Analisis dan Pemodelan Prediksi Pergantian Karyawan Menggunakan Model Machine Learning Klasifikasi

Muhammad Rifqi Ma'ruf[1], Adyuta Prajahita Murdianto[2], Delai Resgista Setyawan [3]

I. PENDAHULUAN

1.1 Latar Belakang

Pergantian karyawan adalah salah satu tantangan utama yang dihadapi oleh departemen sumber daya manusia di berbagai organisasi atau perusahaan. Pergantian karyawan yang tinggi tidak hanya berdampak pada biaya rekrutmen dan pelatihan yang meningkat, tetapi juga dapat mengganggu produktivitas, moral tim, dan stabilitas organisasi secara keseluruhan. Oleh karena itu, memahami faktor-faktor yang berkontribusi terhadap *attrition employee* atau pergantian karyawan dan mengembangkan model prediktif untuk mengidentifikasi karyawan yang berisiko tinggi untuk meninggalkan perusahaan adalah langkah penting dalam strategi manajemen sumber daya manusia.

Dalam upaya ini, Synthetic Employee Attrition Dataset telah dikembangkan sebagai dataset simulasi yang dirancang untuk analisis dan prediksi *attrition* karyawan. Dataset ini mencakup informasi terperinci tentang berbagai aspek profil karyawan, termasuk demografi, fitur terkait pekerjaan, dan kondisi pribadi. Dengan total 64,498 sampel yang dibagi menjadi set pelatihan dan pengujian. Dataset ini memberikan basis yang kuat untuk pengembangan dan evaluasi model machine learning yang bertujuan untuk memprediksi attrition.

Setiap baris dalam dataset mencakup ID Karyawan yang unik serta fitur-fitur yang mempengaruhi attrition karyawan, Tujuannya adalah untuk memahami faktor-faktor yang berkontribusi terhadap attrition dan mengambangkan model prediktif untuk mengidentifikasi karyawan yang berisiko. Dataset ini sangat ideal untuk analitik HR, pengembangan model machine learning, dan demonstrasi teknik analisis data tingkat lanjut. Dengan memberikan pandangan yang komprehensif dan realistis tentang faktor-faktor yang mempengaruhi retensi karyawan, dataset ini menjadi sumber daya yang berharga bagi peneliti dan praktisi di bidang sumber daya manusia dan pengembangannya.

Penggunaan dataset ini memungkinkan perusahaan untuk mengidentifikasi pola dan tren yang mungkin tidak terlihat melalui analisis tradisional. Dengan menggunakan metode machine learning, perusahaan dapat membuat prediksi yang lebih akurat tentang karyawan mana yang berisiko tinggi untuk meninggalkan perusahaan dan mengambil tindakan proaktif untuk mempertahankan mereka yang berharga. Hal ini tidak hanya dapat mengurangi biaya yang terkait dengan pergantian karyawan, tetapi juga dapat membantu menciptakan lingkungan kerja yang lebih stabil dan produktif.

1.2 Rumusan Masalah

1. Apa model yang paling optimal antara random forest, naive bayes, svm, knn, decision tree, svc, logistic regression, XGBoost, dan AdaBoost dalam klasifikasi attrition karyawan
2. Bagaimana penggunaan normalisasi dan hyperparameter tuning mampu meningkatkan hasil klasifikasi

1.2 Batasan Masalah

1. Projek akhir ini menggunakan dataset sintetik yang tidak sepenuhnya mencerminkan kompleksitas dan variasi kondisi nyata di berbagai tempat. Sehingga hasil dari penelitian tidak dapat digeneralisasi ke semua sektor industri
2. Data yang digunakan mencerminkan kondisi pada saat dataset tersebut dibuat. Sehingga tidak mencerminkan perubahan tren atau kondisi pasar tenaga kerja yang terjadi pada saat setelah dataset dibuat.

1.3 Tujuan Penelitian

1. Mengidentifikasi model terbaik dalam klasifikasi attrition karyawan menggunakan model random forest, naive bayes, svm, knn, decision tree, logistic regression, XGBoost, dan AdaBoost.
3. Mengetahui pemodelan machine learning yang paling akurat untuk memprediksi attrition karyawan.

1.4 Manfaat Penelitian

1. Sebagai mahasiswa, penelitian ini membantu pemahaman penggunaan machine learning dalam tugas klasifikasi untuk penerapan di dunia nyata.
2. Sebagai pemangku manajemen perusahaan, penelitian ini membantu perusahaan memahami faktor-faktor utama yang mempengaruhi attrition karyawan dan menggunakan model prediktif untuk mengoptimalkan strategi retensi karyawan.
3. Sebagai peneliti, penelitian ini menjadi dasar bagi penelitian lebih lanjut di bidang attrition karyawan, baik dalam konteks metodologi machine learning maupun eksplorasi faktor-faktor tambahan yang mempengaruhi retensi karyawan.

II. TINJAUAN PUSTAKA

2.1 State of The Art

Tabel 2.1 merupakan penelitian terdahulu yang berhubungan dengan perancangan model dengan konsep *attrition employee*.

Judul	Hasil dan Pembahasan
Explaining and Predicting Employees Attrition: A Machine Learning Approach	Penelitian ini menggunakan dataset yang mencakup berbagai fitur terkait karyawan, seperti data demografis dan informasi

	pekerjaan. Metodologi yang diterapkan berfokus pada analisis univariat dan bivariat serta penggunaan svm, decision tree, dan random forest untuk memahami dan memprediksi pergantian karyawan. Hasil penelitian menunjukkan bahwa random forest memberikan akurasi mencapai 99% dibandingkan metode lainnya dalam mengidentifikasi fitur-fitur seperti gaji dan jumlah proyek yang ditangani sebagai indikator penting untuk prediksi attrition. Kelebihan utama dari pendekatan ini adalah visualisasi data yang komprehensif, yang membantu dalam pemahaman mendalam tentang hubungan antar variabel. Namun, kekurangannya adalah bahwa model ini mungkin kurang akurat dalam situasi dengan data yang sangat dinamis atau tidak terstruktur dengan baik.
Predicting Employee Attrition Using Machine Learning Techniques	Penelitian ini memanfaatkan dataset IBM yang berisi 35 fitur dan sekitar 1500 sampel. Beberapa algoritma klasifikasi diterapkan, termasuk Gaussian Naïve Bayes, yang memberikan performa terbaik dengan recall rate sebesar 54 persen dan false negative rate sebesar 4.5%. Kelebihan dari algoritma ini adalah kemampuannya dalam mengidentifikasi semua instance positif, namun recall rate yang relatif rendah menunjukkan bahwa beberapa instance positif mungkin terlewat.
A Machine Learning Approaches for Employee Retention Prediction	Penelitian ini menggunakan dataset yang berisi berbagai atribut terkait karyawan seperti pendidikan, pengalaman dan lainnya. Menggunakan knn, random forest dan svm, menghasilkan hasil terbaik dengan akurasi training, testing dan overall berturut turut 99.1%, 84.6% dan 91.8%. Kelebihan dari penelitian ini adalah eksplorasi mendalam terhadap berbagai teknik machine learning dan deep learning untuk meningkatkan akurasi prediksi. Namun, kekurangannya terletak pada kurangnya variasi dalam jenis data yang dapat membatasi generalisasi model ke industri lain.
Turnover Prediction in a Call Center: Behavior Evidence of Loss Aversion Using Random Forest and Naive Bayes Algorithm	Penelitian ini menggunakan dataset dari call center yang fokus pada perilaku karyawan. Random Forest dan Naive Bayes digunakan cukup kompetitif untuk memprediksi turnover berdasarkan data perilaku, dengan random forest berhasil memberikan hasil terbaik dalam hal akurasi sebesar 85% pada random forest. Kelebihan utama dari model ini adalah efektivitasnya dalam menangkap pola perilaku yang kompleks, meskipun analisis perilaku membutuhkan data yang sangat terperinci yang mungkin tidak terlalu tersedia.
Prediction of Employee Turnover in Organization Using Machine Learning Algorithms	Penelitian ini menggunakan dataset yang berisi berbagai informasi terkait demografi dan atribut pekerjaan. Algoritma Support Vector Machine digunakan untuk memprediksi niat karyawan untuk keluar, menunjukkan performa yang baik dengan akurasi yang tinggi. XGBoost sangat efektif untuk dataset dengan margin yang jelas antara kelas-kelas yang berbeda, tetapi bisa menjadi kurang efektif pada dataset yang sangat besar atau tidak terstruktur dengan baik. Performa terbaik AUC XGBoost mencapai 88%.

Tabel 2.1

2.2 Dasar Teori

Adapun istilah-istilah yang dipakai dalam penelitian ini antara lain.

A. Encoding
Encoding adalah proses mengubah informasi dari satu bentuk ke bentuk lain yang dapat diproses oleh sistem komputer atau algoritma machine learning. Ini melibatkan transformasi data mentah menjadi format yang sesuai untuk analisis lebih lanjut atau pelatihan model. Library encoder yang umum digunakan adalah, label encoder, one hot encoder, dan robust encoder.

B. Feature Scaling
Feature scaling adalah teknik statistik yang digunakan untuk menormalisasi rentang nilai pada fitur-fitur data sehingga seluruh fitur berada pada skala yang sama. Algoritma machine learning bekerja lebih baik ketika variabel numerik yang dimasukkan berada dalam skala yang serupa.

C. Correlation Matrix
Correlation Matrix atau korelasi matriks adalah tabel yang menampilkan koefisien korelasi antara berbagai variabel dalam suatu dataset. Setiap sel dalam matriks berisi nilai koefisien korelasi yang mengatur hubungan antara dua variabel. Nilai koefisien berkisar antara -1 hingga 1. Matriks korelasi ini dapat membantu memahami ketergantungan antara variabel-variabel.

D. *Oversampling* dan *Undersampling*
Ketidakseimbangan data dapat diatasi dengan metode *Oversampling* dan *Undersampling*. *Oversampling* melibatkan menggandakan sampel dari kelas minoritas, sehingga jumlah total sampel di kedua kelas menjadi lebih seimbang. Sebaliknya, *undersampling* menghapus sampe secara acak dari dataset, mengurangi jumlah sampel pada kelas mayoritas agar sebanding dengan kelas minoritas.

E. PCA
Principal Component Analysis (PCA) adalah metode pengurangan dimensi yang sering digunakan dalam pengolahan data. Tujuan metode ini adalah untuk mengurangi jumlah dimensi dalam dataset besar dengan ketentuan tertentu tanpa kehilangan informasi penting didalamnya. Dengan teknik tertentu PCA dapat membantu meningkatkan performa model dan mengurangi noise dalam data.

F. Logistic Regression
Logistik regression adalah model statistik yang digunakan untuk memprediksi probabilitas kejadian biner. Model ini menggunakan fungsi logit untuk mengubah nilai input menjadi probabilitas. Fungsi logit tersebut didefinisikan sebagai berikut.

Persamaan 2.1

$$logit(p) = ln(\frac{p}{1-p}) = \beta_0 + \beta_1 X_1 + ... + \beta_n X_n$$

G. KNN
K-Nearest Neighbor menyimpan semua instance dari training data untuk mengklasifikasikan instance baru. KNN menghitung jarak antara instance baru dan semua instance dalam training data. Instance baru akan diklasifikasikan berdasarkan mayoritas kelas dari k tetangga terdekat.

H. Naive Bayes
Algoritma berbasis probabilitas yang mengasumsikan independensi antar fitur dengan menghitung probabilitas masing - masing kelas. Menggunakan Teorema Bayes untuk menghitung probabilitas posterior dari setiap kelas untuk instance baru dan memprediksi kelas untuk probabilitas tertinggi.

I. Decision Tree
Model prediktif yang menggunakan struktur pohon keputusan untuk membuat prediksi. Cara kerja model ini memisahkan fitur yang memaksimalkan *information gain* atau mengurangi *impurity gini* atau entropy. Kemudian, membangun tree berdasarkan pemisahan dan membuat prediksi berdasarkan fitur input.

J. Random Forest
Salah satu metode ensemble learning yang mengadopsi konsep Decision Tree. Model ini akan mengambil sampel acak dari dataset. Setelah itu, membangun decision tree pada setiap sampel dengan tiap tree dilatih pada subset acak dan menggabungkan prediksi dari semua tree melalui *majority vote* untuk klasifikasinya.

K. XGBoost
Termasuk algoritma boosting yang meningkatkan akurasi model dengan menggabungkan banyak model sederhana. Pertama XGBoost akan membangun model secara bertahap dan menggabungkannya. Lalu, setiap model baru akan memberikan perbaikan dari kesalahan pada model sebelumnya. Terakhir model akan menggunakan fungsi objektif untuk mengoptimalkan *loss function* dan *regularization* untuk mencegah *overfitting*.

L. AdaBoost
Termasuk algoritma boosting yang mengkombinasikan banyak weak learners. Model ini bekerja dengan membangun model secara bertahap, di setiap model baru berfokus pada instance yang salah diklasifikasikan oleh model sebelumnya. Kemudian memberikan bobot lebih pada instance yang sulit untuk meningkatkan akurasi keseluruhan. Terakhir algoritma ini akan menggabungkan semua model dengan bobot yang ditentukan berdasarkan kinerjanya.

III. METODOLOGI

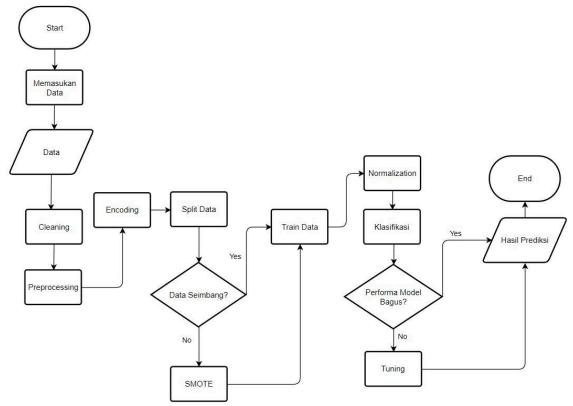


Diagram 3.1

Adapun metode penelitian yang digunakan pada penelitian ini akan berisi tentang pengumpulan data, cleaning, preprocessing, training, testing, dan evaluasi.

3.1 Pengumpulan Data

Dataset diambil dari platform kaggle dengan nama employee attrition. Dataset dengan format .csv ini berisi train dan test set untuk melakukan prediksi. Dataset memiliki ukuran 9.55 MB, setiap baris dari dataset ini menyimpan tentang detail informasi dari variasi atribut profil karyawan, termasuk demografi, dan kebutuhan personal. Dataset memiliki total 74,498 sampel dan 24 kolom dengan detail fitur disajikan dalam tabel berikut

Kolom	Penjelasan
Employee ID	Angka untuk identifikasi setiap karyawan.
Age	Umur karyawan
Gender	Jenis kelamin (Laki-laki / perempuan)
Years at Company	Jumlah tahun telah bekerja
Monthly Income	Gaji bulanan dalam dolar
Job Role	Peran karyawan.
Work - Life Balance	Keseimbangan dalam bekerja
Job Satisfaction	Kepuasan dalam bekerja
Performance Rating	Tingkat kinerja
Number of Promotions	Total promosi yang diterima
Overtime	Kerja lembur
Distance from Home	Jarak dari rumah
Education level	Jenjang pendidikan
Marital Status	Status perkawinan
Number of Dependents	Jumlah tanggungan
Job Level	Tingkat pekerjaan
Company Size	Ukuran pekerjaan
Company Tenure	Jumlah total tahun karyawan bekerja
Remote Work	Karyawan bekerja jarak jauh
Leadership Opportunities	Karyawan memiliki peluang kepemimpinan
Innovation Opportunities	Karyawan memiliki peluang untuk berinovasi
Company Reputation	Tingkat pengakuan perusahaan
Attrition	Apakah karyawan telah keluar perusahaan

Tabel 3.1

Dataset memiliki 0 nilai null dan 0 nilai duplicated value. Dengan masing masing persebaran nilai unik dan tipe data masing-masing kolom adalah sebagai berikut:

Kolom	Tipe Data	N-Unique
Employee ID	Int64	74498
Age	Int64	42
Gender	Object	2
Years at Company	Int64	51
Monthly Income	Object	5
Job Role	Int64	9842
Work - Life Balance	Object	4
Job Satisfaction	Object	4
Performance Rating	Object	4
Number of Promotions	Int64	5
Overtime	Object	2

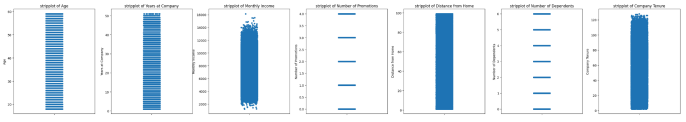
Distance from Home	Int64	99
Education level	Object	5
Marital Status	Object	3
Number of Dependents	Int64	7
Job Level	Object	3
Company Size	Object	3
Company Tenure	Int64	127
Remote Work	Object	2
Leadership Opportunities	Object	2
Innovation Opportunities	Object	2
Company Reputation	Object	4
Employee Recognition	Object	4
Attrition	Object	2

Tabel 3.2

3.2 Data Cleaning

3.2.1 Outlier

Pada bagian data cleaning, kita akan langsung melakukan cleaning outlier karena pada dasarnya data sudah bersih dari *null* dan *duplicated value*. Pertama, kita akan mengecek kolom bertipe numerik mana yang memiliki outlier,



Gambar 3.1

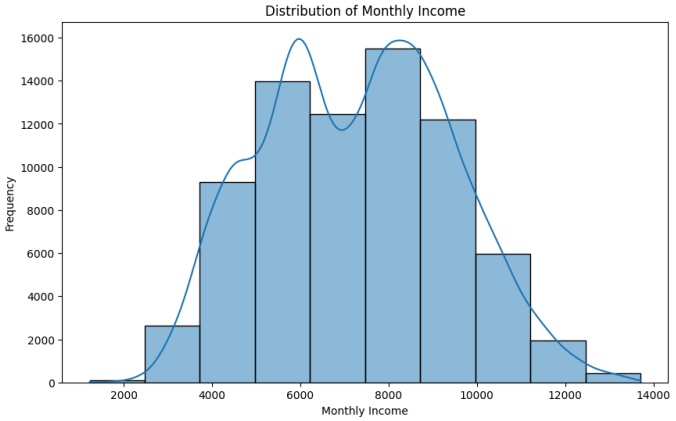
Terlihat pada visualisasi persebaran data menggunakan *stripplot* yang disediakan oleh library *seaborn*, kolom *monthly income* memiliki sedikit outlier. Oleh karena itu, kita akan membersihkannya menggunakan metode *IQR*.

3.3 Preprocessing

3.3.1 Continuous Value Handling

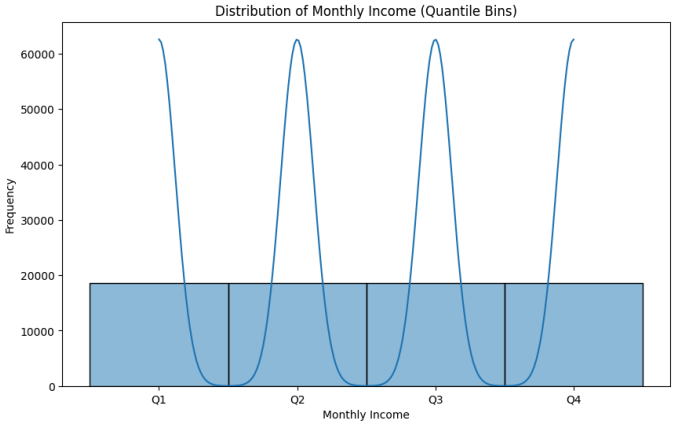
Ada beberapa kolom yang memiliki value yang continuous yang dimana memiliki sangat banyak *unique value*, diantaranya *Age, Monthly Income, Distance from home, Years at Company, dan Company Tenure.* Untuk menangani beberapa kolom tersebut, kita mengelompokkan persebarannya menggunakan *quartile* dari masing-masing persebaran data. Untuk setiap kelompok akan diberi label Q1 hingga Q4, tergantung pada interval berapa data tersebut berada. Untuk visualisasi, disini akan diwakilkan oleh kolom *Monthly Income,*

Sebelum pengelompokan



Gambar 3.2

Setelah pengelompokan



Gambar 3.3

3.3.2 Label Encoding

Dalam menangani beberapa kolom yang memiliki tipe data kategorik, kita akan menggunakan metode *label encoding* untuk mengubah tipe datanya menjadi numerik. Hal ini berguna agar model lebih mudah dalam melakukan prediksi. Cara kerja dari encoding sendiri adalah dengan mengubah setiap value kategorial yang unik kedalam angka yang unik. Sebagai contoh pada kolom *Monthly Income* yang sudah kita kelompokkan tadi value nya akan berubah dari Q1 menjadi 1, Q2 menjadi 2, Q3 menjadi 3, dan Q4 menjadi 4.

3.3.3 Train Test Split

Tahap selanjutnya, kita akan membagi data menjadi dua bagian, yaitu *Train* dan *Test. Train* merupakan bagian yang digunakan untuk melatih model dan *Test* merupakan bagian yang digunakan untuk menguji model yang sudah dilatih. Untuk permbagiannya sendiri, disini *Train* memiliki bagian 80% dan *TestI* memiliki bagian 20%.

3.3.4 Standardisation

Dilakukan standarisasi sedemikian rupa sehingga memiliki rata-rata (mean) 0 dan standar deviasi 1. Hal ini dilakukan agar fitur-fitur data memiliki skala yang sama dan tidak ada fitur yang mendominasi yang lain hanya karena perbedaan skala. Salah satu metode yang umum digunakan untuk standardisasi adalah dengan menggunakan StandardScaler dari library scikit-learn.

StandardScaler bekerja dengan mengurangi nilai mean dari setiap fitur dan kemudian membaginya dengan standar deviasi fitur tersebut. Rumus yang digunakan oleh StandardScaler adalah sebagai berikut:

$$z = \frac{x - \mu}{\sigma}$$

Persamaan 3.1

Dimana:

- z adalah nilai standar dari fitur setelah standardisasi.
- x adalah nilai asli dari fitur.
- μ adalah nilai rata-rata dari fitur.
- σ adalah nilai standar deviasi dari fitur.

Proses ini membantu dalam meningkatkan performa algoritma machine learning yang sensitif terhadap skala fitur, seperti algoritma berbasis gradien (misalnya, regresi logistik, SVM, dan neural networks).

3.3.5 Modeling

Tahap terakhir adalah modeling. Tahap ini merupakan tahap dimana kita akan menggunakan beberapa model machine learning terhadap data yang sudah di *split*. Untuk model yang digunakan diantaranya, *Logistic Regression, KNN, Naive Bayes, Decision Tree, Random Forest, SVM, XGBoost, dan AdaBoost.* Dalam menggunakan model-model tersebut, kita membuat beberapa skenario untuk melihat dan menentukan skenario mana yang paling bagus untuk data yang kita pilih. Skenario pertama kita akan menggunakan data yang belum di standarisasi, skenario kedua kita menggunakan data yang sudah distandarisasi, skenario ketiga kita menggunakan data yang sudah standarisasi dengan menambahkan *feature selection,* dan skenario keempat kita menggunakan data yang sudah distandarisasi beserta *hyperparameter tuning.*

IV. HASIL DAN PEMBAHASAN

4.1 Skenario 1

Pada skenario pertama, pengujian evaluasi dilakukan training model dengan dataset tanpa melalui normalisasi. Berikut adalah hasil metrics evaluasi yang didapat

Model	Accuracy	Precision	Recall	F1
Logistic Regression	0.6539	0.6563	0.7350	0.6906
KNN	0.5152	0.5383	0.5549	0.5465
Naive Bayes	0.7251	0.7615	0.7579	0.7270
SVC	0.5264	0.5264	1	0.6897
Decision Tree	0.6642	0.6838	0.6736	0.6786
Random Forest	0.7464	0.7641	0.0.7496	0.7568
XGBoost	0.7489	0.7656	0.7537	0.7596
AdaBoost	0.7603	0.7727	0.7717	0.7722

Tabel 4.1

Berikut adalah hasil perbandingan matriks evaluasi model skenario 1

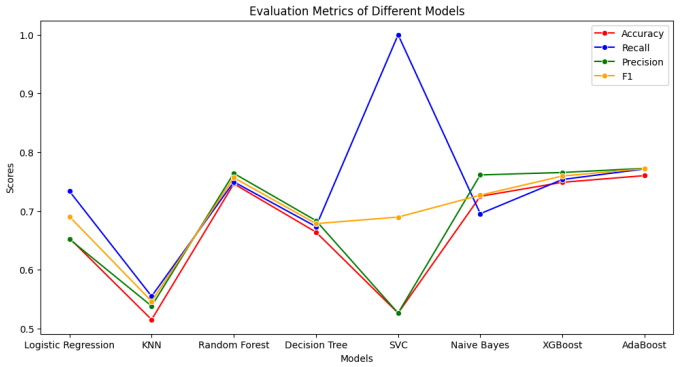


Diagram 4.1

4.2 Skenario 2

Pada skenario kedua, pengujian evaluasi dilakukan training model dengan dataset yang sudah melalui normalisasi. Berikut adalah hasil metrics evaluasi yang didapat

Model	Accuracy	Precision	Recall	F1
Logistic Regression	0.7285	0.7433	0.7395	0.7414
KNN	0.6912	0.7041	0.7129	0.7085
Naive Bayes	0.7233	0.7612	0.7590	0.7245
SVC	0.7465	0.7784	0.7708	0.7506
Decision Tree	0.6629	0.6820	0.6738	0.6779
Random	0.7470	0.7636	0.7523	0.7579

Forest				
XGBoost	0.7489	0.7656	0.7537	0.7596
AdaBoost	0.7603	0.7727	0.7717	0.7722

Tabel 4.2

Berikut adalah matriks evaluasi dari pemodelan skenario 2

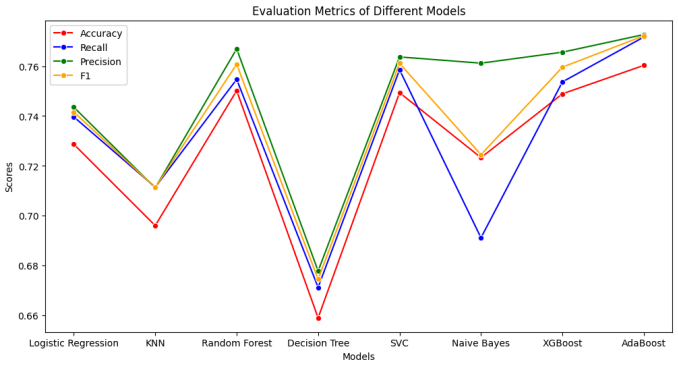


Diagram 4.2

4.3 Skenario 3

Pada skenario ketiga, pengujian evaluasi dilakukan training model dengan dataset yang sudah melalui normalisasi dan feature selection dengan correlation matrix > 0.1 pada kolom attrition. Berikut adalah hasil metrics evaluasi yang didapat

Model	Accuracy	Precision	Recall	F1
Logistic Regression	0.6964	0.7119	0.7109	0.7114
KNN	0.6318	0.6442	0.6395	0.7085
Naive Bayes	0.6904	0.7388	0.7497	0.6964
SVC	0.6983	0.7171	0.7047	0.7109
Decision Tree	0.6191	0.6424	0.6237	0.6329
Random Forest	0.6746	0.6932	0.6849	0.6890
XGBoost	0.7040	0.7263	0.7060	0.7140
AdaBoost	0.7125	0.7324	0.7150	0.7236

Tabel 4.3

Berikut adalah hasil matriks evaluasi yang didapatkan pada skenario 3

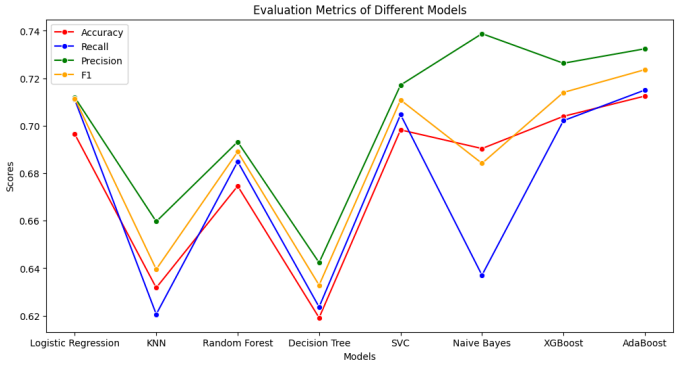


Diagram 4.3

4.4 Skenario 4

Pada skenario keempat, pengujian evaluasi berfokus pada 3 model yang memiliki performa paling stabil. Model tersebut dilakukan hyperparameter tuning model dengan dataset yang sudah melalui normalisasi tanpa menggunakan feature selection. Berikut adalah best parameter yang digunakan

Pengujian	Model	Best Parameter
1	Random Forest	max_depth = 13 max_features = log2 min_samples_leaf = 2 min_samples_split = 18 n_estimators = 995
2	XGBoost	colsample_bytree = 0.7944 learning_rate = 0.1255 max_depth = 1 n_estimators = 355
3	AdaBoost	learning_rate = 0.3768 n_estimators = 363
4	AdaBoost	estimator = DecisionTreeClassifier(max_depth=1) learning_rate = 0.01 n_estimators = 50
5	AdaBoost	estimator__class_weight = {0: 1, 1: 1} estimator__max_depth = 1 learning_rate = 0.01 n_estimators = 50

Tabel 4.4

Berikut adalah hasil evaluasi metrics yang didapat

Pengujian	Accuracy	Precision	Recall	F1
1	0.7655	0.7708	0.7713	0.7709
2	0.7676	0.7731	0.7711	0.7721

3	0.7604	0.7723	0.7719	0.7727
4	0.6610	0.7403	0.8609	0.7322
5	0.6628	0.7321	0.8614	0.7307

Tabel 4.5

Berikut adalah grafik hasil matriks evaluasi skenario 4

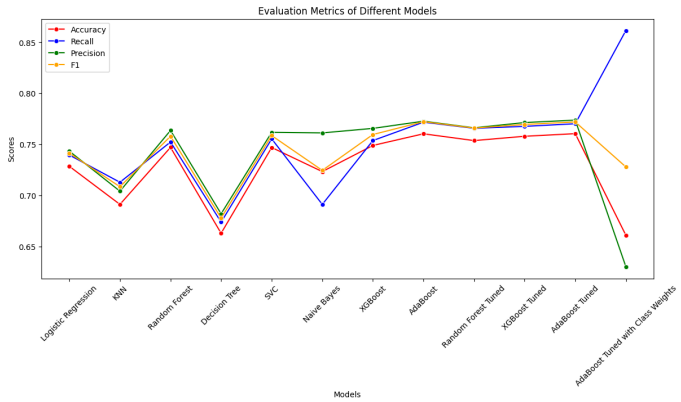


Diagram 4.4

4.5 Analisis Hasil

Kami telah melakukan data cleaning dan preprocessing pada dataset sebelum melakukan pemodelan agar dataset bersih dari outlier dan missing value, selanjutnya dilakukan pengkuartilan pada kolom yang memiliki data type int atau continous, setelah itu dilakukan coding agar dataset dapat dilakukan pemodelan. Kami memutuskan untuk tidak dilakukan SMOTE pada dataset ini karena data sudah balance dari awal.

Pemodelan dilakukan dengan 8 model yang berfokus pada nilai recall. Nilai recall dipilih karena dapat mengukur sejauh mana model dapat mengidentifikasi semua kasus positif yang sebenarnya, karena kita tidak ingin ada kasus negatif yang terprediksi positif yang mana akan merugikan perusahaan. Model dengan hasil yang paling optimal dan stabil ada pada tiga model yaitu, Random Forest, XGBoost, AdaBoost baik di bagian skenario I dan skenario II. Kami meyakini dilakukannya scaling dapat meningkat akurasi dan recall karena terjadinya penurunan overfitting, peningkatan keandalan statistik, peningkatan kebisingan data dan peningkatan kemampuan menemukan fitur penting, uniknya di dalam skenario I, kami menemukan adanya anomali bahwa recall dengan model SVC mendapatkan nilai 1, yang artinya model berhasil mengidentifikasi semua kasus positif yang sebenarnya dengan benar.

Selanjutnya pada skenario III, kami ingin mencoba menggunakan metode feature selection namun ternyata didapat hasil matriks evaluasi yang cenderung turun daripada pemodelan dengan seluruh kolom. Hal ini dikarenakan kurangnya informasi yang berkorelasi dengan kelas target. akurasi dari pemodelan dengan melakukan hyperparameter tuning. Pada pengujian satu hingga tiga peningkatan dari akurasi dan recall nampak tidak jauh terlihat. Pada saat ini model sudah mencapai titik maksimal, sehingga pada pengujian empat dan lima, hyperparameter yang digunakan di set untuk memaksimalkan nilai recall. Dampaknya nilai metrics lainnya berkurang. Pada Model AdaBoost yang telah di tuning dengan parameter *weight class,* model mencapai titik recall terbaiknya yaitu sebesar 0.8614.

V. KESIMPULAN

5.1 Kesimpulan

Setelah dilakukan percobaan didapat kesimpulan yang menjawab permasalahan antara lain:

a. Model Random Forest, XGBoost, dan AdaBoost cukup bersaing untuk melakukan prediksi.
b. Normalisasi berpengaruh pada hasil akurasi dan recall model dan Hyperparameter Tunning tidak serta merta akan memberikan peningkatan akurasi model.

5.2 Saran

Adapun saran yang kami berikan dari penelitian ini adalah:

a. Menggunakan teknik feature engineer
b. Menggunakan dataset yang lebih beragam
c. Menggunakan model deep learning untuk mendapatkan model yang optimal

.REFERENSI

[1] Fallucchi, F., Coladangelo, M., Giuliano, R., & De Luca, E. W. (2020). Predicting Employee Attrition Using Machine Learning Techniques. *Computers, 9*(4), 86. https://doi.org/10.3390/computers9040086

[2] G. Marvin, M. Jackson and M. G. R. Alam, "A Machine Learning Approach for Employee Retention Prediction," *2021 IEEE Region 10 Symposium (TENSYMP)*, Jeju, Korea, Republic of, 2021, pp. 1-8, doi: 10.1109/TENSYMP52854.2021.9550921.

[3] S. Zhang, "Challenges in KNN Classification," *IEEE Transactions on Knowledge and Data Engineering*, vol. 34, no. 10, pp. 4663, Oct. 2022.

[4] Wickramasinghe, I., Kalutarage, H. Naive Bayes: applications, variations and vulnerabilities: a review of literature with code snippets for implementation. *Soft Comput* **25**, 2277–2293 (2021). https://doi.org/10.1007/s00500-020-05297-6

[5] Jijo, B. T., & Abdulazeez, A. M. (2021). Classification Based on Decision Tree Algorithm for Machine Learning. *IT Department, Technical College of Informatics Akre, Duhok Polytechnic University, Duhok, Kurdistan Region, Iraq*. Vol. 02, No. 01, pp. 20–28. ISSN: 2708-0757.

[6] Fallucchi, F., Coladangelo, M., Giuliano, R., & William De Luca, E. (2020). Predicting employee attrition using machine learning techniques. *Computers*, *9*(4), 86.

[7] Jin, Z., Shang, J., Zhu, Q., Ling, C., Xie, W., & Qiang, B. (2020). RFRSF: Employee turnover prediction based on random forests and survival analysis. In *Web Information Systems Engineering–WISE 2020: 21st International Conference, Amsterdam, The Netherlands, October 20–24, 2020, Proceedings, Part II 21* (pp. 503-515). Springer International Publishing.

[8] Alsheref, F. K., Fattoh, I. E., & M. Ead, W. (2022). Automated prediction of employee attrition using ensemble model based on machine learning algorithms. *Computational Intelligence and Neuroscience*, *2022*(1), 7728668.

[9] Lazzari, M., Alvarez, J. M., & Ruggieri, S. (2022). Predicting and explaining employee turnover intention. *International Journal of Data Science and Analytics*, *14*(3), 279-292.

[10] Pessach, D., Singer, G., Avrahami, D., Ben-Gal, H. C., Shmueli, E., & Ben-Gal, I. (2020). Employees recruitment: A prescriptive analytics approach via machine learning and mathematical programming. *Decision Support Systems*, *134*, 113290.

[11] Zhu, H. (2021). Research on human resource recommendation algorithm based on machine learning. *Scientific programming*, *2021*(1), 8387277.

[12] Muncie, T. (2020). *Using machine learning models to predict student retention: Building a state-wide early warning system* (Doctoral dissertation, Morehead State University).

[13] Paul, T., & Bommu, R. (2024). Strategic Employee Performance Analysis in the USA: Leveraging Intelligent Machine Learning Algorithms. *Unique Endeavor in Business & Social Sciences*, *3*(1), 113-124.

[14] Dutta, S., & Bandyopadhyay, S. K. (2020). Employee attrition prediction using neural network cross validation method. *International Journal of Commerce and Management Research*, *6*(3), 80-85

[15] Sainju, B., Hartwell, C., & Edwards, J. (2021). Job satisfaction and employee turnover determinants in Fortune 50 companies: Insights from employee reviews from Indeed. com. *Decision Support Systems*, *148*, 113582.