

Memprediksi Daftar Ulang Mahasiswa Baru Menggunakan Algoritma Bayesian Classification Di Universitas XYZ

Wahju Tjahjo Saputro¹, Hamid Muhammad Jumasa¹

¹Teknologi Informasi, Universitas Muhammadiyah Purworejo, Purworejo 54111, Indonesia

Abstrak

Penelitian ini menganalisa data PMB tahun 2016. Proses PMB di universitas XYZ memiliki beberapa jalur yaitu Reguler, Transfer dan Bidikmisi. Ketika proses PMB kesulitan yang dijumpai salah satunya adalah jumlah calon mahasiswa yang tidak daftar ulang semakin meningkat.

Berdasarkan identifikasi permasalahan dan informasi dari petugas hal yang paling penting yaitu jumlah calon mahasiswa yang tidak daftar ulang semakin bertambah setiap tahun, sehingga mempengaruhi dampak jumlah penerimaan PMB. Penelitian ini menyelesaikan permasalahan tersebut menggunakan Algoritma Bayesian Classification agar peluang daftar ulang mahasiswa dapat diketahui lebih awal.

Berdasarkan penelitian yang dilakukan diperoleh kesimpulan bahwa class TDU 55% lebih besar dari pada class DU 45%. Artinya calon mahasiswa yang tidak daftar ulang secara keseluruhan semua program studi lebih besar. Dijelaskan bahwa program studi PBSJ 100% tidak ada calon mahasiswa yang daftar ulang. Program studi Pendidikan Ekonomi masih ada peluang 75% calon mahasiswa yang DU. Program studi Pendidikan Fisika 100% tidak ada calon mahasiswa yang mendaftar. Program studi Teknik Sipil ada peluang 25% calon mahasiswa akan DU. Program Studi Agribisnis ada peluang 75% calon mahasiswa melakukan DU. Program studi Peternakan 90% tidak ada peluang calon mahasiswa yang mendaftar. Program studi Psikologi memiliki peluang sama dengan program studi Peternakan yaitu 90% tidak ada peluang calon mahasiswa mendaftar. Untuk program studi Hukum memiliki peluang 90% calon mahasiswa mendaftar.

Kata kunci: Bayesian classification, Mahasiswa baru, Daftar ulang

Abstract

This study analyzes the PMB data in 2016. The PMB process at XYZ university has several paths, namely Regular, Transfer and Bidikmisi. When the PMB process difficulties are encountered one of them is the number of prospective students who do not register again is increasing.

Based on the identification of problems and information from officers the most important thing is the number of prospective students who do not re-register is increasing every year, thus affecting the impact of the number of PMB receipts. This study resolves these problems using the Bayesian Classification Algorithm so that the opportunity to re-register students can be known earlier.

Based on the research conducted, it was concluded that the TDU class was 55% greater than the 45% DU class. This means that prospective students who do not completely re-register all study programs are larger. It was explained that PBSJ study programs 100% no prospective students who re-register. The Economic Education study program still has the opportunity of 75% of prospective students who are DU. 100% Physics Education study program no prospective students register. The Civil Engineering study program has a 25% chance that prospective students will be DU. The Agribusiness Study Program has the opportunity of 75% of prospective students doing DU. The 90% Animal Husbandry Study Program has no chance for prospective students to register. The Psychology study program has the same opportunities as the Animal Husbandry study program which is 90%, there is no chance that prospective students register. For the Law study program, it is possible for 90% of prospective students to register.

Keywords: Bayesian classification, New students, Re-registration

1. PENDAHULUAN

Universitas XYZ merupakan salah satu PTS di kota Purworejo. Saat ini universitas XYZ memiliki jumlah mahasiswa aktif 4.190 (<http://forlap/ristekdikti.go.id>, 2017) yang tersebar pada 5 (lima) fakultas dan 15 (lima belas) program studi sebagai berikut:

Tabel 1. Daftar Fakultas dan Program Studi

Fakultas	Program Studi	Jml
FKIP	P. Bhs. dan Sastra Indonesia	447
	P. Bhs. dan Sastra Jawa	64
	P. Ekonomi	280
	P. Bahasa Inggris	595
	P. Matematika	893
	P. Fisika	108
	P. Teknik otomotif	310
	PGSD	157
FE	Manajemen	767
FT	Teknik Sipil	230
	Teknologi Informasi	37
Pertanian	Agribisnis	110
	Peternakan	46
FIS	Psikologi	32
	Hukum	16

Sistem penerimaan mahasiswa baru di universitas XYZ saat ini telah menggunakan SIMPMB (Sistem Informasi Manajemen Penerimaan Mahasiswa Baru). Dimana salah satu fitur SIMPMB telah tersedia sistem ujian yang berbasis komputer (CBT). Proses penerimaan mahasiswa baru (PMB) di universitas XYZ pertama kali menggunakan sistem tahun 2013. Sebelum tahun 2013 proses PMB menggunakan aplikasi Microsoft Excel yang dikelola UPT PMB. Penelitian ini akan menganalisa data PMB tahun 2016. Proses PMB di universitas XYZ memiliki beberapa jalur yaitu Reguler, Transfer dan Bidikmisi. Ketika proses PMB beberapa kendala yang dialami yaitu setelah mengikuti seleksi mahasiswa melaporkan keinginan pindah program studi, ketidaktahuan calon mahasiswa dalam memilih program studi, latarbelakang pendidikan yang ditempuh di SMK berbeda jalur dengan program studi yang dipilih, paksaan dari pihak keluarga.

Berdasarkan identifikasi permasalahan dan informasi dari pengalaman petugas di lapangan dampak dari temuan masalah tersebut yaitu masa studi mahasiswa menjadi lama, terjadi pindah program studi setelah menempuh beberapa

semester di program studi asal, putus sekolah, nilai indek prestasi yang rendah, tidak mampu menunjukkan prestasi di bidang karya ilmiah dan temuan dampak lainnya. Dengan penelitian ini diharapkan temuan kasus-kasus tersebut dapat diminimalisir oleh pimpinan perguruan tinggi dengan melakukan tindakan yang tepat guna mempertahankan calon mahasiswa baru maupun mahasiswa yang telah menempuh studi (Kusrini dan Luthfi, 2009).

Pada temuan kasus-kasus PMB dapat diatasi dengan menggali bongkahan data berkapasitas besar yang telah tersimpan di dalam perangkat teknologi dengan data mining (Susanto dan Suryadi, 2010). Dalam penelitian ini peneliti ingin menggali data PMB yang tersedia di universitas XYZ guna mengetahui informasi yang tersembunyi, sehingga menghasilkan pengetahuan (Larose, 2005). Hasil dari penelitian ini berupa pengetahuan informasi yang memiliki makna sehingga dapat digunakan sebagai bahan dalam pengambil kebijakan bagi pimpinan.

Data mining memiliki berbagai teknik dan algoritma yang mampu digunakan untuk memecahkan permasalahan. Menurut Susanto dan Suryadi, 2010 disebutkan data mining memiliki teknik Deskripsi, Estimasi, Prediksi, Klasifikasi, Klustering dan Asosiasi. Disebutkan oleh Wu dan Kumar, 2009 beberapa algoritma yang ada pada data mining diantaranya Algoritma C.45, *K-Means*, *Super Vector Machine (SVM)*, Apriori, *k-Nearest Neighbors* dan *Naiv Bayes*. Pada penelitian ini peneliti akan menggunakan Teknik Bayesian Classification untuk menemukan informasi berbasis pengetahuan dari data PMB (Sugianti, 2012).

Bayesian Classification adalah pengklasifikasian atau pengelompokan statistik yang dapat digunakan untuk memprediksi keanggotaan suatu kelas tertentu. *Bayesian Classification* didasarkan pada *Teorema Bayes* yang memiliki kemampuan klasifikasi seperti *decision tree* dan *neural network* (Kusrini dan Luthfi, 2009). Pada penelitian ini data yang digunakan yaitu data calon mahasiswa baru yang tersimpan di server. Kemudian algoritma yang digunakan untuk proses klasifikasi menggunakan *Bayesian Classification* merupakan teknik prediksi berbasis probabilistik sederhana yang berdasarkan pada penerapan teorema Bayes atau aturan Bayes dengan asumsi independensi yang kuat pada fitur lain dalam data yang sama.

Diidentifikasi permasalahan terkait data PMB sebagai berikut:

1. Memanfaatkan kumpulan atau bongkahan data yang besar (*big data*) menjadi informasi berbasis pengetahuan (*discovery knowledge in data*).
2. Informasi pemetaan klasifikasi data calon mahasiswa baru belum pernah dilakukan berdasarkan elemen-elemen tertentu.
3. Pengunduran diri calon mahasiswa baru sulit diketahui lebih dini.

Penelitian ini dibatasi pada data yang diperoleh dari server dengan obyek data calon mahasiswa baru TA 2016/2017. Dengan data sample satu periode tahun ajaran akademik dapat mewakili periode-periode sebelumnya. Hal ini didasarkan pada wilayah sebaran yang setiap tahun sama dan jumlah calon mahasiswa baru yang relatif naik turun tidak jauh berbeda.

Algoritma yang digunakan untuk menyelesaikan permasalahan tersebut menggunakan Algoritma *Bayesian Classification*. Merupakan salah satu algoritma dalam data mining yang memiliki kemampuan memprediksi dan berbasis probabilitas. Data yang diperoleh akan diolah menggunakan Ms. Excel, dimana data tersebut diproses sesuai tahapan data mining. Sumber data awal berasal dari server dengan format .MYD yaitu format database MySQL dengan sistem operasi Windows Server 2000 di PMB.

Pokok dari penelitian ini yaitu bagaimana mengetahui prediksi daftar ulang mahasiswa baru menggunakan Algoritma Bayesian Classification di universitas XYZ. Supaya dapat diketahui sebaran pola prediksi mahasiswa yang melakukan daftar ulang maupun yang tidak daftar ulang beserta penyebabnya.

Tujuan dari penelitian ini yaitu untuk mengetahui sejauh mana sebaran pola prediksi mahasiswa yang melakukan daftar ulang maupun yang tidak daftar ulang beserta penyebabnya. Sehingga pada periode PMB berikutnya dapat dicari solusinya, ada kebijakan dari pimpinan dan pengambilan keputusan yang tepat. Supaya jumlah calon

mahasiswa baru yang tidak daftar ulang mengalami penurunan.

Hasil penelitian ini diharapkan dapat membantu pimpinan perguruan tinggi dalam menentukan kebijakan dan mengambil keputusan yang tepat. Untuk UPT PMB manfaatnya adalah memiliki informasi berbasis pengetahuan tentang calon mahasiswa baru yang tidak daftar ulang. Berdasarkan informasi tersebut dapat digunakan sebagai pedoman dalam mengambil strategi atau langkah selanjutnya yang tepat dan sesuai. Sehingga penerimaan tahun berikutnya tidak mengulang kesalahan yang sama.

2. METODE

Bayesian classification adalah pengklasifikasi statistik yang dapat digunakan untuk memprediksi probabilitas keanggotaan suatu class (Wu dan Kumar, 2009). Bayesian classification didasarkan pada teorema bayes yang memiliki kemampuan klasifikasi serupa dengan decision tree dan neural network. Bayesian classification terbukti memiliki akurasi dan kecepatan yang tinggi saat diaplikasikan ke dalam database dengan data yang besar. Teorema Bayes memiliki bentuk umum sebagai berikut (Wu dan Kumar, 2009):

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

yang mana:

X = data dengan class yang belum diketahui

H = hipotesis data X suatu class spesifik

P(H|X) = probabilitas hipotesis H berdasar kondisi X (posteriori probability)

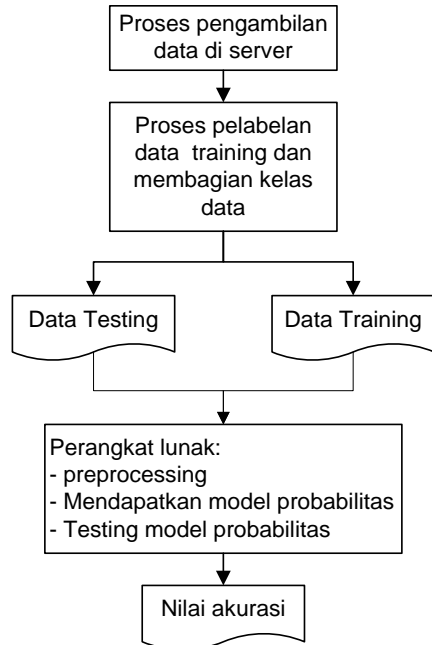
P(H) = probabilitas hipotesis H (prior probability)

P(X|H) = probabilitas X berdasar kondisi hipotesis H

P(X) = probabilitas dari X

Penelitian ini diawali dengan mengumpulkan data PMB tahun 2016 yang terletak di server PPTIK. Proses pengumpulan data pada tahap awal melibatkan semua atribut. Selanjutnya data yang telah siap diberi pelabelan sesuai dengan kelasnya yaitu kelas mahasiswa daftar ulang (DU) dan mahasiswa tidak melakukan daftar ulang. Proses daftar ulang (DU) diberikan setelah calon mahasiswa membayar biaya angsuran pertama. Sedangkan mahasiswa tidak daftar ulang

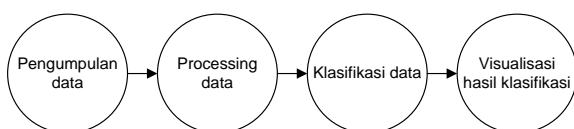
menjelaskan membayar biaya pendaftaran, sudah ujian namun tidak melakukan pembayaran angsuran pertama. Dimana angsuran pertama dijadikan indikator sebagai mahasiswa yang dipastikan kuliah dan registrasi.



Gambar 1. Alur Penelitian

Selanjutnya pembersihan data dilakukan dengan membuang atribut yang tidak dipakai. Selanjutnya data dalam format MySQL dikonversi ke format Ms. Excel untuk dilakukan proses cleaning data dan transformasi. Proses cleaning data yaitu memilah data baik sesuai kriteria dalam data mining dan data kotor. Dimana data kotor biasanya memiliki record yang tidak lengkap atau memiliki nilai data tidak sesuai. Setelah data bersih kemudian dilakukan implementasi dengan Algoritma Bayesian Classification pada proses training untuk membangun model probabilitas dari data training. Langkah berikutnya pengujian model klasifikasi yang dihasilkan pada proses training menggunakan data training.

Secara umum sistem ini terdiri dari empat bagian yaitu pengambilan data, processing data, klasifikasi data dan visualisasi hasil klasifikasi. Adapun desain arsitektur sistem terlihat pada Gambar 2.



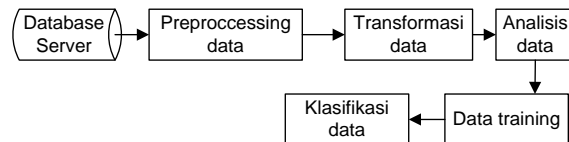
Gambar 2. Arsitektur Sistem

Penelitian ini dilaksanakan pada semester Ganjil 2017/2018 selama enam bulan dengan lokasi di

universitas XYZ. Kegiatan penelitian dilakukan di laboratorium Sistem Informasi Program Studi Teknologi Informasi. Data yang digunakan sebagai penelitian menggunakan data calon mahasiswa baru tahun 2016 yang berada di server. Format data tersimpan pada aplikasi database server MySQL. Data ini nanti perlu diolah menggunakan teknik data mining agar dapat digunakan untuk proses klasifikasi.

Dari data yang diperoleh di server untuk mengetahui sesuai tujuan penelitian ini maka ditentukan beberapa variabel. Dalam memprediksi daftar ulang mahasiswa baru dibutuhkan variabel: kota, program studi, status mendaftar dan gelombang.

Preprocessing data perlu dilakukan karena data yang diperoleh masih dalam keadaan asli belum sesuai untuk proses data mining, sehingga perlu diolah dahulu. Proses ini disebut transformasi data. Setelah data diolah sesuai kaidah data mining maka dapat diteruskan ke proses selanjutnya. Setelah data di transformasi menuju ke tahap berikutnya analisis data. Dari data training yang telah diolah kemudian dibuat prediksi klasifikasi untuk kemungkinan-kemungkinan yang akan datang bagi data testing jenis yang sama.



Gambar 3. Proses klasifikasi

3. PEMBAHASAN DAN HASIL

Dari data pendaftaran peneliti mengambil data calon mahasiswa baru tahun 2016/2017 sebagai sampel penelitian. Data diambil dari database MySQL yang berada di server. Format awal data berupa .MYD merupakan format default dari MySQL. Berikut tampilan data awal dari server.

ID	TahunID	Gelombang	Nama	NoKTP	TanggalLahir	TanggalLahir	JenisKecamin	WargaNegara	Agama
1	2016	1	KHIA MARDIANA	3306054212940002	PURWOREJO	1994-12-02	P	WNI	ISLAMI
2	2016	1	KHAWANIL KAREM	3306030308960002	BEKASI	1996-09-03	L	WNI	ISLAMI
4	2016	1	CAHYO PRIYAMBODO	3305962405940002	PURWOREJO	1994-05-24	L	WNI	ISLAMI
8	2016	1	ANDRIKHA BERKHA NUGROHO	3306067804870004	PURWOREJO	1997-04-28	L	WNI	ISLAMI
13	2016	1	BERLIAN CAHYANING SAPUTRI	0000000000000000	PURWOREJO	1999-02-08	P	WNI	ISLAMI
16	2016	1	NADIAWENGG ROCHMETSARI	1234567891011123	PALEMBANG	1996-09-09	P	WNI	ISLAMI
23	2016	1	NGATNE	3306124511940003	PURWOREJO	1994-11-06	P	WNI	ISLAMI
27	2016	1	KHAWANIL KAREM	3306030308960002	BEKASI	1996-09-03	L	WNI	ISLAMI

Gambar 4. Data awal PMB

Data tersebut diambil dari file table pmb_mahasiswadaftar2016 pada database pmb_8_2016. Data yang terekam dari table pmb_mahasiswadaftar2016 sebanyak 1003 record dari gelombang I sampai dengan gelombang ke III.

Tabel pmb_mahasiswadaft2016 dibuat dengan struktur engine MyISAM, format dynamic, memiliki satu primary key bernama id dan memiliki 72 atribut/field. Dari tabel tersebut nantinya data akan diolah, selanjutnya di mining menggunakan Bayesian Classification.

Action	Keyname	Type	Unique	Packed	Field	Cardinality	Colla	
		PRIMARY	BTREE	Yes	No	id	326	A

Space usage		Row Statistics	
Type	Usage	Statements	Value
Data	529.9 KIB	Format	dynamic
Index	12,288 B	Collation	utf8_unicode_ci
Overhead	360.7 KIB	Rows	326
Effective	181.2 KIB	Row length ø	531
Total	541.9 KIB	Row size ø	1,702 B
		Next Autoindex	1,004
		Creation	Aug 23, 2016 at 10:57 AM
		Last update	Aug 23, 2016 at 11:06 AM

Gambar 5. Table pmb_mahasiswadaftar2016

Proses selanjutnya yaitu data di konversi ke format Ms. Excel. Proses konversi ini membutuhkan waktu yang cukup lama, tergantung jumlah data yang tersimpan dalam tabel pmb_mahasiswadaftar2016. Proses konversi sendiri memiliki beberapa tahapan untuk merubah file dari format .MYD ke .XLSX.

Pertama file .MYD di proses dengan fitur import milik MySQL. Kemudian pilih fitur .CSV, ini merupakan format standar dari Ms. Excel supaya data mudah dibaca dalam bentuk tabel atau sel. Langkah kedua melakukan proses konversi ke format .CSV. Berikut hasil konversi dari format .MYD ke .CSV pad Ms. Excel. Dari format .CSV kemudian disimpan dalam format standar Ms. Excel yaitu .XLSX supaya lebih mudah dalam proses klasifikasi.

	B	C	D	E	F	G	H	I	J	K	L	M
1	1	P	ISLAM	BELUM KA	33	PURWORE	Reg	61201	61201	61201	TATA BUS	Daftar
2	1	L	ISLAM	BELUM KA	33	PURWORE	Reg	87203	61201	88201	MULTIMEI	Daftar
4	1	L	ISLAM	BELUM KA	33	PURWORE	Reg	88201	61201	83204	TEKNIK KC	Lulus
8	1	L	ISLAM	BELUM KA	33	PURWORE	Reg	61201	22201	87203	PEMESIN	Daftar
13	1	P	ISLAM	BELUM KA	33	PURWORE	Reg	84202	84202	84202	IPS	Gagal
16	1	P	ISLAM	BELUM KA	16	OGAN KCI	Reg	61201	88201	86206	SMA	Daftar
23	1	P	ISLAM	BELUM KA	33	PURWORE	Reg	54231	86206	86206	IPA	Daftar
27	1	L	ISLAM	BELUM KA	33	BANJARINI	Reg	87203	88201	61201	MULTIMEI	Daftar
28	1	P	ISLAM	BELUM KA	33	PURWORE	Reg	73201	86206	87203	AKUNTAN	Lulus
32	1	P	ISLAM	BELUM KA	33	PURWORE	Reg	88203	83204	73201	TEHNIK KE	Lulus
35	1	L	ISLAM	BELUM KA	33	PURWORE	Reg	61201	87203	88201	MULTIMEI	Lulus
36	1	L	ISLAM	BELUM KA	33	PURWORE	Reg	22201	22201	22201	TEHNIK SI	Lulus
40	1	L	ISLAM	BELUM KA	18	LAMPUNG	Reg	83204	22201	22201	IPS	Lulus
48	1	P	ISLAM	BELUM KA	33	PURWORE	Reg	86206	88203	88201	IPA	Lulus
50	1	P	ISLAM	BELUM KA	33	KEBUMEN	Reg	88203	88203	88203	IPA	Daftar
51	1	L	ISLAM	BELUM KA	33	WONOSO	Reg	86206	74201	88203	IPS	Daftar
52	1	P	ISLAM	BELUM KA	33	BANYUM	Reg	61201	86206	73201	IPS	Daftar
53	1	L	ISLAM	BELUM KA	33	KEBUMEN	Reg	54231	54231	54231	TEKNIK MI	Lulus
54	1	L	ISLAM	BELUM KA	33	WONOSO	Reg	61201	61201	61201	AKUNTAN	Lulus
65	1	P	ISLAM	BELUM KA	33	KEBUMEN	Reg	86206	73201	61201	IPS	Daftar
68	1	P	ISLAM	BELUM KA	33	KEBUMEN	Reg	84203	86206	73201	IPA	Daftar
74	1	P	ISLAM	BELUM KA	33	BANJARINI	Reg	86206	86206	86206	IPA	Daftar
75	1	L	ISLAM	BELUM KA	33	PURWORE	Reg	84202	61201	86206	TEKNIK KE	Daftar
84	1	P	ISLAM	BELUM KA	33	PURWORE	Reg	73201	73201	73201	BIOLOGI	Daftar
90	1	P	ISLAM	BELUM KA	33	PURWORE	Reg	88203	88203	88203	IPS	Daftar

Gambar 6. Hasil konversi format CSV

Tahap pertama dimulai dari *data preprocessing*, dimana menurut Larose (2005) bagian ini membutuhkan waktu hampir 60% dari proses data mining mulai dari aspek waktu sampai aspek aktifitas selama proses data mining secara keseluruhan. Tahap *data preprocessing* tampak pada gambar 7 dimana meliputi semua aktifitas untuk menghasilkan satu himpunan data yang akan dilakukan mining. Selama tahap *data preprocessing* dilakukan, data yang diperoleh mengalami berbagai perubahan format data. Dimana setiap perubahan juga dilakukan proses *data cleaning* pada semua data.



Gambar 7. Tahap *data preprocessing*

Proses seleksi data dilakukan setelah melewati tahap konversi. Proses seleksi disini yang dilakukan adalah melakukan pemotongan atribut atau field yang tidak digunakan. Beberapa field yang dipotong adalah: noktp, jenis kelamin, agama, stkawin, telphp, rt, rw, tempatlahirayah, tempatlahiribu, alamatrumah, pekerjaanayah, pekerjaanibu, pendidikanayah, pendidikanibu, gajiayah, gajiibu, waktuujian, nilai, petugas dan data lainnya yang tidak terkait dengan penelitian. Hasil seleksi data ditunjukkan pada Gambar 8. Dari 72 field pada data sumber, setelah di seleksi ada 10 field yang digunakan untuk klasifikasi. Diantaranya: field gelombang, jk, alamat, kota/kabupaten, jenis, prodi, infodaftar.

Gambar 8. Hasil seleksi data

Cleaning data adalah proses membersihkan data calon mahasiswa yang tidak terpakai. Karena data yang diperoleh dalam kondisi tidak siap digunakan maka proses cleaning data sangat diperlukan, supaya proses klasifikasi dapat dicapai. Pembersihan dilakukan agar data yang dipakai siap digunakan dalam proses data mining. Misal jk terinput K, padahal yang benar L (laki-laki) atau data lain yang tidak teridentifikasi dengan baik. Selain pembersihan, deskripsi pengkodean perlu dilakukan untuk mempermudah pemahaman selama klasifikasi. Seperti kode program studi perlu dideskripsikan dengan memberi nama program studi.

H	I	J	K	L
jenis	prodi	info dari		
Reg	61201	WEBSITE UMP	DATA 2013	
Reg	87203	INTERNET/GOOGLE	Kode Prodi	
Reg	88202	TEMAN	88201	PBSI
Reg	88201	BROSUR	88202	PBSJ
Reg	83204	WEBSITE UMP	88203	PBI
Reg	88203	BALIHU REKLAME	87203	PEKO
Tra	88203	BROSUR	84202	PMAT
Reg	61201	BALIHU REKLAME	84203	PFIS
Reg	54201	BROSUR	83204	POTO
Reg	86206	TEMAN	86206	PGSD
Reg	88201	MAHASISWA UMP	61201	MANAJ
Reg	61201	TEMAN	54201	AGRI
Reg	84202	MAHASISWA UMP	54231	PETER
Reg	22201	TEMAN	22201	SIPIL
Reg	83204	TEMAN	73201	PSIKOLOG
Reg	61201	WEBSITE UMP	74201	HUKUM
Reg	83204	GURU SEKOLAH		
Reg	61201	BALIHU REKLAME		

Gambar 9. Proses cleaning data

Langkah awal mengkonversi data dari MYD ke format XLS. Selanjutnya dalam file XLS ini dilakukan keseragaman seperti satuan barang, nama barang, id member. Keseragaman record akan mempermudah dalam proses data mining. Dari format XLS kemudian di konversi ke CSV untuk di impor ke dalam database MySQL. Keseragaman data dilakukan selama proses perubahan format data. Tahap melakukan perubahan format data tampak pada gambar 10.



Gambar 10. Tahap perubahan format data

Setelah dilakukan *cleaning data* hasilnya tampak pada gambar 20 dimana semua field data calon mahasiswa baru sudah mengalami keseragaman dalam penulisan. Pada data *pmb_mahasiswadaftar2016* proses *cleaning* dilakukan hampir semua record dengan cara melakukan tracer satu persatu atau menggunakan fitur “Find Replace” pada Ms Excel.

Proses cleaning data yang sering dilakukan terdapat pada atribut agama, jenis, prodi dan *info_daftar*. Misal agama, pada sistem PMB record agama, jenis, prodi diinput menggunakan kode. Sehingga ada beberapa user atau calon mahasiswa tidak melakukan pengisian data karena tidak mengetahui agama tertentu berkode apa, program studi yang dituju berkode apa dan lainnya. Sehingga perlu ada petunjuk dalam sistem PMB dilain waktu supaya proses data mining lebih cepat dilakukan.

Gambar 11. Hasil cleaning data yang sudah rapi dan seragam

Transformasi data berguna untuk menstrukturkan data transaksi pendaftaran menjadi bentuk yang mudah di proses oleh data mining. Beberapa data dalam database transaksi masih perlu dilakukan transformasi. Setelah dilakukan identifikasi ulang terhadap semua field, langkah berikutnya yaitu integrasi data. Contoh proses transformasi dari kode menjadi data yang lebih mudah dipahami tampak pada field prodi. Dimana di dalam sistem PMB data prodi diinputkan atau disimpn oleh sistem dalam bentuk kode. Sehingga diperlukan transformasi ke bentuk kode yang mudah dipahami.

Data yang telah bersih dari *missing value*, tidak konsisten dan tidak lengkap (Larose, 2005) dilakukan integrasi data di MySQL. Proses data *integrated* dilakukan dari berbagai data yang digabungkan untuk menghasilkan data yang disimpan pada tabel baru. Dikarenakan data yang

diperoleh berada dalam satu file maka proses integrasi data tidak perlu dilakukan. sehingga ketika mengambil data calon mahasiswa baru dari server data dalam kondisi telah terintegrasi.

TahunID	Gelombang	Nama	NoKTP	TempatLahir	TanggalLahir	JenisKeturunan	WargaNegara	Agama	StatusKawin
2016	1	KHA MARDIANA	3306054212940002	PURWOREJO	1994-12-02	P	WNI	ISLAM	BELUM KAWIN
2016	1	KHWANIL KAREM	3306030308960002	BEKASI	1996-08-03	L	WNI	ISLAM	BELUM KAWIN
2016	1	CAHYO PRYAMBODO	3306052405940002	PURWOREJO	1994-05-24	L	WNI	ISLAM	BELUM KAWIN
2016	1	ANDIKA WAHYU NUSROHO	3306067804970004	PURWOREJO	1997-04-28	L	WNI	ISLAM	BELUM KAWIN
2016	1	BERLIAN CAHYANING SAFUTRI	0000000000000000	PURWOREJO	1999-02-08	P	WNI	ISLAM	BELUM KAWIN
2016	1	NADIAJENS ROCHITASARI	1234567891011123	PALEMBANG	1998-09-09	P	WNI	ISLAM	BELUM KAWIN
2016	1	NGATIN	3306124611940003	PURWOREJO	1994-11-06	P	WNI	ISLAM	BELUM KAWIN
2016	1	KHWANIL KAREM	3306030308960002	BEKASI	1996-08-03	L	WNI	ISLAM	BELUM KAWIN
2016	1	KA FITRIANI NURJANAH	3306054303960003	PURWOREJO HARET	1996-03-03	P	WNI	ISLAM	BELUM KAWIN
2016	1	REKHA OKTAFIA PURNAMASARI	3306086910960003	PURWOREJO	1996-10-29	P	WNI	ISLAM	BELUM KAWIN
2016	1	KHWANIL KAREM	3306030308960002	BEKASI	1996-08-03	L	WNI	ISLAM	BELUM KAWIN
2016	1	WAWAN SETYAWAN	3306090212960003	PURWOREJO	1996-12-02	L	WNI	ISLAM	BELUM KAWIN

Gambar 12. Hasil integrasi data

Selanjutnya database hasil integrasi yang baru yang bernama siap dilakukan proses klasifikasi menggunakan algoritma Bayesian Classification.

Data yang dianalisis sebanyak 1003 yang berasal dari data calon mahasiswa baru TA 206/2017 dari Gelombang I – III. Dari tabel dibawah menunjukkan kondisi data yang telah siap dilakukan analisis dengan data mining dalam format XLS pada Ms Excel. Proses analisis dilakukan sesuai dengan kebutuhan klasifikasi penelitian ini yaitu melibatkan atribut atau variabel: kota, program studi, status daftar dan gelombang. Karena data yang diambil adalah TA 2016/2017 maka ada program stud baru yang tidak masuk dalam penelitian yaitu data program studi Teknologi Informasi. Sehingga dari 15 program studi, ada 14 program studi yang terlibat dalam proses analisis data mining.

I	J	K	L	M	N	O	P
ID Prodi	ID Prodi	ID Prodi	Jurusan Asal Sekolah	Ujian	Test	DU	TanggalDU
88201	61201	83204	TEKNIK KOMPUTER DAN JARINGAN	Lulus	Y		0000-00-00
61201	22201	87203	PEMESINAN	Daftar	Y		0000-00-00
84202	84202	84202	IPS	Gagal	Y	Y	14/03/2016
61201	88201	86206	SMA	Daftar	Y		0000-00-00
88203	83204	73201	TEHNIK KENDARAAN RINGAN	Lulus	Y	Y	05/04/2016
61201	87203	88201	MULTIMEDIA	Lulus	Y	Y	26/05/2016
22201	22201	22201	TEHNIK SIPIL	Lulus	Y	Y	08/04/2016
83204	22201	22201	IPS	Lulus	Y		0000-00-00
86206	88203	88201	IPA	Lulus	Y	Y	28/05/2016
88203	88203	88203	IPA	Daftar	Y	Y	03/05/2016
61201	86206	73201	IPS	Daftar	Y		0000-00-00
54231	54231	54231	TEKNIK MEKANIK OTOMOTIF	Lulus	Y	Y	03/06/2016
61201	61201	61201	AKUNTANSI	Lulus	Y		0000-00-00
86206	73201	61201	IPS	Daftar	Y	Y	12/05/2016
84203	86206	73201	IPA	Daftar	Y	Y	31/05/2016
84202	61201	86206	TEKNIK KENDARAAN RINGAN	Daftar	Y		0000-00-00
73201	73201	73201	BIOLOGI	Daftar	Y	Y	27/04/2016
88203	88203	88203	IPS	Daftar	Y	Y	10/05/2016

Gambar 13. Data calon mahasiswa siap proses analisis data mining

Sistem pendaftaran calon mahasiswa baru di universitas XYZ memiliki 3 jalur yaitu: jalur test, bebas test dan bidikmisi. Untuk jalur bebas test diberlakukan bagi calon mahasiswa yang memiliki prestasi di SMA/SMK dengan menunjukkan bukti prestasi yang diraih. Adapun

kriteria prestasi yang ditentukan adalah peringkat I sampai III baik tingkat sekolah sampai propinsi. Untuk program bidikmisi pada penelitian ini dimasukkan dalam kategori bebas test. Dimana peserta ujian cukup mengisi biodata kemudian dilanjutkan proses seleksi administrasi di BAAK. Bagi calon mahasiswa yang masuk melalui jalur transfer berasal dari D3, pada penelitian ini dimasukkan dalam kategori bebas test.

Dari data pendaftaran pada atribut kota setelah di analisis ditemukan ada 30 kota di Indonesia. Artinya calon mahasiswa baru di universitas XYZ tersebar pada 30 kota, sehingga dapat dikatakan unviersitas XYZ cukup dikenal di pulau Jawa khususnya dengan sebaran jumlah yang mendaftar terbanyak propinsi Jawa Tengah. Dari Tabel 2 tampak lima besar beberapa kota menjadi asal mahasiswa baru di universitas XYZ. Dimana kota-kita tersebut berada di wilayah kedu. Diantaranya adalah kota Purworejo, Magelang, Wonosobo, Banyumas, Temanggung.

Tabel 2. Pendaftar berdasarkan kota asal

Kota	Tidak DU	DU	Jml
Purworejo	52	722	774
Magelang	3	69	72
Wonosobo	6	54	60
Banyumas	2	14	16
Temanggung	1	12	13
Kulonprogo	0	11	11
Banjarnegara	5	4	9
Cilacap	3	3	6
Jakarta Barat	1	5	6
Brebes	3	1	4
Bekasi	0	3	3
Indramayu	2	1	3
Tangerang	2	1	3
Dumai	1	1	2
Semarang	1	1	2
Kuningan	1	1	2
Lampung	1	1	2
Rembang	2	0	2
Tebo	2	0	2
Bengkulu Utara	0	1	1
Bogor	0	1	1
Cirebon	0	1	1
Jember	1	0	1
Kerawang	1	0	1
Musi Banyuasin	0	1	1
Nganjuk	0	1	1
Ogan Komering Ulu	0	1	1
Pekalongan	0	1	1

Pemalang	0	1	1
Tanjung Jabung Barat	0	1	1

Proses analisa berikutnya yaitu atribut pendaftaran berdasarkan program studi yang dipilih calon mahasiswa yang penentuan program studi sebelum mengisi dari SIMPMB. Namun data tersebut tentu berbeda dengan data yang ada di BAAK dikarenakan ada proses data cleaning dan data selection. Seperti program studi Teknologi Informasi tidak masuk pada proses data mining. Ada perpindahan mahasiswa baru setelah proses KRS dan ada mahasiswa yang mengundurkan diri sebelum masa perkuliahan dan faktor lainnya.

Tabel 3. Pendaftar berdasarkan PS

Program Studi	Tidak DU	DU	Jml
P. Bhs Sastra Indonesia	4	77	81
P. Bhs Sastra Jawa	0	11	11
P. Ekonomi	3	29	32
P. Bahasa Inggris	10	115	125
P. Matematika	8	89	97
P. Fisika	0	9	9
P. Teknik Otomotif	7	95	102
P. Guru SD	14	134	148
Manajemen	21	212	233
Teknik Sipil	6	48	54
Agribisnis	6	37	43
Peternakan	3	14	17
Psikologi	7	22	29
Hukum	1	21	22

Tabel 4. Pendaftar berdasarkan ujian masuk

Program Studi	Tidak DU	DU	Jml
Bebas test	5	54	66
Tes	85	859	937

Tabel 5. Pendaftar berdasarkan gelombang

Program Studi	Tidak DU	DU	Jml
I	23	158	181
II	29	436	465
III	38	319	357

Setelah proses analisis data mining dilakukan berdasarkan atribut yang telah ditentukan, selanjutnya menghitung setiap item pada atribut tersebut. Perhitungan menggunakan algoritma Bayesian Classification. Penghitungan dimulai dari atribut kota asal, program studi yang diambil, status tes dan gelombang.

Berdasarkan data atribut nomial, kemudian identifikasi data calon mahasiswa dihitung $P(H_i)$ pada setiap atribut. Pada kasus penelitian dataset ini terdapat dua *class* atribut yaitu TDU (Tidak Daftar Ulang) = “tidak” dan DU (Daftar Ulang) = “ya”. Kemudian hitung $P(X|H_i)$ dimana $i=1,2$ untuk setiap *class* atau atribut. Setelah perhitungan selesai, kemudian dibandingkan. Jika $P(X|H_1) > P(X|H_2)$ maka kesimpulan H_1 adalah *class* DU = “ya”. Jika $P(X|H_1) < P(X|H_2)$ maka kesimpulan H_2 adalah *class* TDU = “tidak”. Dari data-data analisis diatas terdapat 2 *class* dari klasifikasi yang dibentuk yaitu:

- $H_1 = (class \text{ daftar ulang}) = \text{“ya”}$
- $H_2 = (class \text{ tidak daftar ulang}) = \text{“tidak”}$

Dari kedua *class* klasifikasi diatas maka untuk menghitung *class* dari TDU dan DU berdasarkan klasifikasi yang terbentuk (*prior probability*). Maka dibutuhkan untuk memaksimalkan diperoleh $P(X|H_i)$ dimana $i=1,2$. Selanjutnya $P(H_i)$ merupakan *prior probability* untuk setiap *class* berdasarkan data berikut dan diperoleh:

- $P(H_1 \text{ class DU = “ya”}) = 913/1003 = 0.910$
- $P(H_2 \text{ class TDU = “tidak”}) = 90/1003 = 0.089$

Kemudian hitung $P(X|H_i)$ dimana $i=1,2$ dari tabel 1, tabel 2, tabel 3 dan tabel 4. Untuk tabel 1 proses klasifikasi *Bayesian* dilakukan pada 5 data teratas. Hal ini mempertimbangkan fokus terhadap jumlah calon mahasiswa potensial. Untuk tabel 2 proses klasifikasi dilakukan pada program studi dengan jumlah pendaftar dibawah 70.

Berikutnya membandingkan hasil *class* atribut. Perhitungan diatas $P(X|H_i) P(H_i) / P(H_i)$ ditunjukkan pada tabel dibawah. Dimana bila $P(X|H_1) > P(X|H_2)$ maka kesimpulannya H_1 adalah *class* yang daftar ulang.sebaliknya bila $P(X|H_1) < P(X|H_2)$ maka H_2 adalah *class* yang tidak daftar ulang.

Tabel 6. Perbandingan hasil class atribut

PS Pilihan	Kota Calon Mhs	Class DU
P. Bhs Jawa	Purworejo	Tidak
	Magelang	Tidak
	Wonosobo	Tidak
	Banyumas	Tidak
	Temanggung	Tidak
P. Ekonomi	Purworejo	Tidak
	Magelang	Ya
	Wonosobo	Ya
	Banyumas	Tidak

P. Fisika	Temanggung	Ya
	Purworejo	Tidak
	Magelang	Tidak
	Wonosobo	Tidak
	Banyumas	Tidak
Teknik Sipil	Temanggung	Tidak
	Purworejo	Tidak
	Magelang	Ya
	Wonosobo	Tidak
	Banyumas	Tidak
Agribisnis	Temanggung	Tidak
	Purworejo	Ya
	Magelang	Ya
	Wonosobo	Tidak
	Banyumas	Ya
Peternakan	Temanggung	Ya
	Purworejo	Tidak
	Magelang	Ya
	Wonosobo	Tidak
	Banyumas	Tidak
Psikologi	Temanggung	Ya
	Purworejo	Tidak
	Magelang	Tidak
	Wonosobo	Tidak
	Banyumas	Ya
Hukum	Temanggung	Ya
	Purworejo	Ya
	Magelang	Ya
Hukum	Wonosobo	Tidak
	Banyumas	Ya
	Temanggung	Ya

4. SIMPULAN

Berdasarkan perhitungan data mining menggunakan algoritma Bayesian Classification untuk data calon mahasiswa dapat diperoleh kesimpulan bahwa class TDU 55% lebih besar dari pada class DU 45%. Artinya calon mahasiswa yang tidak daftar ulang secara keseluruhan semua program studi lebih besar.

Berdasarkan proses perhitungan data mining menggunakan Algoritma Bayesian classification dihasilkan suatu informasi baru seperti tabel 5. Berdasarkan tabel 5 dijelaskan bahwa program studi PBSJ 100% tidak ada calon mahasiswa yang daftar ulang. Program studi Pendidikan Ekonomi masih ada peluang 75% calon mahasiswa yang DU. Program studi Pendidikan Fisika 100% tidak ada calon mahasiswa yang mendaftar. Program studi Teknik Sipil ada peluang 25% calon mahasiswa akan DU. Program Studi Agribisnis

ada peluang 75% calon mahasiswa melakukan DU. Program studi Peternakan 90% tidak ada peluang calon mahasiswa yang mendaftar. Program studi Psikologi memiliki peluang sama dengan program studi Peternakan yaitu 90% tidak ada peluang calon mahasiswa mendaftar. Untuk program studi Hukum memiliki peluang 90% calon mahasiswa mendaftar.

Algoritma Bayesian Classification dapat digunakan untuk menemukan permasalahan pada data calon mahasiswa mengenai peluang apakah seorang calon mahasiswa akan DU atau TDU pada program studi tertentu.

DAFTAR PUSTAKA

- A.G. Mabur dan R. Lubis, 2012, Penerapan Data Mining Untuk Memprediksi Kriteria Nasabah Kredit, Jurnal Komputa, Vol. 1 pp. 53 – 57
- Budi S., 2007, Data Mining: Teknik Pemanfaatan Data Untuk Keperluan Bissnis, Graha Ilmu, Yogyakarta
- Dewi S., 2012, Algoritma Bayesian Classification Untuk Memprediksi Heregristrasi Mahasiswa Baru STMIK Widya Pratama, Jurnal ICTech, Vol. 10 No. 2 Mei
- Kusrini dan Luthfi T. Emha, 2009, Algoritma Data Mining, Andi, Yogyakarta
- Larose T. Daniel, 2005, Discovering Knowledge In Data, An Introduction to Data Mining, Wiley-Interscience a John Wiley and Sons, Inc., Publication Taylor and Francis Group, London
- Muwardah F.R., dan Pramunendar R.A., 2014, Penentuan Penerimaan Mahasiswa Baru Menggunakan Decision Tree, Penelitian Skripsi, Universitas Dian Nuswantoro, Semarang
- Nugroho, Y.S., 2011, Data Mining Menggunakan Algoritma Naive Bayes Untuk Klasifikasi Kelulusan Mahasiswa, Penelitian Skripsi, Universitas Dian Nuswantoro, Semarang

- Rodiyansyah S.F. dan Winarko E., 2012, Klasifikasi Posting Twitter Kemacetan Lalu Lintas Kota Bandung Menggunakan Naive Bayesian Classification, IJCCS, Volume 6 Nomor 1 Januari 2012, pp 91 – 100
- Sandi Fajar R., dan Winarko E., 2012, Klasifikasi Posting Twitter Kemacetan Lalu Lintas kota Bandung Menggunakan Naive Bayesian Classification, Jurnal IJCCS Vol. 6 No. 1 pp. 91 – 100 Januari
- Sugianti S., 2012, Algoritma Bayesian Classification Untuk Memprediksi Heregristrasi Mahasiswa Baru, Penelitian Skripsi, STMIK Widya Pratama
- Susanto S., dan Suryadi D., 2010, Pengantar Data Mining, Menggali Pengetahuan Dari Bongkahan Data, Andi, Yogyakarta
- Wu Xindong, dan Kumar Vipin, 2009, The Top Ten Algorithm in Data Mining, CRC Press
- Widiastuti D., 2010, Analisa Perbandingan Algoritma SVM, Naive Bayes dan Decision Tree Dalam Mengklasifikasikan Serangan Pada Sistem Pendeteksi Intrusi, Penelitian Skripsi, Universitas Gunadharma, Jakarta