APLIKASI PREDIKSI PEMINATAN SMAN 8 BANDUNG MENGGUNAKAN METODE KLASIFIKASI DENGAN ALGORITMA ID3

INTEREST PREDICTION APPLICATION OF NEGERI 8 BANDUNG SENIOR HIGH SCHOOL USING CLASSIFICATION METHOD WITH ID3 ALGORITHM

Fatimah Zahraeni Talib, RA Paramita Mayadewi, S.Kom., M.T.[2], Ely Rosely, Ir., M.B.S.[3]

[123] Program Studi D3 Manajemen Informatika, Fakultas Ilmu Terapan, Universitas Telkom
[1]zahradiyo41@gmail.com, [2]paramita@tass.telkomuniversity.ac.id, [3]ely.rosely@tass.telkomuniversity.ac.id

Abstrak

Proyek akhir ini dimaksudkan untuk melakukan prediksi peminatan siswa SMAN 8 Bandung dengan menggunkan metode klasifikasi dengan data mining menggunakan algoritma ID3. Berdasarkan aturan kurikulum pendidikan 2013 peminatan siswa SMAN dilakukan di semester genap di kelas X sehingga data yang digunakan untuk melakukan prediksi adalah data tahun ajaran 2014-2015 yang berjumlah 442 data. Data yang menjadi atribut adalah data ujian nasioanal, rata-rata raport dan test potensi akademik. Weka dengan versi 3.6.13 merupakan tools yang digunakan untuk membangun *rule. Rule yang terbentuk* kemudian diimplementasikan dalam pembuatan aplikasi prediksi peminatan siswa SMAN 8 Bandung berbasis web.

Kata kunci : Datamining, Klasifikasi, Weka, ID3.

Abstract

This final project is intended to predict students' interest at SMAN 8 Bandung by using classification method with data mining, that's algorithms ID3. Based on Curricullum 2013, the interest of students at senior high school conducted in the second semester in the class X so that the data used to make predictions is data 2014-2015 among 442 data. The data attributes is national examination data, the average report and their academic potential. Weka with version 3.6.13 is a tool that is used to build the rule. Then, the rule that is formed is implemented in making application about predictions of students' interest at SMAN 8 Bandung based on web.

Keywords: Data Mining, Classification, Weka, ID3

1. Pendahuluan

perlu dibuat suatu teknik untuk memprediksi penentuan jurusan yaitu dengan *data mining. Data mining* adalah disiplin ilmu yang mempelajari metode ilmu untuk mengekstrak pegetahuan atau pola dari suatu data, sehingga *data mining* juga sering disebut dengan *knowledge discovery in database*.[1]

Berdasarkan kasus di atas, metode klasifikasi dengan algoritma ID3 digunakan untuk memprediksi jurusan. Klasifikasi merupakan metode yang sangat kuat dan terkenal dalam hal membuat pohon keputusan. Metode pohon keputusan mengubah fakta yang sangat besar menjadi pohon keputusan yang mempresentasikan aturan. Banyak algoritma yang dapat dipakai dalam membuat pohon keputusan, ID3 merupakan salah satu algoritma dapat men-*generate* pohon keputusan yang *simple* dan spesifik, serta eliminasi perhitungan yang tidak diperlukan karena ketika menggunakan pohon keputusan maka sampel diuji hanya berdasarkan kriteria atau kelas tertentu, Pohon keputusan yang dihasilkan ID3 mudah dimengerti, pembentukan pohon keputusan cepat dan pendek serta dapat memproses *data mining* jika data atribut terdapat pada data set berupa data nominal.[2]

2. Metode Pengerjaan

Metode yang digunakan dalam pengerjaan proyek akhir ini adalah metode CRIPS-DM. CRIPS-DM menyediakan standar proses baku untuk data mining yang dapat diterapkan dalam strategi pemecahan masalah umum pada bisnis atau pada unit penelitian. CRIPS-DM membandingkan metodologi *data mining* lain lebih lengkap dan terdokumentasi dengan baik. Setiap fase terstruktur dan terdefenisi dengan jelas sehingga mudah diaplikasikan bahkan bagi pemula sekalipun.[1]
Tahapan *data mining* adalah sebagai berikut.

A. Pemahaman Bisnis
Pemahaman bisnis merupakan tahap awal yaitu pemahaman penelitian, penentuan tujuan dan rumusan masalah *data mining*. Tahap awal ini berfokus pada pemahaman tujuan dan persyaratan proyek dari perspektif bisnis, kemudian mengubah pengetahuan ini ke dalam definisi masalah data mining dan rencana awal dirancang untuk mencapai tujuan [3].

B. Pemahaman Data
Pemahaman data mempertimbangkan data yang dibutuhkan. Langkah ini bisa meliputi pengumpulan data awal, deskripsi data, esplorasi data, dan verifikasi kualitas data. Model-model seperti analisis pengelompokan (*cluster analysis*) dapat pula diterapkan dalam tahap ini, dengan tujuan mengidentifikasi pola dalam data tersebut [4].

C. Pengolahan Data
Setelah sumber data yang tersedia diidentifikasi, sumber data tersebut perlu diseleksi, dibersihkan, dibangun kedalam wujud yang dikehendaki, dan dibentuk. Pembersihan dan transformasi data dalam persiapan pembuatan model data perlu dilakukan pada tahap ini. Eksplorasi data secara lebih mendalam juga dapat diterapkan dalam tahap ini, dan penggunaan model-model tambahan sekali lagi memberikan peluang untuk melihat berbagai pola berdasarkan pemahaman bisnis [4].

D. Pemodelan
Metode penggalian data, seperti visualisasi serta analisis pengelompokan (untuk mengidentifikasi variabel mana yang berhubungan satu sama lain) bermanfaat bagi analisis awal. Alat bantu seperti induksi aturan yang digeneralisasi dapat mengembangkan aturan-aturan asosiasi awal. Begitu pemahaman data yang lebih luas diperoleh maka model-model yang lebih rindah sesuai dengan jenis data tersebut bisa diterapkan [4].

E. Evaluasi

Hasil model kemudian dievaluasi dalam konteks tujuan bisnis yang ditetapkan pada tahap awal. Hal ini untuk mengarahkan pada identifikasi kebutuhan [4].

F. Penyebaran
Pada tahap hasil yang sudah terbentuk berupa *rule* kemudian didokumentasikan dan diimplementasikan kedalam aplikasi yang kemudian akan dibentuk menjadi sistem yang sesuai dengan kebutuhan pengguna.

3. Tinjauan Pustaka

A. Definisi Data Mining, Klasifikasi, ID3, Metode Validasi, Weka

1. Data Mining
Data mining adalah proses yang menggunakan teknik statistik, matematika, kecerdasan buatan dan *machine learning* untuk mengekstraksi dan mengidentifikasi informasi yang bermanfaat dan pengetahuan yang terkait dari berbagai database besar [2].
Selain definisi di atas, berikut merupakan beberapa definisi dari *data mining*.
    1. *Data mining* adalah serangkaian proses untuk menggali nilai tambah dari suatu kumpulan data berupa pengetahuan

yang selama ini tidak diketahui secara manual [2].

2. *Data mining* merupakan analisis dari peninjauan kumpulan data untuk menemukan hubungan yang tidak diduga dan meringkas data dengan cara yang berbeda dengan sebelumnya yang dapat dipahami dan bemanfaat bagi pemilik data [2].

2. Klasifikasi Data Mining

Klasifikasi adalah teknik data mining yang paling terkenal. Contoh pengaplikasian klasifikasi mengandung gambar dan pola yang akurat, diagnosa medis, persetujuan kredit, mendeteksi kesalahan dalam aplikasi industri dan mengklasifikasikan *trend* pasar keuangan [5].
Proses untuk menyatakan salah satu kategori untuk didefinisikan sebelumnya. Proses pembelajaran fungsi target yang memetakan setiap sekumpulan atribut ke salah satu kelas yang didefinisikan sebelumnya [1]. Adapun *preparing* data untuk klasifikasi adalah sebagai berikut.

1. *Data Cleaning*

   Hal ini melakukan preprocessing data dengan menghapus dan menghilangkan *noisy* (kebisingan data) dengan menggunakan teknik *smooting* dan membersihkan *missing values* dengan mengganti data yang kosong dengan menghitung mode atau median pada data [5].
2. *Revance analysis*
   Banyak atribut pada data yang mungkin tidak relevan untuk diklasifikasikan, contohnya adalah data nama, tempat tanggal lahir, dan alamat [5].
3. *Data Transformation*
   Transformasi data merupakan cara untuk mengubah data ke dalam bentuk yang tepat atau cocok untuk proses *data mining* [5].

Adapun cara untuk menghitung akurasi hasil *data mining*, yaitu dengan menggunakan model *confision matrix*.

Tabel 2- 1 Confusion Matrix

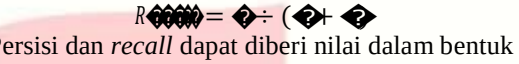
Correct Classification	Classified as	
	+	-
+	True positives (A)	False negatives (B)
-	False positives (C)	True negatives (D)

Akurasi dengan tabel *confision matrix* adalah sebagai berikut.

Persisi didefenisikan sebagai rasio item relevan yang dipilih terhadap semua item yang terpilih.

Persisi dapat diartikan sebagai kecocokan antara permintaan informasi dengan jawaban terhadap permintaan tersebut. Adapun rumus presisi adalah sebagai berikut.

Recall didefenisikan sebagai rasio dari item relevan yang dipilih terhadap total jumlah item relevan yang tersedia. Adapun rumus dari *recall* adalah sebagai berikut.



Persisi dan *recall* dapat diberi nilai dalam bentuk angka dengan menggunakan perhitungan presentase (1-100%) atau dengan menggunakan bilangan antara 0-1. Sistem rekomendasi akan dianggap baik jika nilai presisi dan recallnya tinggi.

Kurva ROC menunjukkan akurasi dan membandingkan klasifikasi secara visual. ROC mengekspresikan *confusion matrix*. ROC adalah grafik dua dimensi dengan *false positive* sebagai garis horizontal dan *truepositive* sebagai garis vertikal. AUC (*the area under curve*) dihitung untuk mengukur perbedaan performansi metode yang digunakan. ROC memiliki tingkat nilai diagnosa sebagai berikut [4].

a. Akurasi bernilai 0,90 – 1,00 = *excellent classification*
b. Akurasi bernilai 0,80 – 0,90 = *good classification*
c. Akurasi bernilai 0,70 – 0,80 = *fair classification*
d. Akurasi bernilai 0.60 – 0,70 = *poor classification*
e. Akurasi bernilai 0.50 – 0.60 = *failure*

3. ID3

ID3 adalah salah satu model *decision tree*. Dalam ID3 kita gunakan kriteria informasi *gain* untuk memilih atribut yang akan digunakan untuk pemisahan objek. Atribut yang mempunyai information *gain* paling tinggi dibanding atribut yang lain relatif terhadap set y dalam suatu data, dipilih untuk melakukan pemecahan [6].

Secara umum algoritma ID3 membangun pohon keputusan adalah sebagai berikut.

a. Pilih atribut akar
b. Buat cabang untuk masing-masing nilai
c. Bagi kasus dalam cabang

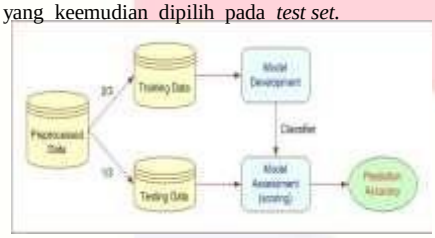
Ulangi proses untuk setiap cabang sampai semua kasus pada cabang memiliki nilai.

4. Metode Evaluasi

Ukuran kinerja dari model pada *test set* sering kali berguna karena ukuran tersebut memberikan estimasi yang tidak biasa dari *error* generalisasinya. Akurasi dari tingkat *error* yang dihitung dari *test set* dapat juga digunakan untuk membandingkan kinerja relatif dari *classifier-classifier* pada domain yang sama. Dalam melakukan perbandingkan ini, label kelas dari test *record* haruslah diketahui. Berikut adalah metode yang umum digunakan untuk mengevaluasi kinerja *classifier* [7].

1. Simple Split

Simple split atau pengukuran pada sampel test adalah metode yang membagi dua subset data yang saling eklusif satu sama lain yang disebut dengan training set yaitu data yang digunakan untuk pelatihan dan test set untuk menguji kecocokan data. Biasanya yang umum digunakan adalah memilih data yang akan digunakan. Dua per tiga. Data digunakan untuk *training* dan sepertiga data digunakan untuk *testing. Training set* digunakan oleh model yang membentuknya dan classifier yang keemudian dipilih pada *test set.*



Gambar 2- 1 Tampilan Methodology Simple Split

Sumber: https://beritati.blogspot.co.id, 2013

2. *Random Subsampling*

Metode *holdout* dapat diulangi beberapa kali untuk meningkatkan estimasi dari kinerja *classifier*. Pendekatan ini dikenal sebagai *random subsampling*. *Random sampling* masih menjumpai beberapa masalah yang terkait dengan metode *holdout* karena tidak menggunakan sebanyak mungkin data untuk *training*. Metode ini juga tidak memiliki kontrol terhadap berapa kali setiap *record* digunakan untuk *training* dan testing. Dengan demikian, beberapa *record* dapat digunakan untuk *training* lebih sering dibanding dengan *record-record* yang lain.

3. *K-Fold Cross-Validation*

Dalam *k-fold cross-validation*, yang disebut juga dengan rotation estimation, dataset yang utuh di pecah secara random menjadi 'k' subset dengan size yang hampir sama dan saling eksklusif satu sama lain. Model dalam *classification* di-latih dan di-tes sebanyak 'k' kali. Setiap kali pelatihan semua dilatih pada semua fold kecuali hanya satu fold saja yang disisakan untuk pengujian. Penilaian cross-validation terhadap akurasi model secara keseluruhan dihitung dengan mengambil rerata dari semua hasil akurasi individu 'k', seperti yang ditunjukkan dengan persaman berikut:

$$CVA = \frac{1}{k}\sum_{i=1}^{k} A_i$$

Dimana CVA adalah akurasi *cross-validation*, k adalah jumlah fold yang digunakan, dan A adalah ukuran akurasi (misalnya, *hit-rate*, *sensitivity*, *specifity*) dari masing-masing fold.

4. Weka

Weka adalah perangkat lunak data mining yang memiliki sekumpulan algoritma standar data mining. Weka dapat dijalankan berbasis GUI dan secara langsung melalui Command Line. Weka dapat digunakan untuk melakukan preprocessing, klasifikasi, clustering (pengelompokan), regresi, association *rule mining* (ARM) dan visualisasi.
Dalam weka setiap dataset merupakan instance dari *class* : weka.core.instance, setiap instance memiliki atribut (field). Domain atribut dapat berupa nominal, numerik, string, date, relasional [8].

5. Interval kelas

Data pertama yang diperoleh pada suatu observasi disebut dengan data mentah. Data ini belum tersusun secara nominal. [12] Dalam bekerja dengan jumlah data yang cukup besar, biasanya lebih menguntungkan jika data disajikan dalam kelas-kelas atau kategori. Beberapa istilah yang digunakan yatu:

1.Batas kelas

Batas kelas adalah bilangan terkecil dan terbesar sesungguhnya yang masuk dalam kelas interval tertentu. Dari batas kelas inilah terbentuk range.

$$R = \text{[illegible]}$$

2. Interval Kelas

Interval kelas adalah interval yang diberikan untuk mementukan kelas-kelas.
I : interval
R : range
K : banyaknya kelas

$$I = R/K$$

B. Proses Data mining

Metode yang digunakan untuk melakukan pengolahan *data mining* yaitu CRIPS-DM atau sikus hidup pengembangan *data mining* sebagai *framework* dari proyek *data mining* [3].

1. Fase pemahaman bisnis

SMAN 8 Bandung merupakan salah satu sekolah percontohan yang menggunakan kurikulum 2013 sebagai acuan pembelajaran. Sehingga, proses peminatan dilakukan di semester genap di kelas x SMA. Oleh karena itu SMAN 8 Bandung membutuhkan aplikasi untuk memprediksi peminatan siswa. Dengan Metode klasifikasi *data mining* menggunakan algoritma ID3 aplikasi tersebut bisa dibangun dengan menggunakan *tools* weka untuk membangun *rule.*

2. Fase pemahaman Data

Data yang digunakan oleh SMAN 8 Bandung untuk melakukan penentuan peminatan adalah data siswa angkatan 2014-2015 berjumlah 475 data yang meliputi no pendaftaran, nama siswa, gugus, nilai ujian nasional, nilai raport SMP semester 1-6, rata-rata nilai raport dan nilai test potensi akademik, dan minat jurusan siswa.

3. Fase Pengolahan data

Berdasarkan data yang sudah ada tersebut kemudian akan diolah kedalam metode *data mining*. Namun sebelumnya data tersebut akan diproses untuk mendapatkan data yang sesuai dengan keinginan. Sehingga dilakukanlah proses berikut:

1) Seleksi data

Memilih data yang akan digunakan dalam proses *data mining.* Data yang digunakan adalah data yang menjadi acuan pokok untuk melakukan peminatan yaitu nilai ujian nasional matematika dan IPA, rata-rata raport SMP semester 1-5, dan nilai test potensi akademik matematik, IPA, dan IPS serta hasil peminatan. Data tersebutlah yang kemudian menjadi atribut yang digunakan pada proses klasifikasi *data mining*. Dalam proyek akhir ini, data yang digunakan terbagi dua, yaitu data dengan menggunakan nilai TPA dan data tanpa nilai TPA. Hal ini dilakukan untuk mengetahui tingkat akurasi data tertinggi, sehingga *rule* dari data tersebut yang digunakan untuk membangun aplikasi.

2) *Preprocessing Data*

Proses pembersihan data dilakukan dengan mencari dan memperbaiki kesalahan atau permasalahan dalam data seperti *missing values* dan *noisy*. Ada beberapa cara untuk mengatasi missing values diantaranya mengisi dengan modus atau rata-rata. Modus digunakan pada data yang berskala kontinu. Median/modus digunakan pada data yang berskala kategorik, misal jenis kelamin. Hal tersebut dilakukan untuk memastikan kualitas data. Pada peminatan siswa SMAN 8 Bandung terdapat beberapa data yang missing values sehingga dilakukan penanganan dengan mencari nilai rata- rata. Data yang digunakan adalah sebanyak 475 data dan terdapat 24 data yang tidak bisa digunakan. Dikategorikan sebagai data yang tidak bisa digunakan jika jumlah missing valuesnya melebihi 2 semester. Terdapat 7 baris data *missing values* pada beberapa nilai semester dan
26 data *noisy* pada nilai ujian nasional. Sehingga data yang digunakan adalah sebanyak 442.

Gambar 2- 2Tampilan Hasil Penambahan Data pada Missing Values

3) *Transformation*

Mengelompokkan atribut-atribut atau *field* yang telah terpilih berdasarkan *range* nilai siswa. Penentuan tersebut menggunakan rumus statistika dasar dengan tahapan sebagai berikut.

1. Menentukan jumlah kelas.

Jumlah kelas yang ingin dibentuk adalah sebanyak tiga kelas yaitu: rendah, sedang, dan tinggi.

2. Menentukan range

1. *R*[illegible]= [illegible] – [illegible] *Range* data dengan nilai TPA

$$R = 100 - 13 = 87$$

2. *R□□□= □□□□ □□□□□□□□– □□□□□□□□□□Range* data *non* TPA

$$R\square\square\square = 100 - 45 = 55$$

3. Menentukan Interval

Setelah menentukan jumlah kelas dan range, langkah selanjutnya adalah menentukan interval (rentang skala) untuk masing-masing kelas. Interval kelas adalah selisih antara batas atas nyata dengan batas bawah nyata. Rumus interval adalah sebagai berikut.

a. Interval data dengan nilai TPA

$$I = \square\square/K = \frac{87}{3} = 29$$

b. Interval data dengan nilai *non* TPA

$$I = \square\square/K$$

$$I = \frac{55}{3} = 18$$

Tabel 2- 2 Atribut Data yang terpilih Data dengan Nilai TPA

Atribut	Atribut Linguistik	Keterangan
MAT_UN	rendah	13-41
	sedang	42-70
	tinggi	71-100
IPA_UN	rendah	13-41
	sedang	42-70
	tinggi	71-100

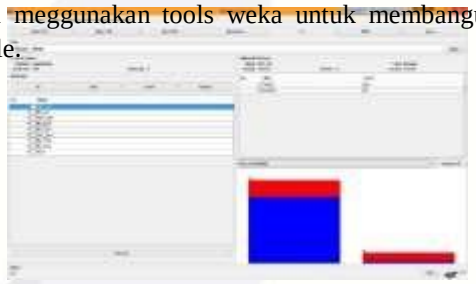
Atribut	Atribut Linguistik	Keterangan
MAT_RAT	rendah	13-41
	sedang	42-70
	tinggi	71-100
IPA_RAT	rendah	13-41
	sedang	42-70
	tinggi	71-100
IPS_RAT	rendah	13-41
	sedang	42-70
	tinggi	71-100
MAT_TPA	rendah	13-41
	sedang	42-70
	tinggi	71-100
IPA_TPA	rendah	13-41
	sedang	42-70
	tinggi	71-100
IPS_TPA	rendah	13-41
	sedang	42-70
	tinggi	71-100

MAT_UN	IPA_UN	MAT_RAT	IPA_RAT	IPS_RAT	MAT_TPA	IPA_TPA	IPS_TPA	HASIL
TINGGI	TINGGI	TINGGI	TINGGI	TINGGI	SEDANG	RENDAH	RENDAH	A
SEDANG	TINGGI	TINGGI	TINGGI	TINGGI	SEDANG	SEDANG	SEDANG	A
TINGGI	TINGGI	TINGGI	TINGGI	TINGGI	RENDAH	SEDANG	RENDAH	A
TINGGI	TINGGI	TINGGI	TINGGI	TINGGI	TINGGI	TINGGI	RENDAH	A
SEDANG	SEDANG	TINGGI	TINGGI	TINGGI	SEDANG	RENDAH	RENDAH	A
TINGGI	TINGGI	TINGGI	TINGGI	TINGGI	TINGGI	TINGGI	RENDAH	A
TINGGI	TINGGI	SEDANG	SEDANG	TINGGI	RENDAH	RENDAH	RENDAH	A
SEDANG	TINGGI	TINGGI	TINGGI	TINGGI	SEDANG	RENDAH	RENDAH	A
RENDAH	RENDAH	SEDANG	TINGGI	SEDANG	RENDAH	RENDAH	RENDAH	A
SEDANG	TINGGI	TINGGI	TINGGI	TINGGI	SEDANG	SEDANG	RENDAH	A
TINGGI	TINGGI	TINGGI	TINGGI	TINGGI	TINGGI	SEDANG	RENDAH	A
TINGGI	TINGGI	TINGGI	TINGGI	TINGGI	TINGGI	SEDANG	SEDANG	A
TINGGI	SEDANG	TINGGI	TINGGI	TINGGI	SEDANG	RENDAH	RENDAH	A
TINGGI	TINGGI	TINGGI	TINGGI	TINGGI	TINGGI	SEDANG	RENDAH	A
TINGGI	TINGGI	TINGGI	TINGGI	TINGGI	SEDANG	SEDANG	SEDANG	A

Gambar 2- 3 Data Setelah Transformasi

4. Fase Pemodelan

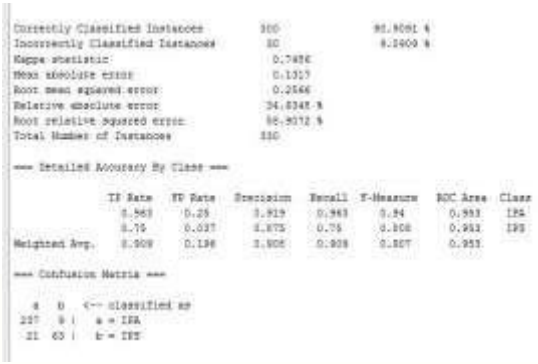
Metode yang digunakan pada penelitiian ini adalah klasifikasi dengan algoritma ID3. Proses ini meggunakan tools weka untuk membangun rule.



Gambar 2- 4 Tampilan Menu Preprocess Weka 3.6.13

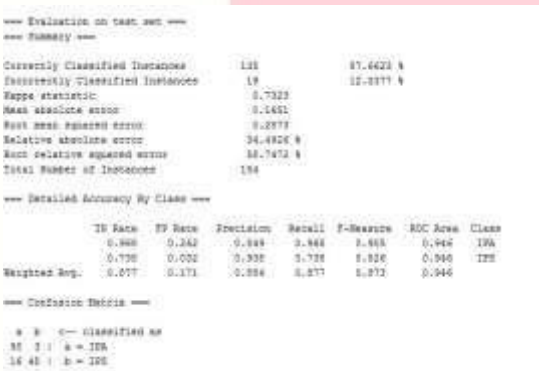
Data yang sudah dipilih kemudian diuji dengan metode validasi dengan algoritma id3. Pengujian ini digunakan untuk mengetahui rule yang terbaik. Hasil validasi model pada weka 3.6.13. Pada awalnya sebelum dilakukan pengujian pada weka data dipecah menjadi dua bagian yaitu data *training* sebanyak 75% dan data *testing* sebanyak 35 % hal ini dilakukan supaya dapat terbentuk suatu model dengan menggunakan data *training*, selanjutnya data yang terbentuk dengan menggunakan data *training* akan diujikan kembali menggunakan data *testing*. Khusus untuk data yang akan di uji dengan model Supplied test set.

Hasil klasifikasi data *training* data dengan nilai TPA sebanyak 75% dengan menggunakan WEKA 3.6.13 dapat dilihat pada gambar dibawah ini:



Gambar 2- 5 Hasil Klasifikasi Data Training dengan Weka 3.6.13

Hasil klasifikasi data *testing* data dengan nilai TPA sebanyak 35% dengan menggunakan WEKA 3.6.13 dapat dilihat pada gambar dibawah ini:



Gambar 2- 6 Hasil Klasifikasi Data Testing dengan Weka 3.6.13

Tabel 2- 8 Hasil Perbandingan Correctly classified pada Weka 3.6.13

Atribut	Use Training set	cross-validation			percentage split 66%	Supplied test set	
		10	15	25		Training	Testing
Nilai	82.5%	79.6%	80.3%	79.8%	79.3%	83.1%	75.4%

Dari tabel di atas diketahui bahwa *Correctly classified* tertinggi adalah data dengan nilai TPA dengan metode validasi *supplied test set*. Berikut merupakan tampilan *visualization error.*

Gambar 3- 1 Tampilan Visualization Error Data Testing dengan Nilai TPA

5. Fase Evaluasi

Setelah melakukan pemodelan hasil weka akan dievaluasi. Evaluasi dilakukan secara manual berdasarkan confution matrix pada weka 3.6.13.

Tabel 2- 9 Tampilan Hasil Perhitungan Model Test Pada Data Non TPA

Mode Test		Confution Matrix		Akurasi	Recall	Presisi
Use Training Set		310	13	0.824	0.852	0.827
		65	54			
Supplied Test Set	Training	237	10	0.834	0.859	0.840
		45	39			
	Testing	88	6	0.755	0.752	0.733
		32	29			
Cross Validation	10	304	29	0.780	0.864	0.813
		70	48			
	15	300	23	0.791	0.860	0.813
		69	49			
	25	303	20	0.800	0.858	0.817
		68	50			
Precentage Split 66%		99	6	0.793	0.832	0.798
		25	20			

Tabel 2- 10 Tampilan Hasil Perhitungan Mode Test Pada Data dengan TPA

Mode Test		Confution Matrix		Akurasi	Recall	Presisi
Use Training Set		306	17	0.88235	0.785	
		35	84			
Supplied Test Set	Training	237	10	0.90634	0.79	0.9186
		21	63			
	Testing	90	4	0.87097	0.667	0.8491
		16	45			
Cross Validation	10	300	22	0.85227	0.8	0.8746
		43	75			
	15	299	22	0.85649	0.795	0.8794
		41	77			
	25	299	22	0.84966	0.802	0.8717
		44	74			
		96	8			

Dari kedua tabel di atas, terlihat bahwa hasil perhitungan model yang tertinggi adalah pada Supplied Test Set pada data dengan nilai TPA. Oleh karena itu data testing dan training pada data dengan nilai TPA akan dibandingkan hasilnya.

b. Data *Training*

$$\text{Akurasi} = \frac{\text{TP}+\text{TN}}{\text{TP}+\text{TN}+\text{FP}+\text{FN}} = = 0.906$$

$$\text{Recall} = \frac{\text{TP}+\text{TN}}{\text{TP}+\text{TN}+\text{FP}+\text{FN}} = \frac{31}{331} = 0.9$$

Persentase Akurasi = 0.906 x 100% = 90.6%

a. Data *testing*

$$\text{Akurasi} = \frac{\text{TP}+\text{TN}}{\text{TP}+\text{TN}+\text{FP}+\text{FN}} = 0.8709$$

$$\text{Recall} = \frac{\text{TP}+\text{TN}}{\text{TP}+\text{TN}+\text{FP}+\text{FN}} = \frac{20}{115} = 0.17$$

Persentase Akurasi = 0.877 x 100% = 87.7%

Hasil perbadingan akurasi dan error rate data testing dan training dapat dilihat pada tabel berikut.

Tabel 2- 11 Perbandingan Akurasi dan Error Rate

Dataset	Akurasi (%)	Kategori	Error rate (%)
Training	90.6%	Excellent classification	9%
Testing	87.7%	Good classification	17%

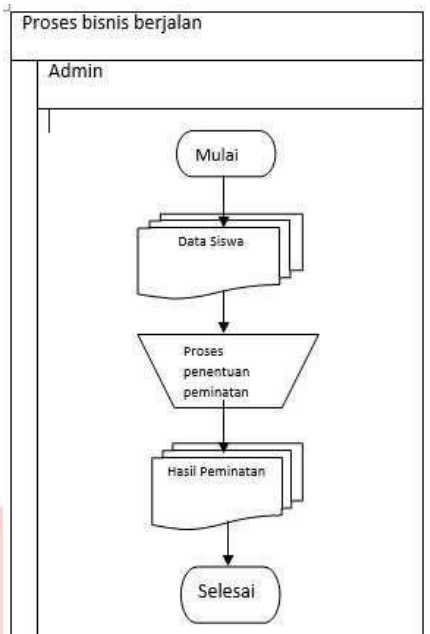
Dari tabel diatas dapat dilihat nilai akurasi serta *error rate data training* serta data *testing*, dimana data *training* memiliki nilai akurasi yang lebih tinggi daripada data *testing.* Dengan menggunakan algoritma ID3 nilai akurasi dari *data training* lebih dari 90% (*excellent classification)*. Hal ini menunjukkan bahwa algoritma ID3 dapat digunakan pada dataset prediksi peminatan SMAN 8 Bandung. Jika hasil akurasi berada pada tingkat *poor classification atau failure classification* maka kembali ke tahap preprocessing untuk menganalisa ulang atribut yang harus dihilangin untuk mencari akurasi tertingi pada masing-masing model validasi untuk mendapatkan rule yang bisa diimplementasikan.

6. Fase Penyebaran

Pada tahap ini pengetahuan atau informasi yang telah diperoleh akan diatur dan dipresentasikan dalam bentuk khusus sehingga dapat digunakan oleh pengguna. Informasi yang sudah ada akan diimplementasikan untuk membuat aplikasi prediksi peminatan siswa SMAN 8 Bandung menggunakan bahasa pemrograman PHP.

4. Pembahasan

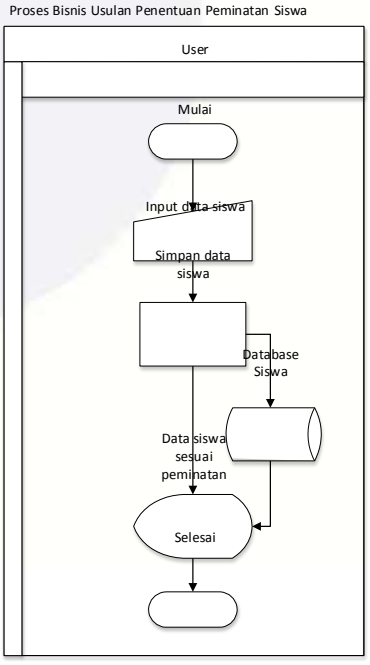
A. Sistem Saat Ini



Penentuan Peminatan SMAN 8 Bandung dilakukan pada semester genap di kelas X. Peminatan ini dilakukan setelah tahap pengumuman penerimaan siswa baru. peminatan dilakukan berdasarkan nilai rata-rata raport SMP pelajaran matematika, IPA dan IPS semester 1-5, nilai matematika dan IPA ujian nasional, dan nilai TPA.

1. Proses Bisnis Usulan pengguna

Berikut merupakan proses bisnis usulan dari peminatan siswa.



B. Antarmuka Tampilan

Gambar 4- 1 Tampilan Halaman Awal Aplikasi

Gambar 4- 2 Tampilan Antarmuka Halaman Dasboard

Gambar 4- 3 Tampilan Antarmuka Input Data Peminatan

Gambar 4- 4 Tampilan Antarmuka Hasil Prediksi Peminatan

1. Kesimpulan

Kesimpulan dari proyek akhir ini adalah aplikasi prediksi peminatan siswa dengan menggunakan metode klasifikasi dengan algoritma ID3 berhasil terbentuk sesuai dengan tujuan. Yaitu mengimplementasikan *rule* yang dibentuk oleh aplikasi weka kedalam aplikasi web dan *rule* yang terbentuk memiliki keakuratan 90% sehingga bisa diimplementasikan.

Daftar Pustaka:

[1] Kusni, Algoritma Data Mining, Yogyakarta: Andi, 2009.

[2] A. P. Fadillah, "Penerapan Metode CRISP-DM untuk Prediksi," *Jurnal Teknik Informatika dan Sistem Informasi* , vol. 1, p. 2, 2015.

[3] P. Chapman, Step by step data mining guide.

[4] B. Santosa, Data Mining Teknik Pemanfaatan Data Untuk Keperluan Bisnis, Yogyakarta: Graha Ilmu, 2007.

[5] M. H. Dunham, Data Mining Introductory and Advance Topics, Upper Saddle River: Pearson Education Inc, 2002.

[6] M. B. Romney and P. J. Steinbart, Accounting Information Systems, USA: Pearson Education, 2012.

[7] A. S. Rosa and M. Shalahuddin, Rekayasa Perangkat Lunak Terstruktur dan Berorientasi Objek, Bandung: Informatika Bandung, 2015.

[8] S. Y. David Olson, Pengantar Ilmu Penggalian data Bisnis (Introduction to Business Data Mining), Jakarta: Salemba Empat, 2008.

[9] T. F. Abidin, "Pengantar WEKA," pp. 3 - 4, 14 januari 2012.

[10] Anonymous, "Metode-metode dalam Data Mining - Seri Data Mining for Business Intelligence (6)," p. 3, 1 Agustus 2013.

[11] mendiknas, "Model Peminatan, Lintas Minat dan Pendalaman Minat Kurikulum 2013," p. 2, 2013.

[12] J. Hernadi, "Statistika Dasar," p. 2, 2009.