis kelamin, asal sekolah, jalur masuk, nilai ujian nasional, gaji orangtua, IP semester 1-4, dan IPK semester 1-4. Setelah proses *mining*, data ini akan memiliki kelas berdasarkan tabel probabilitas yang diperoleh dari data *training*.

* 1. Data riwayat matakuliah

Data ini digunakan untuk mengevaluasi data target ketika diklasifikasikan lulus tidak tepat waktu. Data ini akan dianalisis untuk memberikan rekomendasi dalam proses perkuliahan berikutnya. Data ini memiliki atribut NIM, kode matakuliah, dan nilai.

1. **PENERAPAN ALGORITMA C4.5 UNTUK KLASIFIKASI PREDIKAT KELULUSAN MAHASISWA FAKULTAS KOMUNIKASI DAN INFORMATIKA UNIVERSITAS MUHAMMADIYAH SURAKARTA.**

IPS dan LAIN (selain IPA dan IPS), variabel X2 terdiri dari 2 nilai *class* dengan tipe *polynomial*, yaitu PRIA dan WANITA, Variabel X3 dibuat menjadi 2 nilai *class* yang bertipe *polynomial,* yaitu SURAKARTA (jika asal sekolah se-Karesidenan Surakarta) dan LUAR (jika asal sekolah di luar Karesidenan Surakarta), Variabel X4 dibagi menjadi 2 nilai *class* yang bertipe *polynomial*, yaitu SKS

18 dan SKS > 18, Variabel X5 terdiri dari 2 nilai *class* dengan tipe *polynomial*, yaitu YA (jika pernah menjadi asisten) dan TIDAK (jika belum pernah menjadi asisten).

Tahap kelima adalah pengumpulan data. Banyaknya data mahasiswa yang diperoleh dari IT UMS dihitung dengan metode Slovin (terlihat pada persamaan 1) untuk menentukan jumlah sampel yang digunakan sebagai data pengujian proses data mining.

|  |  |  |  |
| --- | --- | --- | --- |
| *n* = |  | *N* | ................................................................................(1) |
| 1 + *Ne*2 | |
|  |

Keterangan:

*n =* jumlah sampel

*N =* jumlah keseluruhan data / populasi

*e =* galat kesalahan (ditentukan sebesar 5%)

Data mahasiswa Fakultas Komunikasi dan Informatika yang diperoleh dari IT UMS seluruhnya berjumlah (N) 2358 mahasiswa. Sesuai dengan rumus Slovin, dapat ditentukan jumlah sampel yang diambil yaitu sebanyak 341 data mahasiswa.

Tahap keenam adalah mengolah data. Olah data yang dilakukan menurut meliputi pemisahan atribut-atribut yang diperlukan untuk proses data mining, standarisasi data (*preprocessing*), hingga pengubahan data-data *real* menjadi data-data nominal dengan tipe *binomial* maupun *polynomial* sesuai dengan kebutuhan data mining.

Tahap terakhir adalah melakukan analisis data. Klasifikasi mahasiswa berdasarkan predikat kelulusannya dilakukan dengan metode C4.5 berdasarkan entropi dari masing-masing atribut yang telah ditentukan dengan persamaan 2 dan 3. Lesmana (2012) menjelaskan bahwa *decision tree* merupakan salah satu metode belajar yang sangat populer dan banyak digunakan secara praktis. Metode ini merupakan metode yang berusaha menemukan fungsi–fungsi pendekatan yang bernilai diskrit dan tahan terhadap data – data yang memiliki kesalahan *(noisy data)* serta mampu mempelajari ekspresi – ekspresi disjunctive seperti ekspresi OR.

|  |  |  |  |
| --- | --- | --- | --- |
| *gain*( *y*, *A*)= *entropi*( *y*)−∑ | *yc* | *entropi*( *yc* )............................................... | (2) |
| *y* |
| *c* ∈*nilai*( *A*) |  |  |
| *Entropi*( *y*)= − *p*1log2 *p*1− *p*2log2 | *p*2− − *pn* log2 *pn* ........................................ | | (3) |

klasifikasi predikat kelulusan mahasiswa Fakultas Komunikasi dan Informatika UMS. Variabel yang paling tinggi pengaruhnya terhadap terhadap predikat kelulusan adalah partisipasi mahasiswa menjadi asisten. Interpretasi hasil penelitian mengindikasikan bahwa variabel yang perlu digunakan sebagai pertimbangan bagi Fakultas Komunikasi dan Informatika UMS untuk memperoleh tingkat predikat kelulusan yang maksimal adalah peran serta mahasiswa untuk menjadi asisten. Secara umum probabilitas predikat “Cumlaude” pada kelompok mahasiswa yang pernah menjadi asisten lebih tinggi dibandingkan dengan yang tidak pernah menjadi asisten. Seorang mahasiswa dari kelompok yang pernah menjadi asisten jika berasal dari jurusan IPA semasa sekolah menengah atas memiliki probabilitas predikat kelulusan “Cumlaude” yang lebih tinggi dibandingkan dengan mahasiswa dari jurusan lainnya.

1. ***DATA MINING* MENGGUNAKAN ALGORITMA *NAÏVE BAYES* UNTUK KLASIFIKASI KELULUSAN MAHASISWA UNIVERSITAS DIAN NUSWANTORO**

Deployment merupakan tahapan akhir dalam pembuatan laporan hasil kegiatan data mining. Laporan akhir yang berisi mengenai pengatahuan yang diperoleh atau pengenalan pola pada data dalam proses data mining.

Berdasarkan penelitian yang dilakukan, telah dihasilkan suatu pola, informasi, dan pengetahuan baru dalam proses data mining untuk klasifikasi kelulusan mahasiswa berdasarkan data mahasiswa Fakultas Ilmu Komputer Udinus angkatan 2009. Dari penelitian tersebut dihasilkan suatu pola, informasi, dan pengetahuan baru sesuai dengan tujuan data mining yaitu pola perhitungan data mining yang berisi data training dan data testing serta mencari probabilitas dari setiap atribut berdasarkan data training dan data testing untuk menghasilkan suatu informasi baru, apakah pada data mahasiswa Fakultas Ilmu Komputer angkatan 2009 lebih banyak kelas tahun lulus yang tepat waktu atau kelas tahun lulus tidak tepat waktu. Kemudian untuk menguji tingkat keakurasiannya maka digunakan Rapidminer sebagai alat bantu dalam proses pengujian tingkat akurasi dari klasifikasi tersebut.

Dari proses perhitungan data mining menggunakan algoritma naïve bayes dan tingkat keakurasian, dihasilkan suatu informasi baru yaitu perhitungan data mining berdasarkan mahasiswa Fakultas Ilmu Komputer angkatan 2009, menunjukkan kelas tahun lulus “yes” / tepat waktu dengan total perkalian prior probability senilai 0, sedangkan kelas tahun lulus “no” / tidak tepat waktu dengan total perkalian prior probability senilai 0.00055. Untuk tingkat akurasi berdasarkan proses klasifikasi menggunakan algoritma naïve bayes, dengan melalui semua tahapan dipastikan tidak ada bagian – bagian penting yang terlewatkan, dihasilkan tingkat akurasi sebesar 82.08 %.

Berdasarkan hasil perhitungan data mining dan proses pengujian tingkat akurasi dengan menggunakan Rapidminer, dapat ditarik kesimpulan bahwa angkatan 2009 kelas tahun lulus “no” / tidak tepat waktu lebih besar dari kelas tahun lulus “yes” / tepat waktu. Sedangkan analisa yang dilakukan terhadap tingkat akurasi menggunakan algoritma naïve

bayes menunjukkan bahwa nilai yang dihasilkan oleh algoritma naïve bayes memiliki tingkat kekuatan yang cukup tinggi. Hal ini di buktikan dengan hasil perhitungan yang mencapai nilai 82.08 %, Nilai 82. 08 % membuktikan bahwa model yang dibangun dapat digunakan untuk melakukan klasifikasi kelulusan mahsiswa. Nilai 82.08 % bisa juga di sebabkan oleh kekurang kompleksan data yang mengakibatkan model dapat memprediksi dengan akurat.

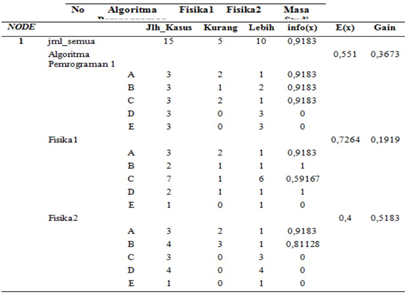
Berdasarkan perhitungann data mining menggunakan algoritma naïve bayes, dapat ditarik kesimpulan bahwa kelas tahun lulus “no” / tidak lulus tepat waktu lebih besar daripada kelas tahun lulus “yes” / lulus tepat waktu.

Dari hasil observasi terhadap dataset mahasiswa Udinus Fakultas Ilmu Komputer angkatan 2009 dan melalui proses perhitungan menggunakan metode klasifikasi naïve Bayes dengan atribut yang telah dijelaskan di pembahasan sebelumnya, didapatkan sebuah hasil bahwa nilai akurasi terhadap klasifikasi kelulusan sebesar 82.08 %. Dimana 82.08 % bisa juga disebabkan oleh kurang kompleksitas data yang mengakibatkan model dapat memprediksi cukup akurat.

1. **ANALISIS DAN PENERAPAN ALGORITMA C4.5 DALAM DATA MINING UNTUK MEMPREDIKSI MASA STUDI MAHASISWA BERDASARKAN DATA NILAI AKADEMIK**

Analisis Algoritma C4.5

Perhitungan berdasarkan kasus data nilai mata kuliah mahasiswa untuk menentukan masa studi dilakukan dengan mengimplementasikan Algoritma C4.5 terlihat pada tabel di bawah ini. Perhitungan akan dilakukan berdasarkan 3 nilai mata kuliah yaitu : Algoritma Pemrograman 1, Fisika 1 dan Fisika 2. Dengan nilai masing-masing mata kuliah adalah A, B, C, D dan E. Mata kuiah ini dijadikan atribut awal dalam implementasi Algoritma C4.5 dan atribut tujuannya adalah masa studi dengan nilai dari atribut masa studi adalah “kurang” dan “lebih”. Kurang artinya lulus kurang dari lima tahun dan lebih artinya lulus lebih dari lima tahun.



Dari tabel di atas terdapat gain untuk masing- masing mata kuliah. Dimana hasil gain dari masing-masing mata kuliah tersebut diantaranya adalah: Algoritma Pemrograman 1 = 0,3673, Fisika 1 = 0,1919 dan Fisika 2 = 0,5183

Sesuai dengan ketentuan dalam algoritma C4.5, setiap atribut yang memiliki gain tertinggi akan menjadi root atau node. Dari ketiga mata kuliah tersebut, matakuliah yang memiliki gain tertinggi adalah Fisika 2, maka mata kuliah Fisika 2 menjadi root. Setelah diperoleh root maka akan dilihat infoa(x) dari masing nilai atribut mata kuliah yang jadi root yaitu Fisika 2.

dapat dilihat nilai infoa(x) untuk nilai mata kuliah Fisika 2 diantaranya adalah:

1. Fisika 2 dengan nilai A : 0,9183

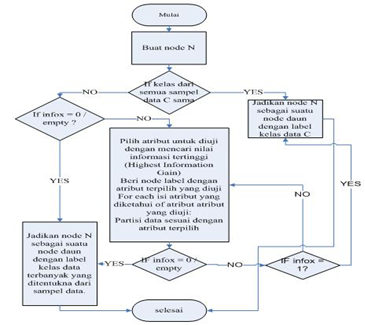
2. Fisika 2 dengan nilai B : 0,81128

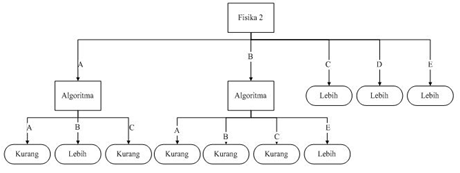
3. Fisika 2 dengan nilai C : 0

4. Fisika 2 dengan nilai D : 0

5. Fisika 2 dengan nilai E : 0

**Flowchart Sistem**





Gambar di atas merupakan hasil akhir dari proses pembentukan tree. Setelah selesai pembentukan tree merupakan proses pembuatan aturan berdasarkan tree yang terbentuk. Aturan dari tree di atas adalah sebagai berikut:

1. Jika Fisika 2 nilai A dan Algoritma Pemrograman 1 nilai A maka masa studinya adalah kurang
2. Jika Fisika 2 nilai A dan Algoritma Pemrograman 1 nilai B maka masa studinya adalah lebih
3. Jika Fisika 2 nilai A dan Algoritma Pemrograman 1 nilai C maka masa studinya adalah kurang
4. Jika Fisika 2 nilai B dan Algoritma Pemrograman 1 nilai A maka masa studinya adalah kurang
5. Jika Fisika 2 nilai B dan Algoritma Pemrograman 1 nilai B maka masa studinya adalah kurang
6. Jika Fisika 2 nilai B dan Algoritma Pemrograman 1 nilai C maka masa studinya adalah kurang
7. Jika Fisika 2 nilai B dan Algoritma Pemrograman 1 nilai E maka masa studinya adalah lebih
8. Jika Fisika 2 nilai C maka masa studinya adalah lebih
9. Jika Fisika 2 nilai D maka masa studinya adalah lebih
10. Jika Fisika 2 nilai E maka masa studinya adalah lebih

Aplikasi untuk memprediksi masa studi mahasiswa berhasil dirancang dan diimplementasikan menggunkan algoritma C4.5. Aplikasi tersebut dapat memprediksi masa studi mahasiswa.

Berdasarkan training dan pengujian kemudian dilakukan analisis maka dapat diambil beberapa kesimpulan bahwa jumlah *data training* mempengaruhi persentase kecocokan atau keakurasian. Persentase kecocokan hasil pengujian yang didapatkan dari *data training* 112 lebih kecil di bandingkan dengan *data training* 70 untuk berapapun kombinasi atribut. Demikian pula untuk semua konfigurasi atribut, persentase kecocokan hasil pengujian dengan *data training* 112 lebih kecil di bandingkan dengan *data training* 70.

1. **IMPLEMENTASI DATA MINING UNTUK MEMPREDIKSI MASA STUDI MAHASISWA MENGGUNAKAN ALGORITMA C4.5 (STUDI KASUS: UNIVERSITAS DEHASEN BENGKULU)**

Ada beberapa tahap dalam menggunakan Rapid Miner:

1) Untuk menganalisa, dibutuhkan data training. Data tarining yang akan dimasukkan kedalam Rapid Miner bisa dalam format .csv, .xls, .mdb dan laini-lain. Data yang digunakan penulis adalah data dalam format .csv.

2) Buka program Rapid Miner, kemudian akan muncul tampilan awal. Untuk memasukkan data training yang telah dibuat sebelumnya, pilih menu File – Import Data – Import CSV file.

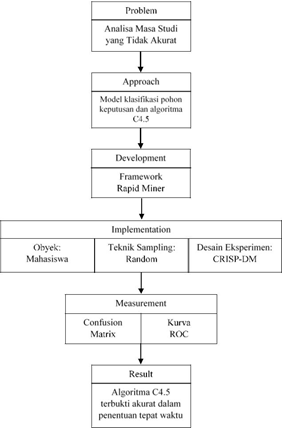
3) Tampilan jendela Data wizard dengan total 5 langkah. Pada langkah ke-1 ini tentukan nama file yang berisi data training dalam direktori kemudian pilih Next

4) Data training yang sebelumnya disimpan, akan tersimpan otomatis ke dalam Repositories. Pilih tab Repositories – NewLocalRepository – data\_training. Geser data\_training ke area Main Process. Untuk menambahkan model, pilih tab

Operators- Modelling-Classification and Regression-Tree Induction-Decision Tree. Geser Decision Tree ke area Main Process dan hubungkan

5) Untuk melihat hasilnya, pilih process – Run maka akan tampil hasil berupa pohon keputusan.

Gambar Kerangka pemikiran.



Berdasarkan hasil pengujian sistem yang telah dilakukan di Universitas Dehasen Bengkulu yang dilakukan penulis dapat diketahui bahwa terdapat kelebihan dan kekurangan dari sistem lama dan sistem baru tersebut.

Dengan menggunakan sistem lama, masih menggunakan sistem perkiraan saja sehingga tingkat kesalahan untuk memprediksi tingkat kelulusan mahasiswa tepat waktu atau tidak masih besar

Sedangkan dengan menggunakan teknik data mining ini tingkat kesalahan dalam memprediksi masa studi mahasiswa tersebut lulus tepat waktu atau tidak dapat dikurangi dengan tingkat kesalahan 5 % .

Dengan menggunakan sistem lama, pengambilan keputusannya masih kompleks dan global. Sedangkan setelah menggunakan sistem baru daerah pengambilan keputusan yang sebelumnya kompleks dan sangat global, dapatndiubah menjadi lebih simpel dan spesifik.

Dengan menggunakan sistem lama, seorang penguji masih kesulitan dalam menganalisis untuk mengestimasikan baik itu distribusi dimensi tinggi ataupun parameter tertentu dari distribusi kelas tersebut. Sedangkan dengan sistem baru, dalam analisis, 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 pohon keputusan dapat menghindari munculnya permasalahan ini dengan menggunakan criteria yang jumlahnya lebih sedikit pada setiap node internal tanpa banyak mengurangi kualitas keputusan yang dihasilkan.

Dengan menggunakan sistem lama, membutuhkan waktu lama dalam penentuan keputusan. Sedangkan dengan sistem baru waktu yang diperlukan dalam penentuan keputusan lebih cepat dengan menggunakan teknik data mining tersebut.