INFOTECH journal
ISSN : 2460-1861 (Print), 2615-4250 (Online)
Vol. 10 No. 1 Juni 2024, pp. 84-89

PERBANDINGAN METODE K-NN DAN *RANDOM FOREST* PADA KLASIFIKASI MAHASISWA BERPOTENSI *DROPOUT*

Muhammad Maulana Rofi[1], Foni Agus Setiawan[2], Freza Riana[3]
[1,2,3]Teknik Informatika, Fakultas Teknik & Sains, Universitas Ibn Khaldun
Email: muhammadrofi726@gmail.com

ABSTRACT
Universities are responsible for providing the best education to produce quality individuals. A high dropout rate can damage accreditation. A model was developed using K-Nearest Neighbor (K-NN) and Random Forest to classify dropout cases. Random Forest has higher accuracy (99.05%) than K-NN (98.10%). The Active Percentage attribute stands out as the most influential factor in classifying potential dropouts, according to the Random Forest algorithm. This indicates the importance of active engagement in minimizing the risk of dropping out.

Keywords: Dropout, K-Nearest Neighbor(K-NN), Random Forest.

ABSTRAK
Perguruan tinggi bertanggung jawab memberikan pendidikan terbaik untuk menghasilkan individu berkualitas. Tingginya angka drop out dapat merusak akreditasi. Model dikembangkan menggunakan K-Nearest Neighbor (K-NN) dan Random Forest untuk mengklasifikasikan kasus drop out. Random Forest memiliki akurasi lebih tinggi (99.05%) dibanding K-NN (98.10%). Atribut Persentase Aktif menonjol sebagai faktor paling berpengaruh dalam mengklasifikasikan siswa yang berpotensi putus sekolah, menurut algoritma Random Forest. Ini menandakan pentingnya keterlibatan aktif dalam meminimalkan risiko drop out.

Kata Kunci: *Dropout, K-Nearest Neighbor(K-NN), Random Forest.*

Riwayat Artikel :
Tanggal diterima : 26-02-2024
Tanggal revisi : 05-03-2024
Tanggal terbit : 07-03-2024

DOI :
https://doi.org/10.31949/infotech.v10i1.8856

1. PENDAHULUAN

1.1. Latar Belakang

Perguruan tinggi berperan sebagai lembaga pendidikan bagi mahasiswa. Perguruan tinggi memiliki tanggung jawab untuk menyelenggarakan pendidikan yang unggul bagi mahasiswa agar dapat menghasilkan individu yang berkualitas dalam hal sumber daya manusia[1]. Keberadaan mahasiswa yang mengalami keterlambatan dalam menyelesaikan studi atau tidak lulus tepat waktu merupakan salah satu hambatan bagi kemajuan perguruan tinggi tersebut[2].

Program Studi Teknik Informatika Universitas Ibn Khaldun merupakan salah satu jurusan yang memiliki nilai angka drop out yang cukup tinggi. Dari data yang diketahui bahwa jumlah mahasiswa drop out yang cukup tinggi bisa mengakibatkan terjadinya penurunan akreditas di Program Studi tersebut. Klasifikasi diartikan sebagai suatu metode pengelompokkan objek berdasarkan karakteristik yang dimiliki oleh objek tersebut[3]. Dari banyaknya kategori, klasifikasi digunakan untuk mendeskripsikan dan memisahkan data satu dengan data yang lainnya[4]. Cara ini untuk mendapatkan informasi penting dan bermanfaat bagi suatu organisasi, seperti perguruan tinggi.

Dalam penelitian ini, algoritma model yang akan digunakan yaitu K-Nearest Neighbor (K-NN) dan Random Forest dalam hal klasifikasi. Penelitian ini bertujuan untuk mengetahui algoritma terbaik berdasarkan performa tertinggi dalam mengklasifikasikan mahasiswa berpotensi drop out dan juga mengetahui faktor yang paling berpengaruh sehingga dapat dijadikan acuan untuk mengurangi mahasiswa drop out.

1.2. Tinjuan Pustaka

1. Mahasiswa

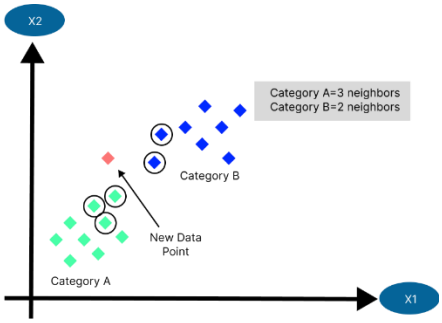
Mahasiswa adalah orang yang belajar di perguruan tinggi, baik di universitas, atau institut. Mahasiswa merupakan orang yang telah lulus dari pendidikan sekunder (SMA atau sederajat) dan memilih untuk melanjutkan pendidikan ke jenjang yang lebih tinggi. Sebagaimana tercantum dalam Undang-Undang Republik Indonesia Nomor 12 Tahun 2012 tentang Pendidikan Tinggi Pasal 1 Ayat (1) yang berbunyi: “Mahasiswa adalah peserta didik pada jenjang Pendidikan Tinggi”. Mahasiswa dapat menyelesaikan program Pendidikan sesuai dengan kecepatan belajar masing-masing dan tidak melebihi ketentuan batas waktu yang ditetapkan oleh Perguruan Tinggi[5].

2. Klasifikasi

Klasifikasi merupakan suatu teknik menemukan kumpulan pola atau fungsi yang mendeskripsikan serta memisahkan kelas data yang satu dengan yang lainnya untuk menyatakan objek tersebut masuk pada kategori tertentu dengan melihat pada kelakuan dan atribut dari kelompok yang telah didefinisikan. Teknik ini dapat memberikan klasifikasi pada data baru dengan memanipulasi data yang ada yang telah diklasifikasi dan dengan menggunakan hasilnya untuk memberikan sejumlah aturan[6].

3. K-Nearest Neighbor

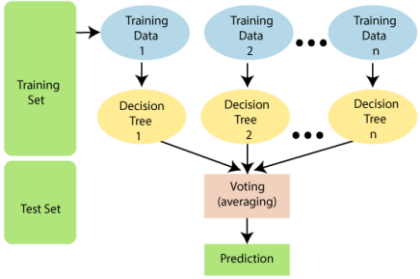
Algoritma K-Nearest Neighbor (K-NN) adalah metode yang sering digunakan untuk klasifikasi, meskipun juga dapat digunakan untuk estimasi dan prediksi. KNN merupakan metode berdasarkan analogi penatihan, di mana himpunan data pelatihan disimpan, sehingga klasifikasi untuk rekaman baru yang tidak diklasifikasikan dapat ditemukan dengan membandingkannya dengan rekamanrekaman paling mirip dalam himpunan pelatihan. Penentuan nilai k memegang peran sentral dalam proses klasifikasi K-NN. Dalam konteks K-NN, nilai k mengindikasikan jumlah tetangga terdekat yang ikut berpartisipasi dalam memprediksi label kelas pada data uji. Fungsi dalam menentukan jarak yang paling umum digunakan adalah euclidean distance[7].



Gambar 1. Konsep sederhana K-NN

4. Random Forest

Random forest merupakan algoritma dalam machine learning yang digunakan untuk pengklasifikasian dataset dalam jumlah besar. Random forest adalah pengklasifikasi yang berisi sejumlah pohon keputusan pada berbagai subset dari kumpulan data yang diberikan dan mengambil rata-rata untuk meningkatkan akurasi prediktif dari kumpulan data tersebut. Random forest bekerja dalam dua fase, pertama adalah membuat hutan acak dengan menggabungkan N pohon keputusan, dan kedua membuat prediksi untuk setiap pohon yang dibuat pada fase pertama. Proses ini terkadang disebut "feature bagging". Alasan untuk melakukan ini adalah korelasi dari pohon-pohon dalam sampel bootstrap biasa. Jika satu atau 9 beberapa fitur adalah prediktor yang sangat kuat untuk kelas keluaran, fitur-fitur ini akan dipilih dalam banyak pohon, menyebabkan mereka menjadi berkorelasi[8].



Gambar 2. Cara kerja Random Forest

5. Confusion Matrix

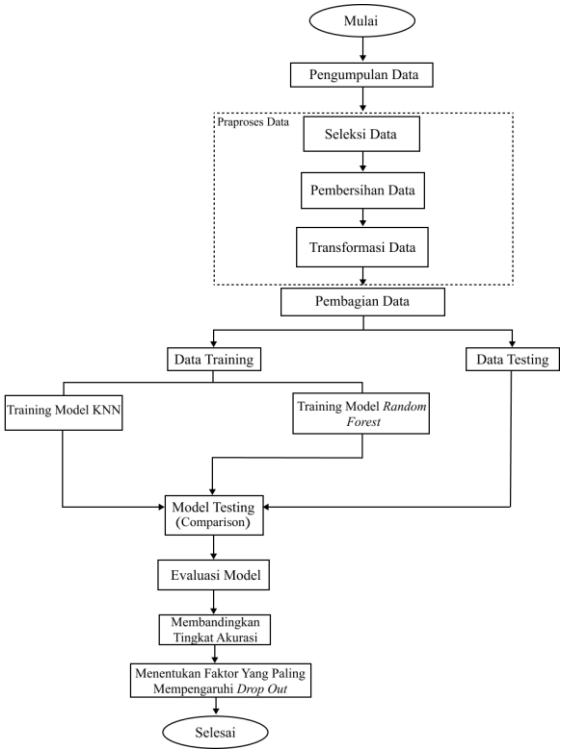
Confusion matrix adalah suatu teknik yang digunakan untuk mengevaluasi dan menggambarkan kinerja dari proses klasifikasi yang dihasilkan oleh model prediksi. Ada beberapa metode yang termasuk dalam confusion matrix, seperti akurasi, tingkat kesalahan (error rate), sensitivitas atau recall, spesifisitas, presisi, dan tingkat false positive[9].

Tabel 1. Confusion Matrix

Data Aktual	Klasifikasi	
	DO	**Tidak**
DO	TP (*True Positive*)	FN (*False Negative*)
Tidak	FP (*False Positive*)	TN (*True Negative*)

2. METODOLOGI PENELITIAN

Adapun tahapan penelitian pada penelitian ini dapat dilihat pada Gambar 3.



Gambar 3. Tahapan Penelitian

2.1. Pengumpulan Data

Pada tahap ini, proses pengumpulan data dilakukan dengan mengambil data dari UKSI Universitas Ibn Khaldun Bogor yang terdiri dari data akademik mahasiswa S1 Program Studi Teknik Informatika dari angkatan 2016-2021 pada awal tahun 2022.

2.2. Praproses Data

Data Preprocessing merupakan proses untuk mempersiapkan data yang akan digunakan pada penelitian ini, agar data yang diperoleh dapat diproses dengan baik oleh model yang akan dibuat menggunakan algoritma K-Nearest Neighbor dan Random Forest. Beberapa tahapan yang akan dilakukan dalam praproses data adalah sebagai berikut:

a. Data Selection

Data selection atau seleksi data adalah tahap pemilihan data yang sesuai untuk proses klasifikasi. Data dalam database yang diperoleh tidak semuanya dipakai, hanya data yang sesuai untuk diklasifikasi yang akan diambil.

b. Data Cleaning

Data cleaning merupakan tahap pembersihan data yang mana dalam penelitian ini terdapat data kosong (missing value) sehingga perlu dihilangkan (dihapus).

c. Data Transformation

Data transformation atau transformasi data adalah tahap pengubahan skala data asli menjadi bentuk lain. Pada penelitian ini terdapat data kategorikal sehingga perlu diubah kedalam bentuk angka agar dapat diproses pada tahap klasifikasi.

2.3. Pembagian Data

Pada tahap ini data akan dibagi menjadi 2 bagian dengan proporsi tertentu diantaranya yaitu:

1. Data training merupakan data latih yang akan digunakan sebagai data pembelajaran algoritma dalam menentukan pola/model data.
2. Data testing merupakan data uji yang akan digunakan untuk menghasilkan peforma dari algoritma yang digunakan.

2.4. Variabel Penelitian

Variabel penelitian dapat dilihat pada Tabel 2.

Tabel 2. Variabel Penelitian

No	Variabel	Keterangan
1	Indeks Prestasi Kumulatif (IPK)	Nilai prestasi belajar secara komulatif yang memiliki rentang nilai 0-4
2	Satuan Kredit Semester (SKS) Total	Jumlah mata kuliah yang sudah di kontrak
3	Penghasilan Orang Tua	Keterangan penghasilan orang tua mahasiswa
4	Jalur Bayaran	Keterangan pembayaran mahasiswa melalui mandiri atau beasiswa
5	Status Bayaran	Keterangan status bayaran mahasiswa
6	Persentase Aktif	Keterangan persentase aktif pada setiap masing-masing mahasiswa

3. PEMBAHASAN

Pada bab ini, dijelaskan urutan proses klasifikasi mahasiswa berpotensi drop out menggunakan algoritma K-Nearest Neighbor dan Random Forest, mengukur hasil akurasi dari model yang dibuat, mengukur performa model yang dibangun serta menentukan faktor yang paling berpengaruh dari algoritma yang memiliki akurasi terbaik. Proses pembuatan model klasifikasi dibantu dengan library python yaitu Scikit-learn.

3.1. Skenario Percobaan Model

Dalam penelitian ini, dilakukan percobaan dan skenario pada Model K-NN dan Random Forest yang telah dibuat. Skenario-skrenario ini dirancang untuk mengidentifikasi model optimal melalui penentuan hyperparameter yang dioptimalkan.

a. K-Nearest Neighbor

Pada percobaan ini, digunakan hyperparameter yang mengacu pada jumlah tetangga κ terdekat yang akan digunakan untuk mengambil keputusan prediksi. Pemilihan nilai κ yang tepat bisa mempengaruhi performa model. Dalam menentukan jarak pada penelitian ini menggunakan Euclidean Distance. Nilai k yang digunakan dapat dilihat pada Tabel dibawah ini.

Tabel 3. Parameter K-NN

Parameter	Nilai Parameter
κ	1, 3, 5

b. Random Forest

Pada percobaan ini, *hyperparameter* yang digunakan ialah *n_estimators* dan *max_depth. n_estimators* ini untuk menentukan berapa banyak pohon keputusan yang akan dibangun dalam *ensemble*, sedangkan *max_depth* ini untuk mengatur kedalaman maksimal dari setiap pohon keputusan dalam *ensemble.*

Tabel 4. Parameter Random Forest

Parameter	Nilai Parameter
n_estimators	100, 200, 300
max_depth	1, 2, 3, 4, 5, 6

3.2. Evaluasi Model

Pada evaluasi model ini, peneliti menggunakan confusion matrix untuk mengukur tingkat ketinggian dari performa masing-masing model. Confusion matrix menghasilkan berupa nilai akurasi, presisi, recall dan f1-score.

a. Evaluasi Model K-NN

Evaluasi model K-NN ini dilakukan dengan melihat confusion matrix dan evaluasi performa yang terdiri dari akurasi, presisi, recall dan f1-score. Confusion matrix menggunakan data testing sebesar 20% dari total keseluruhan data, yang mana jumlah data testing yang digunakan sebanyak 105 data mahasiswa untuk setiap masing-masing kelas baik drop out maupun tidak drop out. Nilai tetangga κ terdekat yang digunakan untuk perbandingan merupakan dari percobaan parameter yang menghasilkan performa tertinggi. Untuk evaluasi performa dari masingmasing parameter dapat dilihat pada Tabel dibawah ini.

Tabel 5. Evaluasi Performa K-NN

Nilai κ =	Performa			
	Akurasi	**Presisi**	***Recall***	***F1-Score***
1	98.10	94.47	94.47	94.47
3	98.10	94.47	94.47	94.47
5	96.19	86.96	93.42	89.85

Dari hasil evaluasi performa diatas, didapat bahwa parameter dengan nilai κ =1 dan 3 memiliki nilai performa paling tinggi. Namun demikian, κ =1 lebih rentan terhadap pengaruh *outlier*, karena prediksi hanya didasarkan pada satu titik data terdekat tanpa mempertimbangkan tetangga yang lain yang mungkin memiliki nilai target yang lebih sesuai. Maka dari itu, κ =3 dipilih sebagai parameter untuk model yang telah dibangun. Model evaluasi yang digunakan ialah *confusion matrix* yang menghasilkan nilai TP (*True Positive*), TN (*True Negative*), FP (*False Positive*), dan FN (*False Negatif*).

Tabel 6. Confusion Matrix K-NN

Data Sebenarnya	Klasifikasi	
	DO	**Tidak**
DO	9 (TP)	1 (FN)
Tidak	1 (FP)	94 (TN)

Pada Tabel 6 didapat hasil confusion matrix data testing dengan parameter κ =3. Dari hasil tersebut diketahui bahwa sebanyak 9 data mahasiswa DO berhasil diprediksi dengan benar oleh model, sedangkan hanya ada 1 data mahasiswa DO yang salah diprediksi oleh model dan menjadikannya masuk kedalam kelas Tidak DO. Sedangkan untuk kelas mahasiswa yang Tidak DO sebanyak 94 data mahasiswa yang berhasil diprediksi oleh model dan sebanyak 1 data mahasiswa 29 salah diprediksi oleh model yang menjadikannya masuk kedalam kelas mahasiswa DO.

b. Evaluasi Model Random Forest

Evaluasi model Random Forest ini dilakukan dengan melihat confusion matrix dan evaluasi performa yang terdiri dari akurasi, presisi, recall dan f1-score. Confusion matrix menggunakan data testing sebesar 20% dari total keseluruhan data, yang mana jumlah data testing yang digunakan sebanyak 105 data mahasiswa untuk setiap masing-masing kelas baik drop out maupun tidak drop out. n_estimators dan max_depth yang digunakan untuk perbandingan merupakan dari percobaan setiap parameter yang menghasilkan performa tertinggi. Untuk evaluasi performa dari masing-masing parameter dapat dilihat pada Tabel dibawah ini.

Tabel 7. Evaluasi Performa Random Forest

n_estimators	*max_depth*	Performa			
		Akurasi	**Presisi**	***Recall***	***F1-Score***
100	1	87.62	64.99	66.32	65.61
	2	95.24	90.82	79.47	84
	3	95.24	87.33	83.95	85.53
	4	95.24	87.33	83.95	85.53
	5	96.19	88.95	88.95	88.95
	6	97.14	90.38	93.95	92.06
200	1	90.48	45.24	50	47.50
	2	96.19	92.20	84.47	87.85
	3	95.24	87.33	83.95	85.53
	4	95.24	87.33	83.95	85.53
	5	97.14	90.38	93.95	92.06
	6	99.05	99.48	95	97.11
300	1	90.48	45.24	50	47.50
	2	96.19	92.20	84.47	87.85
	3	95.24	87.33	83.95	85.53
	4	95.24	87.33	83.95	85.53
	5	96.19	88.95	88.95	88.95
	6	99.05	99.48	95	97.11

Dari hasil evaluasi performa diatas, didapat bahwa parameter dengan *n_estimators* = 200, 300 dan *max_depth* = 6 memiliki performa paling tinggi. Namun demikian, *n_estimators* = 300 kurang unggul pada pengujian di parameter *max_depth* = 5 dan menjadikan *n_estimators* = 200 sebagai parameter terbaik untuk model yang telah dibangun. Untuk *confusion matrix* dari parameter yang telah ditentukan dapat dilihat pada Tabel 8.

Tabel 8. Confusion Matrix Random Forest

Data Sebenarnya	**Klasifikasi**	
	DO	**Tidak**
DO	9 (TP)	1 (FN)
Tidak	0 (FP)	95 (TN)

Pada Tabel 8 didapat hasil *confusion matrix* data *testing* dengan parameter *n_estimators* = 200 dan *max_depth* = 6. Dari hasil tersebut diketahui bahwa sebanyak 9 data mahasiswa DO berhasil diprediksi dengan benar oleh model, sedangkan hanya ada 1 data mahasiswa DO yang salah diprediksi oleh model dan menjadikannya masuk kedalam kelas Tidak DO. Sedangkan untuk kelas mahasiswa yang Tidak DO sebanyak 95 data mahasiswa yang berhasil diprediksi oleh model dan tidak ada data mahasiswa Tidak DO yang salah diprediksi oleh model yang menjadikannya masuk kedalam kelas mahasiswa DO.

4. KESIMPULAN

Dari penelitian yang telah dilakukan telah diperoleh kesimpulan diantaranya:

1. Perbandingan algoritma K-NN dan *Random Forest* menghasilkan nilai akurasi masing-masing sebesar 98.10% dan 99.05%. Dari hasil yang diperoleh, algoritma *Random Forest* memiliki hasil peforma lebih tinggi dari algoritma K-NN, sehingga algoritma *Random Forest* lebih baik daripada algoritma K-NN dalam mengklasifikasi mahasiswa berpotensi *drop out*.
2. Berdasarkan perhitungan seberapa banyak pengurangan impurity, diperoleh hasil bahwa atribut Persentase Aktif merupakan nilai tertinggi yang berarti Persentase Aktif merupakan faktor yang paling berpengaruh dalam klasifikasi mahasiswa berpotensi *drop out.*

Berisi berbagai kesimpulan yang diambil berdasarkan penelitian yang telah dilakukan.Berisi pernyataan singkat tentang hasil yang disarikan dari pembahasan. Saran dapat dituliskan pada bagian paling akhir.

PUSTAKA

[1] I. P. Ramayasa, “Perancangan Sistem Klasifikasi Mahasiswa Drop Out Menggunakan Algoritma K-Nearest

Neighbor," *Semin. Nas. Sist. Inf. dan Teknol. Inf. 2018*, vol. 1, no. 1, pp. 585–589, 2018.

[2] S. Samasil, Y. Yuyun, and H. Hazriani, "Klasifikasi Mahasiswa Berpotensi Drop Out Menggunakan Algoritma Naive Bayes Dan Decision Tree," *J. Ilm. Ilmu Komput.*, vol. 8, no. 2, pp. 108–114, 2022, doi: 10.35329/jiik.v8i2.242.

[3] F. A. D. Aji Prasetya Wibawa, Muhammad Guntur Aji Purnama, Muhammad Fathony Akbar, "Metode-metode Klasifikasi," *Pros. Semin. Ilmu Komput. dan Teknol. Inf.*, vol. 3, no. 1, p. 134, 2018.

[4] R. T. Wahyuni, D. Prastiyanto, and E. Supraptono, "Penerapan Algoritma Cosine Similarity dan Pembobotan TF-IDF pada Sistem Klasifikasi Dokumen Skripsi," *J. Tek. Elektro Univ. Negeri Semarang*, vol. 9, no. 1, pp. 18–23, 2017, [Online]. Available: https://journal.unnes.ac.id/nju/index.php/jte/article/download/10955/6659

[5] Kementrian Hukum dan HAM, "UU RI No. 12/2012 tentang Pendidikan Tinggi," *Undang Undang*, p. 18, 2012.

[6] Aprilla Dennis, "Belajar Data Mining dengan RapidMiner," *Innov. Knowl. Manag. Bus. Glob. Theory Pract. Vols 1 2*, vol. 5, no. 4, pp. 1–5, 2013.

[7] D. T. Larose, *Discovering Knowledge in Data: an Introduction to Data Mining*, vol. 100, no. 472. 2005. doi: 10.1198/jasa.2005.s61.

[8] M. Kantardics, *Data mining: Concept, models, methods, and algorithms*. 2020.

[9] S. Adinugroho and Y. A. Sari, *Implementasi Data Mining Menggunakan Weka*. UB Press, 2018. [Online]. Available: https://books.google.co.id/books?id=p91qDwAAQBAJ&printsec=frontcover&hl=id&source=gbs_ge_summary_r&cad=0#v=onepage&q&f=false