PREDIKSI *CHURN* NASABAH BANK MENGGUNAKAN KLASIFIKASI *RANDOM FOREST* DAN *DECISION TREE* DENGAN EVALUASI *CONFUSION MATRIX*

Azel Fabian Azmi[1], Apriade Voutama[2]

[1,2] Sistem Informasi, Universitas Singaperbangsa Karawang
Jl. HS.Ronggo Waluyo, Puseurjaya, Telukjambe Timur, Karawang
E-mail : azelfa65@gmail.com[1]

Abstrak

Dalam lanskap perbankan yang terus berkembang, customer churn menjadi tantangan yang signifikan, yang berdampak pada pendapatan dan reputasi. Penelitian ini mengeksplorasi efektivitas pengklasifikasi *Random Forest* dan *Decision Tree* dalam memprediksi *churn* bank. Memanfaatkan dataset yang bersumber dari *Kaggle*, yang berisi informasi mengenai 10.000 nasabah bank, penelitian ini menggunakan teknik *preprocessing* dan pemilihan fitur untuk menyaring data. Selanjutnya, dataset tersebut dibagi menjadi set pelatihan dan pengujian untuk evaluasi model. Model *Random Forest* dan *Decision Tree* dilatih dan dievaluasi menggunakan analisis *Confusion matrix*. Hasilnya menunjukkan bahwa *Random Forest* mengungguli *Decision Tree*, mencapai rata-rata *precision* yang lebih tinggi (79% vs 72%), *recall* (78% vs 72%), *F1-score* (78% vs 72%), dan *accuracy* (78% vs 72%). Penelitian ini menyoroti keampuhan *Random Forest* dalam memprediksi *churn* nasabah secara akurat, sehingga memberikan wawasan yang berharga bagi bank dalam menerapkan strategi retensi nasabah yang efektif. Penelitian ini berkontribusi dalam memajukan analisis prediktif di sektor perbankan, memberdayakan institusi untuk mengurangi churn dan meningkatkan kepuasan pelanggan.
Kata kunci : *Churn, Random Forest, Decision Tree, Confusion Matrix,* Klasifikasi

Abtract

In the ever-evolving banking landscape, customer churn is a significant challenge, impacting revenue and reputation. This research explores the effectiveness of Random Forest and Decision Tree classifiers in predicting bank churn. Utilizing a dataset sourced from Kaggle, which contains information on 10,000 bank customers, this research uses preprocessing and feature selection techniques to filter the data. Next, the dataset was divided into training and testing sets for model evaluation. Random Forest and Decision Tree models were trained and evaluated using confusion matrix analysis. The results showed that Random Forest outperformed Decision Tree, achieving higher average precision (79% vs 72%), recall (78% vs 72%), F1-score (78% vs 72%), and accuracy (78% vs 72%). This research highlights the efficacy of Random Forest in accurately predicting customer churn, thus providing valuable insights for banks in implementing effective customer retention strategies. This research contributes to advancing predictive analytics in the banking sector, empowering institutions to reduce churn and improve customer satisfaction.
Keywords : *Churn, Random Forest, Decision Tree, Confusion Matrix, Classifications*

1. PENDAHULUAN

Seiring berjalannya waktu, kebutuhan hidup akan terus meningkat. Selain itu, rasa ingin memiliki membuat orang ingin membeli sesuatu. Bank menawarkan fasilitas untuk mempermudah proses tersebut. Banyak bank bekerja keras untuk memenuhi kebutuhan nasabah yang cerdas, sadar harga, dan memiliki informasi produk lain yang mudah diakses melalui internet [1]. Salah satu ukuran umum dari kehilangan nasabah adalah perpindahan nasabah. Tingkat retensi pelanggan berdampak kuat pada nilai seumur hidup nasabah, dan mengetahui nilai sebenarnya dari kemungkinan churn nasabah akan membantu bank dalam mengelola hubungan nasabah [2].

Bank churn atau kegiatan pelanggan yang berpindah ke bank lain merupakan salah satu persoalan penting dalam industri perbankan. Churn bank dapat mempengaruhi pendapatan dan reputasi bank, karena pelanggan yang berpindah ke bank lain dapat menyebabkan biaya yang tinggi untuk mendapatkan pelanggan baru. Dalam industri perbankan, analisis perilaku pelanggan dan perkiraan perpindahan

pelanggan berdasarkan perilaku ini merupakan topik penelitian yang sangat penting. Hasil analisis perpindahan pelanggan memiliki dampak yang signifikan terhadap kebijakan bank karena memungkinkan bank untuk membuat strategi pelanggan baru atau meningkatkan strategi yang sudah ada [3].

Klasifikasi adalah proses menemukan model untuk digunakan dalam label yang belum diketahui kelasnya untuk memprediksi kelas. Ini dapat menjelaskan dan membedakan konsep atau kelas dalam data [4]. Klasifikasi dapat digunakan untuk berbagai tujuan, mulai dari penggunaan dasar hingga penggunaan yang lebih lanjut. Misalnya, klasifikasi dapat digunakan dalam manajemen data perbankan untuk menemukan pelanggan yang akan *churn*, atau dalam manajemen data medis untuk menemukan penyakit.

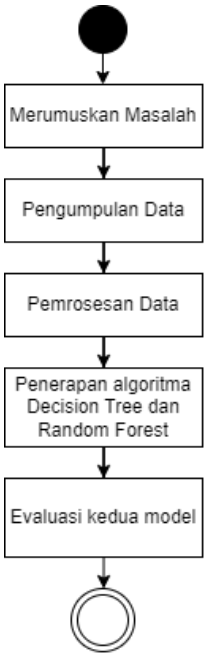
Klasifikasi *Random Forest* menggunakan pendekatan bagi-dan-taklukkan yang menggunakan metode subruang acak. Beberapa pohon dibuat, dan setiap pohon keputusan dilatih dengan memilih atribut dari kumpulan atribut prediktor secara acak. Setiap pohon menjadi matang berdasarkan karakteristik atau parameter yang tersedia, dan rata-rata tertimbang digunakan untuk membuat keputusan akhir. Khususnya, metode ini memiliki kemampuan untuk mengelola banyak parameter input tanpa menghapus apa pun, dan memiliki kemampuan untuk menangani nilai yang hilang dalam kumpulan data pelatihan untuk pelatihan model prediktif [5].

Metodologi *Decision Tree* adalah teknik data mining yang digunakan untuk membuat sistem klasifikasi berdasarkan faktor-faktor tertentu atau untuk membuat algoritme prediksi untuk variabel target. Metode ini mengkategorikan populasi ke dalam segmen seperti cabang yang membangun pohon terbalik dengan *node* akar, *node* internal, dan *node* daun. [6]. Dalam klasifikasi *Decision Tree*, setiap jalur mulai dari akar dijelaskan oleh data yang memisahkan urutan hingga hasil Boolean pada node daun tercapai. Ini menunjukkan hubungan pengetahuan hirarkis dengan node dan koneksi. Saat relasi digunakan untuk mengklasifikasikan, node berfungsi sebagai tujuan [7].

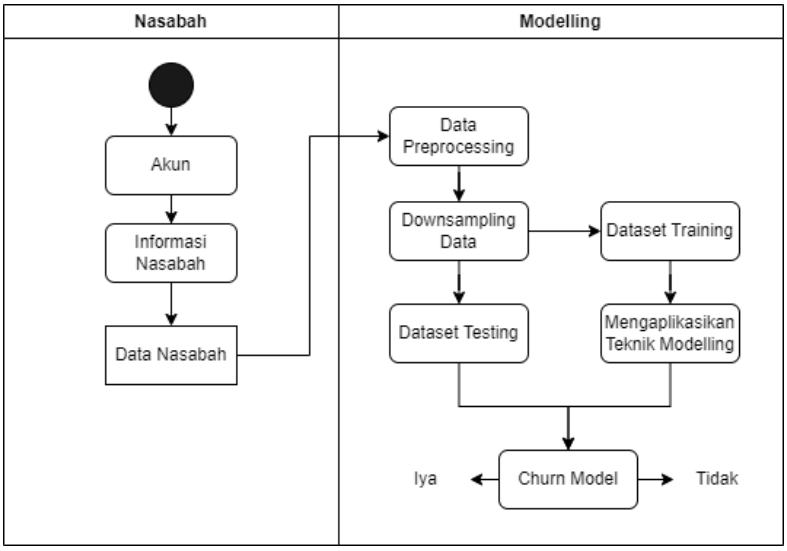
Penelitian ini akan menggunakan klasifikasi random forest dan decision tree untuk mengidentifikasi pelanggan yang akan churn dalam dataset churn bank yang didapatkan dari website Kaggle. Untuk mengevaluasi kinerja model yang diperoleh, hasil tersebut mencoba menemukan model yang dapat membantu bank menemukan pelanggan yang berpindah ke bank lain dan mengambil tindakan yang tepat untuk mempertahankan pelanggan tersebut. Kemudian menggunakan metode evaluasi yang tepat, menggunakan klasifikasi hutan random dan pohon keputusan.

2. METODOLOGI

Tahapannya terdiri dari *preprocessing*, *feature selection*, pembuatan model prediksi dengan algoritma *Random Forest* dan *Decision Tree*, pengujian, dan terakhir , membandingkan evaluasi dengan model satu sama lain.



Gambar 1. Diagram Flowchart Metodologi Penelitan



Gambar 2. *Activity Diagram* dari sistem yang diusulkan

2.1 Dataset

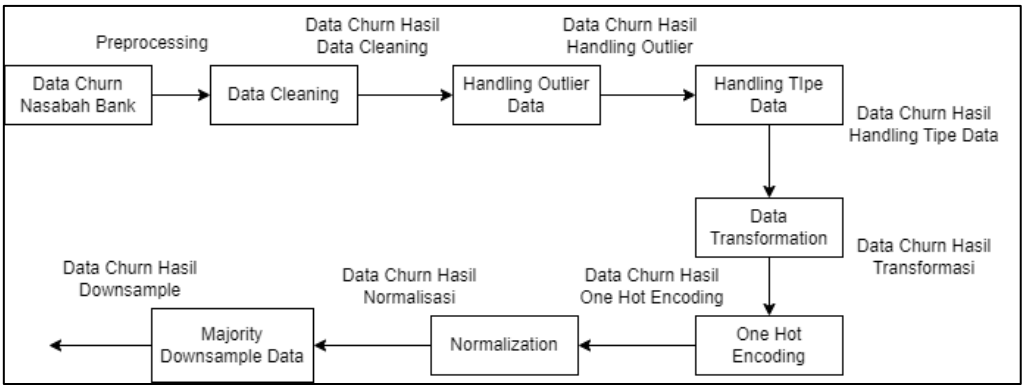
Pada penelitian ini, menggunakan data terkait data bank *churn* yang tersedia secara publik, *open-source* yang berasal dari website *Kaggle*. Dataset ini mencakup informasi dari 10.000 nasabah bank, dan parameter targetnya adalah variabel biner yang merepresentasikan apakah nasabah telah meninggalkan bank atau masih menjadi nasabah. Variabel variabel target mencerminkan bendera biner 1 ketika nasabah memiliki rekening bank ditutup, dan 0 ketika klien dipertahankan. Dataset berisi 13 vektor fitur (prediktor) yang dilaporkan dari data pelanggan dan transaksi yang diproses oleh nasabah. Rincian fitur-fitur ini diberikan pada Tabel 1.

Tabel 1. Deskripsi Dataset

Nama Fitur	Deskripsi Fitur
Row Number	Angka baris dari 1 sampai 10000
Customer ID	Angka unik untuk mengidentifikasi nasabah
Surname	Nama belakang nasabah
Credit Score	Kredit Skor pada nasabah
Geography	Negara asal nasabah
Gender	Jenis kelamin nasabah
Age	Umur nasabah
Tenure	Lamanya tahun nasabah bersama bank
Balance	Saldo nasabah
Num of Products	Jumlah produk bank yang digunakan nasabah (rekening tabungan, mobile banking, internet banking, dll.).
Has Cr Card	Angka biner untuk mengetahui apakah nasabah memiliki kartu kredit di bank atau tidak
Is Active Member	Angka biner untuk mengetahui apakah nasabah masih aktif di bank atau tidak
Estimated Salary	Perkiraan gaji nasabah dalam dolar
Exited	Angka biner 1 jika nasabah menutup rekening di bank dan 0 jika nasabah dipertahankan

2.2 Preprocessing

Preprocessing data adalah teknik paling awal sebelum melakukan *data mining* dan digunakan untuk menghilangkan masalah yang mungkin terjadi selama pemrosesan data karena format data yang tidak konsisten. Namun, proses *preprocessing* juga mencakup beberapa proses seperti membersihkan, mentransformasikan, dan mereduksi data. Tahapannya dapat dilihat pada gambar 3



Gambar 3. Proses *preprocessing*

2.2.1 Data Cleaning

Peneliti menemukan 3 data yang terduplikasi dan 6 data yang tidak lengkap. Peneliti melakukan operasi *data cleaning* untuk membersihkan data tersebut. Untuk data yang terduplikasi, dilakukannya penghapusan baris data yang terduplikat. Kemudian, data yang tidak lengkap juga dilakukannya penghapusan baris data yang tidak lengkap karena penghapusan data tersebut hanyalah sedikit sehingga tidak terlalu berpengaruh untuk hasil penerapan model nantinya.

2.2.2 Handling Outlier Data

Outlier data adalah data yang nilainya jauh dari nilai yang ditujukan. Data ini ditemukannya beberapa data outlier pada 2 kolom, yaitu *Gender* dan *Age*. Gender memiliki perbedaan nilai dimana mayoritas nilainya berupa "*Male*" dan "*Female*", dimana nilai yang berbeda, yaitu "M" dan "F". Dilakukannya penggantian semua nilai yang berbeda menjadi nilai yang sesungguhnya ("*Male*" dan "*Female*").

2.2.3 Handling Tipe Data

Tipe data dalam dataset dapat berupa nominal dan *string*, Ditemukannya dalam data ini yaitu pada kolom *Age* itu berupa *string*, dimana data tersebut bukanlah nominal sehingga tidak dapat digunakan dalam perhitungan. Dilakukannya penggatian tipe data terhadap kolom *Age* menjadi nominal.

2.2.4 Data Transformation

Data transformation adalah mengubah data dari satu bentuk ke bentuk lainnya [8].

1. *Drop Column*

   Menghapus kolom pada dataset yang tidak akan digunakan pada *modelling*. Kolom yang dihapus adalah *RowNumber, CustomerID,* dan *Surname*.
2. *One Hot Encoding*

   One hot encoding adalah teknik pengubahan kategori kategorikal menjadi format numerik yang dapat digunakan dalam algoritma *machine learning*. Proses ini berguna untuk mempermudah pengolahan dan analisis data dalam algoritma *machine learning*, serta memperbaiki kualitas data dan mempermudah kompatibilitas antara aplikasi, sistem, dan tipe data.
3. *Normalization*

   Proses normalisasi membantu pengolahan dan analisis data untuk algoritma pengajaran mesin. Ini juga meningkatkan kualitas data dan mempermudah kompatibilitas antara aplikasi, sistem, dan tipe data [10]. Metode skala normalisasi yang dipakai adalah *min-max scaling*. Metode ini berfungsi untuk menetapkan rantang batas minimum dan maksimum, biasanya dengan batas 0 dan 1, dengan rumus perhitungan sebagai berikut :

$$Xnorm = \frac{x^1 - \min(x)}{\max(x) - \min(x)} \left(new_{max}(x) - new_{min}(x)\right) + new_{\min}(x) \quad (1)$$

Dimana:

X*norm*	: nilai hasil normalisasi
x	: nilai asli dalam dataset
min(x)	: nilai minimum dari semua nilai dalam dataset
max(x)	: nilai maksimum dari semua nilai dalam dataset
$new_{min}(x)$	: nilai minimum yang diinginkan setelah normalisasi

$new_{max}(x)$: nilai maksimum yang diinginkan setelah normalisasi

4. *Downsampling*

Downsampling dilakukan dengan menghapus atau mengurangi data dari kelas majoritas, sehingga menghasilkan dataset yang lebih seimbang [9]. Proses ini berguna untuk menangani masalah data *imbalanced*, yang terjadi ketika jumlah data dalam kelas minoritas lebih kecil dari kelas majoritas. Dimana pada kasus penelitian dataset ini, pada nilai biner kolom *Exited* = 0 merupakan nilai mayoritas, dan *Exited* = 1 merupakan nilai minoritas. Dapat dilihat pada tabel 2. Data mayoritas disesuaikan dengan dat minoritas dengan menggunakan metode *resample*, yang artinya mengurangi data mayoritas sama dengan data minoritas. Dapat dilihat pada tabel 3.

Tabel 2. Banyaknya nilai variabel sebelum *downsampled*

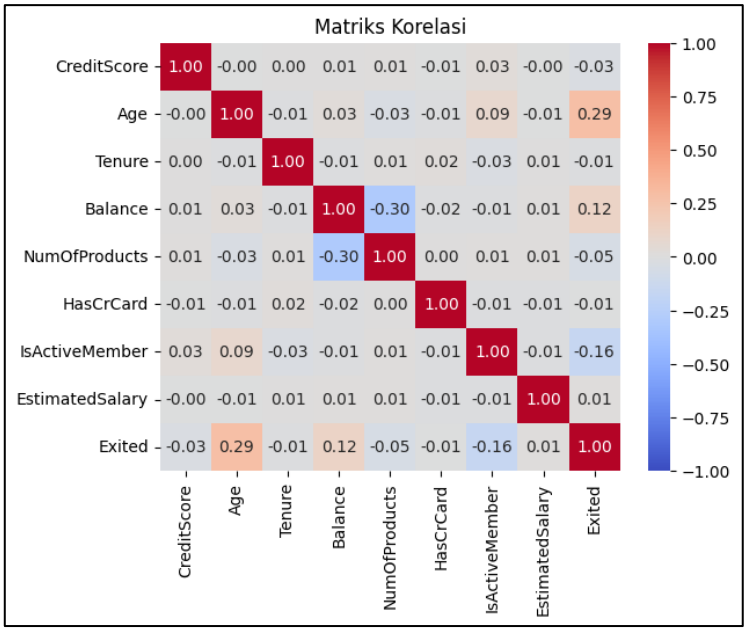
Exited	Value
0	7,958
1	2,036

Tabel 3. Banyaknya nilai variabel sesudah *downsampled*

Exited	Value
0	2,036
1	2,036

2.3 Feature Selection

Setelah dilakukannya data *preprocessing*, untuk mencari fitur yang akan digunakan dalam *modelling* pada penelitian ini, peneliti menggunakan matriks korelasi. Dapat dilihat hasil matriks korelasi pada gambar 4.



Gambar 4. Plot Matriks Korelasi

Korelasi adalah pengukuran statistik yang menunjukkan seberapa erat dua variabel berkaitan satu sama lain. Koefisien korelasi berkisar antara -1 dan 1. Nilai 1 menunjukkan korelasi positif sempurna, seperti *Age* dan *Exited* memiliki korelasi yang positif dimana jika nilai *Age* itu naik, maka nilai *Exited* juga naik. Nilai -1 menunjukkan korelasi negative sempurna, seperti *Balance* dan *NumOfProducts* memiliki korelasi negative dimana jika nilai *Balance* naik, maka nilai *NumOfProducts* akan turun. Nilai 0 menunjukkan tidak adanya korelasi.

Berdasarkan Plot tersebut pada gambar 4, didapatkan fitur yang dapat dipakai untuk *modelling*, yaitu *CreditScore*, *Age*, *Tenure*, *Balance*, *NumOfProducts*, *HasCrCard*, *IsActiveMember*, *EstimatedSalary*, dan *Exited*.

2.4 Model

Setelah data tersebut diproses, data sebelumnya dibagi menjadi dua, yaitu *data training* dan *data testing* dengan rasio 20% untuk testing. *Data testing* adalah data yang digunakan untuk memeriksa kinerja model, sedangkan *data training* adalah data yang digunakan untuk mengajar model.

Random Forest sebagai pengklasifikasi grup pembelajar pohon. Metode ini menggunakan beberapa pohon keputusan, sehingga setiap pohon bergantung pada nilai vektor acak yang dipilih secara khusus dengan distribusi yang sama untuk setiap pohon ini merupakan pilihan yang tepat untuk kecenderungan pohon keputusan yang overfit dengan koleksi pelatihan mereka [8]. Ini memungkinkan model yang lebih tepat dan stabil.

Decision Tree adalah sebuah prosedur yang mengiris kumpulan data menjadi beberapa segmen seperti cabang. *Decision Tree* mudah dibaca. Keuntungan ini membuat penjelasan untuk model menjadi sederhana [8]. Ini memungkinkan model yang lebih mudah dipahami dan diterapkan, tetapi dapat memiliki efek variasi yang lebih tinggi.

Pada tahap ini, peneliti membangun model *Random Forest* dan *Decision Tree* dengan menggunakan *data training* yang telah dihasilkan. Setelah model dibangun, peneliti menggunakan *data testing* untuk memeriksa kinerja model. Kinerja model akan diukur menggunakan metrik seperti *accuracy*, *precision*, *recall*, dan *F1-score*. Berikut dibawah ini rumus-rumusnya:

$$Accuracy = \frac{TP+FN}{TP+FP+TN+FN} \quad (2)$$

$$Precision = \frac{TP}{TP+FP} \quad (3)$$

$$Recall = \frac{TP}{TP+FN} \quad (4)$$

$$F1\ Score = \frac{Precision*Recall}{Precision+Recall} \quad (5)$$

Dimana:

True Positive (TP) : mendefinisikan pelanggan yang melakukan *churn* (positif) dan diklasifikasikan sebagai *churn* (positif).

True Negative (TN) : mendefinisikan nasabah yang tidak melakukan *churn* (negatif) dan diklasifikasikan sebagai tidak melakukan *churn* (negatif).

False Positive (FP) : mengidentifikasi pelanggan yang tidak melakukan *churn* (negatif) dan diklasifikasikan sebagai *churn* (positif).

False Negative (FN) : mengidentifikasi nasabah yang melakukan *churn* (positif) dan diklasifikasikan sebagai tidak melakukan *churn* (negatif)

3. HASIL DAN PEMBAHASAN

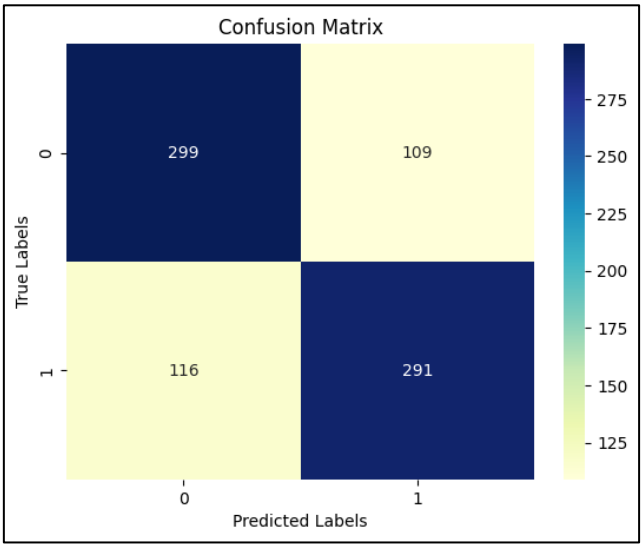
Ketika *preprocessing* data selesai, data akan berada dalam bentuk operasional, dan 9 fitur yang diperoleh setelah *preprocessing* diambil untuk sisa studi. Diantaranya, 80% data akan digunakan untuk pelatihan dan 20% sisanya akan digunakan untuk pengujian secara acak. Didapatkan data pelatihan sebanyak 3.257, dan data pengujian sebesar 815. Untuk memvalidasi model yang dikembangkan *Decision Tree* dan *Random Forest*. Bagian ini dibagi berdasarkan model-model yang dikembangkan yang telah dijelaskan sebelumnya di bagian metodologi.

Tabel 4. Hasil Model *Decision Tree*

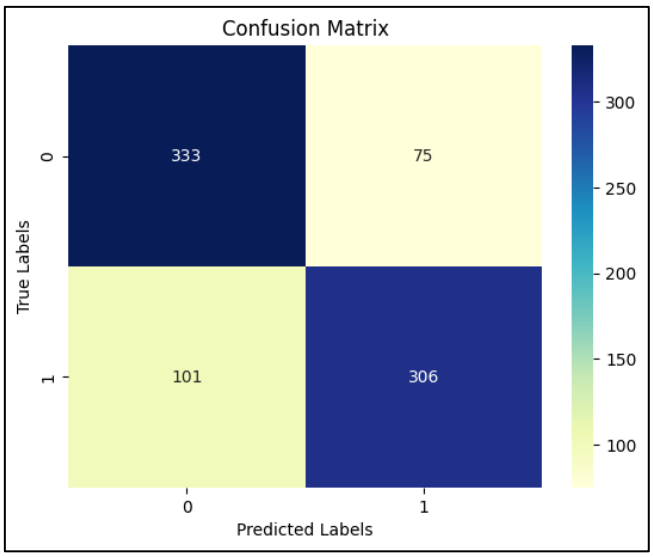
Decision Tree Classification Report				
	Precision	***Recall***	***F1-Score***	***Accuracy***
0	72%	73%	73%	72%
1	73%	71%	72%	
Average	72%	72%	72%	

Tabel 5. Hasil Model *Random Forest*

Decision Tree Classification Report				
	Precision	***Recall***	***F1-Score***	***Accuracy***
0	77%	82%	79%	78%
1	80%	75%	78%	
Average	79%	78%	78%	



Gambar 5. *Confusion Matrix Decision Tree*



Gambar 6. *Confusion matrix Random Forest*

Hasil dari pengujian kinerja kedua model *Decision Tree* dan *Random Forest* menggunakan *confusion matrix* menunjukkan kinerja model yang baik. Terlihat pada tabel 4 dan 5 merupakan hasil kedua klasifikasi tersebut. Rata-rata *precision* yang didapatkan oleh *Decision Tree* adalah 72%, sedangkan *Random Forest* mendapatkan 79%, yang artinya rata-rata tersebut nasabah yang diprediksi akan *churn* benar-benar *churn* dan sebaliknya. Rata-rata *recall* yang didapatkan oleh *Decision Tree* adalah 72%, sedangkan *Random Forest* mendapatkan 78%, yang artinya rata-rata tersebut nasabah yang benar *churn* atau tidak benar diprediksi oleh model. Rata-rata *F1-score* yang didapat oleh *Decision Tree* adalah 72%, sedangkan *Random Forest* mendapatkan 78%, yang juga artinya besarnya rata-rata kinerja model dalam melakukan prediksi sebesar. Dan rata-rata *accuracy* yang didapatkan oleh model *Decision Tree* adalah 72%, sedangkan *Random*

Forest adalah 78%, yang artinya persentase akurasi dari semua pelanggan yang diprediksi oleh model benar.

Hasil dari *confusion matrix* pada gambar 5 dan 6, hasil yang terlihat lebih bagus ialah model *Random Forest*. Untuk penjelasan bagian atas *confusion matrix* dari kiri ke kanan adalah *True Negative* dan *False Negative*, yang artinya *True Negative* adalah hasil prediksi yang benar bahwa data yang diprediksi sebagai negatif ternyata benar negative. Sedangkan *False Negative* adalah hasil prediksi yang salah bahwa data yang negatf yang diprediksi oleh model sebagai positif .Kemudian bagian bawahnya dari kiri ke kanan adalah *False Positive* dan *True Positive*, yang artinya *False Positive* adalah hasil prediksi yang salah dimana data yang benar adalah positif tetapi diprediksi sebagai negative. Sedangkan *True Positive* adalah hasil prediksi yang benar dimana data yang positif diprediksi sebagai positif juga. Terlihat dari kedua gambar tersebut dimana hasil prediksi bagian *True* itu lebih memiliki angka yang lebih besar daripada *Random Forest*. Dari hasil persentase juga terlihat bahwa *Random Forest* memiliki angka yang lebih tinggi dibandingkan *Decision Tree*. Model *Random Forest* telah menunjukkan kinerja yang tinggi dalam melakukan prediksi churn bank.

4. PENUTUP

Dalam penelitian ini, peneliti berhasil menggunakan metode klasifikasi *Decision Tree* dan *Random Forest* untuk memprediksi kehilangan nasabah dalam industri perbankan. Setelah melakukan proses *preprocessing data*, pembagian data pelatihan, dan pengujian, peneliti menemukan bahwa kedua model berhasil memprediksi perilaku kehilangan nasabah dengan baik. Namun, *Random Forest* sedikit lebih baik daripada *Decision Tree.*

Hasil evaluasi menggunakan *confusion matrix* menunjukkan bahwa *Random F*orest memiliki *precision*, *recall*, *F1-score*, dan *accuracy* yang lebih tinggi dibandingkan dengan *Decision Tree*. Ini menunjukkan bahwa model *Random Forest* mampu memprediksi pelanggan yang benar-benar churn dengan lebih baik dan meminimalkan kesalahan dalam memprediksi pelanggan yang tidak churn.

Meskipun *Decision Tree* menunjukkan hasil yang dapat diterima, peningkatan signifikan dalam kinerja *Random Forest* menunjukkan bahwa ada potensi untuk mengoptimalkan strategi manajemen risiko churn nasabah yang digunakan oleh industri perbankan.

Untuk penelitian selanjutnya, peneliti merekomendasikan beberapa langkah untuk meningkatkan pemahaman dan kinerja model prediksi churn nasabah:

a. Penyesuaian Parameter: Untuk meningkatkan kinerja prediksi, ubah parameter kedua model.
b. Pemrosesan Data Lanjutan: Untuk mengatasi ketidakseimbangan kelas dalam dataset, pelajari metode pemrosesan data khusus seperti *oversampling* atau *undersampling*.
c. Penggunaan Fitur Tambahan: Tingkatkan fitur yang mungkin relevan untuk memprediksi kehilangan pelanggan.
d. Validasi Eksternal: Untuk menguji generalisasi model pada data baru, gunakan dataset eksternal untuk melakukan validasi model.

Dengan mengambil langkah-langkah ini, penelitian selanjutnya diharapkan dapat memberikan wawasan yang lebih mendalam dan solusi yang lebih efisien untuk menangani masalah churn nasabah dalam industri perbankan.

DAFTAR PUSTAKA

[1] Miryam Clementine and Arum, “Prediksi Churn Nasabah Bank Menggunakan Klasifikasi NaÃ¯ve Bayes dan ID3,” *Jurnal Processor*, vol. 17, no. 1, pp. 9–18, May 2022, doi: 10.33998/processor.2022.17.1.1170.

[2] M. Kaur, K. Singh, and N. Sharma, “International Journal on Recent and Innovation Trends in Computing and Communication Data Mining as a tool to Predict the Churn Behaviour among Indian bank customers”, [Online]. Available: http://www.ijritcc.org

[3] H. Guliyev and F. Yerdelen Tatoğlu, “Customer churn analysis in banking sector: Evidence from explainable machine learning models,” *Journal of Applied Microeconometrics*, vol. 1, no. 2, pp. 85–99, Dec. 2021, doi: 10.53753/jame.1.2.03.

[4] F. R. Lumbanraja, W. Mudyaningsih, B. Hermanto[3], and A. Syarif, “IMPLEMENTASI METODE RANDOM FOREST UNTUK PREDIKSI POSISI METILASI PADA SEKUENS PROTEIN.”

[5] M. A. Hambali and I. Andrew, “Bank Customer Churn Prediction Using SMOTE: A Comparative Analysis,” *Qeios*, Mar. 2024, doi: 10.32388/H82XTW.

[6] Y. Y. Song and Y. Lu, "Decision tree methods: applications for classification and prediction," *Shanghai Arch Psychiatry*, vol. 27, no. 2, pp. 130–135, Apr. 2015, doi: 10.11919/j.issn.1002-0829.215044.

[7] B. Charbuty and A. Abdulazeez, "Classification Based on Decision Tree Algorithm for Machine Learning," *Journal of Applied Science and Technology Trends*, vol. 2, no. 01, pp. 20–28, Mar. 2021, doi: 10.38094/jastt20165.

[8] M. Rahman and V. Kumar, "Machine Learning Based Customer Churn Prediction in Banking," in *Proceedings of the 4th International Conference on Electronics, Communication and Aerospace Technology, ICECA 2020*, Institute of Electrical and Electronics Engineers Inc., Nov. 2020, pp. 1196–1201. doi: 10.1109/ICECA49313.2020.9297529.

[9] C. Chen and A. Liaw, "Using Random Forest to Learn Imbalanced Data."

[10] M. Ohsaki, P. Wang, K. Matsuda, S. Katagiri, H. Watanabe, and A. Ralescu, "Confusion-matrix-based kernel logistic regression for imbalanced data classification," *IEEE Trans Knowl Data Eng*, vol. 29, no. 9, pp. 1806–1819, Sep. 2017, doi: 10.1109/TKDE.2017.2682249.

[11] F. Agung, J. Ayomi, and K. E. Dewi, "ANALISIS EMOSI PADA MEDIA SOSIAL TWITTER MENGGUNAKAN METODE MULTINOMIAL NAÏVE BAYES DAN SYNTHETIC MINORITY OVERSAMPLING TECHNIQUE," *KOMPUTA : Jurnal Ilmiah Komputer dan Informatika*, vol. 12, no. 2, 2023, [Online]. Available: https://www.statista.com

[12] D. AL-Najjar, N. Al-Rousan, and H. AL-Najjar, "Machine Learning to Develop Credit Card Customer Churn Prediction," *Journal of Theoretical and Applied Electronic Commerce Research*, vol. 17, no. 4, pp. 1529–1542, Dec. 2022, doi: 10.3390/jtaer17040077.

[13] M. Zheng, "Customer Churn Prediction based on Multiple Algorithms," *Scientific Journal of Economics and Management Research*, vol. 2, p. 2020.

[14] S. D. Oleh and A. Allaam, "Prediksi Churn Konsumen Menggunakan Algoritma Random Forest dengan Fuzzy C-Means untuk Meningkatkan Produktivitas Penjualan Bisnis," 2023.

[15] P. Verma, "Churn prediction for savings bank customers: A machine learning approach," *J Stat Appl Probab*, vol. 9, no. 3, pp. 535–547, 2020, doi: 10.18576/JSAP/090310.