MJRICT 2019;2(1) : (26-34)

MJRICT : Musamus Journal Of Research Information and Communication Technology

ISSN 2655-5735 (Online) | ISSN 2654-4083 (Cetak)

https://ejournal.unmus.ac.id/index.php/mjrict

Penerapan Metode *Mutual Information* Dan *Bayes Network* Untuk Klasifikasi Penyelesaian Studi

Chusnul Chotimah
Teknik Informatika, Universitas Musamus, Merauke, Indonesia
email : cchotimah@unmus.ac.id

Abstrak. Data jumlah mahasiswa yang lulus tiap tahunnya tidak sebanding dengan jumlah yang mendaftar. Penyelesaian studi setiap mahasiswa dapat disebabkan atau dipengaruhi oleh banyak faktor. Kemungkinan pada beberapa faktor memiliki hubungan (kausalitas) satu dengan yang lain. *Mutual Information* digunakan untuk menghitung kausalitas antar faktor yang mempengaruhi penyelesaian studi. Hubungan antar faktor untuk membangun model pada penelitian ini digunakan metode *Bayes Networs* (BN). Metode *Bayes Networks* merupakan metode pemodelan data berbasis probabilitas yang merepresentasikan suatu himpunan variabel dan *conditional dependency* melalui *Directed Acyclic Graph* (DAG). Hasil pengujian dari sistem yang dikembangkan mengunakan data uji sebanyak 128 memiliki tingkat akurasi sebesar 71,09%. Hasil akurasi sistem lebih tinggi dibanding dengan menggunakan metode *Naive Bayes Classifier* yaitu sebesar 67,97%.

Kata kunci: Penyelesain Studi, Bayes Network, Mutual Information.

1. Pendahuluan

Mahasiswa menghadapi banyak tantangan dalam menyelesaikan studi. Hampir dibanyak lembaga pendidikan tinggi terjadi kesenjangan antara jumlah pendaftar dengan jumlah lulusan. Data jumlah yang lulus tiap tahunnya tidak sebanding dengan jumlah yang mendaftar. Permasalahan tersebut menjadi perhatian yang harus diselesaikan oleh para *stakeholder*, sehingga pemantauan dan evaluasi terhadap kecenderungan penyelesaian studi tepat waktu (TW) atau tidak tepat waktu (TTW) perlu dilakukan.

Salah satu indikator keberhasilan proses belajar mengajar mahasiswa tiap semester adalah indeks prestasi (IP) maupun indeks kumulatif (IPK) [1]. Faktor-faktor lain yang mempengaruhi penyelesaian studi mahasiswa adalah faktor non akademis diantaranya jenis kelamin, asal suku [2], asal sekolah, jurusan [3], dan status mahasiswa [4]. Kumpulan data kelulusan akan diklasifikasi menurut kelas masing-masing. Klasifikasi akan menentukan apakah penyelesaian studi atau kelulusan tersebut merupakan kelulusan TW atau TTW.

Mutual Information digunakan untuk menghitung kausalitas antar faktor yang mempengaruhi penyelesaian studi. Hubungan antar faktor untuk membangun model pada penelitian ini digunakan metode *Bayes Networs* (BN). Metode *Bayes Networks* merupakan metode pemodelan data berbasis probabilitas yang merepresentasikan suatu himpunan variabel dan *conditional dependency* melalui *Directed Acyclic Graph* (DAG) [5]. Pendekatan ini digunakan karena dapat menggambarkan

hubungan sebab-akibat antara ketepatan waktu penyelesaian studi dengan faktor-faktor yang diduga berpengaruh terhadap penyelesaian studi.

2. Metode Penelitian

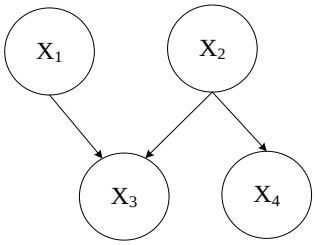
2.1 Teori Bayes

Dalam probabilitas terdapat teori penting yang mendasari berbagai macam analisa statistika dalam berbagai bidang yaitu teori *bayes*. Teori *bayes* merupakan teori yang mengacu pada konsep probabilitas bersyarat [6]. Prinsip dalam teori *bayes* adalah digunakan *prior probability* ketika tidak ada informasi lain yang dapat digunakan untuk melihat kemungkinan terjadinya suatu kejadian, tetapi begitu informasi baru diketahui maka probabilitas yang baru harus dilihat berdasarkan informasi yang baru diketahui tersebut. Probabilitas jenis ini disebut probabilitas bersyarat (*conditional probability*). Pernyataan probabilitas bersyarat ditulis dalam notasi *P(Y | X)* dengan persamaan :

$$\boldsymbol{P(Y|X)} = \frac{\boldsymbol{P(X|Y)P(Y)}}{\boldsymbol{P(X)}} \quad (1)$$

2.2 Bayes Network

Bayes Network merupakan *probabilistic graphical model* (PGM) dengan *edge* berarah yang digunakan untuk merepresentasikan pengetahuan tentang hubungan kebergantungan *(dependency)* atau kebebasan *(independency)* di antara variabel-variabel dari domain persoalan yang dimodelkan [7]. Variabel direpresentasikan dengan *node* dari suatu graf yakni *directed acyclic graph* (DAG) dan *edge* yang merepresentasikan hubungan ketergantungan antar *node* seperti pada Gambar 1. Komponen utama *Bayes Network* meliputi *Direct Acyclic Graf* (DAG) dan *Conditional Probability Table* (CPT).



Gambar 1 Contoh struktur sebuah DAG

Struktur ketergantungan yang digambarkan dengan DAG tersebut diinterpresentasikan dengan fungsi kepekatan bersama (*join probability distribution*/JPD), di mana struktur tersebut kemungkinan terdiri dari *n* variabel/*node*. Jika keseluruhan *node* yang terdapat pada *Bayes Network* adalah $\{x_i, i=1,...,n\}$ dan *parent*(x_i) mengambarkan himpunan *parent* dari x_i maka JPD P(x_i, i=1,...,n) adalah dengan mengalikan semua probabilitas berdasarkan *parent*-nya [6]:

$$\boldsymbol{P(x_1, \dots x_n)} = \prod_{\boldsymbol{i=1}}^{\boldsymbol{n}} \boldsymbol{P\left(x_i \mid parent(x_i)\right)} \quad (2)$$

Sebagai contoh JPD untuk struktur yang terdapat pada Gambar 1 adalah :

$$P(X_1, X_2, X_3, X_4) = P(X_1).P(X_2).P(X_3| X_1, X_2).P(X_4| X_2)$$

Kekuatan hubungan diantara variabel (*node*) diperlihatkan dengan nilai probabilitas. Terdapat beberapa langkah dalam membuat *Bayes Network*, antara lain :

1. Menentukan variabel yang relevan dan menyusunnya. Kemudian tentukan hubungan antar variabel pada *Bayes Network* yang disebut dengan *parent nodes.*
2. Menentukan nilai pada setiap variabel
3. Memperkirakan *conditional probability* dari setiap hubungan yang ada.

2.3 Mutual Information

Mutual information (MI) merupakan bagian dari *information theory* yang digunakan untuk menentukan besaran aliran informasi antar variabel. *Mutual information* dapat merepresentasikan interaksi yang terjadi antar variabel yang ada pada suatu sistem. *Mutual information* antara dua *node* A dan B didefinisikan [8] dengan persamaan :

$$\boldsymbol{I(A,B) = \sum_{a,b} p(a,b) log \frac{p(a,b)}{p(a)p(b)}} \quad (3)$$

Sedangkan *conditional mutual information* adalah *mutual information* berdasarkan nilai yang terdapat pada *node* C dapat ditulis dalam bentuk persamaan :

$$\boldsymbol{I(A,B|C) = \sum_{a,b,c} p(a,b|c) log \frac{p(a,b|c)}{p(a|c)p(b|c)}} \quad (4)$$

Secara formal, *mutual information* membandingkan peluang pengamatan *a* dan *b* secara bersama dengan peluang pengamatan *a* dan *b* secara bebas. Apabila nilai I(*a*,*b*) ≈ 0 (diwakilkan dengan sebutan nilai *threshold*). Nilai *threshold ε* yang dianjurkan [8] adalah 0,01, sehingga algoritma ini mengatakan bahwa A dan B akan *independent* jika I(a,b) < *ε*, dimana *ε*=0,01. Semakin tinggi nilai *mutual information* antara dua *node*/variabel, mengindikasikan adanya interaksi yang ditimbulkan dari *node-node* yang dihubungkan tersebut.

2.4 Akurasi

Evaluasi pada penelitian ini menggunakan *confusion matrix* [6], sehingga bisa dihitung besarnya tingkat akurasi dari proses klasifikasi. *Confusion matrix* terbagi menjadi 4, yaitu *True Negative, False Positive, False Negative, dan True Positive*. Penjelasan masing-masing bagian terdapat pada Tabel 1.

Tabel 1 *Confusion matrix*

	Predicted	
	Negative	*Positive*
Actual negative	*TN*	*FP*
Actual positive	FN	*TP*

keterangan :

- **TN adalah jumlah prediksi negatif untuk data uji negatif**
- **FP adalah jumlah prediksi positif untuk data uji negatif**
- **FN adalah jumlah prediksi negatif untuk data uji positif**
- **TP adalah jumlah prediksi positif untuk data uji positif**

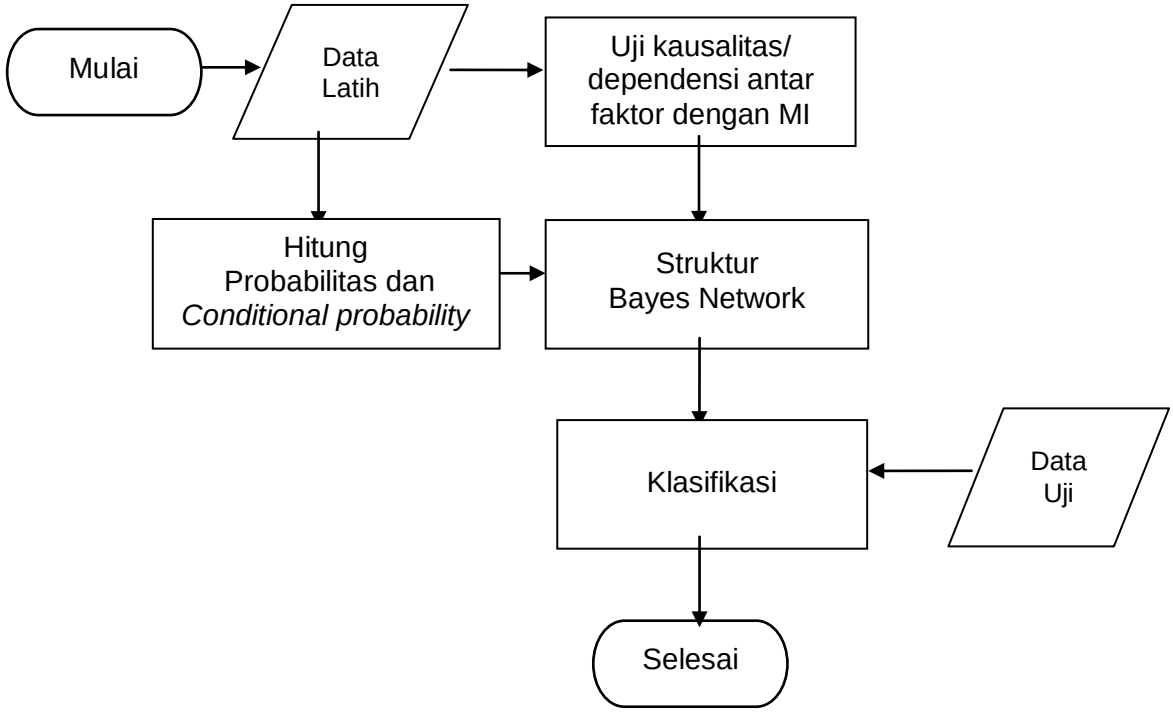
Akurasi merupakan proporsi dari jumlah total prediksi benar terhadap semua data dan dihitung dengan persamaan

$$\boldsymbol{Akurasi = \frac{TP + TN}{TN + FP + FN + TP}} \quad (5)$$

3. Hasil dan Pembahasan

3.1 Deskripsi Sistem

Implementasi sistem menggunakan pendekatan *Bayes Network* untuk klasifikasi. Data yang digunakan perlu melalui proses pembersihan dan pengubahan bentuk. Tujuan dilakukannya kegiatan ini adalah untuk mentransformasikan data ke format yang sesuai, sehingga data tersebut siap digunakan. Setelah didapat data training, selanjutnya akan dihitung kausalitas antar faktor/variabel menggunakan *Mutual Information*. Penghitungan kausalitas dengan *Mutual Information* akan menghasilkan struktur *Bayes Network* atau model penyelesaian studi untuk proses klasifikasi. Proses rancangan sistem digambarkan pada Gambar 2 berikut.



Gambar 2 Rancangan Sistem *Bayes Network*

3.2 Penjelasan Rancangan Sistem

1. Mutual Information (MI)

Uji validitas dependensi antar faktor dengan data empiris menggunakan *Mutual Information*. Proses perhitungan menggunakan dengan persamaan (3) dan persamaan (4) dicontohkan sebagai berikut:

Prior probability :

Prior jkel+(jkel=laki-laki) = 0,48

Prior jkel-(jkel=perempuan) = 0,52

Prior suku++(suku=non papua) = 0,87

Prior suku+-(suku=papua) = 0,11

Prior suku--(suku=marind) = 0,02

Conditional probability :

jkel+suku++(jkel=laki-laki, suku=non papua)=0,41

jkel+suku+-(jkel=laki-laki, suku=papua)=0,05

jkel+suku--(jkel=laki-laki, suku=marind)=0,01

jkel-suku++(jkel=perempuan, suku=non papua)=0,45

jkel-suku+-(jkel= perempuan, suku=papua)=0,06

jkel-suku--(jkel= perempuan, suku=marind)=0,01

I(jkel,suku) = I(jkel+,suku++) + I(jkel+, suku+-) + I(jkel+, suku--)
+ I(jkel-,suku++) + I(jkel-, suku+-) + I(jkel-, suku--)

$$= \left(0{,}41 \times \log\left(\frac{0{,}41}{0{,}48\times 0{,}87}\right)\right) + \left(0{,}05 \times \log\left(\frac{0{,}05}{0{,}48\times 0{,}11}\right)\right)$$

$$+ \left(0{,}01 \times \log\left(\frac{0{,}01}{0{,}48\times 0{,}02}\right)\right) + \left(0{,}45 \times \log\left(\frac{0{,}45}{0{,}52\times 0{,}87}\right)\right)$$

$$+ \left(0{,}06 \times \log\left(\frac{0{,}06}{0{,}52\times 0{,}11}\right)\right) + \left(0{,}01 \times \log\left(\frac{0{,}01}{0{,}52\times 0{,}02}\right)\right)$$

$$= (-0{,}00044) + (-0{,}00020) + (0{,}00068)$$
$$+ (0{,}00044) + (0{,}00020) + (-0{,}00059)$$

= 0,00005

Hasil Uji dependesi antar faktor ditunjukkan pada Tabel 2 berikut.

Tabel 2 Hasil uji dependensi dengan *Mutual Information*

Faktor	as	js	st	jkel	suku	ipk	ps
as		0,02056**	0,00027*	0,00025*	0,00030*	0,00003*	0,00004*
js	0,02056**		0,00016*	0,00054*	0,00038*	0,00193*	0,01212**
st	0,00027*	0,00016*		0,00000*	0,00076*	0,00297*	0,01219**
jkel	0,00025*	0,00054*	0,00000*		0,00005*	0,00025*	0,01313**
suku	0,00030*	0,00038*	0,00076*	0,00005*		0,00122*	0,01526**
ipk	0,00003*	0,00193*	0,00297*	0,00025*	0,00122*		0,01820**

*. Nilai MI dibawah *threshold*

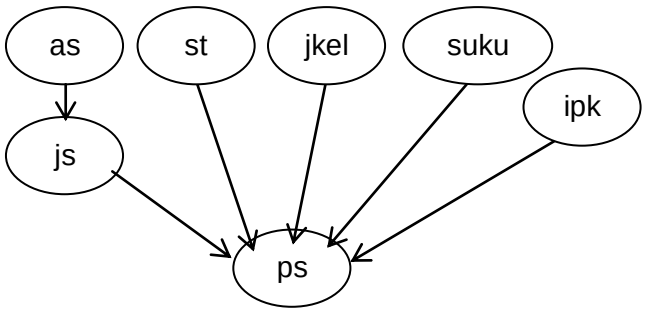
**. Nilai MI sesuai *threshold*

Berdasar Tabel 2 dapat disimpulkan bahwa faktor jurusan sekolah, status mahasiswa, jenis kelamin, suku dan ipk memiliki interaksi dengan penyelesaian studi hal tersebut dapat dilihat dari hasil nilai *mutual information* bernilai lebih dari nilai *threshold*. Sedang faktor asal sekolah tidak memiliki interaksi dengan penyelesaian studi, hal tersebut dilihat dari nilai *mutual information* bernilai kurang dari nilai *threshold*. Namun asal sekolah memiliki interaksi dengan jurusan.

2. Bayes Network (BN)

Directed Acyclic Graph (DAG) yang merupakan representasi dari *node* dan *edge. Node* adalah variabel-variabel yang terdapat pada model. Pada setiap *node* terdapat CPD (*Conditionally Probabilistic Distribution*) atau nilai probabilitas pada variabel pada *node* yang dipengaruhi oleh

variabel orang tuanya. Sedangkan *edge* adalah anak panah yang menunjuk dari *node* ke *node* yang lain, dimana *node* asal adalah orang tua dari *node* yang ditunjuk. DAG ini tidak boleh *Cyclic* atau berputar kembali. Berdasar analisis perhitungan dengan MI maka DAG (*Directed Acyclic Graph*) penyelesaian studi ditunjukkan pada Gambar 3. Nilai pada setiap variabel merupakan suatu nilai real antara 0 sampai dengan 1.



Gambar 3 DAG Rancangan Bayes Network

Pada DAG diatas menunjukkan **ps** merupakan variabel kelas *output*, perhitungan setiap *node input* menggunakan persamaan (2) nantinya akan mempengaruhi hasil dari kelas yang akan dicari.

3. Penghitungan nilai probabilitas

Probabilitas setiap variabel akan dihitung dengan melihat frekuensi nilai pada tiap variabel (*node*) dan juga nilai dari kombinasi variabel mempengaruhinya. Untuk menghitung probabilitas pada variabel yang tidak dipengaruhi adalah dengan membagi jumlah frekuensi dengan jumlah totalnya, sementara pada variabel yang dipengaruhi cara menghitung probabilitas adalah menggunakan persamaan (2).

Tabulasi nilai probabilitas dengan menggunakan data empiris disajikan pada tabel 3 dan tabel 4.

Tabel 3 Nilai Probabilitas (P) setiap *node*

	P(st)
st=bl	0,90
st=me	0,10

	P(suku)
suku=np	0,87
suku=pa	0,11
Suku=ma	0,02

	P(as)
as=sma	0,83
as=smk	0,17

	P(jkel)
jkel=l	0,48
jkel=p	0,52

	P(ipk)
ipk=mm	0,17
ipk=sm	0,76
Ipk=cum	0,06

Tabel 4 Nilai Probabilitas bersyarat pada *node* P(**js|as**) dan P**(ps|js,st,jkel,suku,ipk)**

js	**as**	**P(js|as)**
ipa	sma	0,47
ipa	smk	0,44
ips	sma	0,44
ips	smk	0,44
lain	sma	0,09
lain	smk	0,39

node					**P(ps=tw)**	**P(ps=ttw)**
js	**st**	**jkel**	**suku**	**ipk**		
ipa	bl	l	np	mm	0,36	0,64
ips	bl	l	np	mm	0,33	0,67
lain	bl	l	np	mm	0,33	0,67
:	:	:	:	:	:	:
:	:	:	:	:	:	:
:	:	:	:	:	:	:
ipa	bl	p	np	sm	0,51	0,49
ips	bl	p	np	sm	0,55	0,45
lain	bl	p	np	sm	0,47	0,53
:	:	:	:	:	:	:
:	:	:	:	:	:	:
:	:	:	:	:	:	:
ipa	bl	p	np	cum	1,00	0,00
ips	bl	p	np	cum	1,00	0,00
lain	bl	p	np	cum	1,00	0,00

4. Pengukuran Akurasi

Pengukuran akurasi sistem dilakukan dengan menggunakan persamaan (5). Akurasi adalah ketepatan sistem melakukan proses klasifikasi dengan benar.

3.3 Pengujian Sistem Penyelesaian Studi

Pengujian sistem merupakan proses perhitungan nilai probabilitas penyelesaian studi. Hasil perhitungan menunjukkan bahwa nilai probabilitas penyelesaian studi tepat waktu “tw” sebesar 49%. Hasil perhitungan tersebut mengindikasikan bahwa peluang penyelesaian studi tidak tepat waktu “ttw” lebih besar.

3.4 Pengujian dengan Data Uji

Pengujian ini merupakan proses klasifikasi untuk mengetahui nilai akurasi model. Proses klasifikasi menggunakan data yang berjumlah 182. Data diambil secara acak dengan label kelas yang sudah diketahui. Hasil klasifikasi ditabulasikan dengan *confusion matrix* ditunjukkan pada Tabel 5.

Tabel 5 Hasil Klasifikasi *Bayes Network* dengan *Confusion Matrix*

		Penyelesaian Studi	
		Tepat Waktu	Tidak Tepat Waktu
Prediksi	Tepat Waktu	48	18
	Tidak Tepat Waktu	19	43

Hasil klasifikasi untuk penyelesaian studi yang dilakukan menggunakan *Bayes Network* memiliki nilai akurasi sebesar 71,09%.

Metode lain yang digunakan untuk melakukan proses klasifikasi adalah metode *naive bayes classifiers* (NBC). Proses perhitungan nilai probabilitas pada metode NBC hampir sama dengan *bayes network*, namun metode NBC mengabaikan dependensi antar variabel. Hasil klasifikasi penyelesaian studi pada data mahasiswa menggunakan NBC ditunjukkan pada Tabel 6.

Tabel 6 Hasil Klasifikasi *Naive* B*ayes Classifier* dengan *Confusion Matrix*

		Penyelesaian Studi (Aktual)	
		Tepat Waktu	Tidak Tepat Waktu
Prediksi	Tepat Waktu	45	19
	Tidak Tepat Waktu	22	42

Hasil klasifikasi dengan metode *naive bayes classifier* mendapatkan nilai akurasi sebesar 67,97%.

4. Kesimpulan

Berdasarkan hasil penelitian dapat disimpulkan bahwa struktur *bayes network* yang terbentuk sangat dipengaruhi oleh kondisi *dataset* yang digunakan, sehingga struktur yang dihasilkan dari satu *dataset* belum tentu berlaku pada dataset yang lain. Dalam proses konstruksi struktur *Bayes Netwoks* kuantitas dan distribusi nilai variabel dalam *dataset* yang digunakan memiliki pengaruh yang signifikan terhadap hasil akurasi. Hasil pengujian dari sistem yang dikembangkan mengunakan data uji sebanyak 128 memiliki tingkat akurasi sebesar 71,09%. Hasil akurasi sistem lebih tinggi dibanding dengan menggunakan metode *Naive Bayes Classifier* yaitu sebesar 67,97%.

Daftar Pustaka

[1] M. H. Meinanda, M. Annisa, N. Muhandri, and K. Suryadi, "Prediksi masa studi sarjana dengan artificial neural network," *Internetworking Indones. J*, vol. 1, no. 2, pp. 31–35, 2009.

[2] M. Laugerman, D. T. Rover, M. C. Shelley, and S. K. Mickelson, "Determining graduation rates in engineering for community college transfer students using data mining," *Int. J. Eng. Educ.*, vol. 31, no. 6A, p. 1448, 2015.

[3] N. I. K. D. ARIANI, I. W. SUMARJAYA, and T. B. OKA, "ANALISIS FAKTOR-FAKTOR YANG MEMENGARUHI WAKTU KELULUSAN MAHASISWA DENGAN MENGGUNAKAN METODE GOMPIT (Studi Kasus: Mahasiswa Fakultas MIPA Universitas Udayana)," *E-Jurnal Mat.*, vol. 2, no. 3, pp. 40–45, 2013.

[4] J. Yingkuachat, P. Praneetpolgrang, and B. Kijsirikul, "An Application of the Probabilistic Model to the Prediction of Student Graduation Using Bayesian Belief Network," *ECTI Trans. Comput. Inf. Technol.*, vol. 3, no. 1, pp. 63–71, 2007.

[5] A. L. Madsen and U. B. Kjærulff, *Bayesian networks and influence diagrams: a guide to construction and analysis*. New York: Springer. ISSN, 2013.

[6] G. Shmueli, P. C. Bruce, I. Yahav, N. R. Patel, and K. C. Lichtendahl Jr, *Data mining for*

business analytics: concepts, techniques, and applications in R. John Wiley & Sons, 2017.

[7] C. Chotimah, "Klasifikasi Penyelesaian Studi Mahasiswa Berbasis Bayes Network (Studi Kasus: Universitas Musamus Merauke)." Universitas Gadjah Mada, 2017.

[8] C. Wang, H. Wang, L. Liu, W. Song, and M. Yuan, "Learning Bayesian Network Structure Based on Topological Potential," *J. Inf. &COMPUTATIONAL Sci.*, vol. 12, no. 9, pp. 3383–3393, 2015.