Jurnal Teknologi Informasi dan Ilmu Komputer (JTIIK) DOI: 10.25126/jtiik.202072585
Vol.7, No. 3, Juni 2020, hlm. 555-564 p-ISSN: 2355-7699
Akreditasi KEMENRISTEKDIKTI, No.30/E/KPT/2018 e-ISSN: 2528-6579

PEMANFAATAN DATA PDDIKTI SEBAGAI PENDUKUNG KEPUTUSAN MANAJEMEN PERGURUAN TINGGI

Ngatmari[1], Muhammad Bisri Musthafa[2], Cahya Rahmad[3], Rosa Andrie Asmara[4], Faisal Rahutomo*[5]

[1,2]Jurusan Teknik Elektro, Politeknik Negeri Malang, [3,4,5]Jurusan Teknologi Informasi, Politeknik Negeri Malang
Email:[1]marymalangcity@gmail.com,[2]ichsancomp@gmail.com, [3]cahya.rahmad@polinema.ac.id,
[4]rosa.andrie@polinema.ac.id, [5]faisal@polinema.ac.id
*Penulis Korespondensi

(Naskah masuk: 16 Oktober 2019, diterima untuk diterbitkan: 27 April 2020)

Abstrak

Pangkalan Data Pendidikan Tinggi (PDDIKTI) merupakan sebuah sistem penyimpan data yang dikelola Pusat Data dan Informasi (Pusdatin) Kementrian Ristek dan Pendidikan Tinggi. Data yang tersedia di PDDIKTI merupakan data yang akurat, karena proses pelaporan data akademik secara berkala dua kali setiap semester. Data yang telah berlimpah tersebut, tentu sangat disayangkan jika tidak digunakan untuk keperluan yang lebih bermanfaat, misal untuk mengetahui pola akademik kelulusan mahasiswa dan prestasi akademik mahasiswa. Untuk memperoleh informasi-informasi penting tersebut bisa dilakukan dengan cara penggalian informasi (knowledge discovery). Teknik dalam memberikan solusi masalah tersebut adalah teknik klasifikasi untuk membantu pengambilan keputusan, misalkan Decission Tree (C4.5, ID3, CHAID, rule induction) dan teknik peramalan (forecasting) menggunakan metode simple moving average (SMA). Tujuan dari penambangan data PDDIKTI adalah untuk melakukan deteksi dini terhadap mahasiswa, sehingga dosen bisa memberikan masukan-masukan ketika mahasiswa tersebut telah diklasifikan sebagai mahasiswa yang lulus tidak tepat waktu serta memprediksi jumlah mahasiswa yang akan masuk pada perguran tinggi pada salah satu prodi X, sehingga manajemen baik tingkat program studi maupun universitas bisa melakukan langkah-langkah yang dianggap penting guna meningkatkan jumlah mahasiswa. Pengujian pada 2.601 record akademik mahasiswa dengan atribut ipk_sem1, ipk_sem2, ipk_sem3, ipk_sem4, pekerjaan_ortu, ket_lulus, rerata_ipk, penghasilan_ayah, untuk klasifikasi kelulusan mahasiswa menghasilkan nilai *accuracy* 86,54% nilai *precission* 93,37% dan nilai *recall* 89,27% serta pengujian prediksi jumlah peminat program studi diperoleh nilai *MAPE* sebesar 21,75 %.

Kata kunci: *PDDIKTI, lulus tepat waktu, peminat, C4.5, simple moving average*

UTILIZATION OF PDDIKTI DATA AS A HIGHER EDUCATION MANAGEMENT DECISSION SUPPORT

Abstract

The Higher Education Database (PDDIKTI) is a data storage system managed by the Center for Data and Information (Pusdatin) of the Ministry of Research and Technology and Higher Education. The data available at PDDIKTI is accurate data, because the process of reporting academic data regularly twice each semester. The abundant data is certainly unfortunate if not used for more useful purposes, for example to find out the academic patterns of student graduation and student academic achievement. To obtain important information can be done by extracting information (knowledge discovery). Techniques in providing solutions to these problems are classification techniques to assist decision making, for example Decission Tree (C4.5, ID3, CHAID, rule induction) and forecasting techniques using simple moving average (SMA) methods. The purpose of PDDIKTI data mining is to conduct early detection of students, so that lecturers can provide input when the students have been classified as students who graduate not on time and predict the number of students who will enter the tertiary institutions in one of the X study programs, so that management both the level of study program and university can take steps that are considered important to increase the number of students. Tests on 2601 student academic records with the attributes ipk_sem1, ipk_sem2, ipk_sem3, ipk_sem4, occupation_ortu, graduated, average_ipk, income_ayah, for the graduation classification of students resulted in an accuracy value of 86.54% a value of 93.37% and a recall value of 89.27% and a test of 89.27% and a test of graduation prediction of the number of study program enthusiasts obtained a MAPE value of 21.75%.

Keywords: *PDDIKTI*, *on time graduation, enthusiasts, C4.5 algorithm, simple moving average*

1. PENDAHULUAN

Pangkalan Data Pendidikan Tinggi (PDDIKTI) merupakan sebuah sistem penyimpan data yang dikelola Pusat Data dan Informasi (Pusdatin) Kementrian Ristek dan Pendidikan Tinggi (Kemristekdikti)(Kementrian Riset, 2017). Perguruan tinggi dapat mengolah data yang menunjukkan profil perguruan tinggi tersebut melalui sebuah antarmuka aplikasiPDDIKTI FEEDER. Kemudian data tersebut disinkronisasi dengan data yang ada di Pusdatin Kemristekdikti.

Data yang tersedia di PDDIKTI merupakan data yang akurat, karena proses pelaporan data akademik secara berkala dilakukan dua kali setiap semester dan perkembangan akademik setiap mahasiswa dapat ditampilkan pada aplikasi Forlap yang bisa diketahui oleh masyarakat luas.

Data yang telah berlimpah tersebut, tentu sangat disayangkan ketika tidak digunakan untuk keperluan yang lebih bermanfaat, misalkan sebagai bahan penilaian akreditasi, penilaian kinerja dosen, pola akademik kelulusan mahasiswa, strategi pemasaran perguruan tinggi dan lain-lain, untuk memperoleh informasi-informasi penting tersebut bisa dilakukan dengan cara penggalian informasi (*knowledge discovery*).

Dalam dunia pendidikan sebagai salah satu tujuan kegiatan pendidikan dan pengajaran di perguruan tinggi adalah mencetak mahasiswa yang berkualitas, beberapa cara yang dilakukan antara lain : memonitoring nilai akademik mahasiswa, meningkatkan kualitas dosen dan tenaga kependidikan, memberikan *reward* (beasiswa) kepada mahasiswa yang berprestasi dan memastikan mahasiswa bisa lulus tepat waktu. Pemberian *reward* tersebut diharapkan bisa memicu kepada mahasiswa untuk semakin memicu nilai akademik. Setiap perguruan selalu menekankan kepada dosen asuh (dosen wali) untuk selalu mendampingi serta mendorong anak asuhnya dalam menempuh studi dengan melihat perkembangan indeks prestasi yang diperoleh agar nilai akademik anak didiknya selalu berpredikat baik, dengan nilai akademik yang baik tersebut diharapkan mahasiswa bisa lulus sesuai dengan ketentuan, yakni 8 semester untuk jenjang S1, namun terkadang masih ada beberapa mahasiswa yang masih terlambat menyelesaikan studi, yang berakibat menurunnya kualitas program studi, karena kelulusan mahasiswa tepat waktu menjadi tolok ukur tingkat keberhasilan perguruan tinggi dalam mendidik mahasiswa.

Data mining merupakan penambangan atau penemuan informasi baru dengan cara menemukan pola atau aturan tertentu dari sejumlah data dalam skala besar, dengan cara ini diharapkan bisa memberikan solusi terhadap permasalahan tersebut, yakni dengan cara mempelajari pola akademik mahasiswa, dengan cara ini pihak perguruan tinggi bisa melakukan deteksi dini terhadap mahasiswa yang berpeluang lulus tidak tepat waktu dan memberikan rekomendasi-rekomendasi atau masukan agar bisa lulus tepat waktu. Disamping itu keberadaan calon pendaftar pada perguruan tinggi menjadi perhatian yang sangat penting, karena jika hal ini tidak diperhatikan maka akan berdampak pada pencapaian harapan setiap perguruan tinggi dalam peningkatan mutu pendidikan, pelayanan dan nilai akreditasi perguruan tinggi. Salah satu teknik untuk memberikan solusi masalah tersebut adalah teknik klasifikasi untuk membantu dalam pengambilan keputusan, misalkan *Decission Tree* (C4.5, ID3, CHAID, *rule induction*) dan teknik peramalan jumlah mahasiswa yang mendaftar pada salah satu program studi. Penelitian dilakukan oleh (Rohman, 2015) untuk memprediksi kelulusan mahasiswa menggunakan teknik klasifikasi data mining algoritma *K-Nearest Neighbor* (K-NN) dengan parameter usia, jenis kelamin, indeks prestasi semester 1-4 dengan *accuracy* 85,15% . penelitian terkait klasifikasi kelulusan mahasiswa dilakukan oleh (Yunianita & others, 2018) menggunakan algoritma C4.5 dengan data latih mahasiswa tahun 2010-2013 dievaluasi dengan *confusion matrix* yang menghasilkan *accuracy* 73,99%. Penelitian lain di bidang pendidikan (Alverina, Chrismanto, & Santosa, 2018) untuk memprediksi kategori IP Mahasiswa Baru pada jalur prestasi, Algoritma C4.5 dan CART menghasilkan *accuracy* yang sama yakni sebesar 86,86%, sedangkan untuk memprediksi kategori IP Mahasiswa baru pada jalur non-prestasi (data numerik), Algoritma CART menghasilkan *accuracy* yang lebih baik dibandingkan algoritma C4.5 yaitu 63,16% dan 61,54%. Penelitian terkait tentang prediksi jumlah mahasiswa registrasi per semester menggunakan teknik linear regresi dan *MAPE* yang dilakukan oleh Amirudin (2018), dalam mengatasi permasalahan mahasiswa tidak melakukan herregistrasi semakin meningkat sehingga mahasiswa yang herregistrasi semakin menurun dengan mengambil *sample* 2 program studi yaitu Teknik Informatika didapatkan hasil tingkat error sebesar 4,24% atau tingkat akurasi 95,76%, sedang untuk program studi yang lain adalah program studi Ilmu Hukum dengan tingkat *error* sebesar 7,69% atau tingkat akurasi sebesar 92,31%.

Dari uraian tersebut, bahwa penelitian di bidang pendidikan untuk klasifikasi dan prediksi (*forecasting)* menggunakan algoritma C4.5 danalgoritma *Moving Average (*rata-rata bergerak*)* dengan memanfaatkan data akademik dan profil mahasiswa, pada penelitian ini mengeksploitasi data PDDIKTI yang telah ditransformasi menjadi *Data warehouse*(Rahutomo, Rahmad, Musthafa, & Ngatmari, 2019) untuk bisa dimanfaatkan oleh pihak manajemen. Penelitian ini akan membahas 2 studi kasus : Mengklasifikasikan mahasiswa berdasarkan riwayat akademik, dalam hal ini adalah IPK semester 1-4 dan jumlah SKS yang sudah ditempuh serta latar belakang orang tua dan memprediksi jumlah

mahasiswa baru yang melakukan daftar ulang pada masing-masing program studi yang dikehendaki.

Data training yang digunakan adalah bersumber dari PDDIKTI yang ditransformasi dengan atribut nipd, ipk1, ipk2, ipk3, ipk4, jumlah SKS, keterangan lulus, rerata IPK 4 semester terakhir, pekerjaan ayah, pendidikan ayah, pendidikan ibu, penghasilan ayah, penghasilan ibu, status ayah, status ibu. Permodelan data dilakukan dengan data .csv, tidak secara otomatis diambil langsung dari data PDDIKTI.

2. METODE PENELITIAN

2.1 Pemanfaatan Data PDDIKTI

PDDIKTI menjadi salah satu instrument pelaksanaan penjaminan mutu (Kementrian Riset, 2017). Dalam pasal 56 ayat 2 UU No. 12 Tahun 2012 tentang Pendidikan Tinggi menyebutkan bahwa Pangkalan Data Pendidikan Tinggi sebagaimana dimaksud pada ayat (1) berfungsi sebagai sumber informasi bagi:Lembaga akreditasi, Pemerintah, dan Masyarakat (Depdiknas, 2012).

Mahasiswa dinyatakan lulus ketika sudah menyelesaikan seluruh kewajiban akademik, yakni telah menyelesaikan seluruh SKS yang telah ditentukan oleh program studi dengan nilai atau IPK minimal yang telah ditentukan oleh program studi. Selain kedua persyaratan tersebut setiap program studi mempunyai kriteria-kriteria bahwa mahasiswa tersebut dinyatakan telah selesai / lulus masa studi, dengan mengikuti ketentuan yang tercantum dalam peraturan SNDIKTI nomor 44 tahun 2015, sehingga mahasiswa yang sudah lulus tersebut berhak mendapatkan Nomor Ijazah Nasional (NINA).

Jumlah Mahasiswa baru yang melakukan daftar ulang di Universitas XYZ dengan 33 Program Studi dari berbagai daerah dan latar belakang, untuk pemetaan wilayah, asal sekolah, latar belakang keluarga dalam menetapkan skala prioritas pada kegiatan promosi penerimaan mahasiswa baru. Jumlah data mahasiswa diambil mulai angkatan 2007 sampai angkatan 2018 dengan melihat *trend* perkembangan setiap tahunnya. Dengan melihat pola perkembangan jumlah mahasiswa baru ini diharapkan bisa dijadikan dasar untuk kepentingan manajemen dalam memprediksi keberadaan mahasiswa yang akan datang dengan melihat beberapa variabel yang terlibat didalamnya.

2.2 Algoritma C4.5

Algoritma C4.5 merupakan algoritma yang digunakan untuk membentuk pohon keputusan (*Decision Tree*). Pohon keputusan merupakan metode klasifikasi dan prediksi yang terkenal. Ada beberapa tahapan dalam membuat sebuah pohon keputusan dalam algoritma C4.5 (Larose & Larose, 2014). Mempersiapkan data training. Kemudian menghitung akar dari pohon. Akar akan diambil dari atribut yang akan terpilih, dengan cara menghitung nilai *gain* dari masing-masing atribut, nilai *gain* yang paling tinggi yang akan menjadi akar pertama. Sebelum menghitung nilai *gain* dari atribut, hitung dahulu nilai *entropy*. Untuk menghitung nilai entropy dan *gain* digunakan persamaan (1) dan (2)(Larose & Larose, 2014).

$$Entropy\ (S) = \sum_{i=1}^{n} - pi\ log_2\ pi \quad (1)$$

$$Gain=entropy\ (S) - \sum_{i=1}^{n} * \frac{|Si|}{S} * Entropy\ (Si) \quad (2)$$

2.3 Simple Moving Average(SMA)

Metode ini merupakan bagian dari model deret berkala (*time series*), yaitu metode yang digunakan untuk menganalisis serangkaian data yang merupakan fungsi waktu.Persamaan (3) menunjukkan rumus dasar metode ini.

$$MA = \frac{\sum X}{jumlah\ periode} \quad (3)$$

dimana :
MA= Moving Average
Σx= Penjumlahan semua data yang diperhitungkan

Atau dapat ditulis dengan persamaan (4).

$$MA = (n_1 + n_2 + n_3 + \ldots) / n \quad (4)$$

dimana :
MA= Moving Average
n_1= Data nilai pertama
n_2= Data nilai kedua
n_3= Data nilai ketiga
n= Jumlah nila rata-rata bergerak

Sebagai pengujian performa berdasarkan model prediksi yang telah dibuat dengan input data testing adalah menggunakan *Mean Absolut Percentage Error (MAPE)* merupakan metode yang digunakan untuk menilai tingkat akurasi(Tannady & Andrew, 2013). Persamaan (5) menunjukkan rumus tersebut.

$$MAPE = \frac{\sum \frac{|y-y'|}{y} x100\%}{n} \quad (5)$$

dimana :
y' = hasil prediksi
y = data aktual
n = jumlah data

2.4 PenerapanData Mining

Penerapan data mining memiliki enam fase *CRISP-DM* (*Cross Industry Standard Prosess for Data Mining*) (Larose & Larose, 2014)(Han, Pei, & Kamber, 2011)(Aggarwal, 2015)(Zaki, Meira Jr, & Meira, 2014)yaitu :

1. *Business Understanding*, mendefinisikan objek bisnis sehingga memperoleh model terbaik sesyau dengan tujuan bisnis.
2. *Data Understanding*, memeriksa data sehingga memperoleh permasalahan-permasalahan yang ada dalam data.

3. *Data Preparation*, memperbaiki permasalahan data.
4. *Modeling*, membuat permodelan
5. *Evaluation*, memberikan penilaian antara hasil permodelan dengan tujuan bisnis.
6. *Deployment*, penerapan permodelan.

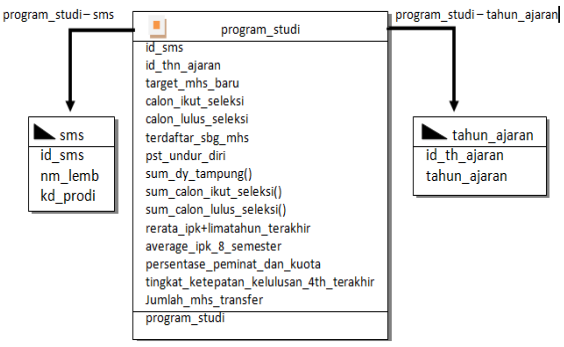
2.5 Data

Data yang digunakan untuk proses mining adalah Data *warehouse* PDDIKTI Universitas XYZ pada tahun akademik 2007-2018 serta didukung dengan Sisfo (Sistem Informasi) internal kampus. Beberapa keputusan manajemen yang dapat diperoleh secara otomatis berdasarkan data *warehouse* PDDIKTI (Rahutomo et al., 2019), explorasi penggunaan data *warehouse* dijelaskan pada Tabel 1.

Tabel 1. Eksplorasi penggunaan Data warehouse

No	Tabel Fakta	Eksplorasi Fungsi
1	program_studi	Ramalan peminat mahasiswa pada program studi
2	program_studi	Prediksi jumlah mahasiswa transfer pada program studi
3	program_studi	Pengembangan Program Studi
4	program_studi	Klasifikasi Akreditasi Program Studi
5	mahasiswa	Klaster Mahasiswa layak mendapatkan bantuan pendidikan
6	mahasiswa	Strategi Promosi mahasiswa
7	mahasiswa	Klasifikasi Kelulusan tepat waktu
8	mahasiswa	Klasterisasi mahasiswa berprestasi
9	mahasiswa	Klasifikasi mahasiswa potensi DO
10	pengajaran_dosen	Klasterisasi Kinerja Dosen
11	status_mahasiswa	Dashboard statistik mahasiswa
12	persen_ipk	Dashboard statistik IPK mahasiswa

2.6 Fakta Program_Studi



Gambar 1. Fakta program_studi

Fakta program studi seperti ditunjukkan pada Gambar 1 berfungsi untuk menampung data-data terkait dengan program studi pada setiap tahun dengan atribut-atribut seperti ditunjukkan pada Tabel 2, yang dapat dieksplorasi penggunaannya pada bagian 2.6.1 s/d 2.6.4.

2.6.1 Ramalan Peminat Mahasiswapada Program Studi

Berguna untuk meramal peminat calon mahasiswa baru berdasarkan data jumlah peminat pada tahun sebelumnya, dengan tujuan agar pihak manajemen bisa menentukan kebijakan ketika jumlah mahasiswa diprediksi naik maupun turun. Fungsi ini dapat dilakukan dengan teknik *forecasting* menggunakan algoritma s*ingle moving average*(*SMA*). Atribut yang digunakan adalah calon_ikut_seleksi pada setiap tahun.

Tabel 2. Deskripsi atribut fakta program_studi

Nama Atribut	Deksripsi
id_sms	kode program studi
id_th_ajaran	kode tahun sebagai identifikasi tahun
target_mhs_baru	Target kuota pendaftaran mahasiswa baru setiap program studi
calon_ikut_seleksi	Jumlah riil pendaftar yang mengikuti seleksi penerimaan mahasiswa baru.
calon_lulus_seleksi	Jumlah riil pendaftar yang lulus seleksi penerimaan mahasiswa baru.
terdaftar_sbg_mhs	Jumlah riil pendaftar yang lulus seleksi dan melakukan registrasi.
pst_undur_diri	Jumlah riil yang lulus seleksi, tetapi tidak melakukan registrasi
rerata_ipk_limatahun_terakhir	Rata-rata IPK seluruh mahasiswa dalam program studi pada 5 tahun terakhir
average_ipk_8_semester	Rata-rata IPK seluruh mahasiswa dalam program studi pada 8 semester terakhir
persentase_peminat_dan_kuota	Persentase antara jumlah pendaftar dengan kuota program studi.
tingkat_ketepatan_kelulusan_4th_terakhir	Tingkat ketepatan kelulusan program studi selama 4 tahun terakhir.
jumlah_mhs_transfer	Jumlah mahasiswa baru yang masuk melalui jalur transfer pada setiap program studi

2.6.2 Prediksi Jumlah Mahasiswa Transfer Pada Program Studi

Fungsi ini berguna untuk memprediksi jumlah mahasiswa yang masuk pada program studi melalui jalur transfer. Fungsi ini dapat dilakukan dengan teknik *forecasting* menggunakan algoritma s*ingle moving average*(SMA). Atribut yang digunakan adalah jumlah_mhs_transfer pada setiap tahun.

2.6.3 Pengembangan Program Studi

Prediksi pengembangan prodi berdasarkan peminat pada prodi tertentu semakin meningkat, dengan variabel *average* IPK selama 8 semester terakhir, tingkat ketepatan kelulusan mahasiswa selama 3-4 tahun terakhir, persentase peminat prodi terhadap kuota, menggunakan pendekatan *clustering* dengan menggunakan *Library SciKit Learn* dengan *alternative* algoritma *K-Means*, DBSCAN. *Library* yang digunakan pandas, matplotlib, numpy, sklearn (algoritma *K-Means* dan DBSCAN serta *Preprocessing*). atribut-atribut yang digunakan :

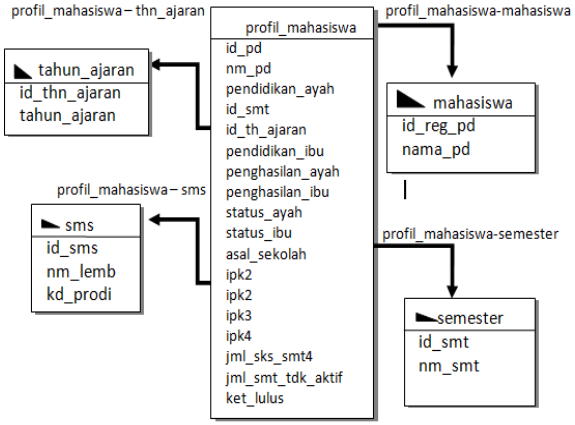
id_sms, id_th_ajaran, prosentase_peminat, average_ipk_8_semester, tingkat_ketepatan_kelulus-an_4th_terakhir.

Adapun pengembangan yang dimaksud bisa meliputi penyediaan lokal ruangan untuk proses belajar mengajar (PBM), laboratorium, jumlah dosen/ pengajar, tenaga administrasi maupun untuk sarana-sarana yang lainnya.

2.6.4 Klasifikasi Akreditasi Program Studi

Dalam menunjang kebijakan manajemen pada tingkat program studi, maka perlu dilakukan kajian-kajian serta melakukan analisis tentang klasifikasi akreditasi program studi, dimana variabel-variabel yang akan digunakan terdapat 4 variabelyaitu : rasio mahasiswa peminat dengan kuota prodi, *average* IPK lima tahun terkahir, rasio pendaftar dengan daya tampung, rasio mahasiswa yang melakukan registrasi dengan mahasiswa yang lulus seleksi dan prosentase kelulusan tepat waktu. Menggunakan pendekatan *classification* dengan menggunakan *Library SciKit Learn*serta*alternative* algoritma KNN, Naïve Bayes, *Decission Tree*. Library yang digunakan pandas, matplotlib, numpy, sklearn (algoritma KNN, Naïve Bayes, *Decission Tree* serta *Preprocessing*). Atribut-atribut yang digunakan : rasio_minat_kuota, avg_ipk, rasio_lulus_registrasi, pros_lulus_tepat_waktu. Semua atribut tersebut digunakan untuk pemodelan. Untuk label yang digunakan, yaitudata riwayat nilai_akreditasi program studi yang diperoleh dari Kantor Jaminan Mutu (KJM) Universitas XYZ. Atribut-atribut yang digunakan adalah: rasio_minat_ kuota, avg_ipk, rasio_lulus_ registrasi, pros_lulus_ tepat_waktu

2.7 Fakta Mahasiswa



Gambar 2. Fakta mahasiswa

Fakta mahasiswa yang ditunjukkan pada Gambar 2 berfungsi untuk menampung data-data yang terkait dengan profil dan akademik mahasiswa pada setiap tahun, dengan detail atribut dijabarkan pada Tabel 3. Fakta tersebut dieksplorasi untuk penggunaan yang dibahas pada Bagian 2.7.1 s/d 2.7.4.

2.7.1 Strategi Promosi Mahasiswa

Pengelompokan asal sekolah mahasiswa (SLTA/SMK/MA), serta kelulusan tepat waktu untuk skala prioritas pemetaan wilayah pendaftar, hal ini dilakukan untuk menentukan strategi promosi perguruan tinggi terhadap suatu tempat (sekolah, wilayah kab/kota, khalayak umum), menggunakan pendekatan *clustering* dengan menggunakan *LibrarySciKit Learn* dengan *alternative* algoritma *K-Means*, DBSCAN. *Library* yang digunakan pandas, matplotlib, numpy, sklearn (algoritma *K-Means* dan DBSCAN serta *Preprocessing*). atribut-atribut yang digunakan adalah : asal_sekolah, ipk, ket_lulus yang telah ditunjukkan pada Tabel 3 berikut :

Tabel 3. Deskripsi atribut fakta mahasiswa

Nama Atribut	Deksripsi
id_pd	Kode mahasiswa
nama_pd	Nama mahasiswa
id_smt	Kode Semester
id_th_ajaran	Kode tahun ajaran
id_sms	Kode program studi
pendidikan_ayah	Pendidikan terakhir ayah
pendidikan_ibu	Pendidikan terakhir ibu
penghasilan_ayah	Penghasilan ayah
penghasilan_ibu	Penghasilan ibu
status_ayah	Status Ayah (Hidup/Mati)
status_ibu	Status Ibu (Hidup/Mati)
asal_sekolah	Asal sekolah sebelum masuk perguruan tinggi (SMAN /SMA/MA/MAN/SMK/SMKN)
ipk1	IPK pada semester 1
ipk2	IPK pada semester 3
ipk3	IPK pada semester 3
ipk4	IPK pada semester 4
jml_sks_smt4	Jumlah SKS yang sudah ditempuh sampai semester 4
jml_smt_tidak_aktif	Jumlah semester yang tidak aktif (telah melakukan registrasi, tetapi tidak mengikuti kuliah)

2.7.2 Klasifikasi Kelulusan Tepat Waktu

Berdasarkan IPK semester 1-4 dan profil orang tua. Prediksi ini dilakukan untuk melakukan deteksi dini terhadap mahasiswa yang terklasifikasikan mahasiswa lulus tidak tepat waktu, sehingga dari manajemen tingkat program studi dapat memberikan tindakan/masukan yang dianggap perlu untuk dilakukan. Menggunakan pendekatan *Classification* dengan menggunakan *Library SciKit Learn*serta*alternative* algoritma KNN, naïve bayes, *decision tree*. *Library* yang digunakan adalah pandas, matplotlib, numpy, sklearn (algoritma KNN, naïve bayes, *decision tree* serta *Preprocessing*). Atribut-atribut yang dipakai adalah : ipk1, ipk2, ipk3, ipk4, pendidikan_ayah, pendidikan_ibu, penghasilan_ ayah, penghasilan_ibu, status_ayah, status_ibu. Kemudian untuk label yaitu keterangan_lulus. Semua atribut yang telah disebutkan diatasakan digunakan untuk pemodelan.

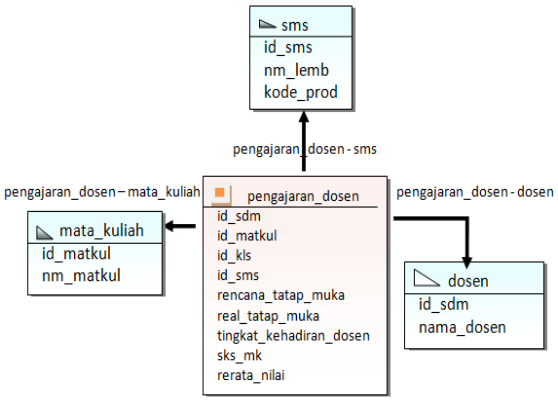
2.7.3 Klasterisasi Mahasiswa Berprestasi

Klasterisasi mahasiswa berprestasi berdasarkan IPK 4 semester terakhir, latar belakang keluarga (pendidikan orang tua, penghasilan orang tua, penanggung_biaya), dengan tujuan untuk mempermudah proses seleksi penerimaan bantuan pendidikan.

2.7.4 Klasifikasi Mahasiswa Potensi *Drop Out*

BerdasarkanIPK semester 1-4 dan jumlah semester yang tidak aktif, prediksi ini dilakukan untuk melakukan deteksi dini terhadap mahasiswa yang terklasifikasikan mahasiswa berpotensi *Drop Out*. Menggunakan pendekatan c*lassification* dengan menggunakan *LibrarySciKit Learn*sertaa*lternative* algoritma KNN, naïve bayes, *decision tree*.

2.8 Fakta Pengajaran_Dosen



Gambar 3. Fakta pengajaran_dosen

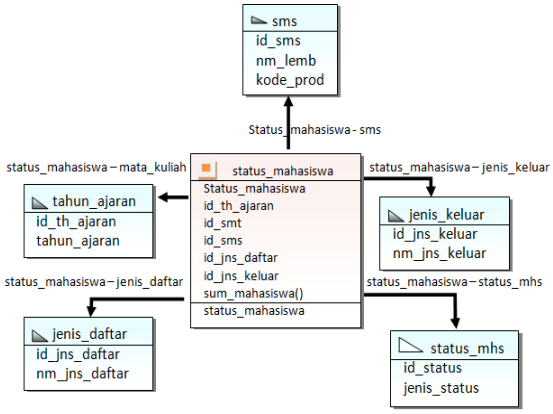
Fakta pengajaran_dosenseperti ditunjukkan pada Gambar 3 berfungsi untuk menampung data-data yang terkait dengan dosen dan pengajaran dosen pada setiap program studi dan matakuliah dengan detail atribut dijabarkan pada Tabel 4.

Tabel 4. Deskripsi atribut fakta pengajaran_dosen

Nama Atribut	Deksripsi
id_sdm	Kode dosen
Id_matkul	Kode matakuliah
id_kls	Kode kelas
id_sms	Kode program studi
Rencana_tatap_muka	Rencana tatap muka matakuliah
Real_tatap_muka	Realisasi tatap muka matakuliah
Tingkat_kehadiran_do sen	Tingkat kehadiran dosen (real/target)
Sks_mk	Jumlah SKS matakuliah
Rerata_nilai	Nilai rata-rata matakuliah pada setiap matakuliah dan dosen

Fakta tersebut dieksplorasi untuk penggunaan klusterisasi kinerja dosen. Dilakukan untuk mengelompokkan dosen menjadi beberapa kluster berdasarkan kinerja dosen. Variabel yang digunakan adalah tingkat kehadiran di perkuliahan dan rata-rata nilai per semester. Hasil klusterisasi bisa digunakan sebagai bahan masukan pimpinan untuk melaksanakan pelatihan bagi dosen, menggunakan pendekatan *clustering* dengan menggunakan *LibrarySciKit Learn* dengan *alternatif* algoritma *K-Means, Hierarchial Clustering,* DBSCAN. *Library* yang digunakan pandas, matplotlib, numpy, sklearn (algoritma *K-Means, Hierarchial Clustering* dan DBSCAN serta *Preprocessing*). Atribut-atribut yang digunakan : rata-rata nilai per semester, tingkat kehadiran dosen per mata kuliah. atribut data latih adalah id_dosen, id_matkul, id_kelas, tingkat_kehadiran_dosen, rerata_nilai. 2 atribut yang digunakan untuk permodelan yaitu : tingkat_kehadiran_dosen, rerata_nilai.

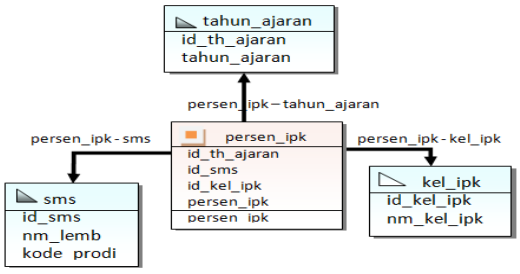
2.9 Fakta Status_Mahasiswa



Gambar 4. Fakta status_mahasiswa

Pada Gambar 4 adalah fakta status_mahasiswa berfungsi untuk menampung informasi status mahasiswa, jumlah mahasiswa berdasarkan status dalam program studi, tahun akademik dan semester(Rahutomo et al., 2019). Fakta tersebut dieksplorasi untuk penggunaan *dashboard* statistik mahasiswa yang menampilkan visualisasi kuantitas mahasiswa.

2.10 Fakta Persen_Ipk



Gambar 5. Fakta persen_ipk

Fakta persen_ipk yang ditunjukkan pada Gambar 5 berfungsi untuk menampilkan klasifikasi IPK Mahasiswa. IPK dengan nilai < 2,75 dan antara 2,75-3,5 serta IPK dengan nilai > 3,5 dalam tahunan dan program studi(Rahutomo et al., 2019). Fakta tersebut dieksplorasi untuk penggunaan dashboard statistik IPK mahasiswa yang menampilkan persentase IPK mahasiswa berdasarkan kategori

(kategori 1, 2, 3. Manfaat dari *view* ini adalah untuk memberikan visualisasi perkembangan akademis prodi setiap semester.

3. HASIL DAN PEMBAHASAN

Dari beberapa skema kegunaan *data warehouse* PDDIKTI, akan dilakukan pembahasan metode data mining pada dua skema, yaitu klasifikasi kelulusan tepat waktu dan prediksi minat mahasiswa pada program studi yang dituju. Tahapan-tahapan data mining menggunakan metode CRISP-DM (Larose and Larose, 2014) adalah sebagai berikut :

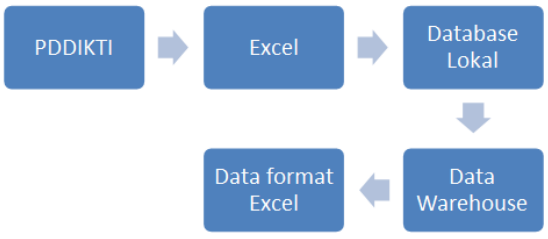
3.1 Pemahaman Terhadap Bisnis

Tujuan dari penambangan data PDDIKTI adalah untuk menemukan pola data akademik mahasiswa, penambangan data ini dilakukan untuk melakukan deteksi dini terhadap mahasiswa, sehingga dosen bisa memberikan masukan-masukan ketika mahasiswa tersebut telah diklasifikan sebagai mahasiswa yang lulus tidak tepat waktu serta mengkategorikan mahasiswa berdasarkan prestasi akademik dan latar belakang orang tua sehingga membantu pihak manajemen untuk memberikan beasiswa kepada mahasiswa yang berprestasi sebagai *reward*.

3.2 Pemahaman Terhadap Data

Dataset diperoleh dari data PDDIKTI pada Universitas XYZ untuk seluruh program studi jenjang S1.

1. Pengumpulan Data awal, dataset akademik yang digunakan adalah antara tahun akademik 2007-2018. Data diperoleh melalui aplikasi *webservice* PDDIKTI *Feeder,* dimanadata tersebut diperoleh dalam format *excel* kemudian ditansformasikan dalam bentuk *MySQL* dan selanjutnya dilakukan d*ata warehouse* sesuai dengan kebutuhan. Pada Gambar 6 dibawah iniadalah ilustrasi proses pengambilan data yang dilakukan(Rahutomo *et al.*, 2019).
2. Pengevaluasian data, evaluasi pada data dilakukan untuk memastikan data-data tersebut bisa diproses dan tidak ada data yang kosong (*null*). Untuk mengatasi data yang bernilai *null*, dilakukan pengisian data dengan nilai *average* dari IPK lainnya.
3.



Gambar 6. Proses Pengumpulan Data

3.3 Persiapan Data

Pada tahap ini dilakukan proses pemilihan dan pengolahan data yang diperlukan untuk tahap permodelan sehingga didapatkan hasil yang maksimal sesuai dengan target yang diinginkan. Data yang akan diolah adalah sebagai berikut :

1. Data keterangan ketepatan kelulusan mahasiswa,
2. Data akademik mahasiswa (IPK semester 1-4, rerata IPK 4 semester terakhir dan jumlah SKS yang sudah ditempuh),
3. Profil Mahasiswa (pekerjaan ayah, pendidikan ayah, pendidikan ibu, penghasilan ayah, penghasilan ibu, status ayah, status ibu).

Dari data tersebut dilakukan transformasi, sehingga diperoleh data dalam satu tabel memuat atribut nipd, ipk1, ipk2, ipk3, ipk4, jumlah SKS, keterangan lulus, rerata IPK 4 semester terakhir, pekerjaan ayah, pendidikan ayah, pendidikan ibu, penghasilan ayah, penghasilan ibu, status ayah, status ibu seperti yang ditunjukkan pada Tabel 5 berikut :

Tabel 5. Dataset akademik mahasiswa

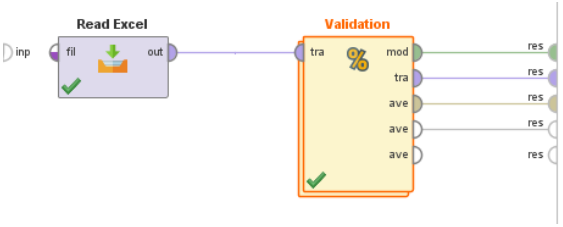
No	NIPD	IPK1	IPK2	IPK3	IPK4	Tot SKS	Pek. Ortu	Lulus
1	2070210009	3.71	3.61	3.57	3.51	88	PNS	Tepat
2	2070210011	3.43	3.26	3.45	3.5	89	PNS	Tepat
3	2070210012	3.62	3.31	3.36	3.36	89	Pedagang	Tepat
4	2070210014	2.43	2.26	2.01	2.12	77	PNS	Tidak
5	2070210017	3.48	3.33	3.44	3.45	87	Petani	Tepat
6	2070210022	2.62	2.56	2.71	1.36	81	Pedagang	Tidak
..	..	..	..	..	..	..	..	..
..	..	..	..	..	..	..	..	..
2600	2070211027	2.76	2.76	2.79	2.79	93	Pedagang	Tidak
2601	2070211055	2.76	2.79	2.79	2.79	93	Guru	Tidak

3.4 Permodelan

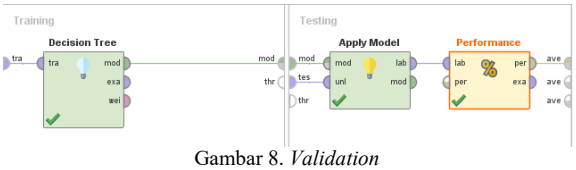
3.4.1 Klasifikasi Kelulusan Mahasiswa

Proses klasifikasi dilakukan berdasarkan kriteria tertentu sehingga proses pengklasifikasian dapat dilakukan. Pada data akademik mahasiswa dilakukan proses klasifikasi terhadap ketepatan kelulusan mahasiswa. Teknik tersebut dilakukan permodelan dengan menggunakan algoritma C4.5 dengan menggunakan Tool Rapidminner. Jumlah dataset yang digunakan adalah 1.610 *record*. Desain model yang digunakan adalah sebagai berikut :

1. *Retreiving Data,* operator ini digunakan untuk import data set yakni berupa file excel (.xlsx) seperti Gambar 7.
2. *Validation,* metode validasi yang digunakan adalah *Splitting Validation* seperti Gambar 8.
3. *Decision Tree,* Metode klasifikasi yang digunakan
4. *Apply Model*
5. *Performance,* Operator yang digunakan untuk mengukur performance akurasi dari model.



Gambar7. *Retreiving data* dan *Validation*



Gambar 8. *Validation*

3.4.2 Prediksi Jumlah Peminat Program Studi

Proses peramalan dilakukan berdasarkan data peminat (calon mahasiswa) yang mendaftar di universitas XYZ setiap tahun. Teknik tersebut dilakukan permodelan dengan menggunakan algoritma *Simple Moving Average* (*SMA*) dengan menggunakan Python. Jumlah dataset yang digunakan adalah 12 *record* untuk peramalan peminat mahasiswa di tingkat universitas dan 10 *record* untuk masing-masing program studi. Desain model yang digunakan ditunjukkan *screenshoot*Gambar 9seperti berikut :

1. *Read Data*, melakukan pembacaan data dari file .csv

	id_sms	tahun	peminat
0	ebad46f1-a8f6-4cb8-8133-ac6983aeed08	2011	75
1	ebad46f1-a8f6-4cb8-8133-ac6983aeed08	2012	61
2	ebad46f1-a8f6-4cb8-8133-ac6983aeed08	2013	77
3	ebad46f1-a8f6-4cb8-8133-ac6983aeed08	2014	100
4	ebad46f1-a8f6-4cb8-8133-ac6983aeed08	2015	100

Gambar 9. Import data dengan library pandas

2. *Preprocessing,* tahapan ini dilakukan untuk memastikan bahwa data di setiap *row* tidak ada yang kosong seperti ditunjukkan pada Gambar 10.

```
dt.isna().sum()

id_sms      0
tahun       0
peminat     0
dtype: int64
```

Gambar 10. Memastikan tidak ada data yang bernilai *Null*

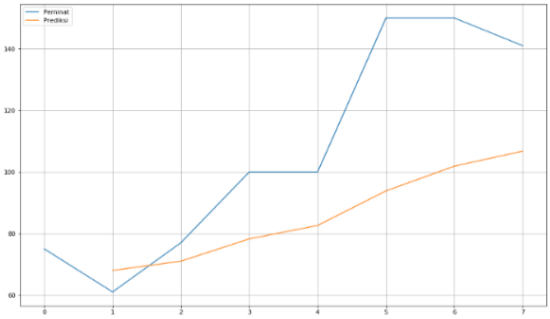
Dari proses tersebut diperoleh bahwa tidak ada data yang mengandung nilai *Null.*

3. Melakukan permodelan dengan menghitung nilai *Average* pada range tahun tertentu, yakni dengan range 3 tahun seperti ditunjukkan pada Gambar 11.

	peminat	next
0	75	NaN
1	61	NaN
2	77	71.000000
3	100	78.250000
4	100	82.600000
5	150	93.833333
6	150	101.857143
7	141	106.750000

Gambar 11. Permodelan dengan SMA

Perhitungan pada Average 3 *row* (*index* ke 0,1,2) atribut peminat disimpan pada *row* ke-3, atribut *next* begitu seterusnya sampai *row* terakhir, kolom *next* pada 2 baris yang awal bernilai *Null*. Pada Gambar 12 berikut adalah hasil pemodelan dengan *SMA*.



Gambar 12. Hasil Permodelan dengan *SMA*

3.5 Evaluasi

3.5.1 Klasifikasi Kelulusan Mahasiswa

Evaluasi dilakukan agar hasil pada tahap permodelan sesuai dengan target yang ingin dicapai pada tahap *business understanding* (Budiman *et al.*, 2012). Evaluasi dilakukan dengan *Confusion Matrix* seperti ditunjukkan pada Tabel 6. Evaluasi ini menghasilkan nilai *accuracy, precission* dan *recall*, untuk memastikan bahwa permodelan yang dilakukan telah mencapai *accuracy* terbaik. Nilai *accuracy* merupakan persentase jumlah record data yang diklasifikasikan benar oleh algoritma dan pada data bernilai benar (Han, Pei and Kamber, 2011).

Tabel 6. Model *Confusion Matrix*

Correct Classification	***Classification as***	
	+	-
+	*True Positive (TP)*	*False Negative (FN)*
-	*False Positive (FP)*	*True Negative (TN)*

Untuk memperoleh nilai tersebut dilakukan dengan :

1. *Accuracy :* Jumlah yang diklasifikasikan benar / total sampel data yang diuji coba
2. *Precission :* Jumlah data berkategori positif dikategorikan benar / total data yang diklasifikasikan benar ($\frac{TP}{TP+FP}$)

3. *Recall :* Jumlah data diklasifikasi positif / total data testing diklasifikasikan positif ($\frac{TP}{TP+FN}$)

Pada uji coba ini, diberikan data *training* untuk membentuk model dan data *testing* untuk menguji model yang telah terbentuk seperti ditunjukkan pada Tabel 7 berikut :

Tabel 7.HasilEvaluasi hasil klasifikasi kelulusan

	True Tepat	*True* Tidak Tepat
Prediksi Tepat	549	66
Prediksi Tidak Tepat	39	126

Dari tabel tersebut, diperoleh nilai *accuracy* 86,54% nilai *precission* 93,37% dan nilai *recall* 89,27%.

3.5.2 Prediksi Jumlah Peminat Program Studi

Evaluasi terhadap prediksi peminat (calon mahasiswa yang mendaftar) di program studi perguruan tinggi adalah dengan menghitung nilai error menggunakan *MAPE* (*Mean Average Percentage Error*) seperti pada Gambar 13.

```
[ ]  y_true = [3, -0.5, 2, 7]
     y_pred = [2.5, -0.3, 2, 8]
     expected = [0.0, 0.5, 0.0, 0.5, 0.0]
     predictions = [0.2, 0.4, 0.1, 0.6, 0.2]

     mape=mean_absolute_percentage_error(dt_T['peminat'], dt_T['next'])

[ ]  print('MAPE: ', mape)

     MAPE:  21.749725758542837
```

Gambar 13. Evaluasi dengan menghitung *MAPE*

Dari evaluasi yang dilakukan diperoleh nilai tingkat error sebesar 21,75% seperti ditunjukkan apda Tabel 8 berikut :

Tabel 8.Perhitungan *MAPE*

Tahun Akademik	Jumlah Peminat	Hasil Prediksi	Selisih	*MAPE*
2011	75			
2012	61	68	-7	11.47541
2013	77	71	6	7.792208
2014	100	78.25	21.75	21.75
2015	100	82.6	17.4	17.4
2016	150	93.83	56.16	37.44444
2017	150	101.85	48.14	32.09524
2018	141	106.75	34.25	24.29078
Rata-rata				21.74973

3.6 Penyebaran

Membuat laporan proses data mining, yakni pengetahuan baru yang diperoleh berupa pola proses data mining dan ditampilkan dalam bentuk grafik. *Deployment* ini dilakukan dengan membuat purwarupa dari hasil permodelan yang menghasilkan nilai *accuracy* terbaik, dengan menggunakan python ataupun dengan PHP yang melakukan pengambilan data secara otomatis dari *database*.

4. KESIMPULAN

Pengujian klasifikasi kelulusan mahasiswa dengan tools rapidminer menggunakan algoritma C4.5 pada data 1.610 *record,* diperoleh nilai *accuracy* 86,54%, *precission* 93,37 dan nilai *recall* 89,27%. Hasil data mining ini digunakan untuk melakukan klasifikasi tingkat kelulusan mahasiswa, sehingga dosen bisa memberikan masukan pada mahasiswa berdasarkan IPK semester 1-4 ketika mahasiswa tersebut dinyatakan lulus tidak tepat waktu, serta menjadi bahan masukan kepada pimpinan dalam memberikan *reward* (beasiswa) kepada mahasiswa. Pengujian prediksi jumlah peminat program studi diperoleh nilai error rata-rata (*MAPE*) sebesar 21,75%. Hasil peramalan menunjukkan dibawah dari data aktual, hal ini menunjukkan bahwa peminat mahasiswa melebihi dari peramalan. Data *training* menjadi acuan dalam menentukan pola, sehingga sangat diperlukan keakuratan data dan adanya *split* antara data *training* dan data *testing*. Hal itu bisa mempengaruhi terhadap nilai *accuracy, recall* dan *precission,* selain itu kuantitas data *training* juga mempengaruhi pola yang dihasilkan.

DAFTAR PUSTAKA

AGGARWAL, C. C. , 2015. Data mining: the textbook. Springer.

ALVERINA, D., CHRISMANTO, A. R. and SANTOSA, R. G. , 2018. Perbandingan Algoritma C4. 5 dan CART dalam Memprediksi Kategori Indeks Prestasi Mahasiswa, *Jurnal Teknologi dan Sistem Komputer. Departemen Teknik Sistem Komputer, Fakultas Teknik, Universitas Diponegoro, 6(2), pp. 76–83.*

BUDIMAN, I. et al. , 2012. Data Clustering menggunakan metodologi Crisp-DM untuk pengenalan pola proporsi pelaksanaan tridharma. Universitas Diponegoro.

DEPDIKNAS, D. D. , 2012. tahun 2012 tentang Pendidikan Tinggi, Jakarta: Depdiknas, Ditjen Dikdasmen.

HAN, J., PEI, J. and KAMBER, M. , 2011. Data mining: concepts and techniques. Elsevier.

HEIZER, J. and RENDER, B. , 2009. Manajemen operasi, Jakarta: Salemba Empat.

JAMES, G. et al. , 2013. An introduction to statistical learning. Springer.

KEMENTRIAN RISET, T. dan P. T. R. I. , 2017. Sistem Informasi Manajemen Akademik Modul Pangkalan Data Perguruan Tinggi. Jakarta: Direktorat Jenderal Pendidikan Tinggi Republik Indonesia.

LAROSE, D. T. and LAROSE, C. D. , 2014. Discovering knowledge in data: an introduction to data mining. John Wiley & Sons.

RAHUTOMO, F. et al. , 2019. Desain Skema Data Warehouse PDDIKTI sebagai Pendukung

Keputusan Perguruan Tinggi, INOVTEK Polbeng-Seri Informatika, 4, pp. 90–100.

ROHMAN, A. , 2015. Model Algoritma K-Nearest Neighbor (K-NN) Untuk Prediksi Kelulusan Mahasiswa, Neo Teknika, 1(1).

TANNADY, H. and ANDREW, F. , 2013. Analisis Perbandingan Metode Regresi Linier dan Exponential Smothing Dalam Parameter Tingkat Error, Teknik dan Ilmu Komputer, 2(7).

TURBAN, E., SHARDA, R. and DELEN, D. , 2014. Decision support and business intelligence systems. Pearson.

YUNIANITA, S. and others , 2018. SISTEM KLASIFIKASI KELULUSAN MAHASISWA DENGAN ALGORITMA C4. 5. Universitas Islam Indonesia.

ZAKI, M. J., MEIRA JR, W. and MEIRA, W. , 2014. Data mining and analysis: fundamental concepts and algorithms. Cambridge University Press.

BENGNGA, A, and REZQIWATI ISHAK, 2018. Prediksi Jumlah Mahasiswa Registrasi Per Semester Menggunakan Linier Regresi Pada Universitas Ichsan Gorontalo. ILKOM Jurnal Ilmiah 10.2 : 136-143.

THIRAFI, M. F. S., & RAHUTOMO,2018. Implementation of Naïve Bayes Classifier Algorithm to Categorize Indonesian Song Lyrics Based on Age. In 2018 International Conference on Sustainable Information Engineering and Technology (SIET). (pp. 106-109). IEEE.