IMPLEMENTASI TEKNIK SELEKSI FITUR *INFORMATION GAIN* PADA ALGORITMA KLASIFIKASI *MACHINE LEARNING* UNTUK PREDIKSI PERFORMA AKADEMIK SISWA

Betha Nurina Sari

Informatika, Fakultas Ilmu Komputer, Universitas Singaperbangsa Karawang
Jl. HS. Ronggowaluyo Teluk Jambe Timur Karawang
Email : betha.nurina@staff.unsika.ac.id[1)]

Abstrak

Masalah utama dalam dalam proses discovering knowledge dari data di bidang pendidikan adalah mengidentifikasi data yang representatif. Penelitian ini dilakukan untuk mengidentifikasi faktor relevan yang mempengaruhi performa akademik siswa dengan mengimplementasikan teknik seleksi fitur Information Gain pada algoritma klasifikasi machine learning. Algoritma klasifikasi machine learning yang digunakan adalah Decision Tree, Random Forest, ANN, SVM, dan Naïve Bayes. Data yang digunakan adalah 395 data akademik dan personal siswa di wilayah Alentejo Portugal.

Eksperimen implementasi teknik seleksi fitur Information Gain dibagi dengan dua macam, yaitu menggunakan pemilihan atribut dengan batas threshold (threshold > 0.01) dan rangking dalam jumlah fitur tertentu (n=5, 10, 15). Selain itu skenario klasifikasi juga dilakukan dalam dua macam skenario, yaitu binary classification (lulus atau gagal) dan 5-level classification. Eksperimen dilakukan menggunakan data mining library pada Java dalam lingkungan Weka. Eksperimen dalam penelitian ini menggunakan 10 fold cross validation dengan. Hasil evaluasi berupa akurasi prediksi yang didapatkan dari matriks konfusi.

Hasil eksperimen menunjukkan bahwa dengan implementasi teknik pemilihan fitur information gain dapat meningkatkan performa algoritma klasifikasi machine learning (J48, Random Forest, MLP, SVM (SMO), dan Naïve Bayes) untuk memprediksi performa akademik siswa pada mata pelajaran Matematika.

Kata kunci: *seleksi fitur, information gain, klasifikasi, prediksi, performa akademik siswa.*

1. Pendahuluan

Masalah utama dalam proses *discovering knowledge* dari data di bidang pendidikan adalah mengidentifikasi data yang representatif sebagai basis model klasifikasi yang dibangun [1]. Penelitian mengenai prediksi performa akademik siswa telah dilakukan untuk mengidentifikasi faktor apa saja yang mempengaruhi hasil akhir akademik siswa. Paulo Cortez dan Alice Silva memprediksi performa akademik siswa dan mengidentifikasi faktor yang relevan dengan memperhatikan hasil model klasifikasi yang tertinggi tingkat akurasinya [2]. Komparasi akurasi prediksi performa akademik siswa dari beberapa algoritma klasifikasi juga telah dilakukan oleh Pedro Strecht dkk, dengan menghasilkan Decision Tree dan SVM sebagai algoritma terbaik klasifikasi. Pedro dkk membagi ke dalam dua skenario eksperimen, yaitu prediksi klasifikasi nilai akhir lulus atau gagal dan prediksi regresi nilai berdasarkan level tingkatan [3].

Prediksi performa akademik siswa tingkat menengah juga dilakukan Ramaswami dan Rathinasabapathy dengan menerapkan metode Bayesian Networks dan tiga teknik seleksi fitur (Information Gain, Chi Square, dan Consistency Subset Evaluation) [4]. Ramesh dkk juga mengidentifikasi faktor yang mempengaruhi performa akademik siswa di ujian akhir dan menerapkan beberapa algoritma klasifikasi (Naïve Bayes, Multi Layer Perception, SMO, J48, REPTree) dengan 10 atribut yang didapatkan dari rata-rata rangking terbaik dari lima teknik seleksi fitur yang diterapkan (Chi Square, Information Gain, OneR, Symmetrical Uncertainty (SU), dan ReliefF) [5]. Osmanbegovic,dkk menerapkan teknik seleksi fitur Information Gain dan Gain Ratio dengan batas *threshold* 0.01 untuk mereduksi dimensi data siswa lalu diterapkan pada teknik klasifikasi algoritme *rule based*, *tree-based*, *function based*, dan *bayes-based* [1].

Tujuan penelitan ini adalah mengidentifikasi faktor relevan yang mempengaruhi performa akademik siswa dengan mengimplementasikan teknik seleksi fitur Information Gain. Penelitian ini diterapkan pada beberapa algoritma klasifikasi *Machine Learning*, yaitu Decision Tree, Random Forest, ANN, SVM, dan Naïve Bayes agar bisa dilakukan komparasi performa dari hasil klasifikasi sebelum dan sesudah dilakukan seleksi fitur pada data akademik siswa.

Pada penelitian ini, digunakan Information Gain untuk seleksi fitur dan beberapa algoritma klasifikasi *Machine Learning*, yaitu Decision Tree, Random Forest, ANN, SVM, dan Naïve Bayes. Algoritma ini juga digunakan

oleh Paulo Cortez dan Alice Silva untuk memprediksi tingkat performa akademik siswa.

Teknik seleksi fitur dilakukan untuk mengurangi fitur yang tidak relevan dan mengurangi dimensi fitur pada data. Dinakaran dkk menerapkan teknik seleksi fitur Information Gain dengan metode perangkingan fitur terbaik dan algoritma klasifikasi Decision Tree J48 [6]. Ramaswami dan Bhaskaran melakukan studi komparasi lima teknik seleksi fitur dan menerapkan empat algoritma klasifikasi pada data mining pendidikan. Hasil eksperimen menunjukkan bahwa teknik seleksi fitur Information Gain menunjukkan hasil yang terbaik [7].

Untuk menghitung Information gain, dihitung dengan rumus di bawah ini [8]:

$$Info\ (D) = -\sum_{i=1}^{c} p_i log_2(p_i) \quad (1)$$

Keterangan dari rumus tersebut adalah:
c : jumlah nilai yang ada pada atribut target (jumlah kelas klasifikasi)
pi : jumlah sampe untuk kelas i

$$Info_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} x\ Info(D_j) \quad (2)$$

Keterangan dari rumus tersebut adalah:
A : atribut
|D| : jumlah seluruh sampel data
$|D_j|$: jumlah sampel untuk nilai j
v : suatu nilai yang mungkin untuk atribut A

Kemudian nilai *information gain* yang digunakan untuk mengukur efektifitas suatu atribut dalam pengklafikasian data dapat dihitung dengan rumus di bawah ini :

$$Gain\ (A) = |Info(D) - Info_A(D)| \quad (3)$$

Implementasi algoritma seleksi fitur Information Gain dan beberapa algoritma klasifikasi ini dijalankan menggunakan *data mining library* pada Java dalam lingkungan Weka [9]. Lingkungan komputasi yang digunakan dalam melakukan eksperimen ini adalah prosesor Intel® Core™ i3-3110M CPU 2,40 GHz dengan RAM 2 GB dengan Windows 7 Ultimate 32 bit.

Data yang digunakan adalah data akademik dan personal siswa dari 2 sekolah di wilayah Alentejo Portugal, yang diambil selama 2005-2006 [2]. Data yang digunakan adalah data 395 siswa untuk evaluasi nilai akhir pada mata pelajaran Matematika. Data tersebut sebagian besar berupa kategori deskriptif sehingga peneliti melakukan proses transformasi menjadi data kategori dalam bentuk numerik. Pada gambar 1 ditunjukkan contoh proses transformasi data yang dilakukan pada tahap *preprocessing*.

Mjob	Fjob	reason	guardian
at_home	teacher	course	mother
at_home	other	course	father
at_home	other	other	mother
health	services	home	mother
other	other	home	father
services	other	reputatio	mother
other	other	home	mother
other	teacher	home	mother
services	other	home	mother
other	other	home	mother
teacher	health	reputatio	mother
services	other	reputatio	father

➡

Mjob	Fjob	reason	guardian
1	1	1	1
1	4	1	2
1	4	4	1
2	3	2	1
5	4	2	2
3	4	3	1
5	4	2	1
5	1	2	1
3	4	2	1
5	4	2	1
4	2	3	1
3	4	3	2

Gambar 1. Tahap *Preprocessing* Data

Eksperimen implementasi teknik seleksi fitur Information Gain dibagi dengan dua macam, yaitu menggunakan pemilihan atribut dengan batas *threshold* (*threshold* > 0.01) dan rangking dalam jumlah fitur tertentu (n=5,10,15). Selain itu skenario klasifikasi juga dilakukan dalam dua macam skenario, yaitu :

1) *binary classification* (klasifikasi biner) : hasil akhir ujian Matematika dikategorikan menjadi dua, yaitu lulus jika G3 ≥ 10 dan gagal jika G3 < 10.
2) *5-level classification* (klasifikasi 5 tingkat) berdasarkan sistem konversi pada tingkat penilaian Erasmus, yaitu 1: 16-20, 2: 14-15, 3: 12-13, 4: 10-11, dan 5: 0-9. Lima tingkat itu adalah (1) sangat bagus, (2) bagus, (3) memuaskan, (4) cukup (5)gagal

Dua skenario klasifikasi hasil nilai ujian Matematika ini juga dilakukan Paulo Cortez dan Alice Silva dalam prediksi performa akademik siswa [2].

Eksperimen dalam penelitian ini menggunakan 10 *fold cross validation* dengan. Hasil evaluasi akan ditampilkan dengan matriks konfusi dengan membandingkan nilai yang diprediksi dengan nilai yang sebenarnya. *Confusion* matrix memberikan penilaian kinerja klasifikasi berdasarkan objek dengan benar atau salah [10]. Matriks konfusi terdiri dari 4 bagian, yaitu dapat dilihat pada Tabel 1.

Tabel 1. Matriks Konfusi

		Prediksi	
		Negatif	Positif
Aktual	Negatif	A	B
	Positif	C	D

Keterangan :
- a dalah jumlah prediksi yang benar bahwa yang diprediksi nilai negatif (TN)
- b adalah jumlah prediksi yang salah, yang seharusnya nilai negatif diprediksi positif (FP)
- c adalah jumlah prediksi yang salah, yang seharusnya nilai positif diprediksi negatif (FN)
- d adalah jumlah prediksi yang benar bahwa yang diprediksi nilai positif (TP)

Akurasi adalah proporsi dari jumlah prediksi yang benar dari semua data yang diprediksi, yaitu dengan rumus :

$$Akurasi\ prediksi(\%) = \frac{A+D}{A\ |\ B\ |\ C\ |\ D} = \frac{TN+TP}{TN\ |\ FP\ |\ FN\ |\ TP}$$

2. Pembahasan

Hasil eksperimen dengan menggunakan semua atribut pada dataset menggunakan lima algoritma klasifikasi *machine learning* untuk memprediksi performa akademik siswa dapat dilihat pada Tabel 2.

Tabel 2. Akurasi Sebelum Seleksi Fitur

Algoritma Klasifikasi	Akurasi	
	Skenario A	Skenario B
Decision Tree (J48)	88.58	100
Random Forest	90.43	93.04
Neural Network (MLP)	88.15	73.65
SVM (SMO)	89.14	66.2
Naïve Bayes	85.67	80.46

Dari tabel 2 dapat dilihat perbedaan tingkat akurasi dari setiap algoritma klasifikasi *machine learning* setelah diterapkan dua skenario target klasifikasi. Pada skenario A algoritma random forest menunjukkan perfoma terbaik dalam klasifikasi biner (lulus atau gagal) sedangkan pada skenario B algoritma decision tree (J48) dapat 100% memprediksi perfoma akademik siswa dalam lima tingkat.

Teknik seleksi fitur information gain diterapkan dengan 2 macam eksperimen yang berbeda, yaitu menggunakan perangkingan atribut dengan menggunakan batas *threshold* (*threshold* > 0.01) dan batas rangking dalam jumlah fitur tertentu (n = 5,10,15). Hasil dari implementasi teknik seleksi fitur information gain pada dataset dapat dilihat pada tabel 3.

Tabel 3. Hasil Seleksi Fitur

Seleksi Fitur	Fitur / atribut yang terpilih	
threshold	A	B
threshold > 0.01	G2, G1, failures, goout, higher	G2, G1, failures
jumlah fitur (n)	A	B
5	G1, G2, failures, goout, higher	G1, G2, failures, reason, Mjob
10	G1, G2, failures, goout, higher, Mjob, Fjob, traveltime, reason, guardian	G1, G2, failures, reason, Mjob, Fjob, studytime, guardian, traveltime, age
15	G1, G2, failures, goout, higher, Mjob, Fjob, traveltime, reason, guardian, age, address, school, sex, Medu	G1, G2, failures, reason, Mjob, Fjob, studytime, guardian, traveltime, age, address, school, sex, Medu, Fedu

Fitur yang terpilih dengan menggunakan *threshold* >0.01 pada skenario A dan B, keduanya terpililh lagi pada eksperimen pemilihan fitur menggunakan perangkingan berdasar jumlah (n=5,10,dan 15). Hal ini menunjukkan bahwa pada skenario A, lima fitur terpilih tersebut relevan terhadap hasil akhir nilai siswa pada ujian ketiga (G3), yaitu G2 (nilai ujian kedua), G1 (nilai ujian pertama), failures (jumlah kegagalan yang pernah dialami siswa pada kelas sebelumnya), goout (tingkat frekuensi bermain bersama teman), higher (ada tidaknya motivasi melanjutkan sekolah tinggi). Pada skenario B, ada tiga fitur yang selalu terpilih, baik berdasarkan *threshold* maupun rangking untuk jumlah yang ditentukan (n=5,10,15), yaitu G2 (nilai ujian kedua), G1 (nilai ujian pertama), failures (jumlah kegagalan yang pernah dialami siswa pada kelas sebelumnya).

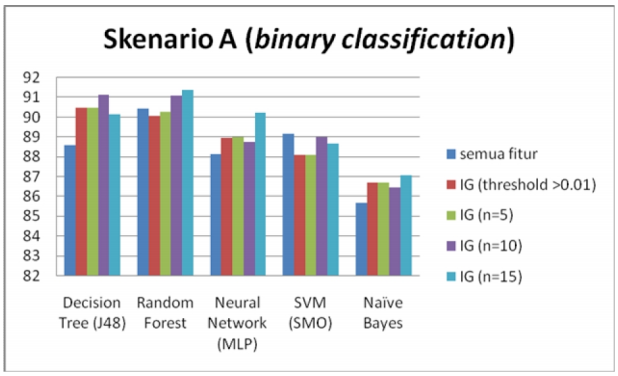
Tabel 4 merupakan tabel komparasi tingkat akurasi dari lima algoritma klasifikasi *machine learning* untuk prediksi performa akademik siswa setelah seleksi fitur pada skenario A. Dari tabel 4 dapat dilihat nilai akurasi tertinggi untuk skenario A adalah Decision Tree J48 yang diterapkan pada sepuluh fitur yang terpilih, yaitu sebesar 91.14 %. Dari tabel 4 juga ditunjukkan bahwa ada pengaruh tingkat akurasi pada semua algoritma klasifikasi yang digunakan pada penelitian ini. Hasil eksperimen pada tabel 3 menunjukkan performa terbaik dari J48 dan SVM (SMO) adalah saat jumlah fitur berjumlah sepuluh, sedangkan performa Random Forest, MLP dan Naïve Bayes mencapai akurasi terbaiknya pada saat fitur yang terpilih berjumlah 15 fitur.

Tabel 4. Akurasi setelah Seleksi Fitur pada Skenario A

Algoritma Klasifikasi	Skenario A (*binary classification*)			
	IG $\theta > 0.01$	IG (n=5)	IG (n=10)	IG (n=15)
J48	90.48	90.48	91.14	90.13
Random Forest	90.05	90.05	91.09	91.37
MLP	88.96	88.96	88.73	90.23
SVM (SMO)	88.1	88.1	89.01	88.68
Naïve Bayes	86.68	86.68	86.46	87.06

Tabel 4 juga menunjukkan bahwa fitur yang terpilih dengan menggunakan batas *threshold* > 0,01 pada eksperimen skenario A tidak bisa dijadikan dasar yang terbaik untuk mengoptimalkan tingkat akurasi dari kelima algoritma klasifikasi *machine learning*. Hal ini dapat dilihat pada saat jumlah fitur berjumlah sepuluh atau lima belas, nilai akurasi dari kelima algoritma klasifikasi menjadi lebih tinggi bila dibandingkan dengan akurasi menggunakan fitur yang dipilih dari hasil teknik seleksi fitur Information Gain dengan batas *threshold* > 0,01.

Gambar 2 menunjukkan grafik komparasi tingkat akurasi prediksi sebelum dan setelah teknik seleksi fitur information gain diimplementasikan pada skenario A.



Gambar 2. Komparasi Akurasi pada Skenario A

Apabila dibandingkan tingkat akurasi sebelum dan setelah teknik seleksi fitur information gain diimplementasikan yaitu dengan memperhatikan tabel 2 dan tabel 4 serta gambar 2, maka didapatkan bahwa empat dari lima algoritma klasifikasi *machine learning* mengalami peningkatkan akurasi, yaitu J48, Random Forest, MLP, dan Naïve Bayes. Dari lima algoritma, hanya SVM (SMO) yang turun tingkat akurasinya yang awalnya sebesar 89.14% dengan semua atribut dari data menjadi 89.01% dengan sepuluh fitur yang terpilih dari data.

Reduksi data dengan pemilihan fitur biasanya meningkatkan performa model prediksi karena fitur yang tidak relevan terhadap target klasifikasi telah berkurang. Hasil eksperimen menunjukkan tidak selalu terjadi kenaikan tingkat akurasi, tetapi menurunkan tingkat akurasi. Hal ini juga terjadi pada eksperimen yang dilakukan Dinakaran, dkk setelah menerapkan teknik pemilihan fitur Information Gain justru menurunkan tingkat akurasi dari Decicion Tree [6]. Eksperimen serupa juga dilakukan Jozef Zurada untuk menguji apakah reduksi fitur meningkatkan akurasi dari algoritma klasifikasi untuk prediksi skor kredit pada dataset German. Hasil eksperimen menunjukkan tingkat akurasi setelah reduksi fitur mengalami penurunan bila dibandingkan dengan tingkat akurasi menggunakan semua fitur pada data [11].

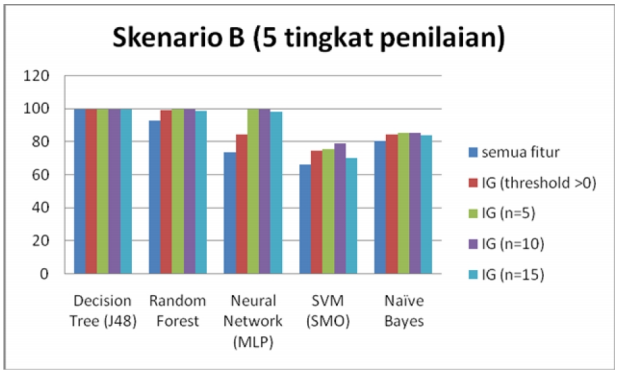
Tabel 5 merupakan tabel komparasi tingkat akurasi dari lima algoritma klasifikasi *machine learning* untuk prediksi performa akademik siswa setelah seleksi fitur pada skenario B. Dari tabel 4 dapat dilihat nilai akurasi tertinggi untuk skenario B adalah Decision Tree J48, yaitu sebesar 100% pada semua kondisi jumlah fitur, baik hasil rangking dengan batas threshold >0.01 maupun rangking 5,10 dan 15 tertinggi. Selain Decision Tree J48, tingkat akurasi Random Forest juga mencapai 100% saat diterapkan pada sepuluh fitur yang terpilih.

Tabel 5. Akurasi setelah seleksi fitur pada skenario B

Algoritma Klasifikasi	Skenario B (5 tingkat penilaian)			
	IG $\theta > 0.01$	IG (n=5)	IG (n=10)	IG (n=15)
Decision Tree (J48)	100	100	100	100
Random Forest	99.27	99.95	100	98.99
Neural Network (MLP)	84.38	99.95	99.95	98.51
SVM (SMO)	74.33	75.39	79.09	69.9
Naïve Bayes	84.38	85.39	85.39	84.13

Tabel 5 juga menunjukkan bahwa fitur yang terpilih dengan menggunakan batas *threshold* > 0,01 pada eksperimen skenarion B tidak bisa dijadikan dasar yang terbaik untuk mengoptimalkan tingkat akurasi dari kelima algoritma klasifikasi *machine learning*. Sifat ini sama dengan hasil eksperimen pada skenario A untuk penggunaan batas threshold > 0.01 pada teknik seleksi fitur information gain.

Gambar 3 menunjukkan grafik komparasi tingkat akurasi prediksi sebelum dan setelah teknik seleksi fitur information gain diimplementasikan pada skenario B.



Gambar 3. Komparasi Akurasi pada Skenario B

Gambar 3 menunjukkan bahwa tingkat akurasi semua algoritma klasifikasi machine learning meningkat bila dibandingkan sebelum dan setelah dilakukan pemilihan fitur. Algoritma Decision Tree J48 sebelum dan setelah direduksi jumlah fiturnya akurasinya tetap 100 % . Keempat algoritma klasifikasi yang lain, yaitu Random Forest, MLP, SVM (SMO) dan Naïve Bayes mencapai puncak tingkat akurasinya saat diterapkan dengan sepuluh fitur yang terpilih.

3. Kesimpulan

Adapun kesimpulan yang didapatkan dari penelitian adalah bahwa dengan implementasi teknik pemilihan fitur information gain dapat mempengaruhi tingkat akurasi algoritma klasifikasi *machine learning* (J48, Random Forest, MLP, SVM (SMO), dan Naïve Bayes) untuk memprediksi performa akademik siswa pada mata pelajaran Matematika. Hasil dari penelitian ini menunjukkan bahwa implementasi teknik seleksi fitur bisa meningkatkan tingkat akurasi karena fitur yang tidak relevan terhadap target klasifikasi telah berkurang. Teknik seleksi fitur information gain dengan memilih

sepuluh fitur pada rangking teratas menunjukkan hasil yang terbaik dalam penelitian ini, ditunjukkan pada tingkat akurasi terbaik dicapai oleh semua algoritma klasifikasi *machine learning* yang digunakan untuk prediksi.

Daftar Pustaka

[1] E. Osmanbegovic, M. Suljic, H. Agic, "Determining Dominant Factor For Student Performance Prediction by Using Dta Mining Classification Algorithms", Original Scientific Paper, July-Des 2014.
[2] P. Cortez, A. Silva, "Using Data Mining to Predict Secondary School Student Performance", in 5th Future Business Technology Conference (FUBUTEC), pp. 5-12, 2008.
[3] P. Stecht, L.Cruz, C. Soares, J.Mendes-Moreira, R. Abreu, "A Comparative Study of Classification and Regression Algorithms for Modelling Student's Academic Performance", in 8th International Conference on Educational Data Mining, Madrid, Spain, June 2015.
[4] M. Ramaswami, R.Rathinasabapathy, "Student Performance Prediction Modeling : A Bayesian Networks Approach", International Journal of Computational Intelligence and Informatics, vol. 1, no.4, pp 231-235. January-March 2012.
[5] V. Ramesh, P. Parkavi, K. Ramar, "Predicting Student Performance : A Statistical and Data Mining Approach", International Journal of Computer Application, vol. 63, no. 8, February 2013.
[6] S. Dinakaran, Dr. P. R. J. Thangaiah, "Role of Attribute Selection in Classification Algorithms", International Journal of Scientific & Engineering Research, vol. 4, issue 6, pp. 67-71. June 2013.
[7] M. Ramaswami, R. Bhaskaran, "A Study on Feature Selection Techniques in Educational Data Mining", Journal Of Computing, vol. 1, Issue 1, pp. 7-11. December 2009.
[8] J. Han, M. Kamber, Data mining concepts and techniques: Morgan Kaufman Publishers, Elsevier, pp. 297- 298. 2006.
[9] I.H. Witten, E. Frank, Data Mining-Practical Machine Learning Tools and Techniques in Java Implementation, San Fransisco: Morgan Kaufmann, 2000.
[10] F. Gorunescu, Data Mining: Concepts, Models and Techniques, Berlin: Springer-Verlag, 2011.
[11] J. Zurada, "Does Feature Reduction Help Improve the Classification Accuracy Rates? A Credit Scoring Case Using a German Data Set", in Review of Business Information Systems, vol. 14, no. 2, pp. 35-40, Second Quarter 2010.

Biodata Penulis

Betha Nurina Sari,memperoleh gelar Sarjana Komputer (S.Kom), Jurusan Matematika program studi Ilmu Komputer di Universitas Brawijaya, lulus tahun 2012. Memperoleh gelar Magister Komputer (M.Kom) Program Pasca Sarjana Magister Ilmu Komputer Universitas Indonesia, lulus tahun 2015. Saat ini menjadi Dosen di program studi Informatika, Fakultas Ilmu Komputer, Universitas Singaperbangsa Karawang.

ISSN : 2302-3805