Jurnal Infotekmesin
Vol.11, No.02, Juli 2020
p-ISSN: 2087-1627, e-ISSN: 2685-9858
DOI: 10.35970/infotekmesin.v11i2.255, pp.80-86

Penerapan Algoritma Klasifikasi dengan Fitur Seleksi *Weight By Information Gain* pada Pemodelan Prediksi Kelulusan Mahasiswa

Avira Budianita[1*], Fandy Indra Pratama[2]
[1]Program Studi Ilmu Komputer, Universitas Muhammadiyah Kudus
[2]Program Studi Teknik Informatika, Universitas Wahid Hasyim
[1]Jln. Ganesha 1 Purwosari, Kabupaten Kudus, 59316, Indonesia
[2]Jln. Menoreh Tengah X / 22 Sampangan Gajahmungkur, Kota Semarang, 50232, Indonesia
E-mail: avirabudianita@umkudus.ac.id[1], fandy@unwahas.ac.id[2]

Info Naskah:
Naskah masuk: 13 Juni 2020
Direvisi: 29 Juli 2020
Diterima: 6 Agustus 2020

Abstrak

Permasalahan yang dihadapi institusi perguruan tinggi adalah tidak tepatnya kelulusan mahasiswa. Hal tersebut yang menjadi tugas institusi perguruan tinggi terutama program studi dalam memantau akademik mahasiswanya. Dewasa ini, banyak sekali metode atau cara untuk menyelesaikan berbagai permasalahan teknologi informasi salah satunya dengan penggunaan machine learning pada data mining. Penelitian ini menggunakan algoritma Naive Bayes, Linear Regression, dan MultiLayer Perceptron serta fitur seleksi Weight by Information Gain untuk optimalisasi akurasi dalam memprediksi waktu kelulusan mahasiswa. Hasil pengolahan dataset dengan atribut jalur pendaftaran, asal sekolah, asal kota, dan Indeks Prestasi semester 1 sampai 4 pada RapidMiner dengan menerapkan ketiga algoritma beserta fitur seleksi tersebut menghasilkan akurasi yang tergolong baik. Naive Bayes menghasilkan akurasi sebesar 81.66% dengan waktu eksekusi 1,16 detik, Linear Regression sebesar 80.70% dalam 2,44 detik dan MultiLayer Perceptron sebesar 82.16% dalam 1 jam 57 menit.

Keywords:
Linier Regression;
MultiLayer Perceptron;
Naive Bayes;
Time of Graduation;
Weight by Information Gain.

Abstract

The problem faced by institutions of higher education is not exactly of graduation. It is the task of institution of higher education, especially courses in the students academic monitoring. Today, there are many methods or ways to solve various problems of information technology either by machine learning at data mining. This study uses Naive Bayes. Linear Regression, dan Multi Layer Perceptron algorithm also Weight by Information Gain as Features Selection to optimize accuracy in predicting the time of students graduation. The dataset processing with certain attributes including registration line, origin school, origin city, and the semester grade point 1 to 4 in RapidMiner by implementing the three algorithms along with the selection feature produces relatively good accuracy. Naive Bayes produces an accuracy of 81.66% with an execution time of 1.16 seconds, Linear Regression of 80.70% in 2.44 seconds and Multi Layer Perceptron of 82.16% in 1 hour 57 minutes.

***Penulis korespondensi:**
Avira Budianita
E-mail: avirabudianita@umkudus.com

1. Pendahuluan

Tingkat kelulusan mahasiswa pada suatu perguruan tinggi menjadi tolak ukur keberhasilan dari perguruan tinggi itu sendiri. Pada Peraturan Menteri Pendidikan dan Kebudayaan Republik Indonesia (Permendikbud), studi S1 harus menempuh program belajar paling sedikit 144 sks atau menyelesaikan pendidikannya selama 4 tahun [1]. Tingkat kelulusan mahasiswa program studi XYZ pada suatu universitas berada di bawah standar yang telah ditetapkan oleh BAN-PT yaitu sebesar 50% dari mahasiswa yang mendaftar. Hal ini menjadi tugas tersendiri untuk program studi dalam mengevaluasi kinerja akademiknya guna mencapai standar lulusan mahasiswa yang telah ditetapkan oleh BAN-PT.

Penelitian-penelitian terdahulu saat ini terdapat beberapa yang mencoba untuk membandingkan beberapa lagoritma [3]. Pada penelitian Xhemali dkk [2], peneliti membandingkan 3 algoritma antara lain *Naive Bayes*, *Decision Tree*, dan *Neural Network* untuk klasifikasi web. Dari penelitian tersebut *Naive Bayes* menghasilkan akurasi yang lebih unggul daripada 2 algoritma lainnya. Sehingga Xhemali dkk menyimpulkan bahwa Algoritma *Naive Bayes* bekerja baik dalam melakukan proses klasifikasi dengan *dataset* web dibanding dengan algoritma lainnya pada penelitiannya.

Penelitian Arsad dkk dua tahun berturut-turut pada tahun 2013 dan 2014 [4][5][6] membandingkan algoritma *Linear Regression* dan *Neural Network* untuk memprediksi performa mahasiwa teknik pada fakultas Electrical Engineering, Universiti Teknologi MARA (UiTM) Malaysia. Pada penelitian ini, Arsad et al menggunakan IPK (Indeks Prestasi Kumulatif) mahasiswa sebagai label atau output pada datasetnya, sedangkan atribut yang digunakan sebagai variabel prediktor input adalah nilai mata kuliah pada semester-semester awal dan model yang dihasilkan adalah berbentuk MSE (Mean Square Error). Hasil penelitian dari kedua model menunjukkan korelasi yang kuat antara hasil yang mendasar untuk mata pelajaran inti di semester satu atau semester tiga dengan IPK akhir.

MultiLayer Perceptron digunakan oleh Houari dkk [7] pada penelitiannya yang dikembangkan dengan algoritma training *Levenberg-Marquardt* untuk prediksi suhu udara di wilayah Meknes di Maroko. Atribut yang digunakan pada penelitian Houri et al ini adalah tekanan atmosfer (Pr), kelembaban (H), visibility (Vis), kecepatan angin (V), titik embun (Tr) dan curah hujan (P). Dataset yang digunakan berisi sejarah meteorologi parameter scovering 3288 hari, dari tahun 2004 sampai 2012. MLP pada penelitian ini digunakan untuk mendekati hubungan antara parameter ini dengan *Minimum Square Error*. Untuk prediksi suhu udara yang lebih baik, Houari dkk mengembangkan model neural stokastik. *Mean Square Error* (MSE) dan koefisien korelasi (R) digunakan untuk mengevaluasi kinerja model yang dikembangkan. Studi indikator statistik ini menunjukkan bahwa prediksi suhu udara yang kuat dengan algoritma *Levenberg-Marquard*.

Beberapa penelitian juga menggunakan fitur seleksi untuk meningkatkan dan mengoptimalisasi hasil akurasi dari algoritma klasifikasi yang antara lain yaitu *Weight by Information Gain*, yang mana fitur seleksi tersebut berfungsi untuk memberikan nilai bobot pada setiap atribut yang relevan untuk proses prediksi.

Berdasarkan penelitian terkait tersebut, peneliti membandingkan keakuratan tiga algoritma klasifikasi ketika dilakukan seleksi *Weight by Information Gain*. Tiga algoritma tersebut adalah *Naive Bayes*, *Linier Regression* dan *Multi Layer Perceptron* yang mana algoritma tersebut memiliki keuanggulan pada kasus-kasus tertentu. Serta penggunaan variable Indeks Prestasi Semester (IPS) yang telah ditempuh hingga semester IV, jenis kelamin, kota lahir, asal kota, dan asal sekolah [8][9], dapat dijadikan sebagai prediktor untuk permasalahan prediksi kelulusan mahasiswa dalam penelitian ini.

Sehingga hasil dari penelitian ini yang membandingkan algoritma *Naive Bayes*, *Linier Regression* dan *Multi Layer Perceptron* ketika data dilakukan seleksi *Weight by Information Gain* menghasilkan kesimpulan baru yang dapat digunakan sebagai dasar membangun sistem prediksi yang optimal.

2. Metode Penelitian

Pada penelitian ini menggunakan *dataset* salah satu universitas di indonesia. Atribut dari *dataset* tersebut antara lain perlu dilakukan *preprocessing* karena terdapat data yang kiranya tidak digunakan. Berikut ini langkah-langkah dalam preprocessing yang dilakukan pada penelitian ini:

1) Penghapusan Record

Beberapa record dari data yang diperoleh akan dihapus pada proses ini yaitu data mahasiswa transfer dan pindahan serta data mahasiswa yang aktif, mangkir, ataupun cuti.

2) Penghapusan Atribut

Pada pemrosesan mining, atribut yang tidak memiliki pengaruh atau tidak digunakan harus dihapus seperti atribut NIM dan status akademik dan atribut yang digunakan pada proses menggunakan RapidMiner adalah jalur pendaftaran, asal sekolah, asal kota, ips1, ips2, ips3, dan ips4.

3) Pemberian Label

Pemberian label “1” untuk mahasiswa yang lulus tepat waktu yaitu yang menempuh pendidikan selama 3,5 tahun dan 4 tahun. Sedangkan label “2” untuk mahasiswa yang lulus tidak tepat waktu atau yang menempuh pendidikan lebih dari 4 tahun.

4) Inisialisasi

Proses inisialisasi ini dilakukan untuk memberikan inisial pada atribut asal sekolah dan asal kota. Asal sekolah diinisialisasi menjadi 2 kategori yaitu SMA dan SMK. Sedangkan asal kota akan diinisialisasi menjadi 2 kategori juga yaitu DALAM_KOTA dan LUAR_KOTA.

Selanjutnya data tersebut dilakukan seleksi atribut kembali dengan menggunakan *Weight by Information Gain* sebelum dilakukan klasifikasi oleh algoritma Naive Bayes, Linier Regression dan *Multi Layer Perceptron*

2.1 *Weight by Information Gain*

Weight by Information Gain merupakan salah satu algoritma yang mempunyai fungsi untuk menghitung relevansi atribut terhadap variabel target atau atribut label berdasarkan rasio gain informasi dan memberikan bobot yang sesuai pada atribut tersebut. Semakin tinggi nilai berat atribut, semakin atribut tersebut dianggap relevan. Bobot dari setiap atribut yang dianggap relevan atau memiliki mempengaruhi terhadap atribut label adalah kisaran 0-1[14][15], sehingga dihasilkan atribut yang digunakan pada penelitian ini yaitu pada Tabel 1. Pada Gambar 1 adalah contoh dataset yang digunakan pada penelitian ini.

Tabel 1. Atribut Dataset

NO	Atribut	Tipe Data	Kategori	Keterangan
1.	Jalur Pendaf-taran	Integer	Reguler = 1 Khusus = 2 PMDK = 3 Beasiswa Unggulan = 4	Jalur pendaftaran yang ditempuh mahasiswa pada saat pertama kali masuk universitas.
2.	Asal SLTA	Integer	SMA = 1 SMK = 2	Asal sekolah mahasiswa
3.	Kota Asal	Integer	Dalam Kota = 1 Luar Kota = 2	Asal kota mahasiswa
4.	IPS1	Numeric	0 – 4	IPS mahasiswa pada semester 1
5.	IPS2	Numeric	0 – 4	IPS mahasiswa pada semester 2
6.	IPS3	Numeric	0 – 4	IPS mahasiswa pada semester 3
7.	IPS4	Numeric	0 – 4	IPS mahasiswa pada semester 4

	A	B	C	D	E	F	G	H	I
1	jalur_daftar	asal_SLTA	kota_asal	ips1	ips2	ips3	ips4	status	
2	1	1	2	2,68	2,30	3,00	3,00	1	
3	1	1	2	3,20	3,13	3,13	2,75	1	
4	1	1	2	2,50	3,30	3,83	3,57	1	
5	1	1	2	3,65	3,43	3,39	3,61	1	
6	1	1	2	3,45	3,48	4,00	3,87	1	
7	1	1	1	2,25	3,14	3,39	3,10	1	
8	1	1	2	3,05	3,58	3,83	3,82	1	
9	1	2	2	2,20	3,81	3,62	3,74	1	
10	1	1	1	2,95	3,29	3,91	3,91	1	
11	1	1	2	2,40	2,57	3,09	2,62	1	
12	1	2	1	2,65	3,00	2,89	3,48	1	
13	1	1	2	3,25	3,21	3,10	3,05	1	
14	1	1	2	2,95	3,24	3,29	3,24	1	
15	1	1	2	3,05	3,00	3,71	3,17	1	

Gambar 1. Contoh Dataset

Selanjutnya dari data tersebut dilakukan perbandingan antara klasifikasi menggunakan algoritma *Naive Bayes*, *Linear Regression*, dan *Multi Layer Perceptron* (MLP) berdasarkan *Weight by Information Gain* sebagai pembobot atribut.

2.2 Naive Bayes Classifier

Naive Bayes Classifier (NBC) merupakan salah satu algoritma data mining yang menerapkan teorema Bayes dalam proses klasifikasi. *Naive Bayes Classifier* sendiri memiliki definisi pengklasifikasian dengan teknik probabilitas dan statistik untuk memprediksi kejadian di masa depan berdasarkan kejadian yang sudah ada sebelumnya [10][11]. Persamaan *Teorema Bayes* dan penjelasannya seperti pada persamaan 1 dan Tabel 2.

$$P(H|X) = \frac{P(X|H).P(H)}{P(X)} \quad (1)$$

Tabel 2. Keterangan Persamaan Teorema Bayes

X	**Data dengan class yang belum diketahui**
H	Hipotesis data X yang merupakan suatu *class* yang spesifik
$P(H\|X)$	Probabilitas H berdasarkan data X (posteriori)
$P(X\|H)$	Probabilitas X berdasarkan kondisi H
$P(H)$	Probabilitas H (prior)
$P(X)$	Probablitas X

2.3 Linear Regression

Linear Regression merupakan algoritma yang digunakan ketika akan memprediksi nilai dari suatu variabel berdasarkan nilai variabel yang lain. Bila ada lebih dari satu variabel input, *Linear Regression* akan menjadi *Multiple Linear Regression* atau *Multiple Regression* [4]. Persamaan Linear Regression seperti pada persamaan 1, serta penjelasannya pada tabel 3.

$$Y = a + bX \quad (2)$$

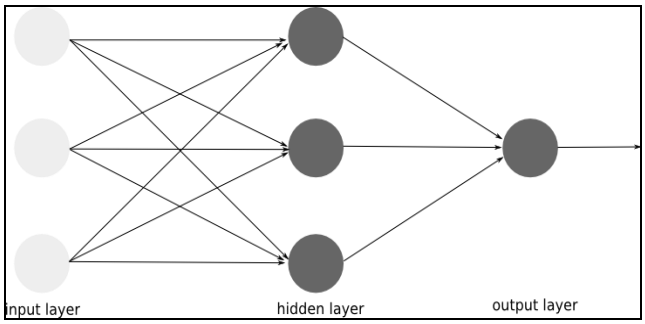
Tabel 3. Keterangan Persamaan Linear Regression

Y	**Variabel Response atau Variabel Akibat (Dependent)**
a	Konstanta
b	koefisien regresi (kemiringan); besaran Response yang ditimbulkan oleh Predictor.
X	Variabel Predictor atau Variabel Faktor Penyebab (Independent)

2.4 Multi Layer Perceptron

Multi Layer Perceptron (MLP) adalah suatu model feed-forward jaringan saraf tiruan yang memetakan set input data ke satu set output yang sesuai. MLP terdiri dari beberapa lapisan *node* dalam sebuah grafik yang diarahkan, dengan masing-masing lapisan sepenuhnya terhubung ke lapisan berikutnya. Kecuali untuk *node input*, setiap *node* adalah *neuron* (atau elemen pengolahan) dengan fungsi aktivasi nonlinier.

MLP menggunakan back propagation untuk pelatihan jaringan. Kelas jaringan ini terdiri dari beberapa lapisan unit komputasi, biasanya berhubungan dengan cara *feed-forward*. Dalam banyak aplikasi unit jaringan ini menerapkan *fungsi sigmoid* sebagai fungsi aktivasi [12][13].

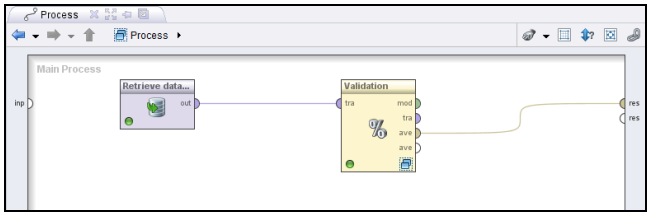


Gambar 2. Jaringan Multi Layer Perceptron

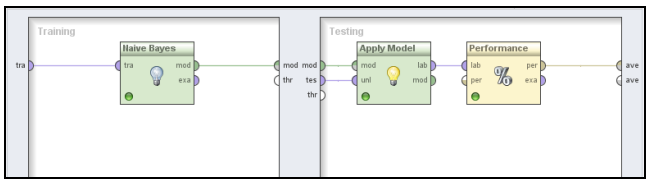
3. Hasil dan Pembahasan

3.1 Implementasi Algoritma Klasifikasi

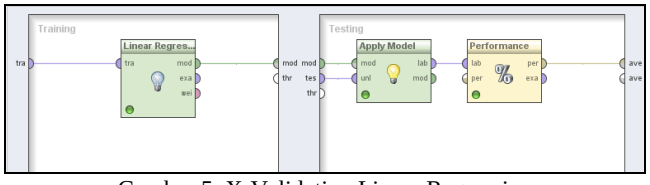
Implementasi algoritma dilakukan dengan menguji dataset mahasiswa menggunakan tools RapidMiner dengan 10 *fold number of validation* dan linear sampling untuk tipe samplingnya. Pada Gambar 3 menunjukkan dataset mahasiswa diuji menggunakan *X-Validation* pada RapidMiner. Pada Gambar 4 menunjukkan klasifikasi dengan menggunakan algoritma *Naive Bayes* pada *X-Validation*, serta Gambar 5 menunjukkan klasifikasi dengan menggunakan algoritma *Linear Regression* pada *X-Validation*.



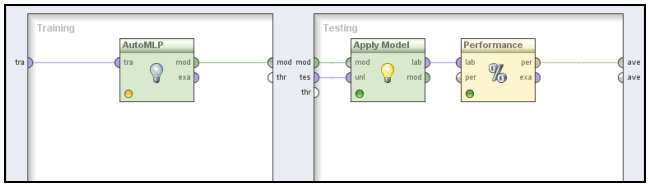
Gambar 3. Implementasi Algoritma



Gambar 4. X-Validation Naive Bayes



Gambar 5. X-Validation Linear Regression



Gambar 6. X-Validation Multi Layer Perceptron

Pada Gambar 6 menunjukkan klasifikasi dengan menggunakan algoritma *Multi Layer Perceptron* pada *X-Validation*. Validasi dengan *X-Validation* menggunakan RapidMiner menghasilkan tingkat akurasi dari ketiga algoritma tersebut dan tabel *confusion matrix*.

Tabel 4. Confusion Matrix [16]

Classification	***Predicted Class***	
	Class = Yes	*Class = No*
Class = Yes	*A (True Positive-TP)*	*B (False Negative-FN)*
Class = No	*C (False Positive-FP)*	*D (True Negative-TN)*

Keterangan dari tabel confusion matrix tersebut adalah:

a. A (True Positive-TP) : proporsi benar dalam dataset kategori benar.
b. B (False Negative-FN) : proporsi salah dalam dataset kategori salah.
c. C (False Positive-FP) : proporsi salah dalam dataset kategori benar.
d. D (True Negative-TN) : proporsi benar dalam dataset kategori salah.

Dari implementasi validasi algoritma Naive Bayes pada RapidMiner menggunakan X-Validation didapatkan hasil seperti pada Gambar 7. Perhitungan *accuracy*, *error rate*, *precision*, *recall*, dan *f-measure* berdasarkan confusion matrix algoritma Naive Bayes yang didapatkan dari proses validasi.

$$\textit{Accuracy} \quad \frac{TP+TN}{TP+TN+FP+FN} = \frac{232+266}{232+266+63+61} = 0.8007$$

$$\textit{Error Rate} \quad \frac{FP+FN}{TP+TN+FP+FN} = \frac{63+61}{232+266+63+61} = 0.1993$$

$$\textit{Precision} \quad \frac{TP}{TP+FP} = \frac{232}{232+63} = 0.7864$$

$$\textit{Recall} \quad \frac{TP}{TP+FN} = \frac{232}{232+51} = 0.8198$$

$$\textit{F-Measure} \quad \frac{2(PRECISION*RECALL)}{PRECISION+RECALL} = \frac{2(0.7864*0.8189)}{0.7864+0.8189} = 0.8023$$

Dari implementasi validasi algoritma Linear Regression menggunakan X-Validation didapatkan hasil seperti Gambar 8.

Table View | Plot View

accuracy: 80.07% +/- 5.70% (mikro: 80.06%)

	true 1	true 2	class precision
pred. 1	232	61	79.18%
pred. 2	63	266	80.85%
class recall	78.64%	81.35%	

Gambar 7. Akurasi & Confusion Matrix Naive Bayes

Table View | Plot View

accuracy: 79.27% +/- 6.83% (mikro: 79.26%)

	true 1	true 2	class precision
pred. 1	230	64	78.23%
pred. 2	65	263	80.18%
class recall	77.97%	80.43%	

Gambar 8. Akurasi & Confusion Matrix Linear Regression

Berikut adalah perhitungan *accuracy*, *error rate*, *precision*, *recall*, dan *f-measure* berdasarkan confusion matrix algoritma *Linear Regression* yang didapatkan dari proses validasi.

Accuracy	$\frac{TP+TN}{TP+TN+FP+FN} = \frac{230+263}{230+263+65+64} =$		0.7927
Error Rate	$\frac{FP+FN}{TP+TN+FP+FN} = \frac{65+64}{230+263+65+64} =$		0.2074
Precision	$\frac{TP}{TP+FP} = \frac{230}{230+65} =$		0.7997
Recall	$\frac{TP}{TP+FN} = \frac{230}{230+64} =$		0.7823
F-Measure	$\frac{2(PRECISION*RECALL)}{PRECISION+RECALL} = \frac{2(0.7997*0.7823)}{0.7997+0.7823} =$		0.7909

Dari implementasi validasi algoritma *Multi Layer Perceptron* menggunakan *X-Validation* didapatkan hasil pada Gambar 9.

Table View | Plot View

accuracy: 79.75% +/- 4.53% (mikro: 79.74%)

	true 1	true 2	class precision
pred. 1	240	71	77.17%
pred. 2	55	256	82.32%
class recall	81.36%	78.29%	

Gambar 9. Akurasi & Confusion Matrix Multi Layer Perceptron

Berikut adalah perhitungan *accuracy*, *error rate*, *precision*, *recall*, *dan f-measure* berdasarkan confusion matrix algoritma Multi Layer Perceptron yang didapatkan dari proses validasi.

Accuracy	$\frac{TP+TN}{TP+TN+FP+FN} = \frac{240+256}{240+256+55+71} =$		0.7975
Error Rate	$\frac{FP+FN}{TP+TN+FP+FN} = \frac{55+71}{240+256+55+71} =$		0.2026
Precision	$\frac{TP}{TP+FP} = \frac{240}{240+55} =$		0.8135
Recall	$\frac{TP}{TP+FN} = \frac{240}{240+71} =$		0.7717
F-Measure	$\frac{2(PRECISION*RECALL)}{PRECISION+RECALL} = \frac{2(0.8135*0.7717)}{0.8135+0.7717} =$		0.7920

Hasil akurasi yang didapat pada implementasi ketiga algoritma tersebut menunjukkan hasil akurasi yang berbeda, hasil akurasi ditunjukkan pada Tabel 5.

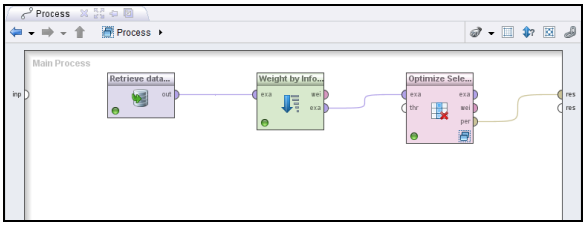
Tabel 5. Perbandingan Akurasi & Waktu Eksekusi Ketiga Algoritma

Algoritma	Akurasi	Waktu Eksekusi
Naive Bayes	80.06 %	0,63 detik
Linear Regression	79.27 %	0,63 detik
Multi Layer Perceptron	79.75 %	3 menit 38 detik

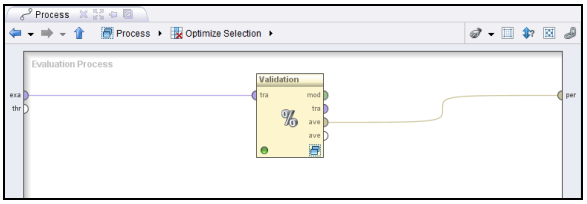
Tabel 5 menunjukkan hasil akurasi dan waktu eksekusi yang dibutuhkan dalam memproses data mahasiswa untuk prediksi kelulusan mahasiswa dengan *10 fold number of validation* dan *linear sampling* pada validasinya.

3.2 Implementasi Fitur Seleksi pada Algoritma Klasifikasi

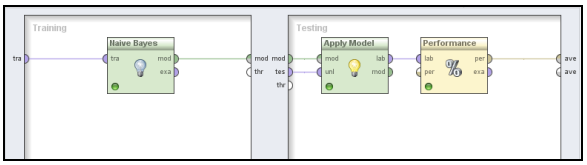
Implementasi algoritma yang dilakukan pada bagian ini sama dengan implementasi algoritma pada bagian sebelumnya tetapi ditambah dengan implementasi fitur seleksi Weight by Information Gain, telrihat pada Gambar 10. Di dalam operator Optimize Selection ditambahkan operator validasi untuk mengetahui hasil akurasi yang dihasilkan dari implementasi algoritma dan fitur seleksi Weight by Information Gain. Gambar 11 menunjukkan dataset mahasiswa diuji menggunakan *X-Validation*. Kemudian pada operator validasi diisikan algoritma yang akan diketahui hasil akurasinya sama seperti pada implementasi algoritma pada gambar 3. Pada Gambar 12 menunjukkan klasifikasi dengan menggunakan algoritma Naive Bayes + fitur seleksi Weight by Information Gain pada X-Validation. Gambar 13 menunjukkan klasifikasi dengan menggunakan algoritma Linier Regression + fitur seleksi *Weight by Information Gain* pada *X-Validation*.



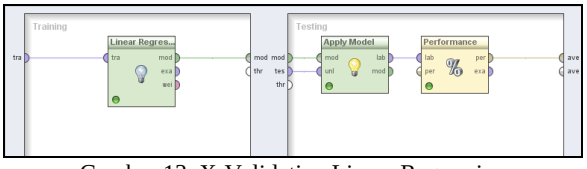
Gambar 10. Implementasi Algoritma + Fitur Seleksi



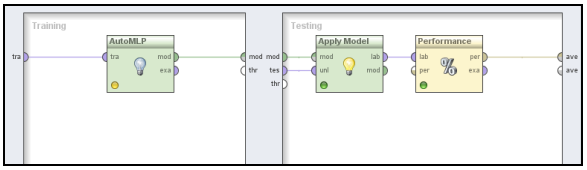
Gambar 11. Operator X-Validation pada Implementasi Algoritma + Fitur Seleksi



Gambar 12. X-Validation Naive Bayes



Gambar 13. X-Validation Linear Regression



Gambar 14. X-Validation Multi Layer Perceptron

Pada Gambar 14 menunjukkan klasifikasi dengan menggunakan algoritma *Multi Layer Perceptron* + fitur seleksi *Weight by Information Gain* pada *X-Validation*. Berdasarkan validasi dengan *X-Validation* menghasilkan tingkat akurasi dari ketiga algoritma ditambah dengan fitur seleksi Weight by Information Gain tersebut dan tabel confusion matrixnya.

Dari implementasi validasi algoritma Naive Bayes dengan Weight by Information Gain menggunakan X-Validation didapatkan hasil seperti pada Gambar 15.

Table View Plot View

accuracy: 81.66% +/- 7.17% (mikro: 81.67%)

	true 1	true 2	class precision
pred. 1	230	49	82.44%
pred. 2	65	278	81.05%
class recall	77.97%	85.02%	

Gambar 15. Akurasi & Confusion Matrix Naive Bayes

Berikut adalah perhitungan *accuracy*, *error rate*, *precision*, *recall*, dan *f-measure* berdasarkan confusion matrix algoritma *Naive Bayes* + *Weight by Information Gain*.

Accuracy	$\frac{TP+TN}{TP+TN+FP+FN} =$	$\frac{230+278}{230+278+65+49} =$	0.8166
Error Rate	$\frac{FP+FN}{TP+TN+FP+FN} =$	$\frac{65+49}{230+278+65+49} =$	0.1833
Precision	$\frac{TP}{TP+FP} =$	$\frac{230}{230+65} =$	0.7797
Recall	$\frac{TP}{TP+FN} =$	$\frac{230}{230+49} =$	0.8244
F-Measure	$\frac{2(PRECISION*RECALL)}{PRECISION+RECALL} =$	$\frac{2(0.7797*0.8244)}{0.7797+0.8244} =$	0.8108

Dari implementasi validasi algoritma *Linear Regression* dengan *Weight by Information Gain* menggunakan *X-Validation* didapatkan hasil seperti pada Gambr 16.

Table View Plot View

accuracy: 80.70% +/- 7.26% (mikro: 80.71%)

	true 1	true 2	class precision
pred. 1	234	59	79.86%
pred. 2	61	268	81.46%
class recall	79.32%	81.96%	

Gambar 16. Akurasi & Confusion Matrix Linear Regression

Berikut adalah perhitungan *accuracy*, *error rate*, *precision*, *recall*, dan *f-measure* berdasarkan *confusion matrix* algoritma *Linear Regression* + *Weight by Information Gain*.

Accuracy	$\frac{TP+TN}{TP+TN+FP+FN} =$	$\frac{234+268}{234+268+61+59} =$	0.8070
Error Rate	$\frac{FP+FN}{TP+TN+FP+FN} =$	$\frac{61+59}{234+268+61+59} =$	0.1929
Precision	$\frac{TP}{TP+FP} =$	$\frac{234}{234+61} =$	0.7932
Recall	$\frac{TP}{TP+FN} =$	$\frac{234}{230+59} =$	0.7986
F-Measure	$\frac{2(PRECISION*RECALL)}{PRECISION+RECALL} =$	$\frac{2(0.7932*0.7986)}{0.7932+0.7986} =$	0.7959

Dari implementasi validasi algoritma *Multi Layer Perceptron* dengan *Weight by Information Gain* menggunakan *X-Validation* didapatkan hasil seperti pada Gambar 17.

Table View Plot View

accuracy: 82.16% +/- 4.83% (mikro: 82.15%)

	true 1	true 2	class precision
pred. 1	239	55	81.29%
pred. 2	56	272	82.93%
class recall	81.02%	83.18%	

Gambar 17. Akurasi & Confusion Matrix Multi Layer Perceptron

Berikut adalah perhitungan *accuracy*, *error rate*, *precision*, *recall*, dan *f-measure* berdasarkan confusion matrix algoritma *Multi Layer Perceptron* + *Weight by Information Gain*.

Accuracy	$\frac{TP+TN}{TP+TN+FP+FN} =$	$\frac{239+272}{239+272+56+55} =$	0.8216
Error Rate	$\frac{FP+FN}{TP+TN+FP+FN} =$	$\frac{56+55}{239+272+56+55} =$	0.1785
Precision	$\frac{TP}{TP+FP} =$	$\frac{239}{239+56} =$	0.8102
Recall	$\frac{TP}{TP+FN} =$	$\frac{239}{239+55} =$	0.8129
F-Measure	$\frac{2(PRECISION*RECALL)}{PRECISION+RECALL} =$	$\frac{2(0.8102*0.8129)}{0.8102+0.8129} =$	0.8115

Hasil akurasi yang di dapat pada implementasi ketiga algoritma tersebut menunjukkan hasil akurasi yang berbeda, hasil akurasi ditunjukkan pada Tabel 6. Tabel 6 menunjukkan hasil akurasi dan waktu eksekusi dalam memproses data mahasiswa untuk prediksi kelulusan mahasiswa dengan algoritma *Naive Bayes*, *Linear Regression*, *Multi Layer Perceptron* dengan fitur seleksi *Weight by Information Gain*.

Tabel 6. Perbandingan Akurasi & Waktu Eksekusi Ketiga Algoritma + Fitur Seleksi

Algoritma	Akurasi	Waktu Eksekusi
Naive Bayes	81.66 %	1,16 detik
Linear Regression	80.70 %	2,44 detik
Multi Layer Perceptron	82.16 %	1 jam 57 menit

Tabel 7. Perbandingan Akurasi Tanpa dan Dengan Fitur Seleksi

Algoritma	Tanpa Fitur Seleksi	Dengan Fitur Seleksi
Naive Bayes	80.06 %	81.66 %
Linear Regression	79.27 %	80.70 %
Multi Layer Perceptron	79.75 %	82.16 %

Dari implementasi algoritma serta implementasi algoritma dengan fitur seleksi Weight by Information Gain didapatkan hasil perbandingan akurasi yang disajikan pada tabel 7. Tabel 7 menunjukkan perbandingan akurasi dari algoritma *Naive Bayes*, *Linear Regression* dan *Multi Layer Perceptron* tanpa atau dengan fitur seleksi *Weight by Information Gain*. Hasil yang diperoleh adalah akurasi dengan menggunakan fitur seleksi *Weight by Information Gain* lebih tinggi daripada tenpa menggunakan fitur seleksi *Weight by Information Gain*.

Penerapan algoritma *Naive Bayes*, *Linear Regression*, dan *Multi Layer Perceptron* menghasilkan akurasi yang tergolong baik. Algoritma *Naive Bayes* menghasilkan akurasi sebesar 80.06% dengan waktu eksekusi 0,63 detik , algoritma *Linear Regression* sebesar 79.27% dengan waktu eksekusi 0,63 detik dan algoritma *Multi Layer Perceptron* sebesar 79.75% dengan waktu eksekusi 3 menit 38 detik. Tetapi pada penerapannya, hasil akurasi tersebut masih dapat ditingkatkan dengan menggunakan fitur seleksi dan fitur seleksi yang dipilih pada penelitian ini adalah *Weight by Information Gain*. Dengan menggunakan fitur seleksi pada algoritma tersebut didapatkan bahwa algoritma *Naive Bayes* menghasilkan akurasi sebesar 81.66% dengan waktu eksekusi 1,16 detik , algoritma *Linear Regression* sebesar 80.70% dengan waktu eksekusi 2,44 detik dan algoritma *Multi Layer Perceptron* sebesar 82.16% dengan waktu eksekusi 1 jam 57 menit.

4. Kesimpulan

Dari hasil penelitian, disimpulkan bahwa penerapan fitur seleksi *Weight by Information Gain* pada algoritma *Naive Bayes*, *Linear Regression*, dan *Multi Layer Perceptron* menghasilkan peningkatan akurasi meskipun tidak terlalu signifikan dan waktu eksekusi yang tidak terpaut jauh jika menggunakan ataupun tanpa fitur seleksi kecuali *Multi Layer Perceptron* yang membutuhkan waktu lebih lama dalam proses eksekusi. Kesimpulan lain dari penelitian ini adalah algoritma Naive Bayes dengan menggunakan dataset mahasiswa dalam kasus prediksi kelulusan lebih unggul tingkat akurasinya dibanding dengan menggunakan algoritma *Linear Regression* dan *Multi Layer Perceptron* serta akurasi *Multi Layer Perceptron* lebih tinggi dibanding Linear *Regression* meskipun waktu eksekusi yang dibutuhkan terhitung lama jika tanpa fitur seleksi *Weight by Information Gain*. Sedangkan jika menggunakan fitur seleksi *Weight by Information Gain*, algoritma *Multi Layer Perceptron* lebih unggul hasil akurasinya dibanding algoritma *Naive Bayes* dan *Linear Regression* tetapi meskipun lebih unggul hasil akurasinya, waktu eksekusi yang dibutuhkan tergolong lama sehingga algoritma tersebut belum layak dikatakan optimal karena menurut beberapa penelitian, algoritma yang optimal yaitu algoritma dengan akurasi baik dan waktu eksekusi yang singkat.

Untuk penelitian penerapan algoritma pada pemodelan prediksi kelulusan mahasiswa kedepannya dapat menggunakan algoritma lain selain algoritma *Naive Bayes*, *Linear Regression*, dan Multi Layer Perceptron. Selain itu dapat pula menggunakan fitur seleksi maupun algoritma optimasi lainnya untuk meningkatkan hasil akurasi agar dapat lebih optimal. Penelitian ini juga dapat dikembangkan di bidang lain selain bidang pendidikan misalnya perbankan, kedokteran, dan lain-lain.

Daftar Pustaka

[1] M. Pendidikan, D. A. N. Kebudayaan, and R. Indonesia, "Peraturan Menteri Pendidikan Dan Kebudayaan Nomor 03 Tahun 2020 Tentang Standar Nasional Perguruan Tinggi," 2020.

[2] D. Xhemali, C. J. Hinde, and R. G. Stone, "Naive Bayes vs. Decision Trees vs. Neural Networks in the Classification of Training Web Pages," *Int. J. Comput. Sci.*, vol. 4, no. 1, pp. 16–23, 2009, [Online]. Available: http://cogprints.org/6708/.

[3] P. M. Barnaghi, V. A. Sahzabi, and A. A. Bakar, "A Comparative Study for Various Methods of Classification," *Int. Conf. Inf. Comput. Networks*, vol. 27, no. Icicn, pp. 62–66, 2012.

[4] P. M. Arsad, N. Buniyamin, and J. A. Manan, "A neural network students' performance prediction model (NNSPPM)," *Smart Instrumentation, Meas. Appl. (ICSIMA), 2013 IEEE Int. Conf.*, no. November, pp. 1–5, 2013, doi: 10.1109/ICSIMA.2013.6717966.

[5] P. M. Arsad, N. Buniyamin, and J. L. A. Manan, "Prediction of engineering students' academic performance using artificial neural network and linear regression: A comparison," *2013 IEEE 5th Int. Conf. Eng. Educ. Aligning Eng. Educ. with Ind. Needs Nation Dev. ICEED 2013*, pp. 43–48, 2014, doi: 10.1109/ICEED.2013.6908300.

[6] P. Mohd Arsad, N. Buniyamin, and J. L. Ab Manan, "Neural Network and Linear Regression methods for prediction of students' academic achievement," *IEEE Glob. Eng. Educ. Conf. EDUCON*, no. April, pp. 916–921, 2014, doi: 10.1109/EDUCON.2014.6826206.

[7] M. Ben, E. Houari, O. Zegaoui, and A. Abdallaoui, "Prediction of Air Temperature using Multi-Layer Perceptrons with Levenberg-Marquardt Training Algorithm," *Int. Res. J. Eng. Technol.*, vol. 02, no. 08, pp. 2395–56, 2015.

[8] A. Jananto, "Algoritma Naive Bayes untuk Mencari Perkiraan Waktu Studi Mahasiswa," *Teknol. Inf. Din.*, vol. 18, no. 1, pp. 9–16, 2013.

[9] M. Ridwan, H. Suyono, and M. Sarosa, "Penerapan Data Mining Untuk Evaluasi Kinerja Akademik Mahasiswa Menggunakan Algoritma Naive Bayes Classifier," *Eeccis*, vol. 7, no. 1, pp. 59–64, 2013.

[10] S. Budi, *Data Mining Teknik Pemanfaatan Data untuk Keperluan Bisnis*. 2007.

[11] H. Goker, H. I. Bulbul, and E. Irmak, "The Estimation of Students' Academic Success by Data Mining Methods," *2013 12th Int. Conf. Mach. Learn. Appl.*, vol. 2, pp. 535–539, 2013, doi: 10.1109/ICMLA.2013.173.

[12] D. Echegaray-Calderon, O A. Barrios-Aranibar, "Optimal selection of factors using Genetic Algorithms and Neural Networks for the prediction of students' academic performance," *2015 Lat. Am. Congr. Comput. Intell.*, pp. 1–6, 2015, doi: 10.1109/LA-CCI.2015.7435976.

[13] B. Rahmani and H. Aprilianto, "Early Model of Student's Graduation Prediction Based on Neural Network," *TELKOMNIKA (Telecommunication Comput. Electron. Control.*, vol. 12, no. 2, p. 465, 2014, doi: 10.12928/TELKOMNIKA.v12i2.1603.

[14] S. Chormunge and S. Jena, "Efficient feature subset selection algorithm for high dimensional data," *Int. J. Electr. Comput. Eng.*, vol. 6, no. 4, pp. 1880–1888, 2016, doi: 10.11591/ijece.v6i4.9800.

[15] A. A. Syafitri Hidayatul AA, Yuita Arum S, "Seleksi Fitur Information Gain untuk Klasifikasi Penyakit Jantung Menggunakan Kombinasi Metode K-Nearest Neighbor dan Naïve Bayes," *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 2, no. 9, pp. 2546–2554, 2018.

[16] F. Gorunescu, "Data mining: Concepts, models and techniques," *Intell. Syst. Ref. Libr.*, 2011, doi: 10.1007/978-3-642-19721-5.