PENERAPAN ALGORITMA KNN PADA PREDIKSI PRODUKSI MINYAK MENTAH

Willmen TB Panjaitan

Magister Teknik Informatika, Fakultas Ilmu Komputer, Universitas AMIKOM Yogyakarta
Jl. Ring Road Utar, Condong Catur, Depok, Sleman, Yogyakarta 55281
Email: willmenpanjaitan@gmail.com[1)]

Abstrak

Minyak mentah merupakan salah satu komoditas utama dalam dunia energi dan termasuk salah satu sumber daya alam yang tidak dapat diperbaharui. Menurut Energy Information Administration (EIA), dunia saat ini mengkonsumsi 85.640.000 barel minyak mentah setiap hari. Ini merupakan proporsi terbesar dari konsumsi energi dunia dibandingkan dengan sumber-sumber lain.

Pada penelitian kali ini akan dibahas mengenai penerapan salah satu metode datamining dalam proses prediksi produksi minyak mentah. Data set yang digunakan berasal dari EIA. Adapun metode yang digunakan yaitu K-Nearest Neighbor. Pengujian hasil dari prediksi menggunakan RMSE.

Hasil dari penelitian ini yaitu penerapan metode K-Nearest Neighbor sederhana dapat memprediksi produksi minyak mentah dengan K=2 untuk nilai RMSE dan absolute error.

Kata kunci: *Energy Information Administration (EIA), K-Nearest Neighbor, RMSE, Absolute Error*

1. Pendahuluan

Berawal pada tahun 1970-an, deregulasi yang terjadi secara dramatis mengakibatkan terjadinya peningkatan harga dipasar energi yang sulit diprediksi. Hal ini mendorong terjadinya pengembangan energi turunan dari sumber daya alam yang ada. Keberhasilan dalam pengembangan sumber energi turunan tersebut telah menarik perhatian pelaku pasar didunia industri. Saat ini, banyak bursa pasar modal di seluruh dunia menawarkan kerja sama dalam produksi baik dalam bentuk kontrak jangka pendek, ataupun kerja sama jangka panjang di berbagai produk energi, termasuk minyak mentah, bahan bakar minyak, gasoil, heating oil, bensin tanpa timbal, dan gas alam(natural gas) [1].

Minyak mentah merupakan komoditas penting di pasar dunia dan merupakan komoditas yang mahal di pasar internasional.[2]. Menurut Energy Information Administration (EIA), dunia saat ini mengkonsumsi 85.640.000 barel minyak mentah setiap hari. Ini merupakan proporsi terbesar dari konsumsi energi dunia dibandingkan dengan sumber-sumber lain. Harga minyak mengalami nilai volatilitas yang tinggi dan fluktuasi yang sulit untuk diprediksi. Di pasar global, minyak mentah adalah komoditas yang paling aktif dan sering diperdagangkan dengan volatilitas mencapai 25% per tahun. Tingkat volitalitas tersebut tidak dapat diabaikan karena memiliki pengaruh dalam perekonomian dunia, khususnya di negara-negara berkembang. Lonjakan drastis dari harga minyak, ketidakpastian keadaan perekonomian dunia dan tren dalam perubahan harga minyak dapat berdampak pada politik dunia, ekonomi, militer dan semua sektor masyarakat [3].

Terdapat beberapa definisi berbeda yang diusulkan dalam berbagai literature ilmiah sebagai pengertian dari data mining [6]. Beberapa definisi yang paling umum antara lain:

a. Data mining meliputi deteksi pola valid, baru, dan mudah dipahami dalam set data; dengan kata lain, itu adalah proses yang ekstrak pengetahuan dari set data dengan menggunakan teknik cerdas.

b. Data mining adalah bidang interdisipliner yang telah terintegrasi berbagai bidang seperti database, statistik, pembelajaran mesin dan bidang terkait lainnya, sehingga informasi berharga dan pengetahuan yang tersembunyi dalam jumlah besar data dapat diekstraksi

Algoritma kNN adalah salah satu algoritma klasifikasi yang paling terkenal digunakan untuk memprediksi kelas dari catatan atau (sampel) dengan kelas yang tidak ditentukan berdasarkan kelas dari catatan tetangganya. algoritma ini terbuat dari tiga langkah sebagai berikut [6]:

1. Menghitung jarak record masukan dari semua catatan pelatihan.
2. Mengatur catatan pelatihan berdasarkan jarak dan pemilihan K-tetangga terdekat.
3. Menggunakan kelas yang memiliki mayoritas diantara k-tetangga terdekat (metode ini menganggap kelas sebagai kelas record input yang diamati lebih dari semua kelas-kelas lain antar K-tetangga terdekat)

Classifier berasumsi jarak catatan dari satu sama lain sebagai kriteria untuk kedekatan mereka dan memilih catatan paling mirip. Ada banyak metode untuk menghitung jarak seperti fungsi jarak Euclidean, Manhattan, dll, di antaranya fungsi jarak Euclidean

adalah salah satu yang paling umum didefinisikan sebagai Persamaan yaitu:

$$d(x_i, x_j) = \sqrt{\sum_{r=1}^{n}(a_r(x_i) - a_r(x_j))^2}$$

Keterangan :
$d(x_i,x_j)$: Jarak *Euclidean (Eucledean Distance).*
(x_i) : *record* ke- *i*
(x_j) : *record* ke- *j*
(a_r) : data ke-*r*
i,j :1,2,3,…n

Metode k-NN adalah metode yang menentukan nilai jarak pada pengujian data testing dengan data training berdasarkan nilai terkecil dari nilai ketetanggaan terdekat[7], didefinisikan sebagi berikut:

$$D_{nn}(C_1, C_2) = min_{1 \le i \le r, 1 \le j \le s} d(y_i, z_j)$$

Minyak Mentah (*Crude Oil*)

Minyak mentah adalah kandungan yang unik dan merupakan campuran kompleks dari ribuan senyawa. Sebagian besar senyawa dalam minyak mentah adalah hidrokarbon (senyawa organik terdiri dari atom karbon dan hidrogen)[8]. Senyawa lain yang terkandung dalam minyak mentah tidak hanya karbon dan hidrogen, tetapi juga senyawa kecil lainnya yang terdiri dari unsur- unsur terutama sulfur, serta nitrogen dan juga logam (misalnya, nikel, vanadium dan logam lainnya). Senyawa yang terbentuk berbagai minyak mentah tardiri dari molekul hidrokarbon terkecil dan paling sederhana (CH4-metana) serta molekul kompleks yang mengandung hingga 50 atau lebih atom karbon (hidrogen dengan baik dan hetero-elemen).

Karakteristik dari minyak mentah dibedakan dari komposisi API Gravity dan Kandungan Sulfur[8].

1. Api Gravity

Kepadatan dari minyak mentah menunjukkan bagaimana komposisi ringan atau berat kandungan yang terdapat didalamnya, secara keseluruhan. Minyak mentah ringan mengandung proporsi yang lebih tinggi dari molekul kecil, dimana kilang dapat memproses menjadi bensin, bahan bakar jet, dan solar. Minyak mentah berat mengandung proporsi yang lebih tinggi dari molekul besar, dimana kilang akan memprosesnya dan dapat digunakan dalam bahan bakar berat industri, aspal, dan produk berat lainnya serta produk bahan bakar transportasi dimana telah dilakukan proses menjadi lebih kecil molekul(lanjutan).

2. Kandungan Sulfur

Dari semua elemen hetero yang terdapat dalam komposisi minyak mentah, sulfur memiliki efek yang paling penting dalam proses penyulingan.

2. Pembahasan

Literature Review

Pada penelitian yang dilakukan oleh Omekara et al [2], melakukan penelitian mengenai penerapan model ARIMA pada produksi minyak mentah di Nigeria. Dataset yang digunakan berasal dari website Central Bank of Nigeria (CBN). Penelitian ini menggunakan aplikasi minitab dengan parameter set yaitu Ø =0.3813, Ø1 =0.6931, Ø2 =0.8649.

Pada penelitian yang dilakukan oleh Augustine dan Saratha [4],melakukan penelitian mengenai prediksi produks minyak mentah menggunakan metode Quadratic Regression dan Layer recurrent Neural Network. Adapun data yang digunakan berasal dari dataset Nigerian National Petroleum Corporation (NNPC). Diperoleh hasil bahwa model Recurrent Neural Network lebih baik dalam melakukan forecasting dengan RMSEs dan MEAs lebih kecil baik untuk data 50-200hari maupun 400-800hari.

Pada penelitian yang dilakukan oleh Ibrahim et al [5], melakukan penelitian mengenai forecasting produksi minyak dunia menggunakan model Multicyclic Hubbert. Adapun dataset yang digunakan bersumber dari berbagai sumber seperti : Twentieth Century Petroleum Statistics,36 Oil and Gas Journal37 (OGJ) database, World Oil Journal38 (WOJ), Energy Information Administration39 (EIA), and OPEC40 official Internet database Web site. Hasil yang diperoleh dari 47 negara yang dievaluasi, Hubbert model menghasilkan 17% model excellent, 70% model sangat bagus dan 13% model baik.

Pada penelitian yang dilakukan Xue Bao dan Xin Guan[12], melakukan penelitian tentang prediksi keluaran minyak mentah dengan menggunakan metode RS-C4.5. Dataset yang digunakan yaitu 9 kategori dimulai dari Januari 1990 s.d Juni 2015 yang terdiri dari:OPEC, GEC, SEC, WIEC, OWTN, OWON, OWOR, WIR, COO. Adapun tool yang digunakan dalam penelitian ini adalah weka dan hasilnya yaitu Efisiensi RS-C4.5 MAE 0.0495 dan RMSE 0.1573.

Pada Penelitian ini, menggunakan jenis penelitian eksperimen, dimana beberapa tahapan penelitian antara lain:

1. Pengumpulan data (Data Gathering)

Pengumpulan data merupakan proses awal dari suatu penelitian. Jenis pengumpulan data terdiri dari dua jenis yaitu, pengumpulan data primer dan pengumpulan data sekunder. Data primer merupakan data utama yang dijadikan objek penelitian. Sedangkan data sekunder merupakan data yang diperoleh dari hasil studi literature yang dilakukan.

2. Pengolahan awal data (Data pre-processing)

Pengolahan awal data merupakan proses mempersiapkan data yang telah diperoleh dari tahap

pengumpulan data sebelumnya dimana akan dilakukan proses selanjutnya.

3. Model/Metode yang diusulkan

Metode yang diusulkan adalah strategi yang digunakan untuk menyelesaikan permasalahan yang dihadapi.

4. Eksperimen dan pengujian metode

Eksperiment dan pengujian metode adalah proses penghitungan dan simuasi untuk menguji metode yang digunakan.

Metode Pengumpulan Data

Pengumpulan data primer pada penelitian kali ini yaitu menggunakan dataset yang diperoleh dari US Energy Information Administration (EIA) dimana terdapat 504 data yang tersedia dan dapat diakses dari http://www.eia.gov/totalenergy/data/monthly/dataunits.cfm. Pengolahan awal data yang dimaksud dengan pengolahan data dalam penelitian ini adalah proses pengelompokan data-data yang telah dikumpulkan sebelumnya dengan tujuan untuk menemukan variable-variabel yang akan digunakan beserta himpunan yang termasuk ke variable-variabel yang digunakan dengan merujuk junal yang ada. Adapun hasil dari preprocessing yang diharapkan adalah seperti pada table.1 sebagai berikut:

Table 1. *Dataset setelah preprocessing*

No	Bulan-Tahun	Produksi
1	1920-January	34008
2	1920-February	33193
3	1920-March	36171
…	…	…
1173	2017-September	284441

Metode yang diusulkan

Adapun metode yang didimaksud yaitu menggunakan *k-nearest neighbor* (k-NN atau KNN) dimana akan dibandingkan hasil produksi dengan jarak antar objek berdasarkan data pembelajaran yang jaraknya paling dekat dengan objek tersebut

Algoritma KNN didasarkan pada pembelajaran dengan analogi yaitu dengan membandingkan contoh tes yang diberikan dengan contoh-contoh pelatihan yang mirip. Contoh pelatihan dijelaskan oleh atribut *n*. setiap contoh merupakan titik dalam ruang n-dimensi. Dengan cara ini, semua contoh pelatihan disimpan diruang pola *n* dimensi. Ketika diberikan contoh yang tidak diketahui, algoritma KNN mencari ruang pola untuk contoh *k* pelatihan yang paling dekat dengan contoh yang tidak diketahui. Contoh *k* pelatihan ini adalah *k* "Nearest Neighbour" dari contoh diketahui "kedekatan" yang didefinisikan dalam hal jarak metrik, seperti pada jarak Euelidean.

KNN merupakan algoritma non parametric lazy learning. Hal ini dikarenakan algoritma KNN tidak membuat asumsi apapun pada distribusi data pokok. Keuntungan ini karena mayoritas data praktis tidak mematuhi asumsi teoritis yang dibuat dan disinilah algoritma non parametric seperti KNN digunakan. KNN juga merupakan algoritma lazy learning dikarenakan tidak menggunakan generalisasi sehingga fase training sangat cepat. Kurangnya generalisasi artinya KNN menyimpan semua data training. KNN menghasilkan keputusan berdasarkan seluruh training dataset.[10]
Misalkan setiap sample pada dataset memiliki atribut *n* yang digabungkan untuk membentuk vector berdimensi *n*:

$$X=(x1, x2, x3,….xn)$$

Atribut *n* dianggap sebagai variable independen yang nilainya bergantung pada *n* yang lain atribut *x*. asumsikan y adalah variable kategori dan terdapat fungsi scalar *f*, yang mewakili kelas untuk setiap vector. Kita tidak mengetahui apapun tengan *f*(jika tidak ada kebutuhan untuk data mining) kecuali kita asumsikan sangat halus. Kita mengira bahwa satu set T vector tersebut diberikan bersamaan dengan kelas yang sesuai.

$$X(i), Y(i) \text{ untuk i} = 1, 2, 3, …. T$$

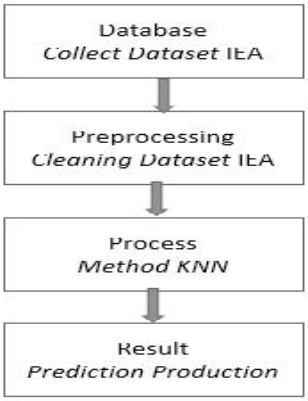
Set ini disebut sebagai training set. Kita anggap kita memberikan sample baru dimana x = u. kita harus menemukan kelas dimana sample ini berasal. Bila kita mengetahui fungsi f, maka dengan mudah kita dapat menghitung $v = f(u)$ untuk mengetahui bagaimana kita mengklasifikasikan sample baru ini tetapi tentu saja kita tidak mengetahui apapun tentang f kecuali f cukup halus.

Ide dari metode k-Nearest Neighbour adalah untuk mengidentifikasi k sampel dalam training set yang independen variable x mirip dengan y, dan menggunakan sample k ini untuk mengklasifikasi sample baru ini kedalam kelas v. Jika kita mengasumsikan bahwa f adalah fungsi halus, sebuah ide yang masuk akal adalah untuk mencari sample dalam data training yang paling dekat(dalam hal variable independen) dan untuk menghitung v dari nilai-nilai y untuk sampel. Ketika kita membahas tentang Neighbour kita menyiratkan adanya jarak atau mengukur perbedaan yang dapat kita hitung antar sample berdasarkan pada variable independen. Measure of Distance yang paling popular adalah Eucludian Distance. Eucludian of Distance antar poin x dan u dapat dirumuskan sebagai[11]:

$$D(x,u) = \sqrt{\sum_{i=1}^{n}(xi - ui)^2}$$

Eksperimen dan pengujian metode

Adapun konsep dalam melakukan eksperimen dan pengujian antara lain:



Gambar 1. *Diagram Eksperimen dan pengujian*

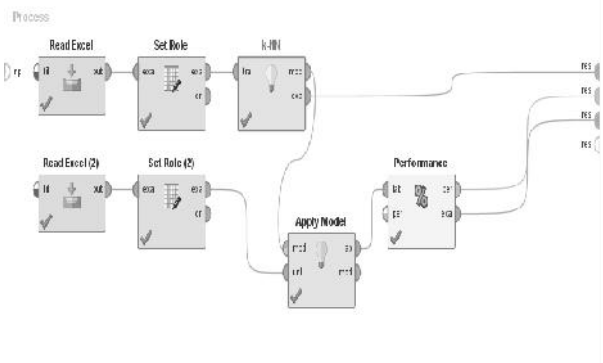
Root Mean Square Error (RMSE) telah digunakan sebagai metrik statistik standar untuk mengukur kinerja model meteorologi, kualitas udara, dan studi penelitian iklim. RMSE adalah metode alternatif untuk mengevaluasi teknik peramalan yang digunakan untuk mengukur tingkat akurasi hasil prakiraan suatu model [9]. Adapun persamaan dapat diperoleh dengan rumus:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}$$

Hasil dan Pembahasan

Eksperimen dilakukan dengan menentukan nilai k dan penentuan metode untuk menentukan jarak(distance). Nilai k untuk menetapkan berapa k data dengan jarak terdekat yang akan dihitung atau diuji dalam percobaan ini. Nilai dari k biasanya adalah angka positif dan kecil. Dalam eksperimen ini, angka k yang diujikan berkisar dari 1 sampai dengan 10. Untuk measure_types akan dipilih numerical measures hal ini dikarenakan dataset dan data prediksi berupa angka.

Pada penelitian ini, digunakan tool *rapidminer studio* 8.0 dengan spesifikasi perangkat yaitu: *Intel* i5 2.5GHz, 4GB of RAM dan sistem operasi Windows 7.



Gambar 2. *Desain model rapidminer yang diterapkan*

Table. 2 *Data perhitungan untuk 12bulan terakhir dengan nilai k=1*

No	Data Testing (x)	Data Produksi (y)	Prediksi (z)
1	Oct-2016	272520	272520
2	Nov-2016	266282	266282
3	Dec-2016	271896	271896
4	Jan-2017	273569	273569
5	Feb-2017	253267	253267
6	Mar-2017	282307	282307
7	Apr-2017	272792	272792
8	May-2017	283169	283169
9	Jun-2017	272035	272035
10	Jul-2017	285465	285465
11	Aug-2017	284909	284909
12	Sep-2017	284441	284441

Table. 3 *Data perhitungan untuk 12bulan terakhir dengan nilai k=3*

No	Data Testing (x)	Data Produksi (y)	Prediksi (z)
1	Oct-2016	272520	265134.667
2	Nov-2016	266282	270232.667
3	Dec-2016	271896	270582.333
4	Jan-2017	273569	266244
5	Feb-2017	253267	269714.333
6	Mar-2017	282307	269455.333
7	Apr-2017	272792	279422.667
8	May-2017	283169	275998.667
9	Jun-2017	272035	280223
10	Jul-2017	285465	280803
11	Aug-2017	284909	284938.333
12	Sep-2017	284441	284938.333

Table. 4 *Data perhitungan untuk 12bulan terakhir dengan nilai k=5*

No	Data Testing (x)	Data Produksi (y)	Prediksi (z)
1	Oct-2016	272520	267501.400
2	Nov-2016	266282	267506.800
3	Dec-2016	271896	267506.800
4	Jan-2017	273569	269464.200
5	Feb-2017	253267	270766.200
6	Mar-2017	282307	273020.800
7	Apr-2017	272792	272714
8	May-2017	283169	279153.600

9	Jun-2017	272035	279674
10	Jul-2017	285465	282003.800
11	Aug-2017	284909	282003.800
12	Sep-2017	284441	282003.800

Table. 5 *Data perhitungan untuk 12bulan terakhir dengan nilai k=7*

No	Data Testing (x)	Data Produksi (y)	Prediksi (z)
1	Oct-2016	272520	268600.857
2	Nov-2016	266282	266334.714
3	Dec-2016	271896	268063.286
4	Jan-2017	273569	270376.143
5	Feb-2017	253267	271897.429
6	Mar-2017	282307	272719.286
7	Apr-2017	272792	274657.714
8	May-2017	283169	276277.714
9	Jun-2017	272035	280731.143
10	Jul-2017	285465	280731.143
11	Aug-2017	284909	280731.143
12	Sep-2017	284441	280731.143

Table. 6 *Data perhitungan untuk 12bulan terakhir dengan nilai k=9*

No	Data Testing (x)	Data Produksi (y)	Prediksi (z)
1	Oct-2016	272520	286511
2	Nov-2016	266282	287862.222
3	Dec-2016	271896	286719.556
4	Jan-2017	273569	286274.222
5	Feb-2017	253267	283146.778
6	Mar-2017	282307	283511.111
7	Apr-2017	272792	280815.222
8	May-2017	283169	279025.556
9	Jun-2017	272035	276643.222
10	Jul-2017	285465	274377.444
11	Aug-2017	284909	273377.778
12	Sep-2017	284441	270099.333

Table. 7 *Hasil perhitungan RMSE dan Absolute error untuk k=1 sampai dengan k=10*

Nilai k	Nilai RMSE	Nilai Absolute Error
1	0.000 +/- 0.000	0.000 +/- 0.000
2	6150.207 +/- 0.000	5032.714 +/- 3535.086
3	6639.264 +/- 0.000	5665.325 +/- 3461.779
4	6356.907 +/- 0.000	5141.042 +/- 3738.978
5	6378.268 +/- 0.000	5116.719 +/- 3808.082
6	6949.853 +/- 0.000	5596.736 +/- 4120.316
7	6450.232 +/- 0.000	5052.350 +/- 4009.893
8	6538.091 +/- 0.000	5113.924 +/- 4073.624
9	6643.558 +/- 0.000	5212.218 +/- 4119.423
10	6733.158 +/- 0.000	5151.467 +/- 4335.644

3. Kesimpulan

Metode KNN dapat digunakan untuk melakukan prediksi produksi minyak mentah. Adapun penerapan algoritma masih cukup sederhana dan untuk mendapatkan nilai RMSE dan absolute error yang baik, perlu dilakukan pengujian untuk nilai-K yang akan digunakan serta dilakukan perbandingan dengan metode lainnya sehingga didapat model terbaik.

Daftar Pustaka

[1] J. Fleming and B. Ostdiek, “THE IMPACT OF ENERGY DERIVATIVES ON THE CRUDE OIL MARKET,” THE JAMES A.BAKER III INSTITUTE FOR PUBLIC POLICY OF RICE UNIVERSITY, 2008.

[2] C. . Omekara, O. E. Okereke, K. . Ire, and C. . Okamgba, “ARIMA Modeling of Nigeria Crude Oil Production,” J. Energy Technol. Policy, vol. 5, no. 10, pp. 1–5, 2015.

[3] L. Gabralla and A. Abraham, “Computational Modeling of Crude Oil Price Forecasting: A Review of Two Decades of Research,” Mirlabs.Org, vol. 5, pp. 729–740, 2013.

[4] A. Pwasong and S. Sathasivam, “Forecasting crude oil production using quadratic regression and layer recurrent neural network models,” vol. 20001, p. 20001, 2016.

[5] I. S. Nashawi, A. Malallah, and M. Al-Bisharah, “Forecasting world crude oil production using multicyclic Hubbert model,” Energy and Fuels, vol. 24, no. 3, pp. 1788–1800, 2010.

[6] M. Kuhkan, “A Method to Improve the Accuracy of K-Nearest Neighbor Algorithm,” Int. J. Comput. Eng. Inf. Technol., vol. 8, no. 6, pp. 90–95, 2016.

[7] N. Krisandi, B. Prihandono, and Helmi, “Algoritma K - Nearest Neighbor Dalam Klasifikasi Data Hasil Produksi Kelapa Sawit Pada PT. MINAMAS Kecamatan Parindu,” Bul. Ilm. Math.Stat.dan Ter., vol. 2, no. 1, pp. 33–38, 2013.

[8] International Council on Clean Transportion, "An Introduction To Petroleum Refining and the Production of Ultra Low Sulfur Gasoline," pp. 1–33, 2011.
[9] T. Chai and R. R. Draxler, "Root mean square error (RMSE) or mean absolute error (MAE)? -Arguments against avoiding RMSE in the literature," Geosci. Model Dev., vol. 7, no. 3, pp. 1247–1250, 2014.
[10] Prajesh P Anchalia, Kaushik Roy, "The k-Nearest Neighbor Algorithm Using MapReduce Paradigm", Fifth International Conference on Intelligent System, Modeling and Simulation, 2014.
[11] K. Ming Leung, "k-Nearest Neigbor Algorithm for Classification", NYU Politechnic Scool of Enginering, 2007.
[12] Xue Bao and Xin Guan, "A Method of predicting crude oil output based on RS-C4.5 Algorithm", 3rd International Conference on Information Science and Control Engineering, ICISCE, 2016.

Biodata Penulis

Willmen TB Panjaitan, memperoleh gelar Sarjana Teknik (S.T), Jurusan Teknik Informatika Universitas AtmaJaya Yogyakarta dengan peminatan *mobile computation*, lulus pada tahun 2010. Saat ini menempuh pendidikan Program Pasca Sarjana Magister Teknik Informatika Universitas Amikom Yogyakarta dengan peminatan *intellegent business*.