**Manajemen Teknologi Informasi**

**Penerapan Knowledge Discovery in Database (KDD)**

**Pada Penentuan Jurusan Siswa SMA**

****

**Oleh:**

**Nama : Elly Kurniati**

**NIM : 09031281621034**

**Kelas : SI Reg 5B**

**PROGRAM STUDI SISTEM INFORMASI REGULER**

**FAKULTAS ILMU KOMPUTER**

**UNIVERSITAS SRIWIJAYA**

**2018**

**Knowledge Discovery in Database**

Knowledge Discovery in Database (KDD) berhubungan dengan teknik integrasi dan penemuan ilmiah, interprestasi dan visualisasi dari pola-pola sejumlah kumpulan data.

Knowledge Discovery in Database adalah keseluruhan proses non-trivial untuk mencari dan mengidentifikasi pola (pattern) dalam data, dimana pola yang ditemukan bersifat sah, baru, dapat bermanfaat dan dapat dimengerti.

Knowledge Discovery in Database (KDD) adalah proses menentukan informasi yang berguna serta pola-pola yang ada dalam data. Informasi ini terkandung dalam basis data yang berukuran besar yang sebelumnya tidak diketahui dan potensial bermanfaat (Han & Kamber, 2006 dalam Baskoro, 2010). Berikut merupakan tahapan dari KDD:

****

1. **Seleksi Data (Data Selection)**

Menciptakan himpunan data target, pemilihan himpunan data, atau memfokuskan pada subset variabel atau sampel data, dimana penemuan (discovery) akan dilakukan.

*Selection* (seleksi / pemilihan) data dari sekumpulan data operasional perlu dilakukan sebelum tahap penggalian informasi dalam *Knowledge Discovery Database (KDD)* dimulai. Data seleksi yang akan digunakan untuk proses *Data Mining,* disimpan dalam suatu berkas, terpisah dari basis data operasional.

1. **Pemilihan Data *(Preprocessing / Cleaning)***

Pemprosesan pendahuluan dan pembersihan data merupakan operasi dasar seperti penghapusan noise dilakukan.  Sebelum proses data mining dapat dilaksanakan, perlu dilakukan proses cleaning pada data yang menjadi fokus KDD.

Proses Preprocessing mencakup antara lain membuang duplikasi data, memeriksa data yang inkonsisten dan memperbaiki kesalahan pada data, seperti kesalahan cetak *(tipografi).* Juga dilakukan proses *enrichment*, yaitu proses “memperkaya” data yang sudah ada dengan data atau informasi lain yang relevan dan diperlukan untuk *KDD,* seperti data atau informasi *eksternal*.

1. **Transformasi *(Transformation)***

Pada fase ini yang dilakukan adalah mentransformasi bentuk data yang belum memiliki entitas yang jelas ke dalam bentuk data yang *valid* atau siap untuk dilakukan proses *Data Mining.*

1. ***Data Mining***

Pada fase ini yang dilakukan adalah menerapkan algoritma atau metode pencarian pengetahuan. Ini adalah langkah penting di mana teknik kecerdasan diterapkan untuk mengekstrak pola informasi yang berpotensi berguna dari data yang dipilih.

Teknik, metode, atau algoritma dalam data mining sangat bervariasi. Pemilihan metode atau algoritma yang tepat sangat bergantung pada tujuan dan proses KDD secara keseluruhan.

1. **Interpretasi / Evaluasi *(Interpratation / Evaluation)***

Pada fase terakhir ini yang dilakukan adalah proses pembentukan keluaran yang bersumber pada proses *Data Mining* pola informasi (Nofriansyah, 2014).Pola informasi yang dihasilkan dari proses data mining perlu ditampilkan dalam bentuk yang mudah mimengerti oleh pihak yang berkepentingan. Tahap ini merupakan bagian dari proses KDD yang mencakup pemeriksaan apakah pola atau informasi yang ditemukan bertentangan dengan fakta atau hipotesa yang ada sebelumnya.

Dalam salah satu tahap KDD terdapat data mining, data mining merupakan suatu kegiatan yang meliputi pengumpulan, pemakaian data historis untuk menentukan keteraturan, pola atau hubungan dalam set data berukuran besar (Santosa, 2007). Beberapa teknik yang sering disebut-sebut dalam literatur data mining antara lain yaitu *association rule mining, clustering,* klasifikasi, *neural network, genetic algorithm* dan lain-lain.

**Analisis Jurnal Penerapan KDD pada Penjurusan Siswa SMA**

1. Klasifikasi Penjurusan Program Studi Sekolah Menengah Atas dengan Algoritma Naive Bayes Classifier pada SMA Negeri 1 Subah (**Nikmatul Hidayah**)

 Penjurusan siswa kelas X SMA yang akan naik ke kelas XI bertujuan mengarahkan peserta didik agar lebih fokus mengembangkan kemampuan dan minat yang dimiliki. Strategi ini diharapkan dapat memaksimalkan potensi, bakat atau talenta individu, sehingga juga akan memaksimalkan nilai akademisnya. Penentuan jurusan akan berdampak terhadap kegiatan akademik selanjutnya dan mempengaruhi pemilihan bidang ilmu atau studi bagi siswa-siswi yang ingin melanjutkan ke perguruan tinggi nantinya. Jurusan yang tidak tepat bisa sangat merugikan siswa dan masa depannya.

 Atas dasar permasalahan tersebut, maka dilakukan penelitian untuk menerapkan metode data mining yaitu algoritma Naïve Bayes Classifier untuk mengklasifikasikan jurusan program studi. Naïve Bayes adalah suatu metode pengklasifikasian data dengan model statistic yang dapat digunakan untuk memprediksi probabilitas keanggotaan pada suatu kelas dan digunakan ntuk menganalisis dalam membantu tercapainya hasil keputusan terbaik suatu permasalahan dari sejumlah alternatif.

 Penelitian ini merupakan penelitian eksperimen. Untuk melakukan eksperimen dibutuhkan alat bantu berupa spesifikasi software dan hardware. Dalam pengujian model eksperiman dengan jumlah 11 atribut prediktor yaitu Kimia, Fisika, Biologi, Ekonomi, Sosiologi, Ekonomi, Geografi, Sosiologi, Jumlah IPA, Jumlah IPS , IQ dan Minat. Semua atribut tersebut akan dihitung peluangnya oleh naïve bayes sehingga menghasilkan prediksi jurusan program studi.

 Pada Percobaan pertama adalah dengan menerapkan algoritma naïve bayes untuk memprediksi jurusan dengan data set yang terdiri dari 300 examples, 4 special attributes meliputi final, confidence IPA, confidence IPS, Prediction Final, 13 regular attributes meliputi Kimia, Fisika, Biologi, Jumlah IPA, Ekonomi, Geografi, Sosiologi, Jumlah IPS, IQ, dan Minat. Dengan total 17 atribut yang akan diproses dan dimodelkan dengan algoritma naïve bayes classifier. Kemudian inputan proses klasifikasi dengan 8 atribut berupa nilai Kimia, nilai Fisika, nilai Biologi, nilai Ekonomi, nilai Geografi, nilai
Sosiologi, IQ, dan Minat.

 Berdasarkan data training sebanyak 300 exampleset dengan 8 atribut berupa nilai Kimia, nilai Fisika, nilai Biologi, nilai Ekonomi, nilai Geografi, nilai Sosiologi, IQ, dan Minat yang dimodelkan dengan algoritma naïve bayes diperoleh hasil akurasi sebanyak 98.00 % dengan rincian sebagai berikut:

* Jumlah true positif (tp) adalah 142 jumlah siswa yang diklasifikasikan ke dalam kelas IPA dan false negative (fn) sebanyak 3 jumlah siswa yang diklasifikasikan ke dalam kelas IPA tetapi masuk kelas IPS. Jumlah true negative (tn) adalah sebanyak 152 jumlah siswa yang diklasifikasikan ke dalam kelas IPS dan false positif (fp) sebanyak 3 jumlah siswa yang diklasifikasikan IPS tetapi masuk kelas IPA. Berdasarkan hasil confusion matrix, menunjukkan bahwa hasil akurasi yang didapat dengan menggunakan algoritma naïve bayes classifier adalah sebesar 98.00% .

1. **Metode Klasifikasi dengan Algoritma Naive Bayes Untuk Rekomendasi Penjurusan SMA Terang Bangsa (Andrew Yova Kencana, Setia Astuti)**

Penelitian yang dilakukan oleh Andrew Yova Kencana dan Setia Astuti ini dilakukan di SMA Terang Bangsa, dimana sekolah ini menerapkan kurikulum 2013, penjurusan dilakukan mulai dari kelas X. Berbeda dengan penelitian sebelumnya penelitian ini dilakukan dengan nilai-nilai SMP mereka. Penelitian ini bertujuan untuk membuat aplikasi data mining klasifikasi menggunakan algoritma Naive Bayes. Algoritma Naive Bayes ini adalah algoritma yang menggunakan teknik probabilitas sederhana berdasarkan penerapan aturan bayes digunakan untuk mencari nilai probabilitas tertinggi untuk mengklasifikasi data uji pada kategori paling tepat. Atribut yang di gunakan dalam penelitian ini adalah NIS (No Induk Siswa), nilai rata-rata pelajaran IPA, nilai rata rata Matematika , nilai rata-rata Bahasa Inggris, dan minat.

 Data yang di dapat di SMA Terang bangsa sebanyak 205. Data terlebih dahulu dikelompokkan dengan kategori: *High* apabila nilai diatas 84, *Middle* nilai antara 60-84, kemudian *Low* apabila nilai dibawah 60.

 Ketika Mendapat probabilitas IPA dan IPS bernilai 0 maka dapat di klasifikasian sebagai kelas IPS. Di dapat hasil 81 Percobaan yang menyatakan Siswa mendapat kelas IPA terdapat 18 percobaan dan 63 percobaan menghasilkan Nilai IPA.

 Berbeda dengan penelitian yang dilakukan oleh Nikmatul Hasanah dengan metode yang sama penelitian ini tidak mendeskripsikan hasil dengan persentase dan pengujian yang dilakukan terbilang cukup sederhana.

1. **Penerapan Algoritma Klasifikasi Data Mining ID3 Untuk Menentukan Penjurusan Siswa SMAN 6 Semarang (Obbie Kristanto)**

 Penelitian yang dilakukam oleh Obie Kristanto ini dilakukan pada SMAN 6 Semarang. Ada beberapa permasalahan yang melatar-belakangi penelitian ini dari hal yang sepele sampai masalah penting yang nantinya berdampak pada pemilihan minat pada saat di perguruan tinggi. Tidak jarang juga siswa-siswi yang asal-asalan dalam menentukan jurusan yang akan mereka ambil, tidak mempunyai tujuan setelah mereka lulus sekolah menengah atas, dan lain-lain berakibat pada menurunnya prestasi belajar mereka. Untuk mengatasi masalah tersebut siswa hanya berkonsultasi secara langsung kepada guru wali kelas, BP atau dengan orangtua masing-masing yang dilakukan secara manual sehingga memakan waktu yang cukup lama.

 Tetapi seiring dengan perkembangan teknologi hal itu dapat diatasi dengan teknik pengelompokan data didapat dari hasil data mining. Informasi dan knowledge yang didapat tersebut dapat digunakan untuk mengetahui suatu pola dalam suatu data yang banyak, terlebih lagi besarnya kebutuhan untuk mengubah data tersebut menjadi informasi yang berguna.

 Pada kasus ini penulis menggunakan metode data mining ID3. Dalam pengimplementasian algoritma ID3 dengan dataset siswa SMAN6 Semarang yang telah diolah, terlebih dahulu dilakukan proses perhitungan entropi dan information gain secara manual. Hail ini dikarenakan agar aplikasi yang digunakan sudah mempunyai dan memenuhi algoritma ID3 yang dipakai dalam klasifikasi untuk penjurusan SMAN 6 Semarang. Dari perhitungan tersebut diketahui bahwa atribut yang paling besar dalam menentukan jurusan adalah peminatan diri. Atribut ini sekaligus menjadi root (akar) dalam decision tree tersebut. Atribut ini memiliki nilai information gain yang paling besar yaitu 0.157314902. Setelah mengetahui entropi dari masing-masing nilai atribut yang ada, entropi tersebut digunakan untuk menghitung nilai information gain dari setiap atribut yang ada.

 Dengan hasil root yang telah diperoleh, yaitu atribut peminatan diri maka dilakukan pengujian algoritma ID3 pada aplikasi dengan cara mengambil sampel dari dataset sebanyak 20 record untuk diujikan dalam aplikasi yang telah dibuat. Label yang telah ada dihilangkan sehingga target yang akan diujikan class / labelnya tidak diketahui agar dapat mengetahui tingkat keakuratan dari aplikasi yang telah dibuat. Diambil 20 data uji dari 371 record yang ada dalam dataset peminatan siswa kelas X kurikulum 2013. Dari 20 data uji yang telah diujikan terdapat 4 data uji yang tidak cocok dengan hasil pada data yang ada pada Guru BP. 16 data uji lainnya cocok dengan hasil data yang ada pada Guru BP. Sehingga akurasi pada aplikasi penjurusan siswa SMAN 6 Semarang adalah sebesar 80%.

1. **Penerapan Algoritma C4.5 Untuk Prediski Jurusan Siswa SMA Negeri 3 Rembang (Faid Ari Prasatya)**

Penelitian yang dilakukan oleh Faid Ari Prasatya ini dilakukan pada SMA Negeri 3 Rembang. Penggunaan metode didasarkan pada tingkat akurasi Algoritma C4.5 terbilang tinggi seperti penelitian yang dilakukan oleh Obie Kristanto (2014) dalam penelitiannya menerapkan algoritma ID3 yang menunjukkan hasil akurasi penelitian sebesar 80%.

 Kemudian penelitian yang dilakukan Liliana Swastina, Penerapan algoritma C4.5 untuk penentuan jurusan mahasiswa. Algoritma C4.5 sering digunakan, Memiliki tingkat akurasi yang tinggi. Algoritma C4.5 prediksi lebih akurat dari algoritma Naive Bayes. Tingkat keakuratan 93,31% dan Akurasi Jurusan 82,64%

 Secara umum algoritma C4.5 untuk membangun keputusan meliputi pilih atribut sebagai akar, buat cabang untuk masing – masing nilai, bagi kasus dalam cabang, ulangi proses untuk masing – masing cabang sampai semua kasus pada cabang memiliki kelas yang sama.

 Data yang digunakan dalam penelitian ini adalah sumber data primer. Data diperoleh dari Sekolah Menengah Atas Negeri 3 Rembang. Data yang dikumpulkan yaitu data siswa baru tahun ajaran 2014/2015 dengan attribute Nama, Nomor Pendaftaran, Asal Sekolah, Tanggal Lahir, Jenis Kelamin, Rata Nilai Mat, Rata Nilai IPA, Nilai Ujian Mat, Nilai Ujian IPA, Piagam, Peminatan, Jurusan. Dan selanjutnya adalah pemilihan atribut data yang digunakan.

 Jumlah true positif (tp) sebanyak 107 record, false positif (fp) sebanyak 34 record, jumlah true negative (tn) sebanyak 108 record dan jumlah false negative (fn) sebanyak 39 record. Akurasi yang dihasilkan dari datashet Siswa Baru SMAN 3 Rembang Tahun Ajaran 2014/2015 dengan pengujian metode algoritma C4.5 pada Rapid Miner sebesar 74,65%

1. Implementasi Data Mining dengan Algoritma C4.5 untuk Penjurusan Siswa (Studi Kasus: SMA Negeri 1 Pontianak)

 Penelitian yang dilakukan oleh Beti Novianti, Tedy Rismawan dan Syamsyul Bhari ini dilakukan pada SMA Negeri 1 Pontianak. Penjurusan sangat penting untuk mengelompokkan siswa sesuai dengan kemampuan (nilai), bakat dan minat yang relatif sama agar selanjutnya, pelajaran yang diberikan kepada siswa lebih fokus dan terarah.

 Dalam sistem penjurusan ini, siswa diberi kesempatan memilih jurusan, baik itu jurusan “IPA” atau “IPS”, sebelum nantinya di klasifikasi keputusan jurusannya menurut nilai dan minat masing masing siswa, sehingga nanti pilihan siswa dan hasil akhir keputusan jurusan siswa dapat berbeda karena menyesuaikan kemampuan (nilai) siswa tersebut.

 Tes penjurusan diikuti oleh seluruh siswa kelas X yang kemudian akan disortir bersama dengan nilai rapor dan nilai UN SMP masing-masing siswa. Dalam proses penjurusan, siswa diberi kesempatan memilih jurusan, baik itu jurusan IPA atau IPS, sebelum nantinya di klasifikasi keputusan jurusannya menurut nilai dan minat masing-masing siswa, sehingga nanti pilihan siswa dan hasil akhir keputusan jurusan siswa dapat berbeda karena menyesuaikan kemampuan (nilai) siswa tersebut.

 Algoritma C4.5 merupakan salah satu algoritma yang digunakan untuk membentuk pohon keputusan. Metode pohon keputusan mengubah fakta yang sangat besar menjadi pohon keputusan yang merepresentasikan aturan. Aturan dapat dengan mudah dipahami dengan bahasa alami. Dan mereka juga dapat diekspresikan dalam bentuk bahasa basis data seperti *Structured Query Language* untuk mencari *record* pada kategori tertentu. Algoritma Pohon Keputusan C4.5 atau *Classification version* 4.5 adalah pengembangan dari algoritma ID3. Oleh karena pengembangan tersebut, algoritma C4.5 mempunyai prinsip dasar kerja yang sama dengan algoritma ID3.

 Data penelitian yang digunakan adalah data siswa kelas X SMA Negeri 1 Pontianak Tahun Ajaran 2015/2016. Data didapatkan dari Wakil Kepala Sekolah (WaKa)
Kurikulum SMA Negeri 1 Pontianak. Kemudian data tersebut diklasifikasi sebagai berikut: A untuk nilai antara 86-100, B antara 71-85, C antara 56-70, D antara 41-55, dan E kurang dari sama dengan 40.

 Dari hasil pengujian aplikasi, semua dataset berhasil diklasifikasikan ke kelas keputusan “IPA” dan “IPS” Jumlah seluruh data yang dinilai ialah sebanyak 39 data. Untuk kolom negatif diisi sebagai nilai “IPS”, sedangkan untuk kolom positif diisi sebagai nilai “IPA”. Hasil keputusan awal sebelum dilakukan proses C4.5 diperoleh keputusan kelas IPS sebanyak 8 siswa sedangkan keputusan kelas IPA sebanyak 31 siswa. Untuk hasil setelah proses C4.5 ialah diperoleh keputusan kelas IPS sebanyak 10 Siswa dan Keputusan kelas IPA sebanyak 29 siswa. Dari data diatas diketahui tingkat akurasinya sebesar 89,74%.

**Kesimpulan:**

Berdasarkan kelima jurnal diatas Penerapan Knowledge Discovery in Database pada Penjurusan Siswa SMA menunjukkan tingkat akurasi hasil yang cukup tinggi yaitu 74,65% keatas. Diantara semua metode yang digunakan metode Algoritma C4.5 menunjukkan tingkat akurasi yang paling tinggi. Hal ini dikarenakan Algoritma C4.5 merupakan pengembangan dari Algoritma ID3.