JASIEK

Jurnal Aplikasi Sains, Informasi, Elektronika dan Komputer

ISSN- 2685-497X Vol 6, Issue 1, 2024

Penerapan Metode *Decision Tree* C4.5 untuk Klasifikasi Data Kandidat Tenaga Kerja pada Perusahaan *Outsourcing*

Ahmad Ardhy Ansyah [a,1*], Tresna Maulana Fahrudin [a,2], Dwi Arman Prasetya [a,3]

[a] Sains Data UPN "Veteran" Jawa Timur, Jl. Rungkut Madya Kec. Gunung Anyar, Kota Surabaya, Indonesia
[1] 20083010008@student.upnjatim.ac.id*; [2] tresna.maulana.ds@upnjatim.ac.id;
[3]arman.prasetya.sada@upnjatim.ac.id
* Penulis Koresponden

INFO ARTIKEL

Histori Artikel
Pengajuan 2024-03-19
Diperbaiki 2024-06-19
Diterima 2024-06-25

Kata Kunci
Klasifikasi
Tenaga Kerja
Outsourcing
Perusahaan
Decision Tree C4.5

ABSTRAK

Sebuah perusahaan biasanya melakukan *screening* menggunakan metode konvensional dalam mencari kandidat tenaga kerja. Hal tersebut berdampak pada proses pemilihan kandidat tenaga kerja yang menghabiskan waktu yang cukup lama dan ketidakkonsistenan dalam pengambilan keputusan. Oleh karena itu, dalam proses pemilihan kandidat tenaga kerja dapat menggunakan sebuah model Machine Learning yang dapat melakukan klasifikasi berdasarkan profil dan kompetensi kandidat tenaga kerja. Model machine learning yang digunakan salah satunya adalah Decision Tree C4.5 yang mampu menghasilkan sebuah keputusan pemilihan kandidat tenaga kerja secara otomatis berdasarkan data. Hasil pengujian menunjukkan bahwa model ini memiliki akurasi dan *precision* yang tinggi, terutama pada *validation* model menggunakan *holdout* atau *percentage split* dengan proporsi data training dan testing sebesar 70:30% masing-masing yakni mencapai akurasi terbaik sebesar 0.99, dan *precision* sebesar 0.9. Dengan demikian, model ini dapat dimplementasikan dalam sistem seleksi kandidat untuk meningkatkan efisiensi proses seleksi dan pengelolaan data kandidat.

1. Pendahuluan

Dalam mencari kandidat sebuah perusahaan perlu melakukan beberapa tahap untuk mengidentifikasi kemampuan dari seorang kandidat, apakah sesuai dengan kriteria yang dibutuhkan pada sebuah posisi. Perusahaan *outsourcing* ini adalah sebuah perusahaan yang bergerak pada bidang penyedia jasa *outsourcing* tenaga ahli IT. Untuk mendapatkan kandidat yang sesuai, perlu dilakukan dengan proses yang cepat dan tepat. Perkembangan teknologi saat ini sudah sangat pesat, apalagi di zaman sekarang sudah sering terdengar istilah kecerdasan buatan. Kecerdasan buatan dapat memecahkan masalah yang hampir mengungguli manusia dengan tingkat ketidakpastian dan kompleksitas rendah, sehingga dapat menyelesaikan tugas manusia dengan pembelajaran yang lebih dalam [1]. Dengan memanfaatkan kecerdasan buatan tersebut dapat dimanfaatkan untuk mendukung pengambilan keputusan. Salah satu bidang kecerdasan buatan adalah Machine Learning, Machine

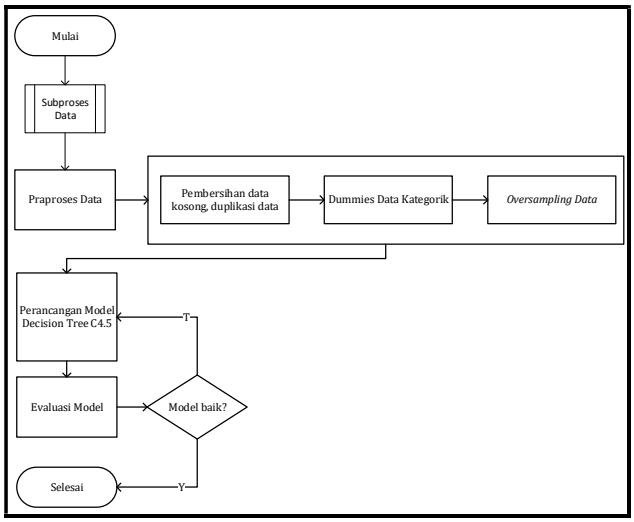
Learning adalah sebuah mesin yang mempelajari dan meningkatkan kinerjanya secara otomatis tanpa campur tangan manusia. Teknologi ini memanfaatkan berbagai disiplin ilmu, seperti statistik, matematika, dan data mining, untuk mencapai kemampuan belajar mandiri [2]. Dalam Machine Learning terdapat bidang klasifikasi yang bertujuan untuk mengategorikan data berdasarkan kesamaan karakteristik yang dimiliki oleh data [3].

Saat ini perusahaan *outsourcing* ini dalam melakukan *screening* untuk mencari tenaga kerja masih menggunakan cara manual, yaitu dengan menggunakan proses *review* daftar riwayat hidup kandidat, lalu yang sesuai dengan kualifikasi akan dikelompokkan sendiri, sedangkan yang tidak sesuai kualifikasi akan dihapuskan. Hal tersebut menjadikan proses kurang efisien, karena proses *review* akan menghabiskan waktu lama, dan juga kandidat yang terpilih ataupun tidak terpilih tidak ada rekap datanya. Oleh karena itu akan dibuat sebuah model yang dapat melakukan klasifikasi data kandidat. Pada penelitian terdahulu, Decision Tree C4.5 pernah digunakan untuk klasifikasi calon debitur LPD Desa Adat Anggungan dengan menggunakan metode Decision Tree C4.5 dan dinilai dengan menggunakan Confusion Matrix menghasilkan nilai akurasi kebenaran model dalam klasifikasi data adalah 0.9, lalu nilai *precision* 0.938 dan *recall* sebesar 0.834, serta nilai AUC dari penelitian tersebut sebesar 0.834 sehingga termasuk ke dalam *Good Classification* [4]. Dari penelitian yang pernah dilakukan, maka dalam penelitian ini akan menggunakan metode tersebut.

Tujuan dari penelitian ini yaitu untuk membuat sebuah model Machine Learning yang dapat melakukan klasifikasi kandidat tenaga kerja. Selain itu, *record* dari kandidat yang pernah di-*review* akan disimpan ke dalam basis data sistem, harapannya ketika nanti ada kebutuhan dengan posisi yang sama, masih dapat menggali potensi dari kandidat yang pernah di-*review* pada proses sebelumnya. Diharapkan dari penelitian ini dapat mempercepat proses seleksi kandidat *outsourcing* di perusahaan *outsourcing*.

2. Metode penelitian

Pada penelitian ini akan mengacu pada diagram alir berikut



Gambar 1. Metode Penelitian

2.1. Persiapan Data

Persiapan data adalah tahapan dalam mencari data yang akan digunakan untuk proses klasifikasi. Dataset di dalam penelitian ini diperoleh dari perusahaan *outsourcing*, data tersebut adalah data kandidat yang pernah diterima dan dipromosikan ke perusahaan yang sedang membutuhkan kandidat tenaga kerja. Berikut adalah data yang digunakan.

Tabel 1. Data Olah Untuk Pembuatan Model Klasifikasi

Nama Kandidat	Posisi Harapan	Usia	Lama Pengalaman	Kesesuaian Posisi dengan Skills	Kesesuaian Posisi dengan Pengalaman	Label
P1	Data Analyst	23	1	1	1	1
P2	Data Scientist	25	1	1	1	1
P3	Java Developer	30	1	1	1	1
P4	C++ Dev	18	0	1	0	0
P5	iOS Developer	27	1	1	1	1

Data tersebut adalah sampel data, total data terdapat 300 baris dengan 7 kolom, dengan penjelasan sebagai berikut:

1. Usia: Usia dari kandidat *outsourcing* saat mendaftar.
2. Lama Pengalaman: berisi data kategorik 0 dan 1, 0 menjelaskan bahwa lama pengalaman yang sesuai dengan posisi harapan kurang dari 1 tahun. Dan apabila 1 menjelaskan bahwa lama pengalaman yang sesuai dengan posisi harapan lebih dari 1 tahun.
3. Kesesuaian Posisi Dengan Skills: berisi data kategorik 0 dan 1, di mana 0 berarti posisi harapan dari kandidat sesuai dengan skills (keahlian) yang dimiliki.
4. Kesesuaian Posisi Dengan Pengalaman: berisi data kategorik 0 dan 1, di mana 0 berarti posisi harapan dari kandidat sesuai dengan pengalaman kerja yang dimiliki.
5. Label: berisi data hasil klasifikasi, yang berisi nilai 0 dan 1, di mana 0 adalah data dengan kriteria "*Not Recommended*" dan nilai 1 adalah data dengan kriteria "*Recommended*".

Data di atas adalah data mentah yang perlu dilakukan pembersihan. Selanjutnya data tersebut akan dilakukan pengolahan melalui proses *data preparation.*

2.2. Data Preparation

Data preparation adalah sebuah proses untuk mempersiapkan data agar dapat digunakan untuk pengujian secara baik, sehingga model klasifikasi yang dibuat juga bagus. Proses yang dilakukan yaitu *encode data. Encode data* adalah proses mengubah data dari satu bentuk ke bentuk lain yang sesuai dengan kebutuhan analisis atau aplikasi tertentu [5].

2.3. Perancangan Decision Tree C4.5

Decision Tree adalah sebuah pohon yang dibangun dari akar, node, dan daun, di mana setiap cabangnya memiliki makna yaitu sebagai hasil dari pengujian data, node daun memiliki makna mewakili kelas tertentu [6]. Model yang digunakan dalam klasifikasi di sini adalah Decision Tree Algoritma C4.5. Algoritma tersebut merupakan pengembangan dari algoritma ID3. Berikut adalah beberapa formula yang digunakan untuk menghitung nilai di dalam Decision Tree C4.5 [7].

Rumus Entropi

$$Entropy\,(S) = \sum_{i=1}^{n} p_i * log_2\, p_i \qquad (1)$$

Keterangan:
S: himpunan kasus
pi: proporsi dari Si terhadap S

Rumus untuk menghitung nilai gain

$$Gain\,(S, A) = Entropy\,(S) - \sum_{i=1}^{n} \frac{|S_j|}{|S|} * Entropy\,(S_i) \qquad (2)$$

Keterangan:
S: himpunan kasus

A: atribut
|Si|: jumlah kasus pada partisi ke-i
|S|: jumlah kasus dalam S

Rumus untuk menghitung nilai *splitinfo*

$$Splitinfo\ (S, A) = \ -\sum_{j=1}^{k} \frac{S_j}{S} * log^2 \frac{S_j}{S} \quad (3)$$

Keterangan:
S: himpunan kasus
A: atribut
Sj: jumlah sample untuk atribut ke j

Rumus untuk menghitung nilai *gainratio*

$$GainRatio\ (S, A) = \frac{Gain\ (S,A)}{Split\ (S,Ai)} \quad (4)$$

Keterangan:
Gain(S,A): information gain pada atribut (S,A)
A: atribut
Split(S,A): split information pada atribut (S,A)

Algoritma dari Decision Tree C4.5 adalah sebagai berikut :

a. Pilih atribut yang akan digunakan sebagai data pengujian.
b. Hitung Entropi dari data.
c. Hitung nilai gain.
d. Hitung nilai split info.
e. Hitung nilai gain ratio.
f. Nilai gain yang paling tinggi akan menempati cabang tertinggi.
g. Ulangi langkah hingga mendapatkan semua cabang memiliki kelas yang serupa.

Setelah semua cabang mendapatkan kelas, maka akan didapatkan sebuah rule dari pohon tersebut [8], *rule* tersebut akan digunakan untuk pedoman klasifikasi data dari kandidat tenaga kerja.

2.4. Evaluasi Model

Evaluasi model adalah tahap untuk membagi data menjadi data latih dan data uji. Pembagian data akan menggunakan beberapa skema, untuk mencari hasil terbaik dari setiap model. Dalam mengevaluasi model akan digunakan *Confussion Matrix* untuk mengetahui hasil akurasi atau kebenaran model dalam prediksi data [9]. Matriks ini menyediakan rincian label prediksi dan aktual untuk tugas klasifikasi. Metrik kunci yang didefinisikan di dalam matriks kebingungan dan untuk mengevaluasi performa model yang dilatih, nilai metrik performa dapat dihitung menggunakan persamaan berikut [10]:

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

$$Precision\ = \frac{TP}{TP+FP} \quad (6)$$

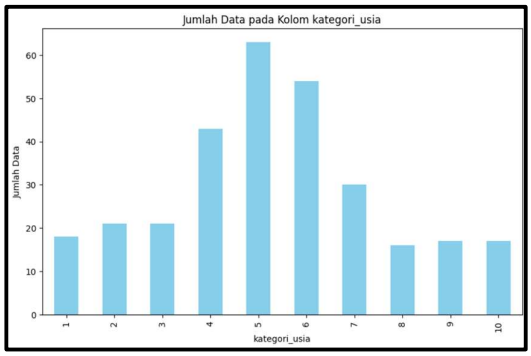
$$Recall\ = \frac{TP}{TP+FN} \quad (7)$$

$$F1-\ Score\ = \frac{2\times(\ Precision \times Recall\ )}{(\ Precision + Recall\ )} \quad (8)$$

3. Hasil dan Analisis

3.1. Hasil *Data Preparation*

Data yang telah dilakukan *preparation* dengan urutan, mengecek nilai null, mengecek duplikasi data, dan yang terakhir dummies data kategorik. Berikut adalah tampilan data awal. Berikut adalah jumlah kategori umur hasil kategorisasi



Gambar 2. Hasil Kategorisasi Data Usia

Data di gambar 2 adalah hasil kategorisasi umur menggunakan metode *binning,* apabila data tersebut langsung digunakan untuk memodelkan data Machine Learning, maka model akan mengalami kesalahan dalam memahami data, untuk itu perlu dilakukan proses *dummies data,* proses ini dapat menggunakan *library* Pandas. Berikut adalah tabel hasil proses *dummies data.*

Tabel 2. Hasil *Dummies* Kategori Data Usia

Kategori _Usia_1	Kategori _Usia_2	Kategori _Usia_3	Kategori _Usia_4	Kategori _Usia_5	Kategori _Usia_6	Kategori _Usia_7	Kategori _Usia_8	Kategori _Usia_9	Kategori _Usia_10
1	0	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0	0

Hasil data tersebut mendeskripsikan keberadaan dari nilai data, misal pada data di atas, kolom kategori usia 1 terdapat nilai 1, berada data pada row tersebut kandidat terdapat pada kategori usia 1. Begitu seterusnya pada kolom – kolom selanjutnya. Berikut ini pada Gambar 3 adalah tabel hasil data setelah dilakukan *data preparation.*

Kategori_Umur_1	Kategori_Umur_2	Kategori_Umur_3	Kategori_Umur_4	Kategori_Umur_5	Kategori_Umur_6	Kategori_Umur_7	Kategori_Umur_8	Kategori_Umur_9	Kategori_Umur_10	Pendidikan	Lama_Pengalaman	Kesesuaian_Posisi_Skill	Kesesuaian_Posisi_Pengalaman
1	0	0	0	0	0	0	0	0	0	0	1	1	1
0	0	0	0	1	0	0	0	0	0	1	1	1	0
0	0	1	0	0	0	0	0	0	0	1	0	1	1
1	0	0	0	0	0	0	0	0	0	0	1	0	1
0	0	0	0	0	1	0	0	0	0	1	1	1	1
...	...	...	...	...	...	...	...	...	...	...	...	...	...
0	0	0	0	1	0	0	0	0	0	1	0	0	0
0	0	0	0	0	0	0	0	0	1	1	0	0	1
0	0	0	0	0	1	0	0	0	0	1	1	1	1
0	0	0	1	0	0	0	0	0	0	1	1	1	1
0	0	0	0	1	0	0	0	0	0	0	0	0	0

Gambar 3. Data Akhir Hasil Praproses Data

Setelah itu akan dilakukan *oversampling* untuk menyeimbangkan data. Untuk menyelesakan itu dapat menggunakan library di Python yaitu menggunakan imblearn.oversampling. Berikut adalah jumlah data sebelum dilakukan *oversampling*

```
data_olah['label'].value_counts()

label
0    253
1     47
Name: count, dtype: int64
```

Gambar 4. Jumlah Label data sebelum *oversampling data*

Berikut adalah jumlah data setelah dilakukan proses *oversampling*

```
y_resampled.value_counts()

label
0    253
1    253
Name: count, dtype: int64
```

Gambar 5. Hasil *Oversampling* data

Oversampling data tersebut berfungsi agar model dapat seimbang mempelajari data dengan label 0 dan 1. Sehingga nantinya ketika model mempelajari data dapat mengerti karakteristik kedua jenis label data.

3.2. Pembuatan Model Decision Tree C4.5

Untuk membangun model Decision Tree, dalam penelitian ini akan menggunakan kode *scratch* Decision Tree C4.5. Sebelum membuat model, data akan dibagi dulu menjadi data latih dan data uji, dalam proporsi data tersebut akan diuji dengan beberapa skenario untuk dicari hasil model terbaik. Selanjutnya adalah proses untuk melatih model dengan data latih yang sudah ditetapkan sebelumnya, dalam melatih data akan menggunakan data X_train dan y_train. Dalam penelitian ini akan menggunakan beberapa proporsi data yaitu data latih : data validasi : data ujinya adalah 70:15:15, 80:10:10, dan 90:5:5. Lalu dibandingkan hasil dari beberapa model data tersebut, mana yang terbaik untuk diimplementasikan menjadi sistem klasifikasi.

3.2. Evaluasi Model Decision Tree C4.5

Hasil dari model yang telah dibuat setelah diuji dengan beberapa skema tercantum di dalam tabel berikut:

Tabel 3. Hasil Uji Model Decision Tree C4.5

Classification Report Model Decision Tree C4.5				
		70:15:15	80:10:10	90:5:5
CM	TP	41	26	16
	TN	34	24	9
	FP	1	1	1
	FN	0	0	0
0	Precision	1.00	1.00	1.00
	Recall	0.97	0.96	0.90
	F1-Score	0.99	0.98	0.95
	Support	35	25	10
1	Precision	0.98	0.96	0.94
	Recall	1.00	1.00	1.00
	F1-Score	0.99	0.98	0.97
	Support	41	26	16
	Akurasi	0.99	0.98	0.96

Classification Report Model Decision Tree C4.5

	70:15:15	80:10:10	90:5:5
Macro AVG	0.99	0.98	0.96
Weighted Avg	0.99	0.98	0.96

Berdasarkan Tabel 3, hasil model berbeda pada beberapa skema pengujian. Model dengan proporsi pengujian data latih 70%, data validasi 15%, dan data uji sebanyak 15% menghasilkan nilai akurasi sebesar 0.99, *precision* sebesar 1, *recall* sebesar 0.97 dan F-1 Score sebesar 0.99, sedangkan model dengan data latih 80%, data validasi 10%, dan data uji sebanyak 10% menghasilkan nilai akurasi sebesar 0.98, *precision* sebesar 1, *recall* sebesar 0.96 dan F-1 Score sebesar 0.98, sedangkan model dengan proporsi data latih 90%, data validasi 5%, dan data uji sebanyak 5% yang menghasilkan nilai akurasi sebesar 0.96, *precision* sebesar 1, *recall* sebesar 0.90 dan F-1 Score sebesar 0.95. Hasil dari model yang telah ditulis di atas memiliki kelebihan dan kekurangannya masing – masing, apabila ditinjau dari kebutuhan perusahaan yang membutuhkan keakuratan model dalam klasifikasi, maka dapat mengambil sebuah model yang memiliki nilai akurasi dan *precision* yang tinggi.

4. Kesimpulan

Berdasarkan hasil yang telah didapatkan, model Decision Tree C4.5 yang telah dibuat, menghasilkan hasil klasifikasi yang bagus, dan mendapatkan nilai *precision* yang tinggi. Model terbaik untuk digunakan yaitu model menggunakan pengujian data dengan proporsi 70:15:15, lalu terbaik kedua yaitu dengan proporsi data 80:10:10, lalu terbaik ketiga yaitu model dengan proporsi data 90:5:5. Pemilihan model terbaik tersebut dikarenakan dilihat berdasarkan mendapat nilai akurasi yang tinggi, serta nilai *precision* yang tinggi dari hasil pengujiannya. Oleh karena itu, selanjutnya model tersebut dapat digunakan untuk membuat sebuah sistem yang dapat digunakan untuk klasifikasi data kandidat tenaga kerja bagi perusahaan *outsourcing.* Apabila dilihat dari perbandingan pada penelitian sebelum – sebelumnya yang menggunakan model Decision Tree yang mendapatkan akurasi 0.9.

References

[1] J. B. Bullock, "Artificial Intelligence, Discretion, and Bureaucracy," *Am. Rev. Public Adm.*, vol. 49, no. 7, pp. 751–761, 2019, doi: 10.1177/0275074019856123.

[2] P. R. Sihombing and I. F. Yuliati, "Penerapan Metode Machine Learning dalam Klasifikasi Risiko Kejadian Berat Badan Lahir Rendah di Indonesia," *MATRIK J. Manajemen, Tek. Inform. dan Rekayasa Komput.*, vol. 20, no. 2, pp. 417–426, 2021, doi: 10.30812/matrik.v20i2.1174.

[3] A. Amrulloh and A. P. Wibowo, "Implementasi Algoritma Decission Tree Untuk Mengklasifikasi Kondisi Kesuburan Pria," *JASIEK (Jurnal Apl. Sains, Informasi, Elektron. dan Komputer)*, vol. 1, no. 1, pp. 7–11, 2019, doi: 10.26905/jasiek.v1i1.3096.

[4] I. G. A. M. P. Dewi, W. G. S. Parwita, and I. M. D. Setiawan, "Algoritma Decision Tree untuk Klasifikasi Calon Debitur LPD Desa Adat Anggungan," *J. Krisnadana*, vol. 1, no. 1, pp. 23–36, 2021, doi: 10.58982/krisnadana.v1i1.79.

[5] A. S. Lombu, S. Hidayat, and A. F. Hidayatullah, "Pemodelan Klasifikasi Gaji Menggunakan Support Vector Machine," *J. Comput. Syst. Informatics*, vol. 3, no. 4, pp. 363–370, 2022, doi: 10.47065/josyc.v3i4.2137.

[6] A. H. Nasrullah, "Implementasi Algoritma Decision Tree Untuk Klasifikasi Produk Laris," *J. Ilm. Ilmu Komput.*, vol. 7, no. 2, pp. 45–51, 2021, doi: 10.35329/jiik.v7i2.203.

[7] T. M. Fahrudin, I. Syarif, and A. R. Barakbah, "Data Mining Approach for Breast Cancer Patient Recovery," *Emit. Int. J. Eng. Technol.*, vol. 5, no. 1, pp. 36–71, 2017, doi: 10.24003/emitter.v5i1.190.

[8] L. Bachtiar and M. Mahradianur, "Analisis Data Mining Menggunakan Metode Algoritma C4.5 Menentukan Penerima Bantuan Langsung Tunai," *J. Inform.*, vol. 10, no. 1, pp. 28–36, 2023, doi: 10.31294/inf.v10i1.15115.

[9] Normah, B. Rifai, S. Vambudi, and R. Maulana, "Analisa Sentimen Perkembangan Vtuber Dengan Metode Support Vector Machine Berbasis SMOTE," *J. Tek. Komput. AMIK BSI*, vol. 8, no. 2, pp. 174–180, 2022, doi: 10.31294/jtk.v4i2.

[10] M. Idhom, D. A. Prasetya, P. A. Riyantoko, T. M. Fahrudin, and A. P. Sari, "Pneumonia Classification Utilizing VGG-16 Architecture and Convolutional Neural Network Algorithm for Imbalanced Datasets," *TIERS Inf. Technol. J.*, vol. 4, no. 1, pp. 73–82, 2023, doi: 10.38043/tiers.v4i1.4380.

Ahmad Ardhy Ansyah lahir di Surabaya, 03 Juli 2002. Sedang menempuh gelar sarjana di Program Studi Sains Data di UPN "Veteran" Jawa Timur. Minat penelitian pada bidang Machine Learning.

Alamat Email: 20083010008@student.upnjatim.ac.id

Tresna Maulana Fahrudin lahir di Gresik, menyelesaikan jenjang Diploma-4 dan Magister Terapan di Politeknik Elektronika Negeri Surabaya. Saat ini bekerja sebagai dosen tetap di program studi S-1 Sains Data UPN "Veteran" Jawa Timur berperan aktif sebagai Tim Kurikulum, dan PIC MBKM Program Studi Sains Data. Minat penelitian pada bidang Big Data, Data Mining, dan Machine Learning.

Alamat Email: tresna.maulana.ds@upnjatim.ac.id

Dwi Arman Prasetya lahir di Surabaya. memperoleh gelar Sarjana dari Institut Teknologi Sepuluh November pada 2004, Magister dari Universitas Brawijaya pada 2010, Doktor dari University of Tokushima pada 2013 . Saat ini, bekerja di Program Studi Sains Data, UPN "Veteran" Jawa Timur, sebagai Koordinator Program Studi dan Dosen.

Alamat Email: arman.prasetya.sada@upnjatim.ac.id