Penerapan Metode Content-Based Filtering Pada Sistem Rekomendasi Kegiatan Ekstrakulikuler (Studi Kasus di Sekolah ABC)

Firmahsyah[1], Tiur Gantini[2]

Fakultas Teknologi Informasi, Universitas Kristen Maranatha Jl. Suria Sumantri 65, Bandung

[1]fns.find16@gmail.com

[2]tiur.gantini@it.maranatha.edu

Abstract— ***ABC School is an educational organization. The "ABC" used as alias of the original organization's name. They operate their routine activities without information systems, especially for extracurricular activies. System recommendation was made to help the school. It provides recommendations of extracurricular activities***
which is more suitable with student interest. Primary data source is obtained by interview and observation with headmaster of ABC School. Primary data used for analysis and design sytem. The analysis use one of data mining technique, which is content based recommendation. The content based recommendation method that used for this research is Naïve Bayes. The result of this research is a recommendation system to show probability the extracurricular of each student with some chosen attributes.

Keywords— **attribute, information gain, Naïve Bayes, recommendation system**

I. PENDAHULUAN

Sekolah ABC adalah sebuah lembaga pendidikan yang berada pada naungan sebuah yayasan. Sekolah ini melayani beberapa jenjang pendidikan mulai dari *Play Group*, Taman Kanak-kanak (TK), Sekolah Dasar (SD), Sekolah Menengah Pertama (SMP), dan Sekolah Menengah Atas (SMA). Setiap tahun, setiap jenjang pendidikan menerima siswa baru, baik siswa dari sekolah yang sama maupun siswa dari sekolah lain.

Selain mengikuti kegiatan belajar mengajar secara normal, siswa pun diupayakan untuk memiliki kegiatan ekstrakurikuler di luar jam pelajaran sekolah. Kegiatan ekstrakurikuler telah ada sejak jenjang pendidikan terendah, dalam hal ini adalah *Play Group*. Agar menumbuhkan rasa disiplin dan komitmen kepada ekstrakurikuler yang dipilihnya, siswa harus berkomitmen untuk menekuni satu atau lebih ekstrakurikuler setiap semesternya dengan tidak berubah ekstrakurikuler di pertengahan semester. Namun pada prakteknya banyak orang tua siswa yang memaksa untuk berhenti dari ekstrakurikuler tertentu karena dirasakan putera/i mereka kurang cocok di dalam mengikuti ekstrakurikuler yang telah dipilih.

Pada penelitian kali ini akan dirancang sebuah sistem yang dapat memberikan rekomendasi untuk pemilihan ekstrakurikuler siswa. Dengan harapan dapat membantu pihak orang tua agar memilih ekstrakurikuler yang lebih tepat bagi putera/i mereka dan tidak ada lagi yang keluar atau berhenti dari ekstrakurikuler sebelum masa berakhir yang telah ditentukan.

Oleh karena itu tujuan dari penelitian ini adalah sebagai berikut:

1) Menggunakan data mining dalam menganalisis karakteristik siswa yang telah mengikuti ekstrakulikuler dari jenjang pendidikan PG, TKA dan TKB.

2) Menggunakan penerapan metode Naïve Bayes dengan Information Gain sebagai seleksi fitur. Seleksi fitur digunakan sebagai teknik untuk meningkatkan akurasi rekomendasi ekstrakurikuler. Hasil analisis akurasi akan dipergunakan sebagai model data mining pada aplikasi. Mode tersebut digunakan untuk melakukan rekomendasi ekstrakurikuler kepada setiap siswa.

II. LANDASAN TEORI

A. Sistem Rekomendasi

Sistem rekomendasi adalah fitur-fitur dan teknik-teknik pada perangkat lunak yang menyediakan sesuatu hal yang berguna untuk *user* [1].Sistem rekomendasi juga menyediakan rekomendasi-rekomendasi dari beberapa item yang berpotensi menarik untuk pengguna. Rekomendasi-rekomendasi yang diberikan erat kaitannya dengan pengambilan keputusan, seperti item apa saja yang harus dibeli, musik seperti apa yang harus didengarkan, dan berita apa yang harus dibaca [2]. Dalam hal ini, item adalah sebuah objek yang direkomendasikan [1].

Sebuah sistem rekomendasi harus dapat membangun dan memelihara *user model* atau *user profile* yang berisi ketertarikan pengguna. Sebagai contoh, pada sebuah toko

buku, sistem menyimpan buku apa saja yang pengunjung lihat atau beli di masa lalu. Hal ini untuk memprediksi buku-buku lainnya yang mungkin diminati oleh pengunjung [3].

Terdapat tiga teknik rekomendasi utama yaitu: *collaborative filtering*, *content-based filtering,* dan *knowledge-based recommendation*. *Collaborative filtering* merupakan metode yang merekomendasikan sebuah item yang berdasarkan pada kemiripan ketertarikan antar pengguna [2]. Sistem rekomendasi *content-based* merekomendasikan item yang mirip dengan yang disukai user sebelumnya. Nilai kesamaan antar item dihitung berdasarkan fitur yang ada pada setiap konten [1]. Sistem rekomendasi *knowledge-based* merekomendasikan item berdasarkan domain pengetahuan yang spesifik tentang bagaimana fitur-fitur yang ada pada suatu item dapat memenuhi kebutuhan pengguna dan berguna bagi pengguna.Nilai kesamaan dihitung berdasarkan seberapa besar nilai kesamaan antara kebutuhan pengguna dengan rekomendasi yang ada [1].Terdapat dua pendekatan dalam metode *knowledge-based recommendation*, yaitu *case-based* dan *constraint-based recommendation*.Kesamaan dari kedua pendekatan ini adalah pengguna harus memberikan permintaan terlebih dulu. Kemudian sistem akan mengidentifikasi solusi yang sesuai dengan permintaan pengguna [1].

Salah satu metode yang sering digunakan pada sistem *content-based recommendation* adalah metode Naïve Bayes [1].Algoritma Klasifikasi Naïve Bayes adalah pengklasifikasi statistik.Algoritma ini dapat memprediksi kemungkinan-kemungkinan anggota kelas.Klasifikasi Naïve Bayes mengasumsikan pengaruh dari sebuah nilai atribut pada kelas yang diberikan adalah independen dari nilai-nilai pada atribut lainnya [4].Naïve Bayes banyak digunakan untuk klasifikasi teks dalam *machine learning* yang didasarkan pada fitur probabilitas [5].Pendekatan Bayesian digunakan untuk mengukur probabilitas dari asumsi-asumsi yang ada. Pada statistik Bayesian, parameter-parameter dianggap sebagai variabel acak; dan data dianggap sebagai sesuatu yang akan diketahui klasifikasinya. Parameter-parameter dianggap datang dari sebuah distribusi yang memiliki nilai kemungkinan, dan Bayesian bertujuan untuk mengobservasi data untuk memberikan informasi pada parameter yang memiliki nilai kemungkinan yang besar [6].

Seleksi fitur merupakan sebuah bagian penting untuk meningkatkan kinerja dari pengklasifikasi data [7].Banyaknya fitur secara optimal tereduksi menurut sebuah kriteria evaluasi tertentu [8]. Seleksi fitur dapat dibedakan menjadi tiga, yaitu *filter model*, *wrapper model*, dan *embedded model*[9] [10]. Metode Filter mengevaluasi kualitas dari fitur yang diseleksi secara independen dari algoritma klasifikasi. Metode Wrapper membutuhkan penerapan dari algoritma klasifikasi untuk mengevaluasi kualitas klasifikasi. Metode Embedded menerapkan seleksi fitur selama pembelajaran dari parameter-parameter yang optimal [11].

Metode filter menyeleksi atribut yang relevan sebelum berpindah pada fase pembelajaran selanjutnya. Atribut yang terlihat paling signifikan dipilih untuk pembelajaran, sementara sisanya yang lain disisihkan. Salah Satu metode yang digunakan untuk melakukan Filter adalah Information Gain [10].

B. Data Mining

Menurut Kamber, Hian, dan Pei [4], *Data Mining* adalah sebuah proses untuk menemukan pengetahuan yang menarik seperti pengelompokan, perubahan-perubahan pola, dari sebuah basis data, data warehouse atau tempat penyimpanan informasi lainnya. Sementara itu menurut Larose [6]*Data Mining* sebagai suatu proses eksplorasi dan analisis secara otomatis maupun semiotomatis terhadap data dalam jumlah besar dengan tujuan menemukan pola dan aturan yang berarti. Sedangkan menurut Kantardzic [9], *Data Mining* adalah keseluruhan proses dari pengaplikasian sebuah metodologi berbasis komputer, termasuk teknik-teknik baru, untuk menemukan pengetahuan dari sebuah data. Saat ini banyak yang mengartikan *data mining* dengan istilah *knowledge discovering* ‘temu pengetahuan’.

Terdapat beberapa proses untuk mendapatkan sebuah pemodelan data yang tepat dan memiliki tingkat akurasi yang baik. Beberapa proses pada *data mining* menurut Kamber, Hian, dan Pei [4] di antaranya adalah:

1) Pembersihan data: untuk menghapus data yang mengganggu atau inkonsisten;

2) Pengintegrasian data: menggabungkan beberapa sumber data;

3) Penyeleksian data: memilih data yang relevan untuk dianalisis yang didapatkan dari basis data

4) Transformasi data: data ditransformasikan dan dikonsolidasikan ke format yang cocok untuk melakukan proses penggalian informasi dengan menggunakan operasi agregasi.

5) Data mining: proses terpenting yang menggunakan suatu metode untuk mengekstraksi pola-pola yang terdapat pada data.

6) Evaluasi pola: mengidentifikasi pola-pola yang merepresentasikan basis pengetahuan atau ukuran.

C. Naïve Bayes

Menurut Kamber, Hian, dan Pei [4]; dalam permasalahan klasifikasi diperlukan untuk menentukan nilai $P(X \mid H)$ yang merupakan peluang dari hipotesis (H) seperti *data tuple* (X) yang dimiliki oleh kelas (C). Menurut Kamber, Hian, dan Pei [4]; dan Larose [6]; Teorema Bayes direpresentasikan pada persamaan:

$$P(H \mid X) = \frac{P(X \mid H)P(H)}{P(X)}$$

$P(X \mid H)$ adalah nilai *posterior probability* dari X yang memiliki kondisi H; $P(H)$ adalah nilai *prior probability* dari H; $P(X)$ adalah nilai *prior probability* dari (X) dan $P(H \mid X)$ adalah nilai *posterior probability* dari H yang memiliki kondisi X.

Sebuah perhitungan peluang $P(X \mid C_i)$ sebagai perkalian dari peluang-peluang $P(x_1 \mid C_i).P(x_2 \mid C_i).(...).P(x_k \mid C_i)$ berdasarkan asumsi dari *class-conditional independence*. Peluang-peluang tersebut memungkinkan untuk bernilai nol sehingga perkalian yang didapat pun bernilai nol. Peluang nol akan menghanguskan *posterior probability* pada atribut yang lainnya.

Menurut Larose [6] terdapat penyesuaian untuk frekuensi bernilai nol, yakni dengan rumus:

$$\frac{n_c + n_{equiv} p}{n + n_{equiv}}$$

Dengan n_c adalah frekuensi dari atribut tersebut, n_{equiv} adalah konstanta yang mewakilkan besarnya ukuran sampel, p adalah *prior probability*. Nilai p dapat dicari dengan persamaan $p = \frac{1}{k}$ dengan k adalah banyaknya kelas target dan n adalah banyaknya keseluruhan data.

D. Information Gain

Information Gain adalah suatu cara untuk mengukur seberapa efektif suatu atribut tersebut dalam mengklasifikasikan sebuah kelas. Menurut Kamber, Han, dan Pei [4], *Information Gain* adalah sebuah pengukuran yang digunakan untuk menyeleksi atribut. Sebelum mencari nilai dari *Information Gain*, nilai *Entropy* harus dicari terlebih dulu. Menurut Suyanto [12], nilai *Entropy* digunakan sebagai suatu parameter untuk mengukur heterogenitas dari suatu kumpulan sampel data. Jika kumpulan sampel data semakin heterogen, maka nilai *Entropy*-nya semakin besar. Secara matematis nilai *Entropy* dirumuskan dengan:

$$Entropy(S) = \sum_{i}^{c} -p_i \log_2 p_i$$

Dengan c adalah jumlah nilai yang ada pada kelas klasifikasi dan p_i adalah jumlah sampel untuk kelas i

Menurut Suyanto [12], *Information Gain* adalah sebuah metode untuk mengukur efektivitas suatu atribut dalam mengklasifikasikan data. Secara matematis, *Information Gain* dari suatu atribut A, dituliskan:

$$Gain(S, A) \equiv Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_v)$$

Dengan A adalah atribut atau tupel, V menyatakan suatu nilai yang mungkin untuk atribut A, $Values(A)$ adalah himpunan nilai-nilai yang mungkin untuk A, $|S_v|$ adalah jumlah sampel untuk nilai V, $|S|$ adalah jumlah seluruh sampel data, dan $Entropy(S_v)$ adalah *Entropy* untuk sampel-sampel yang memiliki nilai v.

E. Penelitian Terkait

Beberapa peneliti telah melakukan beberapa penelitian yang berkaitan dengan system rekomendasi dengan menggunakan algoritma Naïve Bayes.Penelitian yang dilakukan Tewari, Kumar, dan Barman [13] mengenai teknik merekomendasikan buku berdasarkan opini. Calon pembeli buku merasa kesulitan untuk membaca *review* dari sebuah buku. Oleh karena itu, dalam penelitian tersebut digunakan Algoritma Naïve Bayes untuk merekomendasikan buku yang disusun berdasarkan peringkat terbaik.

Penelitian yang dilakukan oleh Ghazanfar dan Prugel-Bennett [14] mengenai percobaan beberapa algoritma untuk mengukur akurasi dari Data Testing yang akan digunakan untuk merekomendasikan sesuatu. Salah satu algoritma yang digunakan adalah Naïve Bayes. Dalam penelitian tersebut juga dibahas mengenai *Content-Based Filtering: Feature Extraction and Selection*. *Feature Selection* mengeliminasi banyaknya fitur dengan mengeliminasi kata-kata yang tidak berguna atau tidak memiliki pembeda yang kuat pada saat mengklasifikasikan data. Salah Satu pendekatan *Feature Selection* yang dapat digunakan adalah Information Gain.

III. METODE PENELITIAN

Penelitian ini dilakukan dengan langkah-langkah sebagai berikut seperti digambarkan pada Gambar 1. Data mentah yang digunakan adalah data siswa dan data peserta ekstrakurikuler. Setelah data tersebut disiapkan maka akan dilanjutkan ke data *preprocessing*. *Preprocessing* yang dilakukan adalah sebagai berikut[4]:

A. Memvalidasi Data

Proses ini merupakan aktivitas untuk mengidentifikasi dan menghapus data yang ganjil, data yang tidak konsisten, serta mengisi data yang tidak lengkap.

B. Mengintegrasikan Data

Proses ini merupakan aktivitas untuk menggabungkan kedua jenis data yang dimiliki. Data siswa dan data siswa mengikuti Ekstrakurikuler digabungkan agar mendapatkan informasi tambahan.

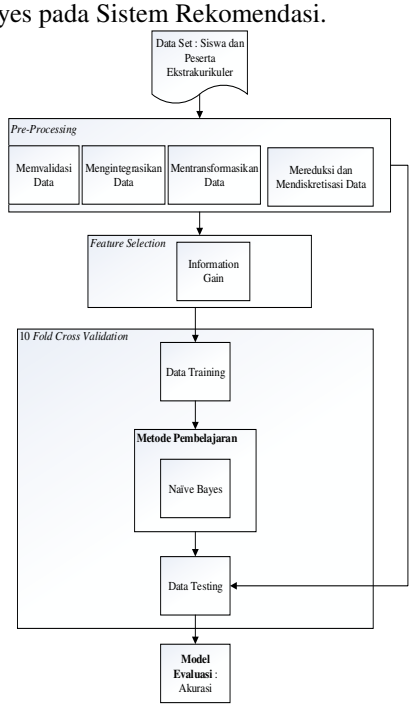
C. *Mentransformasikan Data*

Setelah diintegrasikan, beberapa nilai dari data gabungan diganti dengan nilai yang lebih informatif dan diharapkan dapat meningkatkan akurasi dari algoritma Naïve Bayes.Misalnya, nilai Tempat dan Tanggal yang semula menyatu dalam satu atribut dipecah menjadi atribut apakah siswa tersebut lahir di Bandung dan atribut bulan lahir siswa.

D. *Mereduksi dan Mendiskretisasi Data*

Atribut nomor induk, nama siswa, alamat lengkap, nomor telepon, nama orang tua, nomor identitas dirasa kurang informatif karena setiap siswa memiliki nomor induk, nama siswa, alamat lengkap, nomor telepon, nama orang tua yang berbeda satu sama lain. Sedangkan nomor identitas ekstrakurikuler redudan karena nilai tersebut sebenarnya sudah diwakili oleh nama ekstrakurikuler.

Feature Selection yang diusulkan adalah metode dengan jenis filter yakni Information Gain. Penelitian akan menghasilkan nilai akurasi dan membandingkannya untuk didapatkan satu model terbaik. Model terbaik ini akan dipergunakan untuk melakukan rekomendasi ekstrakurikuler masing-masing siswa. Sehingga terdapat dua proses yang diusulkan yakni proses analisis data untuk mendapatkan model dan proses implementasi Algoritma Naïve Bayes pada Sistem Rekomendasi.



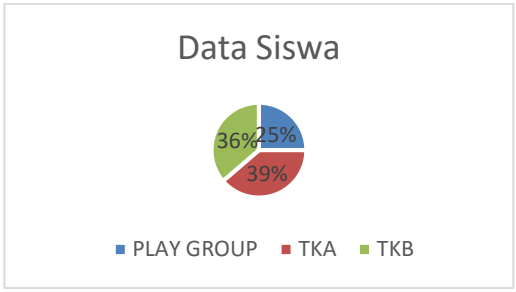
Gambar1. Proses *Tahapan Penelitian*

Proses analisis pada gambar 1 dengan diagram alir; yang terdiri dari melakukan *preprocessing* data, melakukan feature selection, melakukan proses *10- fold cross validation*, dan melakukan evaluasi dengan melihat akurasi yang dihasilkan. Proses 10-*fold cross validation* merupakan proses untuk memisahkan 10 persen data untuk *data testing* dan *90* persen data untuk *data training* [15]. Hasil evaluasi akan menentukan model yang dipilih untuk merekomendasikan ekstrakurikuler pada siswa baru. Implementasi Algoritma Naïve Bayes untuk Melakukan Rekomendasi Ekstrakurikuler Siswa digambarkan pada gambar 1.

IV. HASIL PENELITIAN

A. *Pengumpulan Data Mentah*

Data mentah yang digunakan untuk penelitian disajikan di dalam bentuk grafik pada gambar 2 yang menggambarkan sebaran data berdasarkan jenjang pendidikan. Dan gambar 3 yang menggambarkan banyaknya siswa yang mengikuti setiap ekstrakurikuler yang ada di sekolah ABC.

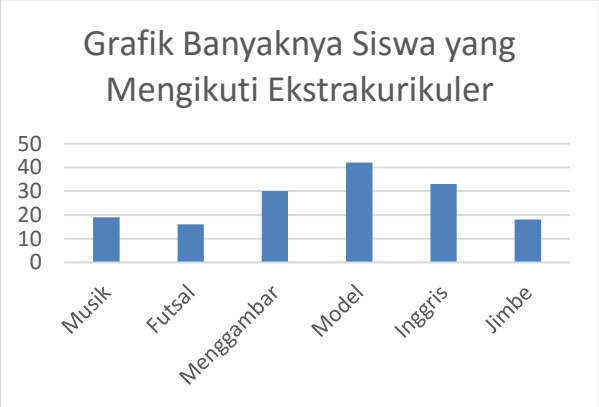


Gambar 2. Sebaran data siswa menurut jenjang pendidikan

Data yang digunakan dalam penelitian kali ini diambil dari tiga jenjang pendidikan yang ada di sekolah ABC, yaitu Play Group/PG (Taman Bermain), Taman Kanak-kanak Kecil yang disebut TKA dan Taman Kanak-kanak besar yan disebut TKB. Sebaran data siswa dapat dilihat pada gambar 2, yang terdiri dari 25% siswa atau 31 siswa adalah siswa play group; 36% atau 45 siswa adalah siswa TKB; dan 39% atau 48 siswa adalah siswa TKA. Data yang dikumpulkan ada 124 baris data. Dan setiap baris data tersebut memiliki 11 atribut yang terdiri dari atribut nomor urut, nomor induk siswa, nama siswa, jenis kelamin, tempat dan tanggal lahir, agama, alamat lengkap, nomor telepon, nama orang tua, pendidikan terakhir orang tua, dan pekerjaan orang tua. Dan sesuai dengan permintaan pihak sekolah, maka untuk menjaga privasi data, atribut nama siswa, alamat, telepon, dan nama orang tua berturut-turut diganti dengan nilai SISWA-n, ALAMAT-n, TELEPON-n, dan ORANG TUA-n. Nilai n adalah urutan dari data tersebut.

Selain jenjang pendidikan, data penelitian ini juga membutuhkan data peserta ekstrakurikuler. Data rekapitulasi peserta ekstrakurikuler dapat dilihat pada gambar 3, yang terdiri dari 19 orang siswa mengikuti ekstrakurikuler 417musik; 16 orang siswa mengikuti ekstrakurikuler futsal;

30 orang siswa mengikuti ekstrakurikuler menggambar; 42 orang siswa mengikuti ekstrakurikuler model; 33 siswa mengikuti ekstrakurikuler Inggris; dan 18 orang siswa mengikuti ekstrakurikuler jimbe. Total keseluruhan data peserta ekstrakurikuler adalah 158 data.



Gambar 3. Grafik banyaknya siswa yang mengikuti Ekstrakurikuler

Hasil rekap yang diperoleh pada gambar 3, merupakan rangkuman dari data set setiap ekstrakulikuler yang dikumpulkan. Karena Data Siswa yang mengikuti ekstrakurikuler sangat banyak, maka di dalam jurnal ini hanya disajikan contoh untuk ekstrakurikuler musik. Data ekstrakulikuler musik dapat dilihat pada TABEL I. Di dalam data peserta ekstrakurikuler.

TABEL I.

DATA SET SISWA YANG MENGIKUTI EKSKUL MUSIK

A	B	C	D	E	F	G
1	E1	SISWA-MUSIK-1	L	4 Thn	PG 1	Rp170,000
2	E1	SISWA-MUSIK-2	P	5 Thn	PG2	Rp170,000
3	E1	SISWA-MUSIK-3	L	5 Thn	PG 2	Rp170,000
4	E1	SISWA-MUSIK-4	L	4 Thn	PG2	Rp170,000
5	E1	SISWA-MUSIK-5	P	5 Thn	A1	Rp170,000
6	E1	SISWA-MUSIK-6	P	5 Thn	A1	Rp170,000
7	E1	SISWA-MUSIK-7	L	5 Thn	A1	Rp170,000
8	E1	SISWA-MUSIK-8	L	5 Thn	A1	Rp170,000
9	E1	SISWA-MUSIK-9	L	5 Thn	A2	Rp170,000
10	E1	SISWA-MUSIK-10	L	6 Thn	A2	Rp170,000
11	E1	SISWA-MUSIK-11	P	5 Thn	A2	Rp170,000
12	E1	SISWA-MUSIK-12	P	6 Thn	A2	Rp170,000
13	E1	SISWA-MUSIK-13	P	5 Thn	A2	Rp170,000
14	E1	SISWA-MUSIK-14	L	6 Thn	B1	Rp170,000
15	E1	SISWA-MUSIK-15	L	6 Thn	B1	Rp170,000
16	E1	SISWA-MUSIK-16	L	6 Thn	B2	Rp170,000
17	E1	SISWA-MUSIK-17	L	6 Thn	B2	Rp170,000
18	E1	SISWA-MUSIK-18	L	6 Thn	B2	Rp170,000
19	E1	SISWA-MUSIK-19	L	6 Thn	B2	Rp170,000

Pada Tabel I, dicatat tujuh atribut. Ketujuh atribut tersebut yakni: A = nomor urut, B = nomor identitas, C =nama siswa, D =jenis kelamin, E =umur, F =kelas, dan G = harga. Untuk melindungi privasi, nama siswa diganti nilainya dengan SISWA-<JENIS EKSTRAKURIKULER>-n. <JENIS EKSTRAKURIKULER> diganti dengan nama ekstrakurikuler yang diikuti oleh siswa yang bersangkutan dan n adalah nomor urut untuk setiap baris data.

B. Memvalidasi Data

Berdasarkan 124 data siswa dan 158 data peserta ekstrakurikuler, langkah selanjutnya adalah memvalidasi data. Memvalidasi data merupakan aktivitas untuk mengidentifikasi dan menghapus data yang ganjil, data yang tidak konsisten, dan data yang tidak lengkap.

Dalam proses ini, dilakukan penghapusan data untuk atribut yang kosong. Penyeragaman data untuk data pekerjaan orang tua, karena terdapatnilai yang sama namun direpresentasikan dengan berbeda. Contohnya adalah data pada pekerjaan orang tua. Atribut dengan nilai "pegawai swasta" direpresentasikan dengan beberapa nilai yakni: "peg. Swasta", "Karyawan swasta, "pegawai swasta", dan "swasta". Hal ini tentu menambah keberagaman data meskipun pada hakikatnya data tersebut bernilai sama. Sehingga data tersebut diseragamkan nilainya menjadi "pegawai swasta".

Atribut pendidikan terakhir orang tua terdapat nilai yang tidak konsisten. Contohnya adalah nilai "S-1" yang diartikan sebagai strata satu memiliki nilai yang beragam. Nilai atribut pendidikan terakhir yang merujuk pada arti strata satu memiliki nilai "S-1", "S1", dan "sarjana". Nilai-nilai tersebut akhirnya diseragamkan menjadi "S-1". Selain itu, terdapat pula nilai pendidikan terakhir "SMA" yang merupakan kepanjangan dari Sekolah Menengah Atas yang memiliki nilai beragam. Sekolah Menengah Atas dapat direpresentasikan dengan nilai "SMA", "SMU", dan "SMEA". Pada akhirnya nilai pendidikan terakhir "Sekolah Mengengah Atas" direpresentasikan dengan nilai "SMA".

C. Mengintegrasikan dan mentransformasikan data

Teknik mengintegrasikan data adalah dengan menggabungkan jenis data yang dimiliki. Data yang diintegrasikan adalah data siswa dan data peserta ekstrakurikuler. Penggabungan tersebut diidentifikasi dengan nama siswa. Hasil pengintegrasian data menghasilkan17 atribut yang terdiri dari nomor urut, nomor induk siswa, nama siswa, jenis kelamin, tempat dan tanggal lahir, agama, alamat lengkap, nomor telepon, nama orang tua, pendidikan

terakhir orang tua, pekerjaan orang tua, nomor identitas, nama siswa, jenis kelamin, umur, kelas, dan harga.

Setelah diintegrasikan, data gabungan tersebut diganti beberapa nilainya dengan nilai yang lebih informatif dan diharapkan dapat meningkatkan akurasi dari algoritma Naïve Bayes.Semisal nilai dari Tempat dan Tanggal yang semula menyatu dalam satu atribut dipecah menjadi atribut apakah siswa tersebut lahir di Bandung dan atribut bulan lahir siswa. Atribut ekstrakurikuler yang semula tidak ditambahkan pada saat pengintegrasian data ditambahkan. Nantinya atribut ekstrakurikuler menjadi sebuah atribut kelas yang akan diklasifikasikan oleh algoritma Naïve Bayes.

Atribut "kelas" juga mengalami transformasi data. Pada mulanya atribut tersebut memuat kelas berdasarkan jenjang yang berbeda-beda: "TKA-1", "TKA-2", "TKB-1", "TKB-2", "PG-1", "PG-2". Keenam nilai tersebut ditransformasikan sehingga hanya nilai jenjangnya saja yang digunakan. Nilai "TKA-1" dan "TKA-2" menjadi "TKA"; nilai "TKB-1" dan "TKB-2" menjadi "TKB"; dan nilai "PG-1" dan "PG-2" menjadi "PG".

D. Mereduksi dan mendiskretisasi data

Tujuan dari mereduksi data adalah untuk mendapatkan sebuah data set seminimal mungkin namun tetap informatif. Dari data yang telah diintegrasikan dan ditransformasikan maka atribut yang digunakan untuk memulai pengklasifikasian data yaitu atribut jenis kelamin, lahir di bandung, bulan lahir, agama, pendidikan terakhir orang tua, pekerjaan orang tua, umur, jenjang pendidikan, dan ekstrakurikuler.

Sedangkan atribut yang dirasa kurang informatif dan redudan tidak digunakan, yaitu atribut nomor urut, nomor induk, alamat lengkap, nomor telepon, nama orang tua, nomor identitas ekskul, dan harga.

Atribut nomor urut tidak digunakan karena nomor urut hanya menandakan urutan data siswa saja. Atribut nomor induk, nama siswa, alamat lengkap, nomor telepon, nama orang tua, nomor identitas dirasa kurang informatif karena setiap siswa memiliki nomor induk, nama siswa, alamat lengkap, nomor telepon, nama orang tua yang berbeda satu sama lain. Sedangkan nomor identitas ekstrakurikuler dianggap redudan karena nilai tersebut sebenarnya sudah diwakili oleh jenis ekstrakurikuler.

Data yang siap digunakan adalah 158 data. Data ini siap digunakan untuk analisis algoritma Naïve Bayes. Karena data ini cukup banyak maka beberapa baris data saja yang disajikan di sini. Data dapat diliha pada Tabel II.

TABEL II.

DATA SET SISWA UNTUK ANALISIS NAÏVE BAYES

H	I	J	K	L	M	N	O	P
L	4	PG	Katolik	Ya	Januari	S1	Wiraswasta	Musik
P	5	PG	Katolik	Ya	Agustus	S1	Pegawai	Musik
L	5	PG	Katolik	Ya	Januari	S1	Pegawai	Musik
L	4	PG	Kristen	Ya	April	S1	Pegawai	Musik
P	5	TKA	Kristen	Ya	Juni	S1	Pegawai	Musik
P	5	TKA	Kristen	Ya	Oktober	S1	Pegawai	Musik
L	5	TKA	Kristen	Ya	April	S1	Pegawai	Musik
L	5	TKA	Kristen	Ya	September	S1	TNI/POLRI	Musik
L	5	TKA	Kristen	Ya	Oktober	S2	Dosen	Musik
L	6	TKA	Kristen	Ya	November	S1	Pegawai	Musik
P	5	TKA	Kristen	Ya	Desember	SMA	Wiraswata	Musik
P	6	TKA	Katolik	Ya	Juli	S1	Pegawai	Musik
P	5	TKA	Kristen	Ya	September	S1	PNS	Musik
L	6	TKB	Katolik	Ya	September	SMA	TNI/POLRI	Musik
L	6	TKB	katolik	Tidak	Juni	S1	Pegawai	Musik
L	6	TKB	Katolik	Ya	Juli	S1	Pegawai	Musik
L	6	TKB	katolik	Ya	April	SMA	TNI/POLRI	Musik
L	6	TKB	Kristen	Tidak	Mei	S1	Pegawai	Musik
L	6	TKB	Katolik	Tidak	Januari	S1	PNS	Musik
L	6	TKA	Kristen	Ya	Desember	S1	Pelaut	Futsal
L	5	TKA	Katolik	Ya	Juli	D3	Radiosrafer	Futsal
…	..	…	…….	….	…….	….	……	…..
L	6	TKB	Katolik	Ya	April	S1	Pegawai	Futsal

Keterangan Tabel II:
H= Jenis Kelamin; I = umur; J= Jenjang Pendidikan
K = Agama; L = Lahir di Bandung; M = Bulan Lahir
N =Pendidikan Terakhir Orang Tua;
O = Pekerjaan Orang Tua; P= Jenis Ekskur yang dipilih

E. Rancangan Sistem Rekomendasi

Rancangan sistem rekomendasi akan membahas metode pemilihan atribut dan penerapan algoritma naïve bayes. Berikut pembahasannya:

E.1 Metode Pemilihan Atribut

Berdasarkan 158 data pada table II, selanjutnya data tersebut dibagi menjadi dua menjadi Data Training dan Data Testing. Sebanyak 78 Data dijadikan Data Training dan 18 data dijadikan data testing. Data tersebut dipilih secara acak. Dari setiap Data Training maupun Data Testing memiliki panjang kelas target atau jenis ekstrakurikuler yang sama

panjang. Data Training yang digunakan terdapat pada TABEL III.

TABEL III.

DATA TRAINING

Q	R	S	T	U	V	W	X	Y	Z
1	L	4 th	PG	Katolik	Ya	Januari	S1	Wiraswasta	Musik
2	P	5 th	PG	Katolik	Ya	Agustus	S1	Pegawai	Musik
3	L	5 th	PG	Katolik	Ya	Januari	S1	Pegawai	Musik
4	P	5 th	TKA	Kristen	Ya	Juni	S1	Pegawai	Musik
5	P	5 th	TKA	Kristen	Ya	Oktober	S1	Pegawai	Musik
6	L	5 th	TKA	Kristen	Ya	April	S1	Pegawai	Musik
7	L	5 th	TKA	Kristen	Ya	September	S1	TNI/POLRI	Musik
8	L	5 th	TKA	Kristen	Ya	Oktober	S2	Dosen	Musik
9	P	6 th	TKA	Katolik	Ya	Juli	S1	Pegawai	Musik
10	P	5 th	TKA	Kristen	Ya	September	S1	PNS	Musik
11	L	6 th	TKB	Katolik	Ya	September	SMA	TNI/POLRI	Musik
12	L	6 th	TKB	Katolik	Ya	Juli	S1	Pegawai	Musik
13	L	6 th	TKB	Katolik	Tidak	Januari	S1	PNS	Musik
14	L	6 th	TKB	Katolik	Ya	April	S1	Pegawai	Futsal
15	L	6 th	TKA	Kristen	Ya	Desember	S1	Pelaut	Futsal
…		…	…..	…..	…	….	….	….	…
76	L	7 th	TKB	Kristen	Ya	Desember	S1	Pegawai	Jimbe
77	L	6 th	TKB	Kristen	Tidak	Mei	S1	Pegawai	Jimbe
78	L	6 th	TKB	Katolik	Tidak	Januari	S1	PNS	Jimbe

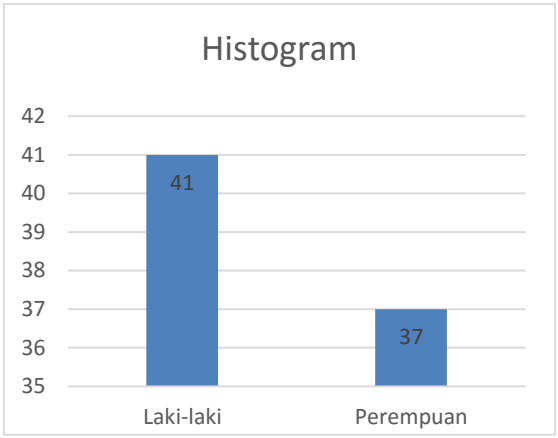
Keterangan Tabel III:

Q = No Urut; R = Jenis Kelaminn; S = umur; T= Jenjang Pendidikan; U = Agama; V = Lahir di Bandung; W= Bulan Lahir; X =Pendidikan Terakhir Orang Tua;

Y = Pekerjaan Orang Tua; Z= Jenis Ekskur yang dipilih

Dari data training tersebut dibuatlah histogram untuk melihat persebaran data dari masing-masing atribut. Adapun histogram dari masing-masing atribut pada data training terdiri dari histogram jenis kelamin, usia, jenjang pendidikan, agama, lahir di bandung, bulan lahir, pendidikan orang tua, pekerjaan orang tua, dan ekstrakurikuler. Masing-masing histogram dapat dilihat pada gambar 4 sampai dengan gambar 12.

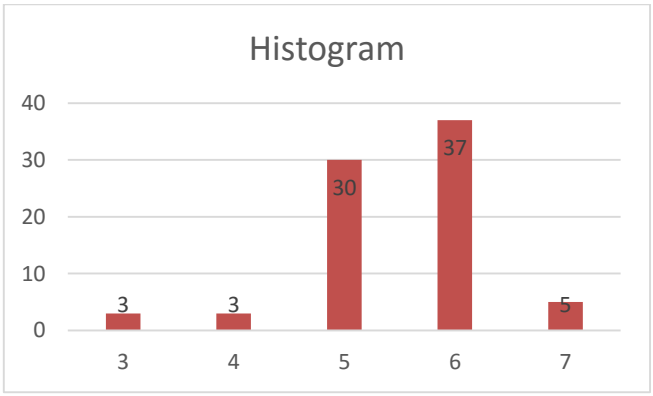
1. Histogram Jenis Kelamin



Gambar 4. Histogram Jenis Kelamin

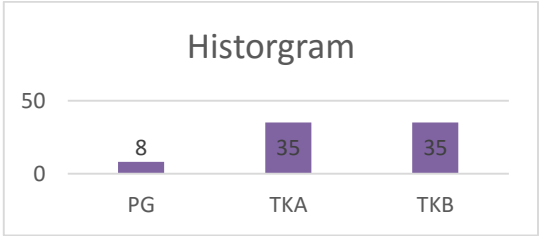
Dalam Histogram Jenis Kelamin didapatkan dua nilai yakni "L" dan "P". Nilai "L" pada atribut Jenis Kelamin berjumlah 40, sedangkan nilai "P" berjumlah 38. Total seluruh data adalah 78.

2. Histogram Usia



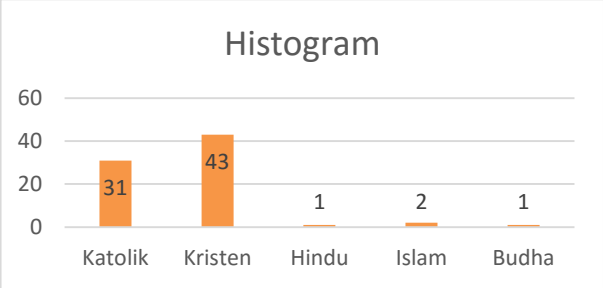
Gambar 5. Histogram Usia

3. Histogram Jenjang Pendidikan



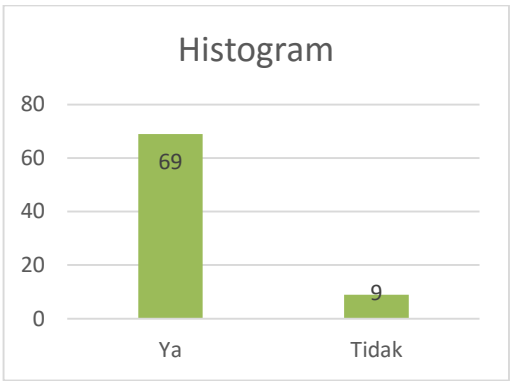
Gambar 6. Histogram Jenjang Pendidikan

4. Histogram Agama



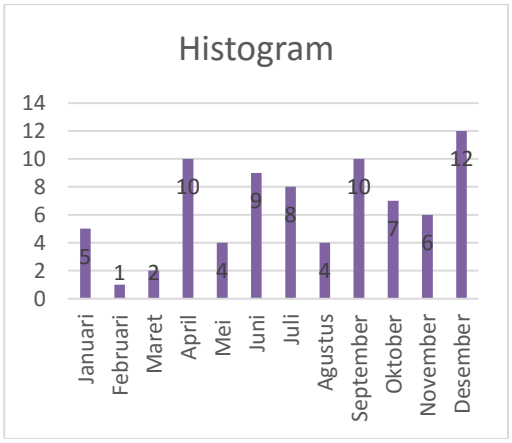
Gambar 7. Histogram Agama

5. Histogram Lahir di Bandung



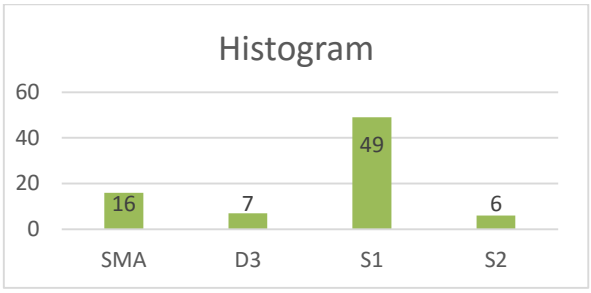
Gambar8. Histogram Lahir Di Bandung

6. Histogram Bulan Lahir



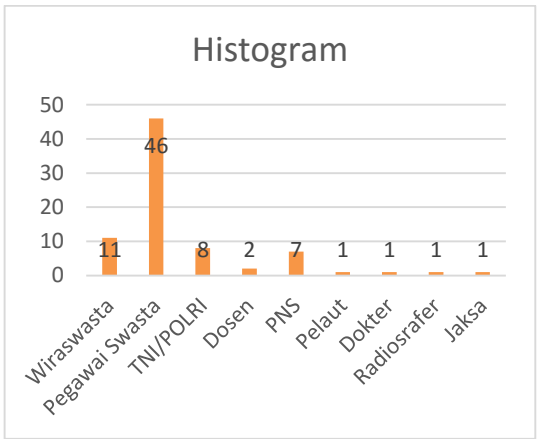
Gambar 9. Histogram Bulan Lahir

7. Histogram Pendidikan Orang Tua



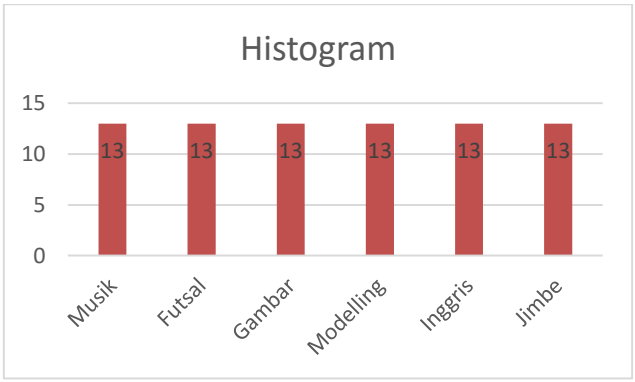
Gambar10. Histogram Pendidikan Orang Tua

8. Histogram Pekerjaan Orang Tua



Gambar11. Histogram Pekerjaan Orang Tua

9. Histogram Ekstrakurikuler



Gambar12. Histogram Ekstrakurikuler

E.2 Perhitungan Information Gain

Setelah diketahui frekuensi masing-masing data yang muncul dalam sebuah atribut yang dijasikan pada bagian E1. Selanjutnya dihitung nilai dari *Information Gain* dan Gain rationya dengan menggunakan persamaan …1) dan 2).

$$Gain(S,A) \equiv Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_v) \ldots\ldots\ldots 1)$$

Dengan nilai dari $Entropy(S)$ adalah

$$Entropy(S) = \sum_{i}^{c} -p_i \log_2 p_i \quad \ldots.2)$$

Adapun perhitungan *Information Gain* dari masing-masing atribut yang ada akan dijelaskan pada sub bab berikut.

1. Perhitungan Nilai *Entropy*

Berdasarkan histrogram pada atribut ektrakurikuler yang menjadi kelas target. Maka nilai dari *Entropy* dari kumpulan data ini adalah:

$$Entropy(S) = \sum_{i}^{c} -p_i \log_2 p_i$$

$$= -(P(musik)\log_2 P(musik) + P(futsal)\log_2 P(futsal) + P(inggris)\log_2 P(inggris) + P(gambar)\log_2 P(gambar) + P(\mathrm{mod}el)\log_2 P(\mathrm{mod}el) + P(jimbe)\log_2 P(jimbe))$$

$$= -(P(musik)\log_2 P(musik) + P(futsal)\log_2 P(futsal) + P(inggris)\log_2 P(inggris) + P(gambar)\log_2 P(gambar) + P(\mathrm{mod}el)\log_2 P(\mathrm{mod}el) + P(jimbe)\log_2 P(jimbe))$$

$$= -(\frac{13}{78}\log_2\frac{13}{78} + \frac{13}{78}\log_2\frac{13}{78} + \frac{13}{78}\log_2\frac{13}{78} + \frac{13}{78}\log_2\frac{13}{78} + \frac{13}{78}\log_2\frac{13}{78} + \frac{13}{78}\log_2\frac{13}{78} + \frac{13}{78}\log_2\frac{13}{78})$$

$$= 2.585$$

2. Perhitungan Nilai *Information Gain* untuk atribut Jenis Kelamin

Adapun perhitungan nilai *Information Gain* dari atribut Jenis Kelamin adalah sebagai berikut.

$$Gain(S,A) \equiv Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_v)$$

$$Gain(S, JenisKela\,\min) = 2.585 - 2.21$$

$$Gain(S, JenisKela\,\min) = 0.375$$

Nilai *Information Gain* untuk atribut jenis kelamin adalah 0.375

3. Perhitungan nilai *Information Gain* untuk atribut Usia

Adapun perhitungan nilai *Information Gain* dari atribut usia adalah sebagai berikut.

$$Gain(S,A) \equiv Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_v)$$

$$Gain(S, Usia) = 2.585 - 2.276$$

$$Gain(S, Usia) = 0.309$$

Nilai *Information Gain* untuk atribut usia adalah 0.309.

4. Perhitungan nilai *Information Gain* untuk atribut Jenjang Pendidikan

Adapun perhitungan nilai *Information Gain* dari atribut jenjang pendidikan adalah sebagai berikut.

$$Gain(S,A) \equiv Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_v)$$

$$Gain(S, JenjangPendidikan) = 2.585 - 2.390$$

$$Gain(S, JenjangPendidikan) = 0.195$$

Nilai *Information Gain* untuk atribut jenjang pendidikan adalah 0.195.

5. Perhitungan nilai *Information Gain* untuk atribut Agama

Adapun perhitungan nilai *Information Gain* dari atribut agama adalah sebagai berikut.

$$Gain(S,A) \equiv Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_v)$$

$$Gain(S, Agama) = 2.585 - 2.422$$

$$Gain(S, Agama) = 0.163$$

Nilai *Information Gain* untuk atribut agama adalah 0.163.

6. Perhitungan nilai *Information Gain* untuk atribut Lahir di Bandung

Adapun perhitungan nilai *Information Gain* dari atribut lahir di Bandung adalah sebagai berikut.

$$Gain(S,A) \equiv Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_v)$$

$$Gain(S, LahirDiBandung) = 2.585 - 2.574$$

$$Gain(S, LahirDiBandung) = 0.011$$

Nilai *Information Gain* untuk atribut lahir di Bandung adalah 0.011.

7. Perhitungan nilai *Information Gain* untuk atribut bulan lahir.

Adapun perhitungan nilai *Information Gain* dari atribut bulan lahir adalah sebagai berikut.

$$Gain(S,A) \equiv Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_v)$$

$$Gain(S, BulanLahir) = 2.585 - 2.022$$

$$Gain(S, BulanLahir) = 0.563$$

Nilai *Information Gain* untuk atribut bulan lahir adalah 0.563.

8. Perhitungan nilai *Information Gain* untuk atribut pendidikan orang tua.

Adapun perhitungan nilai *Information Gain* dari atribut pendidikan orang tua adalah sebagai berikut.

$$Gain(S,A) \equiv Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_v)$$

$$Gain(S, PendidikanOrangTua) = 2.585 - 2.370$$

$$Gain(S, PendidikanOrangTua) = 0.215$$

Nilai *Information Gain* untuk atribut pendidikan orang tua adalah 0.215.

9. Perhitungan nilai *Information Gain* untuk Atribut Pekerjaan Orang Tua

Adapun perhitungan nilai *Information Gain* dari atribut pekerjaan orang tua adalah sebagai berikut

$$Gain(S,A) \equiv Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_v)$$

$$Gain(S, Pe\,\text{ker}\, jaanOrangTua) = 2.585 - 2.287$$

$$Gain(S, Pe\,\text{ker}\, jaanOrangTua) = 0.298$$

Nilai *Information Gain* untuk atribut pekerjaan orang tua adalah 0.298.

10. Analisis *Information Gain*

Adapun hasil dari perhitungan *Information Gain*dari setiap aribut digabungkan ke dalam table IV.

TABEL IV.

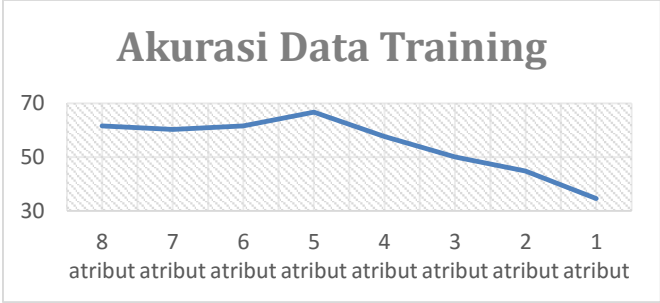
INFORMATION GAIN UNTUK SETIAP ATRIBUT

Atribut	Information gain
Jenis Kelamin	0.375
Usia	0.309
Jenjang Pendidikan	0.195
Agama	0.163
Lahir di Bandung	0.11
Bulan Lahir	0.563
Pendidikan orang tua	0.215
Pekerjaan orang tua	0.298

Dalam grafik pada table IV terlihat bahwa Bulan lahir, menempati nilai *Information Gain* paling besar, disusul dengan usia, pekerjaan orang tua, jenis kelamin, pendidikan orang tua, jenjang pendidikan, agama, dan atribut lahir di Bandung. Atribut lahir di Bandung mendapatkan nilai *Information Gain* terendah yaitu dengan nilai 0.011.

Setelah menghitung nilai *Information Gain* dari setiap atribut, percobaan selanjutnya adalah menghapus satu per satu atribut yang memiliki nilai *Information Gain* paling rendah ke atribut yang memiliki nilai *Information Gain* lebih tinggi. Percobaan ini dengan menggunakan data pada data training itu sendiri. Percobaan dimulai dengan menghapus atribut lahir di bandung, kemudian menghapus atribut agama sebagai pemilik nilai *Information Gain* terendah kedua. Begitu seterusnya hingga tersisa satu atribut dan catat akurasi yang dihasilkan.

Hasil akurasi data training setelah satu per satu atribut dihapus sampai menyisakan satu atribut dapat dilihat pada gambar 13.



Gambar 13. Akurasi Data Training Setelah Dilakukan Penghapusan Atribut Satu per Satu

Dari grafik tersebut dapat dijelaskan pada awal data training diuji coba dengan menggunakan atribut naïve bayes tanpa menghapus satu atribut pun, akurasi yang dihasilkan adalah 61,538%. Setelah salah satu atribut yang memiliki nilai *Information Gain* terendah yakni Atribut Lahir di Bandung dihapus, akurasi turun menjadi 60,256%.

Akurasi tertinggi terjadi pada saat menghapus tiga atribut yakni atribut lahir di Bandung, agama, dan jenjang pendidikan yakni mencapai 66.667%. Setelah menghapus tiga atribut, akurasi yang didapat berangsur-angsur turun.

Akurasi terendah adalah ketika hanya atribut bulan lahir tersisa yaitu menghasilkan akurasi sebesar 34.615%. Dari hal ini dapat disimpulkan sementara kelima atribut yang cukup signifikan dalam meningkatkan akurasi data training yaitu atribut bulan lahir, jenis kelamin, usia, pekerjaan orang tua, dan pendidikan orang tua. Hal ini sesuai dengan perhitungan Nilai *Information Gain*.

E.3 Penerapan Algoritma Naïve Bayes

Dalam tahapan rekomendasi selanjutnya diterapkan algoritma Naïve Bayes.Percobaan pertama yakni menguji kumpulan data tersebut dengan Algoritma Naïve Bayes.Pengujian dilakukan dengan metode 10-*fold cross validation* yang ada pada aplikasi *WEKA*.Hasil pengujian ini berupa data akurasi, *confusion matrix*dan nilai AUC (Area Under Curve) disajikanpada Tabel V.

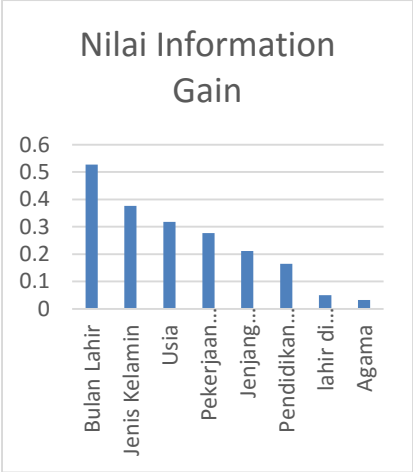
Dalam percobaan ini dilakukan pengujian dalam beberapa prediksi. Satu rekomendasi diartikan jika satu kelas label terdapat pada satu nilai probabilitas tertinggi dari hasil klasifikasi Naïve Bayes. Dua rekomendasi diartikan jika satu kelas label terdapat pada dua nilai probabilitas tertinggi dari hasil klasifikasi Naïve Bayes. Tiga rekomendasi diartikan jika satu kelas label terdapat pada tiga nilai probabilitas tertinggi dari hasil Klasifikasi Naïve Bayes. Tabel V menunjukan akurasi untuk satu prediksi sebesar 19,23%, dua prediksi sebesar 43,59 %, dan tiga prediksi sebesar 79,49%. Nilai AUC (*Area Under Curve*) yang didapat adalah 0.660.

TABEL V.

NILAI AKURASI UNTUK KLASIFIKASI DATA DENGAN ALGORITMA NAÏVE BAYES

Metode	Akurasi			Nilai AUC
	satu prediksi	dua prediksi	tiga prediksi	
Naïve Bayes	19.23%	43.59%	79.49%	0.660

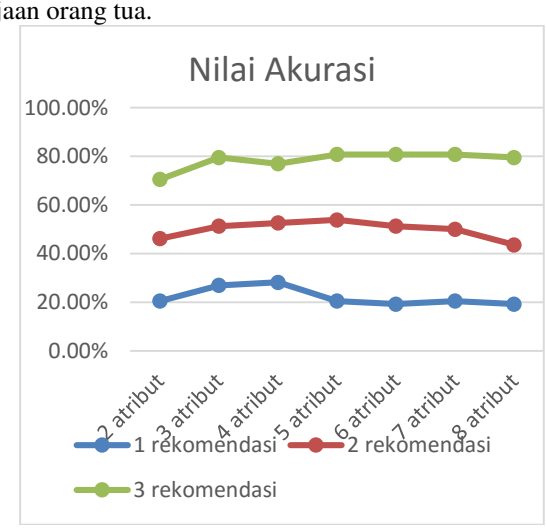
Percobaan kedua adalah dengan melakukan perhitungan nilai Information Gain yang akan digunakan sebagai *feature selection*. Nilai dari Information Gain ini diurutkan dari yang memiliki nilai Information Gain tertinggi ke nilai Information Gain Terendah.Nilai Information Gain dari masing-masing atribut Data Training berturut-turut disajikan pada Gambar 14.



Gambar 14. Nilai Information Gain untuk Setiap Atribut

Percobaan dilakukan dengan melakukan iterasi atribut yang dilakukan yang memiliki nilai Information Gain tertinggi yakni Atribut Bulan Lahir.Iterasi Atribut berhenti setelah sampai pada atribut Lahir di Bandung yang memiliki nilai Information Gain terendah. Hasil percobaan berupa Hasil Akurasi dapat dilihat pada Gambar 15, sedangkan Hasil nilai AUC dapat dilihat pada TABEL V.

Dalam Gambar 15 diperlihatkan nilai akurasi percobaan Algoritma Naïve Bayes dengan Seleksi Fitur Information Gain. Terlihat bahwa dengan menggunakan empat atribut dengan nilai Information Gain tertinggi mendapatkan peningkatan akurasi untuk satu rekomendasi dari 19,23% menjadi 28,21%. Peningkatan akurasi juga terjadi untuk dua rekomendasi yakni dari 43,59% menjadi 52,56%. Namun untuk tiga rekomendasi, nilai akurasi mengalami penurunan dari 79,49% menjadi 76,92%. Meski demikian, dengan memperhatikan peningkatan niaai akurasi untuk satu dan dua rekomendasi yang mencapai 8,97%, maka dipergunakan model dengan empat atribut tersebut untuk melakukan rekomendasi ekstrakurikuler pada Aplikasi. Keempat atribut tersebut adalah atribut bulan lahir, usia, jenis kelamin, dan pekerjaan orang tua.



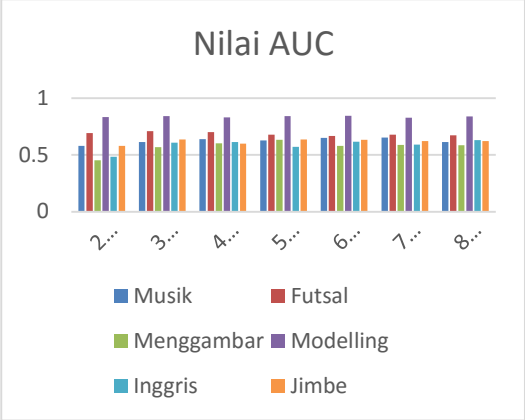
Gambar 15. Nilai Akurasi Percobaan dengan Seleksi Fitur Information Gain untuk Tiga Rekomendasi Tertinggi

Dalam percobaan ini dilakukan pula analisis mengenai nilai AUC untuk masing-masing kelas target dan rata-rata nilai AUC secara keseluruhan. Nilai AUC untuk masing masing kelas target dapat dilihat pada Gambar 16. Sementara untuk Nilai rata-rata AUC dapat dilihat pada TABEL V.

Nilai AUC dipergunakan untuk menganalisis hasil prediksi klasifikasi. Penentuan hasil prediksi klasifikasi dapat dilihat pada batasan nilai AUC sebagai berikut [16]:

1. Nilai AUC 0.90 – 1.00 = *excellent classification*
2. Nilai AUC 0.80 – 0.90 = *good classification*
3. Nilai AUC 0.70 – 0.80 = *fair classification*
4. Nilai AUC 0.60 – 0.70 = *poor classification*
5. Nilai AUC 0.50 – 0.60 = *failure classification*

Dalam Gambar 16 dijelaskan bahwa nilai AUC untuk masing-masing kelas target berbeda-beda. Nilai AUC tertinggi untuk empat atribut terdapat pada kelas target modelling dengan nilai 0.831 yang masuk ke dalam kategori *good classification*, kelas target yang lainnya tersebar di kategori *fair classification* untuk kelas target futsal, sisa kelas target masuk pada *poor classification*.



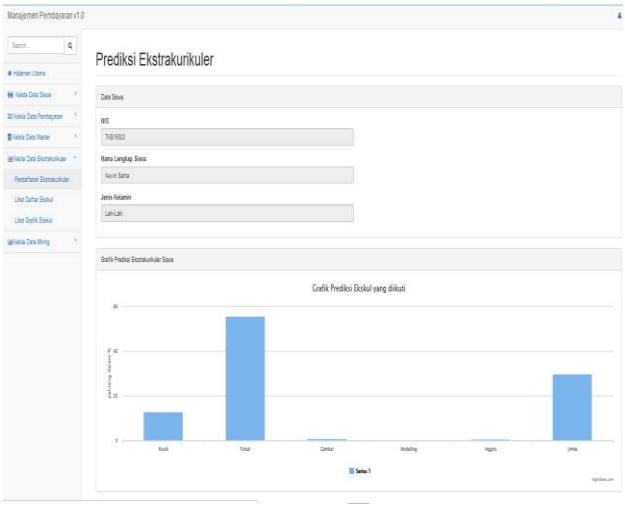
Gambar 16. Nilai AUC untuk Masing-masing Kelas Target

TABEL VI.

NILAI AUC YANG DIHASILKAN

Banyaknya Atribut	Rata-rata Nilai AUC
2 Atribut	0.604
3 Atribut	0.662
4 Atribut	0.664
5 Atribut	0.664
6 Atribut	0.664
7 Atribut	0.660
8 Atribut	0.660

Dari Tabel VI, Nilai rata-rata AUC secara keseluruhan pada empat, lima dan enam atribut adalah 0.664. Nilai tersebut digolongkan pada *poor classification.* Namun nilai AUC tersebut meningkat 0.004 poin dibanding dengan Nilai AUC dengan Algoritma Naïve Bayes tanpa pemilihan seleksi fitur di table V yaitu 0.660.



Gambar 17. Tampilan Antarmuka Prediksi Ekstrakurikuler

Hasil analisis digunakan untuk menentukan atribut yang dipakai untuk melakukan rekomendasi ekstrakurikuler pada aplikasi. Proses yang dilakukan untuk menghitung probabilitas dengan Naïve Bayes dan menerjemahkan menjadi suatu rekomendasi dapat dilihat pada gambar 17. Hasil tampilan rekomendasi mengimplempentasikan penerapan algoritma Naïve Bayes dengan pemilihan 4 atribut yakni bulan lahir, jenis kelamin, usia, dan pekerjaan orang tua.

Keempat atribut tersebut diperoleh dari hasil percobaan yang telah dibahas dan dilakukan sebelumnya pada bagian E.1 dan E.2. Setelah aplikasi ini dibuat, maka dilakukan pengujian terhadap penerapan Algoritma Naïve Bayes dengan aplikasi seperti pada gambar 17 dan juga penerapan Algoritma Naïve Bayes dengan menggunakan WEKA. Metode pengujian pada aplikasi adalah metode *Black box*.

Untuk keperluan pengujian, digunakan data yang sama. Data dipisahkan menjadi 78 data untuk data training dan 18 data untuk data testing. Berikut salah satu contoh pengujian yang dilakukan untuk data training dengan kelas target ekstrakulikuler musik.

Dari gambar 18, hasil klasifikasi pada aplikasi yang dibuat untuk Kelas Target Ekstrakurikuler Musik terlihat bahwa terdapat lima data yang tidak tepat diprediksi (Lihat kolom akurat yang bernilai "Tidak"). Delapan data berhasil diklasifikasikan dengan benar. Pengklasifikasian tersebut menggunakan Algoritma Naïve Bayes yang terdapat pada program. Total seluruh data yang memiliki kelas target ekstrakurikuler berjumlah tiga belas.

Nama Siswa	Jenis Kelamin	Usia	Bulan Lahir	Pendidkan ortu	Pekerjaan ortu	Ekskul	Prediksi	Akurat
S-TR-01	L	4 th	Januari	S1	Wiraswasta	Musik	Musik	Ya
S-TR-02	P	5 th	Agustus	S1	Pegawai	Musik	Gambar	Tidak
S-TR-03	L	5 th	Januari	S1	Pegawai	Musik	Musik	Ya
S-TR-04	P	5 th	Juni	S1	Pegawai	Musik	Gambar	Tidak
S-TR-05	P	5 th	Oktober	S1	Pegawai	Musik	Musik	Ya
S-TR-06	L	5 th	April	S1	Pegawai	Musik	Musik	Ya
S-TR-07	L	5 th	September	S1	TNI/POLRI	Musik	Musik	Ya
S-TR-08	L	5 th	Oktober	S2	Dosen	Musik	Musik	Ya
S-TR-09	P	6 th	Juli	S1	Pegawai	Musik	Modelling	Tidak
S-TR-10	P	5 th	September	S1	PNS	Musik	Musik	Ya
S-TR-11	L	6 th	September	SMA	TNI/POLRI	Musik	Inggris	Tidak
S-TR-12	L	6 th	Juli	S1	Pegawai	Musik	Futsal	Tidak
S-TR-13	L	6 th	Januari	S1	PNS	Musik	Musik	Ya

Gambar 18. Hasil Klasifikasi pada aplikasi yang dibuat untuk kelas target ekstrakurikuler musik

Sedangkan hasil pengklasifikasian data menurut WEKA adalah sebagai berikut.

inst#	actual	predicted	error
1	1:Musik	1:Musik	
2	1:Musik	3:Gambar	+
3	1:Musik	1:Musik	
4	1:Musik	3:Gambar	+
5	1:Musik	1:Musik	
6	1:Musik	1:Musik	
7	1:Musik	1:Musik	
8	1:Musik	1:Musik	
9	1:Musik	4:Modellin	+
10	1:Musik	1:Musik	
11	1:Musik	5:Inggris	+
12	1:Musik	2:Futsal	+
13	1:Musik	1:Musik	

Gambar19. Hasil Klasifikasi pada WEKA untuk Kelas Target Ekstrakurikuler Musik

Dari gambar19 hasilklasifikasi pada WEKA untuk Kelas Target Ekstrakurikuler Musik terlihat bahwa terdapat lima data yang tidak tepat diprediksi (lihat tanda +). Delapan data berhasil diklasifikasikan dengan benar. Total seluruh data yang memiliki kelas target ekstrakurikuler berjumlah tiga belas. Bila dibandingkan antara gambar 18 dan 19, hasil pengklasifikasian atau prediksi yang terdapat pada aplikasi dan WEKA seluruhnya tepat. Sehingga dapat disimpulkan bahwa penerapan Algoritma Naïve Bayes pada aplikasi untuk memberikan rekomendasi ekstrakurikuler telah sesuai.

V. SIMPULAN

Data set yang digunakan dalam penelitian ini adalah data siswa dan data ekstrakulikuler di Sekolah ABC khusus tingkat PG, TKA dan TKB. Data yang digunakan adalah 158 data, yang dibagi menjadi data training sebanyak 78 dan data testing sebanyak 18. Data mining digunakan dalam menganalisis karakteristik siswa yang telah mengikuti ekstrakulikuler dari jenjang pendidikan PG, TKA dan TKB.

Hasil penelitian berfokus kepada sistem rekomendasi dengan penerapan Algoritma Naïve Bayes, yang dapat disimpulkan bahwa hasil penelitian ini sejalan dengan penelitian yang dilakukan oleh Rozzaqi [17] yang melakukan penelitian serupa dengan studi kasus ketepatan kelulusan mahasiswa. Hasil penelitian menunjukan bahwa Algoritma Naïve Bayes dan metode Filtering Feature Selection Information Gain berpengaruh pada akurasi dan nilai AUC untuk prediksi kelulusan mahasiswa. Sedangkan berdasarkan hasil eksperimen dalam penelitian ini, dapat ditarik kesimpulan bahwa algoritma Naïve Bayes dengan pemilihan atribut menghasilkan hasil akurasi rekomendasi yang lebih baik dengan nilai rata-rata 0.664 dibandingkan algoritma Naïve Bayes tanpa Filtering Feature Selection sebesari 0.660. Dan atribut yang dipilih dan digunakan sebagai data untuk sistem rekomendasi ekstrakulikuer di Sekolah ABC adalah atribut bulan lahir, jenis kelamin, usia, dan pekerjaan orang tua.

DAFTAR PUSTAKA

[1] F. Ricii, L. Rokach, B. Shapira and P. B. Kantor, Recommender System Handbook, New York: Springer, 2011.

[2] M. P. Robiliard, W. Maalej, R. J. Walker and T. Zimmerman, Recommendation System pada Software Engineering, Heidelberg: Springer, 2014.

[3] D. Jannach, M. Zanker, A. Felfernig and G. Friedrich, Recommender an Introduction System, New York: Cambridge University Press, 2011.

[4] J. Hian, M. Kamber and J. Pei, Data Minning Concept and Technique, Elsevier, 2012.

[5] W. Zhang and F. Gao, "An Improvement to Naive Bayes for Text Classification," *Procedia Engineering,* vol. 15, no. Advanced in Control Engineering and Information Science, pp. 2160-2164, 2011.

[6] D. T. Larose, Discovery Knowledge in Data: An Introduction to Data Minning, John Willey & Sons, Inc, 2006.

[7] S. Wang, D. Li, X. Song, Y. Wei and H. Lie, "A Feature Selection Method Based on Improved Fischer's Discriminant Ratio for Text Sentiment Classification," *Expert System with Applications,* pp. 8696-8702, 2011.

[8] J. Novakovic, "The Impact of Feature Selection on the Accuracy of Naive Bayes," *18th Telecomunication forum TELFOR 2012,* pp. 1113-1116, 2010.

[9] M. Kantardzic, Data Minning Concepts, Models, Methods, and Algorithm, Wiley Publication, 2011.

[10] C. Vercellis, Business Intelligence: Data Mining and Optimization for Decision Making, Wiley Publisher, 2009.

[11] M. Naseriparsa, A.-M. Bidgoli and T. Varaee, "A Hybrid Feature Selection Method to Improve Performance of a Group of Classification Algorithms," *International Journal of Computer Applications,* vol. 69, pp. 28-35, 2013.

[12] Suyanto, Artificial Intelligence, Bandung: Informatika, 2014.

[13] A. S. Tewari, A. Kumar and A. G. Barman, "Opinion Based Book Recommendation Using Naive Bayes Classification," *International Conference on Contemporary Computing and Informatics (IC3i),* pp. 139-144, 2014.

[14] M. A. Ghazanfar and A. Prugel-Bennet, "An Improved Switching Hybrid Recommender System Using Naive Bayes Classifier and Collaborative Filtering," *The 2010 IAENG International Conference on Data*

Mining and Applications, 2010.

[15] A. R. Khadafy and R. S. Wahono, "Penerapan Naive Bayes untuk Mengurangi Data Noise pada Klasifikasi Multikelas dengan Decision Tree," *Journal of Intelligent System,* vol. 1, pp. 136-142, 2015.

[16] F. Gorunescu, Data Mining: Concepts, Models, and Techniques, Berlin, 2011.

[17] A. R. Rozzaqi, "Naive Bayes dan Filtering Feature Selection Information Gain untuk prediksi Ketepatan Kelulusan Siswa," *Jurnal Informatika UPGRIS,* pp. 30-41, 2015.