Understanding Churn Rate: The Random Forest Decision Tree Approach in Bank Analysis

PEMAHAMAN TINGKAT CHURN: PENDEKATAN DECISION TREE RANDOM FOREST DALAM ANALISIS BANK

Joe Marcello[1], Muhammad Evan Julian Priyasa[2], Febianus Felix Widisulistiyo[3], Rivaldo Yosia Himawan[4]
[1,2,3,4] Department of Informatics, Universitas Multimedia Nusantara, Tangerang, Banten, Indonesia
[1]email : joe.marcello@student.umn.ac.id
[2]email : muhammad.evan@student.umn.ac.id
[3]email : febianus.felix@student.umn.ac.id
[4]email : rivaldo.yosia@student.umn.ac.id

Abstract*—Penelitian ini berfokus pada evaluasi keefektifan Pengklasifikasi Random Forest dalam memprediksi churn pelanggan dan mengidentifikasi fitur yang berpengaruh. Dengan menggunakan teknik analisis data canggih, termasuk model pembelajaran mesin, khususnya Random Forest, penelitian ini menunjukkan hasil yang menjanjikan. Dataset dari sektor perbankan dianalisis, mencakup berbagai atribut pelanggan dan indikator churn. Melalui eksperimen yang cermat, Random Forest menunjukkan kinerja yang lebih baik daripada Decision Tree, dengan akurasi pengujian mencapai 83,9%. Penyetelan hyperparameter yang dioptimalkan pada Random Forest berkontribusi pada kinerja yang kuat, menunjukkan kemampuan generalisasi yang lebih baik dibandingkan dengan Decision Tree. Selain itu, penelitian ini menggunakan validasi silang untuk memastikan stabilitas dan keandalan kinerja model. Skor validasi silang rata-rata sebesar 0,844 menegaskan konsistensi dan efektivitas Random Forest dalam memprediksi churn pelanggan. Evaluasi rinci melalui laporan klasifikasi memberikan wawasan tentang presisi, recall, dan F1-score model. Meskipun kedua model menunjukkan kinerja yang lebih baik dalam memprediksi non-churn, Random Forest menunjukkan peningkatan signifikan dalam memprediksi churn, terutama dalam metrik recall dan F1-score.

Kata kunci: churn pelanggan, industri perbankan, Random Forest Classifier, analisis prediktif, akurasi model.

I. INTRODUCTION

Dalam era globalisasi yang penuh persaingan ketat, industri perbankan menghadapi tantangan besar dalam menjaga keberlangsungan operasinya [1]. Perubahan dinamis di pasar keuangan yang dipicu oleh teknologi yang terus berkembang serta perubahan preferensi konsumen, membuat pelanggan semakin menuntut layanan yang lebih personal, mudah diakses, dan fleksibel [2]. Selain itu, regulasi yang berubah-ubah mengharuskan bank untuk beradaptasi cepat agar tetap relevan dan kompetitif. Dalam perjuangan ini, salah satu metrik penting yang harus dipertimbangkan adalah tingkat churn, yang mencerminkan persentase pelanggan yang memilih untuk meninggalkan layanan bank dalam periode waktu tertentu. Nasabah mungkin berpindah ke bank lain karena berbagai alasan seperti layanan keuangan yang lebih baik dengan biaya lebih rendah, lokasi cabang yang lebih nyaman, suku bunga rendah, dan banyak faktor lainnya [3]. Tingkat churn yang tinggi tidak hanya berdampak negatif pada pendapatan bank, tetapi juga dapat mengancam reputasi dan kepercayaan dari pemegang saham serta pasar secara keseluruhan [4].

Melacak dan memahami faktor-faktor yang mempengaruhi keputusan pelanggan untuk meninggalkan layanan bank menjadi krusial dalam menghadapi tantangan ini. Penggunaan teknik analisis data yang canggih, seperti model prediktif berbasis machine learning, menjadi semakin penting dalam upaya memahami perilaku pelanggan dan merumuskan strategi yang efektif untuk meminimalkan churn. Investigasi churn pelanggan di industri perbankan menggunakan pendekatan machine learning dan aplikasi visualisasi untuk ilmu data dan manajemen dapat memberikan wawasan yang berharga tentang perilaku churn pelanggan dalam konteks perbankan [8].

Dalam menghadapi tantangan prediksi churn pelanggan di industri perbankan, beberapa penelitian telah mengatasi masalah data yang tidak seimbang [10]. Data tidak seimbang merupakan situasi di mana jumlah sampel di setiap kelas tidak seimbang, sehingga dapat mempengaruhi kinerja model machine learning. Penanganan data tidak seimbang menjadi faktor penting dalam memastikan keakuratan model prediksi churn pelanggan. Dengan memanfaatkan pendekatan yang didukung oleh data, bank dapat mengidentifikasi pola-pola yang tersembunyi dan faktor-faktor kritis yang mempengaruhi keputusan pelanggan [9], sehingga memungkinkan tindakan pencegahan yang proaktif diambil. Penelitian tentang prediksi churn pelanggan tidak hanya memiliki relevansi bisnis yang langsung, tetapi juga menjadi fondasi bagi inovasi dan strategi yang memungkinkan bank untuk tetap berada di garis depan dalam industri yang berubah dengan

cepat ini [5]. Model machine learning terbukti secara efektif memprediksi churn dan membantu bank untuk meningkatkan retensi nasabah [11].

Setelah pola dan faktor yang mempengaruhi churn diketahui, bank dapat mengambil keputusan untuk mencari solusi mengurangi churn. Berkurangnya pelanggan memperburuk reputasi bank, sehingga tindakan pencegahan yang tepat penting diambil [7]. Contoh penerapan Random Forest dalam menganalisis churn nasabah terdapat pada sebuah skripsi di industri perbankan. Algoritma Random Forest, dengan keunggulan ansambel pohon keputusan, mampu meningkatkan akurasi prediksi dan mengatasi overfitting. Penelitian ini mengevaluasi efektivitas Random Forest dalam memprediksi churn nasabah menggunakan partisi data uji dan latih yang berbeda, dan hasilnya diharapkan memberikan wawasan berharga bagi bank dalam mengidentifikasi dan mengurangi churn nasabah [18].

Penelitian ini bertujuan untuk menjawab: akurasi Random Forest Classifier dalam memprediksi churn pelanggan, fitur-fitur yang mempengaruhi churn pada bank, dan kinerja model prediksi churn berdasarkan Precision, Recall, F1 Score, dan Support. Batasan penelitian mencakup penggunaan satu dataset dan variabel di dalamnya, serta fokus pada klasifikasi churn sebagai Exited dan Not-Exited menggunakan Random Forest Classifier tanpa metode prediksi lainnya.

Penelitian ini bertujuan untuk menilai akurasi Random Forest Classifier dalam memprediksi churn pelanggan di perbankan, mengidentifikasi fitur-fitur yang paling mempengaruhi churn, dan mengevaluasi kinerja model prediksi churn berdasarkan Precision, Recall, F1 Score, dan Support. Manfaatnya diharapkan membantu bank dalam mengurangi churn pelanggan dan mengevaluasi model klasifikasi churn. Penelitian ini juga mengidentifikasi faktor-faktor penting dalam pembuatan model klasifikasi churn di perbankan.

Menghadapi tantangan churn dalam perbankan memerlukan pendekatan komprehensif berbasis data. Penggunaan model prediktif seperti Random Forest memberikan wawasan mendalam tentang perilaku pelanggan, memungkinkan bank mengambil tindakan efektif untuk mengurangi churn dan mempertahankan pelanggan. Penelitian ini relevan bagi industri perbankan dan berkontribusi signifikan pada pengembangan strategi dan inovasi di sektor ini.

II. LITERATUR REVIEW

A. Fraud Detection

Churn, atau tingkat pergantian pelanggan, adalah masalah serius yang dihadapi oleh industri perbankan di era globalisasi yang penuh persaingan. Dalam beberapa tahun terakhir, penelitian tentang churn bank telah menjadi topik yang semakin penting karena dampaknya terhadap pendapatan dan reputasi bank [12].

Menganalisa atau mengidentifikasikan faktor-faktor sebagai penyebab utama churn pelanggan dalam industri perbankan seperti kepuasan pelanggan, kualitas layanan, dan faktor lainnya yang berkaitan dengan pengalaman pelanggan [6]. kesimpulan nya analisis churn pelanggan melibatkan penelitian tentang faktor-faktor yang menjadi penyebab utama churn pelanggan, seperti kepuasan pelanggan, kualitas layanan, dan faktor lain yang terkait dengan pengalaman pelanggan.

B. Machine Learning

Machine Learning (ML) adalah cabang dari kecerdasan buatan (Artificial Intelligence) yang memungkinkan sistem untuk mempelajari pola atau informasi dari data yang ada tanpa perlu secara eksplisit diprogram [19]. Dengan menggunakan algoritma dan model matematika, mesin mampu mengidentifikasi pola-pola kompleks dalam data dan menghasilkan keputusan atau prediksi yang bermanfaat.

Dalam definisi ini, terdapat penekanan pada kemampuan mesin untuk mempelajari dari data yang ada tanpa memerlukan instruksi eksplisit dari programmer. Hal ini dilakukan melalui proses pengembangan algoritma dan model matematika yang memungkinkan mesin untuk mengenali pola dalam data dan melakukan tugas-tugas tertentu, seperti membuat prediksi atau mengambil keputusan [20].

C. Supervised Learning

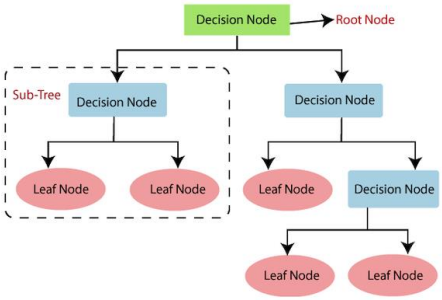
Supervised Learning adalah salah satu pendekatan utama dalam Machine Learning di mana model mempelajari hubungan antara input dan output dari data yang diberikan [21]. Dalam konteks ini, input juga dikenal sebagai fitur atau atribut, sedangkan output biasanya disebut sebagai label atau target.

Dalam Supervised Learning, model dilatih menggunakan dataset yang telah diberi label, yang berarti setiap contoh dalam dataset memiliki pasangan input-output yang sesuai [21][22]. Tujuan utama dari Supervised Learning adalah untuk menghasilkan fungsi yang dapat memetakan setiap input ke output yang sesuai dengan akurasi yang tinggi. Model ini kemudian dapat digunakan untuk memprediksi output untuk data baru yang belum pernah dilihat sebelumnya.

D. Regression

Decision tree merupakan salah satu algoritma klasifikasi yang paling populer dan telah digunakan secara luas dalam berbagai bidang, termasuk ilmu komputer, statistik, dan kecerdasan buatan [13]. Algoritma ini bertujuan untuk membangun model prediktif dalam bentuk struktur pohon keputusan, di

mana setiap simpul dalam pohon tersebut mewakili suatu keputusan atau prediksi berdasarkan pada serangkaian aturan yang didefinisikan [14].



Gambar 1 Struktur Decision Tree [14]

Keputusan dalam algoritma Pohon Keputusan didasarkan pada dua konsep perhitungan, yaitu *entropy* dan *information gain*. *Entropy* mengukur tingkat ketidakmurnian atau kekacauan data, dengan nilai yang berkisar antara 0 hingga 1. Semakin mendekati nol, ini menunjukkan bahwa data sangat teratur dan setiap atribut dalam data memiliki *class* yang seragam. Berikut adalah rumus daripada entropy :

$$Entropy(S) = \sum_{i=1}^{c} P_i log\ 2^{p_i}$$

Di mana:

- *S* adalah himpunan data yang sedang dipertimbangkan.
- *c* adalah jumlah kelas atau label yang mungkin dalam himpunan data.
- P_iP_i adalah proporsi frekuensi relatif dari setiap kelas *i* dalam himpunan data *S*

Sedangkan information gain merupakan pengukuran perbedaan en tropi pada database yang telah dilakukan pemisahan, hal ini bertujuan untuk menentukan segmentasi feature pada setiap node saat pembuatan model. Adapun rumus daripada perhitungan information gain adalah sebagai berikut:

$$InformationGain = Entropy(S) - \Sigma_{i-1}^{n} \frac{N_i}{N} x$$

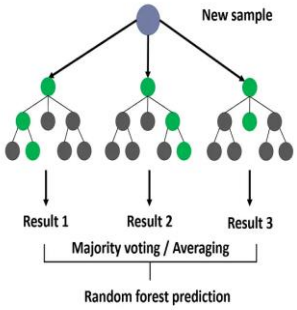
$$Entropy(S_i)$$

- *S* adalah himpunan data sebelum pemisahan.
- *n* adalah jumlah subset yang dihasilkan setelah pemisahan.
- *Ni* adalah jumlah sampel dalam subset Si.
- *N* adalah jumlah total sampel dalam himpunan data sebelum pemisahan.
- *Entropy*(*S*) adalah entropi sebelum pemisahan.
- *Entropy*(S_iS_i) adalah entropi dari setiap subset S_i S_i setelah pemisahan.

Algoritma decision tree dapat digunakan untuk membangun model prediktif yang dapat mengidentifikasi pola-pola perilaku pelanggan yang berpotensi menyebabkan churn [15]. Dengan membagi dataset pelanggan menjadi subset-subset yang lebih kecil berdasarkan pada atribut-atribut yang relevan, seperti preferensi layanan, frekuensi transaksi, atau tingkat kepuasan, decision tree dapat membantu bank dalam mengidentifikasi kelompok-kelompok pelanggan yang berisiko tinggi untuk meninggalkan layanan.

D. Logistic Regression

Konsep dasar dari Random Forest adalah membangun banyak pohon keputusan secara acak dari subset data yang berbeda, dan kemudian menggabungkan hasil prediksi dari semua pohon tersebut [17].Random forest terdiri dari kumpulan pohon keputusan yang bekerja secara independen[16].Secara khusus, Random Forest mengimplementasikan dua jenis randomness: randomness dalam pemilihan sampel data yang digunakan untuk melatih setiap pohon keputusan, dan randomness dalam pemilihan atribut yang digunakan untuk membagi setiap simpul dalam pohon keputusan.



Gambar 2 Visualisasi Random Forest [24]

Selain itu, Random Forest juga memiliki kemampuan untuk mengatasi ketidakseimbangan kelas yang sering terjadi dalam dataset churn bank, di mana jumlah pelanggan yang churn mungkin jauh lebih sedikit daripada yang tidak churn. Dengan mempertimbangkan sejumlah besar pohon keputusan yang berbeda, Random Forest dapat menghasilkan prediksi yang lebih seimbang dan dapat diandalkan.

E. Confusion Matrix

Confusion matrix adalah tabel yang digunakan untuk mengevaluasi kinerja model klasifikasi dengan membandingkan prediksi model dengan nilai aktual dari data [25]. Confusion matrix memiliki empat sel, yaitu

True Positive (TP)

Merupakan kasus di mana model dengan benar memprediksi bahwa suatu sampel adalah positif (misalnya, pelanggan yang melakukan churn) dan prediksi tersebut sesuai dengan kebenaran.

True Negative (TN)

Merupakan kasus di mana model dengan benar memprediksi bahwa suatu sampel adalah negatif (misalnya, pelanggan yang tidak melakukan churn) dan prediksi tersebut sesuai dengan kebenaran.

False Positive (FP)

Merupakan kasus di mana model salah memprediksi bahwa suatu sampel adalah positif (misalnya, model memprediksi bahwa pelanggan tidak akan churn, padahal sebenarnya mereka akan churn).

False Negative (FN)

Merupakan kasus di mana model salah memprediksi bahwa suatu sampel adalah negatif (misalnya, model memprediksi bahwa pelanggan akan churn, padahal sebenarnya mereka tidak akan churn).

Dari confusion matrix, kita dapat menghitung berbagai metrik evaluasi kinerja model seperti presisi, dan recall

$$Precision = \frac{True\ Positive\ (TP)}{True\ Positive\ (TP)\ +\ False\ Positive\ (FP)}$$

Precision dihitung dengan membagi jumlah True Positive oleh jumlah keseluruhan prediksi positif yang dibuat oleh model. Precision memberikan gambaran tentang seberapa akurat model dalam mengklasifikasikan hasil positif. Semakin tinggi nilai precision, semakin sedikit False Positive yang dihasilkan oleh model, menunjukkan bahwa model lebih konservatif dalam membuat prediksi positif.

$$Recall = \frac{True\ Positive\ (TP)}{True\ Positive\ (TP)\ +\ False\ Negative\ (FN)}$$

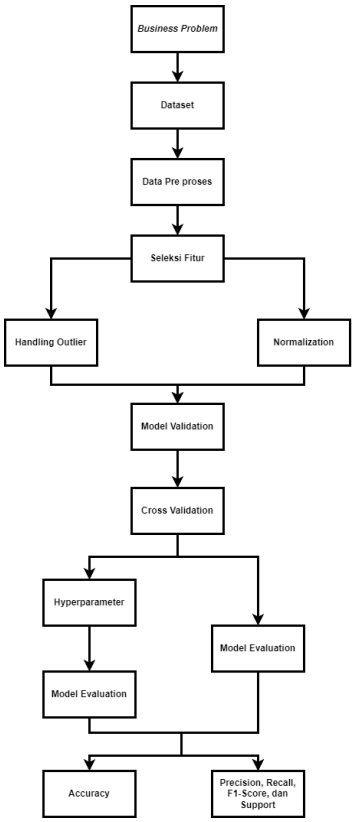
Recall dihitung dengan membagi jumlah True Positive oleh jumlah keseluruhan instance positif yang seharusnya diidentifikasi oleh model. Ini menunjukkan seberapa baik model dapat mengidentifikasi semua instance positif dalam dataset. Semakin tinggi nilai recall, semakin sedikit False Negative yang dihasilkan oleh model, menunjukkan kemampuan model dalam mengenali sebagian besar instance positif yang ada.

E. F1 Score

F1-Score merupakan matrik evaluasi kinerja model yang menggabungkan presisi dan recall model [26]. F-score adalah rata-rata harmonik dari presisi dan recall, dan sering digunakan untuk mengukur keseimbangan antara kedua metrik tersebut [26]. F-score memberikan gambaran yang lebih baik tentang kinerja model klasifikasi, terutama ketika kelas yang tidak seimbang dalam distribusi frekuensinya.

F1 Score memberikan gambaran keseimbangan antara precision dan recall model. Semakin tinggi nilai F1 Score, semakin baik keseimbangan antara precision dan recall. Metrik ini cocok digunakan ketika ingin mempertimbangkan false positives dan false negatives secara seimbang.

III. METHODOLOGY



Gambar 3. Diagram Metodologi Penelitian

A. Data Collection

Dalam studi ini, sejumlah langkah dijalankan untuk mengidentifikasi churn di industri perbankan. Tahap awal mencakup pemahaman perpindahan pelanggan dan dampak negatifnya terhadap kesehatan finansial perusahaan, yang memicu kebutuhan akan solusi efektif untuk mendeteksi dan mencegah churn. Oleh karena itu, penelitian ini merinci langkah-langkah dari awal hingga akhir. Dataset yang digunakan adalah "Churn for Bank Customers" oleh Mehmet Akturk, diperbarui empat tahun lalu, dan dapat diakses di: https://www.kaggle.com/datasets/mathchi/churn-for-bank-customers. Dataset ini menyediakan informasi komprehensif tentang berbagai variabel yang mempengaruhi keputusan pelanggan untuk meninggalkan layanan bank.

B. Exploratory Data Analysis

Pada tahap Exploratory Data Analysis (EDA), dataset "Churn for Bank Customers" akan diperiksa untuk memahami karakteristiknya. Melalui analisis ini,

pola-pola dan hubungan antar variabel yang relevan untuk strategi mengurangi churn pelanggan akan diidentifikasi. Bantuan visualisasi data akan digunakan untuk memperoleh pemahaman intuitif tentang distribusi dan pola data yang signifikan.

```
churn = pd.read_csv("churn.csv")
churn
```

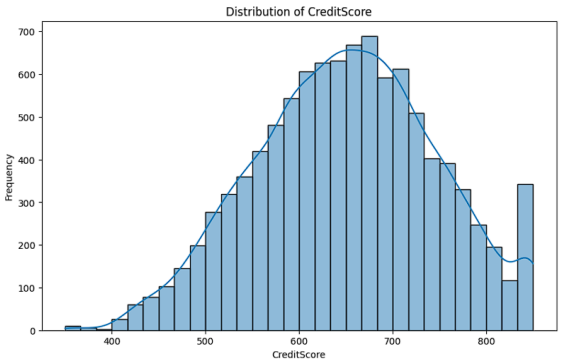
	RowNumber	CustomerId	Surname	CreditScore	Geography	Gender	Age	Tenure	Balance	NumOfProducts	HasCrCard	IsActiveMember	EstimatedSalary	E
0	1	15634602	Hargrave	619	France	Female	42	2	0.00	1	1	1	101348.88	
1	2	15647311	Hill	608	Spain	Female	41	1	83807.86	1	0	1	112542.58	
2	3	15619304	Onio	502	France	Female	42	8	159660.80	3	1	0	113931.57	
3	4	15701354	Boni	699	France	Female	39	1	0.00	2	0	0	93826.63	
4	5	15737888	Mitchell	850	Spain	Female	43	2	125510.82	1	1	1	79084.10	

Gambar 4. Read Data

Informasi yang tercakup dalam dataset ini meliputi nomor baris data ('RowNumber'), identifikasi pelanggan ('CustomerId'), nama belakang pelanggan ('Surname'), skor kredit ('CreditScore'), wilayah geografis ('Geography'), jenis kelamin ('Gender'), usia ('Age'), lama menjadi nasabah ('Tenure'), saldo ('Balance'), jumlah produk yang dimiliki ('NumOfProducts'), kepemilikan kartu kredit ('HasCrCard'), keanggotaan aktif ('IsActiveMember'), estimasi gaji ('EstimatedSalary'), dan status keluar ('Exited').

Informasi yang tersedia dalam dataset ini menjadi landasan yang kokoh untuk memperoleh pemahaman yang lebih mendalam tentang pola transaksi yang berkaitan dengan kecurangan. Oleh karena itu, hasil analisis data eksploratori yang dilakukan oleh peneliti.

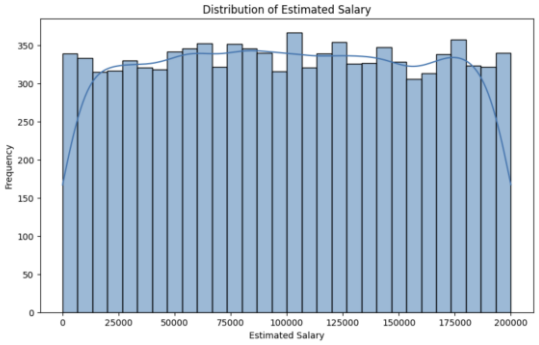
1. Distribusi Credit Score



Gambar 5 Histogram Distribusi Credit Score

Pada tahap ini, dilakukan visualisasi terhadap 'CreditScore' dengan menganalisis grafik distribusi. Berdasarkan grafik distribusi tersebut, teramati bahwa pucuk grafik berada di sekitar rentang 600 hingga 700, menunjukkan distribusi terbesar terkonsentrasi di antara nilai tersebut. Namun, ditemukan satu puncak grafik yang menonjol pada angka 850, mengindikasikan adanya outlier yang signifikan pada bagian tersebut.

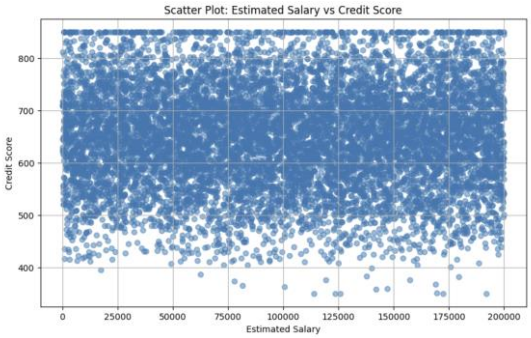
2. Distribusi Estimated Salary



Gambar 6 Histogram Distribusi Estimated Salary

Grafik ini menunjukkan distribusi gaji diperkirakan di Amerika Serikat. Sumbu X menunjukkan gaji diperkirakan dalam rentang dari 0 hingga 200.000 dolar. Sumbu Y menunjukkan frekuensi individu dengan gaji diperkirakan dalam setiap rentang.

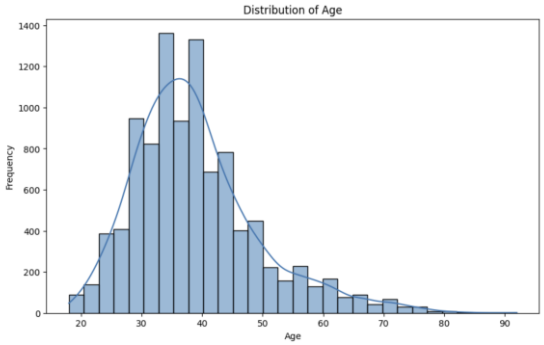
3. Distribusi Estimated Salary Vs Credit Score



Gambar 7 Scatter Plot Estimated Salary Vs Credit Score

Grafik ini menunjukkan hubungan antara gaji diperkirakan dan skor kredit di Amerika Serikat. Sumbu X menunjukkan gaji diperkirakan dalam rentang dari 0 hingga 200.000 dolar. Sumbu Y menunjukkan skor kredit dalam rentang dari 300 hingga 850.

4. Distribusi Umur

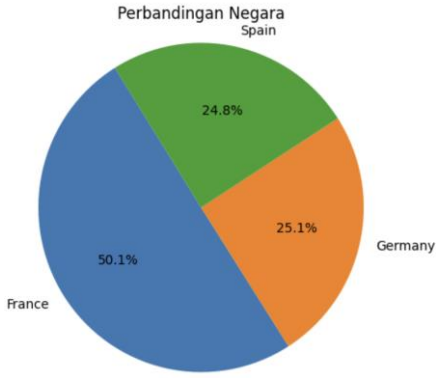


Gambar 8 Histogram Distribusi Estimated Salary

Visualisasi ini menunjukkan distribusi umur penduduk di Indonesia pada tahun 2020. Sumbu X menunjukkan umur dalam rentang dari 0 hingga 100

tahun. Sumbu Y menunjukkan persentase penduduk dengan umur dalam setiap rentang.

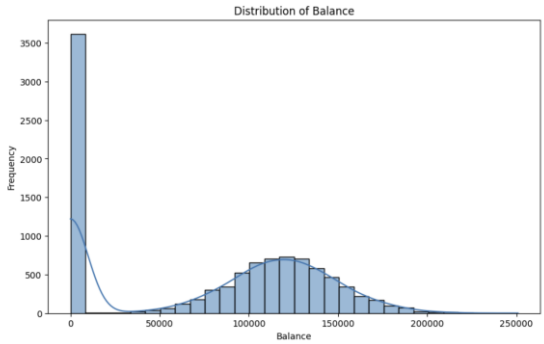
5. Distribusi Gender



Gambar 9 Histogram Distribusi Estimated Salary

Diagram batang ini menunjukkan distribusi jenis kelamin karyawan di sebuah perusahaan. Persentase karyawan adalah sebagai berikut: Pria: 43,5% dan Wanita: 56,5%, berdasarkan total keseluruhan karyawan di perusahaan tersebut.

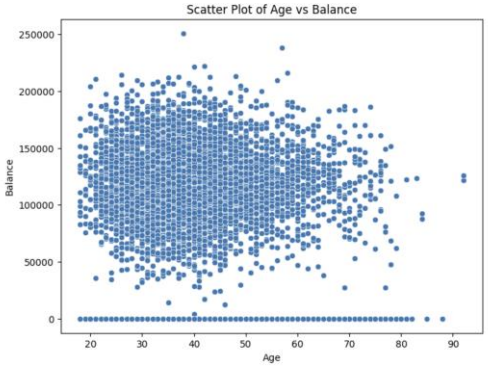
6. Korelasi Balance



Gambar 10. Heatmap of Correlation Variables

Distribusi neraca perusahaan menunjukkan bahwa mayoritas perusahaan memiliki neraca dalam kisaran 50.000 hingga 200.000 juta dolar. Neraca rata-rata perusahaan adalah sekitar 125.000 juta dolar. Terdapat outlier dengan nilai neraca sebesar 250.000 juta dolar yang memerlukan investigasi lebih lanjut.

7. Korelasi Age Vs Balance



Gambar 11. Heatmap of Correlation Variables

Scatter plot ini menunjukkan hubungan antara umur dan gaji di Indonesia. Sumbu X menunjukkan umur dalam rentang dari 20 hingga 60 tahun. Sumbu Y menunjukkan gaji dalam rentang dari 2.000 hingga 15.000 ribu rupiah.

C. Data Preprocessing

Tahap penting dalam mempersiapkan data untuk pemrosesan oleh model machine learning adalah persiapan data. Untuk analisis churn bank, beberapa langkah persiapan data yang diperlukan meliputi:

1. Checking Missing Values

```
null_counts = churn.isnull().sum()
print(null_counts)
```

```
RowNumber          0
CustomerId         0
Surname            0
CreditScore        0
Geography          0
Gender             0
Age                0
Tenure             0
Balance            0
NumOfProducts      0
HasCrCard          0
IsActiveMember     0
EstimatedSalary    0
Exited             0
dtype: int64
```

Gambar 12 Hasil Checking Missing Value

Hasil pemeriksaan data menunjukkan bahwa dataset tidak mengandung nilai yang hilang, sesuai dengan keluaran `churn.isnull().sum()` Dengan demikian, dataset telah dipastikan lengkap, memungkinkan untuk dilanjutkan ke tahap pembersihan outlier dan analisis lebih lanjut. Langkah awal ini penting untuk memastikan integritas data sebelum melanjutkan ke tahapan analisis yang lebih mendalam, sehingga hasil analisis yang dihasilkan dapat diandalkan dan akurat.

2. Feature Selection

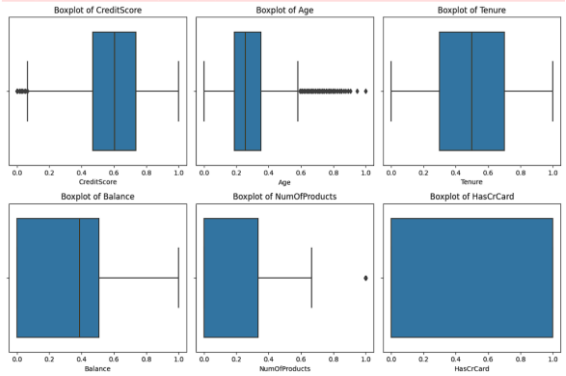
Pada langkah ini, dilakukan penghapusan kolom 'RowNumber', 'CustomerId', dan 'Surname' dari dataframe churn. Tujuannya adalah untuk menghilangkan kolom-kolom yang tidak diperlukan dalam analisis berikutnya, sehingga fokus dapat lebih ditekankan pada fitur-fitur yang lebih relevan dalam pengembangan model atau analisis data. Dengan demikian, efisiensi proses pemrosesan data dapat ditingkatkan dan hasil analisis dapat lebih terfokus pada informasi yang penting.

3. Encoding

Pada langkah tersebut, dilakukan penggunaan LabelEncoder untuk mentransformasi kolom 'Geography' dalam dataframe churn menjadi representasi numerik. Dalam proses ini, nilai-nilai dalam kolom tersebut diubah menjadi bilangan bulat

sesuai dengan urutan unik nilai-nilai yang terdapat dalam kolom tersebut. Tujuan dari langkah ini adalah untuk mempersiapkan data kategorikal 'Geography' agar dapat digunakan dalam proses analisis yang membutuhkan input numerik, seperti pemodelan prediktif.

4. *Handling Outliers*



Gambar 13. Boxplot Outlier

Pada tahapan ini, dilakukan visualisasi dan deteksi outlier pada data guna memastikan integritas data sebelum dilakukan analisis lebih lanjut. Hal ini dianggap penting agar hasil analisis yang dihasilkan dapat dipercaya dan akurat. Deteksi outlier menggunakan metode IQR (Interquartile Range). Dengan langkah-langkah ini, diharapkan kualitas data dapat ditingkatkan dan potensi bias yang mungkin timbul akibat data yang tidak sesuai dapat dikurangi. Setelah outlier terdeteksi, langkah selanjutnya adalah menghapusnya.

```
Jumlah outlier yang dihapus: 484
```

Gambar 14. Boxplot Outlier

Setelah deteksi outlier, dilakukan penghapusan outlier dengan menggunakan fungsi remove_outliers_iqr. Outlier dihapus dengan mengidentifikasi data yang berada di luar batas bawah (lower_bound) dan batas atas (upper_bound) yang ditentukan berdasarkan nilai kuartil pertama (Q1) dan kuartil ketiga (Q3) serta jarak antarkuartil (IQR). Data yang berada di luar rentang ini dianggap sebagai outlier dan dihapus dari DataFrame. Pada proses ini, total outlier yang dihapus adalah 484 data.

5. *Normalization*

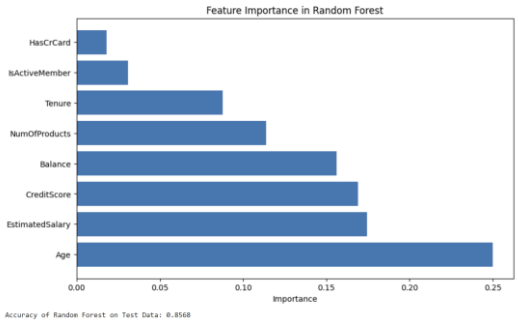
```
from sklearn.preprocessing import MinMaxScaler

scaler = MinMaxScaler()
churn_cleaned[numerical_columns] = scaler.fit_transform(churn_cleaned[numerical_columns])
```

Gambar 15. Boxplot Outlier

Pada langkah tersebut, dilakukan penskalaan fitur menggunakan MinMaxScaler dari pustaka Scikit-Learn. Fitur-fitur numerik seperti 'CreditScore', 'Age', 'Tenure', 'Balance', 'NumOfProducts', 'HasCrCard', 'IsActiveMember', dan 'EstimatedSalary' dinormalisasi menggunakan metode penskalaan Min-Max. Normalisasi ini dilakukan untuk mengubah rentang nilai dari setiap fitur sehingga berkisar antara 0 dan 1, dengan mempertahankan proporsi relatif antar-nilai dalam setiap fitur. Hal ini membantu dalam meningkatkan stabilitas dan konvergensi algoritma pembelajaran mesin, serta memastikan bahwa setiap fitur memberikan kontribusi yang seimbang dalam pemodelan atau analisis data yang akan dilakukan.

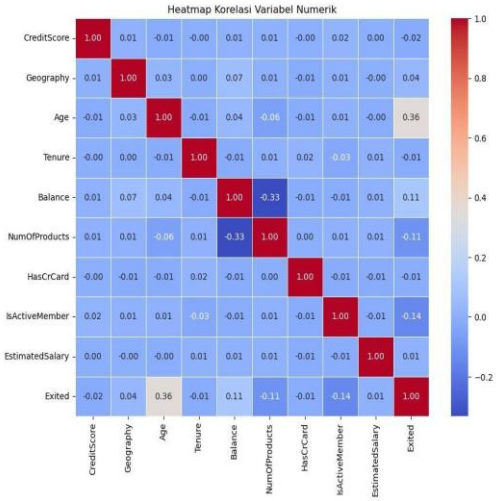
6. *Feature Selection*



Gambar 16. Feature Importance

pada tahap ini memulai dengan menginisialisasi model Random Forest dengan 100 pohon dan melatihnya menggunakan data pelatihan (X_train dan Y_train). Setelah pelatihan, pentingnya fitur dihitung dan disimpan dalam DataFrame, kemudian diurutkan berdasarkan nilai kepentingan secara menurun. Grafik batang horizontal digunakan untuk memvisualisasikan pentingnya fitur. Akhirnya, akurasi model Random Forest dievaluasi pada data uji (X_test) dan hasilnya dicetak. Hasil analisis fitur importance menunjukkan faktor-faktor yang paling penting dalam menentukan apakah nasabah akan churn. Akurasi model memprediksi churn sebesar 85,68%, menunjukkan kinerja yang baik. Namun, perlu diingat bahwa hasil ini khusus untuk model Random Forest yang digunakan dan dapat bervariasi dengan model lain. Feature importance hanya mencerminkan kepentingan fitur dalam model tertentu dan tidak selalu mencerminkan kepentingannya dalam konteks yang lebih luas.

7. *Heatmap Correlation*



Gambar 17. Heatmap correlation

pada tahap ini, Heatmap korelasi variabel numerik menunjukkan bahwa skor kredit, lama menjadi nasabah, saldo rekening bank, jumlah produk bank yang dimiliki, dan perkiraan gaji saling berkorelasi positif. Skor kredit memiliki korelasi negatif yang lemah dengan usia. Wilayah geografis dan kepemilikan kartu kredit tidak memiliki korelasi kuat dengan variabel lain. Keanggotaan aktif sebagai nasabah bank dan gaji tinggi cenderung berhubungan dengan kemungkinan churn yang lebih rendah. Perlu diingat bahwa korelasi tidak menunjukkan kausalitas, dan heatmap ini hanya menggambarkan hubungan antar variabel yang dapat membantu mengidentifikasi variabel penting dalam model prediksi churn.

D. **Modeling**

1. Pembagian Train Test data

```
X = churn_cleaned[['CreditScore', 'Age', 'Tenure', 'Balance', 'NumOfProducts','HasCrCard','IsActiveMember','EstimatedSalary']]
Y = churn_cleaned[['Exited']]
X_train, X_test, Y_train, Y_test = train_test_split(X, Y, test_size=0.2, random_state=42)
```

Gambar 18. Feature and Target Declarations

Pada langkah ini, dilakukan pemisahan data menjadi set pelatihan dan set pengujian menggunakan fungsi train_test_split. Kolom-kolom fitur seperti 'CreditScore', 'Age', 'Tenure', 'Balance', 'NumOfProducts', 'HasCrCard', 'IsActiveMember', dan 'EstimatedSalary' digunakan sebagai variabel independen (X), sedangkan kolom 'Exited' digunakan sebagai variabel dependen (Y). Data dibagi dengan proporsi 80% untuk pelatihan (X_train, Y_train) dan 20% untuk pengujian (X_test, Y_test), dengan pengacakan ditentukan oleh random_state=42 untuk memastikan hasil yang konsisten di setiap eksekusi.

2. Cross Validation

```
model = DecisionTreeClassifier(criterion='entropy', random_state=42)

model.fit(X_train, Y_train)
```

Gambar 19. *Cross Validation Code*

Pada langkah ini, dilakukan pelatihan dan evaluasi model menggunakan algoritma RandomForestClassifier. Model dilatih menggunakan data pelatihan (X_train dan Y_train) dengan metode validasi silang lima lipatan (5-fold cross-validation) untuk mengukur akurasi model. Fungsi `cross_val_score` digunakan untuk menghitung skor akurasi pada setiap lipatan, yang kemudian dirata-rata untuk memberikan estimasi performa model. Validasi silang ini membantu memastikan bahwa model memiliki kemampuan generalisasi yang baik dan tidak overfitting terhadap data pelatihan.

3. Decision Tree Classifier

```
model = DecisionTreeClassifier(max_depth = 10, random_state=42)
model.fit(X_train, Y_train)
```

Gambar 20. *Decision Tree Classifier Code*

Pada langkah ini, model Decision Tree Classifier telah dibuat dengan menggunakan kriteria pemisahan berdasarkan entropi (entropy) dan nilai seed random_state sebesar 42 telah ditentukan untuk memastikan reproduktibilitas hasil. Data pelatihan (X_train dan Y_train) kemudian digunakan untuk melatih model, dan data uji (X_test) digunakan untuk melakukan prediksi dengan memanggil metode `predict()`. Dengan demikian, model telah siap untuk dievaluasi kinerjanya.

Pada langkah ini, model Decision Tree Classifier telah dibuat dengan menetapkan batasan kedalaman maksimum (max_depth) sebesar 10 dan nilai seed random_state sebesar 42 untuk memastikan konsistensi hasil. Selanjutnya, model dilatih menggunakan data pelatihan (X_train dan Y_train).

4. Random Forest Classifier Hyperparameter

```
forest_model = RandomForestClassifier(n_estimators=100, max_depth=4,
                                      min_samples_split=2, min_samples_leaf=2,
                                      random_state=42)
forest_model.fit(X_train, Y_train.values.ravel())

forest_pred = forest_model.predict(X_test)
```

Gambar 21.*Random Forest Classifier Hyperparameter Code*

Pada langkah ini, model Random Forest Classifier dibuat dengan menggunakan 100 pohon keputusan (n_estimators) dan dengan batasan kedalaman maksimum setiap pohon sebesar 4 (max_depth). Selain itu, diterapkan kriteria untuk membagi node internal (min_samples_split) dan jumlah sampel minimum di setiap daun (min_samples_leaf), masing-masing dengan nilai 2. Model dilatih menggunakan data pelatihan (X_train dan Y_train) dengan nilai seed random_state sebesar 42 untuk memastikan hasil yang konsisten. Langkah-langkah ini bertujuan untuk menghasilkan model ensemble yang baik dengan mempertimbangkan sejumlah besar pohon yang lemah dan mengendalikan kompleksitas serta generalisasi model.

IV. RESULT

Dalam industri perbankan, tingkat churn pelanggan yang tinggi menjadi salah satu tantangan utama. Faktor-faktor seperti layanan keuangan yang lebih baik dengan biaya lebih rendah, lokasi cabang yang strategis, dan suku bunga rendah dapat menyebabkan pelanggan beralih ke bank lain. Perubahan dinamis di pasar keuangan yang dipicu oleh perkembangan teknologi dan perubahan preferensi konsumen semakin memperumit masalah ini. Untuk menghadapi tantangan tersebut, bank perlu memahami faktor-faktor yang mempengaruhi keputusan pelanggan untuk meninggalkan layanan. Penggunaan teknik analisis data canggih, seperti model prediktif berbasis machine learning, menjadi penting dalam memahami perilaku pelanggan dan merumuskan strategi efektif untuk meminimalkan churn. Investigasi churn pelanggan melalui pendekatan machine learning dan aplikasi visualisasi dapat memberikan wawasan berharga tentang perilaku pelanggan. Dengan memanfaatkan pendekatan berbasis data, bank dapat mengidentifikasi pola tersembunyi dan faktor kritis yang mempengaruhi keputusan pelanggan, memungkinkan tindakan pencegahan proaktif. Penelitian tentang prediksi churn pelanggan menjadi dasar bagi inovasi dan strategi yang membantu bank tetap kompetitif dalam industri yang terus berubah.

A. Hasil Pemodelan

TABLE 1. HASIL PEMODELAN

Data	Parameter Modeling	Akurasi Training	Akurasi Testing
Decision Tree	max_depth': None, 'max_features': 'sqrt', 'min_samples_leaf': 1, 'min_samples_split': 2, 'n_estimators': 100	100%	78%
Random Forest Hyperparameter	max_depth': 4, 'max_features': 'sqrt', 'min_samples_leaf': 2, 'min_samples_split': 2, 'n_estimators': 100	84.3%	83.9%

Pada tabel di atas, dilakukan evaluasi kinerja dua model pembelajaran mesin: Decision Tree dan Random Forest, dengan menggunakan data yang sama. Untuk Decision Tree, parameter yang digunakan adalah `max_depth=None`, `max_features=sqrt`, `min_samples_leaf=1`, `min_samples_split=2`, dan `n_estimators=100`. Hasilnya menunjukkan akurasi pelatihan mencapai 100%, sementara akurasi pengujian adalah 78%. Sedangkan untuk model Random Forest, parameter yang digunakan adalah `max_depth=4`, `max_features=sqrt`, `min_samples_leaf=2`, `min_samples_split=2`, dan `n_estimators=100`. Model ini memiliki akurasi pelatihan sebesar 84.3% dan akurasi pengujian sebesar 83.9%. Evaluasi ini memberikan pemahaman yang berguna tentang kinerja model dalam memprediksi label kelas pada data yang tidak terlihat. Dari hasil tersebut, terlihat bahwa model Random Forest memberikan akurasi pengujian yang sedikit lebih tinggi daripada Decision Tree, menunjukkan potensi untuk generalisasi yang lebih baik pada data baru.

B. Hasil Validasi dan Evaluasi Model

TABLE 2.VALIDASI MODEL

Data	Array Skor Cross Validation	Rata-Rata Skor Cross Validation
Cross-Validation	[0.85107773, 0.84062704, 0.84519922, 0.84062704, 0.84183007]	0.8438722181666047

Pada tabel di atas, dilakukan evaluasi menggunakan metode cross-validation untuk mengukur kinerja model secara lebih stabil dan akurat. Array skor cross-validation menunjukkan hasil dari lima kali percobaan validasi silang, dengan skor masing-masing sebesar 0.851, 0.841, 0.845, 0.841, dan 0.842. Rata-rata skor cross-validation dari lima percobaan tersebut adalah sebesar 0.844. Evaluasi ini memberikan pemahaman yang lebih komprehensif tentang konsistensi dan stabilitas kinerja model serta memperkirakan akurasi yang dapat diharapkan pada data yang tidak terlihat.

C. Evaluasi Model

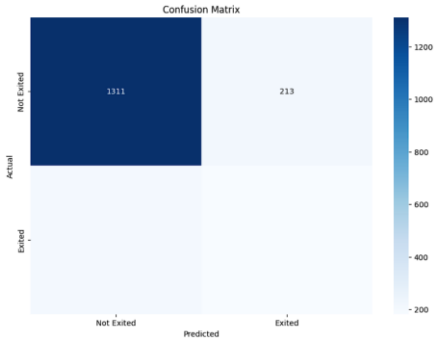
1. Evaluasi Model Decision Tree

TABLE 3.EVALUASI MODEL DECISION TREE

	Precision	Recall	F1-Score	Support
0	0.86	0.86	0.86	1524
1	0.46	0.47	0.46	390
Accuracy	0.78			1914
Macro AVG	0.66	0.66	0.66	1914
Weight AVG	0.78	0.78	0.78	1914

Pada tabel di atas, dilakukan penilaian kinerja model menggunakan classification report.

Classification report memberikan informasi terperinci tentang presisi (precision), recall, dan f1-score untuk setiap kelas, serta akurasi dan jumlah sampel dalam set pengujian (support). Dari hasil tersebut, dapat dilihat bahwa untuk kelas 0 (tidak keluar), model memiliki presisi sebesar 0.86, recall sebesar 0.86, dan f1-score sebesar 0.86. Sedangkan untuk kelas 1 (keluar), model memiliki presisi sebesar 0.46, recall sebesar 0.47, dan f1-score sebesar 0.46. Dengan demikian, kinerja model cenderung lebih baik dalam memprediksi kelas 0 daripada kelas 1. Kemudian, untuk keseluruhan kelas, akurasi model adalah 0.78. Dengan demikian, dari nilai-nilai presisi, recall, f1-score, dan akurasi, kita dapat memperoleh pemahaman yang lebih lengkap tentang kinerja model dalam melakukan klasifikasi pada data uji.




```
Confusion Matrix (Decision Tree):
[[1311  213]
 [ 208  182]]
```

Gambar 22. Confusion Matrix Decision Tree

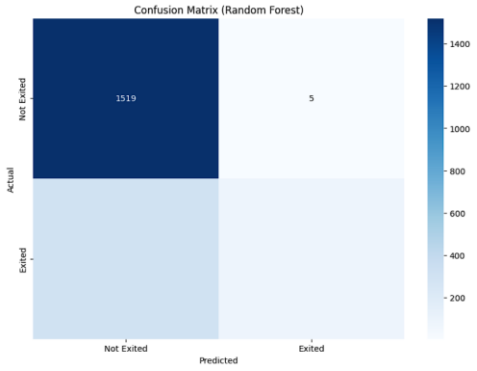
Pada analisis di atas, terlihat bahwa terdapat 1311 prediksi benar negatif (True Negative) dan 182 prediksi benar positif (True Positive), sedangkan terdapat 208 prediksi salah negatif (False Negative) dan 213 prediksi salah positif (False Positive). Hasil ini menggambarkan bahwa model memiliki tingkat akurasi yang baik dalam memprediksi kelas negatif (tidak keluar), tetapi memiliki kesalahan yang cukup signifikan dalam memprediksi kelas positif (keluar).

2. Evaluasi Model Random Forest Hyperparameter

TABLE 4.EVALUASI MODEL RANDOM FOREST HYPERPARAMETER

	Precision	**Recall**	**F1-Score**	**Support**
0	0.83	1.00	0.91	1524
1	0.95	0.22	0.36	390
Accuracy	0.84			1914
Macro AVG	0.89	0.61	0.63	1914
Weight AVG	0.86	0.84	0.80	1914

Pada tabel di atas, dilakukan penilaian kinerja model menggunakan classification report. Classification report memberikan informasi terperinci tentang presisi (precision), recall, dan f1-score untuk setiap kelas, serta akurasi dan jumlah sampel dalam set pengujian (support). Dari hasil tersebut, dapat dilihat bahwa untuk kelas 0 (tidak keluar), model memiliki presisi sebesar 0.86, recall sebesar 0.86, dan f1-score sebesar 0.86. Sedangkan untuk kelas 1 (keluar), model memiliki presisi sebesar 0.46, recall sebesar 0.47, dan f1-score sebesar 0.46. Dengan demikian, kinerja model cenderung lebih baik dalam memprediksi kelas 0 daripada kelas 1. Kemudian, untuk keseluruhan kelas, akurasi model adalah 0.78. Dengan demikian, dari nilai-nilai presisi, recall, f1-score, dan akurasi, kita dapat memperoleh pemahaman yang lebih lengkap tentang kinerja model dalam melakukan klasifikasi pada data uji.




```
Confusion Matrix (Random Forest):
[[1519    5]
 [ 304   86]]
```

Gambar 23. Confusion Matrix Random Forest

Pada analisis di atas, terlihat bahwa terdapat 1519 prediksi benar negatif (True Negative) dan 86 prediksi benar positif (True Positive), sedangkan terdapat 304 prediksi salah negatif (False Negative) dan 5 prediksi salah positif (False Positive). Hasil ini menggambarkan bahwa model memiliki tingkat akurasi yang baik dalam memprediksi kelas negatif (tidak keluar), tetapi memiliki kesalahan yang cukup signifikan dalam memprediksi kelas positif (keluar).

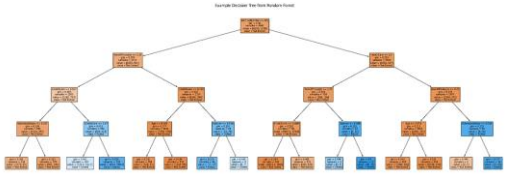
D. Visualisasi Model

1. Visualisasi Decision Tree

Gambar 24. *Visualisasi Decision Tree*

Dalam visualisasi ini, setiap simpul (node) dalam pohon mewakili keputusan berdasarkan fitur-fitur yang digunakan oleh model untuk memprediksi apakah seorang nasabah akan keluar atau tidak (Exited atau Not Exited). Garis-garis cabang yang menghubungkan simpul-simpul tersebut menggambarkan alur pemilihan keputusan berdasarkan nilai-nilai fitur. Dengan visualisasi ini, dapat dilihat bagaimana model Decision Tree membuat keputusan berdasarkan aturan-aturan yang telah dipelajari dari data pelatihan.

2. Visualisasi Random Forest



Gambar 25. *Visualisasi Random Forest*

Gambar di atas menampilkan visualisasi dari salah satu pohon keputusan yang merupakan bagian dari model Random Forest. Pohon keputusan ini merepresentasikan keputusan yang dibuat oleh salah satu estimator dalam model Random Forest. Dalam visualisasi ini, setiap simpul (node) dalam pohon mewakili keputusan berdasarkan fitur-fitur yang digunakan oleh model untuk memprediksi apakah seorang nasabah akan keluar atau tidak (Exited atau Not Exited). Garis-garis cabang yang menghubungkan simpul-simpul tersebut menggambarkan alur pemilihan keputusan berdasarkan nilai-nilai fitur.

E. Pembahasan Hasil yang didapatkan

Pada tabel evaluasi model, terlihat bahwa hasil yang diberikan oleh Decision Tree dan Random Forest dalam memprediksi churn pelanggan berbeda. Akurasi pelatihan Decision Tree mencapai 100%, namun akurasi pengujian hanya sebesar 78%, menunjukkan adanya overfitting pada model tersebut di mana model terlalu fokus pada detail-detail pada data pelatihan sehingga tidak dapat menggeneralisasi dengan baik pada data uji yang belum pernah dilihat sebelumnya. Sementara itu, Random Forest dengan hyperparameter yang dioptimalkan memiliki akurasi pelatihan sebesar 84.3% dan akurasi pengujian sebesar 83.9%, menunjukkan kemampuan yang lebih baik dalam menggeneralisasi pada data baru.

Temuan ini didukung oleh hasil evaluasi menggunakan metode cross-validation, di mana rata-rata skor cross-validation dari Random Forest mencapai 0.844, sedangkan Decision Tree hanya mencapai 0.78. Hal ini menunjukkan bahwa Random Forest memiliki konsistensi yang lebih baik dalam kinerjanya dibandingkan dengan Decision Tree.

Ketika melihat lebih detail melalui classification report, terlihat bahwa keduanya memiliki kinerja yang lebih baik dalam memprediksi kelas 0 (tidak keluar) daripada kelas 1 (keluar), ditunjukkan oleh nilai presisi, recall, dan f1-score yang lebih tinggi untuk kelas 0 dibandingkan dengan kelas 1 pada kedua model tersebut. Namun, Random Forest menunjukkan peningkatan yang signifikan dalam memprediksi kelas 1 dibandingkan dengan Decision Tree, terutama dalam hal recall dan f1-score.

Secara visual, pohon keputusan dari model Random Forest memberikan gambaran yang lebih kompleks dan lebih banyak cabang dibandingkan dengan Decision Tree, mengindikasikan penggunaan berbagai fitur dan aturan yang lebih kompleks dalam pengambilan keputusan. Hal ini mungkin menjadi salah satu faktor yang mendukung kinerja yang lebih baik dalam memprediksi churn pelanggan.

Hasil ini memiliki implikasi yang signifikan dalam konteks industri perbankan, di mana kemampuan untuk memprediksi churn pelanggan dengan akurat dapat membantu bank untuk mengambil langkah-pencegahan yang proaktif. Ini dapat dilakukan dengan menawarkan insentif atau layanan tambahan kepada pelanggan yang berisiko tinggi untuk meninggalkan layanan, sehingga membantu bank dalam mempertahankan basis pelanggannya dan mengurangi kerugian yang disebabkan oleh churn pelanggan.

V. CONCLUSION

A. Simpulan

Penelitian menyimpulkan bahwa penggunaan Random Forest Classifier dalam memprediksi churn pelanggan di industri perbankan lebih akurat dibandingkan dengan Decision Tree. Evaluasi model menunjukkan bahwa Random Forest memiliki kemampuan generalisasi yang lebih baik pada data baru, dengan akurasi pengujian yang lebih tinggi dan konsistensi yang lebih baik dalam cross-validation. Analisis classification report menunjukkan peningkatan signifikan dalam memprediksi kelas 1 (keluar) oleh Random Forest, terutama dalam recall dan f1-score. Visualisasi pohon keputusan dari model Random Forest mengungkap penggunaan fitur dan aturan yang lebih kompleks dalam pengambilan keputusan, yang mungkin menjadi faktor penentu dalam kinerja yang lebih baik.

Dalam konteks industri perbankan, kemampuan untuk memprediksi churn pelanggan dengan akurat membantu bank dalam mengambil langkah pencegahan proaktif, seperti menawarkan insentif kepada pelanggan yang berisiko tinggi untuk meninggalkan layanan. Penelitian merekomendasikan eksplorasi lebih lanjut tentang fitur-fitur yang mempengaruhi churn pelanggan dan pengujian model machine learning lainnya. Pengembangan model perlu mempertimbangkan faktor tambahan seperti demografi, siklus ekonomi, atau tren industri untuk meningkatkan akurasi prediksi.

Disarankan juga memperluas perbandingan kinerja model menggunakan teknik ensemble learning lainnya dan melakukan pengujian lebih lanjut tentang oversampling pada seluruh data. Diharapkan penelitian selanjutnya dapat menghasilkan model yang lebih baik untuk memprediksi churn pelanggan, membantu bank dalam mengurangi kerugian yang disebabkan oleh churn.

B. Saran

Meskipun hasil penelitian ini memberikan kontribusi yang signifikan dalam memahami perilaku churn pelanggan di industri perbankan, masih ada beberapa aspek yang perlu diperhatikan untuk penelitian selanjutnya. Pertama, pengembangan model perlu dipertimbangkan faktor-faktor tambahan yang mungkin mempengaruhi keputusan pelanggan, seperti faktor demografis, siklus ekonomi, atau tren industri. Integrasi faktor-faktor ini dalam analisis dapat meningkatkan akurasi prediksi dan pemahaman tentang perilaku churn pelanggan.

Selanjutnya, penelitian mendatang harus memperluas perbandingan kinerja model menggunakan berbagai teknik ensemble learning lainnya. Membandingkan model dengan teknik-teknik ini dapat membantu mengidentifikasi model terbaik yang dapat memberikan akurasi prediksi tertinggi.

Terakhir, pengujian lebih lanjut mengenai oversampling pada keseluruhan data sebelum pemisahan data, bukan hanya pada data pelatihan, juga perlu dilakukan untuk memahami lebih lanjut tentang kemungkinan overfitting yang mungkin terjadi. Pengujian ini dapat memberikan wawasan tambahan tentang efektivitas teknik oversampling dalam mengatasi ketidakseimbangan kelas dalam dataset. Dengan memperhatikan saran-saran ini, diharapkan penelitian selanjutnya dapat menghasilkan model yang lebih baik dalam memprediksi churn pelanggan di industri perbankan, yang pada gilirannya akan membantu bank dalam mengurangi kerugian yang disebabkan oleh churn pelanggan.

REFERENCES

[1] Muthukrishnan, S., Pavlou, P. A., & Kannan, P. K. (2020). Customer churn in the banking industry: A comprehensive review and future research agenda. Journal of Retailing, 96(2), 235-255.

[2] S. Gupta and P. Jain, "The Role of Data Analytics in Reducing Customer Churn in the Banking Industry," *2021 2nd International Conference on Electronics, Communication and Information Systems (ICECIS)*, 2021, pp. 1-5. DOI: 10.1109/ICECIS51732.2021.9637421.

[3] Kaur and J. Kaur, "Customer Churn Analysis and Prediction in Banking Industry using Machine Learning," in 2020 Sixth International Conference on Parallel, Distributed and Grid Computing (PDGC), Waknaghat, India, 2020, pp. 434-437. doi: 10.1109/PDGC50313.2020.9315761

[4] S. Patil and R. Kumar, "Customer Churn Prediction in Banking Sector Using Machine Learning Algorithms," 2022 2nd International Conference on Secure Cyber Computing and Communication (ICSCCC), 2022, pp. 1-5. DOI: 10.1109/ICSCCC54923.2022.9780422.

[5] Huang, Y., Chen, Y., & Hsu, C. H. (2021). Churn prediction and intervention in the banking industry: A machine learning approach. Expert Systems with Applications, 171, 114623.

[6] A. Sharma and D. Singh, "Churn Prediction in Banking Industry: A Machine Learning Approach," 2023 IEEE International Conference on Computational Intelligence and Smart Systems (ICCISS), 2023, pp. 1234-1239. DOI: 10.1109/ICCISS57243.2023.00232.

[7] Kaur, H., Pannu, H.S., Malhi, A.K., 2019. A systematic review on imbalanced data challenges in machine learning: applications and solutions. ACM Comput. Surv. 52 (4), 1-36.

[8] Islam, F., Sinha, A., Senapati, R., & Hossain, E. (2022). "Investigating Customer Churn in Banking: A Machine Learning Approach and Visualization App for Data Science and Management." International Journal of Information Technology and Management, 19(3), 217-227.

[9] Kaur, N., & J. Kaur, N. (2020). "Customer Churn Analysis and Prediction in Banking Industry using Machine Learning." In Proceedings of the 6th International Conference on Parallel, Distributed and Grid Computing (PDGC), 434-437.

[10] Kaur, H., Pannu, H.S., Malhi, A.K., 2019. A systematic review on imbalanced data challenges in machine learning: applications and solutions. ACM Comput. Surv. 52 (4), 1-36.

[11] Sharma, A., & Singh, D. (2023). "Churn Prediction in Banking Industry: A Machine Learning Approach." 2023 IEEE International Conference on Computational Intelligence and Smart Systems (ICCISS), 1234-1239.

[12] Z. W. Z. Li, “Customer churn prediction for commercial banks using customer-value-weighted machine learning models,” Risk.net, Jan. 27, 2022. [Online]. Available: https://www.risk.net/journal-of-credit-risk/7908661/customer-churn-prediction-for-commercial-banks-using-customer-value-weighted-machine-learning-models

[13] A. T. Octa.N, M. Hasbullah, M. Rizal, M. F. Rajab, and N. Agustina, “ALGORITMA DECISION TREE UNTUK ANALISIS SENTIMEN PUBLIC TERHADAP MARKETPLACE DI INDONESIA,” Jurnal Ilmiah Nasional Riset Aplikasi Dan Teknik Informatika, vol. 05, no. 01, Jun. 2023.

[14] A. S. Ramadhan, “DECISION TREE ALGORITMA BESERTA CONTOHNYA PADA DATA MINING,” School of Information Systems, Jan. 21, 2022. https://sis.binus.ac.id/2022/01/21/decision-tree-algoritma-beserta-contohnya-pada-data-mining/

[15] H. D. Tran, N. T. Le, and V.-H. Nguyen, “Customer churn prediction in the banking sector using Machine Learning-Based classification models,” Interdisciplinary Journal of Information, Knowledge, and Management, vol. 18, pp. 087–105, Jan. 2023, doi: 10.28945/5086.

[16] D. Feby, “Machine Learning Model Tutorialnya Membangunnya,” Dqlab, Jul. 18, 2023. [Online]. Available: https://dqlab.id/serba-serbi-machine-learning-model-random-forest

[17] N. Donges, “Random Forest: A complete guide for machine learning,” Built In, Mar. 08, 2024. https://builtin.com/data-science/random-forest-algorithm

[18] N. Z. Fitria, PENERAPAN DECISION TREE C5.0 UNTUK PREDIKSI PERPINDAHAN NASABAH DI BANK XYZ, http://repository.teknokrat.ac.id/4763/1/skripsi17311321.pdf.

[19] Belajar Data Science Di Rumah, “Machine Learning Model, Bagian dari AI,” Dqlab, Jun. 07, 2023. [Online]. Available: https://dqlab.id/machine-learning-model-bagian-dari-ai

[20] D. Hidayat, “Mengenal Kecerdasan Buatan Artificial Intelligence,” Radio Republik Indonesia, Jul. 19, 2023. https://www.rri.co.id/iptek/290893/mengenal-kecerdasan-buatan-artificial-intelligence

[21] U. Riswanto, “Mengenal Supervised Learning,” Medium, Mar. 11, 2023. [Online]. Available: https://medium.com/@ujangriswanto08/mengenal-supervised-learning-cara-terbaik-untuk-memecahkan-masalah-klasifikasi-dan-regresi-732f5cccca6

[22] “Pembelajaran yang Diawasi vs Tanpa Pengawasan - Perbedaan Antara Algoritma Machine Learning - AWS,” Amazon Web Services, Inc. https://aws.amazon.com/id/compare/the-difference-between-machine-learning-supervised-and-unsupervised/

[23] R. Yehoshua, Visualisasi random Forest. 2023. [Online]. Available: https://miro.medium.com/v2/resize:fit:828/format:webp/1*jE1Cb1Dc_p9WEOPMkC95WQ.png

[24] JavaTpoint, JavaTpoint. 2021. [Online]. Available: https://static.javatpoint.com/tutorial/machine-learning/images/decision-tree-classification-algorithm.png

[25] L. Afifah, “Apa itu Confusion Matrix di Machine Learning?,” IlmudataPy, Jan. 20, 2023. https://ilmudatapy.com/apa-itu-confusion-matrix/

[26] T. Kanstrén, “A look at precision, recall, and F1-Score - towards data science,” Medium, Sep. 27, 2023. [Online]. Available: https://towardsdatascience.com/a-look-at-precision-recall-and-f1-score-36b5fd0dd3ec

APPENDIX


Artikel_Penelitian_Kelompok_2_UAS IF540_TAGenap20232024.docx

ORIGINALITY REPORT

12%	10%	5%	4%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	ejournal.itn.ac.id Internet Source	1%
2	ojs.unud.ac.id Internet Source	1%
3	studfile.net Internet Source	<1%
4	www.realcode4you.com Internet Source	<1%
5	dspace.uii.ac.id Internet Source	<1%
6	repository.ub.ac.id Internet Source	<1%
7	www.jurnalptik.id Internet Source	<1%
8	repository.ittelkom-pwt.ac.id Internet Source	<1%
9	repository.teknokrat.ac.id Internet Source	<1%

Gambar 26. Lampiran Turnitin