IMPLEMENTASI METODE *RANDOM FOREST* PADA KLASIFIKASI *CHURN CUSTOMER*

(Skripsi)

Oleh
LUTFIA HUMAIROSI
1817031037

JURUSAN MATEMATIKA
FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
UNIVERSITAS LAMPUNG
BANDAR LAMPUNG
2022

ABSTRACT

IMPLEMENTATION OF RANDOM FOREST METHOD ON CHURN CUSTOMER CLASSIFICATION

By

Lutfia Humairosi

Banking industry is growing each year, it results in high market competition. The movement of customers from one company to another that generates customers within a certain time interval of customer churn. The case of customer transfer needs to be studied by predicting the behavior of customers who leave the company and the company can retain customers. This problem can be handled by classify customer behavior based on previous available data. This study uses the random forest method with a best composition of 90% used as training data, and 10% of the data used as test data, the best parameters obtained are 400 n_estimators, 40 max_depth, and Synthetic Minority Over-Sampling Technique (SMOTE) is used to handle imbalanced data, the model accuracy is 90,83%, the precision value is 89,29%, and the recall value is 92,07%, and the f1-score value is 90,66%.

Keywords: Churn Customer, Machine Learning, Random Forest, SMOTE

ABSTRAK

IMPLEMENTASI METODE *RANDOM FOREST* PADA KLASIFIKASI *CHURN CUSTOMER*

Oleh

Lutfia Humairosi

Industri perbankan semakin meningkat setiap tahunnya, hal ini mengakibatkan semakin tinggi persaingan perlakuan terhadap *customer*. Perpindahan *customer* dari suatu perusahaan ke perusahaan lain yang mengakibatkan hilangnya *customer* dalam selang waktu tertentu dinamakan *churn customer*. Kasus perpindahan *customer* perlu dikaji dengan memprediksi perilaku *customer* yang berpotensi meninggalkan perusahaan dan peusahaan dapat mempertahankan *customer*. Permasalahan tersebut dapat ditangani dengan cara melakukan klasifikasi perilaku *customer* berdasarkan data yang ada sebelumnya. Penelitian ini menggunakan metode *random forest* dengan komposisi terbaik sebesar 90% digunakan sebagai data latih dan 10% data digunakan sebagai data uji, didapatkan parameter terbaik berupa 400 *n_estimators,* 40 *max_depth,* dan digunakan *Synthetic Minority Over-Sampling Technique* (SMOTE) untuk menangani *imbalanced* data, dihasilkan nilai akurasi model sebesar 90,83%, nilai *precision* sebesar 89,29%, nilai *recall* sebesar 92,07%, dan nilai *f1-score* sebesar 90,66%.

Kata kunci: *Churn Customer, Machine Learning, Random Forest,* SMOTE

IMPLEMENTASI METODE *RANDOM FOREST* PADA KLASIFIKASI *CHURN CUSTOMER*

Oleh

Lutfia Humairosi

Skripsi

Sebagai Salah Satu Syarat untuk Mencapai Gelar
SARJANA MATEMATIKA

Pada

Jurusan Matematika
Fakultas Matematika dan Ilmu Pengetahuan Alam
Universitas Lampung

FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
UNIVERSITAS LAMPUNG
BANDAR LAMPUNG
2022

Judul Skripsi	: **IMPLEMENTASI METODE *RANDOM FOREST* PADA KLASIFIKASI *CHURN CUSTOMER***
Nama Mahasiswa	: Lutfia Humairosi
Nomor Pokok Mahasiswa	: 1817031037
Jurusan	: Matematika
Fakultas	: Matematika dan Ilmu Pengetahuan Alam

MENYETUJUI

1. Komisi Pembimbing

Ir. Warsono, M.Sc., Ph.D.	**Dr. Notiragayu, S.Si., M.Si.**
NIP. 196302161987031003	NIP. 197311092000122001

2. Ketua Jurusan Matematika

Dr. Aang Nuryaman, S.Si., M.Si.
NIP. 197403162005011001

MENGESAHKAN

1. Tim Penguji

Ketua : **Ir. Warsono, M.Sc., Ph.D.** .....

Sekertaris : **Dr. Notiragayu, S.Si., M.Si.** .....

Penguji
Bukan Pembimbing : **Dian Kurniasari, S.Si., M.Sc.** .....

2. Dekan Fakultas Matematika dan Ilmu Pengetahuan Alam

Dr. Eng. Suripto Dwi Yuwono, S.Si., M.T.
NIP. 197407052000031001

Tanggal Lulus Ujian Skripsi : 05 Oktober 2022

PERNYATAAN SKRIPSI MAHASISWA

Yang bertanda tangan di bawah ini:

Nama Mahasiswa	: **Lutfia Humairosi**
Nomor Pokok Mahasiswa	: **1817031037**
Jurusan	: **Matematika**
Judul Skripsi	: **IMPLEMENTASI METODE *RANDOM FOREST* PADA KLASIFIKASI *CHURN CUSTOMER***

Dengan ini menyatakan bahwa penelitian ini adalah hasil pekerjaan saya sendiri dan apabila kemudian hari terbukti bahwa skripsi ini merupakaan salinan atau dibuat oleh orang lain, maka saya bersedia menerima sanksi sesuai dengan ketentuan akademik yang berlaku.

Bandar Lampung, 05 Oktober 2022
Penulis

Lutfia Humairosi
NPM. 1817031037

RIWAYAT HIDUP

Penulis bernama Lutfia Humairosi, anak pertama dari tiga bersaudara yang dilahirkan di Bandar Lampung pada 27 Februari 2000 dari pasangan Bapak Muhammad Santoso dan Ibu Tetti Emina HS.

Penulis menyelesaikan pendidikan sekolah dasar di SD Negeri 1 Sawah Lama pada tahun 2006-2012. Selanjutnya, penulis melanjutkan jenjang pendidikannya di SMP Negeri 23 Bandar Lampung pada tahun 2012-2015, dan sekolah menengah atas di SMA Negeri 2 Bandar Lampung pada tahun 2015-2018.

Pada tahun 2018 penulis terdaftar sebagai mahasiswa S1 Jurusan Matematika, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Lampung melalui jalur SNMPTN. Selama menjadi mahasiswa penulis ikut serta dalam organisasi Himpunan Mahasiswa Matematika (HIMATIKA) FMIPA UNILA sebagai anggota Biro Dana dan Usaha. Pada tahun 2021 penulis melaksanakan Kuliah Kerja Nyata (KKN) di Kelurahan Langkapura, Kota Bandar Lampung dan Kerja Praktik (KP) di Badan Pendapatan Daerah Provinsi Lampung (BAPENDA), serta mengikuti Program Kampus Merdeka yang bernama Kampus Mengajar di SD Negeri 3 Kotakarang dan Studi Independen di PT. Digitalisasi Pemuda Indonesia Digital Skola.

KATA INSPIRASI

Dan apabila hamba-hamba-Ku bertanya kepadamu (Muhammad) tentang Aku, maka sesungguhnya Aku dekat. Aku Kabulkan permohonan orang yang berdoa apabila dia berdoa kepada-Ku. Hendaklah mereka itu memenuhi (perintah)-Ku dan beriman kepada-Ku, agar mereka memperoleh kebenaran
(Q.S. Al-Baqarah: 186)

Janganlah engkau bersedih, sesungguhya Allah bersama kita
(Q.S. At-Taubah: 40)

Knowing the lacking part of yourself is also an amazing strength
(Huang Guanheng Hendery)

To be yourself and love yourself, because only then you will be able to shine your own light and people will like yo more
(Huang Renjun)

PERSEMBAHAN

Alhamdulillahirabbil'alamin.
Puji dan syukur atas kehadirat Allah SWT karna atas segala rahmat dan hidayah-Nya skripsi ini dapat terselesaikan dengan baik. Penulis mempersembahkan skripsi ini kepada:

Ayah Muhammad Santoso dan Ibu Tetti Emina HS.
Terima kasih kepada kedua orang tua saya yang telah memberikan dukungan, kasih sayang, doa, serta semangat dalam setiap langkah yang saya tempuh.

Nabila Atifa dan Naila Amali Mumtaza
Terima kasih kepada adik-adiku yang selalu memberikan doa, mendukung dan memberikan semangat.

Dosen
Terima kasih kepada bapak dan ibu dosen pembimbing dan pembahas yang sangat berjasa dalam membimbing, memberikan arahan, serta ilmu yang bermanfaat.

Sahabat-Sahabatku
Para sahabat yang telah membantu, menemani, serta mendukung setiap langkah dalam hidupku.

Almamater Kebanggaan, Universitas Lampung

SANWACANA

Puji dan syukur penulis ucapkan kehadirat Allah SWT yang telah memberikan rahmat dan hidayah-Nya, sehingga penulis dapat menyelesaikan skripsi yang berjudul "Implementasi Metode *Random Forest* Pada Klasifikasi *Churn Customer*".

Penulis menyadari bahwa dalam penulisan skripsi ini tidak terlepas dari bimbingan, dukungan, bantuan dan doa dari berbagai pihak. Oleh karena itu, penulis ingin menyampaikan terimakasih kepada:

1. Bapak Ir. Warsono, M.S., Ph.D., selaku Dosen Pembimbing I dan Pembimbing Akademik yang telah bersedia membimbing, memberi saran, bantuan, motivasi, dan arahan dalam menyelesaikan skripsi ini.
2. Ibu Dr. Notiragayu, S.Si., M.Si., selaku Pembimbing II yang telah memberikan saran serta masukan kepada penulis sehingga dapat menyelesaikan skripsi ini.
3. Ibu Dian Kurniasari, S.Si., M.Sc., selaku dosen penguji yang telah memberikan kritik dan saran selama proses penyusunan skripsi.
4. Bapak Dr. Aang Nuryaman, S.Si., M.Si., selaku Ketua Jurusan Matematika Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Lampung.
5. Bapak Dr. Eng. Suripto Dwi Yuwono, S.Si., M.T., selaku Dekan Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Lampung.
6. Seluruh dosen, staff, karyawan Jurusan Matematika Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Lampung.
7. Ayah, Ibu, Bila, Naila, dan seluruh keluarga besar yang selalu mendoakan dan memberikan dukungan.

8. Sahabat-sahabat seperjuangan, Caca, Nanda, Shabrina, Shofiyyah, Mutia, Nadya, Zahwa, Reajeng, Putsal, Ranti, Maydia, Ratu, Virda, dan Sofa yang selalu memberikan bantuan, dukungan, dan motivasi selama masa perkuliahan.
9. Sahabat-sahabat terbaikku, Ailsa, Dara, Fisky, Norol, Sekar, Kila, Celsi, dan Yunda Anggita yang selalu memberikan dukungan, serta menemani selama menyusun skripsi ini.
10. Teman-teman Matematika 2018 atas kebersamaan selama masa perkuliahan hingga akhir.
11. Semua pihak yang telah membantu yang tidak bisa penulis sebutkan satu persatu.

Penulis menyadari masih banyak kekurangan dalam penulisan skripsi ini. Oleh karena itu, kritik dan saran sangat diharapkan agar dapat menjadi pelajaran dan perbaikan untuk kedepannya. Semoga skripsi ini dapat memberikan manfaat baik bagi penulis maupun bagi pihak yang membutuhkan.

Bandar Lampung, Oktober 2022
Penulis,

Lutfia Humairosi

DAFTAR ISI

Halaman

DAFTAR TABEL

Tabel Halaman

DAFTAR GAMBAR

Gambar Halaman

I. PENDAHULUAN

1.1 Latar Belakang dan Masalah

Industri perbankan semakin meningkat pada setiap tahun, peningkatan ini ditunjukkan dengan bertambahnya jumlah bank. Hal ini mengakibatkan semakin tinggi persaingan perlakuan terhadap *customer*. Pamina dkk (2019) mengatakan bahwa, perpindahan *customer* dari suatu perusahaan ke perusahaan lain yang mengakibatkan hilangnya *customer* dalam selang waktu tertentu. Secara signifikan kejadian tersebut dapat mengurangi pendapatan perusahaan, perpindahan *customer* tersebut dinamakan *churn customer*. Kasus perpindahan *customer* perlu dikaji dengan memprediksi perilaku *customer* yang berpotensi meninggalkan perusahaan. Prediksi ini harus dilakukan dengan tepat agar dapat memberikan informasi kepada perusahaan untuk dapat mempertahankan *customer* sehingga dapat mengurangi persentase *churn customer*. Permasalahan tersebut dapat ditangani dengan cara melakukan klasifikasi perilaku *customer* berdasarkan data yang ada sebelumnya.

Klasifikasi merupakan proses menemukan model atau fungsi yang membedakan kelas atau konsep data (Utami dkk, 2020). Pada kelas klasifikasi sering terjadi permasalahan *imbalanced data* atau data tidak seimbang yang menyebabkan kesalahan klasifikasi pada kelas minoritas dan juga akurasi yang dihasilkan menjadi kurang maksimal. Permasalahan tersebut dapat ditangani dengan *Synthetic Minority Over-Sampling Technique* (SMOTE). Kasus *imbalanced data* ditangani oleh SMOTE karena data sintetis yang dihasilkan untuk kelas minoritas

merupakan data yang diambil dari tetangga terdekat kelas minoritas (Chawla dkk, 2002).

Metode yang sering digunakan untuk klasifikasi ialah algoritma *C4.5* untuk *decision tree.* Metode ini menghasilkan *rules* dan pohon keputusan, namun sering terjadi *overlapping* ketika data yang dikelola berukuran besar dan juga perlu banyak waktu untuk pengambilan keputusan. Semakin dalam pohon dapat membuat model menjadi *overfitting* (Benediktus dan Oetama, 2020). Salah satu metode yang dapat menangani permasalahan tersebut adalah *random forest.* Metode *random forest* membangun pohon yang lebih kecil dengan menerapkan *bootsrap,* kemudian masing-masing pohon akan menghasilkan keputusan yang selanjutnya dilakukan *majority voting (aggregation)* sebagai keputusan akhir. Breiman (2001) mengatakan bahwa, *random forest* dapat bekerja secara efisien untuk dataset dengan skala besar dan dapat memberikan tingkat akurasi yang baik.

Berdasarkan penelitian yang telah dilakukan oleh Syukron dan Subekti (2018) mengenai penerapan metode *random forest* untuk klasifikasi penilaian kredit dengan *imbalanced data,* diperoleh akurasi sebesar 76,01%. Setelah dilakukan *random over-under sampling* sehingga data menjadi *balanced* diperoleh akurasi sebesar 90,1%. Religia dkk (2021) melakukan penelitian mengenai optimasi model *random forest* untuk klasifikasi data *bank marketing* diperoleh nilai akurasi menggunakan *random forest* sebesar 88,30%.

Berdasarkan penjelasan sebelumnya penelitian yang sudah dilakukan untuk metode *random forest* memiliki nilai akurasi yang baik, oleh karena itu penelitian ini akan melakukan pengklasifikasian perilaku *customer* untuk memprediksi apakah customer tersebut *churn* atau tidak menggunakan metode *random forest* pada data sekunder *churn customer* dengan menerapkan SMOTE untuk menangani *imbalanced data.*

1.2 Tujuan Penelitian

Adapun tujuan dari penelitian ini adalah:

1. Melakukan *balanced* data dengan mengaplikasikan metode SMOTE.
2. Melakukan klasifikasi untuk memprediksi potensi *churn customer* dengan menerapkan metode *random forest.*
3. Mengetahui performa klasifikasi *churn customer* dengan menggunakan metode *random forest*.

1.3 Manfaat Penelitian

Manfaat dari penelitian ini adalah:

1. Menambah pengetahuan statistika di bidang *machine learning* khususnya pada metode *random forest.*
2. Sebagai bahan referensi bagi perusahaan dalam memberikan perlakuan yang tepat untuk mempertahankan *customer*.

II. TINJAUAN PUSTAKA

2.1 *Data Mining*

Data mining adalah proses menemukan pengetahuan yang menarik dari sejumlah besar data dan berkaitan dengan berbagai bidang seperti, teknologi *database, artificial intelligence, machine learning, neural networks,* statistik, *pattern recognition, knowledge based systems, knowledge acquisition, information retrieval, high performance computing,* dan visualisasi data (Han dkk, 2000). Vijayakumar dan Nedunchezhian (2012) mengatakan bahwa, *data mining* adalah proses memanipulasi data dengan mengekstrak informasi yang belum diketahui dari kumpulan data yang besar. Dean (2014) berpendapat bahwa, *data mining* adalah sebuah cara atau langkah dalam proses *Knowledge Discovery In Databases* (KDD). Proses yang dilakukan dalam KDD yaitu sebagai berikut (Han dkk, 2000):

1. *Data integration*

   Data integration merupakan proses menggabungkan data dari beberapa sumber data.

2. *Data selection*

   Data selection merupakan proses seleksi data, data yang relevan digunakan terhadap analisis yang akan dilakukan.

3. *Preprocessing data*

   a. *Data cleansing*

      Umumnya data yang didapatkan memiliki data hilang, ataupun kesalahan pada *input* data. *Data cleansing* merupakan proses menghilangkan *noise* dan data yang tidak relevan.

b. Scaling data

Scaling data merupakan proses transformasi data dari bentuk asli ke dalam bentuk lain yang sesuai untuk *data mining.* *Scaling data* digunakan untuk menyesuaikan data yang diolah berdasarkan algoritma yang digunakan. Terdapat dua cara yang biasanya digunakan untuk *scaling data* yaitu:

a. *Min-max normalization* merupakan proses transformasi data yang bekerja dengan cara menempatkan data dalam *range* 0 sebagai nilai terkecil dan 1 sebagai nilai terbesar (Hanifa dkk, 2017). Rumus perhitungan pada *min-max normalization* yaitu:

$$x' = \frac{x - x_{min}}{x_{max} - x_{min}} \quad (2.1)$$

dengan:

x' = nilai x baru

x_{min} = nilai x minimum

x_{max} = nilai x maksimum

b. *Z-score normalization (standardscaler)* merupakan metode transformasi data bedasarkan nilai rata-rata dan standar deviasi yang bertujuan untuk mencegah adanya data yang memiliki nilai terlalu besar dibandingkan dengan nilai yang lain (Prasetyo dkk, 2022). Li dan Liu (2011) mengatakan bahwa, metode ini sangat berguna pada data yang tidak diketahui nilai minimum dan maksimum yang sebenarnya. Rumus perhitungan pada *Z-score normalization* yaitu:

$$Z = \frac{x - \mu}{\sigma} \quad (2.2)$$

dengan:

x = nilai yang diamati

μ = rata-rata nilai (*mean*)

σ = standar deviasi

c. *Handling data categorical*

Dataset dapat terdiri dari data numerik maupun data kategorik, pada data kategorik diberikan label agar lebih mudah dipahami oleh komputer.

Winata dkk (2020) mengatakan bahwa, terdapat dua cara yang biasa digunakan untuk *handling data categorical* yaitu *label encoding* dan *one hot encoding*. *Label encoding* digunakan ketika data memiliki tingkatan yang berbeda, berikut merupakan contoh data yang sudah dilakukan *label encoding.*

Tabel 1. Contoh *label encoding*

Penilaian	*Label encoding*
Buruk	1
Cukup	2
Baik	3

One hot encoding digunakan ketika data tidak memiliki tingkatan yang berbeda berikut merupakan contoh data yang sudah dilakukan *one hot encoding.*

Tabel 2. Contoh *one hot encoding*

Perancis	Jerman	Spanyol
1	0	0
0	1	0
0	0	1

4. *Data mining*

   Data mining merupakan proses penting dimana metode diterapkan untuk mendapatkan informasi baru dalam data menggunakan metode tertentu.

5. *Pattern evaluation*

   Pattern evaluation merupakan proses untuk mengidentifikasi pola yang menarik yang dapat menjelaskan pengetahuan yang didapatkan.

6. *Knowledge presentation*

   Knowledge presentation merupakan teknik visualisasi dan representasi data yang digunakan untuk menggambarkan dan juga mendeskripsikan pengetahuan yang didapatkan.

Davies (2003) berpendapat, *data mining* dapat didefinisikan sebagai berikut :

1. *Data mining* berkaitan dengan penemuan pola data yang tersembunyi dan tidak terduga.
2. *Data mining* biasanya memproses data dalam jumlah besar. Data besar diperlukan untuk menarik kesimpulan yang andal berkaitan dengan model data.
3. *Data mining* membantu dalam pengambilan keputusan, seperti aplikasi dalam penelitian geologi dan meteorologi.

Dari beberapa definisi di atas tentang *data mining*, dapat disimpulkan bahwa *data mining* adalah proses memanfaatkan data yang jumlahnya besar, untuk menemukan informasi atau pola yang menarik yang belum diketahui dan berpotensi menjadi sesuatu yang bermanfaat.

2.2 *Machine Learning*

Machine learning menyelidiki cara-cara dimana komputer dapat memperoleh pengetahuan langsung dari suatu data dan selanjutnya dipelajari untuk menyelesaikan masalah, tidak akan membutuhkan waktu lama bagi *machine learning* untuk mempengaruhi bidang statistik (Ratner, 2011). *Machine learning* adalah mesin yang dirancang untuk dapat belajar sendiri tanpa bimbingan pengguna (Sihombing dan Arsani, 2021).

Machine learning adalah pemrograman komputer untuk memaksimalkan kinerja menggunakan data sampel atau pengalaman masa lalu, dan menggunakan teori statistika dalam membangun model matematis untuk menghasilkan suatu kesimpulan. The royal society (2017) mengatakan bahwa, terdapat 3 cabang utama dalam *machine learning* yaitu :

1. *Supervised Machine Learning*

   Supervised machine learning merupakan sistem yang dilatih menggunakan data berlabel yang mengkategorikan setiap data menjadi satu atau beberapa kelompok. Sistem mempelajari data yang akan dilatih menjadi terstruktur dan digunakan untuk memprediksi data uji.
2. *Unsupervised Learning*

   Unsupervised learning merupakan sistem yang menguji data tidak berlabel, hal ini bertujuan untuk menemukan karakteristik yang membuat titik data kurang lebih mirip satu sama lain, contohnya dengan membuat klaster dan menetapkan data ke klaster tersebut.
3. *Reinforcement Learning*

   Reinforcement learning merupakan sistem dari *machine learning* yang melakukan pendeketan uji dengan *trial and error* untuk mencapai tujuan. Oleh karena itu, diperlukan *reward* dari lingkungannya sebagai pengganti data respon *input* dan *output*. *Reward* digunakan untuk menguji interaksi lingkungan dan pengumpulan jumlah *reward* secara maksimal penting karena *reward* menjadi *signal feedback* dalam proses *learning.*

2.3 Klasifikasi

Klasifikasi adalah proses untuk menemukan sekumpulan model atau fitur yang menggambarkan dan membedakan antara kelas data dan konsep, dengan tujuan menggunakan model untuk memprediksi kelas objek yang labelnya tidak dikenali (Han dkk, 2000). Oktanisa dan Supianto (2018) mengatakan bahwa, klasifikasi adalah teknik dalam *data mining* yang mengelompokkan data berdasarkan keterikatan data terhadap data sampel. Beberapa metode untuk membangun

klasifikasi diantaranya *naïve-bayesian, Support Vector Machine (SVM), random forest, dan neighbor classification.* Tugas klasifikasi adalah membangun model prediksi berdasarkan informasi yang terkandung dalam satu set *training* atau *testing* sampel berlabel (Vluymans, 2019).

Proses klasifikasi memiliki empat komponen dasar, yaitu (Gorunescu, 2011):

1. Kelas

   Kelas atau label kelas adalah variabel terikat dari model yang merupakan variabel kategorik yang mewakili suatu label untuk objek setelah klasifikasi.
2. Prediktor

   Prediktor adalah variabel bebas yang mewakili karakteristik untuk model data yang diklasifikasikan.
3. *Training set*

   Training set berisi kumpulan data yang berisi nilai dari dua komponen sebelumnya (kelas dan prediktor) yang digunakan untuk melatih model agar mengenali kelas, berdasarkan prediktor yang sudah ada.
4. *Testing set*

   Testing set berisi data baru untuk diklasifikasikan dengan model yang telah dibuat, juga untuk mengukur tingkat akurasi klasifikasi, sehingga dapat dilihat efektivitas kinerja dari model klasifikasi.

2.4 *Synthetic Minority Over-Sampling Technique* (SMOTE)

Synthetic Minority Over-Sampling Technique (SMOTE) pertama kali diperkenalkan oleh Chawla dkk pada tahun 2002. Siringoringo (2018) mengatakan bahwa, SMOTE merupakan metode yang mensintesis sampel baru dari kelas minoritas dengan cara melakukan sampling ulang sampel pada kelas minoritas untuk menyeimbangkan dataset. SMOTE bekerja dengan cara mengarahkan sampel sintetis dan mengklasifikasikannya untuk membangun

wilayah keputusan yang lebih besar berdasarkan kelas minoritas terdekat (Hao dkk, 2014).

Ide utama dari SMOTE adalah menambah jumlah sampel pada kelas minoritas agar seimbang dengan kelas mayoritas. Cara yang digunakan adalah membuat data sintetik berdasarkan tetangga terdekat yang dipilih berdasarkan jarak *euclidean* antara kedua data tersebut (Chawla dkk, 2002). Hapsari dan Indriyani (2022) mengatakan bahwa, pemilihan tetangga terdekat bedasarkan jarak *eucledian* antar sepasang data dilakukan sebagai berikut.
Misalkan diberikan data dengan *q* variabel yaitu

$$\mathbf{x}^{\boldsymbol{T}} = \left[\mathrm{x}_{1,}\mathrm{x}_{2,\ldots,}\mathrm{x}_{q}\right] \text{ dan } \mathbf{z}^{T} = \left[\mathrm{z}_{1,}\mathrm{z}_{2,\ldots,}\mathrm{z}_{q}\right] \tag{2.3}$$

dengan:

$\mathbf{x}^{\boldsymbol{T}}$ = data 1

$\mathbf{z}^{T}$ = data 2

Maka jarak *euclidean* d(x,z) secara umum sebagai berikut :

$$d(\mathbf{x},\mathbf{z}) = \sqrt{(\mathrm{x}_1 - \mathrm{z}_1)^{\mathbf{2}} + (\mathrm{x}_2 - \mathrm{z}_2)^{\mathbf{2}} + \cdots + (\mathrm{x}_q - \mathrm{z}_q)^{\mathbf{2}}} \tag{2.4}$$

dengan:

$d(\mathbf{x},\mathbf{z})$ = jarak

Melakukan pembangkitan data *synthetic* dengan menggunakan persamaan berikut:

$$\mathbf{x}_{\mathrm{syn}} = \mathbf{x}_i + (\mathbf{x}_{knn} - \mathbf{x}_i)\gamma \tag{2.5}$$

dengan :

$\mathbf{x}_{\mathrm{syn}}$ = data hasil replikasi

$\mathbf{x}_i$ = data yang akan di replikasi

$\mathbf{x}_{knn}$ = data dari kelas minor yang memiliki jarak terdekat dari $\mathbf{x}_i$

γ = bilangan *random* antara 0 dan 1

2.5 *Boostrap Aggregation*

Leo Breiman pertama kali mengenalkan metode *boostrap aggregation* atau *bagging* untuk mengurangi perbedaan prediktor. Breiman (1996) berpendapat bahwa, *bagging* adalah teknik *ensemble* yang efektif untuk algoritma pembelajaran yang tidak setimbang, karena perubahan kecil dalam kumpulan data *training* menyebabkan perubahan besar dalam prediksi, seperti pohon keputusan, jaringan saraf, dan lain-lain. Syarif dkk (2012) mengatakan bahwa, metode *bagging* sangat berguna pada dataset yang besar. Metode ini umumnya diterapkan pada algoritma *decision tree,* dan algorima klasifikasi lain seperti *random forest, naïve bayes*, dan *nearest neighbour*. Setiap data latih dari proses *bagging* dibuat menjadi pohon klasifikasi dan dilakukan proses *aggregate* untuk mendapatkan suara mayoritas dari kasus klasifikasi. *Bagging* bertujuan meningkatkan akurasi klasifikasi dengan menggabungkan klasifikasi tunggal, sehingga hasil yang diperoleh mejadi lebih baik (Alfaro dkk, 2013).

Implementasi *bagging* dalam algoritma *Tree* dilakukan dengan menggunakan pohon keputusan hasil dari membangun klasifikasi dasar C1, C2,…, Cn pada sampel *bootstrap* D1, D2, .., Dn dengan pengganti untuk data D, kemudian keputusan akhir diturunkan sebagai kombinasi dari semua pengklasifikasi dasar C1, C2,…, Cn dengan suara terbanyak. *Bagging* diterapkan pada metode klasifikasi dengan pohon keputusan seperti *reduced error pruning tree, random forest*, dan C4.5 (Dhakate et al, 2015).

2.6 *Hyperparameter Tuning*

Hyperparameter tuning adalah metode terbaik untuk mendapatkan sejumlah kombinasi parameter yang berbeda untuk menilai kinerja pengklasifikasi (George dan Sumathi, 2020). *Grid search* merupakan salah satu metode alternatif *hyperparameter* yang bertujuan untuk menemukan parameter terbaik dari suatu model yang menghasilkan prediksi data yang akurat. *Grid search* mengeksplorasi masing-masing parameter dengan mengatur jenis nilai prediksi terlebih dahulu untuk menentukan parameter terbaik. Terdapat beberapa parameter yang digunakan untuk melakukan *hyperparameter* pada metode *random forest* yaitu, *n_estimators* merupakan banyaknya pohon dan *max_depth* merupakan kedalaman maksimum pohon (Atei dan Osanloo, 2004). Breiman (1996) berpendapat bahwa, *n_estimator* dengan nilai 50 sudah memberikan hasil yang baik untuk masalah klasifikasi sedangkan Sutton (2005) berpendapat bahwa, nilai *n_estimator* ≥ 100 memberikan tingkat kesalahan klasifikasi yang rendah.

2.7 *Random Forest*

Pada tahun 2001, metode *random forest* diperkenalkan oleh Leo Breiman dengan menunjukan kelebihan dari metode tersebut yang diantaranya dapat secara efisien menangani *training* data dengan jumlah yang banyak, menghasilkan performa yang baik untuk klasifikasi, dan menghasilkan *error* yang rendah. Metode *random forest* adalah kumpulan pohon prediktor dimana setiap pohon secara independen bergantung pada sampel dari vektor acak dan memiliki distribusi yang sama untuk semua pohon di hutan (Breiman, 2001). Metode *random forest* memiliki kemampuan untuk menangani ribuan parameter *input* tanpa penghapusan, hal ini juga dapat menangani data hilang di dalam kumpulan data untuk melatih model prediktif (Lalwani dkk, 2021).

Breiman (2001), mendefinisikan bahwa *random forest* merupakan pengklasifikasi yang terdiri dari kumpulan klasifikasi pohon berstruktur $\{h(x,k), k = 1, \dots\}$ dengan k adalah vektor acak berdistribusi bebas yang identik dan setiap pohon memberikan keputusan untuk kelas mayoritas pada input x.

2.8 Evaluasi Model

Indikator penilaian merupakan hal yang penting untuk mengevaluasi kinerja setiap metode pada *machine learning.* Ada banyak indikator penilaian dalam bidang klasifikasi salah satunya adalah *confusion matrix.* Saputro dan Sari (2019) mengatakan bahwa, *confusion matrix* menggambarkan performa model melalui tabel. Setiap baris dari matriks tersebut mempresentasikan klasifikasi aktual dari data, dan setiap kolom dari matriks tersebut mempresentasikan klasifikasi prediksi dari data atau sebaliknya.

Tabel 3. *Confusion Matrix*

	Kelas Prediksi Positif	Kelas Prediksi Negatif
Kelas Aktual Positif	*True Positive* (TP)	*False Negative* (FN)
Kelas Aktual Negatif	*False Positive* (FP)	*True Negative* (TN)

1. *True Postive* (TP) adalah data diprediksi positif dan data sebenarnya adalah positif.
2. *True Negatif* (TN) adalah data diprediksi negatif dan data sebenarnya adalah negatif.
3. *False Positif* (FP) adalah data diprediksi positif dan data sebenarnya adalah negatif.
4. *False Negatif* (FN) adalah data diprediksi negatif dan data sebenarnya adalah positif.

Dari penyajian tabel *confusion matrix* tersebut, dapat dilakukan perhitungan untuk mengetahui nilai *accuracy, precision, recall,* dan *f1-score* sebagai berikut (Sokolova dan Lapalme, 2009) :

1. *Accuracy* merupakan efektivitas keseluruhan dari hasil klasifikasi

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \quad (2.6)$$

2. *Precision* merupakan presentase dari label data dengan label positif yang diberikan oleh klasifikasi

$$Precision = \frac{TP}{TP+FP} \quad (2.7)$$

3. *Recall* merupakan efektivitas dari pengklasifikasi dalam mengidentifikasi label positif

$$Recall = \frac{TP}{TP+FN} \quad (2.8)$$

4. *F1-score* merupakan hubungan antara data berlabel positif dari hasil klasifikasi yang menunjukan keseimbangan antara *precision* dan *recall*

$$F_1 = 2\frac{precision\,.\,recall}{precision+\,recall} \quad (2.9)$$

III. METODOLOGI PENELITIAN

3.1 Waktu dan Tempat Penelitian

Penelitian ini dilakukan pada semester genap tahun akademik 2021/2022, bertempat di Jurusan Matematika, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Lampung.

3.2 Data Penelitian

Data yang digunakan dalam penelitian ini adalah data sekunder yaitu data *bank customer* yang diperoleh dari *https://www.kaggle.com/datasets/mathchi/churn-for-bank-customers* data di *download* secara langsung dengan *extension csv.* Jumlah data yang digunakan pada penelitian ini sebanyak 10.000 data *bank customer* dan terdapat 14 variabel yaitu *row number, customer id, surname, credit score, geography, gender, age, tenure, balance, num of products, has credit card, is active number, estimated salary,* dan *exited.* Variabel *exited* merupakan variabel target yang terdiri dari 2 klasifikasi yaitu *churn customer* yang dilambangkan dengan "1" berjumlah 2037 *customer* dan *not churn customer* yang dilambangkan dengan "0" berjumlah 7963 *customer.*

	RowNumber	CustomerId	Surname	CreditScore	Geography	Gender	Age	Tenure	Balance	NumOfProducts	HasCrCard	IsActiveMember	EstimatedSalary	Exited
0	1	15634602	Hargrave	619	France	Female	42	2	0.00	1	1	1	101348.88	1
1	2	15647311	Hill	608	Spain	Female	41	1	83807.86	1	0	1	112542.58	0
2	3	15619304	Onio	502	France	Female	42	8	159660.80	3	1	0	113931.57	1
3	4	15701354	Boni	699	France	Female	39	1	0.00	2	0	0	93826.63	0
4	5	15737888	Mitchell	850	Spain	Female	43	2	125510.82	1	1	1	79084.10	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
9995	9996	15606229	Obijiaku	771	France	Male	39	5	0.00	2	1	0	96270.64	0
9996	9997	15569892	Johnstone	516	France	Male	35	10	57369.61	1	1	1	101699.77	0
9997	9998	15584532	Liu	709	France	Female	36	7	0.00	1	0	1	42085.58	1
9998	9999	15682355	Sabbatini	772	Germany	Male	42	3	75075.31	2	1	0	92888.52	1
9999	10000	15628319	Walker	792	France	Female	28	4	130142.79	1	1	0	38190.78	0

10000 rows × 14 columns

Gambar 1. Data *Bank Customer*

3.3 Metode Penelitian

Langkah-langkah yang dilakukan pada penelitian ini adalah sebagai berikut:

1. Melakukan *input* data *bank customer* yang diperoleh dari *kaggle* ke dalam bahasa pemrograman *python* menggunakan *google colab.*
2. Melakukan visualisasi data, visualisasi data digunakan untuk menggambarkan dan juga mendeskripsikan data yang digunakan. Pada tahap ini ditampilkan *pie chart* untuk menunjukan perbandingan jumlah antara *customer* yang keluar dan bertahan, *bar chart* yang menampilkan informasi *customer* berdasarkan variabel yang ada.
3. Melakukan *preprocessing* data, yaitu :

   Tahap yang dilakukan pada *preprocessing* meliputi:

   a. *Cleansing Data*,yaitu memastikan data tidak memiliki data hilang dan data duplikasi.
   b. *Handling Data Categorical*, yaitu memberikan label pada data dalam bentuk kategorik dengan menggunkan *one hot encoding* agar lebih mudah dipahami oleh komputer.
   c. *Scaling data*, yaitu melakukan transformasi data menggunakan *standardscaler.*
4. *Handling Imbalanced Data*

   Menangani *imbalanced data* dengan menggunakan *Synthetic Minority Over-Sampling Technique* (SMOTE), SMOTE membuat sintetik data dimana data yang dibuat merupakan data dari tetangga terdekat dari kelas minoritas.

5. Melakukan pembagian data dengan 4 skema yaitu skema 60% data *training* dan 40% data *testing*, skema 70% data *training* dan 30% data *testing*, skema 80% data *training* dan 20% data *testing*, skema 90% data *training* dan 10% data *testing.*
6. Membangun model *random forest*, dengan menggunakan *hyperparameter tuning* yang berguna untuk memaksimalkan pemilihan parameter terbaik, dan melakukan prediksi.
7. Evaluasi Model

   Pada tahap ini model yang sudah dibuat selanjutnya diuji untuk melihat seberapa baik performa model yang dihasilkan dengan *confusion matrix.*

V. KESIMPULAN

Setelah melakukan proses *machine learning* dengan menggunakan algoritma *random forest* pada klasifikasi *churn customer,* dapat diambil kesimpulan sebagai berikut:

1. Aspek yang digunakan pada penelitian ini diantaranya, SMOTE untuk menangani *imbalanced data,* pembagian data 90% untuk data *training* dan 10% data *testing* menjadi skema terbaik untuk pembagian data, *hyperparameter tuning* untuk menentukan parameter terbaik yang digunakan pada model *random forest* pada klasifikasi *churn customer* yaitu 400 *n-estimator,* dan 40 *maxdepth* .
2. Menggunakan Metode SMOTE untuk menangani *imbalanced data* dapat meningkatkan peforma model *random forest.*
3. Hasil klasifikasi data *customer* untuk memprediksi *churn customer* menggunakan metode *random forest* dengan pembagian data 90% untuk data *training* dan 10% data *testing* menghasilkan nilai *accuracy* sebesar 90,83%, maka metode ini baik dapat melakukan klasifikasi *churn customer* dengan baik.

DAFTAR PUSATAKA

Ataei, A., dan Osanloo, M. 2004. Using A Combination Of Genetic Algorithm and The Grid Searchmethod to Determine Optimum Cutoff Grades Of Multiple Metal Deposits. *International Journal of Surface Mining*. **18**(1): 60-78

Alfaro, E., Gamez, M., dan Garcia, N. 2013. An R Package for Classification with Boosting and Bagging. *Journal of Statistical Software.* **54**(32) : 11-35.

Benediktus, N., dan Oetama, R S. 2020. Algoritma Klasifikasi Decision Tree C5.0 untuk Memprediksi Performa Akademik Siswa. *Jurnal Teknik Informatika.* **12**(1) : 14-19.

Breiman, L. 1996. *Bagging Predictors*. Kluwer Academic Publishers. Boston.

Breiman, L. 2001. *Machine Learning*. Kluwer Academic Publishers. Netherlands.

Chawla, N. A., Bowyer, K. W., Hall, L. O., dan Kegelmeyer, W. P. 2002. SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*. **16**(1): 321-357.

Davies, B. 2003. *Database System.* 3rd Ed. Basingstoke. Palgrave Macmillan.

Dean, J. 2014. *Big Data, Data Mining, and Machine Learning.* John Wiley & Sons . Hoboken:

Dhakate, P., Rajeswari, K., dan Abin, D. 2015. Analysis of Different for Medical Dataset using Various Measures. *International Journal of Computer Applications.* **111**(5): 20-24.

George, S., dan Sumathi, B. 2020. Grid Search Tuning of Hyperparameters in Random Forest Classifier for Customer Feedback Sentiment Prediction. *International Journal of Computer Science and Applications.* **11**(9): 173-178.

Gorunescu, F. 2011. *Data Mining : Concept, Model and Techniques*. Springer. Berlin

Han, J., Kamber, M., dan Pei, J. 2000. *Data Mining: Concepts and Techniques*. Morgan Kaufman. San Fransisco.

Hanifa,T. T., Adiwijaya., dan Faraby,S. 2017. Analisis Churn Prediction pada Data Pelanggan PT. Telekomunikasi dengan Logistic Regression dan Underbagging. *eProceeding of Engineering.* **4**(2): 3210-3225.

Hao, M., Wang, Y., dan Bryant, S. H. 2014. An Efficient Algorithm Coupled with Synthetic Minority Over-sampling Technique to Classify Imbalanced PubChem BioAssay Data. *Analytica Chimica Acta*. **8**(60): 117-127.

Hapsari, K.R., dan Indriyani, T. Implementasi Algoritma SMOTE Sebagai Penyelesaian Imbalance Hight Dimensional Datasets. *Prosiding Seminar Nasional Teknik Elektro, Sistem Informasi, dan Teknik Informatika*, Surabaya: 26 Maret 2022. Hal. 427-432.

Lalwani, P., Mishra, M. K., dan Chadha, J. S. 2021. Customer Churn Prediction System: a Machine Learning Approarch. *Springer*. **10**(4): 271-294.

Li, W., dan Liu, Z. 2011. A Method of SVM with Normalization in Intrrusion Detection. *Elsevier*. **11**: 256-262

Oktanisa, I., dan Supianto, A. A. 2018. Perbandingan Teknik Klasifikasi Dalam Data Mining untuk Bank Direct Marketing. *Jurnal Teknologi Informasi dan Ilmu Komputer*. **5**(5): 567-576.

Pamina, J., Beschi, R. J., Sathya, B. S., Soundarya, S., Sruthi, M. S., dan Kiruthika, S. 2019. An Effective Classifier for Predicting Churn in Telecommunication. *Journal of Advanced Research in Dynamical and Control Systems*. **11**(1): 221-229.

Prasetyo, V.R., Mercifia,M., Averina, A., Sunyoto, L., dan Budiarjo. 2022. Prediksi Rating Film pada Website IMDB Menggunakan Metode Neural Network. *Jurnal Ilmiah NERO.* **7**(1): 1-8.

Ratner, B. 2011. *Statistical and Machine Learning Data Mining*. CRC Press. Florida.

Religia, Y., Nugroho, A., dan Hadikristanto, W. 2021. Analisis Perbandingan Algoritma Optimasi pada Random Forest untuk Klasifikasi Data Bank Marketing. *Jurnal Rekayasa Sistem dan Teknologi Informasi*. **5**(1): 187-192

Saputro, I.W., dan Sari, B.W. 2019. Uji Performa Algoritma Naive Bayes untuk Prediksi Masa Studi Mahasiswa. *Citec Journal*. **6**(1): 1-11.

Sihombing, P. R., dan Arsani, A. M. 2021. Perbandingan Metode Machine Learning Dalam Klasifikasi Kemiskinan Di Indonesia Tahun 2018. *Jurnal Teknik Informatika*. **2**(1):51-56.

Siringoringo, R. 2018. Klasifikasi Data Tidak Seimbang Menggunakan Algoritma SMOTE dan K-NEAREST NEIGHBOR. *Jurnal Information System Development*. **3**(1): 44-49.

Sokolova, M., dan Lapalme, G. 2009. A Systematic Analysis of Performance Measures for Classification Tasks. *Information Processing and Management*. **45**(4): 427-437.

Sutton, C. D. 2005. Classification and Regression Trees, Bagging, and Boosting. *Handbook of Statistics.* **24**(1): 303-329.

Syarif, I., Zaluka, E., Prugel, A., dan Wills, G. 2012. *Application Of Bagging, Boosting and Stacking to Intrusion Detection.* Springer. Berlin.

Syukron, A., dan Subekti, A. 2018. Penerapan Metode Random Over-Under Sampling dan Random Forest untuk Klasifikasi Penilaian Kredit. *Jurnal Informatika.* **5**(2): 175-185.

The Royal Society. 2017. *Machine Learning: The Power and Promise of Computers That Learn by Example.* The Royal Society. London.

Utami, Y. T., Shofiana, D. A., dan Heningtyas, Y. 2020. Penerapan Algoritma C4.5 Untuk Prediksi Churn Rate Pengguna Jasa Telekomunikasi. *Jurnal Komputasi*. **8**(2): 69-76.

Vijayakumar, V., dan Nedunchezhian, R. 2012. A Study on Video Data Mining. *Springer.* **1**(10):153-172.

Vluymans, S. 2019. *Dealing with Imbalanced and Weakly Labelled Data in Machine Learning using Fuzzy and Rough Set Methods.* Springer. Belgia.

Winata, W., Dewi, P. L., dan Tjondrowiguno, N. A. 2020. Prediksi Skor Pertandingan Sepak Bola Menggunakan Neuroevolution of Augmenting Topologies dan Backpropagation. *Jurnal Infra.* **8**(1): 249-254