Walisongo Journal of Information Technology, Vol. 2 No. 1 (2020): 1 - 12
DOI : http://dx.doi.org/10.21580/wjit.2020.2.1.5204

Identifikasi Potensi Keberhasilan Studi Menggunakan *Naïve Bayes Classifier*

Wenty Dwi Yuniarti [1], Achmad Nur Faiz [2], Bagus Setiawan [3]
[1,2,3] Universitas Islam Negeri Walisongo
wenty@walisongo.ac.id, nurfaiz_1808096030@student.walisongo.ac.id,
bagus_1808096008@student.walisongo.ac.id

Abstract

This study aims to predict the success of the study with the naïve bayes classifier classification method. Input variables that are estimated to influence study success are Entrance Pathways (1), City of Origin (2), Province Origin (3), Parents 'Earnings (4), Parents' Work (5), Cumulative Performance Index (6) and Student Status History (7). Knowledge of the potential for success of the study was obtained from a variation of four target variables (class) namely the First and Second Year GPA, Current Student Status, Non-Prodi Subject GPA and Prodi Subject GPA. The process begins with preprocessing data and 5,934 data is obtained cleanly. The data is divided into 80% training, 20% testing, with Correctly Classified Instances 97.53%. Knowledge gathering with the naïve bayes classifier obtained an accuracy of 99.41% for predictive input variables 1,2,3,4,5,6,7 with a target of success in the first and second GPA, 96.96%, for the current Student Status target, 95.87% for Prodi Subject GPA target, and 97.89% for Non-Prodi Subject GPA target. The use of the naïve bayes classifier method in the classification of the potential success of this study provides an accuracy of 95.8% to 99.41% for 4 different targets. It is necessary to strengthen the student recruitment process, and to consider the economic factors of parents to contribute to the continuity of the study process.

Keyword : *the success of studies*, *data mining*, *naïve bayes*, *classification*

Abstrak

Penelitian ini bertujuan melakukan prediksi keberhasilan studi dengan metode klasifikasi *naïve bayes classifier*. Varibel input yang diperkirakan mempengaruhi keberhasilan studi adalah Jalur Masuk (1), Asal Kota (2), Asal Provinsi (3), Penghasilan Orang Tua (4), Pekerjaan Orang Tua (5), Indeks Prestasi Kumulatif (6) serta Riwayat Status Mahasiswa (7). Pengetahuan potensi keberhasilan studi diperoleh dari variasi empat variabel target (*class*) yaitu IPK Tahun Pertama dan Kedua, Status Mahasiswa Terkini, IPK Mata Kuliah (Makul) Non-Keprodian serta IPK Makul Keprodian. Proses diawali *data preprocessing* dan diperoleh 5.934 data bersih. Data dibagi 80% *training*, 20% *testing*, dengan *Correctly Classified Instances* 97,53%. Penggalian pengetahuan dengan *naïve bayes classifier* memperoleh akurasi 99,41% untuk prediksi variabel input 1,2,3,4,5,6,7 dengan target keberhasilan IPK Tahun pertama dan kedua, 96.96 %, untuk target Status Mahasiswa Terkini, 95.87% untuk target IPK Makul Keprodian, dan 97.89 % untuk target IPK Makul Non-Keprodian. Penggunaan metode *naïve bayes classifier* dalam klasifikasi potensi keberhasilan studi ini memberikan akurasi 95.8% sampai dengan 99.41% untuk 4 target berbeda. Bagi perguruan tinggi, perlu penguatan pada proses perekrutan mahasiswa, serta perlu diperhatikan bahwa faktor ekonomi orang tua memberikan andil bagi kelangsungan proses studi.

Kata Kunci : keberhasilan studi, data mining, *naïve bayes classifier*, klasifikasi

ISSN 2715-0143 (online) ISSN 2714-9048 (print)
http://journal.walisongo.ac.id/index.php/jit/index

1. PENDAHULUAN

Undang-Undang Nomor 12 Tahun 2012 menyebutkan bahwa pendidikan tinggi adalah bagian dari sistem pendidikan nasional yang memiliki peran strategis dalam rangka mencerdaskan kehidupan bangsa[1]. Mahasiswa adalah peserta didik pada jenjang pendidikan tinggi. Perguruan tinggi berkewajiban mengembangkan potensi mahasiswa dan sebaliknya, mahasiswa harus mengembangkan potensinya secara aktif melalui pembelajaran, pencarian kebenaran ilmiah dan penguasaan, pengembangan serta pengamalan ilmu[2]. Oleh karena itu, perguruan tinggi berkewajiban mengembangkan potensi mahasiswa untuk meraih keberhasilan studi.

Dalam Kamus Bahasa Indonesia, potensi[3] adalah kemampuan yang mempunyai kemungkinan untuk dikembangkan. Berkaitan dengan keberhasilan studi di perguruan tinggi, kemampuan yang mungkin dikembangkan adalah kemampuan dasar akademik[4]. Diperlukan suatu batas minimal kemampuan dasar akademik sehingga seorang mahasiswa diyakini akan mampu menyelesaikan studi di perguruan tinggi. Secara umum ukuran keberhasilan studi adalah ketika mahasiswa mampu menyelesaikan program pendidikan sesuai kecepatan belajar masing-masing dengan tidak melebihi ketentuan batas waktu yang ditetapkan perguruan tinggi[5].

Dalam rangka memaksimalkan tugas perguruan tinggi dalam mengembangkan potensi untuk keberhasilan studi, identifikasi dini atas potensi mahasiswa perlu dilakukan guna mengetahui kemungkinan keberhasilan mahasiswa dalam penyelesaian studi dan mengantisipasi kemungkinan ketidakberhasilan. Penting untuk memprediksi, apakah seorang mahasiswa yang masuk perguruan tinggi akan dapat keluar atau menyelesaikan studi.

Dalam kenyataannya, pengembangan potensi bukan hal yang mudah dilakukan. Bahkan permasalahan yang berpotensi mempengaruhi keberhasilan studi terjadi, diantaranya tingginya gap jumlah mahasiswa diumumkan diterima dengan jumlah mahasiswa yang registrasi; angka *drop out* yang tinggi (M_{DO}>6%), serta belum idealnya rerata masa studi mahasiswa.

Pengembangan strategi untuk optimalisasi keberhasilan studi mahasiswa dapat dilakukan melalui evaluasi diri (*self evaluation*) dengan melihat secara komprehensif kondisi internal berbasis pada fakta atau data yang dimiliki perguruan tinggi. Data tersebut dapat berupa data akademik seperti jalur masuk, asal sekolah, latar

[1] Republik Indonesia. 2012. Undang-undang Republik Indonesia Nomor 12 Tahun 2012 tentang Pendidikan Tinggi

[2] Republik Indonesia. 2015. Peraturan Menteri Riset, Teknologi dan Pendidikan Tinggi Republik Indonesia Nomor 44 Tahun 2015 tentang Standar Nasional Pendidikan Tinggi

[3] https://kbbi.kemdikbud.go.id/

[4] UI. 2004. Keputusan Rektor Universitas Indonesia tentang Evaluasi Keberhasilan Studi Mahasiswa Universitas Indonesia

[5] UI. 2013. Peraturan Rektor Universitas Indonesia tentang Penyelenggaraan Program Sarjana di Universitas Indonesia.

jurusan, IPK, status akademik per semester, hingga data non akademik seperti pekerjaan dan keadaan ekonomi orang tua.

Identifikasi potensi keberhasilan studi dimaksudkan untuk melakukan deteksi dini potensi ketidakberhasilan, sehingga bisa dicegah atau diminimalkan dengan upaya-upaya akademik. Identifikasi juga memungkinkan pemberian rekomendasi untuk perbaikan atau pengembangan proses akademik.

Oleh karena itu, berkaitan dengan identifikasi potensi keberhasilan studi, perlu diketahui :

a. Parameter atau atribut apa saja yang mempengaruhi potensi keberhasilan studi mahasiswa ?
b. Bagaimana cara mengidentifikasi potensi keberhasilan studi mahasiswa berdasar parameter tersebut dengan teknik data mining metode *Naïve Bayes Classifier* ?

2. METODE

2.1 Model Pendekatan

Penelitian ini menggunakan pendekatan CRISP-DM (*Cross-Industry Standard Process Model for Data Mining*) dengan tahapan[6]:

A. Pemahaman atas permasalahan (*Business Understanding*)

Pada tahap ini dibutuhkan pengetahuan dari objek bisnis, bagaimana membangun atau mendapatkan data serta bagaimana membangun model terbaik.

B. Pemahaman atas data (*Data Understanding*)

Tahap ini fokus pada pemeriksaan serta identifikasi masalah dalam data seperti nilai hilang dan *outlier*[7], sehingga dapat diperbaiki dalam tahap *Data Preparation.*

C. Penyiapan Data (*Data Preparation*)

Tahap penyiapan data berupa pembuatan variabel turunan (*derived*), filtering dan konversi data guna memastikan data bersih serta tepat untuk algoritma yang digunakan. Pada tahap ini dilakukan pembagian data menjadi dua yaitu *data training* dan *data testing*[8].

D. Pemodelan (*Modelling*)

Tahap pemodelan berisi pembuatan model dengan algoritma yang ditetapkan.

E. Evaluasi (*Evaluation*)

Tahap evaluasi digunakan untuk menilai kesesuaian model dengan tujuan yang diharapkan. Evaluasi dilakukan dengan melihat nilai akurasi dan tabel *confusion matrix*.

2.2 Data dan Waktu Penelitian

Data penelitian merupakan data primer yang diperoleh dari Sistem Informasi Akademik dengan subjek data mahasiswa angkatan 2015, 2016, 2017.

Penelitian berlangsung pada bulan April s.d Agustus 2019.

6 North, M. 2016. Data Mining For The Masses With Implementations in Rapidminer and R. Second Edition.

7 Mann, P.S. 2010. Introductory Statistics. John Wiley & Sons, Inc.

8 Kotu, V. and Bale Deshpande. 2015. *Predictive Analytics and Data Mining, Concepts and Practise with RapidMiner*. Morgan Kaufmann, Elseveir Inc.

3. KERANGKA TEORI

3.1 Potensi Keberhasilan Studi

Pada perguruan tinggi, kegiatan akademik yang dilakukan mahasiswa guna meraih keberhasilan studi diatur dalam pedoman akademik. Keberhasilan studi diartikan ketika mahasiswa dapat menyelesaikan program pendidikan yang ditempuhnya, berdasar ukuran tertentu yang ditetapkan yaitu capaian kualifikasi prestasi atau indeks prestasi, serta ketepatan masa studi[9]. Oleh karena itu, harus diperhatikan, pertama, evaluasi status mahasiswa tiap semester, dan kedua, beban dan masa studi mahasiswa. Tidak terpenuhinya salah satu unsur, menyebabkan mahasiswa tidak bisa menyelesaikan studi. Sebagai contoh, jika mahasiswa tidak menyelesaikan beban 144 sks dalam 12 semester yang dipersyaratkan, maka mahasiswa tidak dapat lulus atau mengalami *droup out*.

Ketercapaian beban studi dan masa studi sebagai tolak ukur keberhasilan studi sangat ditentukan oleh penilaian hasil belajar mahasiswa. Hasil capaian pembelajaran di tiap semester dinyatakan dengan indeks prestasi semester (IPs) sedangkan hasil capaian pembelajaran pada akhir studi dinyatakan dengan indeks prestasi kumulatif (IPK). Berkaitan dengan keberhasilan studi, berdasar ketetapan capaian penilaian hasil belajar, seorang mahasiswa dinyatakan lulus apabila 1) lulus seluruh matakuliah, 2) memperoleh Indeks Prestasi (IP) minimal 2.0, 3) jumlah mata kuliah dengan bobot nilai dibawah 2.0 tidak lebih dari 25 persen beban studi wajib.

3.2 Data Mining dengan *Naïve Bayes Classifier*

Saat ini, dalam proses bisnis institusi berbasis teknologi informasi dalam jaringan, termasuk pada perguruan tinggi, sering ditemukan akumulasi data dalam jumlah besar[10]. Teknik analisis konvensional hanya memberikan gambaran umum secara deskriptif tanpa memberi banyak pengetahuan. Dibutuhkan paradigma yang mampu mengelola data dalam jumlah besar, mengamati keterhubungan ratusan variabel, serta merumuskan suatu algoritma *learning* atas data tersebut dalam rangka menemukan pengetahuan.

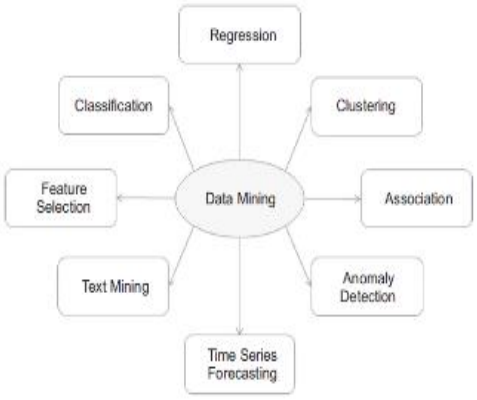
Data mining adalah paradigma pengelolaan data besar dengan banyak variabel yang mampu menyimpulkan pengetahuan atau pola atas data yang dimiliki. Permasalahan dalam data mining dikategorikan menjadi *supervised* dan *unsupervised*. Teknik *supervised* memprediksi keluaran data berbasis pada variabel input yang dimiliki. Suatu model dikembangkan dari suatu data training dimana nilai dari input dan output sebelumnya sudah diketahui. Selanjutnya, model mengeneralisasikan hubungan antara variabel input dan output dan menggunakannya untuk memprediksi

[9] UIN Walisongo Semarang. 2015. Pedoman Akademik UIN Walisongo Tahun 2015.

[10] Yuniarti, W.D. 2019. Dasar-dasar Pemrograman dengan Python. Deepublish Publisher

data dalam keadaan hanya diketahui inputnya saja. *Unsupervised* data mining tidak memerlukan data output untuk diprediksi. Tujuan *unsupervised* lebih pada mengenali pola dalam data berdasarkan hubungan antar record data.

Permasalahan data mining dapat diklasifikasikan menjadi klasifikasi, regresi, asosiasi, *deteksi anomaly time series data*, *text mining* dsb[11].



Gambar 1. Klasifikasi Data Mining

Beberapa metode klasifikasi adalah *Decisioon Tree*, *Rule Induction*, *Naïve Bayes* dan *Support Vector Machine*. Naïve Bayes Classifier adalah teknik prediksi berbasis probabilistik sederhana yang berdasar pada penerapan teorema Bayes. Metode ini memperhatikan asumsi independensi yang kuat (*naif*) dimana model yang digunakan adalah model fitur independen. Independensi yang kuat pada fitur adalah sebuah fitur pada sebuah data tidak ada kaitannya dengan ada atau tidak adanya fitur yang lain dalam data yang sama.

Ide dasar aturan Bayes, hasil dari hipotesis atau peristiwa (H) dapat diperkirakan berdasarkan pada beberapa *evidence* (E) yang diamati. Hal penting dalam Bayes adalah sebuah probabilitas awal/priori H atau $P(H)$ adalah probabilitas dari suatu hipotesis sebelum bukti diamati. Sebuah probabilitas posterior H atau $P(H|E)$ adalah probabilitas dari suatu hipotesis setelah bukti-bukti yang diamati ada.

$$P(H \mid E) = \frac{P(E \mid H) \times P(H)}{P(E)}$$

$P(H|E)$ adalah probabilitas posterior bersyarat (*Conditional Probability*) suatu hipotesis H terjadi jika diberikan *evidence*/bukti E terjadi.
$P(E|H)$ adalah probabilitas sebuah *evidence* E terjadi akan mempengaruhi hipotesis H.
$P(H)$ adalah probabilitas awal (priori) hipotesis H terjadi tanpa memandang *evidence* apapun.
$P(E)$ adalah probabilitas awal (priori) evidence E terjadi tanpa memandang hipotesis/*evidence* yang lain.

3.3 WEKA Machine Learning

Weka adalah aplikasi data mining *open source* berbasis Java dengan koleksi algoritma *machine learning* yang dapat digunakan untuk melakukan generalisasi / formulasi

[11] Kantardzic, M. 2003. Data Mining: Concepts, Models, Methods and Algorithms. John Wiley & Sons, Inc.

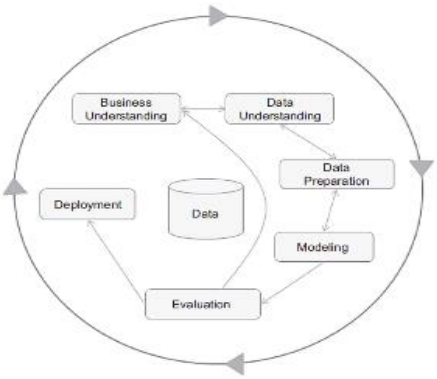
dari sekumpulan data sampling[12]. Pengelolaan data dengan teknik mining menggunakan WEKA dilakukan dengan tahap[13]:

a. Import data dalam format yang diakomodasi WEKA (.csv, .arff dsb).
b. Pemberian metode data mining. Gunakan menu Classify, sehingga tampil ragam metode dan turunan metode seperti *bayes*, *functions*, *lazy*, *meta*, *misc*, *rules* dan *trees.*
c. Setelah metode, selanjutnya ditetapkan sejumah parameter *Test Options* dan split data meliputi *Uses Training Set*, *Supplied Test Set*, *Cross Validation Folds* dan *Percentage Split*.
d. Pengaturan *Classifier Evaluation Options* yang memuat beberapa hal terkait *output model*, *output per-class stats*, *output sentropy evaluation measures*, *output confusion matrix*, *store predictions for visualization*, *error plot point size proportional to margin*, *output predictions*, *cost - sensitive evaluation*, *preserve order for % split* dan *output sourcecode*.
e. Lakukan proses hingga diperoleh report/*summary* dengan atribut *Correctly Classified Instances*, *Incorrectly Classified Instances*, *Kappa Statistics*, *Mean Absolute Error*, *Root Mean Squared Error*, *Relative Absolute Error*, *Root Relative Squared Error* dan *Total Number of Instances*.

Gambar 2. Antar muka WEKA

4. PEMBAHASAN

Penelitian ini dilakukan dengan pendekatan CRISP-DM dengan tahapan seperti gambar berikut:



Gambar 3. CRISP-DM *Framework*

4.1 Pemahaman atas pemasalahan

Sejumlah permasalahan akademik berkaitan dengan standar keberhasilan studi mahasiswa diantaranya, mahasiswa terlambat lulus, mangkir atau tidak mampu menyelesaikan studi (*droup out*).

[12] UoW Machine Learning Group. 2016. WEKA The Workbench for Machine Learning. Diakses pada Juli 2016 melalui laman cs.waikato.ac.nz: https://www.cs.waikato.ac.nz/ml/weka/

[13] Witten, I.H., Eibe Frank, and Mark A. Hall. 2017. *Data Mining Practical Machine Learning Tools and Technique.* 3nd Edition. Morgan Koufmann Publisher, Elsevier, Inc

Tabel 1. Data Mahasiswa *Droup Out*

Tahun/ Angkatan	Jumlah	Thn ke-1	Thn ke-2	Thn ke-3	Thn ke-4	M_{DO}
TS-3	2121	0	0	0	55	2.6%
TS-2	2778	0	0	236	-	8.5%
TS-1	3383	0	224	-	-	6.6%
TS	3777	0	-	-	-	0%

Data perguruan tinggi yang berkaitan dengan keadaan mahasiswa tersimpan dalam sistem informasi akademik dengan sejumlah atribut sebagaimana Tabel 2.

Tabel 2. Ragam Data Diri Mahasiswa

Atribut (Variabel Input)	Atribut (Variabel Input)
Nomor Pendaftaran	Alamat
Nama	Kota_alamat
NIM	Provinsi
Jalur_Masuk	Nomor_telepon
Fakultas	Nama_Ayah
Prodi	Pekerjaan_Ayah
Jenis_kelamin	Penghasilan_ayah
Tempat_lahir	Status_Ayah
Tanggal_lahir	Alamat_ayah
Provinsi_Ayah	Status_Smt4
Nama_Ibu	Status_Smt5
Pekerjaan_Ibu	Status_Smt6
Penghasilan_Ibu	Status_Smt7
Status_Ibu	Status_Smt8
Alamat_Ibu	IPs_1
Provinsi_Ibu	IPs_2
Asal_sekolah	IPs_3
Alamat_Sekolah	IPs_4
Kota_Sekolah	IPs_5
Status_Smt1	IPs_6
Status_Smt2	IPs_7
Status_Smt3	IPs_8

4.2 Identifikasi Data

Data primer yang diperoleh berjumlah 9.963. Penyiapan data diawali dengan penetapan atribut penting untuk proses klasifikasi. Menurut Syafrudin, keberhasilan studi mahasiswa dipengaruhi faktor-faktor yang relatif sulit diukur seperti latar belakang pendidikan sebelumnya, latar belakang keluarga (orang tua), lingkungan belajar dan faktor individu mahasiswa[14]. Mega Khoirunnisak menggunakan rumusan yang menyebutkan bahwa faktor-faktor yang mempengaruhi berhenti studi (*drop out*) mahasiswa adalah intelegensia, penghasilan orang tua, Indeks Prestasi Kumulatif (IPK) dan asal daerah. Faktor-faktor lain yang diduga mempengaruhi mahasiswa *drop out* adalah usia masuk, fakultas, status sekolah asal serta nilai-nilai mata kuliah tertentu[15].

Berdasar ragam atribut atau variabel input yang dirumuskan dalam penelitian sebelumnya serta memperhatikan ketersediaan ragam data maka varibel yang diduga mempengaruhi keberhasilan studi dan menjadi atribut awal adalah

a. Jalur masuk sebagai mahasiswa baru
b. Asal Kota
c. Asal Provinsi
d. Penghasilan Orang Tua
e. Pekerjaan Orang Tua
f. Pernah Mangkir / Cuti
g. IPK

4.3 Penyiapan Data

Penyiapan data dilakukan untuk memperoleh data bersih melalui 1) filtering yaitu meniadakan data tanpa nilai, tidak standar atau tidak sempurna dan 2) konversi data.

Penfilteran (*Filtering*)

[14] Syafrudin. 2006. Analisis Faktor-faktor Yang Mempengaruhi Keberhasilan Studi Mahasiswa Program Sarjana Ekstensi Manajemen Agribisnis Institut Pertanian Bogor

[15] Khoirunnisak, M. dan Nur Iriawan. 2013. Pemodelan Faktor-Faktor Yang Mempengaruhi Mahasiswa Berhenti Studi (*Drop Out*) di Institut Teknologi Sepuluh Nopember menggunakan Analisis Bayesian *Mixture Survival*.

Filtering adalah meniadakan data tak bernilai dengan langkah[16] sebagai berikut:

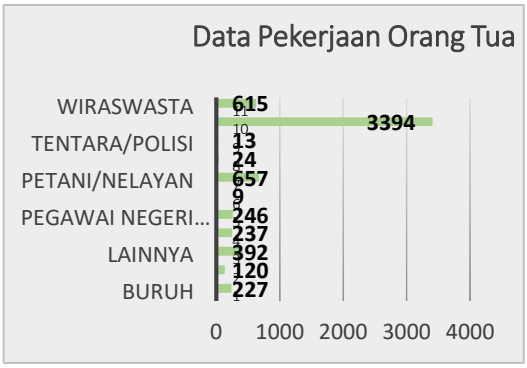
a. Memilih data yang memiliki Indek Prestasi lengkap sejak semester 1.
b. Menghilangkan data yang memiliki status semester DO karena tidak memiliki indek prestasi.
c. Menghilangkan data siswa berstatus pindah.

Dari filtering diperoleh data bersih 5.934 mahasiswa.

Konversi

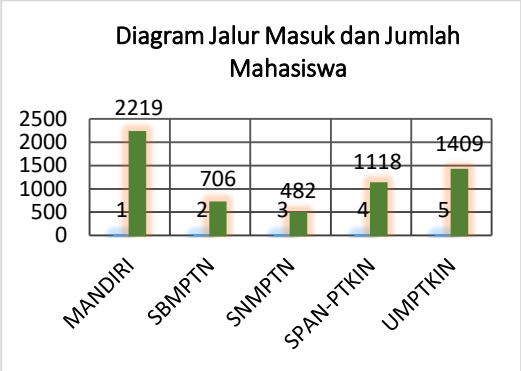
Data mentah dikonversi[17] agar dapat diolah lanjut dengan pendekatan data mining.

a. Pekerjaan, dikonversi dalam 11 kode, yaitu 1-Buruh, 2-Guru dst.



Gambar 4. Diagram Data Pekerjaan

b. Provinsi, dikonversi dalam 33 kode, yaitu 1: Luar Negeri, 2: Aceh dst.
c. Kota, dikonversi dalam 205 kode, yaitu 1: Thailand, 2: Kab Aceh Besar dst. Sebagai catatan, dalam *preprocessing data*, jika ditemukan data kota/kab kosong, diisi dengan kategori Tidak Ada Kota (205).
d. Jalur, dikonversi dalam 5 kode, yaitu 1: Mandiri, 2: SBMPTN dst, dengan diagram data sebagai berikut:



Gambar 5. Diagram Jalur Masuk

e. Status Mangkir-Cuti, dikonversi dalam 4 kode, misal: 1-Pernah Mangkir, 0-Tidak pernah mangkir, 1-Pernah Cuti, 0-Tidak pernah cuti.

Tabel 3. Deskripsi Data Mangkir/Cuti

Status	Kategori	Jumlah
Tidak Pernah Mangkir / Cuti	0	5436
Pernah Mangkir / Cuti	1	498

f. Penghasilan
Data penghasilan berupa data numerik dan tidak diperlukan konversi.

Setelah melakukan identifikasi atribut dan mendapatkan deskripsi data,

16 S. Karthika* and N. Sairam. 2015. A Naïve Bayesian Classifier for Educational Qualification. Indian Journal of Science and Technology, Vol 8(16), July 2015.

17 Redjeki, S. 2013. Identifikasi Penyakit dengan Gejala Awal Demam menggunakan K-Nearest Neighnor (KNN). Jurnal Buana Informatika Vol. 4 No. 1.

selanjutnya ditetapkan atribut kelas. Atribut kelas, atau selanjutnya disebut *Class*, merupakan atribut yang menjadi target dari prediksi. Artinya apakah atribut-atribut yang sudah diidentifikasi sebagai potensi keberhasilan studi memberikan potensi atau tidak memberikan potensi pada keadaan *class*.

Ada 4 variasi *class* yang akan diujicobakan:

a. Dari atribut yang diidentifikasi, memprediksikan atribut **IPs_24** (IP semester 2 dan 4) sebagai atribut target. Atribut *class* berupa data nominal dengan kategori pertama, POTENSI jika IPs_24 < 2.00, dan kategori kedua, TIDAK POTENSI jika IPs_24 ≥ 2.00.
b. Dari atribut yang diidentifikasi, memprediksikan atribut **IPK_NonProdi** (IP mata kuliah Non-Keprodian) sebagai atribut target. Atribut *class* berupa data nominal dengan kategori pertama, POTENSI jika IPK_NonProdi < 2.00 dan kategori kedua, TIDAK POTENSI jika IPK_NonProdi ≥ 2.00.
c. Dari atribut yang diidentifikasi, memprediksikan atribut **IPK_Prodi** (IP mata kuliah Keprodian) sebagai atribut target. Atribut *class* berupa data nominal dengan kategori pertama, POTENSI jika IPK_Prodi < 2.00 dan kategori kedua, TIDAK POTENSI jika IPK_Prodi ≥ 2.00.
d. Dari atribut yang diidentifikasi, memprediksikan atribut **Status** (DO/*drop out*/putus studi) sebagai atribut target. Atribut *class* berupa data nominal dengan kategori pertama, POTENSI jika status terkini adalah DO / *drop out* / putus studi dan kategori kedua, TIDAK POTENSI jika status mahasiswa per semester terkini adalah aktif.

4.4 Pemodelan (Modelling)

Teknik mining yang digunakan dalam pemodelan data ini adalah *Naïve Bayes Classifier* dengan menggunakan *machine learning* dalam tools Weka. Proses diawali dengan diperolehnya data bersih 5.934 dengan komposisi atribut independen meliputi: 1) Jalur, 2) Kota, 3) Provinsi, 4) Penghasilan, 5) Pekerjaan, 6) Mangkir-Cuti, 7) IP_Kumulatif.

Tabel 4. Contoh record dengan variabel Independen dan Target IP Smt 2/4

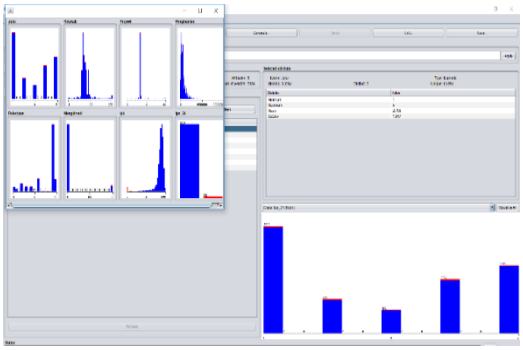
JALUR	KOTA	PROV	PENGHASILAN	PEKERJAAN	Mangkir Cuti	IPK	IP SMT 24
5	198	32	2.043.774	4	0	3.63	2
5	69	12	2.043.774	11	0	3.93	2
1	98	13	750.000	10	0	3.95	2
5	85	12	2.043.774	5	0	3.93	2
4	38	11	700.000	3	0	3.88	2
5	77	12	2.043.774	4	0	3.88	2
3	67	12	1.500.000	7	0	3.91	2
1	72	12	1.000.000	7	0	3.94	2

Modeling dilakukan untuk variasi 4 target: IP Semester 2/4, IP Makul Keprodian, IP Makul Non-Keprodian serta Status Mahasiswa Terkini. Selanjutnya dilakukan pemodelan dengan machine learning dengan *naïve bayes classifier* untuk setiap target. Uji dilakukan dengan persentase data training dan testing sebesar 80%:20%.

Yang pertama, modeling untuk target IP_24 dengan proporsi 80%:20%

1) Pergunakan *machine learning* Weka
2) Buka data dengan target IP_24
3) Pilih model klasifikasi yaitu *naïve bayes* dan tetapkan persentase

training dan testing sebesar 80%:20%



Gambar 6. Deskripsi Data

4) Adapun hasil yang diperoleh sebagai berikut:

```
=== Run information ===
Scheme:       weka.classifiers.bayes.NaiveBayes
Relation:     Versi1-24
Instances:    5934
Attributes:   8
              Jalur
              Kotakab
              Propwil
              Penghasilan
              Pekerjaan
              Mangkircuti
              Ipk
              Ips_24
Test mode:    split 80.0% train, remainder test
=== Classifier model (full training set) ===
Naive Bayes Classifier
                          Class
Attribute       TIDAK POTENSI       POTENSI
                          (0.98)       (0.02)
==========================================
Jalur
  mean                  2.7955       2.8455
  std. dev.             1.6509       1.4025
  weight sum              5824          110
  precision                  1            1
Kotakab
  mean                 78.5158      81.9545
  std. dev.            26.0804      26.5451
  weight sum              5824          110
  precision                  1            1
Propwil
  mean                 12.5125      12.5727
  std. dev.             3.4684       2.9679
  weight sum              5824          110
  precision                  1            1

Penghasilan
  mean           2272991.7198 1813154.6894
  std. dev.      2080527.9742 1925623.2607
  weight sum              5824          110
  precision          40194.8843   40194.8843
Pekerjaan
  mean                  8.375       6.4727
  std. dev.             2.8913       3.4449
  weight sum              5824          110
  precision                  1            1
Mangkircuti
  mean                  0.0666            1
  std. dev.             0.2494       0.1667
  weight sum              5824          110
  precision                  1            1
Ipk
  mean                  3.4422       0.1632
  std. dev.             0.3742       0.3164
  weight sum              5824          110
  precision             0.0144       0.0144

Time taken to build model: 0.03 seconds
=== Predictions on test split ===
inst#,actual,predicted,error,prediction
1,1:'TIDAK POTENSI',1:TIDAK POTENSI,,1
2,1:'TIDAK POTENSI',1:TIDAK POTENSI,,1
3,1:'TIDAK POTENSI',1:TIDAK POTENSI,,1
4,1:'TIDAK POTENSI',1:TIDAK POTENSI,,1
5,1:'TIDAK POTENSI',1:TIDAK POTENSI,,1
6,1:'TIDAK POTENSI',1:TIDAK POTENSI,,1
7,1:'TIDAK POTENSI',1:TIDAK POTENSI,,1
8,1:'TIDAK POTENSI',1:TIDAK POTENSI,,1
9,1:'TIDAK POTENSI',1:TIDAK POTENSI,,1
10,1:'TIDAK POTENSI',1:TIDAK POTENSI,,1
11,1:'TIDAK POTENSI',1:TIDAK POTENSI,,1
12,1:'TIDAK POTENSI',1:TIDAK POTENSI,,1
13,1:'TIDAK POTENSI',1:TIDAK POTENSI,,1
14,1:'TIDAK POTENSI',1:TIDAK POTENSI,,1
15,1:'TIDAK POTENSI',1:TIDAK POTENSI,,1
16,1:'TIDAK POTENSI',1:TIDAK POTENSI,,1
17,1:'TIDAK POTENSI',1:TIDAK POTENSI,,1
18,1:'TIDAK POTENSI',1:TIDAK POTENSI,,1
19,1:'TIDAK POTENSI',1:TIDAK POTENSI,,1
20,1:'TIDAK POTENSI',1:TIDAK POTENSI,,1
21,1:'TIDAK POTENSI',1:TIDAK POTENSI,,1
22,1:'TIDAK POTENSI',1:TIDAK POTENSI,,1
23,1:'TIDAK POTENSI',1:TIDAK POTENSI,,1
24,2:POTENSI,2:POTENSI,,1
25,1:'TIDAK POTENSI',1:TIDAK POTENSI,,1
Dst
=== Evaluation on test split ===
```

```
Time taken to test model on test split: 0.34 seconds
=== Summary ===
Correctly Classified Instances        1180               99.4103 %
Incorrectly Classified Instances         7                0.5897 %
Kappa statistic                          0.8742
Mean absolute error                      0.0058
Root mean squared error                  0.0731
Relative absolute error                 14.6359 %
Root relative squared error             49.0269 %
Total Number of Instances             1187
=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
               0.996     0.074     0.998     0.996     0.997      0.875    0.981     0.999     TIDAK POTENSI
               0.926     0.004     0.833     0.926     0.877      0.875    0.998     0.969     POTENSI
Weighted Avg.  0.994     0.072     0.995     0.994     0.994      0.875    0.981     0.998
=== Confusion Matrix ===
   a    b   <-- classified as
 1155    5 |    a = TIDAK POTENSI
   2   25 |    b = POTENSI
```

Hasil menunjukkan klasifikasi dengan *naïve bayes classifier* untuk variable input dengan target IPs 2/4 memperoleh akurasi sebesar 99.41%.

Sebagai evaluasi, dari *confusion matrix* diperoleh penjelasan terdapat 7 ketidaktepatan data dimana 5 data yang seharusnya berpotensi menunjukkan tidak berpotensi serta 2 data yang menunjukkan tidak berpotensi ternyata berpotensi. Modeling dilakukan untuk 3 target yang lain yaitu IP_NonKeprodian, IP_Keprodian serta Status_Terkini dengan persentase 80%:20%. Hasil akhir modeling untuk 4 target adalah sebagai berikut:

Tabel 5. Akurasi Metode pada 4 Variabel Target

Proporsi Data	IPs 2/4	DO/AKTIF	IP-Makul Prodi	IP-Makul Non Prodi
80% : 20%	99.410	96.9671	95.871	97.89

5. PENUTUP

Berdasarkan penelitian yang dilakukan diperoleh simpulan sebagai berikut

a. Variabel yang dapat menjadi pengindentifikasi potensi keberhasilan studi adalah jalur masuk, IP Kumulatif, riwayat mangkir/cuti, asal kota, pekerjaan orang tua dan penghasilan orang tua.
b. Identifikasi potensi keberhasilan dilakukan dengan berbasis variabel input dengan ragam target (*class*) yang menjadi indikator keberhasilan mahasiswa yaitu IP Semester 2 dan 4, IP mata kuliah keprodian, IP mata kuliah non keprodian serta status terkini mahasiswa. Prediksi potensi keberhasilan dilakukan dengan teknik data mining metode *naïve bayes classifier* dengan persentase kelas *training* dan kelas *testing* sebesar 80% dan 20% dan diperoleh akurasi sebesar 99,41% untuk prediksi variabel input 1,2,3,4,5,6,7 dengan target keberhasilan IPK tahun pertama dan kedua, 96.96 %, untuk target Status Mahasiswa Aktif terkini, 95.87% untuk target IPK Keprodian, dan 97.89 % untuk target IPK Non-Keprodian.

REFERENCES

BAN-PT. 2015. Dokumen Borang Akreditasi Institusi. Badan Akreditasi Nasional Perguruan Tinggi.

Kamus Besar Bahasa Indonesia. 2020. Kamus Besar Bahasa Indonesia (KBBI). Diakses 19 Januari 2020 melalui https://kbbi.kemdikbud.go.id/

Kantardzic, M. 2003. Data Mining: Concepts, Models, Methods and Algorithms. John Wiley & Sons, Inc.

Khoirunnisak, M. dan Nur Iriawan. 2013. Pemodelan Faktor-Faktor yang Mempengaruhi Mahasiswa Berhenti Studi (*Drop Out*) di Institut Teknologi Sepuluh Nopember menggunakan Analisis Bayesian *Mixture Survival*.

Kotu, V. and Bale Deshpande. 2015. *Predictive Analytics and Data Mining, Concepts and Practise with RapidMiner*. Morgan Kaufmann, Elseveir Inc.

Mann, P.S. 2010. Introductory Statistics. John Wiley & Sons, Inc.

North, M. 2016. Data Mining For The Masses With Implementations in Rapidminer and R. Second Edition. ISBN: 1523321431.

Redjeki, S. 2013. Identifikasi Penyakit dengan Gejala Awal Demam menggunakan K-Nearest Neighnor (KNN). Jurnal Buana Informatika Vol. 4 No. 1.

Republik Indonesia. 2012. Undang-undang Republik Indonesia Nomor 12 Tahun 2012 tentang Pendidikan Tinggi.

Republik Indonesia. 2015. Peraturan Menteri Riset, Teknologi dan Pendidikan Tinggi Republik Indonesia Nomor 44 Tahun 2015 tentang Standar Nasional Pendidikan Tinggi.

S. Karthika and N. Sairam. 2015. A Naïve Bayesian Classifier for Educational Qualification. Indian Journal of Science and Technology, Vol 8(16), July 2015.

Syafrudin. 2006. Analisis Faktor-faktor Yang Mempengaruhi Keberhasilan Studi Mahasiswa Program Sarjana Ekstensi Manajemen Agribisnis Institut Pertanian Bogor, Bogor: Fakultas Pertanian, Institut Pertanian Bogor.

UI. 2004. Keputusan Rektor Universitas Indonesia Nomor 478/SK/R/UI/2004 tentang Evaluasi Keberhasilan Studi Mahasiswa Universitas Indonesia.

UI. 2013. Peraturan Rektor Universitas Indonesia Nomor 2198/SK/R/UI/2013 tentang Penyelenggaraan Program Sarjana di Universitas Indonesia.

UIN Walisongo Semarang. 2015. Pedoman Akademik UIN Walisongo Tahun 2015.

UoW Machine Learning Group. 2016. WEKA The Workbench for Machine Learning. Diakses pada Juli 2018 melalui laman cs.waikato.ac.nz: https://www.cs.waikato.ac.nz/ml/weka/

Witten, I.H., Eibe Frank, and Mark A. Hall. 2017. *Data Mining Practical Machine Learning Tools and Technique.* 3nd Edition. Morgan Koufmann Publisher, Elsevier, Inc.

Yuniarti, W.D. 2019. Dasar-dasar Pemrograman dengan Python. Deepublish Publisher ISBN: 9786230203503.