JURNAL REKAYASA INFORMATIKA (JRIT),
Vol. 3, No.1 2026, Hal.55-64

ISSN : 3089-5375

Penerapan Data Mining untuk Prediksi Employee Attrition Menggunakan Naive Bayes dan Decision Tree

Rafly Anugrah[1], Heti Mulyani[2]
[1,2]Teknologi Rekayasa Perangkat Lunak, Politeknik Enjinering Indorama, Indonesia
Email Korespondensi: raflyyanugrahh39@gmail.com

Info Artikel

Histori Artikel:

Dikirim 20-05-2026
Revisi 29-05-2026
Diterima 01-06-2026

Kata Kunci:
Data Mining
Employee Attrition
Decision Tree
Naive Bayes
Klasifikasi

ABSTRAK

Penelitian ini membahas penerapan teknik *data mining* untuk memprediksi *employee attrition* menggunakan dataset IBM HR Employee Attrition. Tingginya tingkat attrition karyawan dapat memberikan dampak negatif bagi perusahaan, seperti meningkatnya biaya rekrutmen dan menurunnya produktivitas kerja. Oleh karena itu, diperlukan sistem prediksi yang mampu membantu perusahaan dalam mengidentifikasi kemungkinan karyawan melakukan resign. Penelitian ini menggunakan dua algoritma klasifikasi, yaitu Decision Tree dan Naive Bayes, untuk membandingkan performa model dalam memprediksi attrition karyawan. Tahapan penelitian meliputi *preprocessing data*, pengecekan *duplicate data*, deteksi *outlier*, visualisasi data menggunakan *heatmap* dan *boxplot*, serta proses pelatihan dan pengujian model. Evaluasi model dilakukan menggunakan *accuracy score* dan *confusion matrix*. Hasil penelitian menunjukkan bahwa algoritma Naive Bayes memiliki performa yang lebih baik dibandingkan Decision Tree dengan tingkat akurasi sebesar 79%, sedangkan Decision Tree menghasilkan akurasi sebesar 73%. Model terbaik kemudian diimplementasikan ke dalam aplikasi berbasis Streamlit untuk mempermudah proses prediksi attrition karyawan. Penelitian ini diharapkan dapat membantu perusahaan dalam pengambilan keputusan terkait pengelolaan sumber daya manusia secara lebih efektif.

Article Info

Article history:
Received 20-05-2026
Revised 29-05-2026
Accepted 01-06-2026

Keywords:
Data Mining
Employee Attrition
Decision Tree
Naive Bayes
Classification

ABSTRACT

This study discusses the application of data mining techniques to predict employee attrition using the IBM HR Employee Attrition dataset. High employee attrition rates can negatively impact companies, including increased recruitment costs and decreased work productivity. Therefore, a prediction system is needed to help companies identify employees who are likely to resign. This research uses two classification algorithms, namely Decision Tree and Naive Bayes, to compare model performance in predicting employee attrition. The research stages include data preprocessing, duplicate data checking, outlier detection, data visualization using heatmaps and boxplots, as well as model training and testing processes. Model evaluation was carried out using accuracy scores and confusion matrices. The results showed that the Naive Bayes algorithm performed better than Decision Tree with an accuracy of 79%, while Decision Tree achieved an accuracy of 73%. The best model was then implemented into a Streamlit-based application to simplify the employee attrition prediction process. This research is expected to assist companies in making more effective human resource management decisions.

1. PENDAHULUAN

Perkembangan teknologi informasi dan digitalisasi di dunia industri menyebabkan perusahaan menghasilkan data dalam jumlah besar yang dapat dimanfaatkan untuk mendukung pengambilan keputusan. Salah satu permasalahan yang sering dihadapi perusahaan adalah tingginya tingkat *employee attrition* atau pengunduran diri karyawan. Tingginya tingkat attrition dapat memberikan dampak negatif bagi perusahaan, seperti meningkatnya biaya rekrutmen, pelatihan karyawan baru, serta menurunnya produktivitas perusahaan [1]. Oleh karena itu, diperlukan suatu metode yang mampu membantu perusahaan dalam memprediksi kemungkinan karyawan melakukan resign sehingga perusahaan dapat mengambil tindakan pencegahan lebih awal. Salah satu pendekatan yang dapat digunakan adalah penerapan teknik *data mining* untuk melakukan analisis dan prediksi berdasarkan data historis karyawan.

Data mining merupakan proses pengolahan data untuk menemukan pola, hubungan, maupun informasi penting dari sekumpulan data berukuran besar [2]. Dalam penelitian ini, teknik *classification* digunakan untuk memprediksi status attrition karyawan berdasarkan atribut tertentu seperti usia, pendapatan bulanan, tingkat jabatan, lembur (*overtime*), dan tingkat kepuasan kerja. Metode klasifikasi dipilih karena mampu mengelompokkan data ke dalam kategori tertentu berdasarkan pola yang dipelajari dari data sebelumnya [3]. Dataset yang digunakan pada penelitian ini adalah IBM HR Employee Attrition Dataset yang memiliki berbagai atribut terkait kondisi dan karakteristik karyawan.

Beberapa penelitian sebelumnya telah menerapkan algoritma klasifikasi untuk memprediksi *employee attrition*. Penelitian menggunakan algoritma Decision Tree menunjukkan bahwa metode tersebut mampu menghasilkan model klasifikasi yang mudah dipahami karena berbentuk pohon keputusan [4]. Selain itu, algoritma Naive Bayes juga banyak digunakan karena memiliki proses komputasi yang cepat dan mampu memberikan performa klasifikasi yang baik pada data tertentu [5]. Namun, setiap algoritma memiliki tingkat performa yang berbeda tergantung pada karakteristik dataset yang digunakan. Oleh karena itu, diperlukan analisis perbandingan performa antar algoritma untuk mengetahui metode yang paling sesuai dalam memprediksi attrition karyawan.

Berdasarkan penelitian sebelumnya, masih terdapat kebutuhan untuk melakukan evaluasi dan perbandingan performa algoritma klasifikasi pada dataset *employee attrition*. Selain itu, penelitian ini juga menerapkan tahapan *preprocessing* seperti pengecekan *duplicate data*, deteksi *outlier*, dan visualisasi data menggunakan *heatmap* dan *boxplot* untuk meningkatkan kualitas data sebelum proses pemodelan dilakukan. Kebaruan penelitian ini terletak pada analisis perbandingan algoritma Decision Tree dan Naive Bayes dalam memprediksi *employee attrition* menggunakan dataset IBM HR yang divisualisasikan melalui implementasi berbasis web menggunakan Streamlit.

Tujuan dari penelitian ini adalah menerapkan teknik *data mining* untuk memprediksi *employee attrition* serta membandingkan performa algoritma Decision Tree dan Naive Bayes berdasarkan nilai akurasi yang dihasilkan. Hasil dari penelitian ini diharapkan dapat membantu perusahaan dalam memahami faktor-faktor yang memengaruhi attrition karyawan dan mendukung pengambilan keputusan yang lebih efektif dalam pengelolaan sumber daya manusia.

2. METODE

Metode penelitian yang digunakan pada penelitian ini adalah metode *Knowledge Discovery in Database* (KDD) dengan pendekatan klasifikasi menggunakan algoritma Decision Tree dan Naive Bayes. Penelitian dilakukan untuk memprediksi *employee attrition* berdasarkan data karyawan pada dataset IBM HR Employee Attrition. Tahapan penelitian dimulai dari pengumpulan data, *preprocessing*, visualisasi data, pembuatan model klasifikasi, evaluasi model, hingga implementasi sistem berbasis web menggunakan Streamlit.

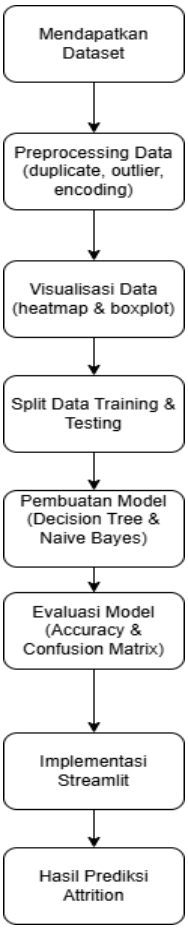
Dataset yang digunakan berasal dari IBM HR Employee Attrition Dataset yang diperoleh melalui platform Kaggle. Dataset tersebut memiliki berbagai atribut yang berkaitan dengan kondisi karyawan, seperti usia, pendapatan bulanan, tingkat jabatan, lembur (*overtime*), kepuasan kerja, dan lama bekerja di perusahaan. Pada penelitian ini, atribut target yang digunakan adalah *Attrition* yang terdiri dari dua

kelas, yaitu *Yes* dan *No*. Sebelum proses pemodelan dilakukan, data terlebih dahulu melalui tahap *preprocessing* untuk meningkatkan kualitas data sehingga hasil prediksi menjadi lebih optimal [6].

Tahapan *preprocessing* meliputi pengecekan *duplicate data*, deteksi *outlier* menggunakan *boxplot*, serta analisis korelasi atribut menggunakan *heatmap*. Selanjutnya dilakukan proses transformasi data kategorikal menjadi numerik menggunakan teknik *encoding* agar dapat diproses oleh algoritma machine learning. Setelah data bersih, dataset dibagi menjadi data *training* dan data *testing* menggunakan metode *train-test split*. Data *training* digunakan untuk melatih model, sedangkan data *testing* digunakan untuk mengukur performa model klasifikasi [7].

Pada tahap pemodelan, penelitian menggunakan dua algoritma klasifikasi, yaitu Decision Tree dan Naive Bayes. Algoritma Decision Tree bekerja dengan membentuk struktur pohon keputusan berdasarkan atribut yang paling berpengaruh terhadap hasil klasifikasi. Sementara itu, Naive Bayes menggunakan pendekatan probabilitas berdasarkan Teorema Bayes untuk menentukan kemungkinan suatu data termasuk ke dalam kelas tertentu [8]. Kedua model kemudian dibandingkan berdasarkan nilai akurasi dan confusion matrix untuk mengetahui algoritma yang memiliki performa terbaik dalam memprediksi *employee attrition*.

Hasil penelitian kemudian diimplementasikan ke dalam aplikasi berbasis web menggunakan Streamlit. Sistem memungkinkan pengguna memasukkan data karyawan dan memperoleh hasil prediksi apakah karyawan berpotensi resign atau tidak resign. Alur penelitian pada penelitian ini dapat dilihat pada Gambar 1.



Gambar 1 Alur Metode Penelitian

Pengujian sistem dilakukan menggunakan confusion matrix dan accuracy score untuk mengetahui performa model klasifikasi. Nilai accuracy digunakan untuk mengukur tingkat ketepatan model dalam

 ISSN : 3089-5375

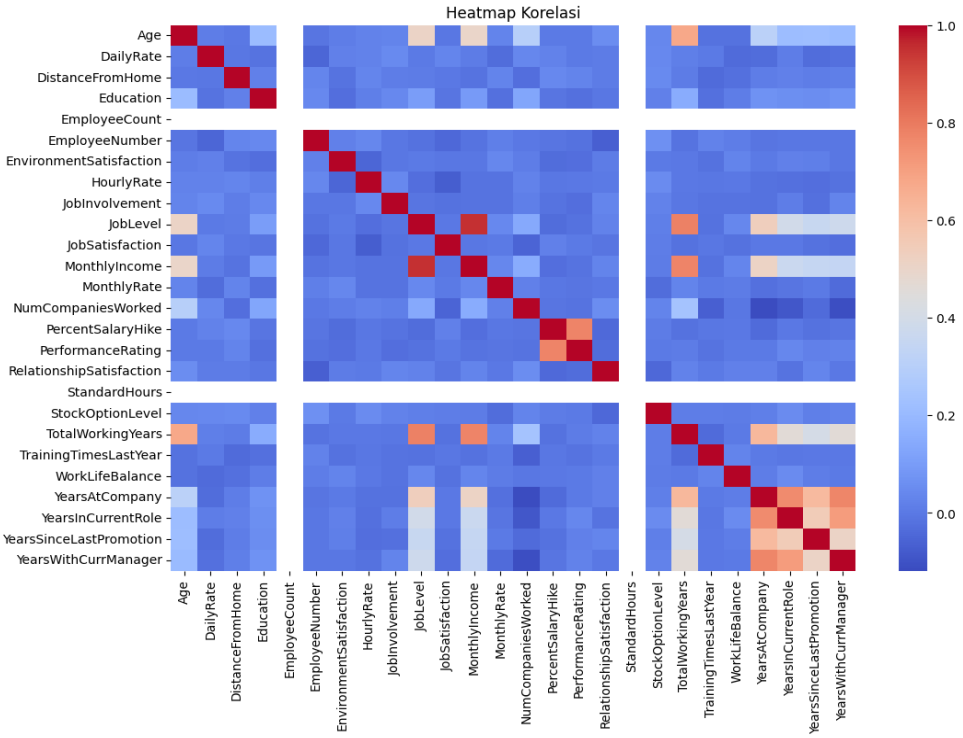
melakukan prediksi, sedangkan confusion matrix digunakan untuk melihat detail prediksi benar dan salah pada masing-masing kelas [9]. Berdasarkan hasil pengujian, model Naive Bayes menghasilkan tingkat akurasi yang lebih tinggi dibandingkan Decision Tree sehingga dipilih sebagai model utama pada implementasi sistem prediksi employee attrition.

3. HASIL DAN PEMBAHASAN

Pada penelitian ini dilakukan proses klasifikasi untuk memprediksi *employee attrition* menggunakan algoritma Decision Tree dan Naive Bayes. Dataset yang digunakan adalah IBM HR Employee Attrition Dataset yang terdiri dari berbagai atribut karyawan seperti usia, pendapatan, tingkat jabatan, kepuasan kerja, dan lembur (*overtime*). Tahapan penelitian dimulai dari proses *preprocessing*, visualisasi data, pembuatan model klasifikasi, hingga evaluasi performa model.

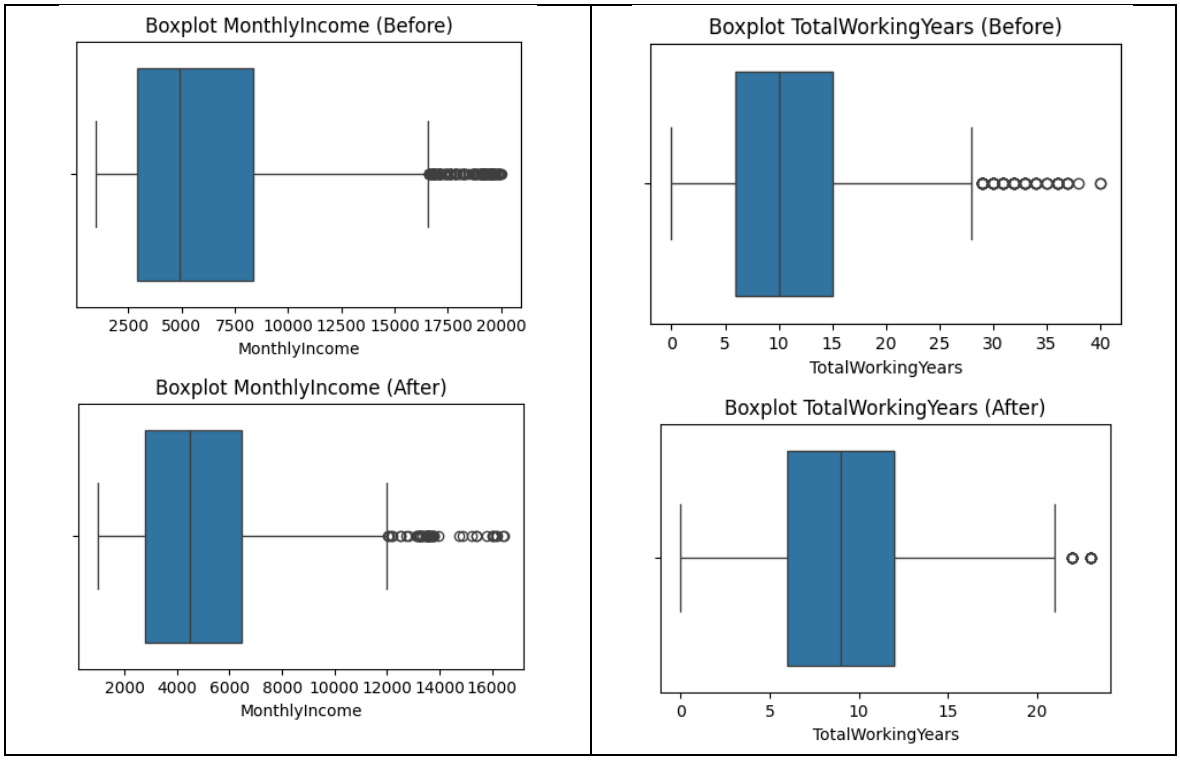
3.1. Hasil Preprocessing dan Visualisasi Data

Tahap awal penelitian dilakukan dengan proses *preprocessing* untuk meningkatkan kualitas data sebelum digunakan pada tahap pemodelan. Proses ini meliputi pengecekan *duplicate data*, deteksi *outlier*, dan transformasi data kategorikal menjadi numerik menggunakan teknik *encoding*. Setelah data dibersihkan, dilakukan visualisasi data menggunakan *heatmap* dan *boxplot*.



Gambar 2 Heatmap Korelasi Antar Atribut

 ISSN : 3089-5375



Gambar 3 Visualisasi Outlier Menggunakan Boxplot

Visualisasi *heatmap* digunakan untuk melihat hubungan korelasi antar atribut pada dataset. Berdasarkan hasil visualisasi, beberapa atribut seperti *MonthlyIncome*, *JobLevel*, dan *TotalWorkingYears* memiliki hubungan terhadap atribut *Attrition*. Sementara itu, visualisasi *boxplot* digunakan untuk mendeteksi keberadaan *outlier* pada atribut numerik. Hasil visualisasi menunjukkan bahwa terdapat beberapa *outlier* pada atribut seperti *MonthlyIncome* dan *TotalWorkingYears* sehingga dilakukan proses penghapusan *outlier* untuk meningkatkan performa model klasifikasi [10].

Selain itu, proses *encoding* dilakukan untuk mengubah data kategorikal seperti *Yes* dan *No* menjadi bentuk numerik agar dapat diproses oleh algoritma machine learning. Dataset kemudian dibagi menjadi data *training* dan *testing* menggunakan metode *train-test split* dengan proporsi data tertentu untuk proses pelatihan dan pengujian model.

3.2. Hasil Implementasi Model Klasifikasi

Pada penelitian ini digunakan dua algoritma klasifikasi, yaitu Decision Tree dan Naive Bayes. Kedua algoritma dilatih menggunakan data *training* dan dievaluasi menggunakan data *testing*. Evaluasi dilakukan menggunakan *accuracy score* dan *confusion matrix* untuk mengetahui performa model dalam memprediksi *employee attrition*.

3.2.1. Hasil Pengujian Decision Tree

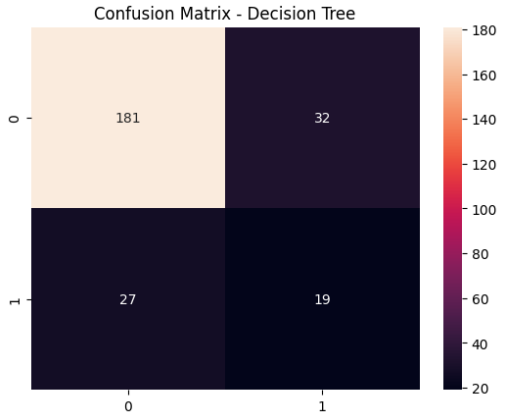
Algoritma Decision Tree bekerja dengan membentuk struktur pohon keputusan berdasarkan atribut yang paling berpengaruh terhadap hasil klasifikasi. Pada penelitian ini, model Decision Tree menghasilkan nilai akurasi sebesar 73%. Nilai akurasi diperoleh menggunakan persamaan 1 berikut:

$$\text{Accuracy} = \frac{Jumlah\ prediksi\ benar}{Total\ Data} \text{ x } 100\% \quad (1)$$

```
from sklearn.tree import DecisionTreeClassifier
from sklearn.metrics import accuracy_score

dt = DecisionTreeClassifier()
dt.fit(X_train, y_train)

y_pred_dt = dt.predict(X_test)

acc_dt = accuracy_score(y_test, y_pred_dt)
print("Accuracy Decision Tree:", acc_dt)
```

```
Accuracy Decision Tree: 0.7799227799227799
```

Gambar 4 Hasil Nilai Akurasi Decision Tree



Gambar 5 Confusion Matrix Decision Tree

Berdasarkan hasil gambar 5 pengujian menggunakan *confusion matrix*, model mampu memprediksi sebagian besar data *non-attrition* dengan baik. Namun, model masih mengalami beberapa kesalahan prediksi pada data *attrition*. Hal ini menunjukkan bahwa Decision Tree cukup baik dalam mengenali pola umum pada dataset, tetapi masih memiliki keterbatasan dalam memprediksi data minoritas.

	Feature	Importance
1	MonthlyIncome	0.295715
0	Age	0.173223
7	YearsAtCompany	0.150899
2	TotalWorkingYears	0.131348
5	JobSatisfaction	0.091036
4	OverTime	0.076995
6	EnvironmentSatisfaction	0.065200
3	JobLevel	0.015584

Gambar 6 Analisis Feature Importance

Selain itu, dilakukan analisis *feature importance* untuk mengetahui atribut yang paling berpengaruh terhadap prediksi *employee attrition*. Hasil analisis menunjukkan bahwa atribut seperti *OverTime*, *MonthlyIncome*, dan *JobLevel* memiliki pengaruh cukup besar terhadap hasil klasifikasi.

3.2.2. Hasil Pengujian Naive Bayes

Algoritma Naive Bayes menggunakan pendekatan probabilitas berdasarkan Teorema Bayes untuk menentukan kemungkinan suatu data termasuk ke dalam kelas tertentu. Persamaan dasar Naive Bayes dapat dituliskan persamaan 2 sebagai berikut:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \quad (2)$$

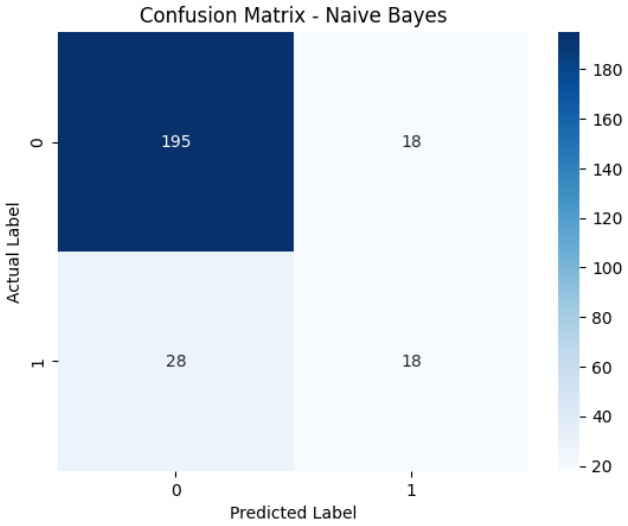
```
from sklearn.naive_bayes import GaussianNB

nb = GaussianNB()
nb.fit(X_train, y_train)

y_pred_nb = nb.predict(X_test)

acc_nb = accuracy_score(y_test, y_pred_nb)
print("Accuracy Naive Bayes:", acc_nb)

Accuracy Naive Bayes: 0.8223938223938224
```

Gambar 7 Hasil Nilai Akurasi Naive Bayes

Keterangan:

- P(A|B)P(A|B)P(A|B) = probabilitas data termasuk ke kelas tertentu
- P(B|A)P(B|A)P(B|A) = probabilitas atribut terhadap kelas
- P(A)P(A)P(A) = probabilitas awal kelas
- P(B)P(B)P(B) = probabilitas total atribut

Berdasarkan hasil pengujian, algoritma Naive Bayes menghasilkan nilai akurasi sebesar 79%, lebih tinggi dibandingkan Decision Tree. Hasil tersebut menunjukkan bahwa Naive Bayes memiliki kemampuan yang lebih baik dalam memprediksi *employee attrition* pada dataset IBM HR Employee Attrition.



Gambar 8 Confusion Matrix Naive Bayes

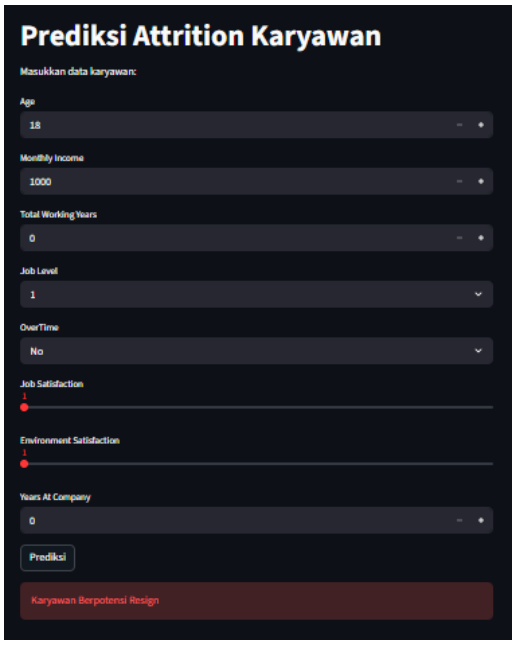
Hasil evaluasi menggunakan *confusion matrix* gambar 8 menunjukkan bahwa model Naive Bayes mampu menghasilkan jumlah prediksi benar yang lebih tinggi dibandingkan Decision Tree. Oleh karena itu, model Naive Bayes dipilih sebagai model utama pada implementasi sistem prediksi berbasis Streamlit.

3.3. Implementasi Sistem Streamlit

Hasil penelitian kemudian diimplementasikan ke dalam aplikasi berbasis web menggunakan Streamlit. Sistem memungkinkan pengguna memasukkan beberapa atribut karyawan seperti usia, pendapatan bulanan, tingkat jabatan, dan lama bekerja untuk memperoleh hasil prediksi attrition secara otomatis.

 ISSN : 3089-5375

Pada implementasi sistem, model Naive Bayes digunakan sebagai model utama karena memiliki tingkat akurasi terbaik. Sistem akan memproses input pengguna dan menghasilkan output berupa prediksi apakah karyawan berpotensi resign atau tidak resign. Implementasi ini diharapkan dapat membantu perusahaan dalam melakukan analisis attrition secara lebih cepat dan efektif.



Gambar 9 Implementasi Sistem

Secara keseluruhan, hasil penelitian menunjukkan bahwa teknik *data mining* menggunakan algoritma klasifikasi mampu membantu proses prediksi *employee attrition*. Berdasarkan hasil evaluasi model, algoritma Naive Bayes memberikan performa yang lebih baik dibandingkan Decision Tree pada dataset IBM HR Employee Attrition.

4. KESIMPULAN

Berdasarkan hasil penelitian yang telah dilakukan, teknik *data mining* dapat diterapkan untuk memprediksi *employee attrition* menggunakan dataset IBM HR Employee Attrition. Penelitian dilakukan melalui beberapa tahapan, yaitu *preprocessing data*, visualisasi data menggunakan *heatmap* dan *boxplot*, pembuatan model klasifikasi menggunakan algoritma Decision Tree dan Naive Bayes, serta evaluasi performa model menggunakan *accuracy score* dan *confusion matrix*. Hasil penelitian menunjukkan bahwa algoritma Naive Bayes memiliki performa yang lebih baik dibandingkan Decision Tree dalam memprediksi attrition karyawan.

Model Decision Tree menghasilkan tingkat akurasi sebesar 73%, sedangkan model Naive Bayes menghasilkan tingkat akurasi sebesar 79%. Berdasarkan hasil tersebut, algoritma Naive Bayes dipilih sebagai model utama dalam implementasi sistem prediksi berbasis Streamlit. Sistem yang dibuat mampu membantu proses analisis data karyawan dan memberikan prediksi apakah seorang karyawan berpotensi resign atau tidak resign berdasarkan atribut tertentu.

Selain itu, penelitian ini menunjukkan bahwa atribut seperti *OverTime*, *MonthlyIncome*, dan *JobLevel* memiliki pengaruh terhadap prediksi *employee attrition*. Dengan adanya sistem prediksi ini, perusahaan diharapkan dapat melakukan tindakan pencegahan lebih awal untuk mengurangi tingkat resign karyawan.

Untuk pengembangan penelitian selanjutnya, metode klasifikasi lain seperti Random Forest, Logistic Regression, atau Support Vector Machine dapat digunakan untuk membandingkan performa

model yang lebih optimal. Selain itu, penelitian selanjutnya juga dapat menggunakan dataset yang lebih besar dan lebih kompleks agar hasil prediksi menjadi lebih akurat dan relevan dengan kondisi nyata di perusahaan.

UCAPAN TERIMA KASIH

Penelitian ini disusun sebagai bagian dari tugas Ujian Tengah Semester (UTS) pada mata kuliah Data Mining. Penulis mengucapkan terima kasih kepada dosen pengampu mata kuliah yang telah memberikan arahan, bimbingan, dan dukungan selama proses penelitian berlangsung. Penulis juga mengucapkan terima kasih kepada berbagai pihak yang telah membantu dalam proses pengumpulan data, pengolahan data, hingga penyusunan jurnal penelitian ini.

Selain itu, penulis juga berterima kasih kepada penyedia dataset IBM HR Employee Attrition yang digunakan dalam penelitian ini sehingga penelitian dapat dilakukan dengan baik. Dukungan dan masukan dari berbagai pihak sangat membantu dalam penyelesaian penelitian ini.

DAFTAR PUSTAKA

[1] G. Negi and V. Nagar, "Abhinav EMPLOYEE ATTRITION : INEVITABLE YET Abhinav," vol. II, no. 1977, pp. 50–59, 2007.
[2] S. Al Faridzi, F. S. Azizah, F. Mustafa, and A. N. Putri, "PENGOLAHAN DATA : Pemahaman Gempa Bumi Di Indonesia Melalui Pendekatan Data Mining," vol. 2, no. 1, pp. 262–270, 2024.
[3] T. N. Phyu, "Survey of Classification Techniques in Data Mining," vol. I, 2009.
[4] J. R. Quinlan, "Induction of Decision Trees," pp. 81–106, 2007.
[5] H. Zhang, "The Optimality of Naive Bayes," 2004.
[6] S. A. A. and W. S. Bhaya, "Review of Data Preprocessing Techniques in Data Mining," *J. Eng. Appl. Sci.*, 2017.
[7] D. H. Kamagi and S. Hansun, "Implementasi Data Mining dengan Algoritma C4 . 5 untuk Memprediksi Tingkat Kelulusan Mahasiswa," vol. VI, no. 1, pp. 15–20, 2014.
[8] H. D. Wijaya and S. Dwiasnati, "Implementasi Data Mining dengan Algoritma Naïve Bayes pada Penjualan Obat," vol. 7, no. 1, pp. 1–7, 2020.
[9] V. M. Patro and M. R. Patra, "Augmenting Weighted Average with Confusion Matrix to Enhance Classification Accuracy," no. August 2014.
[10] G. R. Pandiangan, A. Barus, M. R. Albani, and N. Sipahutar, "Optimasi Naive Bayes dengan Diskritisasi dan Penanganan Outlier untuk Deteksi Diabetes pada Dataset Pima Indians," no. April, pp. 25–37, 2026.