

Data Mining dengan Model Klasifikasi *Naïve Bayes* untuk Pengukuran Pemilihan Matakuliah Peminatan

Iis Istianah dan Detty Purnamasari

Sistem Informasi Bisnis, Universitas Gunadarma

Jl. Margonda Raya No.100, Depok

E-mail : iis.tamad@gmail.com, detty@staff.gunadarma.ac.id

Abstrak

Kurikulum perguruan tinggi mencakup berbagai jenis matakuliah. Sebuah jurusan atau program studi tertentu memiliki struktur kurikulum. Mahasiswa yang belum mengetahui bakat dan minatnya menjadi salah satu faktor belum tercapainya mahasiswa yang lulus tepat waktu. Faktor utama yang mempengaruhi pada pemilihan peminatan antarlain memahami matakuliah inti pada program studi tertentu. Algoritma yang digunakan *naïve bayes*, data yang digunakan adalah tahun angkatan 2013, 2014, 2015 & 2016 mahasiswa teknik informatika, untuk mengetahui hasil klasifikasi dengan menggunakan algoritma *Naïve Bayes* dalam menentukan mata kuliah peminatan, terhadap data mahasiswa program studi Teknik Informatika-S1 STMIK Bani Saleh berdasarkan dari kemampuan akademisnya, sehingga dapat memberikan informasi yang baru dan berguna sebagai rekomendasi dalam pemilihan jalur peminatan. Algoritma yang digunakan untuk mengukur tingkat akurasi peminatan penjurusan Program Studi Teknik Informatika dengan jumlah data training sebanyak 824 dan data testing sebanyak 206 dengan menggunakan *Naïve Bayes*. Dengan memunculkan 2 peminatan yaitu *Software Engineering* dan *Networking*. Adapun aplikasi yang digunakan untuk membantu proses perhitungan *Naïve Bayes Rapidminer* 5.3. Dari hasil pengukuran menggunakan *rapidminer* diperoleh nilai akurasi sebesar 98,06%.

Kata kunci : peminatan, klasifikasi, *Naïve Bayes*, *Rapidminer*

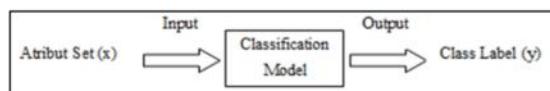
Pendahuluan

Kurikulum perguruan tinggi mencakup berbagai jenis mata kuliah. Sebuah jurusan atau program studi tertentu memiliki struktur kurikulum yang disusun sesuai dengan aturan yang telah ditetapkan. Pada sebuah jurusan atau program studi tersebut kurikulum untuk mata kuliah biasanya terdapat mata kuliah yang wajib untuk diambil dan peminatan. Mata kuliah pilihan (peminatan) biasanya diambil berdasarkan bidang peminatan seorang mahasiswa pada jurusan tersebut. Mahasiswa yang belum mengetahui bakat dan minatnya menjadi salah satu faktor belum tercapainya mahasiswa yang lulus tepat waktu. Faktor utama yang mempengaruhi pada pemilihan peminatan an-

taralain memahami matakuliah inti pada program studi tertentu, dalam memprediksi suatu hal di dalam data mining terdapat beberapa metode antara lain *Naïve Bayes* [3,8,9,10]. Oleh karena itu penelitian ini menggunakan metode *Naïve Bayes*, *Naïve Bayes* merupakan metode yang bersifat sederhana dan efektif dalam pengolahan data yang besar [7] dengan menggunakan algoritma *Naïve Bayes* dalam menentukan mata kuliah peminatan. Prinsip ini kemudian digunakan untuk memecahkan masalah klasifikasi [1]. Penggunaan algoritma ini dinilai sesuai karena *Naïve Bayesian classification* merupakan salah satu algoritma klasifikasi yang sederhana namun memiliki kemampuan dan akurasi tinggi [2].

Tujuan dari hasil dari klasifikasi menggunakan algoritma naive bayes dalam menentukan menggunakan algoritma *Naïve Bayes* dalam menentukan mata kuliah peminatan, terhadap data mahasiswa program studi Teknik Informatika-S1 STMIK Bani Saleh berdasarkan dari kemampuan akademisnya. Tujuan dari penelitian ini adalah menentukan model yang belum ada untuk mata kuliah peminatan bagi mahasiswa STMIK Bani Saleh, sehingga mahasiswa dapat memahami dan mengerti mata kuliah apa yang ambil pada saat menempuh semester berikutnya dan melakukan pengukuran akurasi kehandalan dalam menentukan mata kuliah peminatan.

Klasifikasi merupakan salah satu tugas yang penting dalam data mining. Mengorganisasikan serta mengelompokkan data ke dalam kelas-kelas yang berbeda merupakan tujuan utama dari klasifikasi. Definisi pengklasifikasian adalah sebuah fungsi yang bersifat prediksi dan menggolongkan data item tertentu ke dalam sebuah kelas. Sebuah pengklasifikasian dibuat dari sekumpulan data latih dengan kelas yang ditentukan dan dikenal ciri-cirinya sebelumnya. Performa pengklasifikasian biasanya diukur dengan ketepatan [6].



Gambar 1: Blok diagram model klasifikasi

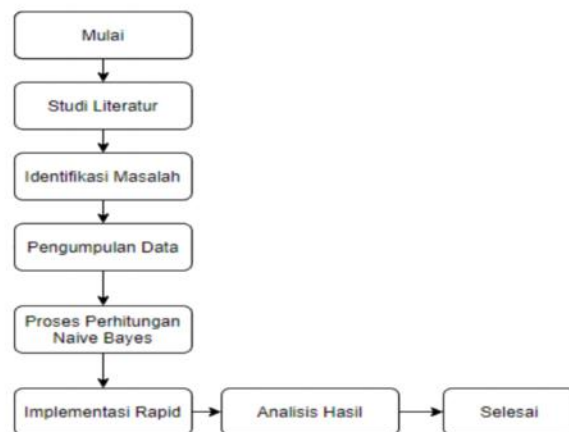
Penelitian ini diharapkan dapat memberikan pengetahuan baru bagi pembaca mengenai data mining dengan metode naive bayes yang diterapkan untuk menentukan mahasiswa yang tepat dan sesuai kemampuannya dalam peminatan. Penelitian yang dilakukan oleh Normah pada tahun 2019, studi ini menganalisis sentimen pengguna aplikasi *Windows Phone Store* dengan secara otomatis mengklasifikasikan review ke dalam kategori opini positif atau negatif [5]. *Naïve Bayes* memiliki potensi yang baik karena kesederhanaan dan performanya sebagai model pengklasifikasian teks pada banyak domain. Model dievaluasi menggunakan 10 *Fold Cross Validation*. Pengukuran dilakukan dengan *Confusion Matrix* dan kurva ROC. Akurasi yang dihasilkan pada

penelitian ini adalah 84,50% yang menunjukkan bahwa *Naïve Bayes* merupakan model yang baik dalam mengklasifikasikan teks terutama dalam hal analisis sentimen.

Penelitian yang dilakukan oleh Wiwit Supriyanti, Kusri, Armadyah Amborowati pada tahun 2016 menggunakan metode forward selection mampu menghasilkan tingkat akurasi yang lebih baik dibandingkan dengan tanpa penambahan seleksi fitur yang hanya mencapai tingkat akurasi pada algoritma C4.5 dari sebelumnya sebesar 84,43% meningkat menjadi 84,98%, sedangkan pada algoritma naive bayes sebelumnya sebesar 78,47% meningkat menjadi 82,01%. Hasil dari komparasi algoritma klasifikasi antara decision tree C4.5, dan naive bayes yang digabungkan dengan metode seleksi fitur forward selection untuk kasus ketepatan pemilihan konsentrasi mahasiswa didapatkan tingkat akurasi tertinggi dengan algoritma terpilih C4.5 dengan nilai akurasi sebesar 84,98% [4].

Metode Penelitian

Alur penelitian terlihat pada Gambar 2.



Gambar 2: Diagram alir tahapan penelitian

Studi Literatur

Peminatan mata kuliah termaksud usaha untuk memperdalam kemampuan mahasiswa tentang mata kuliah yang memang perlu dikembangkan. Penentuan peminatan mata kuliah tentunya tidak terlepas dari penguasaan mahasiswa terhadap mata kuliah yang menjadi inti dari peminatan tersebut.

Identifikasi Masalah

Masalah dalam penelitian ini adalah:

1. Bagaimana menentukan model yang belum ada untuk mata kuliah peminatan bagi mahasiswa STMIK Bani Saleh, sehingga mahasiswa dapat memahami dan mengerti mata kuliah apa yang mereka ambil pada saat mereka menempuh semester berikutnya?
2. Bagaimana melakukan pengukuran akurasi kehandalan dalam menentukan matakuliah peminatan?

Pengumpulan Data

Metode pengumpulan data dilakukan untuk memperoleh informasi yang dibutuhkan dalam rangka mencapai tujuan penelitian. Adapun pengumpulan data yang digunakan dalam penelitian ini adalah: ,digunakan dalam penelitian ini data program studi Teknik Informatika tahun angkatan 2013, 2014, 2015 & 2016. Jumlah sample penelitian untuk kriteria matakuliah peminatan adalah 824 sebagai data training dan 206 sebagai data testing dengan perbandingan data 80/20 yang meliputi mata kuliah : kriteria matakuliah peminatan, logika algoritma, sistem basis data, dan analisa perancangan sistem informasi dan kriteria matakuliah peminatan jaringan adalah : sistem digital, pengantar teknologi informasi, jaringan komputer 1.

Klasifikasi *Naive Bayes* untuk menghitung probabilitas suatu peristiwa dalam langkah-langkah berikut:

1. Hitung probabilitas sebelumnya untuk label kelas yang diberikan
2. Temukan probabilitas Peluang dengan setiap atribut untuk setiap kelas
3. Masukkan nilai ini dalam Formula *Bayes* dan hitung probabilitas posterior.
4. Lihat kelas mana yang memiliki probabilitas lebih tinggi, mengingat input milik kelas probabilitas lebih tinggi

Proses Perhitungan *Naïve Bayes*

Perhitungan *Naïve Bayes* dengan menggunakan data-data di bawah ini :

1. Data sebelum pengolahan data mining

Perhitungan menggunakan dengan contoh data mahasiswa seperti pada Tabel 1 adalah :

Tabel 1: Contoh Data KHS Mahasiswa

NPM	IPK	Kd_Dosen PA
43A87006150xxx	3.09	0410116207
43A87006150xxx	3.38	0410116207
43A87006150xxx	3.52	0410116207
43A87006150xxx	3.16	0410116207
43A87006150xxx	3.03	0410116207
43A87006150xxx	2.57	0410116207
43A87006150xxx	2.89	0410116207

Data pada Tabel 2 diambil objek yaitu data mahasiswa STMIK Bani Saleh angkatan 2013, 2014, 2015 dan 2016. Terdiri dari atribut Nim, IPK, Kd_dosen PA. Data yang digunakan dalam penelitian antara lain diperoleh dari buku, jurnal, internet dan sebagainya. Berikut ini pada Tabel 2 contoh data KHS mahasiswa, Tabel 3 contoh data transkrip nilai mahasiswa, Tabel 4 Inilialisasi Pada Nilai Matakuliah Tabel dibawah ini merupakan contoh data yang didapatkan dari STMIK Bani Saleh.

Tabel 2: Contoh Data KHS Mahasiswa

Tahun Angkatan	NPM	IPS
2015	43A 87006150xxx	2.84
2015	43A 87006150xxx	3.25
2015	43A 87006150xxx	3.26
2015	43A 87006150xxx	3.00
2015	43A 87006150xxx	3.00
2015	43A 87006150xxx	3.10
2015	43A 87006150xxx	3.16

Tabel 3: Contoh Data KHS Mahasiswa

NPM	Kd Matakuliah	Nilai
43A87006150xxx	1506121258	B
43A87006150xxx	1506122251	A
43A87006150xxx	15061111311	B
43A87006150xxx	15061111220	A
43A87006150xxx	15061111411	B
43A87006150xxx	15061111307	C
43A87006150xxx	15061211310	B
43A87006150xxx	12306001	B

Tabel 4: Contoh Data KHS Mahasiswa

Nilai	Angka	Keterangan
A	4	Unggul
B	3	Baik
C	2	Cukup
D	1	Kurang

2. Proses Pengelolaan Data Menjadi Data Mining

Proses data mining terbagi beberapa tahap :

(a) **Data Selection**

Pada tahap selection ini data yang diambil adalah relevan untuk di analisis yang diambil dari database. Yang diperlukan pada data selection meliputi: NPM, Gender, Jurusan Asal Sekolah , Matakuliah peminatan (Sistem Digital, PTI, Pengantar Teknologi Informatika, Jaringan Komputer 1, Logika Algoritma, Sistem Basis Data,Analisa Perancangan Sistem Informasi), Peminatan (Software Engineering & Networking).

(b) **Cleaning Data**

Pada sampel data sebelum dilakukan Cleaning Data terdapat 85 data setelah dilakukan Cleaning menghilangkan noise dan data yang konsisten menjadi 1030 data. Banyaknya data yang tidak diisi dengan semestinya. Data yang awalnya diambil dari NPM, Gender, Jurusan Asal Sekolah, Nilai Mata kuliah, dan Peminatan.

(c) **Data Transformation**

Mengubah data ke dalam bentuk data training dan data testing yang sesuai dengan model yang digunakan dalam rapidminer yang akan di uji.

3. Proses Perhitungan Dengan Algoritma Naïve Bayes

Proses Perhitungan dengan Algoritma Naïve Bayes Adalah metode memprediksi peluang di masa depan berdasarkan pengalaman dimasa sebelumnya. Naïve Bayes merupakan pengklasifikasian statistik yang dapat digunakan un-

tuk memprediksi probabilitas keanggotaan suatu class. Berikut contoh kasus yang akan di selesaikan dengan metode *Naïve Bayes*:

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

a. **Menghitung jumlah kelas/Tabel**

Dari jumlah total data set 824 terdapat 416 label “Networking” dan 408 label “Software Engineering

Tabel 5: Label Peminatan

	Peminatan	Jumlah Mhs
P(Y SoftwareEngineering)	Software Engineering	408
P(Y Networking)	Networking	416

Tabel 6: Jumlah Jenis Kelamin Masing-masing peminatan

Jenis Kelamin	Networking	Software Engineering
Laki – laki (L)	364	322
Perempuan (P)	52	86

Tabel 7: Jurusan Asal Sekolah

Asal Sekolah	Networking	Software Engineering
Administrasi	0	4
Akuntansi	4	12
Elektro	71	64
IPA	83	82
IPS	18	64
Listrik	12	6
Mesin	17	16
Multimedia	1	15
RPL	0	25
TKJ	189	66
TKR	3	11

Tabel 8: Matakuliah Sistem Digital

Sistem Digital	Networking	Software Engineering
1	6	7
2	40	51
3	237	238
4	130	112

Tabel 9: Matakuliah PTI

PTI	Networking	Software Engineering
1	3	6
2	44	50
3	203	213
4	161	137

Tabel 10: Matakuliah Jarkom 1

Jarkom 1	Networking	Software Engineering
1	5	8
2	53	47
3	211	254
4	144	98

Tabel 11: Matakuliah Logika Algoritma

Logika Algoritma	Networking	Software Engineering
1	9	8
2	57	61
3	199	177
4	146	161

Tabel 12: Matakuliah Sistem Basis Data

Tahun Angkatan	NPM	IPS
2015	43A87006150xxx	2.84
2015	43A87006150xxx	3.25
2015	43A87006150xxx	3.26
2015	43A87006150xxx	3.00
2015	43A87006150xxx	3.00
2015	43A87006150xxx	3.10
2015	43A87006150xxx	3.16

Tabel 13: Matakuliah APSI

APSI	Networking	Software Engineering
1	7	4
2	7	4
3	181	204
4	167	141

Tabel 14: Jurusan Asal Sekolah

No	NPM	Gender (L/P)	Jurusan Sekolah	SD	PTI	Jarkom 1	LA	SBD	APSI	Peminatan Sistem
1	43A8700615xxx	L	TKJ	3	2	4	3	2	4	?

Keterangan:

SD = Sistem Digital

PTI = Pengantar Teknologi Informasi

Jarkom1 = Jaringan Komputer1

LA = Logika Algoritma

SBD = Sistem Basis Data

APSI = Analisa Perancangan Sistem Informasi

b. Menghitung $P(X|Ci)$ untuk setiap class I Networking & Software Engineering

Gander ($k = 1$)

= $P(\text{Gender} = L \mid \text{Peminatan} = \text{Networking})$

= $364/416$

= $0,875$

= $P(\text{Gender}=L \mid \text{Peminatan} = \text{Software Engineering})$

= $322/408 = 0,789215686$

Asal Sekolah ($k = 1$) = $P(\text{Asal Sekolah} = \text{TKJ} \mid \text{Peminatan} = \text{Networking})$

= $189/416$

= $0,454326923$

= $P(\text{Asal Sekolah} = \text{TKJ} \mid \text{Peminatan} = \text{Software Engineering}) = 66/408 = 0,161764706$

Matakuliah = $P(\text{Sistem Digital} = 3 \mid \text{Peminatan} = \text{Networking})$

= $237/416$

= $0,569711538$

= $P(\text{Sistem Digital} = 3 \mid \text{Peminatan} = \text{Software Engineering}) = 238/408 = 0,583333333$

= $P(\text{PTI} = 2 \mid \text{Peminatan} = \text{Networking})$

= $44/416 = 0,105769231$

= $P(\text{PTI} = 2 \mid \text{Peminatan} = \text{Software Engineering}) = 50/408 = 0,12254902$

= $P(\text{Jarkom1} = 4 \mid \text{Peminatan} = \text{Networking}) = 144/416 = 0,346153846$

= $P(\text{Jarkom1} = 4 \mid \text{Peminatan} = \text{Software Engineering}) = 98/408 = 0,240196078$

= $P(\text{Logika Algoritma} = 3 \mid \text{Peminatan} = \text{Networking}) = 199/416 = 0,478365385$

= $P(\text{Logika Algoritma} = 3 \mid \text{Peminatan} = \text{Software Engineering}) = 177/408 = 0,433823529$

= $P(\text{Sistem Basis Data} = 2 \mid \text{Peminatan} = \text{Networking}) = 69/416 = 0,165865385$

= $P(\text{Sistem Basis Data} = 2 \mid \text{Peminatan} = \text{Software Engineering}) = 59/408 = 0,144607843$

= $P(\text{APSI} = 4 \mid \text{Peminatan} = \text{Networking}) = 167/416 = 0,401442308$

= $P(\text{APSI} = 4 \mid \text{Peminatan} = \text{Software Engineering}) = 141/408 = 0,345588235$

c. Mengalikan semua cabang

Pada perkalian semua cabang nilai nya dikalikan 10000, karena angka nol koma terlalu banyak :

$$\begin{aligned}
 &1. \text{ Networking} = \text{Gender} * \text{Asal Sekolah} * \\
 &\text{Sistem Digital} * \text{PTI} * \text{Jarkom1} * \text{Logika Algoritma} * \\
 &\text{Sistem Basis Data} * \text{APSI} = 0,875 * 0,454326923 * \\
 &0,569711538 * 0,105769231 * 0,346153846 * \\
 &0,478365385 * 0,165865385 * 0,401442308 \\
 &= 0,000264118 \\
 &= 0,000264118 * 416/824 \\
 &= 0,000133341 \\
 &= 0,000133341 * 10000 \\
 &= 1,333413398
 \end{aligned}$$

$$\begin{aligned}
 &2. \text{ Software Engineering} = \text{Gander} * \text{Asal Sekolah} * \\
 &\text{Sistem Digital} * \text{PTI} * \text{Jarkom1} * \text{Logika Algoritma} * \\
 &\text{Sistem Basis Data} * \text{APSI} = 0,789215686 * 0,161764706 * \\
 &0,583333333 * 0,12254902 * 0,240196078 * 0,433823529 * \\
 &0,144607843 * 0,345588235 \\
 &= 4,75265E-05 \\
 &= 4,75265E-05 * 408 * 824 \\
 &= 2,35325E-05 \\
 &= 2,35325E-05 * 10000 \\
 &= 0,235325441
 \end{aligned}$$

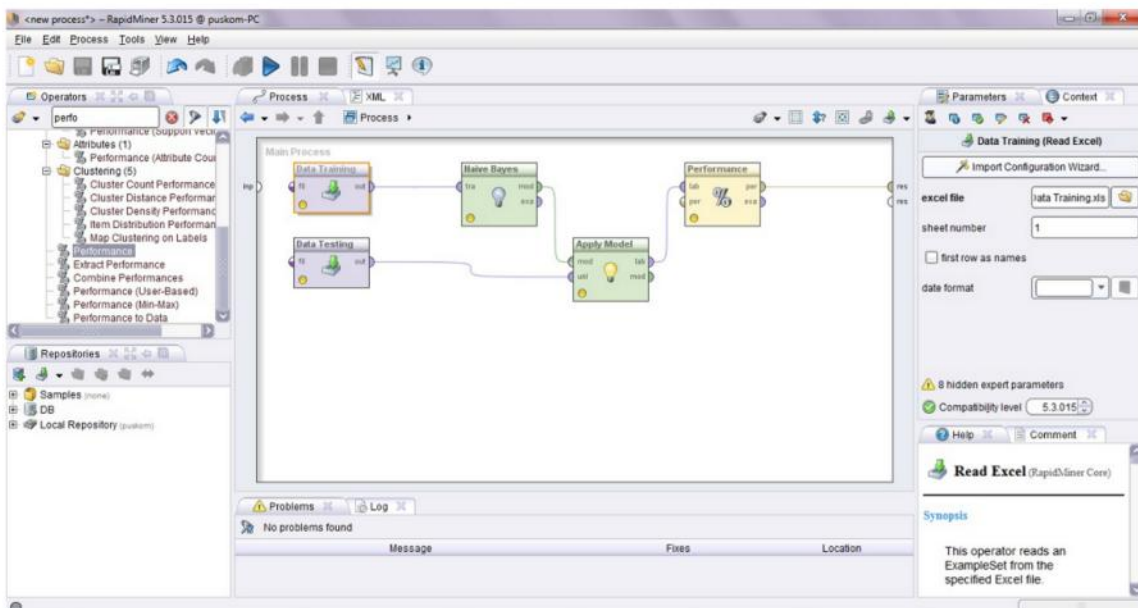
Untuk $P(X \mid \text{Peminatan} = \text{Networking}) = 1,333$ dan $P(X \mid \text{Peminatan} = \text{Software Engineering}) = 0,235$. Karena nilai yang terbesar adalah class Software Engineering, maka untuk data testing, akan memiliki class Networking.

Hasil dan Pembahasan

Dalam tahap penelitian ini hasil dari perhitungan menggunakan Ms. Excel yang di ambil dari mahasiswa jurusan Teknik Informatika berjumlah 1030 dengan perbandingan 80/20 adapun pembagian data 824 record sebagai data training dan 206 record sebagai data testing. Dibawah ini akan dijelaskan akan langkah-langkah dari pemodelan algoritma Naïve Bayes dengan menggunakan aplikasi RapidMiner 5.3.

Memasukan Operator Naïve Bayes

Setelah melakukan import data training dan data testing pada aplikasi rapid miner selanjutnya adalah memasukan operator Naïve Bayes pada halaman design rapid miner. Juga memasukan operator apply model untuk pre-processing model algoritma pada data testing serta memasukan operator performance untuk mengevaluasi keakuratan model algoritma tersebut.



Gambar 3: Model naive bayes

Hasil Dari Data View

Merupakan view sheet untuk menampilkan data pada rapidminer, juga menghasilkan Tabel data testing dengan data hasil prediksi algo-

ritma Naïve Bayes dengan menampilkan 206 contoh, 4 special attribute, dan 10 reguler attribute. Beserta dengan hasil nilai confidence Software Engeengering dan nilai confidence Networking

Gambar 4: Data View

Hasil Algoritma Naïve Bayes

Langkah berikutnya adalah menjalankan proses algoritma Naïve Bayes untuk mengetahui hasil

akurasi pada rapidminer. Berikut merupakan hasil akurasi prediksi yang tepat sebesar 98,06% dari 206 data testing yang dapat dilihat pada gambar 5.

	true Software Engineering	true Networking	class precision
pred. Software Engineering	93	4	95.88%
pred. Networking	0	109	100.00%
class recall	100.00%	96.45%	

Gambar 5: Hasil Akurasi naive bayes

Penutup

Algoritma yang digunakan untuk mengukur tingkat akurasi peminatan penjurusan Program Studi Teknik Informatika dengan jumlah data training sebanyak 824 dan data testing sebanyak 206 dengan menggunakan *Naïve Bayes*. Dengan memunculkan 2 peminatan yaitu *Software Engineering* dan *Networking*. Adapun aplikasi yang digunakan untuk membantu proses perhitungan *Naïve Bayes Rapidminer* 5.3. Dari hasil pengukuran menggunakan rapidminer diperoleh nilai akurasi sebesar 98.06%, Merupakan *view sheet* untuk menampilkan data pada *Rapidminer*, juga menghasilkan Tabel data testing dengan data hasil prediksi algoritma *Naïve Bayes* dengan menampilkan 206 contoh, 4 *special attribute*, dan 10 *reguler attribute*. Beserta dengan hasil nilai *confidence Software Engineering* dan nilai *confidence Networking*.

Daftar Pustaka

- [1] Pan-Ning Tan, Michael Steinbach and Vipin Kumar, "Introduction to Data Mining: Pearson New International Edition", Pearson New International Edition, ISBN-13: 9781292026152, 2013.
- [2] Irina Rish, "An empirical study of The Naive Bayes Classifier", Technical Report T.J. Watson Research Centre, 2001.
- [3] P. Chapman, J. Clinton, R. Kerber, T. Khabaza, T. Reinartz, C. Shearer and R. Wirth, "CRISP-DM v.1.0 Step-by-step data mining guide", The CRISP-DM consortium, 2000.
- [4] W. Supriyanti, Kusri dan A. Amborowati, "Perbandingan Kinerja Algoritma c4.5 Dan Naive Bayes Untuk Ketepatan Pemilihan Konsentrasi Mahasiswa", J. Inf. Politek. Indonusa, vol. 1, no. 3, pp. 61–67, 2016.
- [5] N. Normah, "Naïve Bayes Algorithm For Sentiment Analysis Windows Phone Store Application Reviews," Sinkron, vol. 3, no. 2, p. 13, doi: 10.33395/sinkron.v3i2.242, 2019.
- [6] Edgar Winata, Johan Setiawan, "Analisis dan Perancangan Prototipe Aplikasi Tracking Bis Universitas Multimedia Nusantara pada Platform Android", ULIMA InfoSys, Vol. IV, No. 1 Juni ISSN 2085-4579, 2013
- [7] P. Chandrasekar and K. Qian, "The Impact of Data Preprocessing on the Performance of a Naive Bayes Classifier", IEEE 40th Annual Computer Software and Applications Conference (COMPSAC), 618–619, <http://doi.org/10.1109/COMPSAC.2016.205>, 2016.
- [8] S. Agarwal, "Data mining: Data mining concepts and techniques", In Proceedings - 2013 International Conference on Machine Intelligence Research and Advancement, ICMIRA, 2013.
- [9] J. Han and M. Kamber, "Data mining: Concepts and Techniques", USA:Academic Press. 2001.
- [10] B. Hernández and D. B. Hidalgo, "Findings Seminal Papers Using Data Mining Techniques", Open J. Soc. Sci., vol. 08, no. 09, pp. 293–305, doi: 10.4236/jss.2020.8902, 2020.