ISSN (Online) 2747-0563
Volume 6 Tahun 2026

Perbandingan Algoritma *Gaussian Naïve Bayes* dan *K-Nearest Neighbor* dalam Klasifikasi Dataset *Employee Turnover*

Ridvan Tandiwijaya[1], J.B. Budi Darmawan[2*]

[1] Informatika, Universitas Sanata Dharma Yogyakarta
[1]tandiwijayar@gmail.com

[2] Informatika, Universitas Sanata Dharma Yogyakarta
[2]b.darmawan@usd.ac.id

*Corresponding author email: b.darmawan@usd.ac.id

Abstrak*—*Employee Turnover* merupakan permasalahan penting bagi perusahaan karena berdampak pada peningkatan biaya rekrutmen dan pelatihan serta penurunan produktivitas. Penelitian ini bertujuan untuk membandingkan kinerja algoritma *Gaussian Naïve Bayes* dan *K-Nearest Neighbor* (KNN) dalam klasifikasi dataset *employee turnover*, serta menganalisis pengaruh teknik normalisasi dan standarisasi terhadap performa model. Dataset yang digunakan berasal dari *Kaggle* dengan total 9.540 data karyawan periode 2016-2020. Tahapan penelitian meliputi *preprocessing* data, penyeimbangan kelas menggunakan *Synthetic Minority Over Sampling Technique* (SMOTE), ektraksi fitur dengan *Principal Component Analysis* (PCA), serta optimasi parameter menggunakan *GridSearchCV*. Evaluasi dilakukan menggunakan *Confusion Matrix* dengan metrik akurasi, *precision*, *recall*, dan *F1-Score*. Hasil penelitian menunjukkan bahwa *K-Nearest Neighbor* (KNN) dengan *StandardScaler* merupakan skema terbaik dan secara signifikan mengungguli *Gaussian Naïve Bayes*. Pada skema terbaik tersebut, KNN mencapai akurasi sebesar 85,46%, *precision* 81%, *recall* 92%, *dan F1-Score* 86%, sedangkan *Gaussian Naïve Bayes* hanya mencapai akurasi sebesar 67,18%, *precision* 63%, *recall* 81%, dan *F1-Score* 71%. Penggunaan *StandardScaler* terbukti lebih optimal dibandingkan *Min-Max Scaler* pada kedua algoritma, pada KNN akurasi meningkat dari 80,47% menjadi 85,46%, sedangkan pada *Gaussian Naïve Bayes* meningkat dari 55,86% menjadi 67,18%. Dengan demikian, penggunaan *StandardScaler* pada algoritma KNN terbukti menjadi kombinasi yang paling optimal dalam klasifikasi *employee turnover*.

Kata Kunci— ***employee turnover, klasifikasi, gaussian naïve bayes, k-nearest neighbor, standardscaler.***

I. PENDAHULUAN

Sumber Daya Manusia (SDM) merupakan aset penting bagi perusahaan, sehingga mempertahankan karyawan berkualitas menjadi tantangan utama [1]. Salah satu permasalahan yang sering dihadapi adalah tingginya tingkat *employee turnover*, yaitu kondisi ketika karyawan meninggalkan perusahaan dalam periode tertentu. Tingginya tingkat *employee turnover* dapat memberikan dampak negatif bagi organisasi, seperti meningkatnya biaya perekrutan dan pelatihan serta menurunnya produktivitas.

Beberapa faktor yang mempengaruhi *employee turnover* antara lain kepuasan kerja, komitmen organisasi, stress kerja, dan budaya organisasi [1]. Oleh karena itu, perusahaan perlu mampu memprediksi potensi turnover agar dapat mengambil langkah pencegahan yang tepat.

Prediksi *employee turnover* merupakan proses mengidentifikasi karyawan yang berisiko meninggalkan perusahaan. Salah satu pendekatan yang dapat digunakan adalah klasifikasi, yaitu suatu metode analisis data yang bertujuan untuk mengelompokkan obyek ke dalam kategori tertentu berdasarkan karakteristik atau atribut yang dimilikinya. Prediksi *employee turnover* dapat dilakukan dengan algoritma *machine learning*, seperti *Gaussian Naïve Bayes* dan *K-Nearest Neighbor* (KNN) [2].

Penelitian sebelumnya menunjukkan bahwa *Naïve Bayes* dan KNN memiliki kemampuan yang baik dalam menyelesaikan permasalahan klasifikasi pada bidang pendidikan [3]. Berdasarkan hasil tersebut, menarik untuk dikaji apakah algoritma *Naïve Bayes* dan KNN juga dapat diterapkan secara efektif pada konteks yang berbeda, yakni dalam klasifikasi *employee turnover*.

Dalam penelitian ini penulis tertarik menggunakan algoritma *Gaussian Naïve Bayes* dan KNN untuk mengidentifikasi metode yang paling sesuai dalam klasifikasi data karyawan yang keluar dengan menggunakan dataset karyawan yang keluar dari perusahaan selama rentang waktu 2016-2020.

II. TINJAUAN PUSTAKA DAN LANDASAN TEORI

Penelitian ini tidak terlepas dari berbagai literatur yang digunakan sebagai landasan teori, acuan, serta pedoman dalam pelaksaan dan penyelesaian penelitian.

A. Employee Turnover

Employee turnover dalam organisasi merupakan fenomena alami yang tidak dapat dihindari oleh setiap organisasi, serta tidak terdapat alasan baku yang secara universal menjelaskan penyebab pegawai meninggalkan organisasi. Secara sederhana, *turnover* merupakan keluarnya karyawan dari suatu organisasi. Dalam pengertian yang lebih luas, *turnover* didefinisikan sebagai proses pergantian karyawan dalam perusahaan yang

ISSN (Online) 2747-0563
Volume 6 Tahun 2026

terjadi akibat berbagai faktor, baik yang bersifat sukarela maupun yang tidak bersifat sukarela [1].

B. Machine Learning

Machine learning merupakan cabang kecerdasan buatan (*Artificial Intelligence*) yang berfokus pada pengembangan sistem atau algoritma yang mampu belajar dari data untuk melakukan prediksi atau pengambilan keputusan tanpa pemrograman eksplisit. Teknologi ini mampu mengenali pola dalam data sehingga mendukung efisiensi operasional dan pengambilan keputusan strategis [4]. Secara umum, *machine learning* memiliki dua teknik utama, yaitu *Supervised Learning* dan *Unsupervised Learning*. *Supervised Learning* menggunakan data berlabel untuk mempelajari hubungan input dan output, serta terbagi menjadi metode regresi dan klasifikasi. Pada klasifikasi, data dikelompokkan ke dalam kelas atau kategori tertentu [5].

C. Klasifikasi

Klasifikasi adalah teknik data mining yang digunakan untuk mengalokasikan data rekaman baru ke salah satu kelas yang telah dipilih sebelumnya. Klasifikasi juga digunakan untuk memprediksi label untuk setiap kategori kelas. Selain itu, klasifikasi dapat memodelkan data dengan menggunakan set data pelatihan dan nilai label kelas dalam atribut klasifikasi yang kemudian digunakan untuk mengklasifikasikan data baru [6].

D. Preprocessing

Preprocessing merupakan tahap awal yang penting dalam analisis data untuk memastikan data berada dalam kondisi layak digunakan. Tahap ini bertujuan untuk meningkatkan kualitas data dengan mengurangi *noise*, mengatasi ketidakkonsistenan, menghapus elemen yang tidak relevan, menyeragamkan format, serta mengubah data ke bentuk yang lebih terstruktur. Dengan *preprocessing* yang baik, algoritma pembelajaran mesin dapat bekerja lebih efektif dan menghasilkan performa model yang lebih optimal.
Salah satu proses penting dalam *preprocessing* adalah penyesuaian skala data untuk mengatasi perbedaan rentang nilai antar atribut yang dapat mempengaruhi kinerja algoritma [7]. Teknik umum yang digunakan adalah normalisasi dan standarisasi, yaitu mentransformasikan data ke skala tertentu tanpa menghilangkan informasi penting. Pemilihan teknik yang tepat diharapkan dapat meningkatkan stabilitas perhitungan, mempercepat proses pembelajaran model, serta menghasilkan performa klasifikasi yang lebih optimal.

1) Normalisasi (Min-Max Scaler): Normalisasi dengan *Min-Max Scaler* merupakan teknik *preprocessing* yang digunakan untuk mentransformasi nilai atribut ke rentang 0 hingga 1. Teknik ini bertujuan menyeragamkan skala data tanpa mengubah pola distribusi aslinya, sehingga memudahkan algoritma dalam proses pembelajaran. *Min-Max Scaler* bekerja dengan mengurangi nilai atribut dengan nilai minimum, kemudian membaginya dengan selisih antara nilai maksimum dan minimum, seperti dalam (1) [8].

$$x' = \frac{x - x_{min}}{x_{max} - x_{min}} \quad (1)$$

Keterangan:
x : Nilai data asli
x_{min} : Nilai minimum dari atribut
x_{max} : Nilai maksimum dari atribut
x' : Nilai hasil normalisasi

2) *Standarisasi (StandardScaler)*: Standarisasi dengan menggunakan *StandardScaler* adalah teknik *preprocessing* yang bertujuan untuk mentransformasikan data sehingga memiliki nilai rata-rata (*mean*) sebesar 0 dan standar deviasi sebesar 1. Teknik ini bekerja dengan cara mengurangi setiap nilai atribut dengan nilai rata-rata atribut tersebut, selanjutnya membaginya dengan nilai standar deviasi. Metode ini cukup sensitif terhadap keberadaan nilai pencilan (*outlier*) yang dapat memengaruhi hasil standarisasi karena perhitungannya yang bergantung pada nilai rata-rata dan standar deviasi [8]. Standarisasi dengan *StandardScaler* dapat dirumuskan sebagai (2).

$$y = \frac{x-\mu}{\sigma} \quad (2)$$

Keterangan:
x : Nilai data asli
μ : Nilai rata-rata *(mean)* dari atribut
σ : Nilai standar deviasi dari atribut
y : Nilai hasil standarisasi

E. Balancing (SMOTE)

Synthetic Minority Over-sampling Technique (SMOTE) adalah salah satu metode yang digunakan untuk menangani masalah ketidakseimbangan data (*imbalanced data*). SMOTE bekerja dengan cara membuat data sintetis pada kelas minoritas melalui interpolasi antar data yang berdekatan, sehingga distribusi kelas menjadi lebih seimbang dan model tidak bias terhadap kelas mayoritas. Dengan demikian, SMOTE dapat meningkatkan performa algoritma klasifikasi karena model memiliki kemampuan yang lebih baik dalam mengenali pola pada kelas minoritas [7].

F. Feature Extraction (PCA)

Feature Extraction merupakan suatu proses dalam *machine learning* yang bertujuan mengubah data mentah menjadi representasi yang lebih informatif dan bermakna. Representasi kemudian dapat dimanfaatkan untuk berbagai tugas analisis, termasuk klasifikasi. *Feature Extraction* juga membantu dalam mengurangi dimensi data yang terlalu besar tanpa menghilangkan aspek signifikan yang diperlukan dalam pemodelan [9].
Principal Component Analysis (PCA) merupakan salah satu metode analisis mutivariat yang kerap digunakan dalam

ISSN (Online) 2747-0563
Volume 6 Tahun 2026

Feature Extraction. Dalam penerapannya, PCA menyeleksi atribut yang paling relevan sehingga informasi utama tetap terjaga, sementara variabel yang kurang signifikan dapat dieliminasi. Hal ini menjadikan proses analisis data lebih efisien sekaligus meningkatkan performa algoritma klasifikasi [10].

G. *Grid Search Cross Validation (GridSearchCV)*

GridSearchCV merupakan salah satu teknik optimasi dalam *machine learning* yang digunakan untuk menemukan kombinasi parameter terbaik (optimal) dari suatu algoritma pembelajaran mesin. Metode ini bekerja dengan cara melakukan pencarian menyeluruh (*exhaustive search*) pada ruang parameter (parameter grid) yang telah ditentukan, kemudian mengevaluasi setiap kombinasi parameter menggunakan teknik *cross-validation* [11].

H. *Naïve Bayes*

Naive Bayes merupakan salah satu algoritma klasifikasi yang berlandaskan pada prinsip probabilitas dan statistik. Algoritma ini dikembangkan berdasarkan *Teorema Bayes* yang diperkenalkan oleh *Thomas Bayes*, yang memungkinkan prediksi peluang masa depan berdasarkan pengalaman sebelumnya, seperti dalam (3) [12]. Algoritma ini dikenal sederhana, cepat, dan memiliki akurasi yang cukup baik [6].

$$P(H|X) = \frac{P(X|H)\,P(H)}{P(X)} \qquad (3)$$

Keterangan: *X* adalah data yang belum diketahui kelasnya, *H* adalah hipotesis kelas, *P(H)* probabilitas awal, *P(X|H)* probabilitas data terhadap kelas, dan *P(H|X)* probabilitas akhir.

1) Gaussian Naïve Bayes: Gaussian Naïve Bayes merupakan pengembangan *Naïve Bayes* untuk data numerik atau kontinu. Metode ini mengasumsikan setiap fitur mengikuti distribusi normal (*Gaussian*) pada tiap kelas, sehingga sesuai digunakan pada data kontinu. Perhitungan probabilitas *P(X|H)* dilakukan menggunakan distribusi *Gaussian*, seperti dalam (4) [13].

$$P(x_i \mid y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} exp\left(-\frac{(x_i - \mu_y)^2}{2\sigma_y^2}\right) \qquad (4)$$

Keterangan:
x_i : Nilai fitur ke-i
y : Kelas data
μ_y : Nilai rata-rata fitur pada kelas y
σ_y^2 : Varians fitur pada kelas

Pada *Gaussian Naïve Bayes* parameter yang digunakan adalah *var_smoothing*, yaitu nilai kecil yang ditambahkan pada varians untuk menjaga stabilitas perhitungan dan mencegah pembagian dengan nol [14].

I. *K-Nearest Neighbor (KNN)*

K-Nearest Neighbor (KNN) adalah algoritma klasifikasi dalam data mining yang termasuk metode *supervised learning*. Algoritma ini mengklasifikasikan data baru berdasarkan kedekatan terhadap data latih yang telah memiliki label kelas. KNN termasuk dalam metode *lazy learning*, karena tidak membangun model pada tahap pelatihan, melainkan menggunakan seluruh data latih saat proses prediksi dilakukan. Dengan prinsip ini, KNN menentukan kelas dari data baru dengan mengacu pada tingkat kemiripan fitur antar data [15]. Proses klasifikasi pada KNN dilakukan dengan menentukan nilai *K* sebagai jumlah tetangga terdekat, kemudian menghitung jarak antara data uji dengan seluruh data latih menggunakan metode seperti *Euclidean* atau *Manhattan*. Setelah itu, jarak diurutkan dari yang terkecil, dipilih sebanyak *K* data terdekat, selanjutnya kelas mayoritas dari tetangga tersebut digunakan sebagai hasil prediksi. Persamaan (5) merupakan rumus perhitungan jarak *Euclidean*.

$$d(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \qquad (5)$$

Persamaan (6) merupakan rumus perhitungan jarak *Manhattan*.

$$d(x,y) = \sum_{i=1}^{n}|x_i - y_i| \qquad (6)$$

Keterangan:
d(x, y) = jarak antara data *x* dan data *y*
xi = nilai atribut ke-*i* pada data latih
yi = nilai atribut ke-*i* pada data uji
i = indeks atribut
n = jumlah atribut (dimensi data)

J. *Confusion Matrix*

Confusion Matrix merupakan metode evaluasi untuk menilai kinerja model klasifikasi dengan membandingkan hasil prediksi terhadap kelas sebenarnya [2]. Metode ini menghasilkan empat komponen utama, yaitu *True Positive* (TP), *True Negative* (TN), *False Positive* (FP), dan *False Negative* (FN) seperti pada Tabel I. Berdasarkan komponen tersebut, performa model dapat diukur menggunakan metrik *accuracy*, *precision*, *recall*, dan *F1-score* [16].

TABEL I
CONFUSION MATRIX

	Aktual Ya	**Aktual Tidak**
Prediksi Ya	*True Positive* (TP)	*False Positive* (FP)
Prediksi Tidak	*False Negative* (FN)	*True Negative* (TN)

Keterangan:

TP : Data positif yang diprediksi benar sebagai positif
TN : Data negatif yang diprediksi benar sebagai negatif

ISSN (Online) 2747-0563
Volume 6 Tahun 2026

FP : Data negatif yang salah diprediksi sebagai positif
FN : Data positif yang salah diprediksi sebagai negative

Berikut persamaan untuk menghitung nilai akurasi, presisi, *recall*, dan *F1-score*.

a. Akurasi mengukur seberapa banyak prediksi model yang benar dibandingkan dengan keseluruhan data, seperti dalam (7).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (7)$$

b. Presisi menunjukkan ketepatan model dalam memprediksi kelas positif, yaitu seberapa besar bagian dari data yang diprediksi positif benar termasuk ke dalam kelas positif, seperti dalam (8).

$$Precision = \frac{TP}{TP + FP} \quad (8)$$

c. *Recall* mengukur kemampuan model dalam menemukan seluruh data positif yang sebenarnya ada, seperti dalam (9).

$$Recall = \frac{TP}{TP + FN} \quad (9)$$

d. *F1-score* merupakan rata-rata harmonis antara presisi dan *recall*, seperti dalam (10).

$$F1\text{-}Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (10)$$

III. METODOLOGI PENELITIAN

Penelitian ini menggunakan dataset *employee turnover* yang diperoleh dari *Kaggle* sebanyak 9.540 data periode 2016-2020. Variabel target yang digunakan adalah atribut *left* yang menunjukkan status karyawan keluar atau tetap bekerja. Gambaran skenario penelitian yang akan diterapkan dapat dilihat pada Gbr. 1.

Dataset terdiri atas 10 atribut dengan tipe data campuran antara numerik dan kategorikal. Dataset ini berisi berbagai informasi terkait karyawan, meliputi departemen tempat bekerja, status promosi dalam 24 bulan terakhir, skor evaluasi kinerja terakhir, jumlah proyek yang diikut, tingkat gaji (rendah, sedang, tinggi), masa kerja (*tenure*) dalam tahun, tingkat kepuasan karyawan berdasarkan survei, status penerimaan bonus dalam 24 bulan terakhir, rata-rata jam kerja per bulan, serta status keluar atau tetapnya karyawan di perusahaan. Pada data ini terdapat beberapa atribut yang memiliki nilai data kategorikal, seperti department, salary, dan left.

Agar data kategorial ini dapat diproses secara optimal oleh algoritma *Gaussian Naïve Bayes* dan *K-Nearest Neighbor* (KNN), transformasi data dilakukan dengan menggunakan teknik *one-hot encoding* dan *label encoding*. Teknik *one-hot encoding* diterapkan pada kolom '*department*', dimana setiap kategori dalam kolom tersebut dikonversi menjadi sejumlah fitur biner baru. Selanjutnya, proses *label encoding* dilakukan secara manual pada kolom '*salary*' dan '*left*'. Nilai '*low*', '*medium*', dan '*high*' pada kolom '*salary*' masing-masing diubah menjadi 0, 1, dan 2, sedangkan pada kolom '*left*', nilai '*yes*' diganti dengan 1 dan '*no*' dengan 0.

Pada penelitian ini dilakukan dua percobaan berdasarkan teknik *preprocessing* data yang berbeda. Kedua percobaan menggunakan algoritma *Gaussian Naïve Bayes* dan *K-Nearest Neighbor* (KNN) dengan variasi parameter yang telah ditentukan. Percobaan pertama menggunakan normalisasi dengan *Min-Max Scaler*, sedangkan percobaan kedua menggunakan standarisasi dengan *StandardScaler*. Sebelum proses *scaling*, data melalui tahap data *cleaning* dan *transformation*. Kedua percobaan digunakan untuk membandingkan pengaruh teknik *preprocessing* terhadap performa model klasifikasi.

Data hasil *preprocessing* dari masing-masing percobaan selanjutnya diseimbangkan menggunakan metode SMOTE untuk mengatasi ketidakseimbangan kelas. Setelah itu dilakukan ekstraksi fitur menggunakan *Principal Component Analysis* (PCA) untuk mereduksi dimensi data dengan tetap mempertahankan informasi penting. Data hasil PCA kemudian digunakan pada proses optimasi parameter menggunakan *GridSearchCV* dan pelatihan model.

Pada tahap *GridSearchCV*, diterapkan *K-Fold Cross Validation* dengan nilai k = 3, 5, 7, 9, dan 11 untuk memperoleh model yang stabil. Optimasi dilakukan pada kedua algoritma menggunakan berbagai kombinasi parameter. Rincian parameter dan rentang nilai yang digunakan pada masing-masing algoritma disajikan pada Tabel II.

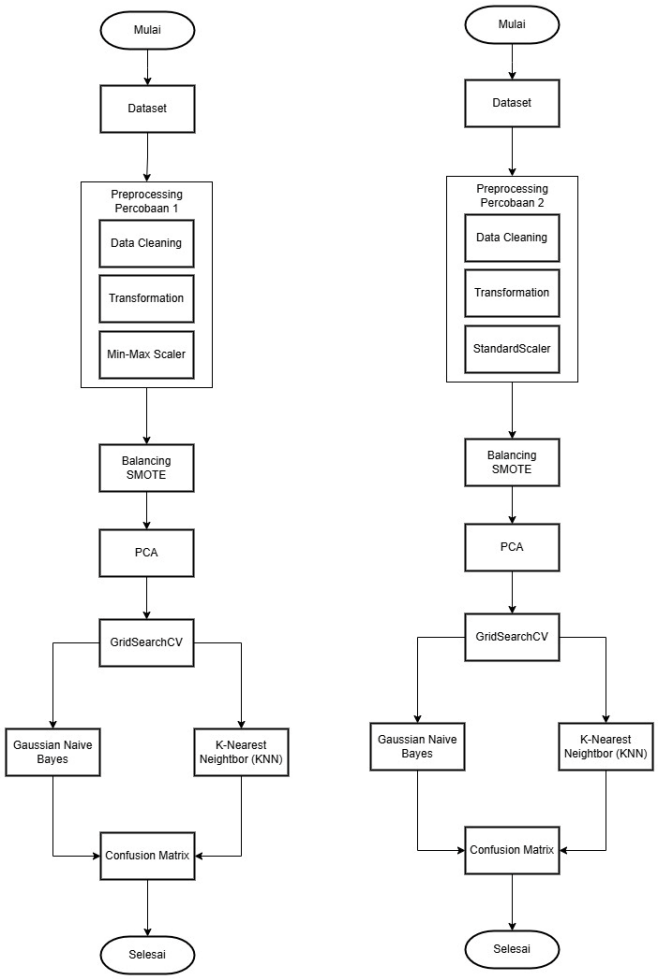
Pada tahap *GridSearchCV*, diterapkan *K-Fold Cross Validation* dengan nilai k = 3, 5, 7, 9, dan 11 untuk memperoleh model yang stabil. Optimasi dilakukan pada kedua algoritma menggunakan berbagai kombinasi parameter. Rincian parameter dan rentang nilai yang digunakan pada masing-masing algoritma disajikan pada Tabel II.

TABEL II
RENTANG NILAI HYPERPARAMETER PADA PROSSES TUNING ALGORITMA

Algoritma	Parameter	Nilai
Gaussian Naïve Bayes	*var_smoothing*	1e-09, 1e-08, 1e-07, 1e-06, 1e-05, 1e-04, 1e-03, 1e-02, 1e-01
K-Nearest Neighbor (KNN)	*n_neighbors*	1, 2, 3, 4, 5, 6, 7, 8, 9, 10
K-Nearest Neighbor (KNN)	*weights*	*distance*
K-Nearest Neighbor (KNN)	*metric*	*Euclidean, Manhattan*

Selanjutnya, model terbaik hasil *GridSearchCV* digunakan untuk melakukan klasifikasi menggunakan *Gaussian Naïve Bayes* dan *K-Nearest Neighbor* (KNN). Hasil prediksi dari kedua metode kemudian dievaluasi menggunakan *confusion matrix* berdasarkan nilai *accuracy*, *precision*, *recall*, dan *F1-score*.

ISSN (Online) 2747-0563
Volume 6 Tahun 2026



Gbr. 1 Gambaran Umum Penelitian.

IV. HASIL DAN ANALISIS

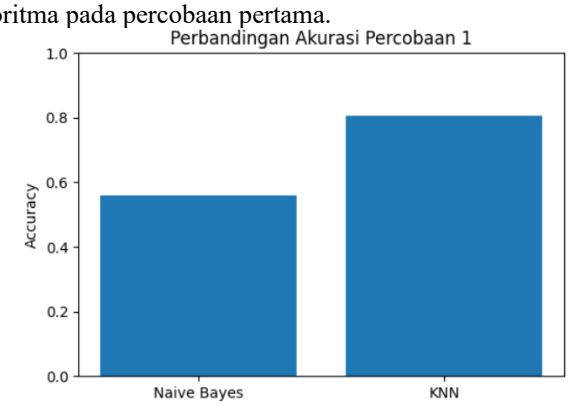
Pada penelitian ini, parameter terbaik dari masing-masing algoritma diperoleh menggunakan *GridSearchCV* pada dua percobaan, yaitu normalisasi dengan *Min-Max Scaler* dan standarisasi dengan *StandardScaler*. Model terpilih kemudian dievaluasi menggunakan data testing untuk menilai kemampuan model dalam mengklasifikasikan data *employee turnover* yang belum pernah dilihat sebelumnya.

Pada percobaan pertama mengggunakan normalisasi *Min-Max Scaler*, hasil pengujian menunjukkan bahwa *Gaussian Naïve Bayes* dengan parameter *var_smoothing* = 1e-09 memperoleh akurasi sebesar 55,86%. Sementara itu, *K-Nearest Neighbor* (KNN) dengan parameter *n_neighbors* = 4, *metric* = *Manhattan*, dan *weights* = *distance* memperoleh akurasi yang lebih tinggi, yaitu sebesar 80,47%. Pada Tabel III adalah hasil pengujian data testing pada percobaan pertama.

TABEL III
HASIL AKURASI DATA TESTING PERCOBAAN 1

Algoritma	Parameter Terbaik	Akurasi
Gaussian Naïve Bayes	*var_smoothing* = 1e-09	55.86%
K-Nearest Neighbor (KNN)	*n_neighbors* = 4 *metric* = *Manhattan*	80.47%

Pada Gbr. 2 merupakan hasil perbandingan performa kedua algoritma pada percobaan pertama.



Gbr. 2 Hasil Perbandingan Akurasi Percobaan 1

Pada percobaan pertama, *K-Nearest Neighbor* (KNN) menunjukkan performa yang lebih baik dibandingkan *Gaussian Naive Bayes*. Hal ini menunjukkan bahwa asumsi independensi pada *Gaussian Naïve Bayes* kurang terpenuhi pada dataset *employee turnover* sehingga berdampak pada menurunnya kemampuan klasifikasi model.

Selanjutnya, pada percobaan kedua menggunakan *StandardSaler*, *Gaussian Naïve Bayes* dengan parameter *var_smoothing* = 1e-09 memperoleh akurasi sebesar 67,18%, sedangkan *K-Nearest Neighbor* (KNN) dengan parameter *n_neighbors* = 4, *metric* = *Euclidean*, dan *weights* = *distance* memperoleh akurasi yang lebih tinggi, yaitu sebesar 85,46%. Pada Tabel IV adalah hasil pengujian data testing pada percobaan kedua.

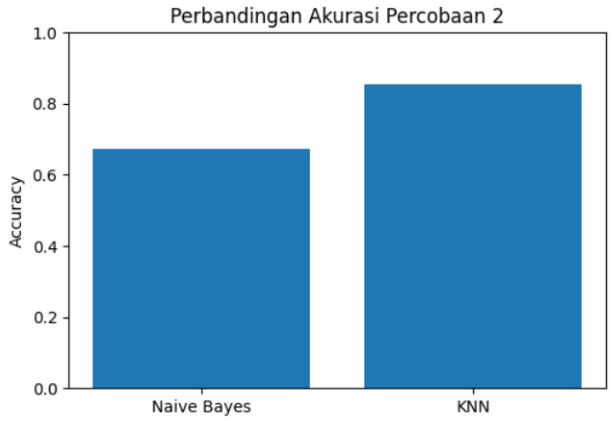
TABEL IV
HASIL AKURASI DATA TESTING PERCOBAAN 2

Algoritma	Parameter Terbaik	Akurasi
Gaussian Naïve Bayes	*var_smoothing* = 1e-09	67.18%
K-Nearest Neighbor (KNN)	*n_neighbors* = 4 *metric* = *Euclidean*	85.46%

Pada Gbr. 3 merupakan hasil perbandingan performa kedua algoritma pada percobaan kedua.

Berdasarkan hasil pengujian pada percobaan kedua, algoritma *K-Nearest Neighbor* (KNN) Kembali menunjukkan performa yang lebih unggul dibandingkan *Gaussian Naïve Bayes*. Peningkatan nilai akurasi pada kedua algoritma, menunjukkan bahwa proses standarisasi menggunakan *StandardScaler* memberikan pengaruh yang lebih optimal dibandingkan dengan normalisasi menggunakan *Min-Max Scaler* terhadap proses klasifikasi.

ISSN (Online) 2747-0563
Volume 6 Tahun 2026



Gbr. 3 Hasil Perbandingan Akurasi Percobaan 1

V. KESIMPULAN

Berdasarkan hasil penelitian dan analisis yang dilakukan dalam klasifikasi dataset *employee turnover*, algoritma *K-Nearest Neighbor* (KNN) dengan *StandardScaler* merupakan kombinasi terbaik dan menunjukkan performa yang jauh lebih unggul dibandingkan *Gaussian Naïve Bayes* dalam klasifikasi dataset *employee turnover* sebanyak 9.540 data. Performa terbaik diperoleh pada kombinasi *K-Nearest Neighbor* (KNN) dengan *StandardScaler*, yang mencapai akurasi sebesar 85,46%, *precision* 0,81%, *recall* 0,92%, dan *F1-score* 0,86%. Nilai tersebut lebih tinggi dibandingkan *Gaussian Naïve Bayes* yang hanya memperoleh akurasi sebesar 67,18%, *precision* 0,63%, *recall* 0,81%, dan *F1-score* 0,71%, sehingga *K-Nearest Neighbor* (KNN) dengan *StandardScaler* terbukti sebagai metode yang paling efektif untuk klasifikasi *employee turnover* pada penelitian ini. Hasil optimal tersebut diperoleh melalui proses *tuning* parameter dan *k-fold cross validation*.

Selain itu, teknik *preprocessing* memberikan pengaruh yang signifikan terhadap performa model. Standarisasi menggunakan *StandardScaler* secara konsisten menghasilkan performa yang lebih baik dibandingkan normalisasi menggunakan *Min-Max Scaler* pada kedua algoritma, sehingga terbukti lebih efektif dalam meningkatkan kinerja klasifikasi.

REFERENSI

[1] M. Miftahurrohman and M. Munifah, "Dampak Turnover Pegawai Terhadap Kinerja Organisasi Perguruan Tinggi di Jawa Tengah," *Jurnal Bisnis Kreatif dan Inovatif*, vol. 1, no. 2, pp. 100–110, 2024.

[2] Y. N. Rumbia et al., "Perbandingan Metode KNN dan Naive Bayes untuk Klasifikasi Kelulusan Mahasiswa Pada Mata Kuliah Probstat," *JURNAL PTI (PENDIDIKAN DAN TEKNOLOGI INFORMASI) FAKULTAS KEGURUAN ILMU PENDIDIKAN UNIVERSITA PUTRA INDONESIA" YPTK" PADANG*, pp. 7–13, 2025.

[3] H. Cahyaningrum, D. Arifianto, and G. Abdurrahman, "Analisis Perbandingan Metode K Nearest Neighbor Dan Gaussian Naive Bayes Pada Klasifikasi Jurusan Siswa (Studi Kasus pada Siswa SMA Muhammadiyah 3 Jember)," *Jurnal Smart Teknologi*, vol. 3, no. 3, pp. 261–266, 2022.

[4] A. R. Pratama, F. Wabula, H. Ilmandry, M. L. Isabela, M. Raharjo, and R. Sianipar, "Literature Review The Impact of Machine Learning in Modern Industries," Nian Tana Sikka: *Jurnal ilmiah Mahasiswa*, vol. 3, no. 1, pp. 177–182, 2025.

[5] R. J. Woodman and A. A. Mangoni, "A comprehensive review of machine learning algorithms and their application in geriatric medicine: present and future," *Aging clinical and experimental research*, vol. 35, no. 11, pp. 2363–2397, 2023.

[6] K. Isyara and A. Kurniasih, "Penggunaan Metode SMOTE pada NaÃ¯ ve Bayes Gaussian untuk Klasifikasi Mahasiswa Drop Out," in *Prosiding Seminar Nasional Mahasiswa Bidang Ilmu Komputer dan Aplikasinya*, 2023, pp. 616–623.

[7] A. Lubis, Y. Irawan, J. Junadhi, and S. Defit, "Leveraging K-Nearest Neighbors with SMOTE and boosting techniques for data imbalance and accuracy improvement," *Journal of Applied Data Sciences*, vol. 5, no. 4, pp. 1625–1638, 2024.

[8] M. Kumar and V. Bhardwaj, "Evaluating label encoding and preprocessing techniques for breast cancer prediction using machine learning algorithms," *International Journal of Computational Intelligence Systems*, vol. 18, no. 1, p. 218, 2025.

[9] N. H. Fadhil, A. M. Mosa, and I. S. Abdulsaheed, "Principal Component Analysis for Feature Extraction," *Vital Annex: International Journal of Novel Research in Advanced Sciences (2751-756X)*, vol. 3, no. 3, pp. 100–103, 2024.

[10] E. Elhaik, "Principal Component Analyses (PCA)-based findings in population genetic studies are highly biased and must be reevaluated," *Scientific reports*, vol. 12, no. 1, p. 14683, 2022.

[11] R. Arifudin, S. Subhan, and Y. N. Ifriza, "Student Adaptability Level Optimization using GridsearchCV with Gaussian Naive Bayes and K-Nearest Neighbor Methods as an Effort to Improve Online Education Predictions," *Jurnal Nasional Pendidikan Teknik Informatika: JANAPATI*, vol. 14, no. 2, 2025.

[12] A. D. Putri, K. I. Nahampun, P. Ritonga, and I. Andre, "PENERAPAN METODE NAIVE BAYES UNTUK MEMPREDIKSI HASIL BELAJAR PESERTA DIDIK 'STUDI KASUS SD SWASTA METHODIST-1 RANTAUPRAPAT,'" *Journal Computer Science and Information Technology (JCoInT)*, vol. 5, no. 2, pp. 238–248, 2024.

[13] I. As' ad, "Advancing healthcare diagnostics: a study on gaussian naive bayes classification of blood samples," *International Journal of Artificial Intelligence in Medical Issues*, vol. 1, no. 2, pp. 115–123, 2023.

[14] M. R. Saputra and P. Parjito, "Analisis sentimen twitter terhadap konflik di papua menggunakan perbandingan naive bayes dan svm," *JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)*, vol. 10, no. 2, pp. 1197–1208, 2025.

[15] N. U. Nia, L. I. Lemi, and D. F. B. Dwi, "Implementation of Machine Learning Based on the K-Nearest Neighbor (KNN) Algorithm in Elderly Activities in Nursing Homes," *JUPITER: Jurnal Penelitian Ilmu dan Teknologi Komputer*, vol. 17, no. 2, pp. 811–820, 2025.

[16] J. C. Obi, "A comparative study of several classification metrics and their performances on data," *World Journal of Advanced Engineering Technology and Sciences*, vol. 8, no. 1, pp. 308–314, 2023.