Perbandingan Tingkat Akurasi Metode KNN Dan *Decision Tree* Dalam Memprediksi Lama Studi Mahasiswa

Endang Etriyanti
Sistem Informasi, STMIK Bina Nusantara Jaya Lubuklinggau
Jl. Yos Sudarso No. 97 A, Kota Lubuklinggau, Sumatera Selatan
Telp : 081996312599
E-mail : endang.etriyanti@gmail.com

Abstract

One of the qualities of graduates from a tertiary institution can be seen from the length of study of students. In addition, the length of study for students illustrates the level of student achievement in their education. The length of study also greatly affects the quality of the study program because the length of study for students is one of the criteria for accreditation assessment. Often the problem faced by a university is the number of students who complete their education in more than the specified time period. STMIK Bina Nusantara Jaya Lubuklinggau also experienced this. To anticipate this, it is necessary to predict the length of study for students because the length of study for students is one of the important things that need to be considered in the study program section of a university. This research contributes theoretically to the implementation of data mining to predict student study time. This research applies data preprocessing to obtain good quality data before the mining process is carried out using the K-Nearest Neighbor and Decision Tree methods on the Rapid Miner Tools, both methods are validated using the K-Fold. Cross Validation (with 10 iterations / repetitions) and Confusion Matrix are used to validate the accuracy of the predicted results. The highest accuracy value from the results of the application of the two methods will be recommended to solve the problem of predicting student study time. From the research results, the accuracy value of the Decision Tree method (60.38%) is better than the accuracy value of the K-Nearest Neighbor method (53.08%).

Keywords: Predictionsi, Lenght of Study, K-Nearest Neighbor, Decision Tree

Abstrak

Kualitas lulusan dari sebuah Perguruan Tinggi salah satunya dapat dilihat dari lama studi mahasiswa. Selain itu lama studi mahasiswa menggambarkan tingkat capaian mahasiswa dalam pendidikannya. Lama studi juga sangat berpengaruh pada kualitas program studi karena lama studi mahasiswa merupakan salah satu kriteria penilaian akreditasi. Seringkali masalah yang dihadapi oleh suatu Perguruan Tinggi adalah banyaknya mahasiswa yang menyelesaikan pendidikannya lebih dari jangka waktu yang ditetapkan. STMIK Bina Nusantara Jaya Lubuklinggau juga mengalami hal tersebut. Untuk mengantisipasi hal tersebut perlu adanya prediksi lama studi mahasiswa karena lama studi mahasiswa menjadi salah satu hal yang penting yang perlu diperhatikan bagian program studi dalam suatu Perguruan Tinggi. Penelitian ini berkontribusi secara teoretis dalam implementasi data mining untuk memprediksi lama studi mahasiswa.Penelitian ini menerapkan *preprocessing* data untuk memperoleh data dengan kualitas baik sebelum dilakukan proses mining menggunakan metode *K-Nearest Neighbor* dan *Decision Tree* pada *Tools RapidMiner*, kedua metode divalidasi menggunakan *K-Fold Cross Validation* (dengan 10 kali iterasi/pengulangan) dan *Confusion Matrix* digunakan untuk memvalidasi nilai akurasi hasil prediksi. Nilai akurasi yang paling tinggi dari hasil penerapan kedua metode akan direkomendasikan untuk menyelesaikan masalah prediksi lama studi mahasiswa. Dari hasil penelitian diperoleh nilai akurasi metode *Decision Tree* (60,38%) lebih baik jika dibandingkan dengan nilai akurasi metode *K-Nearest Neighbor* (53,08%).

Kata kunci: Prediksi, Lama Studi, *K-Nearest Neighbor, Decision Tree*

1. Pendahuluan

Kualitas lulusan dari sebuah Perguruan Tinggi salah satunya dapat dilihat dari lama studi mahasiswa. Selain itu lama studi mahasiswa menggambarkan tingkat capaian mahasiswa dalam pendidikannya. Lama studi juga sangat berpengaruh pada kualitas program studi karena lama studi mahasiswa merupakan salah satu kriteria penilaian akreditasi [1]. Lama studi merupakan jangka waktu yang diperlukan mahasiswa dalam menyelesaikan pendidikanya. Lama studi mahasiswa telah diatur dalam ketetapan Kementerian Pendidikan dan Kebudayaan Direktorat Jenderal Pendidikan Tinggi tentang Sistem Pendidikan Tinggi yang menyatakan bahwa untuk memenuhi standar kompetensi lulusan bagi mahasiswa program sarjana (S1) beban wajib yang harus ditempuh adalah paling sedikit 144 - 160 satuan kredit semester (sks) dengan lama studi selama 8 - 10 semester atau 4 - 5 tahun [2]. Oleh karena itu lama studi mahasiswa menjadi salah satu hal yang penting yang perlu diperhatikan bagian program studi dalam suatu Perguruan Tinggi.

Seringkali masalah yang dihadapi oleh suatu Perguruan Tinggi adalah banyaknya mahasiswa yang menyelesaikan pendidikannya lebih dari jangka waktu yang ditetapkan. Hal tersebut juga dialami oleh STMIK Bina Nusantara Jaya Lubuklinggau. Banyaknya mahasiswa yang menyelesaikan masa pendidikannya lebih dari jangka waktu yang ditetapkan mengakibatkan turunnya mutu lulusan [2]. Untuk mengantisipasi hal tersebut perlu adanya prediksi lama studi mahasiswa. Jika lama studi mahasiswa dapat diprediksi maka bagian program studi dapat mengambil tindakan/keputusan untuk mengantisipasi banyaknya mahasiswa yang diprediksi masa studinya lebih dari waktu yang ditetapkan tersebut.

[3] mendefinisikan prediksi sebagai proses keilmuan untuk mendapatkan *knowledge* secara berurutan berdasarkan bukti-bukti. Sedangkan [4] mendefinisikan prediksi hampir sama dengan klasifikasi dan estimasi, hanya saja prediksi digunakan untuk menduga nilai-nilai tertentu yang akan terjadi dimasa mendatang. Salah satu cara untuk menyelesaikan masalah prediksi adalah menggunakan teknik *data mining*. *Data mining* merupakan cabang ilmu baru untuk mengatasi masalah pengalian informasi atau pola yang penting atau menarik [1] dan pengetahuan abstrak dari sebuah *database* yang besar [5] yang meliputi bentuk dan/atau hubungan antar data. [6] juga mendefinisikan *data mining* sebagai proses ektraksi informasi untuk memperoleh pengetahuan dan menemukan pola pada tumpukan data berskala besar. Dari beberapa definisi di atas dapat disimpulkan bahwa *data mining* merupakan proses menganalisa data dalam jumlah yang besar menjadi suatu informasi yang lebih bermakna.

Ada banyak metode *data mining* yang dapat diterapkan untuk menyelesaikan masalah prediksi. 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 prediksi dan mengukur tingkat akurasi masing-masing metode *data mining.* telah banyak dilakukan sebelumnya. Penelitian [7] dan [8] membandingkan kinerja antara Algoritma *K-Nearest Neighbor* dengan *Decision Tree* dalam melakukan prediksi. Temuan dari penelitian tersebut menunjukan tingkat akurasi metode Algoritma *Decision Tree* lebih baik dibandingkan dengan *K-Nearest Neighbor*. Hasil yang sama diperoleh pada penelitian [9] membandingkan metode *data mining* untuk memprediksi mahasiswa lulus tepat waktu dengan menggunakan tiga metode yaitu *Decision Tree, Naive Bayes Classifier* dan *K-Nearert Neighbor*. Secara berturut-turut diperoleh tingkat akurasi sebesar 98.04%, 96% dan 90%. Pada penelitian [10] juga menunjukkan bahwa menunjukan tingkat akurasi metode Algoritma *Decision Tree* lebih baik dibandingkan dengan *K-Nearest Neighbor*. Dan pada penelitian [4] memprediksi kelulusan tepat waktu mahasiswa, dengan hasil penelitian *Decision Tree* memiliki nilai akurasi 76,69% dan *K-Nearert Neighbor* yaitu 69,82%.

Berdasarkan uraian diatas, penelitian ini bertujuan untuk melakukan prediksi lama studi mahasiswa STMIK Bina Nusantara Jaya dengan 2 metode yaitu *Decision Tree* dan *K-Nearest Neighbor*. Pada penelitian ini data yang digunakan berjumlah 260 data mahasiswa program studi Sistem Informasi tahun angkatan 2013 s.d. 2015 yang telah menyelesaikan masa studinya. Penelitian ini berkontribusi secara teoretis dalam implementasi data mining untuk memprediksi lama studi mahasiswa. Selain itu, penelitian ini diharapkan dapat memberi manfaat bagi institusi untuk dapat mengantisipasi banyaknya mahasiswa yang tidak mampu menyelesaikan pendidikanya sesuai waktu yang ditetapkan.

2. Tinjauan Pustaka

2.1 Data Mining

Data mining merupakan cabang ilmu baru untuk mengatasi masalah penggalian informasi atau pola yang penting atau menarik dari data dalam jumlah besar [1]. Ada pula yang menyatakan bahwa data mining adalah proses ektraksi informasi untuk memperoleh pengetahuan dan menentukan pola pada tumpukan data dalam database berskala besar [6]. Dan [10] mendefinisikan data mining sebagai bidang ilmu untuk proses mendapatkan pengetahuan atau pola dari basis data, sehingga dapat dijadikan solusi pengambilan keputusan.

2.2 *K-Nearest Neighbor*

K-Nearest Neighbor (KNN) merupakan algoritma yang berfungsi untuk mengklasifikasikan suatu data berdasarkan data pembelajaran *(train data sets)*, yang diambil dari K tetangga terdekatnya *(nearest neighbor)*, dengan K adalah banyaknya tetangga terdekat [8]. Dengan kata lain tujuan dari algoritma ini yaitu mengklasifikasikan objek baru berdasarkan atribut dan sample dari data training. Contoh kasus, misalkan untuk memprediksi lama studi mahasiswa digunakan data mahasiswa yang telah menyelesaikan masa studinya.

Tahapan metode KNN [8]:

1) Penentuan nilai K;
2) Perhitungan jarak antar data training dan data testing (uji). Perhitungan jarakke tetangga menggunakan algoritma euclidean seperti pada persamaan 1:
3) Pengurutan data hasil perhitungan;
4) Menentukan kelompok data hasil uji berdasarkan label mayoritas dari K tetangga terdekat.

2.4 *Decision Tree*

Decision Tree atau pohon keputusan merupakan alat pendukung keputusan yang menggunakan model keputusan yang berbentuk seperti pohon [8]. *Decision Tree* memprediksi sebuah kelas (klasifikasi) atau nilai (regresi) berdasarkan aturan-aturan yang dibentuk setelah mempelajari data [7].

Tahapan metode *Decision Tree* [8]:

1) Pilih atribut sebagai simpul akar;
2) Buat cabang untuk tiap-tiap nilai;
3) Bagi kasus dalam cabang;
4) Ulangi proses untuk setiap cabang sampai semua kasus pada cabang memiliki kelas yang sama;
5) Pemilihan atribut sebagai simpul, baik akar *(root)* atau simpul internal didasarkan pada nilai gain tertinggi dari atribut-atribut yang ada.

2.3 RapidMiner

Dalam melaksanakan proses mining biasanya digunakan alat bantu berupa *software*. Salah satu alat bantu tersebut adalah *RapidMiner*. *RapidMiner* digunakan untuk merancang aliran secara visual untuk menganalisis *data science* dan *machine learning* di dalam tim mulai dari analis hingga pakar.

Dalam penggunaannya *Rapidminer* tergolong mudah untuk digunakan mulai dari kempuannya untuk mengumpulkan data dari berbagai sumber, eksplorasi data secara statistik, tersedianya berbagai macam model *mechine learning* dan model validasi.

2.5 Prediksi

[4] mendefinisikan prediksi hampir sama dengan klasifikasi dan estimasi, hanya saja prediksi digunakan untuk menduga nilai-nilai tertentu yang akan terjadi dimasa mendatang. [3] juga mendefinisikan prediksi sebagai proses keilmuan untuk mendapatkan *knowledge* secara berurutan berdasarkan bukti-bukti.

2.6 *K-Fold Cross-Validation*

K-Fold Cross Validation digunakan untuk memvalidasi nilai akurasi kedua metode yang diterapkan. Teknik ini diterapkan dengan membagi data menjadi k bagian *(folds)*, satu bagian digunakan untuk pengujian dan dan sisanya (k-1 *folds)* untuk pemasangan model dengan proses iterasi dilakukan sampai seluruh *folds* digunakan sebagai pengujian [6]. [11] mendefinisikan *K-Fold Cross Validation* sebagai teknik validasi dengan membagi data secara acak ke dalam k bagian dan masing-masing bagian akan dilakukan proses klasifikasi, dengan teknik ini akan dilakukan pengujian sebanyak k. Secara umum pengujian nilai k dilakukan sebanyak 10 kali untuk memperkirakan akurasi estimasi.

2.7 Confusion Matrix

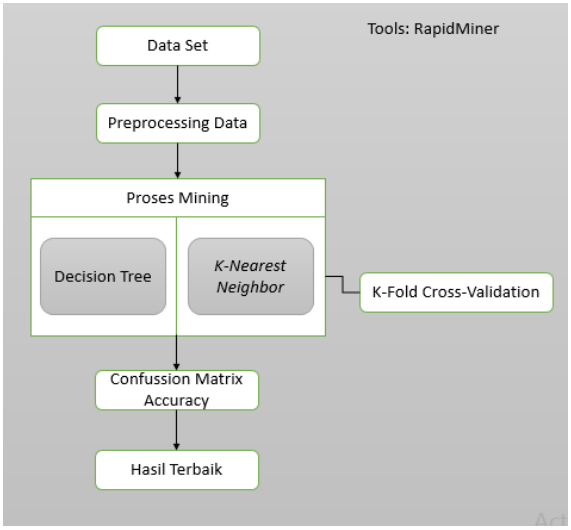
Confusion matrix merupakan matrix dua dimensi yang mengambarkan perbandingan antara hasil prediksi dengan kenyataan [11]. [4] juga menyatakan bahwa *Confusion Matrix* merupakan tabel perhitungan yang digunakan untuk melakukan evaluasi terhadap model klasifikasi berdasarkan perhitungan objek mana yang diprediksi benar dan tidak benar.

Confusion Matrix digunakan untuk mengevaluasi tingkat akurasi perfomansi model klasifikasi. Untuk mengukur performa medel klasifikasi dilakukan dengan membandingkan seluruh data uji yang diklasifikasikan benar dengan banyaknya data uji [6]. Akurasi adalah ukuran rasio prediksi yang benar terhadap total jumlah sampel dievaluasi [12]. Dengan kata lain, akurasi adalah tingkat kedekatan nilai prediksi dengan nilai aktual (sebenarnya).

3. Metodologi Penelitian

Metode penelitian digunakan sebagai acuan atau kerangka proses penelitian sehingga rangkaian proses penelitian dapat dilakukan secara sistematis dan terarah [13]. Tahap pertama yang dilakukan adalah pengumpulan data. Data yang diperoleh adalah sebanyak 326 data set mahasiswa tahun angkatan 2013 s.d. 2015 yang telah menyelesaikan studinya dengan 11 atribut. Tahap kedua dilakukan preprocessing data atau pengolahan data awal untuk mendapatkan data yang baik sebelum data diolah menggunakan menggunakan metode *Decision Tree* dan *K-Nearest Neighbor*. Setelah *preprocessing* data dilakukan maka data set yang diperoleh adalah 260 data mahasiswa dengan 9 atribut. Tahap ketiga, proses mining dilakukan menggunakan metode *Decision Tree* dan *K-Nearest Neighbor* pada *tools RapidMiner*. Penerapan tehnik *K-Fold Cross Validation* digunakan untuk memvalidasi nilai akurasi kedua metode yang diterapkan dan tingkat akurasi dapat dilihat berdasarkan *Confusion Matrix*. Tahap terakhir yaitu hasil pengujian dari metode *Decision Tree* dan *K-Nearest Neighbor* akan dibandingkan, dengan tujuan untuk mengetahui metode terbaik dengan melihat tingkat akurasi yang paling tinggi.

Metode penelitian yang diusulkan tersebut dapat digambarkan seperti pada Gambar 1 berikut:



Gambar 1. Metode Penelitian

3.1 Pengumpulan Data

Penelitian ini menggunakan data mahasiswa program studi Sistem Informasi tahun angkatan 2013 s.d. 2015 yang telah menyelesaikan masa studinya. Data diperoleh dari bagian akademik sebanyak 326 data set mahasiswa dengan 11 atribut atau variabel. Variabel yang digunakan antara lain adalah NIM, Nama, Jenis Kelamin, Status Sekolah, Asal Sekolah, IP semester 1, IP semester 2, IP semester 3, IP semester 4, IPK semester 4 dan Lama Studi. Contoh data set yang digunakan dalam penelitian seperti pada gambar 2 berikut:

No.	NIM	Nama Mahasiswa	Jenis Kelamin	IP-S1	...	IP-S4	IPK-S4	Status Sekolah	Asal Sekolah	Lama Studi
1	2013.01.0001	Ahmad Shalihin	Laki-laki	3,41	...	3,18	3,2	Swasta	SMA	4
2	2013.01.0003	Irma Tilawati	Perempuan	3,23	...	3,1	3,1	Negeri	SMA	4
3	2013.01.0004	Muhammad Hidavatullah	Laki-laki	3,27	...	3,14	3,1	Negeri	SMK	4
4	2013.01.0005	Nurhidayah	Perempuan	3,32	...	3,05	3,1	Negeri	SMK	4
5	2013.01.0006	Sutrisno Raja Guk Guk	Laki-laki	3,18	...	3,06	3,1	Negeri	SMA	4
6	2013.01.0007	Feri Dona Putra	Laki-laki	2,64	...	1,4	2,4	Negeri	SMK	> 4
7	2013.01.0008	Duwi Santoso	Laki-laki	2,91	...	2,29	2,3	Swasta	SMK	> 4
8	2013.01.0009	Hesti Kurnia	Perempuan	3,05	...	3	2,9	Negeri	SMK	> 4
9	2013.01.0010	Edi Lianto	Laki-laki	3,73	...	3,58	3,6	Negeri	SMA	3,5
10	2013.01.0011	Rina	Perempuan	3,59	...	3,75	3,8	Swasta	SMA	4

Gambar 2. Data Set

3.2 *Preprocessing* Data

Dalam penelitian ini prediksi dilakukan menggunakan data mahasiswa yang telah menyelesaikan masa studinya. Sehingga data yang yang digunakan dalam penelitian ini telah memiliki variabel tujuan yaitu lama studi yang dikategorikan menjadi 3 (tiga) yaitu lama studi 3.5 tahun, 4 tahun dan > 4 tahun. Hal ini dimaksudkan agar dapat diketahui nilai akurasi hasil prediksi berdasarkan penerapan dari dua metode *data mining* yang digunakan. Penelitian ini sejalan dengan penelitian [4] penelitian tersebut menggunakan data alumni mahasiswa sebagai data set.

Dari hasil pengumpulan data diperoleh sebanyak 326 data set mahasiswa tahun angkatan 2013 s.d. 2015 yang telah menyelesaikan masa studinya dengan 11 atribut. Namun tidak seluruhnya data *record* dan atribut tersebut dapat digunakan karena perlu dilakukan tahap *preprocessing* data atau pengolahan awal data untuk mendapatkan data set dengan kualitas baik. Adapun rincian 11 atribut yang belum dilakukan *preprocessing* data terlihat seperti dalam Tabel 1 berikut:

Tabel.1 Atribut Sebelum *Preprocessing* 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	Lama Studi	Kategorikal

Pentingnya *preprocessing* data diuraikan pada beberapa penelitian berikut. Seperti penelitian yang dilakukan oleh [4] yang menyatakan bahwa teknik *preproccesing* dilakukan agar kualitas data yang diperoleh lebih baik dengan cara: data *validation* dan data *discretization*. Selanjutnya [7] dalam penelitiannya menerapkan *preprocessing* data yang dilakukan dengan data *cleaning*. Dan dalam penelitian yang dilakukan oleh [14] menyatakan pentingnya *preproccesing* data sebelum data siap diolah. Dalam penelitiannya *preproccesing* meliputi proses *cleaning* yaitu membuang duplikasi data dan memperbaiki kesalahan pada data, transformasi data yaitu mengubah nilai variabel ke format yang sesuai, dan mereduksi data yaitu data yang hanya mempunyai atribut yang berhubungan saja yang akan digunakan.
Berdasarkan pada beberapa penelitian di atas, maka pada penelitian ini *preprocessing* data dilakukan untuk mendapatkan data dengan kualitas baik. *Preprocessing* data yang penulis gunakan antara lain seperti pada tabel 2 berikut:

Tabel.2 *Preprocessing* Data

No	Kegiatan	Tujuan
1	Pembersihan Data	Menghilangkan (menghapus) data yang kosong dan tidak lengkap untuk menghindari adanya *missing value* dalam data set.
2	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.
3	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, *preprocessing* data dilakukan menggunakan operator *Discretize*.

Setelah dilakukan *preprocessing* 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 *Preprocessing* 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	Lama Studi	Kategorikal

3.3 Proses Mining dan Validasi Nilai Akurasi

Proses mining diterapkan pada *Tools RapidMiner 5.3* menggunakan metode perhitungan *K-Nearest Neighbor* dan *Decision Tree*. Dan validasi tingkat akurasi kedua metode dilakukan dengan cara menambahkan operator *X-Validation (k-Fold Cross Validation)* dengan melakukan 10 kali iterasi atau pengulangan (k=10 fold). Dalam 10 kali iterasi 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.

3.4 Evaluasi Tingkat Akurasi

Nilai akurasi dari hasil pengujian dapat dilihat berdasarkan *Confussion Matrik* seperti pada Gambar 7 dan Gambar 11. Tabel matrix menanpilkan 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 3 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 3. *Confusion Matrix*

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

$$Accuracy = \frac{TP+TN}{P+N} \quad (1)$$

$$Precission = \frac{TP}{TP+FP} \quad (2)$$

$$Reccal = \frac{TP}{P} \quad (3)$$

Keterangan:

TP *(True Positive)* : Jumlah variabel positif yang dilabeli dengan benar oleh *classifier*

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

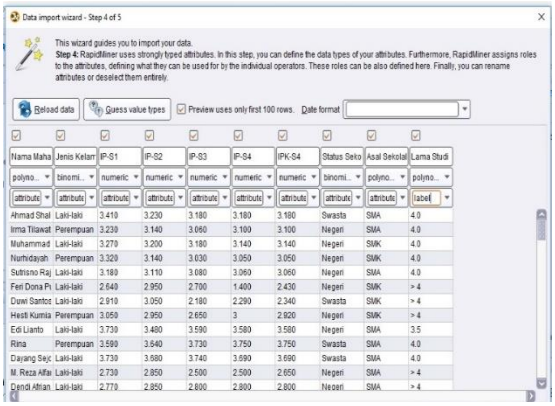
N : Jumlah sampel negatif

4. Hasil dan Pembahasan

4.1 Hasil Implementasi Metode *K-Nearest Neighbor*

Langkah-lanngkah proses mining menggunakan metode *K-Nearest Neighbor* pada *Tools RapidMiner* 5.3 adalah sebagai berikut:

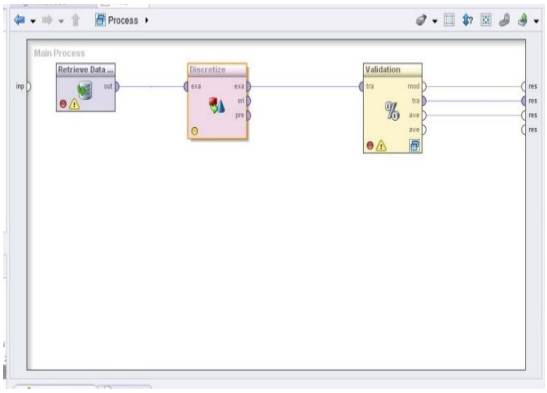
1. Pilih data yang akan diproses (penelitian ini menggunakan data dengan extensi excel).
2. Tentukan kelas data dan label tujuan (kelas tujuan dalam penelitian ini adalah “lama Studi”, sehingga nilai atribut yang digunakan pada variabel tersebut adalah “Label”.



Gambar 4. Penentuan Variabel Tujuan

3. Penerapan Operator yang digunakan

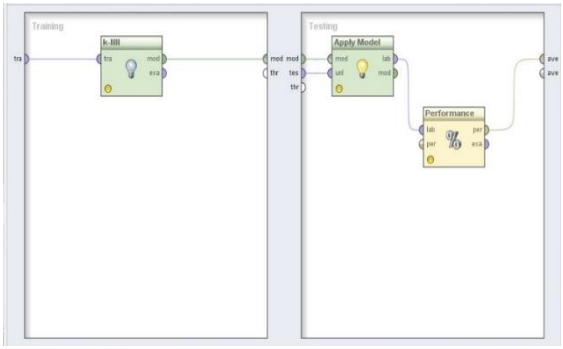
Operator yang digunakan pada tahap ini adalah *Discretize by Binning* dan *X-Validation*. Operator *Discretize by Binning* digunakan untuk mengubah nilai data IP-S1, IP-S2, IP-S3, IP-S4 dan IPK-S4, nilainya dibuatkan interval 2. Sedangkan operator *X-Validation* digunakan untuk menghitung kualitas pemodelan dan menghasilkan tingkat keakurasian berdasarkan data set. Konfigurasi proses implementasi pada *tool RapidMiner* 5.3 menggunakan operator operator *Discretize* dan operator *X-Validation* terlihat seperti pada Gambar 5 berikut ini:



Gambar 5. Penerapan Operator Yang Digunakan

4. Proses Training dan Testing

Tahap ini adalah memilih model klasifikasi yang digunakan, dalam penelitian ini model yang digunakan adalah *K-Nearest Neighbor*. Proses training dan testing dilakukan dengan cara pada kolom *training* diterapkan algoritma yang dipilih yaitu *K-Nearest Neighbor*, sedangkan dikolom *testing* diterapkan *Apply Model* dan *Performance.*

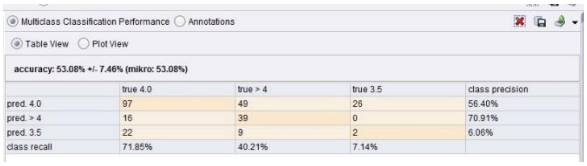


Gambar 6. Penerapan Metode *K-Nearest Neighbor*

Apply Model berfungsi untuk menjalankan model *K-Nearest Neighbor*, tahap ini digunakan untuk memproses data *training* dan data *testing. Performance* digunakan untuk menghubungkan *Apply Model* dan melihat hasil akurasi dari model yang digunakan.

5. Hasil Akurasi Implementasi Metode KNN

Implementasi metode *K-Nearest Neighbor* pada *Tools RapidMiner* menghasilkan nilai akurasi sebesar 53,08% yang dievaluasi menggunakan *Confussion Matrix.* Gambar 7 menampilkan detail hasil proses mining menggunakan metode *K-Nearest Neighbor.*



	true 4.0	true > 4	true 3.5	class precision
pred. 4.0	97	49	26	56.40%
pred. > 4	16	39	0	70.91%
pred. 3.5	22	9	2	6.06%
class recall	71.85%	40.21%	7.14%	

Gambar. 7 Nilai Akurasi Metode *K-Nearest Neighbor*

Dari hasil implementasi metode *K-Nearest Neighbor* diperoleh tingkat akurasi sebesar 53,08% yang dievaluasi berdasarkan *confussion matrix*. Dari 260 data set, terdapat 97 data yang sesuai prediksi yaitu Lama Studi = 4 tahun, 49 data yang diprediksi Lama Studi = 4 tahun ternyata masuk dalam klasifikasi Lama Studi = >4 tahun dan 26 data yang diprediksi Lama Studi = 4 tahun ternyata termasuk dalam klasifikasi Lama Studi = 3.5 tahun. Yang kedua prediksi Lama Studi = >4 tahun terdapat 39 data yang sesuai, yang diprediksi Lama Studi = >4 tahun ternyata masuk dalam klasifikasi Lama Studi = 4 tahun sebanyak 16 data dan yang diprediksi Lama Studi = >4 tahun ternyata masuk dalam klasifikasi Lama Studi = 3.5 tahun tidak ada. Yang terakhir yaitu prediksi Lama Studi = 3.5 tahun terdapat 2 data yang sesuai prediksi, yang diprediksi Lama Studi = 3.5 tahun ternyata masuk dalam klasifikasi Lama Studi = 4 tahun sebanyak 22 data dan yang diprediksi Lama Studi = 3.5 tahun ternyata masuk dalam klasifikasi Lama Studi = >4 tahun sebanyak 9 data.

Akurasi merupakan proporsi tuple positif yang diidentifikasi benar terhadap jumlah semua tuple. Nilai akurasi yang diperoleh dari implementasi metode

K-Nearest Neighbor berdasarkan *confussion matrix* menghasilkan akurasi 53,08% yang diperoleh dari perhitungan berikut ini:

$$Accuracy = \frac{TP+TN}{P+N}$$

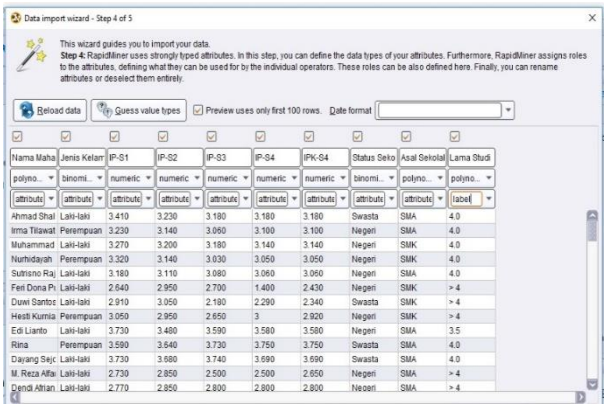
$$Accuracy = \frac{97+39+2}{135+97+28}$$

2. $Accuracy = \frac{138}{260}$
3. $Accuracy = 0{,}5307$
4. $Accuracy = 53{,}07\%$

4.2 Hasil Implementasi Metode *Decision Tree*

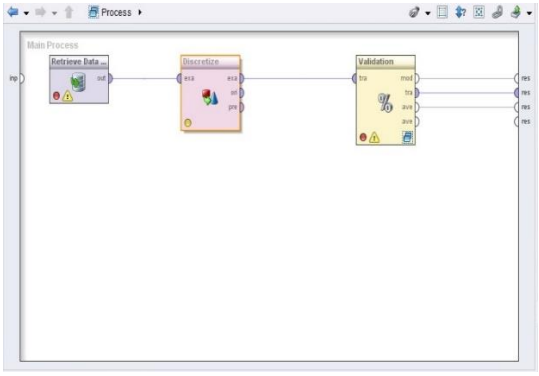
Langkah-lanngkah proses mining menggunakan metode *Decision Tree* pada *Tools RapidMiner* 5.3 adalah sebagai berikut:

1. Pilih data yang akan diproses (penelitian ini menggunakan data dengan extensi excel).
2. Tentukan kelas data dan label tujuan (kelas tujuan dalam penelitian ini adalah "lama Studi", sehingga nilai atribut yang digunakan pada variabel tersebut adalah "Label".



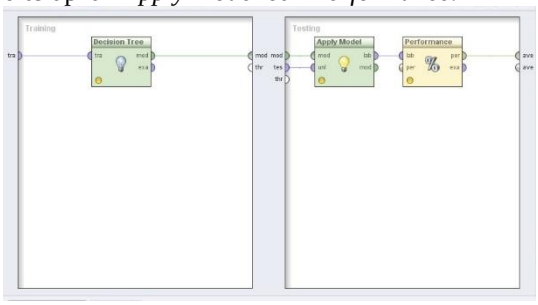
Gambar 8. Penentuan Variabel Tujuan

3. Penerapan Operator yang digunakan
   Operator yang digunakan pada tahap ini adalah *Discretize by Binning* dan *X-Validation*. Operator *Discretize by Binning* digunakan untuk mengubah nilai data IP-S1, IP-S2, IP-S3, IP-S4 dan IPK-S4, nilainya dibuatkan interval 2. Sedangkan operator *X-Validation* digunakan untuk menghitung kualitas pemodelan dan menghasilkan tingkat keakurasian berdasarkan data set. Konfigurasi proses implementasi pada *tool RapidMiner* 5.3 menggunakan operator operator *Discretize* dan operator *X-Validation* terlihat seperti pada Gambar 9 berikut ini:



Gambar 9. Penerapan Operator Yang Digunakan

4. Proses *Training* dan *Testing*
   Tahap ini adalah memilih model klasifikasi yang digunakan, dalam penelitian ini model yang digunakan adalah *Decision Tree*. Proses training dan testing dilakukan dengan cara pada kolom *training* diterapkan algoritma yang dipilih yaitu *Decision Tree*, sedangkan dikolom *testing* diterapkan *Apply Model* dan *Performance.*

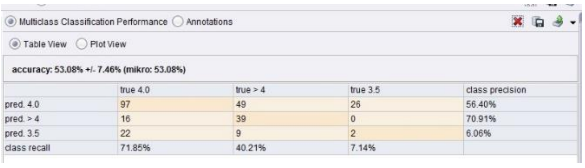


Gambar 10. Penerapan Metode *K-Nearest Neighbor*

Apply Model berfungsi untuk menjalankan model *Decision Tree*, tahap ini digunakan untuk memproses data *training* dan data *testing. Performance* digunakan untuk menghubungkan *Apply Model* dan melihat hasil akurasi dari model yang digunakan.

5. Hasil Akurasi Implementasi Metode *Decision Tree*

Implementasi metode *Decision Tree* pada *Tools RapidMiner* menghasilkan nilai akurasi sebesar 60,38% yang dievaluasi menggunakan *Confussion Matrix.* Gambar 11 menampilkan detail hasil proses mining menggunakan metode *Decision Tree.*



Gambar 11. Nilai Akurasi Metode *Decision Tree*

Dari hasil implementasi metode *Decision Tree* diperoleh tingkat akurasi sebesar 60,38% yang dievaluasi berdasarkan *confussion matrix*. Dari 260 data set,

terdapat 135 data yang sesuai prediksi yaitu Lama Studi = 4 tahun, 75 data yang diprediksi Lama Studi = 4 tahun ternyata masuk dalam klasifikasi Lama Studi = >4 tahun dan 28 data yang diprediksi Lama Studi = 4 tahun ternyata termasuk dalam klasifikasi Lama Studi = 3.5 tahun. Yang kedua prediksi Lama Studi = >4 tahun terdapat 22 data yang sesuai, yang diprediksi Lama Studi = >4 tahun ternyata masuk dalam klasifikasi Lama Studi = 4 tahun tidak ada dan yang diprediksi Lama Studi = >4 tahun ternyata masuk dalam klasifikasi Lama Studi = 3.5 tahun juga tidak ada. Yang terakhir yaitu prediksi Lama Studi = 3.5 tahun, tidak ada data yang sesuai dengan yang diprediksi.

Akurasi merupakan proporsi tuple positif yang diidentifikasi benar terhadap jumlah semua tuple. Nilai akurasi yang diperoleh dari implementasi metode *Decision Tree* berdasarkan *confussion matrix* menghasilkan akurasi 60,38% yang diperoleh dari perhitungan berikut ini:

$$Accuracy = \frac{TP+TN}{P+N}$$

$$Accuracy = \frac{135+22+0}{135+97+28}$$

5. $Accuracy = \frac{157}{260}$
6. $Accuracy = 0{,}6038$
7. $Accuracy = 60{,}38\%$

4.3 Perbandingan Hasil Akurasi Metode *K-Nearest Neighbor* dan *Decision Tree*

Hasil dari implementasi yang telah dilakukan, perbandingan tingkat akurasi antara metode *K-Nearest Neighbor* dan *Decision Tree*:

Tabel 5. Perbandingan Nilai Akurasi Metode *K-Nearest Neighbor* dan *Decision Tree*

No	Metode	Nilai Akurasi
1	*K-Nearest Neighbor*	53,08%
2	*Decision Tree*	63,38%

Berdasarkan tabel diatas, prediksi lama studi mahasiswa menggunakan metode *Decision Tree* memiliki nilai akurasi yang lebih tinggi dibandingkan dengan nilai akurasi metode *K-Nearest Neighbor* yaitu 63,38%. Hal ini sejalan dengan penelitian yang telah dilakukan oleh [7] dan penelitian yang dilakukan oleh [12] dimana nilai akurasi metode *Decision Tree* lebih besar dari metode *K-Nearest Neighbor*.

5. Kesimpulan

5.1 Simpulan

Kesimpulan dari penelitian ini bahwa hasil prediksi lama studi mahasiswa pada STMIK Bina Nusantara Jaya Lubuklinggau berdasarkan data set yang diimplementasikan dengan metode *K-Nearest Neighbor* menunjukan nilai *Accuracy* yaitu 53,08%. dan prediksi menggunakan metode *Decision Tree* diperoleh nilai *Accuracy* yang lebih besar yaitu 60,38%. Karena *Decision Tree* memiliki nilai akurasi yang lebih besar dibandingkan dengan nilai akurasi metode *K-Nearest Neighbor* maka metode *Decision Tree* direkomendasikan untuk digunakan dalam menyelesaikan masalah prediksi lama studi mahasiswa pada STMIK Bina Nusantara Jaya Lubuklinggau.

5.2 Saran

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. 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 Rujukan

[1] A. Azahari, Y. Yulindawati, D. Rosita, and S. Mallala, "Komparasi Data Mining Naive Bayes dan Neural Network memprediksi Masa Studi Mahasiswa S1," *J. Teknol. Inf. dan Ilmu Komput.*, vol. 7, no. 3, p. 443, 2020.

[2] S. Gunawan and P. Palupiningsih, "Pembentukan Model Klasifikasi Data Lama Studi Mahasiswa Stmik Indonesia Menggunakan Decision Tree," pp. 116–121, 2017.

[3] S. Salmu and A. Solichin, "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," no. April, pp. 701–709, 2017.

[4] M. Zainuddin and A. Noercholis, "Studi komparasi algoritma decesion tree (c4.5) dengan algoritma k-nn dalam memprediksi kelulusan tepat waktu mahasiswa," vol. 10, 2019.

[5] D. P. Mulya, "Analisa dan Implementasi

Association Rule Dengan Algoritma FP-Growth," vol. 1, no. 1, pp. 47–57, 2019.
[6] D. P. I. Putri, D. Anggreani, and A. prasetya Wibawa, "PREDIKSI LAMA STUDI MENGGUNAKAN NAÏVE BAYES," vol. 2, no. 1, pp. 41–50, 2020.
[7] D. L. Syarif and K. Umam, "Perbandingan Tingkat Akurasi Prediksi Aritmia dengan Menggunakan Algoritma K-Nearest Neighbor dan," no. Snik, pp. 10–15, 2020.
[8] D. M. Meliala and P. Hasugian, "Perbandingan Algoritma K-Nearest Neighbor Dengan Decision Tree Dalam Memprediksi Penjualan Makanan Hewan Peliharaan Di Petshop Dore Vet Clinic," vol. XV, no. November, pp. 35–39, 2020.
[9] A. Budiyantara, Irwansyah, E. Prengki, P. A. Pratama, and N. Wiliani, "KOMPARASI ALGORITMA DECISION TREE , NAIVE BAYES DAN K-NEAREST NEIGHBOR UNTUK MEMPREDIKSI MAHASISWA LULUS," vol. 5, no. 2, pp. 265–270, 2020.
[10] A. Rohman and M. Rochcham, "KOMPARASI METODE KLASIFIKASI DATA MINING UNTUK PREDIKSI Abstrak," vol. 5, no. 1, pp. 23–29, 2019.
[11] M. S. Maulana, R. Sabarudin, and W. Nugraha, "Prediksi Ketepatan Kelulusan Mahasiswa Diploma dengan Komparasi Algoritma Klasifikasi," vol. 7, no. 3, pp. 202–206, 2019.
[12] S. Widaningsih, "PERBANDINGAN METODE DATA MINING UNTUK PREDIKSI NILAI DAN WAKTU KELULUSAN MAHASISWA PRODI TEKNIK INFORMATIKA," vol. 13, no. 1, pp. 16–25, 2019.
[13] M. Zainuddin, "Perbandingan 4 Algoritma Berbasis Particle Swarm Optimization (pso) Untuk Prediksi Kelulusan Tepat Waktu Mahasiswa," vol. 13, no. 1, pp. 1–12, 2019.
[14] M. Windarti and A. Suradi, "Perbandingan Kinerja 6 Algoritme Klasifikasi Data Mining untuk Prediksi Masa Studi Mahasiswa," *Telematika*, vol. 12, no. 1, p. 14, 2019.