

Analisis Kinerja Algoritma Klasifikasi terhadap Dataset Penerimaan Pegawai *Outsourcing*

Firman Syahputra¹, Dahri Yani Hakim Tanjung², Wiwi Verina³, Ok. Muhammad Ihsan⁴, Rofiqoh Dewi⁵ Andi Sanjaya⁶

¹ Program Studi Manajemen Informatika, AMIK Polibisnis, Perdagangan

^{2,5} Program Studi Bisnis Digital, Universitas Satya Terra Bhinneka, Medan

^{3,4} Program Studi Sistem Informasi, Universitas Potensi Utama, Medan

⁶ Program Studi Sistem Informasi, Universitas Nusa Mandiri, Jakarta

E-mail: 24firman@gmail.com¹, dahritanjung@satyaterrabhinneka.ac.id²,

wiwiverina.azzahra@gmail.com³, okm.ihsan@yahoo.co.id⁴

rofiqohdewi@satyateraabhinneka.ac.id⁵, andisanjaya@gmail.com⁶

Submit : 21 Apr 2025 | **Diterima** : 27 Apr 2025 | **Terbit** : 28 Apr 2025

ABSTRAK

Penelitian ini bertujuan untuk menganalisis dan membandingkan kinerja tiga algoritma klasifikasi, yaitu Decision Tree, Naive Bayes, dan K-Nearest Neighbors (K-NN), dalam memprediksi kelayakan calon pegawai outsourcing berdasarkan data historis. Evaluasi dilakukan menggunakan metrik akurasi, presisi, recall, dan F1-score. Dari hasil pengujian yang telah dilakukan dimana nilai akurasi dari algoritma C4.5 yang dihasilkan adalah 75,00%. Nilai akurasi ini lebih besar daripada model algoritma klasifikasi K-NN sebesar 73,00% dan naïve bayes sebesar 64,00%, namun nilai performa algoritma KNN ini memiliki keunggulan nilai performa akurasi dibandingkan dengan Naïve bayes. Hasil analisis menunjukkan bahwa algoritma Decision Tree memiliki kinerja terbaik dibandingkan dua algoritma lainnya, baik dari sisi akurasi maupun keseimbangan antara presisi dan recall. Hal ini menunjukkan bahwa model Decision Tree cukup efektif dalam menangani data campuran dan menghasilkan prediksi yang andal dalam konteks klasifikasi calon pegawai outsourcing. Dengan hasil ini, diharapkan model klasifikasi berbasis Decision Tree dapat diterapkan dalam sistem pendukung keputusan untuk meningkatkan efisiensi dan akurasi dalam proses rekrutmen tenaga kerja outsourcing. Penelitian ini diharapkan dapat menjadi acuan untuk penelitian lebih lanjut di bidang penerapan data mining dalam manajemen sumber daya manusia.

Kata Kunci : Klasifikasi, C4.5, K-Nearest Neighbors (K-NN), Naïve Bayes, Penerimaan Pegawai Outsourcing.

ABSTRACT

This study aims to analyze and compare the performance of three classification algorithms, namely Decision Tree, Naive Bayes, and K-Nearest Neighbors (K-NN), in predicting the eligibility of prospective outsourcing employees based on historical data. The evaluation was carried out using accuracy, precision, recall, and F1-score metrics. From the results of the tests that have been carried out, the accuracy value of the C4.5 algorithm produced is 75.00%. This accuracy value is greater than the K-NN classification algorithm model of 73.00% and naïve bayes of 64.00%, but the performance value of the KNN algorithm has an advantage in accuracy performance compared to Naïve bayes. The results of the analysis show that the Decision Tree algorithm has the best performance compared to the other two algorithms, both in terms of accuracy and the balance between precision and recall. This shows that the Decision Tree model is quite effective in handling mixed data and producing reliable predictions in the context of classifying prospective outsourcing employees. With these results, it is hoped that the Decision

Tree-based classification model can be applied in decision support systems to improve efficiency and accuracy in the outsourcing workforce recruitment process. This research is expected to be a reference for further research in the field of data mining application in human resource management.

Keywords: Classification, C4.5, K-Nearest Neighbors (K-NN), Naïve Bayes, Outsourcing Employee Acceptance.

PENDAHULUAN

Perusahaan outsourcing memiliki kebutuhan tenaga kerja yang dinamis dan cepat berubah, sehingga proses seleksi pegawai harus dilakukan secara efisien dan akurat. Pemanfaatan teknologi informasi, khususnya data mining, telah menjadi salah satu solusi yang digunakan untuk membantu proses pengambilan keputusan dalam rekrutmen tenaga kerja. Data mining memungkinkan pengguna untuk mengeksplorasi dan menganalisis data dalam jumlah besar guna menemukan pola tersembunyi yang dapat dimanfaatkan untuk prediksi dan klasifikasi [1].

Salah satu metode dalam data mining yang banyak digunakan dalam konteks prediksi dan seleksi adalah teknik klasifikasi. Teknik ini digunakan untuk mengelompokkan data ke dalam kelas-kelas tertentu berdasarkan atribut yang dimiliki. Dalam konteks penerimaan pegawai outsourcing, klasifikasi dapat digunakan untuk memprediksi apakah seorang calon pegawai layak diterima berdasarkan data historis dan atribut tertentu [2].

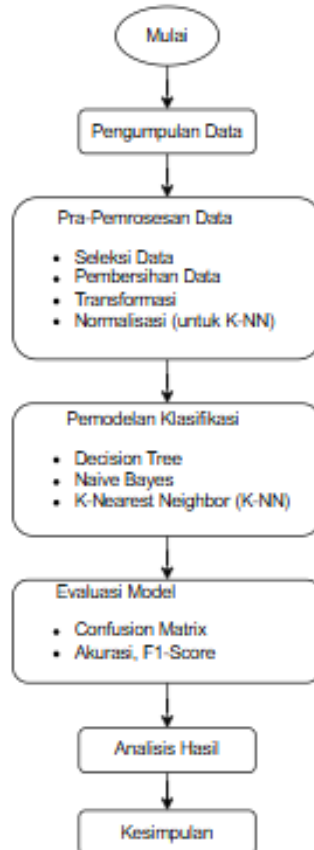
Berbagai algoritma klasifikasi telah digunakan dalam penelitian sebelumnya, di antaranya adalah Decision Tree, Naive Bayes, dan K-Nearest Neighbor (K-NN). Decision Tree membentuk model klasifikasi dalam bentuk pohon keputusan yang mudah diinterpretasikan [3][4]. Naive Bayes menggunakan pendekatan probabilistik berdasarkan Teorema Bayes yang cocok untuk data berskala besar dan kompleksitas rendah [5][6]. Sementara itu, K-NN adalah algoritma berbasis instance learning yang menentukan kelas berdasarkan kedekatan data dengan tetangganya [7][8].

Dari hasil penelitian ini, pengujian yang telah dilakukan terhadap data pengisian ATM membuktikan bahwa algoritma C4.5 mempunyai performa terbaik dengan nilai rata-rata akurasi sebesar 96,00% dibanding algoritma naïve bayes dengan nilai akurasi sebesar 83,00%, selain itu Evaluasi dan validasi pada pengujian ini juga dilihat berdasarkan kurva ROC yang disebut AUC (Area Under the ROC Curve) dimana untuk algoritma C4.5 dengan nilai 0,931 sedangkan naïve bayes 0,894 sehingga algoritma C4.5 dikategorikan dengan klasifikasi Sangat Baik (Excellent Classification) dikarenakan memiliki nilai diantara 0.90-1.00[9] selanjutnya dari hasil penelitian ini menunjukkan bahwa bahwa model klasifikasi C4.5 memiliki nilai performa yang paling baik. Nilai performa yang dihasilkan adalah 54,13% performa akurasi dan 6 detik Execution time[10]

Dalam penelitian ini, dilakukan analisis kinerja terhadap ketiga algoritma tersebut pada dataset penerimaan pegawai outsourcing yang terdiri dari data campuran (numerik dan kategorikal). Evaluasi dilakukan dengan menggunakan metrik akurasi dan F1-score, yang keduanya merupakan ukuran performa yang umum digunakan dalam evaluasi model klasifikasi. Akurasi menunjukkan seberapa banyak prediksi yang benar dibandingkan dengan total data, sedangkan F1-score menggabungkan presisi dan recall untuk memberikan ukuran yang lebih seimbang, khususnya pada data tidak seimbang [11]. Pengujian dilakukan menggunakan perangkat lunak Python karena kemampuannya dalam memfasilitasi proses data mining secara visual serta mendukung berbagai algoritma dan metode evaluasi. Diharapkan hasil dari penelitian ini dapat memberikan gambaran mengenai algoritma klasifikasi yang paling tepat digunakan dalam konteks penerimaan pegawai outsourcing, serta memberikan kontribusi terhadap sistem pendukung keputusan di bidang manajemen sumber daya manusia.

METODE PENELITIAN

Penelitian ini merupakan penelitian kuantitatif dengan pendekatan eksperimental komparatif. Penelitian dilakukan untuk membandingkan kinerja tiga algoritma klasifikasi, yaitu Decision Tree, Naive Bayes, dan K-Nearest Neighbor (K-NN) dalam mengklasifikasikan data penerimaan pegawai outsourcing. Evaluasi dilakukan berdasarkan dua metrik utama, yaitu akurasi dan F1-score.



Gambar 1. Diagram Alir Penelitian

Tahapan penelitian pada Gambar 1 dapat dijelaskan sebagai berikut yaitu :

1. Pengumpulan Data

Data dikumpulkan dari sumber yang relevan, baik dari sistem internal (HR perusahaan) serta data yang berhubungan dengan judul berupa artikel atau jurnal-jurnal terkait. Data ini harus mewakili atribut penting dalam proses penerimaan pegawai outsourcing.

2. Pra-Pemrosesan Data

Tahapan ini sangat krusial karena data mentah sering kali tidak langsung siap digunakan. Proses ini meliputi: *Seleksi Data*: Memilih atribut yang relevan terhadap tujuan klasifikasi; *Pembersihan Data*: Menghapus data duplikat, memperbaiki nilai kosong/missing, atau error input; *Transformasi*: Mengubah format data agar sesuai dengan algoritma; *Normalisasi*: Digunakan khusus untuk algoritma seperti K-NN agar semua atribut memiliki skala yang setara.

3. Pemodelan Klasifikasi

Model dibangun menggunakan algoritma klasifikasi yaitu *Decision Tree*, Menggunakan struktur pohon untuk memisahkan data ke dalam kelas berdasarkan atribut; *Naive Bayes*, Berdasarkan probabilitas dan asumsi independensi antar fitur; *K-Nearest Neighbor (K-NN)*, Mengklasifikasikan data berdasarkan kemiripan dengan k tetangga terdekat.

4. Evaluasi Model

Tahapan evaluasi model bertujuan untuk menilai sejauh mana model klasifikasi yang dibangun mampu melakukan prediksi secara akurat dan seimbang. Evaluasi ini dilakukan dengan menggunakan beberapa metrik utama, yaitu:

a. Confusion Matrix

Confusion matrix merupakan matriks 2x2 yang menggambarkan kinerja model klasifikasi berdasarkan hasil prediksi terhadap nilai aktual. Terdiri atas: *True Positive (TP)*: Data aktual positif dan diprediksi positif; *True Negative (TN)*: Data aktual negatif dan diprediksi negative; *False Positive (FP)*: Data aktual negatif tetapi diprediksi positif (kesalahan tipe I); *False Negative (FN)*: Data aktual positif tetapi diprediksi negatif (kesalahan tipe II) [12]

b. Akurasi (Accuracy)

Akurasi adalah rasio dari jumlah prediksi benar terhadap jumlah total data. Rumusnya:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

Akurasi memberi gambaran umum seberapa banyak prediksi yang tepat dari keseluruhan data, tetapi bisa menyesatkan jika data tidak seimbang (imbalanced)[13]

c. Presisi (Precision)

Presisi mengukur proporsi prediksi positif yang benar-benar relevan. Rumusnya:

$$Precision = \frac{TP}{TP+FP} \quad (2)$$

Presisi penting ketika biaya kesalahan prediksi positif tinggi, seperti pada proses rekrutmen di mana kesalahan memilih kandidat yang tidak layak bisa berdampak pada kualitas kerja[2]

d. Recall (Sensitivity)

Recall atau sensitivitas mengukur seberapa baik model dalam menangkap seluruh kasus positif aktual. Rumusnya:

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

Recall penting ketika tidak boleh melewatkan kasus positif, seperti dalam rekrutmen kandidat yang benar-benar layak[12]

e. F1-Score

F1-score adalah rata-rata harmonis antara presisi dan recall. Rumusnya:

$$F1 = 2 \times \frac{Presisi \times Recall}{Presisi + Recall} \quad (4)$$

F1-score menjadi metrik yang lebih seimbang saat harus mengukur performa model pada dataset yang tidak seimbang.

5. Analisis Hasil

Tahapan ini membahas hasil dari proses klasifikasi dan evaluasi model. Peneliti akan menginterpretasikan kelebihan dan kelemahan masing-masing algoritma.

6. Kesimpulan

Merangkum temuan penelitian, algoritma terbaik, serta rekomendasi penggunaan model klasifikasi untuk mendukung keputusan dalam proses seleksi pegawai outsourcing.

HASIL DAN PEMBAHASAN

Dataset yang digunakan dalam penelitian ini merupakan data riwayat penerimaan pegawai outsourcing yang terdiri dari 500 entri. Setiap entri memiliki sejumlah atribut seperti usia, jenis kelamin, pendidikan terakhir, nilai tes tertulis, hasil wawancara, dan status diterima atau tidak diterima sebagai label kelas. Tipe data dalam dataset terdiri dari:

Tabel 1. Tipe data dalam Dataset

Tipe Data	Keterangan
Numerik	Usia, nilai tes, nilai wawancara
Kategorikal	Jenis kelamin, pendidikan
Label	Status (diterima/tidak diterima)

Data telah melalui proses pra-pemrosesan seperti penanganan *missing values*, normalisasi (khusus untuk K-NN), dan transformasi data kategorikal ke numerik (label encoding). Selanjutnya Evaluasi dilakukan berdasarkan dua metrik utama, yaitu akurasi dan F1-score. Berikut adalah contoh sampel data latih yang ditunjukkan pada Tabel 2 di bawah ini

Tabel 2. Dataset

ID	Usia	Jenis Kelamin	Pendidikan Terakhir	Pengalaman Kerja	Nilai Tes Tertulis	Nilai Wawancara	Status Penerimaan
PEL0001	41	Pria	D3	1	44	85	Tidak Diterima
PEL0002	24	Pria	SMA	4	55	71	Tidak Diterima
PEL0003	21	Pria	D3	7	89	60	Tidak Diterima
PEL0004	44	Pria	S1	10	81	70	Diterima
PEL0005	29	Wanita	D3	6	63	62	Tidak Diterima
PEL0006	28	Pria	S1	7	63	75	Tidak Diterima
PEL0007	28	Pria	D3	9	64	99	Diterima
PEL0008	25	Wanita	D3	7	100	99	Diterima
PEL0009	44	Wanita	S1	6	76	90	Diterima
PEL0010	24	Pria	S1	4	42	86	Tidak Diterima
PEL0011	42	Wanita	SMA	3	78	57	Tidak Diterima
PEL0015	39	Pria	S1	6	100	83	Diterima
...
...
PEL0499	38	Wanita	SMA	0	53	68	Tidak Diterima
PEL0500	34	Pria	S1	2	84	94	Diterima

Dari hasil pengujian yang dilakukan terhadap dataset dengan bantuan perangkat lunak python maka menghasilkan nilai confusion matrix sebagai berikut :

==== Decision Tree ====

Akurasi: 0.75

Classification Report:

	precision	recall	f1-score	support
0	0.06	0.11	0.07	9
1	0.90	0.81	0.86	91
accuracy			0.75	100
macro avg	0.48	0.46	0.46	100
weighted avg	0.83	0.75	0.79	100

==== KNN ====

Akurasi: 0.73

Classification Report:

	precision	recall	f1-score	support
0	0.09	0.22	0.13	9
1	0.91	0.78	0.84	91
accuracy			0.73	100
macro avg	0.50	0.50	0.48	100
weighted avg	0.84	0.73	0.78	100

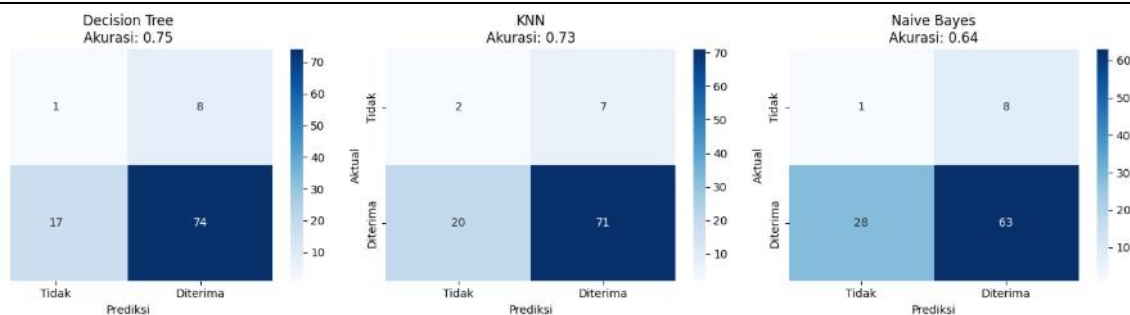
==== Naive Bayes ====

Akurasi: 0.64

Classification Report:

	precision	recall	f1-score	support
0	0.03	0.11	0.05	9
1	0.89	0.69	0.78	91
accuracy			0.64	100
macro avg	0.46	0.40	0.42	100
weighted avg	0.81	0.64	0.71	100

untuk hasil akurasi yang dilakukan dari pengujian ketiga algoritma dapat dilihat pada gambar 2 dimana nilai akurasi dari algoritma C4.5 yang dihasilkan adalah 75,00%. Nilai akurasi ini lebih besar daripada model algoritma klasifikasi K-NN sebesar 73,00% dan naïve bayes sebesar 64,00%, namun nilai performa algoritma KNN ini memiliki keunggulan nilai performa akurasi dibandingkan dengan Naïve bayes



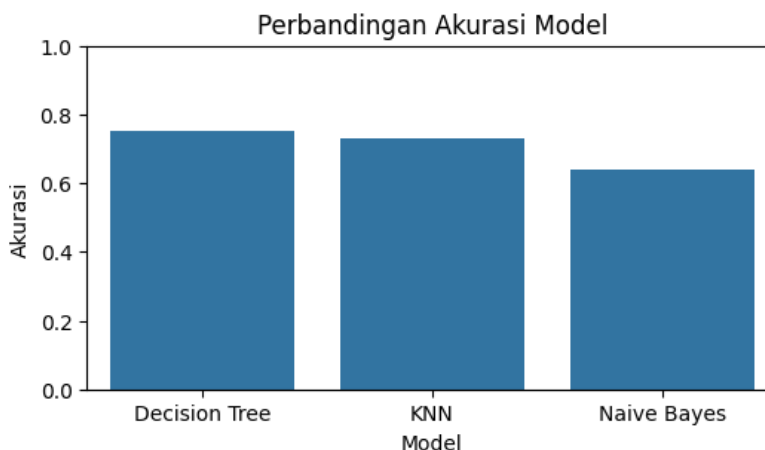
Gambar 2. Hasil Confusion Matrix dari 3 Algoritma Klasifikasi

Hasil Performa Model Algoritma

Tabel 3. Daftar Hasil performa Algoritma

No	Model Algoritma	Parameter Performa			
		Accuracy (%)	Precision	Recall	F1-score
1	Decision Tree	0.75	0.90	0.81	0.85
2	K-NN	0.73	0.91	0.78	0.84
3	Naive Bayes	0.64	0.89	0.69	0.78

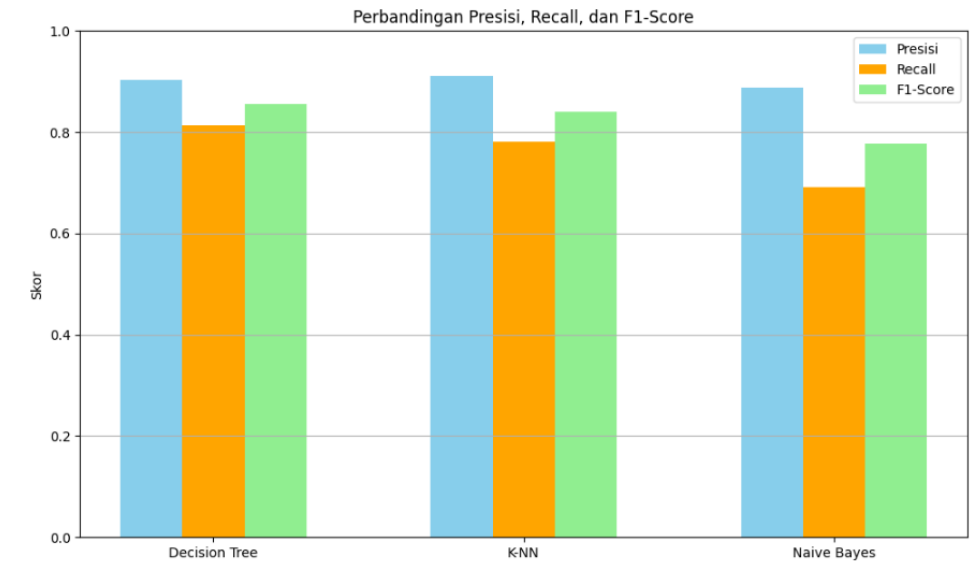
Tabel 3 merepresentasikan hasil pengujian performa algoritma yang dievaluasi berdasarkan nilai performa akurasi, presisi, recall dan f1-score, yang dapat dilihat pada grafik berikut:



Gambar 3. Grafik Hasil Performa Akurasi

Gambar 3 membuktikan bahwa nilai performa accuracy tertinggi didapat dari implementasi algoritma klasifikasi C4.5 dengan nilai akurasi sebesar 75,00%. Sedangkan nilai performa akurasi yang terendah diperoleh dari model algoritma klasifikasi naïve bayes dengan 64,00%. Namun performa akurasi ini memiliki nilai yang berbeda sangat tipis dengan pengimplementasian model algoritma K-NN yaitu 74,00%. Ini membuktikan bahwa dengan jumlah atribut 7 atribut dengan record sebanyak 500, model algoritma C4.5 memiliki nilai performa yang paling baik.

Perbandingan Perfoma Presisi, Recall dan *F1-Score*



Gambar 4. Grafik Hasil Perbandingan Perfoma Presisi, Recall dan *F1-Score*

Evaluasi performa model klasifikasi dilakukan dengan membandingkan nilai presisi, recall, dan F1-score dari tiga algoritma yaitu Decision Tree, K-Nearest Neighbor (K-NN), dan Naive Bayes. Hasil evaluasi ditampilkan pada Gambar 4 yang menunjukkan kinerja masing-masing algoritma terhadap data penerimaan pegawai outsourcing.

1. Decision Tree

Algoritma Decision Tree menunjukkan kinerja yang seimbang antara presisi, recall, dan F1-score. Presisi mencapai nilai tertinggi kedua (0.90), menandakan bahwa prediksi positif yang dihasilkan sebagian besar benar. Recall sebesar 0.82 menunjukkan bahwa model mampu mengenali sebagian besar data pegawai yang benar-benar diterima. F1-score (0.85) yang tinggi juga menunjukkan bahwa keseimbangan antara presisi dan recall cukup optimal. Hal ini menunjukkan bahwa Decision Tree cukup andal dalam klasifikasi data dengan distribusi kelas yang tidak seimbang.

2. K-Nearest Neighbor (K-NN)

K-NN mencatatkan presisi tertinggi (0.91) di antara ketiga algoritma. Hal ini menunjukkan bahwa model sangat efektif dalam menghindari kesalahan dalam memprediksi data pegawai yang diterima (false positive rendah). Namun, recall-nya sedikit lebih rendah (0.78), menandakan bahwa model cenderung lebih selektif dalam mengidentifikasi data positif, sehingga cenderung menghasilkan false negative yang lebih banyak. Meskipun demikian, nilai F1-score (0.84) tetap menunjukkan performa yang kompetitif.

3. Naive Bayes

Naive Bayes menunjukkan kinerja terendah di antara ketiga algoritma, khususnya pada metrik recall (0.70). Hal ini berarti bahwa model sering gagal mengenali data pegawai yang seharusnya diterima (false negative tinggi). Meskipun presisi-nya masih cukup tinggi (0.89), nilai F1-score yang paling rendah (0.78) menunjukkan bahwa model ini kurang mampu menjaga keseimbangan antara presisi dan recall pada dataset ini.

Secara keseluruhan, Decision Tree dapat disimpulkan sebagai algoritma dengan performa paling stabil dan seimbang, cocok untuk digunakan dalam klasifikasi calon pegawai outsourcing berdasarkan data yang digunakan dalam penelitian ini. K-NN bisa menjadi pilihan jika presisi menjadi prioritas utama, sedangkan Naive Bayes kurang direkomendasikan karena recall-nya yang rendah pada dataset ini.

KESIMPULAN

Berdasarkan hasil dan pembahasan yang telah diuraikan pada bab sebelumnya maka dapat diambil beberapa kesimpulan sebagai berikut :

1. Decision Tree merupakan algoritma dengan performa terbaik secara keseluruhan, dengan nilai akurasi, presisi, recall, dan F1-score tertinggi. Model ini mampu menangkap pola kompleks dalam data tanpa membutuhkan banyak pra-proses.
2. K-Nearest Neighbors (K-NN) memberikan hasil yang cukup kompetitif. Meskipun sedikit lebih rendah dari Decision Tree, performanya tetap dapat diandalkan terutama pada kasus dengan distribusi data yang seimbang. Namun, sensitivitas terhadap skala dan jumlah tetangga (k) menjadi perhatian.
3. Naive Bayes menghasilkan performa paling rendah di antara ketiganya. Hal ini diduga karena asumsi independensi antar fitur tidak terpenuhi pada dataset ini, menyebabkan tingginya angka false negative. Meskipun demikian, model ini tetap menarik untuk diterapkan pada skenario dengan kebutuhan proses cepat.

Adapun saran yang dapat diberikan dari penelitian ini untuk pengembangan penelitian berikutnya adalah :

1. Pengembangan Model, Penelitian ini hanya membandingkan tiga algoritma klasifikasi dasar. Untuk penelitian selanjutnya, disarankan untuk menguji algoritma lain yang lebih kompleks seperti Random Forest, Support Vector Machine (SVM), atau Gradient Boosting untuk melihat peningkatan performa klasifikasi.
2. Pengayaan Dataset, Dataset yang digunakan masih terbatas pada sejumlah fitur standar. Akan lebih baik jika data diperluas dengan atribut tambahan seperti soft skill, hasil psikotes, atau riwayat kerja yang lebih detail untuk meningkatkan akurasi prediksi.

DAFTAR PUSTAKA

- J. Han, J. Pei, and M. Kamber, *Data mining: Concepts and techniques*, 3rd ed. Elsevier, 2011.
- I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal, *Data mining: Practical machine learning tools and techniques*, 4th ed. Morgan Kaufmann, 2016.
- J. R. Quinlan, "Improved use of continuous attributes in C4.5," *J. Artif. Intell. Res.*, vol. 4, pp. 77–90, 1996, doi: 10.1613/jair.279.
- F. Syahputra, H. Hartono, and R. Rosnelly, "Penerapan Algoritma C4.5 Dalam Memprediksi Ketersediaan Uang Pada Mesin ATM," *J. Media Inform. Budidarma*, vol. 5, no. 2, p. 556, 2021, doi: 10.30865/mib.v5i2.2933.
- H. Zhang, "The optimality of naive Bayes," in *Proceedings of the Seventeenth International Florida Artificial Intelligence Research Society Conference (FLAIRS)*, 2004, pp. 562–567.
- K. Ahmed and T. Jesmin, "Comparative Analysis of Data Mining Classification Algorithms in Type-2 Diabetes Prediction Data Using WEKA Approach," *Int. J. Sci. Eng.*, vol. 7, no. 2, pp. 155–160, 2014, doi: 10.12777/ijse.7.2.150-154.
- T. Cover and P. Hart, "Nearest neighbor pattern classification," *IEEE Trans. Inf. Theory*, vol. 13, no. 1, pp. 21–27, 1967, doi: 10.1109/TIT.1967.1053964.

"PERBANDINGAN NORMALISASI DATA UNTUK KLASIFIKASI WINE

MENGGUNAKAN ALGORITMA K-NN,” vol. 4, no. 1, pp. 78–82, 2019.

- D. Yani *et al.*, “Analisis Perbandingan Algoritma Klasifikasi Terhadap Data Problem Mesin ATM Dengan Rapidminer Comparative Analysis Of Classification Algorithm On Problem Data Of ATM Machines With Rapid,” vol. 16, no. 2, pp. 188–200, 2024.
- M. Yusa, E. Utami, and E. Luthfi. Taufiq, “Evaluasi Performa Algoritma Klasifikasi Decision Tree ID3, C4.5, dan CART Pada Dataset Readmisi Pasien Diabetes,” *Infosys (Information Syst. J.*, vol. 4, no. 1, pp. 23–34, 2016.
- M. Sokolova and G. Lapalme, “A systematic analysis of performance measures for classification tasks,” *Inf. Process. \& Manag.*, vol. 45, no. 4, pp. 427–437, 2009, doi: 10.1016/j.ipm.2009.03.002.
- P.-N. Tan, M. Steinbach, and V. Kumar, *Introduction to Data Mining*. Pearson Education, 2005.
- J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*, 3rd ed. Morgan Kaufmann, 2012.