Terbit *online* pada laman web jurnal :
http://ejournal.amikompurwokerto.ac.id/index.php/telematika/

Telematika

Terakreditasi Sinta "3" KEMENRISTEKDIKTI, No. 21/E/KPT/2018

Implementasi Data Mining Menggunakan Algoritme *Naive Bayes Classifier* dan *C4.5* untuk Memprediksi Kelulusan Mahasiswa

Endang Etriyanti[1], Dedy Syamsuar[2] dan Yesi Novaria Kunang[3]

[1,2,3]Program Studi Magister Teknik Informatika, Fakultas Ilmu Komputer
Universitas Bina Darma
Email : endang.etriyanti@gmail.com[1], dedy_syamsuar@binadarma.ac.id[2],
yesinovariakunang@binadarma.ac.id[3]

INFO ARTIKEL

Sejarah Artikel:
Menerima 8 Agustus 2019
Revisi 8 Oktober 2019
Diterima 24 Februari 2020
Online 28 Februari 2020

***Keywords*:**
Naïve Bayes Classifier
C4.5 Algorithm
Student Graduation
RapidMiner

Kata kunci:
Naive Bayes Classifier
Algoritme *C4.5*
Kelulusan Mahasiswa
RapidMiner

Korespondensi:
Telepon: +6281996312599
E-mail:
endang.etriyanti@gmail.com

ABSTRACT

The inability of students to complete their studies on time is faced by most of higher education institution. STMIK Bina Nusantara Jaya Lubuklinggau is one of those which is experienced with this matter. In most cases, the students could complete their studies longer than the expected duration. From 162 students of Sistem Informasi study program in the year 2013 and 2014, there were 117 students completed their studies on time, while 45 students were late. As a result, it could prevent new students from joining the institution since the limited student capacity. This study deploys data mining technique in predicting the graduation status of students on time. First, preprocessing is used to obtain a good dataset. Secondly, the data is processed to obtain a set of prediction. In this step, two mining algorithm were applied – Naive Bayes classifier and C4.5 algorithm to be knowing the performance of the two methods, the method has a greater accuracy value will be recommended to solving the problem of prediction of students graduation at STMIK Bina Nusantara Jaya Lubuklinggau. Thirdly, the result then was validated using K-Fold Cross Validation technique. Finally, the Confusion Matrix is deployed to ensure the accuracy of the prediction. The results indicate that the C4.5 Algorithm method can be used to predict student graduation status with an accuracy rate of 79,08% while the accuracy rate of the Naive Bayes Classifier method is only 78,46%. The dominant factor is IPK-S4 variable.

ABSTRAK

Ketidakmampuan mahasiswa untuk menyelesaikan studi tepat waktu dialami oleh sebagian besar Lembaga Pendidikan Tinggi. STMIK Bina Nusantara Jaya Lubuklinggau adalah salah satu perguruan tinggi yang mengalami hal tersebut. Dalam banyak kasus para mahasiswa menyelesaikan studi mereka lebih lama dari rentang waktu yang diharapkan. Dari 162 mahasiswa program studi Sistem Informasi tahun angkatan 2013 dan 2014 terdapat 117 mahasiswa yang menyelesakan studinya tepat waktu, sedangkan 45 mahasiswa terlambat. Akibatnya hal tersebut dapat menghambat mahasiswa baru untuk bergabung dengan lembaga karena kapasitas mahasiswa yang terbatas. Penelitian ini menggunakan teknik data mining dalam memprediksi status kelulusan mahasiswa. Pertama, preprocessing digunakan untuk mendapatkan dataset yang berkualitas. Kedua, data diproses untuk mendapatkan serangkaian prediksi. Pada langkah ini, dua Algoritme data mining diterapkan - Algoritme Naive Bayes Classifier dan Algoritme C4.5 dengan tujuan untuk mengetahui kinerja dari kedua algoritme dengan tingkat akurasi yang lebih besar akan direkomendasikan untuk menyelesaikan masalah prediksi kelulusan mahasiswa pada STMIK Bina Nusantara Jaya Lubuklinggau. Ketiga, hasilnya kemudian divalidasi menggunakan teknik K-Fold Cross Validation. Terakhir, Coffusion Matrix digunakan untuk memvalidasi nilai akurasi hasil prediksi. Hasil penelitian menunjukkan bahwa metode Algoritme C4.5 dapat digunakan untuk memprediksi status kelulusan mahasiswa dengan tingkat akurasi 79,08%

sedangkan metode Naive Bayes Classifier hanya 78,46%. Dengan faktor dominan adalah variabel IPK-S4.

PENDAHULUAN

Mahasiswa merupakan aspek penting yang harus diperhatikan dengan serius dalam evaluasi program studi. Salah satu indikator keberhasilan program studi dapat dilihat dari lama studi mahasiswa. Lama studi mahasiswa adalah rentang waktu bagi mahasiswa untuk menyelesaikan studinya. Selain itu lama studi mencerminkan tingkat pencapaian mahasiswa dalam studinya. Dalam persfektif yang lebih luas rata-rata lama studi mahasiswa mempengaruhi kualitas program studi dan oleh karena itu lama studi mahasiswa dijadikan salah satu kriteria untuk menentukan penilaian akreditasi oleh Badan Akreditasi Nasional Perguruan Tinggi (Zainuddin, 2019). Untuk alasan ini setiap lembaga pendidikan perlu memberikan perhatian terhadap lama studi mahasiswa.

Ketidakmampuan mahasiswa untuk menyelesaikan studi tepat waktu dihadapai oleh sebagian besar lembaga pendidikan tinggi. STMIK Bina Nusantara Jaya Lubuklinggau adalah salah satu perguruan tinggi yang mengalami hal tersebut. Dalam banyak kasus, para mahasiswa menyelesaikan studi mereka lebih lama dari rentang waktu yang diharapkan. Dari 162 data mahasiswa program studi Sistem Informasi tahun angkatan 2013 dan 2014 terdapat 117 mahasiswa yang dapat menyelesaikan studinya tepat waktu sedangkan 45 mahasiswa tidak tepat waktu atau terlambat. Akibatnya hal tersebut dapat menghambat mahasiswa baru untuk bergabung dengan lembaga karena kapasitas mahasiswa yang terbatas. Untuk itu daya tampung mahasiswa baru dan lama studi mahasiswa perlu diperhatikan (Bisri, 2015). Sehingga untuk mengantisipasi hal tersebut maka prediksi perlu dilakukan untuk mengetahui status kelulusan mahasiswa. Jika status kelulusan mahasiswa dapat diprediksi, maka bagian program studi perlu memberi perhatian serius kepada mahasiswa yang diprediksi terlambat untuk dapat meningkatkan IPK pada setiap semester agar dapat dapat menyelesaikan studinya sesuai rentang waktu yang diharapkan.

Prediksi menurut Salmu & Solichin (2017) merupakan proses keilmuan untuk mendapatkan *knowledge* secara berurutan berdasarkan bukti-bukti. Ada berbagai macam cara untuk menyelesaikan masalah prediksi, salah satunya adalah teknik penambangan data (*data mining*). Teknik *data mining* merupakan cara yang mudah dan relatif cepat untuk memperoleh pengetahuan secara otomatis (Suyanto, 2017) dan pengetahuan abstrak dari sebuah *database* yang besar (Mulya, 2019) yang meliputi bentuk dan/atau hubungan antar data. Data mining menurut Juliansa (2019) adalah proses untuk mendapatkan ilmu pengetahuan dari sebuah informasi yang berasal dari gudang basis data.

Bagian penting dalam *data mining* adalah teknik klasifikasi, yaitu cara yang digunakan untuk mempelajari data set agar didapatkan hubungan antar data yang membentuk *pattern* (pola) sehingga dapat diperoleh *knowledge.* Ada banyak metode *data mining* yang bisa diterapkan untuk klasifikasi. Algoritme yang populer antara lain *Artificial Neural Network,* Algoritme *C4.5*, *Nearest Neighbour Rule, Fuzzy Logic, Naive Bayes, K-Mean, Support Vector Machine,* dan lain-lain. Penelitian yang mengangkat topik tentang klasifikasi dan penerapan *data mining* telah banyak dilakukan sebelumnya (Prakoso & Tutik, 2017; Anam & Santoso, 2018; Amalia, 2017; Risqiati & Ismanto, 2017; Septiani, 2017; Zainuddin, 2019).

Beberapa penelitian sebelumnya juga mengukur tingkat akurasi masing-masing metode *data mining.* Penelitian Anam & Santoso (2018) membandingkan kinerja antara Algoritme *Naive Bayes Classifier* dengan Algoritme C4.5 dalam mengklasifikasikan data penerima beasiswa. Temuan dari penelitian tersebut menunjukan tingkat akurasi Algoritme C4.5 (96, 40%) lebih baik dibandingkan dengan *Naive Bayes*

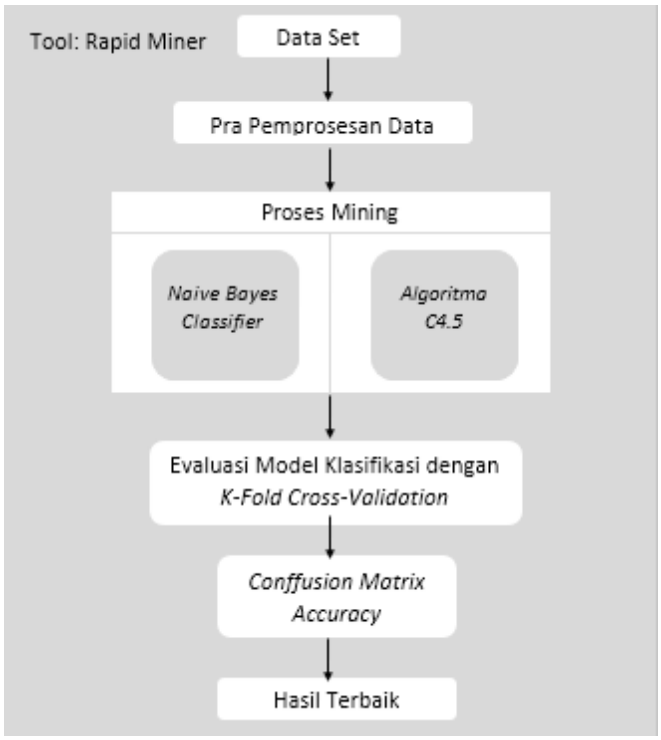
Classifier (95,11%). Hasil yang sama diperoleh pada penelitian selanjutnya (Prakoso & Tutik, 2017; Risqiati & Ismanto, 2017; Zainuddin, 2019) dimana tingkat akurasi C4.5 lebih baik dibandingkan dengan *Naive Bayes Classifier* dengan perbedaan berkisar antara 1-7%. Namun, hasil yang berbeda menjadi temuan dari Septiani (2017) dan Amalia (2017). Pada penelitiannya, Septiani (2017) memprediksi penyakit hepatitis, metode yang digunakan adalah komparasi metode Algoritme C4.5 dan *Naive Bayes Classifier*, dengan hasil penelitian *Naive Bayes Classifier* memiliki nilai akurasi 83,71% dan Algoritme C4.5 yaitu 77,29%. Amalia (2017) membandingkan metode *data mining* untuk memprediksi proses bersalin seorang ibu dengan menggunakan tiga metode yaitu *Neural Network*, *Naive Bayes Classifier* dan Algoritme C4.5. Secara berturut-turut diperoleh tingkat akurasi sebesar 93%, 94% dan 90%.

Berdasarkan uraian diatas, penelitian ini bertujuan untuk melakukan prediksi kelulusan mahasiswa STMIK Bina Nusantara Jaya dengan 2 metode yaitu *Naive Bayes Classifier* dan Algoritme C4.5. Data yang digunakan pada penelitian ini berjumlah 162 data mahasiswa program studi Sistem Informasi tahun angkatan 2013 dan 2014 yang sudah lulus. Secara teoritis penelitian ini berkontribusi dalam penerapan metode data mining untuk memprediksi kelulusan seorang mahasiswa. Manfaat selanjutnya adalah institusi dapat menentukan strategi dengan memberikan perhatian lebih bagi mahasiswa yang diprediksi akan terlambat.

METODE PENELITIAN

Tahap pertama yang dilakukan adalah pengumpulan data. Data yang diperoleh adalah sebanyak 227 data set mahasiswa yang telah menyelesaikan studinya yaitu data set mahasiswa tahun angkatan 2013 dan 2014 dengan 11 atribut. Tahap kedua dilakukan pra-pemrosesan data atau pengolahan data awal untuk mendapatkan data yang baik sebelum data diolah menggunakan menggunakan metode Algoritme C4.5 dan *Naive Bayes Classifier*. Setelah dilakukan pra-pemrosesan data, maka data set yang akan digunakan pada proses mining adalah 162 data mahasiswa dengan 9 atribut. Tahap ketiga dilakukan proses mining menggunakan metode Algoritme C4.5 dan *Naive Bayes Classifier* pada tools RapidMiner. Untuk memvalidasi nilai akurasi kedua metode yang digunakan diterapkan tehnik *K-Fold Cross Validation* dan hasil akurasi dapat dilihat berdasarkan *Confusion Matrix*. Tahap selanjutnya hasil pengujian dari metode Algoritme C4.5 dan *Naive Bayes Classifier* akan dibandingkan, dengan tujuan untuk mengetahui metode yang terbaik dengan tingkat akurasi yang paling tinggi.

Agar lebih jelas desain penelitian yang penulis gunakan dapat dilihat seperti pada Gambar 1.



Gambar 1. Desain Penelitian

1. Pengumpulan Data

Pengumpulan data dilakukan langsung dilapangan yaitu data mahasiswa program studi Sistem Informasi tahun angkatan 2013 dan 2014 yang sudah lulus yang diperoleh dari bagian akademik. Data yang diperoleh yaitu 227 data set mahasiswa dengan 11 atribut atau variabel. Variabel yang digunakan antara lain adalah Jenis Kelamin, Status Sekolah, Asal Sekolah, IP semester 1, IP semester 2, IP semester 3, IP semester 4, IPK semester 4 dan Status Kelulusan. Adapun contoh data yang digunakan dapat dilihat pada Tabel 1.

Tabel 1. Contoh Data

No	*NIM*	*Nama*	*Jenis Kelamin*	*Status Sekolah*	*Asal Sekolah*	*IP-S1*	*IP-S2*	*IP-S3*	*IP-S4*	*IPK-S4*	*Status Kelulusan*
1	2013.01.0001	Ahmad Shalihin	Laki-laki	Swasta	SMA	3,41	3,05	3,09	3,18	3,18	Tepat Waktu
2	2013.01.0003	Irma Tilawati	Perempuan	Negeri	SMA	3,23	3,05	2,89	3,2	3,1	Tepat Waktu
3	2013.01.0004	Muhammad Hidayatullah	Laki-laki	Negeri	SMK	3,27	3,14	3,14	3	3,14	Tepat Waktu
4	2013.01.0005	Nurhidayah	Perempuan	Negeri	SMK	3,32	2,95	2,78	3,1	3,05	Tepat Waktu
5	2013.01.0006	Sutrisno Raja Guk Guk	Laki-laki	Negeri	SMA	3,18	3,05	3	3	3,06	Tepat Waktu
6	2013.01.0008	Duwi Santoso	Laki-laki	Swasta	SMK	2,91	3,05	2,18	2,29	2,34	Terlambat
7	2013.01.0009	Hesti Kurnia	Perempuan	Negeri	SMK	3,05	2,95	2,65	3	2,92	Terlambat
8	2013.01.0010	Edi Lianto	Laki-laki	Negeri	SMA	3,73	3,23	3,82	3,55	3,58	Tepat Waktu
9	2013.01.0011	Rina	Perempuan	Swasta	SMA	3,59	3,68	3,91	3,82	3,75	Tepat Waktu
10	2013.01.0012	Dayang Sejoli	Perempuan	Swasta	SMA	3,73	3,64	3,86	3,55	3,69	Tepat Waktu

2. Alat dan Bahan

Alat dan bahan yang digunakan dalam penelitian ini antara lain adalah:

1) *Ms. Excel* digunakan untuk pengolahan data mentah atau data awal.
2) RapidMiner merupakan *tool* yang dimanfaatkan untuk mengimplementasikan metode *data mining* yang digunakan.

3) *Naive Bayes Classifier* dan Algoritme C4.5 sebagai Algoritme perhitungan untuk menyelesaikan masalah prediksi status kelulusan mahasiswa.
4) Teknik *K-Fold Cross-Validatio*n diterapkan untuk memvalidasi nilai akurasi model yang dibangun.

3. Pra Pemrosesan Data

Pada penelitian ini prediksi dilakukan berdasarkan data-data yang sudah terjadi, maksudnya adalah data yang penulis gunakan berupa data mahasiswa yang sudah menyelesaikan waktu studinya. Jadi data yang akan diolah telah memiliki variabel tujuan yaitu status kelulusan yang dikategorikan tepat waktu dan terlambat. Hal ini dimaksudkan agar dapat diketahui nilai akurasi hasil prediksi berdasarkan penerapan dari dua metode *data mining* yang digunakan. Penelitian ini sejalan dengan penelitian Risqiati & Ismanto (2017) penelitian tersebut menggunakan data kelulusan mahasiswa sebagai data set yang diimplementasikan dengan metode Algoritme C4.5 dan *Naive Bayes Classifier* pada tool RapidMiner.

Hasil dari pengumpulan data didapatkan *record* sebanyak 227 data set mahasiswa yang telah menyelesaikan studinya yaitu data set mahasiswa tahun angkatan 2013 dan 2014 dengan 11 atribut. Mahasiswa yang dikategorikan lulus tepat waktu yakni mahasiswa yang dapat menyelesaikan studinya selama 7 semester (3,5 tahun) atau 8 semester (4 tahun) untuk program sarjana. Sedangkan mahasiswa yang menyelesaikan pendidikannya lebih dari 8 semester, maka dikategorikan terlambat. Namun dari hasil pengumpulan data, data *record* dan atribut tidak seluruhnya bisa digunakan karena perlu dilakukan pra pemrosesan data atau pengolahan data awal untuk mendapatkan data yang baik. Adapun rincian 11 atribut yang belum dilakukan pra pemrosesan data terlihat seperti dalam Tabel 2 berikut:

Tabel 2. Atribut Sebelum Pra Pemprosesan Data

No	*Nama*	*Jenis Data*
1	NIM	Karakter
2	Nama	Karakter
3	Jenis Kelamin	Kategorikal
4	Status Sekolah	Kategorikal
5	Asal Sekolah	Kategorikal
6	IP-S1	Numerik
7	IP-S2	Numerik
8	IP-S3	Numerik
9	IP-S4	Numerik
10	IPK-S4	Numerik
11	Status Kelulusan	Kategorikal

Beberapa penelitian yang telah dilakukan menyatakan bahwa pra pemrosesan data perlu dilakukan untuk mendapatkan data set dengan kualitas baik. Seperti penelitian yang dilakukan oleh Zainuddin (2019) teknik *preproccesing* dilakukan untuk mendapatkan data dengan kualitas baik. Cara yang dilakukan antara lain *validation* data yaitu untuk menghilangkan pencilan, derau, data yang kosong dan yang inkonsisten, serta *discretization* data yaitu dilakukan seleksi atribut kelulusan. Selanjutnya penelitian yang dilakukan oleh Septiani (2017) untuk mendapatkan data dengan kualitas baik beberapa teknik yang dapat dilakukan antara lain *validation, integration and transformation, size reduction/discritization*. Dan dalam penelitian yang dilakukan oleh Prakoso & Tutik (2017) menyatakan pentingnya *preproccesing* data sebelum data set diproses menggunakan teknik *data*

mining. *Preproccesing* meliputi: memeriksa dan membuang data yang inkonsisten, data ganda, data yang perlu diperbaiki dan atau menambah data sesuai dengan kebutuhan.

Berdasarkan pada beberapa penelitian di atas, maka pada penelitian ini pra pemrosesan data dilakukan untuk mendapatkan data dengan kualitas baik. Pra pemrosesan data yang penulis gunakan antara lain:

a. Pembersihan Data, yaitu menghilangkan data yang kosong dan tidak lengkap. Misalnya, *record* mahasiswa dengan status berhenti dan non aktif dihapuskan karena mengandung data nilai mata kuliah yang tidak lengkap. Sehingga data yang awalnya berjumlah 227 menjadi 162 data set saja, jadi sebanyak 28,63% data yang kosong dan tidak lengkap dibersihkan/dihapus pada tahap pra pemrosesan data untuk menghindari adanya *missing value* dalam data set.
b. Reduksi Data, dilakukan guna mendapatkan data set dengan *record* dan jumlah atribut yang bersifat informatif saja. Sebagai contoh atribut NIM dan Nama tidak digunakan pada proses mining karena tidak relevan. Jadi atribut yang digunakan pada proses mining hanya atribut yang bersifat informatif saja yaitu jenis kelamin, status sekolah, asal sekolah, IP-S1, IP-S2, IP-S3, IP-S4, IPK-S4 dan Status Kelulusan.
c. Transformasi Data, digunakan untuk mengubah IP-S1, IP-S2, IP-S3, IP-S4 dan IPK-S4 yaitu nilainya dibuatkan interval yang lebar dan kedalamannya sama. Implementasi dilakukan pada *tool* RapidMiner, pra pemrosesan data dilakukan menggunakan operator *Discretize*.

Setelah dilakukan pra pemrosesan data, maka data set yang digunakan pada proses mining adalah 162 data mahasiswa dengan 9 atribut yang telah dinormalisasi dan *missing value* tidak terdapat pada data set tersebut. Adapun rincian atribut yang digunakan pada proses mining terlihat seperti pada Tabel 3:

Tabel 3. Atribut Data Setelah Pra Pemprosesan Data

No	*Nama*	*Jenis Data*
1	Jenis Kelamin	Kategorikal
2	Status Sekolah	Kategorikal
3	Asal Sekolah	Kategorikal
4	IP-S1	Numerik
5	IP-S2	Numerik
6	IP-S3	Numerik
7	IP-S4	Numerik
8	IPK-S4	Numerik
9	Status Kelulusan	Kategorikal

4. RapidMiner

Pada penelitian ini *tool* RapidMiner yang merupakan *tool data mining* Selain itu *tool* ini menampilkan visualisasi hasil olahan data. *Tool* RapidMiner adalah sebuah *tool* yang bersifat *open source*. RapidMiner menurut Mulya (2019) adalah sebuah alat *data mining* yang digunakan untuk menganalisa informasi. Pada penelitiannya Supriyanti, Kusrini, & Armadyah (2016) menyatakan bahwa RapidMiner merupakan sebuah tool yang dapat digunakan untuk membantu menyelesaikan masalah prediksi, proses *data mining* dan text mining. *Tool* RapidMiner mempunyai banyak operator *data mining* (lebih dari 500 operator), termasuk operator untuk *input*, *output*, data *preproccesing* dan lain-lain.

5. Naive Bayes Classifier

Dalam penelitian ini metode *Naive Bayes Classifier* digunakan sebagai Algoritme perhitungan untuk menyelesaikan masalah prediksi. Metode ini menggunakan teorema Bayes, yang bekerja berdasarkan probabilitas sederhana dinyatakan dengan persamaan berikut:

$$P\ (H|E) = \frac{P(E|H)P(H)}{P(E)} \qquad (1)$$

Keterangan persamaan:

E = Bukti

H = Hipotesis

P(H|E) = Hipotesis H benar untuk bukti E

P(E|H) = Kemungkinan sebuah bukti E terjadi akan memengaruhi hipotesis H atau dengan kata lain kemungkinan bahwa bukti E benar untuk hipotesis H

P(H) = Probabilitas awal (priori) hipotesis H terjadi tanpa memandang bukti apapun

P(E) = Probabilitas awal (priori) bukti E terjadi tanpa memandang hipotesis/bukti yang lain

Maksud dari aturan Bayes yakni berdasarkan bukti-bukti (E) yang diamati maka hasil hipotesis (H) dapat diprediksi.

6. Algoritme C4.5

Metode kedua yang penulis gunakan sebagai Algoritme perhitungan untuk menyelesaikan masalah prediksi kelulusan mahasiswa adalah Algoritme C4.5. Algoritme C4.5 menurut Prakoso & Tutik (2017) yaitu metode yang bisa diterapkan untuk menyelesaikan masalah klasifikasi data dengan atribut kategorial. Sedangkan Anam & Santoso (2018) berpendapat bahwa Algoritme C4.5 diterapkan guna membentuk sebuah pohon keputusan yang mempresentasikan aturan dalam klasifikasi.

Elemen penting yang harus dipahami dalam Algoritme C4.5 yaitu Enteropy dan Gain. Tahapan dalam membangun pohon keputusan antara lain adalah:

1. Memilih atribut untuk dijadikan node/akar
2. Membuat cabang dari setiap nilai
3. Membagi kasus pada setiap cabang
4. Ulangi proses dalam setiap cabang sampai seluruh kasus pada cabang berada pada kelas yang sama.

Pemilihan atribut sebagai node/akar yakni berdasarkan nilai Gain tertinggi dari dari semua atribut. Berikut adalah persamaan untuk menghitung nilai Gain:

$$Gaint\ (S, A) = Entropy\ (S) \sum_{i=i}^{n} \frac{|\ S_i\ |}{|\ S\ |} * \qquad (2)$$

Keterangan:

S = Himpunan Kasus

A = Atribut

n = Jumlah Partisi Atribut A

$|\ S_i\ |$ = Jumlah Kasus Pada Partisi Ke - i

| S | = Jumlah Kasus Dalam S

Nilai Enteropy dapat dihitung menggunakan persamaan berikut:

$$Entropy\ (S) = \sum_{i=i}^{n} Pi * Log_2\ Pi \qquad (3)$$

Keterangan:

S = Himpunan Kasus

A = Fitur

n = Jumlah Partisi S

P_i = Proporsi Dari S_i Terhadap S

http://dx.doi.org/10.35671/telematika.v13i1.881

7. Evaluasi Metode Klasifikasi

Evaluasi metode klasifikasi bertujuan untuk menganalisa perbandingan kinerja dari metode klasifikasi yang digunakan. Dalam penelitiannya Anam & Santoso (2018) menjelaskan bahwa evaluasi kinerja model klasifikasi dimaksudkan untuk mengetahui kinerja model klasifikasi berdasarkan hasil pengujian moedel yang diterapkan. Dalam penelitian ini hasil implementasi metode Algoritme C4.5 akan dibandingkan dengan *Naive Bayes Classifier*. Untuk memvalidasi nilai akurasi model yang dibangun digunakan metode *K-Fold Cross Validation* dan hasil akurasi dapat dilihat berdasarkan *Confusion Matrix*.

a. K-Fold Cross Validation

K-Fold Cross Validation menurut Anam & Santoso (2018) adalah teknik untuk memvalidasi nilai akurasi metode yang diterapkan berdasarkan data set. Suyanto (2017) menyatakan bahwa metode *K-Fold Cross Validation* membagi himpunan data D secara acak menjadi *k-fold* (sub himpunan) yang saling bebas f_1, f_2, ..., f_k, sehingga setiap *fold* berisi 1/k bagian data. Selanjutnya dapat membangun k himpunan data: D_1, D_2, ..., D_k, yang masing-masing berisi (k-1) *fold* untuk *training* data, 1-fold untuk *testing* data. Pada umumnya metode *k-fold cross validation* menggunakan 10 kali iterasi (k=10) dengan tujuan untuk memperoleh akurasi dengan bias dan variansi yang cukup rendah.

b. Confusion Matrix

Tujuan *Confusion Matrix* menganalisa kualitas kinerja model klasifikasi dalam mengenali variabel dari seluruh kelas. *Confussion Matrix* berisi informasi mengenai kelas sebenarnya dan kelas prediksi dari suatu proses klasifikasi (Anam & Santoso, 2018). Tabel matrix digunakan untuk mempresentasikan hasil evaluasi model klasifikasi. Misalnya data set terbagi menjadi kelas A dan kelas B, maka kelas A diasumsikan sebagai variabel positif dan kelas B diasumsikan sebagai variabel negatif. Nilai *accuracy*, *reccal* dan *precission* dapat diperoleh dari hasil evaluasi menggunakan *Confusion Matrix*. Gambar 2 merupakan contoh *Confusion Matrix*:

		Kelas Hasil Prediksi		
		Ya	Tidak	Jumlah
Kelas Aktual	Ya	TP	FN	P
	Tidak	FP	TN	N
	Jumlah	P	N	P + N

Gambar 2. *Confusion Matrix*

Perhitungan nilai akurasi, *precission* dan *reccal* dinyatakan dalam persamaan berikut:

$$Accuracy = \frac{TP+TN}{P+N} \quad (4)$$

$$Precission = \frac{TP}{TP+FP} \quad (5)$$

$$Reccal = \frac{TP}{P} \quad (6)$$

Keterangan:

TP *(True Positive)* : Jumlah variabel positif yang dilabeli dengan benar oleh *classifier*, sebagai contoh variabel dengan label status kelulusan = tepat waktu
TN *(True Negative)* : Jumlah variabel negatif yang dilabeli dengan benar oleh *classifier*
FP *(False Positive)* : Jumlah variabel negatif yang salah dilabeli oleh *classifier*
FN *(False Negative)* : Jumlah variabel positif yang salah dilabeli oleh *classifier*
P : Jumlah sampel positif

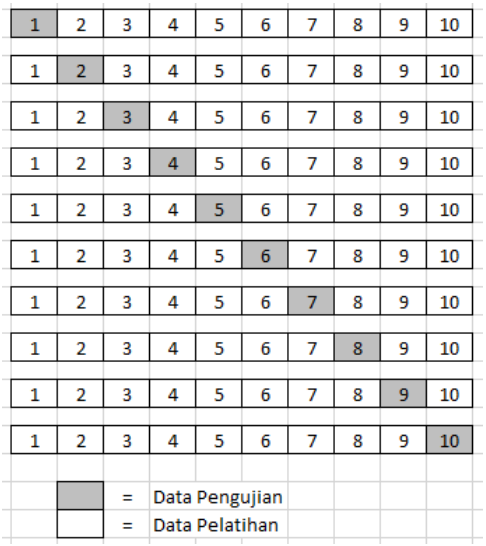
http://dx.doi.org/10.35671/telematika.v13i1.881

N : Jumlah sampel negatif

HASIL DAN PEMBAHASAN

1. Analisa Data

Data yang digunakan pada proses *training* dan *testing* telah melewati proses prapemrosesan data sehingga data telah siap diolah. Dataset yang digunakan dalam proses *training* maupun *testing* sebanyak 162 *record.* Atribut yang digunakan dalam proses *training* dan *testing* terdiri dari atribut Jenis Kelamin, Status Sekolah, Asal Sekolah, IP-S1, IP-S2, IP-S3, IPK-S4 dan Status Kelulusan. Proses *training* dan *testing* diterapkan pada *tool* RapidMiner 5.3 menggunakan metode perhitungan yaitu *Naive Bayes Classifier* dan Algoritme C4.5. Proses pengujian dilakukan dengan cara menambahkan operator *X-Validation (k-Fold Cross-Validation)* dengan melakukan 10 kali iterasi atau pengulangan (k=10 *fold)*. Dalam 10 *Fold Cross-Validation*, data dibagi menjadi 10 subset data. Dari 10 subset data tersebut *Cross-Validation* akan menggunakan 9 *fold* untuk pelatihan dan 1 *fold* untuk untuk pengujian seperti pada Gambar 3.



Gambar 3. Skema 10 *Fold Cross Validation*

Selanjutnya nilai akurasi dari hasil pengujian dapat dilihat berdasarkan *Confussion Matrik* seperti pada Gambar 4 dan Gambar 5.

2. Hasil Implementasi Metode *Naive Bayes Classifier* dan Algoritme C4.5

Dari hasil implementasi metode *Naive Bayes Classifier* didapatkan nilai akurasi sebesar 78,46% yang dievaluasi menggunakan *Confussion Matrix.* Gambar 4 menampilkan detail hasil yang diperoleh dari proses mining yang dilakukan dengan menggunakan *Naïve Bayes Classifier*.

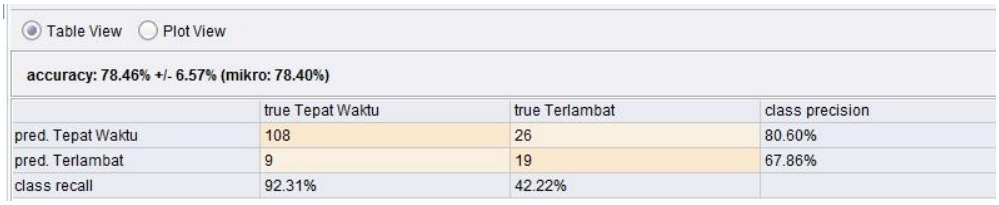


Table View Plot View

accuracy: 78.46% +/- 6.57% (mikro: 78.40%)

	true Tepat Waktu	true Terlambat	class precision
pred. Tepat Waktu	108	26	80.60%
pred. Terlambat	9	19	67.86%
class recall	92.31%	42.22%	

Gambar. 4 Nilai Akurasi Metode *Naive Bayes Classifier*

Gambar 4 menampilkan hasil perhitungan akurasi data set dengan metode *Naive Bayes Classifier* berdasarkan *confussion matrix*. Dari gambar tersebut dapat dilihat bahwa nilai *Accuracy* Rata-rata sebesar 78,46%, nilai *Reccal* kelas tepat waktu sebesar 92,31%, Nilai *Reccal* kelas terlambat sebesar 42,22%, nilai

Preccision kelas tepat waktu sebesar 80,60% dan nilai *Preccision* kelas terlambat sebesar 67,86%. Dari 162 data set, terdapat 108 data yang sesuai prediksi yaitu "tepat waktu", dan 26 yang diprediksi "tepat waktu" ternyata "terlambat". Dan sebanyak 9 data yang diprediksi "terlambat" tenyata termasuk kalsifikasi "tepat waktu" dan sebanyak 19 data sesuai prediksi yaitu "terlambat".

Selanjutya hasil implementasi metode Algoritme C4.5 pada *tool* RapidMiner diperoleh nilai *Accuracy* sebesar 79,08% berdasarkan *Confussion Matrix* seperti pada Gambar berikut:

Table View Plot View

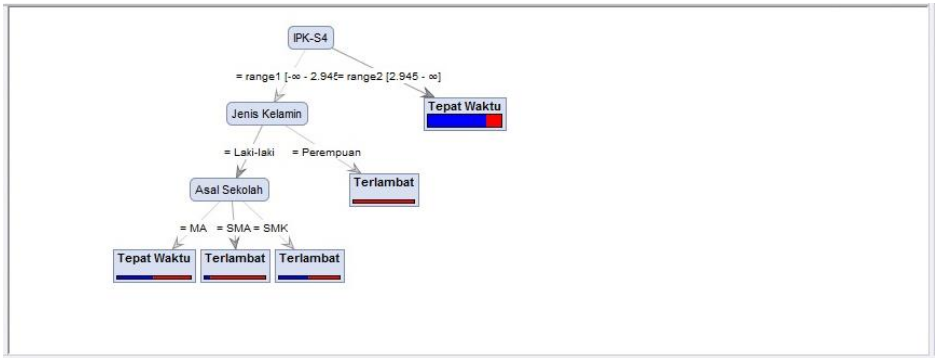
accuracy: 79.08% +/- 7.61% (mikro: 79.01%)

	true Tepat Waktu	true Terlambat	class precision
pred. Tepat Waktu	114	31	78.62%
pred. Terlambat	3	14	82.35%
class recall	97.44%	31.11%	

Gambar 5. Nilai Akurasi Metode Algoritme C4.5

Gambar 5 menampilkan hasil dari perhitungan akurasi data set dengan metode Algoritme C4.5 berdasarkan *confussion matrix*. Dari gambar tersebut dapat dilihat bahwa nilai *Accuracy* dari metode ini sebesar 79,08%, Nilai *Reccal* kelas tepat waktu sebesar 97,44%, Nilai *Reccal* kelas terlambat sebesar 31,11%, nilai *Preccision* kelas tepat waktu sebesar 78,62% dan nilai *Preccision* kelas terlambat sebesar 82,35%. Dari 162 data set, terdapat 114 data yang sesuai prediksi yaitu "tepat waktu", dan 31 yang diprediksi "tepat waktu" ternyata "terlambat". Dan sebanyak 3 data yang diprediksi "terlambat" tenyata termasuk kalsifikasi "tepat waktu" dan sebanyak 14 data sesuai prediksi yaitu "terlambat".

Dari hasil implementasi Algoritme C4.5 menggunakan *tool* RapidMiner maka terbentuk pohon keputusan berdasarkan nilai gain tertinggi adalah sebagai berikut:



Gambar 6. Pohon Keputusan Berdasarkan Informatif Gain

Dari Gambar 6 dapat dilihat bahwa variabel atau kriteria yang berpengaruh dalam prediksi kelulusan mahasiswa di STMIK Bina Nusantara Jaya Lubuklinggau adalah IPK-S4, Jenis Kelamin dan Asal Sekolah. Dari seluruh variabel yang digunakan, variabel IPK-S4 menjadi simpul akar, hal tersebut dikarenakan nilai gain tertinggi ada pada variabel IPK-S4. Variabel nilai IPK-S4 adalah nilai komulatif terakhir yang diambil selama mahasiswa mengikuti proses belajar di STMIK Bina Nusantara Jaya Lubuklinggau. Sehingga dengan IPK-S4 bisa menggambarkan kemajuan proses perkuliahan mahasiswa dan kendala atau hambatan yang dihadapi masing-masing mahasiswa. Data tersebut merupakan data paling dekat yang menggambarkan data prediksi kelulusan mahasiswa dibandingkan variabel-variabel yang lain. Hal ini sejalan dengan penelitian yang sudah dilakukan oleh Romadhona, suprapedi & himawan (2017) dan Priati (2016), dalam penelitiannya salah satu faktor dominan yang berpengaruh dalam prediksi kelulusan mahasiswa adalah IPK-S4.

Selain IPK-S4 yang menarik dari hasil pohon keputusan Algoritme C4.5 adalah jenis kelamin yang menjadi salah satu variabel dominan dalam penelitian ini. Dari pohon keputusan terlihat bahwa jenis kelamin perempuan diprediksi terlambat sedangkan jenis kelamin laki-laki diprediksi tepat waktu jika asal sekolah = MA. Jadi, dapat disimpulkan bahwa mahasiswa dengan jenis kelamin perempuan dan laki-laki yang asal sekolahnya SMK atau SMK perlu diberikan perhatian dan bimbingan yang lebih serius agar dapat dapat memperbaiki IPK pada setiap semester sehingga dapat lulus tepat waktu.

3. Perbandingan Hasil Akurasi Metode *Naive Bayes Classifier* dan Algoritme C4.5

Hasil dari implementasi yang telah dilakukan, perbandingan tingkat akurasi antara metode Algoritme C4.5 dan *Naive Bayes Classifier*:

Tabel 5. Perbandingan Nilai Akurasi Metode Naive Bayes Classifier dan Algoritme C4.5

No	*Metode*	*Nilai Akurasi*
1	*Naive Bayes Classifier*	78,46%
2	Algoritme C4.5	79,08%

Berdasarkan tabel diatas, prediksi kelulusan mahasiswa menggunakan metode Algoritme C4.5 memiliki nilai akurasi yang lebih tinggi dibandingkan dengan nilai akurasi metode *Naive Bayes Classifier* yaitu 79,08%. Selisih nilai akurasi antara kedua metode tersebut adalah sebesar 0,62%. Hal ini sejalan dengan penelitian yang telah dilakukan oleh Anam & Santoso (2018) dan penelitian yang dilakukan oleh Risqiati & Ismanto (2017) dimana nilai akurasi metode Algoritme C4.5 lebih besar dari metode *Naive Bayes Classifier*.

KESIMPULAN DAN SARAN

Kesimpulan dari penelitian ini bahwa hasil prediksi kelulusan mahasiswa pada STMIK Bina Nusantara Jaya Lubuklinggau berdasarkan data set yang diimplementasikan dengan metode *Naive Bayes Classifier* menunjukan nilai *Accuracy* 78,46% dan prediksi menggunakan metode Algoritme C4.5 diperoleh nilai *Accuracy* yang lebih besar yaitu 79,08%. Karena Algoritme C4.5 memiliki nilai akurasi yang lebih besar dibandingkan dengan nilai akurasi metode *Naive Bayes Classifier* maka metode Algoritme C4.5 direkomendasikan untuk digunakan dalam menyelesaikan masalah prediksi kelulusan mahasiswa pada STMIK Bina Nusantara Jaya Lubuklinggau. Dan dari pohon keputusan hasil implementasi Algoritme C4.5 dapat disimpulkan bahwa variabel atau kriteria yang berpengaruh dalam prediksi kelulusan mahasiswa di STMIK Bina Nusantara Jaya Lubuklinggau adalah IPK-S4, Jenis Kelamin dan Asal Sekolah.

Untuk pengembangan dan penelitian selanjutnya, penulis memberikan beberapa saran, yang pertama sebaiknya jumlah data perlu ditambah guna meningkatkan nilai *Accuracy*. Yang kedua, yaitu bukan hanya faktor intern atau faktor akademik saja yang dijadikan sebagai variabel atau kriteria namun faktor eksternal misalnya status bekerja, status pernikahan, faktor pembiayaan, dll perlu dijadikan sebagai variabel atau kriteria. Yang ketiga, penerapan *fitur selection* perlu dilakukan untuk pengembangan penelitian ini, atau untuk penelitian sejenis yang akan dilakukan. Selanjutnya penelitian sejenis dapat dilakukan dengan menerapkan metode *data mining* yang berbeda dengan metode yang telah penulis gunakan. Dan untuk pengembangan penelitian dapat dilakukan dengan mengadopsi hasil prediksi untuk dijadikan sebagai pendukung dalam proses pengambilan keputusan oleh para pemangku keputusan.

DAFTAR PUSTAKA

Amalia, H. E. (2017). Komparasi Metode Data Mining Untuk Penentuan Proses Persalinan Ibu Melahirkan, *13*, 103–109.

Anam, C., & Santoso, H. B. (2018). Perbandingan Kinerja Algoritma C4 . 5 dan Naive Bayes untuk Klasifikasi Penerima Beasiswa, *8*(1), 13–19.

Bisri, A. (2015). Penerapan Adaboost untuk Penyelesaian Ketidakseimbangan Kelas pada Penentuan Kelulusan Mahasiswa dengan Metode Decision Tree, *1*(1).

Juliansa, H. (2019). Data Mining Rought Set Dalam Menganalisa Kinerja Dosen STMIK Bina Nusantara Jaya Lubuklinggau, *4*(1), 11–17.

Mulya, D. P. (2019). Analisa dan Implementasi Association Rule Dengan Algoritma FP-Growth, *1*(1), 47–57.

Prakoso, S. A., & Tutik, E. T. (2017). Komparasi Algoritma C4.5 Dengan Naive Bayes Untuk Klasifikasi Kelulusan Mahasiswa Tepat Waktu Di PTS "KZX," *3*(1).

Priati. (2016). Kajian Perbandingan Teknik Klasifikasi Algoritma C4 . 5 , Naive Bayes Dan Cart Untuk Prediksi Kelulusan Mahasiswa (Studi Kasus : STMIK Rosma Karawang), (July 2016). https://doi.org/10.5281/zenodo.1184054

Risqiati, & Ismanto, B. (2017). Analisis Komparasi Algoritma Naive Bayes Dan C4-5 Untuk Waktu Kelulusan Mahasiswa, *XII*(1), 33–38.

Romadhona, Agus; suprapedi; himawan, H. (2017). Prediksi Kelulusan Tepat Waktu Mahasiswa Stmik-Ymi, *13*, 917.

Salmu, S., & Solichin, A. (2017). Prediksi Tingkat Kelulusan Mahasiswa Tepat Waktu Menggunakan Naïve Bayes : Studi Kasus UIN Syarif Hidayatullah Jakarta Prediction of Timeliness Graduation of Students Using Naïve Bayes : A Case Study at Islamic State University Syarif Hidayatullah Jakarta, (April), 701–709.

Septiani, W. D. (2017). Komparasi Metode Klasifikasi Data Mining Algoritma C4.5 Dan Naive Bayes Untuk Prediksi Penyakit Hepatitis. *Jurnal Pilar Nusa Mandiri*, *13*(1), 76–84.

Supriyanti, W., Kusrini, & Armadyah, A. (2016). Perbandingan kinerja algoritma c4.5 dan naive bayes untuk ketepatan pemilihan konsentrasi mahasiswa, *1*(2012).

Suyanto. (2017). *Data Mining Untuk Klasifikasi dan Klasterisasi Data*. Informatika Bandung.

Zainuddin, M. (2019). Perbandingan 4 Algoritma Berbasis Particle Swarm Optimization (pso) Untuk Prediksi Kelulusan Tepat Waktu Mahasiswa, *13*(1), 1–12.