

PENERAPAN *K-MEANS CLUSTERING* PADA DATA PENERIMAAN MAHASISWA BARU (STUDI KASUS : UNIVERSITAS POTENSI UTAMA)

Fina Nasari¹⁾, Surya Darma²⁾

^{1),2)} Sistem Informasi Universitas Potensi Utama

Jl. K.L. Yos Sudarso Km 6,5 No. 3A Tanjung Mulia-Medan

Email : fina_akhwat17@yahoo.co.id¹⁾, surya.darma_pu@yahoo.com²⁾

Abstrak

Pembentukan cluster merupakan salah satu teknik yang digunakan dalam mengekstrak pola kecenderungan suatu data. Teknik ini digunakan dalam proses Knowledge discovery in database (KDD). Data mining biasanya identik dengan proses penggalian data-data yang cukup besar dan dikelompokkan menjadi data yang tersusun rapi. Dalam hal ini penulis mengelompokkan data mahasiswa baru tahun ajaran 2014/2015 dengan teknik clustering. Pengelompokkan yang penulis terapkan menggunakan algoritma *K-Means Clustering*, algoritma *K-Means Clustering* mampu mengelompokkan data pada kelompok yang sama dan data yang berbeda pada kelompok yang berbeda. Sehingga akan terlihat kelompok data mahasiswa baru tahun ajaran 2014/2015 pada Universitas Potensi Utama yang tidak terstruktur menjadi terstruktur. Tujuan dari penelitian ini adalah menerapkan algoritma *K-Means Clustering* pada data penerimaan mahasiswa baru tahun ajaran 2014/2015 (studi kasus : Universitas Potensi Utama). Hasil *K-Means Clustering* yang diperoleh ada dua kelompok, pusat cluster dengan Cluster 1 = 1 ; 1.75; 1.5 dan Cluster 2 = 2.95; 1.65; 1.4, Cluster pertama jika asal sekolah adalah SMA atau Sekolah Menengah Pertama maka rata-rata jurusan yang diambil adalah Sistem Informasi dan kedua jika asal Sekolahnya adalah SMK rata-rata jurusan yang diambil adalah Teknik Informatika.

Kata kunci : Data Mining, *K-Means Clustering*, Pengelompokkan Data Mahasiswa Baru.

1. Pendahuluan

Pengertian mahasiswa secara umum yaitu suatu peran tertinggi dalam dunia pendidikan yang mengatur pola tingkah laku manusia dari remaja menuju keperan sesungguhnya, bisa dikatakan mahasiswa adalah proses dimana pola pikiran mengarah kelebih tinggi atau lebih serius dalam menjalani peran tersebut.

Data mining adalah proses mencari pola atau informasi menarik dalam data terpilih dengan menggunakan teknik atau metode tertentu. Teknik, metode, atau algoritma dalam *data mining* sangat bervariasi. Pemilihan metode atau algoritma yang tepat sangat bergantung pada tujuan dan proses KDD secara keseluruhan. [jurnal bu fina snif] Salah satu metode yang terdapat dalam data mining yang digunakan dalam penelitian ini adalah pengelompokan (*Clustering*)

dimana metode tersebut mengidentifikasi objek yang memiliki kesamaan karakteristik tertentu, dan kemudian menggunakan karakteristik tersebut sebagai “vektor karakteristik” atau “centroid”. [6]

Beberapa penulis terdahulu telah menerapkan teknik *K-Means Clustering* sebagai penelitian dalam hal Pengelompokkan data, diantaranya :

Ediyanto, dkk (2013) Dalam Penelitiannya yang berjudul “Pengklasifikasian Karakteristik Dengan Metode *K-Means Cluster Analysis*”. Penulis dalam penelitian ini menjelaskan bahwa metode *K-Means Cluster Analysis* cukup efektif diterapkan dalam proses pengklasifikasian karakteristik terhadap objek penelitian. Algoritma *K-Means* juga tidak terpengaruh terhadap urutan objek yang digunakan, hal ini dibuktikan ketika penulis mencoba menentukan secara acak titik awal pusat *cluster* dari salah satu objek pada permulaan perhitungan [1].

Ong Johan Oscar (2013) dengan penelitiannya yang mengangkat judul “Implementasi Algoritma *K-Means Clustering* Untuk Menentukan Strategi Marketing President University”. Tujuan penulisan yang dijelaskan penulis dalam makalah ini menunjukkan bahwa hasil dari pengolahan data mahasiswa membantu pihak marketing President University dalam melakukan pemasaran dan mencari calon mahasiswa baru dari berbagai kota di Indonesia. Dan hasilnya cukup efisien dan efektif [5].

T. Gomasathit (2013) melakukan penelitian untuk mengidentifikasi cakupan awan di atas wilayah Thailand dengan menggunakan data satelit dan Algoritma *K-means clustering*. Namun, hasilnya dijelaskan dalam penelitian memberikan beberapa informasi awal tentang kemungkinan analisis cuaca dari perkiraan awan, dan akan berguna untuk studi lanjut ketika data lainnya tersedia [7].

Dari penelitian yang dilakukan oleh Ediyanto (2013), oleh Ong Johan Oscar (2013), serta penelitian yang dilakukan T. Gomasathit (2013) disimpulkan bahwa Algoritma *K-Means clustering* dapat digunakan dalam mengelompokkan data dengan efisien dan efektif dengan hasil yang diharapkan. Hal inilah yang mendasari penulis dalam melakukan penelitian dalam hal penerapan Algoritma *K-Means Clustering* pada data penerimaan mahasiswa baru tahun ajaran 2014/2015 dengan studi kasus langsung ke target yang ditujuh dalam hal ini melakukan studi kasus pada Universitas Potensi Utama.

Pada penelitian ini penulis menggunakan data analisa seperti No Formulir, No Ujian USM, Nama Mahasiswa Baru, Nilai UAN, dan Asal Sekolah, diharapkan dalam hal ini dapat mengelompokkan kelompok data dengan sangat efektif.

2. Dasar Teori

2.1. KDD (Knowledge Discovery In Database)

Menurut Fayyad dalam buku (kusrini, 2009) Istilah *data mining* dan *knowledge discovery in database* (KDD) sering kali digunakan secara bergantian untuk menjelaskan proses penggalian informasi tersembunyi dalam suatu basis data yang besar. Sebenarnya kedua istilah tersebut memiliki konsep yang berbeda, tetapi berkaitan satu sama lain. Dan salah satu tahapan dalam keseluruhan proses KDD adalah *data mining*. Proses KDD secara garis besar dapat dijelaskan sebagai berikut:

1. Data Selection

Pemilihan (seleksi) data dari sekumpulan data operasional perlu dilakukan sebelum tahap penggalian informasi dalam KDD dimulai. Data hasil seleksi yang akan digunakan untuk proses *data mining* disimpan dalam suatu berkas, terpisah dari basis data operasional.

2. Pre-processing / Cleaning

Sebelum proses *data mining* dapat dilaksanakan, perlu dilakukan proses pembersihan pada data yang menjadi focus KDD. Proses pembersihan mencakup antara lain membuang *duplikasi data*, memeriksa data yang *inkonsisten*, dan memperbaiki kesalahan pada data, seperti kesalahan cetak (*tipografi*).

3. Transformation

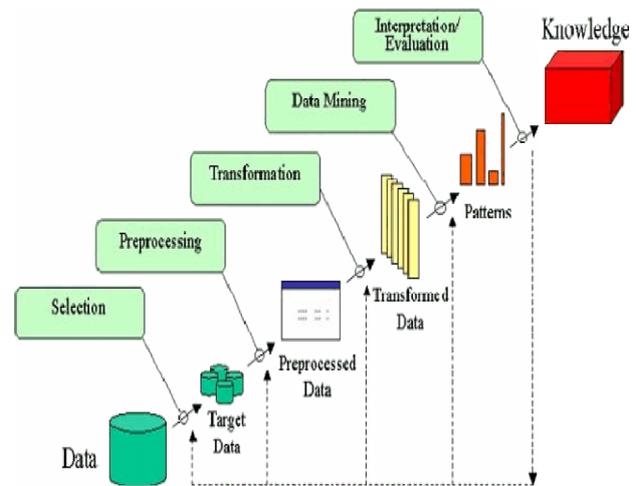
Coding adalah transformasi pada data yang telah dipilih, sehingga data tersebut sesuai untuk proses *data mining*. Proses *coding* dalam KDD merupakan proses kreatif dan sangat tergantung pada jenis atau pola informasi yang akan dicari dalam basis data.

4. Data mining

Data mining adalah proses mencari pola atau informasi menarik dalam data terpilih dengan menggunakan teknik atau metode tertentu. Teknik, metode, atau algoritma dalam *data mining* sangat bervariasi. Pemilihan metode atau algoritma yang tepat sangat bergantung pada tujuan dan proses KDD secara keseluruhan.

5. Interpretation / Evaluation

Pola informasi yang dihasilkan dari proses *data mining* perlu ditampilkan dalam bentuk yang mudah dimengerti oleh pihak yang berkepentingan. Tahap ini merupakan bagian dari proses KDD yang disebut *interpretation*. Tahap ini mencakup pemeriksaan apakah pola atau informasi yang ditemukan bertentangan dengan fakta atau hipotesis yang ada sebelumnya[4].



Gambar 1. Aliran Informasi dalam data mining

2.2. Algoritma K-Means

Pengertian dari *K-Means Clustering* adalah, *K* dimaksudkan sebagai konstanta jumlah *cluster* yang diinginkan, *Means* dalam hal ini berarti nilai suatu rata-rata dari suatu grup data yang dalam hal ini didefinisikan sebagai *cluster*, sehingga *K-Means Clustering* adalah suatu metode penganalisaan data atau metode data mining yang melakukan proses pemodelan tanpa supervisi (*unsupervised*) dan merupakan salah satu metode yang melakukan pengelompokan data dengan sistem partisi. Metode *K-Means* berusaha mengelompokkan data yang ada kedalam beberapa kelompok, dimana data dalam satu kelompok mempunyai karakteristik yang sama satu sama lainnya dan mempunyai karakteristik yang berbeda dengan data yang ada didalam kelompok yang lain. Dasar algoritma *K-means* adalah sebagai berikut : [2]

1. Tentukan nilai *k* sebagai jumlah klaster yang ingin dibentuk.
2. Inisialisasi *k* sebagai centroid yang dapat dibangkitkan secara random.
3. Hitung jarak setiap data ke masing-masing centroid menggunakan persamaan *Euclidean Distance* yaitu sebagai berikut :

$$d(P,Q) = \sqrt{\sum_{j=1}^p (x_j(P) - x_j(Q))^2} \quad (1)$$

4. Kelompokkan setiap data berdasarkan jarak terdekat antara data dengan centroidnya.
5. Tentukan posisi *centroid* baru (*k*)
6. kembali ke langkah 3 jika posisi *centroid* baru dengan *centroid* lama tidak sama.

3. Metodologi Penelitian

Penelitian ini bertujuan untuk mengelompokkan data mahasiswa baru tahun ajaran 2014/2015 dengan menggunakan Algoritma *K-Means Clustering*. Untuk

mencapai tujuan tersebut, penulis akan melakukan pengujian dengan menggunakan data mahasiswa baru tahun ajaran 2014/2015 pada Universitas Potensi Utama.

Adapun Pengelompokan data tersebut digunakan langkah-langkah sebagai berikut :

a. Sumber Data

Sumber data dalam penelitian ini langsung diambil dari Universitas Potensi Utama dimana data tersebut merupakan data sekunder yang terdiri atas data mahasiswa baru tahun ajaran 2014-2015 dengan banyak data 279 data dengan field yaitu data daftar, gel. USM, No. Formulir, No. Ujian USM, Nama Mahasiswa Baru, Prodi, Nilai UAN, Asal Sekolah, berikut adalah contoh data pendaftaran mahasiswa baru.

Tabel 1. Data Pendaftaran Mahasiswa Baru

NO	NO. FORMULIR	NO UJIAN USM	NAMA MAHASISWA BARU	NILAI UAN	ASAL SEKOLAH	Asal Sekolah	PRODI
1	000001	14110001	Alewyah Ramadhani Syahfitri	47,6	SMA Negeri 14 Medan	SMA	Teknik Informatika
2	000002	14110002	Chandra Irawan	43,7	SMA Swasta Swadaya Palm Rakyat	SMA	Teknik Informatika
3	000001	14230001	Muhammad Falmi Sabah	33,9	SMK Negeri 1 Medan	SMK	Sistem Informasi
4	000003	14130001	EkoFrono Wahyu Sumarri	37,4	SMA Swasta Sinar Husni Hebehe	SMA	Teknik Informatika
5	000002	14230002	Deansha Dezanti Pasha	44,65	SMA Swasta Amal Bakti	SMA	Sistem Informasi
6	000003	14220001	Risi Miranti	28,9	SMK Negeri 6 Medan	SMK	Sistem Informasi
7	000004	14130002	M. Faisal	33,8	SMK Swasta Rahmat Istambul Medan	SMK	Teknik Informatika
8	000005	14130003	Adista Amelia	48,6	SMA Swasta Karya Pembangunan Dei Tua	SMA	Teknik Informatika
9	000006	14130004	Feri Kusuma	50,1	SMA Swasta Binaguna Tarah Jawa	SMA	Teknik Informatika
10	000004	14230004	Inam Arjuna Prawira	31,8	SMA Swasta Medan Putri	SMA	Sistem Informasi
11	000005	14230003	Iqbal Hasyim	52,5	SMA Swasta Nabdanti Uluze Medan	SMA	Sistem Informasi

b. Transformasi Data

Pada tahap ini dilakukan proses perubahan data, tujuannya adalah agar data dapat diolah dengan menggunakan algoritma *K-Means Clustering*.

Variabel Yang dipilih pada data Pendaftaran mahasiswa baru yaitu data Nilai UAN, Asal Sekolah dan Program Studi yang dipilih.

Untuk variable Nilai UAN dikelompokkan menjadi 3 Kelompok, yang pertama untuk nilai UAN dengan rata-rata ≤ 5 ditransformasikan dengan nilai 1, > 5 dan data ≤ 7 ditransformasikan dengan nilai 2 dan ≥ 7 ditransformasikan dengan nilai 3.

Untuk Variabel asal Sekolah dikelompokkan menjadi 3 Kelompok yang pertama untuk Asal Sekolah Adalah SMA ditransformasikan dengan nilai 1, Asal Sekolah SMK ditransformasikan dengan nilai 2 dan Asal Perguruan Sebelumnya Selain 2 data diatas ditransformasikan menjadi nilai 3.

Untuk data Program Studi dikelompokkan menjadi 3 kelompok, jika program studi sistem informasi

di transformasikan dengan nilai 1, program studi teknik informatika ditransformasikan dengan nilai 2 dan program studi manajemen informatika ditransformasikan menjadi nilai 3.

Hasil transformasi dapat dilihat pada tabel.2 berikut ini:

Tabel 2. Data Hasil Transformasi

NILAI UAN	Asal Sekolah	PRODI
1	1	1
1	2	1
1	2	2
1	2	2
2	2	2
2	2	2
2	2	2
2	2	1
2	1	1
2	1	2
2	1	2
2	2	1
2	2	2

c. Pengolahan Data

Setelah proses transformasi langkah selanjutnya adalah proses pengolahan data menggunakan algoritma *K-Means Clustering*.

Tahapan proses algoritma *K-Means Clustering* adalah sebagai berikut:

1. Tentukan nilai k dari jumlah cluster yang ingin dibentuk. Cluster yang akan dibuat adalah 2 Cluster.
2. Tentukan titik pusat awal dari setiap cluster. Dalam penelitian ini titik pusat awal ditentukan secara random dan didapat titik pusat dari setiap cluster dapat dilihat pada tabel 3.

Tabel 3. Titik Pusat Awal Setelah Cluster

Titik Pusat Cluster	Nama	Nilai UAN	Asal Sekolah	Prodi
Cluster 1	Muslimana	1	1	1
Cluster 2	Sinta Anggarini	2	2	1

3. Dalam penelitian ini digunakan metode *hard k-means* untuk mengalokasikan setiap data ke dalam suatu cluster yang memiliki jarak paling dekat dengan titik pusat setiap cluster. Untuk mengetahui cluster mana yang paling dekat dengan data, maka perlu dihitung jarak setiap data dengan titik pusat setiap cluster.

Sebagai contoh, akan dihitung jarak dari data mahasiswa pertama ke pusat *cluster* pertama:

$$D(1,1) = \sqrt{(1-1)^2 + (1-1)^2 + (1-1)^2} = 0$$

Dari hasil perhitungan di atas di dapatkan hasil bahwa jarak data mahasiswa pertama dengan pusat *cluster* pertama adalah 0.

4. Jarak data mahasiswa pertama ke pusat *cluster* kedua:

$$D(1,2) = \sqrt{(1-2)^2 + (1-2)^2 + (1-1)^2} = 1.414214$$

Dari hasil perhitungan di atas didapatkan hasil bahwa jarak data mahasiswa pertama dengan pusat *cluster* ketiga adalah 1.414214.

Berdasarkan hasil kedua perhitungan di atas dapat disimpulkan bahwa jarak data mahasiswa pertama yang paling dekat adalah dengan *cluster* 1, sehingga data mahasiswa pertama dimasukkan ke dalam *cluster* 1. Hasil perhitungan selengkapnya untuk 20 data mahasiswa pertama dapat di lihat pada tabel 4 di bawah ini.

Tabel 4. Hasil Perhitungan Setiap Data ke Setiap Cluster Iterasi 1

No	Nim	Nama	Nilai UAN	Asal Sekolah	Prodi	Jarak Ke		Jarak Terdekat Ke Cluster
						C1	C2	
1	14210037	Musliana	1	1	1	0	1	1
2	14210060	Raudhah	1	2	1		1.414214	
3	14110040	Fauzul Azmal Sirait	1	2	2	1	1.732051	1
4	14130024	Handika Putra Effendy	1	2	2		1.732051	
5	14130040	Anzas Putra Ilham	2	2	2	1.414213562	1.414214	1
6	14130033	Deni Setiawan	2	2	2		1.414214	
7	14310041	Rizal Lihadi	2	2	2	1.414213562	1.414214	1
8	14210057	Sinta Anggraini	2	2	1		1	
9	14210043	Erwinskyah Putra	2	1	1	1.732050808	0	2
10	14110023	Simatupang	2	1	2		1	
11	14110038	Ardiansah	2	1	2	1.732050808	1	2
12	14230046	Siti Mirzawati	2	2	1		1	
13	14110044	Cut Nuraini Sari	2	2	2	1.732050808	1.414214	2
14	14220008	Dicky Hardianto	2	2	1		1	
15	14230039	Maulida Yani	2	2	1		1	
16	14130028	Melati Suci Ramadhani	2	1	2	1.414213562	1	2
17	14210031	Alit Benno	2	1	1		0	
18	14210010	Putri Wulandari	2	1	1	1	0	2
19	14130029	Ayu Lestari. P	2	2	2		1.414214	
20	14130029	Devi Nurmayang Sari	2	1	1	1.414213562	0	2
						1.414213562		2
						1.414213562		2
						1.732050808		2
						1.414213562		2
						1.414213562		2
						1.414213562		2
						1		2
						1		2
						1.732050808		2
						1		2

5. Setelah semua data ditempatkan ke dalam *cluster* yang terdekat, kemudian hitung kembali pusat *cluster* yang baru berdasarkan rata-rata anggota yang ada pada *cluster* tersebut. Dalam hal ini jika centroid yang baru konvergen dengan centroid yang lama maka hentikan iterasi, jika tidak maka lanjutkan

iterasi berikutnya. Penghentian iterasi dilakukan karena saat *centroid* baru yang dibangkitkan dengan *centroid* yang lama akan menyebabkan konvergensi pada grup atau *cluster* sehingga tidak perlu menghitung *distance space* (menghitung jarak) data terhadap *centroid*-nya lagi.

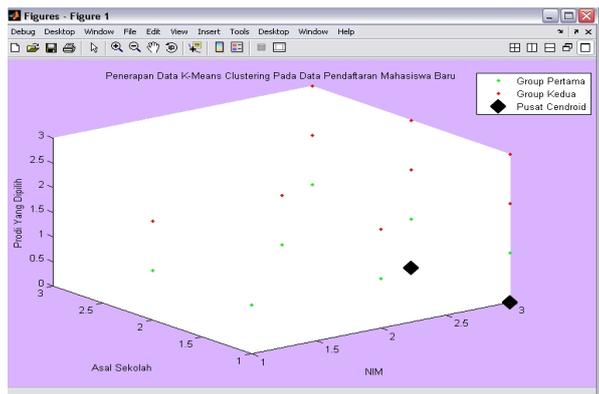
yang berpindah dari satu *cluster* ke *cluster* yang lain.. Maka pusat *cluster* pada iterasi 2 adalah sebagai berikut :

Cluster1 = 1 ; 1.75; 1.5

Cluster2 = 2.907216495; 1.654639175; 1.474226804

3. Analisa dan Pembahasan

Pada tahap ini penulis mengaplikasikan data pendaftaran mahasiswa baru menggunakan algoritma *K-Means Clustering* pada Aplikasi *Matlab 6.1*, untuk 2 kelompok data. Kelompok pertama terlihat pada titik-titik penyebaran yang ditandai dengan warna hijau dan pusat *centroid* pertama terletak pada titik 2.907216495; 1.654639175; 1.474226804 yang ditandai dengan kotak hitam disudut gambar. Kelompok kedua ditandai dengan titik-titik penyebaran warna merah dengan pusat *cluster*nya pada titik 1 ; 1.75; 1.5 dan ditandai dengan kotak hitam sebelah tengah gambar. Hasil *clustering* yang terlihat pada gambar.2 berikut ini:



Gambar.2 Hasil *Clustering* Menggunakan Aplikasi *Matlab*

4. Kesimpulan

Dari hasil penelitian yang dilakukan, maka penulis dapat menarik beberapa kesimpulan, antara lain sebagai berikut :

1. Pengujian yang dilakukan dalam penelitian ini, iterasi *clustering* data mahasiswa terjadi sebanyak 2 kali iterasi.
2. Pengujian yang dilakukan ditemukan pusat *cluster* dengan Cluster 1 = 1 ; 1.75; 1.5 dan Cluster 2 = 2.907216495; 1.654639175; 1.474226804
3. Berhadarkan dari hasil *Cluster* kesimpulan yang dapat diambil adalah bahwa jika asal sekolah adalah SMA atau Sekolah Menengah Pertama maka rata-rata jurusan yang diambil adalah Sistem Informasi dan jika asal Sekolahnya adalah SMK

rata-rata jurusan yang diambil adalah Teknik Informatika.

4. Hasil *cluster* juga dipengaruhi dari nilai *centroid* awal yang dipakai dan jumlah data yang dipakai, perbedaan pengambilan data pusat *centroid* awal yang dipakai juga akan mempengaruhi hasil *centroid* akhirnya.

Daftar Pustaka

- [1] Ediyo, dkk, "Pengklasifikasian Karakteristik Dengan Metode K-Means Cluster Analysis", Buletin Ilmiah Mat. Stat dan Terapannya (Bimaster) Volume 02, No. 2 (2013), hal 133-136
- [2] Himamunanto Agustinus Rudatyo, Mendrova Hendrik, "Perbandingan Algoritma Binerisadi Pada Citra Tulang Abnormal Telapak Tangan Manusia", Prosiding Konferensi Nasional Sistem Informasi 2012 (KNSI 2012), STIKOM Bal 23-25 Pebruari 2012.
- [3] Larose, Daniel, Discovery Knowledge in Data, A Jhon Wiley & Sons, Inc Publication. Canada: 2005
- [4] Nasari Fina, "Penerapan Algoritma c4.5 Dalam Pemilihan Bidang Peminatan Program Studi Sistem Informasi di STMIK Potensi Utama Medan" *Prosiding Seminar Nasional Informatika 2014 (SNIf 2014)*, STMIK Potensi Utama, Medan – Sumatera Utara.
- [5] Ong Johan Oscar, "Implementasi Algoritma K-Means Clustering Untuk Menentukan Strategi Marketing President University", *Jurnal Ilmiah Teknik Industri*, Vol. 12, No. 1, Juni 2013.
- [6] Tanjung Muhammad Rusdi, Erlinda Ningsih, "Perbandingan Penggunaan Data Real dan Data Hasil Normalisasi Pada Pengelompokan Data Dengan Menggunakan Metode Clustering (Studi Kasus PT. PHP Motor)", *Prosiding KeTIK 2014 UIN. Universitas Islam Negeri Medan – Sumatera Utara*.
- [7] T. Gomasathit. "Cloud Coverage Identification Using Satellite Data and K-mean Clustering Algorithm". *Journal of Global Research in Computer Science*, Volume 4, No. 7, July 2013.

Biodata Penulis

Fina Nasari, memperoleh gelar Sarjana Komputer (S.Kom), Program Studi Sistem Informasi Universitas Potensi Utama, lulus tahun 2012. Memperoleh gelar Magister Komputer (M.Kom) Program Pasca Sarjana Magister Sistem Informasi Universitas Putra Indonesia YPTK Padang, lulus tahun 2014. Saat ini menjadi Dosen di Universitas Potensi Utama.

Surya Darma, mahasiswa Fakultas Teknik dan Ilmu Komputer Program Studi Sistem Informasi, Angkatan 2011. Saat ini menjadi mahasiswa Semester VII di Fakultas Teknik dan Ilmu Komputer Program Studi Sistem Informasi Universitas Potensi Utama.