

**OPTIMASI MODEL *ARTIFICIAL NEURAL NETWORK* DALAM
SISTEM KLASIFIKASI *CUSTOMER CHURN* BERBASIS *WEB***

SKRIPSI



**PROGRAM STUDI TEKNIK INFORMATIKA
FAKULTAS SAINS DAN TEKNOLOGI
UNIVERSITAS BUDDHI DHARMA
TANGERANG
2024**

**OPTIMASI MODEL *ARTIFICIAL NEURAL NETWORK* DALAM
SISTEM KLASIFIKASI *CUSTOMER CHURN* BERBASIS *WEB***

SKRIPSI

**Diajukan sebagai salah satu syarat untuk kelengkapan gelar kesarjanaan pada
Program Studi Teknik Informatika
Jenjang Pendidikan Strata 1**



**WILLY WIJAYA
20201000054**

**PROGRAM STUDI TEKNIK INFORMATIKA
FAKULTAS SAINS DAN TEKNOLOGI
UNIVERSITAS BUDDHI DHARMA
TANGERANG
2024**

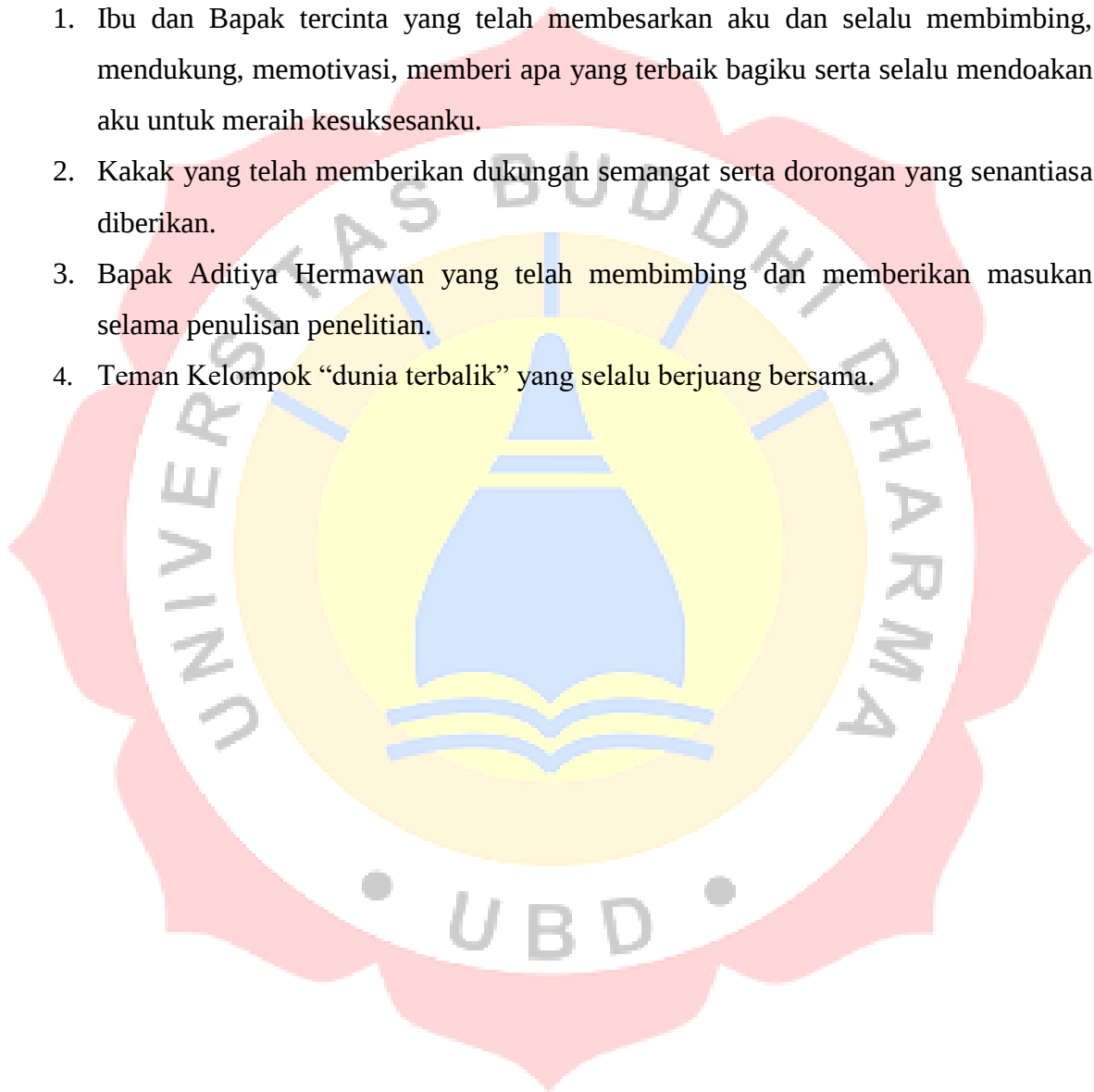
LEMBAR PERSEMBAHAN

“Life isn't fair pal, get used to it”

(Patrick Star)

Dengan mengucapkan puji syukur kepada Tuhan Yang Maha Esa, Skripsi ini kupersembahkan untuk:

1. Ibu dan Bapak tercinta yang telah membesarkan aku dan selalu membimbing, mendukung, memotivasi, memberi apa yang terbaik bagiku serta selalu mendoakan aku untuk meraih kesuksesanku.
2. Kakak yang telah memberikan dukungan semangat serta dorongan yang senantiasa diberikan.
3. Bapak Aditiya Hermawan yang telah membimbing dan memberikan masukan selama penulisan penelitian.
4. Teman Kelompok “dunia terbalik” yang selalu berjuang bersama.



UNIVERSITAS BUDDHI DHARMA
LEMBAR PERNYATAAN KEASLIAN SKRIPSI

Yang bertanda tangan dibawah ini.

NIM : 20201000054
Nama : Willy Wijaya
Jenjang Studi : Strata I
Program Studi : Teknik Informatika
Peminatan : Database

Dengan ini saya menyatakan bahwa:

1. Skripsi ini adalah asli dan belum pernah diajukan untuk mendapat gelar akademik (Diploma/Sarjana) atau kelengkapan studi, baik di Universitas Buddhi Dharma maupun di Perguruan Tinggi lainnya.
2. Skripsi ini saya buat sendiri tanpa bantuan dari pihak lain, kecuali arahan dosen pembimbing.
3. Dalam skripsi ini tidak terdapat karya atau pendapat yang telah ditulis atau dipublikasikan orang lain, kecuali secara tertulis dengan jelas dan dicantumkan sebagai acuan dalam naskah dengan disebutkan nama pengarang dan dicantumkan daftar pustaka.
4. Dalam Skripsi ini tidak terdapat pemalsuan (kebohongan), seperti buku, artikel, jurnal, data sekunder, pengolahan data, dan pemalsuan tanda tangan dosen atau Ketua Program Studi Universitas Buddhi Dharma yang dibuktikan dengan keasliannya.
5. Lembar pernyataan ini saya buat dengan sesungguhnya, tanpa paksaan dan apabila dikemudian hari atau pada waktu lainnya terdapat penyimpangan dan ketidakbenaran dalam pernyataan ini, saya bersedia menerima sanksi akademik berupa pencabutan gelar akademik yang telah saya peroleh karena Skripsi ini serta sanksi lainnya sesuai dengan peraturan dan norma yang berlaku.

Tangerang, 30 Juli 2024
Yang membuat pernyataan,



Willy Wijaya
20201000054

UNIVERSITAS BUDDHI DHARMA
LEMBAR PERSETUJUAN PUBLIKASI KARYA ILMIAH

Yang bertanda tangan dibawah ini.

NIM : 20201000054
Nama : Willy Wijaya
Jenjang Studi : Strata I
Program Studi : Teknik Informatika
Peminatan : Database

Dengan ini menyetujui untuk memberikan izin kepada pihak Universitas Buddhi Dharma, Hak Bebas Royalti Non – Eksklusif (*Non-exclusive Royalty Fee Right*) atas karya ilmiah kami yang berjudul: “OPTIMASI MODEL *ARTIFICIAL NEURAL NETWORK* DALAM SISTEM KLASIFIKASI *CUSTOMER CHURN* BERBASIS *WEB*”.

Dengan Hak Bebas Royalti Non – Eksklusif ini pihak Universitas Buddhi Dharma berhak menyimpan, mengalih-media atau format-kan, mengelolanya dalam pangkalan data (database), mendistribusikannya, dan menampilkan atau mempublikasikannya di internet atau media lain untuk kepentingan akademis tanpa perlu meminta izin dari saya selama tetap mencantumkan nama saya sebagai penulis atau pencipta karya ilmiah tersebut.

Saya bersedia untuk menanggung secara pribadi, tanpa melibatkan Universitas Buddhi Dharma, segala bentuk tuntutan hukum yang timbul atas pelanggaran Hak Cipta dalam karya ilmiah saya ini.

Demikian pernyataan ini saya buat dengan sebenarnya

Tangerang, 30 Juli 2024
Yang membuat pernyataan,



Willy Wijaya
20201000054

UNIVERSITAS BUDDHI DHARMA
LEMBAR PENGESAHAN PEMBIMBING
OPTIMASI MODEL *ARTIFICIAL NEURAL NETWORK* DALAM
SISTEM KLASIFIKASI *CUSTOMER CHURN* BERBASIS *WEB*

Dibuat oleh:

NIM : 20201000054

Nama : Willy Wijaya

Telah disetujui untuk dipertahankan dihadapan Tim Penguji Ujian
Komprehensif

Program Studi Teknik Informatika

Peminatan Database

Tahun Akademik 2023/2024

Disahkan oleh,

Tangerang, 30 Juli 2024

Pembimbing,



Aditiya Hermawan, S.Kom., M.Kom

NIDN. 0406128801

UNIVERSITAS BUDDHI DHARMA
LEMBAR PENGESAHAN SKRIPSI
OPTIMASI MODEL *ARTIFICIAL NEURAL NETWORK* DALAM
SISTEM KLASIFIKASI *CUSTOMER CHURN* BERBASIS *WEB*

Dibuat Oleh:

NIM : 20201000054

Nama : Willy Wijaya

Telah disetujui untuk dipertahankan di hadapan Tim Penguji Ujian
Komprehensif

Program Studi Teknik Informatika

Peminatan Database

Tahun Akademik 2023/2024

Disahkan oleh,

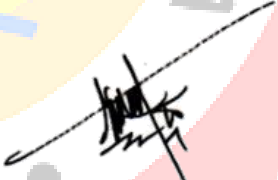
Tangerang, 30 Juli 2024

Dekan,

Ketua Program Studi,


Dr. Yakub, S.Kom., M.Kom., M.M

NIDN. 0304056901

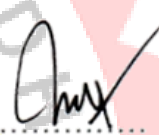
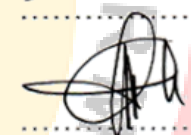

Hartana Wijaya, S.Kom, M.Kom

NIDN. 0412058102

UNIVERSITAS BUDDHI DHARMA
LEMBAR PENGESAHAN TIM PENGUJI

Nama : Willy Wijaya
NIM : 20201000024
Fakultas : Sains dan Teknologi
Judul Skripsi : OPTIMASI MODEL *ARTIFICIAL NEURAL NETWORK* DALAM
SISTEM KLASIFIKASI *CUSTOMER CHURN* BERBASIS *WEB*

Dinyatakan LULUS setelah mempertahankan di depan Tim Penguji Komprehensif pada hari
Selasa, 30 Juli 2024

	Nama penguji:	Tanda Tangan:
Ketua Sidang	Dr. Yakub, S.Kom., M.Kom., M.M NIDN. 0304056901	
Penguji I	Desiyanna Lasut, S.Kom., M.Kom NIDN. 0402128601	
Penguji II	Aditiya Hermawan, S.Kom., M.Kom NIDN. 0406128801	

Mengetahui,
Dekan Fakultas Sains dan Teknologi


Dr. Yakub, S.Kom., M.Kom., M.M
NIDN. 0304056901

KATA PENGANTAR

Dengan mengucapkan Puji Syukur kepada Tuhan Yang Maha Esa, yang telah memberikan Rahmat dan karunia-Nya kepada penulis sehingga dapat menyusun dan menyelesaikan Skripsi ini dengan judul **“OPTIMASI MODEL ARTIFICIAL NEURAL NETWORK DALAM SISTEM KLASIFIKASI CUSTOMER CHURN BERBASIS WEB”**. Tujuan utama dari pembuatan Skripsi ini adalah sebagai salah satu syarat kelengkapan dalam menyelesaikan program pendidikan Strata 1 Program Studi Teknik Informatika di Universitas Buddhi Dharma. Dalam penyusunan Skripsi ini penulis banyak menerima bantuan dan dorongan baik moril maupun materiil dari berbagai pihak, maka pada kesempatan ini penulis menyampaikan rasa terima kasih yang sebesar-besarnya kepada:

1. Ibu Dr. Limajatini, SE., MM., BKP, sebagai Rektor Universitas Buddhi Dharma
2. Bapak Dr. Yakub, S.Kom., M.Kom., M.M., sebagai Dekan Fakultas Sains dan Teknologi
3. Bapak Rudy Arijanto, S.Kom., M.Kom., sebagai Wakil Dekan Fakultas Sains dan Teknologi
4. Bapak Hartana Wijaya, S.Kom., M.Kom., sebagai Ketua Program Studi Teknik Informatika
5. Bapak Aditiya Hermawan, S.Kom., M.Kom., sebagai pembimbing yang telah membantu dan memberikan dukungan serta harapan untuk menyelesaikan penulisan Skripsi ini.
6. Orang tua dan keluarga yang selalu memberikan dukungan baik moril dan materiil.
7. Teman-teman yang selalu membantu dan memberikan semangat

Serta semua pihak yang terlalu banyak untuk disebutkan satu-persatu sehingga terwujudnya penulisan ini. Penulis menyadari bahwa penulisan Skripsi ini masih belum sempurna, untuk itu penulis mohon kritik dan saran yang bersifat membangun demi kesempurnaan penulisan di masa yang akan datang.

Akhir kata semoga ini dapat berguna bagi penulis khususnya dan bagi para pembaca yang berminat pada umumnya.

Tangerang, 30 Juli 2024

Penulis

ABSTRAK

Pelanggan merupakan salah satu aset penting dalam dunia bisnis yang dinamis dan pasar yang kompetitif. Namun, mereka dapat dengan mudah berpindah ke layanan lain, yang biasa disebut *Customer Churn*. Salah satu alasan utama churn adalah ketidakpuasan pelanggan terhadap layanan yang diberikan. Bagi banyak perusahaan, khususnya di sektor telekomunikasi, *customer churn* merupakan masalah besar karena berdampak langsung pada pendapatan perusahaan. Salah satu rencana strategi yang paling menguntungkan adalah mempertahankan pelanggan untuk terus berlangganan karena strategi ini relatif murah dibandingkan mendapatkan pelanggan baru yang biayanya 6 hingga 10 kali lipat. Namun, terciptalah tantangan baru yaitu bagaimana cara untuk mengetahui pelanggan mana saja yang memiliki potensi churn sehingga dapat dipertahankan. Oleh karena itu, banyak peneliti telah membuat model untuk melakukan klasifikasi pelanggan mana yang akan churn, salah satunya menggunakan *Artificial Neural Network*. Penelitian ini bertujuan untuk mengoptimalkan model agar mendapatkan hasil yang lebih baik dengan cara membuat model baru berdasarkan dari penelitian sebelumnya dan menambahkan beberapa teknik pengolahan data. Maka dari itu, penelitian ini menggunakan *Artificial Neural Network* sebagai model algoritmanya, *ROS* dan *SMOTE* sebagai teknik penyeimbangan datanya, dan beberapa algoritma lainnya yaitu *XGBoost*, *Lasso Regression* dan *mRMR* untuk pemilihan fitur. Hasil akurasi model pada penelitian ini mencapai 88.50% dan cukup memuaskan karena meningkat sebesar 0.38% dari model penelitian sebelumnya yang mencapai 88.12%.

Kata Kunci: *Customer Churn, Classification, ANN, ROS, XGBoost*

ABSTRACT

Customers are an important asset in the dynamic business world and competitive market. However, they can easily move to another service, which is commonly called Customer Churn. One of the main reasons for churn is customer dissatisfaction with the services provided. For many companies, especially in the telecommunications sector, Customer Churn is a big problem because it has a direct impact on company revenue. One of the most profitable strategic plans is to retain customers to continue subscribing because this strategy is relatively cheap compared to getting new customers which costs 6 to 10 times more. However, a new challenge is created, namely how to find out which customers have the potential to churn so they can be retained. Therefore, many researchers have created models to carry out classification to predict which customers will churn, one of which uses Artificial Neural Networks. This research aims to optimize the model to get better results by creating a new model based on previous research and adding several data processing techniques. Therefore, this research uses Artificial Neural Networks as the algorithm model, ROS and SMOTE as data balancing techniques, and several other algorithms, namely XGBoost, Lasso Regression and mRMR for Feature Selection. The accuracy results of the model in this study reached 88.50% and were quite satisfactory because they increased by 0.38% from the previous research model which reached 88.12%.

Keyword: Customer Churn, Classification, ANN, ROS, XGBoost



DAFTAR ISI

LEMBAR JUDUL LUAR SKRIPSI

LEMBAR JUDUL DALAM SKRIPSI

LEMBAR PERSEMBAHAN

LEMBAR PERNYATAAN KEASLIAN SKRIPSI

LEMBAR PERSETUJUAN PUBLIKASI ILMIAH

LEMBAR PENGESAHAN PEMBIMBING

LEMBAR PENGESAHAN SKRIPSI

LEMBAR PENGESAHAN TIM PENGUJI

KATA PENGANTAR.....	vii
ABSTRAK	viii
ABSTRACT	ix
DAFTAR ISI	x
DAFTAR TABEL	xiv
DAFTAR GAMBAR.....	xvi
BAB I PENDAHULUAN	1
1.1 Latar Belakang Masalah	1
1.2 Identifikasi Masalah	4
1.3 Ruang Lingkup	4
1.4 Tujuan dan Manfaat.....	5
1.5.1 Tujuan.....	5
1.5.2 Manfaat.....	5
1.5 Sistematika Penulisan.....	5
BAB II LANDASAN TEORI.....	7
2.1 Teori Umum	7
2.1.1 Data.....	7
2.1.2 <i>Data Mining</i>	7
2.1.3 <i>Machine Learning</i>	8
2.1.4 <i>Website</i>	8
2.2 Teori Khusus	9
2.2.1 <i>Customer Churn</i>	9
2.2.2 <i>Artificial Neural Network</i>	9
2.2.3 <i>SMOTE</i>	10
2.2.4 <i>ROS</i>	10

2.2.5	<i>Lasso Regression</i>	11
2.2.6	<i>mRMR</i>	11
2.2.7	<i>XGBoost</i>	12
2.3	Teori Rancangan	12
2.3.1	<i>Python</i>	12
2.3.2	<i>Pandas</i>	13
2.3.3	<i>NumPy</i>	13
2.3.4	<i>Matplotlib</i>	14
2.3.5	<i>Seaborn</i>	14
2.3.6	<i>Scikit-learn</i>	15
2.3.7	<i>Imbalanced-learn</i>	15
2.3.8	<i>Flask</i>	15
2.4	Teori Pengujian	16
2.4.1	<i>Black Box Testing</i>	16
2.4.2	<i>Confusion Matrix</i>	16
2.4.3	<i>Kurva ROC-AUC</i>	18
2.5	Tinjauan Studi	18
2.5.1	<i>Customer Churn prediction in telecom using Machine Learning in big data platform</i>	18
2.5.2	<i>Customer Churn prediction in telecommunication industry using data certainty</i>	19
2.5.3	<i>Customer Churn prediction system: a Machine Learning approach</i>	20
2.5.4	<i>Customer Churn Prediction using Artificial Neural Networks (ANN) in Telecom Industry</i>	21
2.5.5	<i>Improvised_XGBoost Machine Learning Algorithm for Customer Churn Prediction</i>	21
2.5.6	<i>Integrated Churn Prediction and Customer Segmentation Framework for Telco Business</i>	22
2.5.7	<i>B2C E-commerce Customer Churn Prediction Based on K-Means and SVM</i>	22
2.5.8	<i>Customer Churn prediction in telecommunication industry using deep learning</i>	23

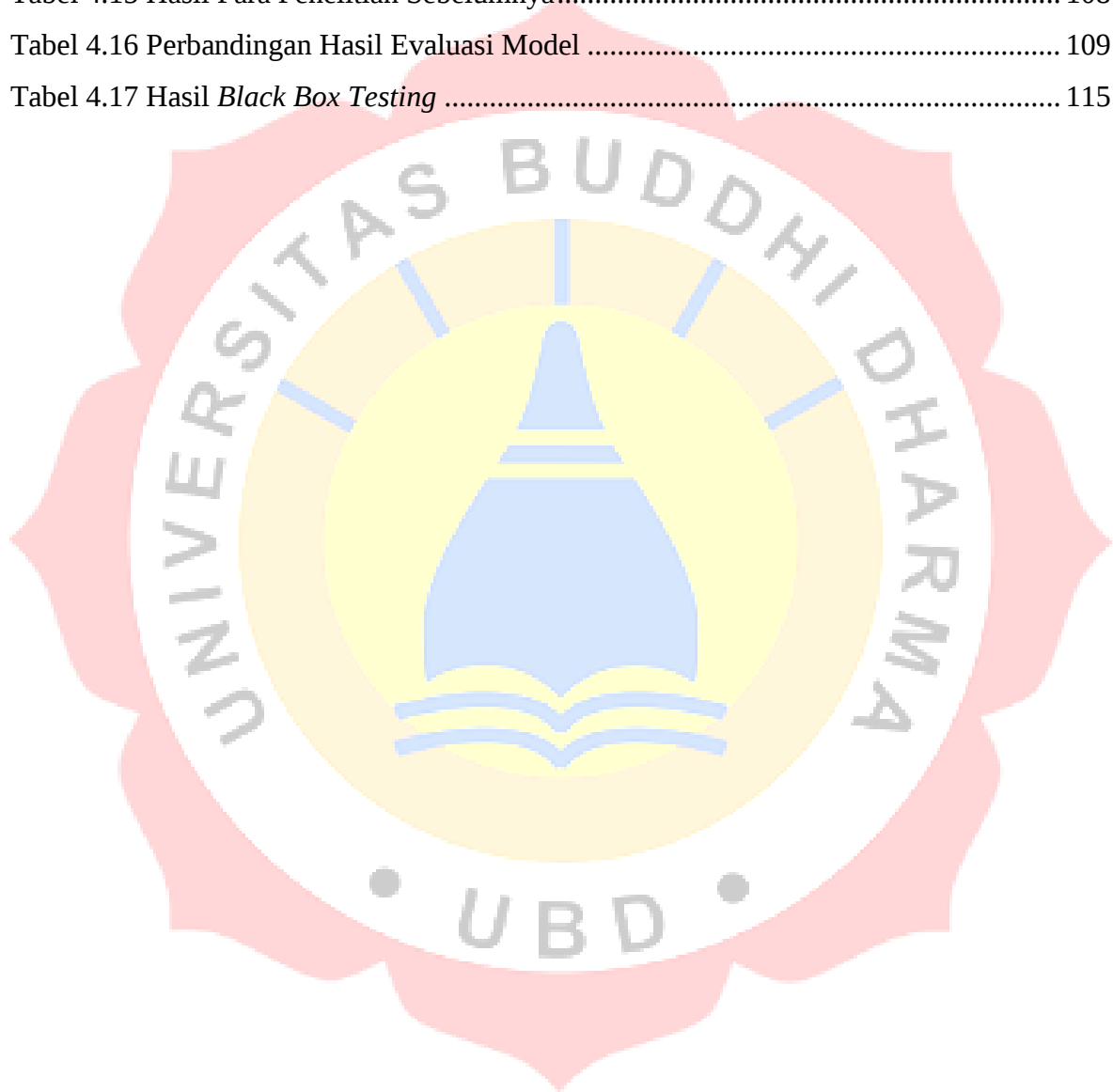
2.5.9	<i>Model Optimization Analysis of Customer Churn Prediction Using Machine Learning Algorithms with Focus on Feature Reduction</i>	24
2.5.10	<i>Customer Churn analysis using XGBoosted decision trees</i>	24
2.5.11	<i>Impact of Hyperparameters on Deep learning Model for Customer Churn Prediction in Telecommunication Sector</i>	25
2.5.12	<i>Customer Churn Prediction in Telecommunication Industry Using Classification and Regression Trees and Artificial Neural Network Algorithms</i>	25
2.5.13	<i>Just-in-time Customer Churn prediction in the telecommunication sector</i>	26
2.5.14	<i>Machine Learning Based Customer Churn Prediction in Banking.</i>	26
2.5.15	<i>Predicting Customer Churn using Artificial Neural Network</i>	27
2.5.16	Rangkuman Model Penelitian	28
2.6	Kerangka Pemikiran	58
BAB III METODOLOGI PENELITIAN		59
3.1	Pengumpulan Data	59
3.2	Pengolahan Data	59
3.2.1	<i>Data Preparation</i>	60
3.2.2	<i>Data Transformation</i>	61
3.2.3	<i>Data Normalization</i>	66
3.2.4	<i>Missing Value</i>	67
3.2.5	<i>Balancing Dataset</i>	70
3.2.6	<i>Feature Selection</i>	76
3.3	Pelatihan Model	77
3.3.1	<i>Splitting Data</i>	77
3.3.2	<i>Flowchart Pelatihan Model</i>	78
3.3.3	Perangkat Keras	78
3.3.4	Lingkungan Pengembangan	79
3.3.5	Konfigurasi Model	79
3.3.6	Kompilasi Model	81
3.4	Metode Evaluasi	87
3.4.1	<i>Confusion Matrix</i>	87
3.4.2	Evaluasi Matrix	87
BAB IV HASIL PEMBAHASAN		88

4.1	Evaluasi Model	88
4.1.1	<i>Holdout ANN-SMOTE-LassoReg</i>	88
4.1.2	<i>Holdout ANN-SMOTE-mRMR</i>	90
4.1.3	<i>Holdout ANN-SMOTE-XGBoost</i>	91
4.1.4	<i>Holdout ANN-ROS-LassoReg</i>	93
4.1.5	<i>Holdout ANN-ROS-mRMR</i>	94
4.1.6	<i>Holdout ANN-ROS-XGBoost</i>	96
4.1.7	<i>10-Fold ANN-SMOTE-LassoReg</i>	97
4.1.8	<i>10-Fold ANN-SMOTE-mRMR</i>	99
4.1.9	<i>10-Fold ANN-SMOTE-XGBoost</i>	100
4.1.10	<i>10-Fold ANN-ROS-LassoReg</i>	102
4.1.11	<i>10-Fold ANN-ROS-mRMR</i>	103
4.1.12	<i>10-Fold ANN-ROS-XGBoost</i>	105
4.2	Perbandingan Studi.....	108
4.3	Implementasi Web	109
4.3.1	Halaman Utama	113
4.3.2	Halaman Hasil	114
4.4	<i>Black Box Testing</i>	115
4.5	Pembahasan	116
BAB V	SIMPULAN DAN SARAN.....	118
5.1	Simpulan.....	118
5.2	Saran	118
	DAFTAR PUSTAKA.....	120
	LAMPIRAN	
	DAFTAR RIWAYAT HIDUP	

DAFTAR TABEL

Tabel 2.1 <i>Confusion Matrix</i>	17
Tabel 2.2 Kategori <i>AUC</i>	18
Tabel 3.1 Hasil Proses <i>One-Hot Encoding</i>	63
Tabel 3.2 Hasil Proses <i>Label Encoding</i>	65
Tabel 3.3 Data Sebelum Proses <i>MinMaxScaler</i>	67
Tabel 3.4 Data Setelah Proses <i>MinMaxScaler</i>	67
Tabel 3.5 Data Sebelum Proses <i>SimpleImputer</i>	69
Tabel 3.6 Data Setelah Proses <i>SimpleImputer</i>	70
Tabel 3.7 Data Sebelum Proses <i>SMOTE</i>	72
Tabel 3.8 Data Setelah Proses <i>SMOTE</i>	73
Tabel 3.9 Data Sebelum Proses <i>ROS</i>	75
Tabel 3.10 Data Setelah Proses <i>ROS</i>	76
Tabel 3.11 Konfigurasi Parameter Model <i>Artificial Neural Network</i>	80
Tabel 3.12 <i>Top 10 Feature (XGBoost + SMOTE + Holdout (90%, 10%))</i>	81
Tabel 3.13 <i>Top 10 Feature (XGBoost + ROS + Holdout (90%, 10%))</i>	81
Tabel 3.14 <i>Feature yang terpilih (Lasso Regression + SMOTE + Holdout (90%, 10%))</i> ..	82
Tabel 3.15 <i>Feature yang terpilih (Lasso Regression + ROS + Holdout (90%, 10%))</i>	82
Tabel 3.16 <i>Top 10 Feature (mRMR + SMOTE + Holdout (90%, 10%))</i>	83
Tabel 3.17 <i>Top 10 Feature (mRMR + ROS + Holdout (90%, 10%))</i>	83
Tabel 3.18 <i>Top 10 Feature (XGBoost + SMOTE + 10-Fold)</i>	84
Tabel 3.19 <i>Top 10 Feature (XGBoost + ROS + 10-Fold)</i>	84
Tabel 3.20 <i>Feature yang terpilih (Lasso Regression + SMOTE + 10-Fold)</i>	85
Tabel 3.21 <i>Feature yang terpilih (Lasso Regression + ROS + 10-Fold)</i>	85
Tabel 3.22 <i>Top 10 Feature (mRMR + SMOTE + 10-Fold)</i>	86
Tabel 3.23 <i>Top 10 Feature (mRMR + ROS + 10-Fold)</i>	86
Tabel 4.1 Hasil Evaluasi Model <i>Holdout ANN-SMOTE-LassoReg</i>	88
Tabel 4.2 Hasil Evaluasi Model <i>Holdout ANN-SMOTE-mRMR</i>	90
Tabel 4.3 Hasil Evaluasi Model <i>Holdout ANN-SMOTE-XGBoost</i>	91
Tabel 4.4 Hasil Evaluasi Model <i>Holdout ANN-ROS-LassoReg</i>	93
Tabel 4.5 Hasil Evaluasi Model <i>Holdout ANN-ROS-mRMR</i>	94
Tabel 4.6 Hasil Evaluasi Model <i>Holdout ANN-ROS-XGBoost</i>	96
Tabel 4.7 Hasil Evaluasi Model <i>10-Fold ANN-SMOTE-LassoReg</i>	97
Tabel 4.8 Hasil Evaluasi Model <i>10-Fold ANN-SMOTE-mRMR</i>	99

Tabel 4.9 Hasil Evaluasi Model <i>10-Fold ANN-SMOTE-XGBoost</i>	100
Tabel 4.10 Hasil Evaluasi Model <i>10-Fold ANN-ROS-LassoReg</i>	102
Tabel 4.11 Hasil Evaluasi Model <i>10-Fold ANN-ROS-mRMR</i>	103
Tabel 4.12 Hasil Evaluasi Model <i>10-Fold ANN-ROS-XGBoost</i>	105
Tabel 4.13 Perbandingan Hasil Evaluasi Model dengan Metode <i>Holdout</i>	106
Tabel 4.14 Perbandingan Hasil Evaluasi Model dengan Metode <i>10-Fold CV</i>	106
Tabel 4.15 Hasil Para Penelitian Sebelumnya.....	108
Tabel 4.16 Perbandingan Hasil Evaluasi Model	109
Tabel 4.17 Hasil <i>Black Box Testing</i>	115



DAFTAR GAMBAR

Gambar 2.1 <i>Artificial Neural Network Architecture</i>	10
Gambar 2.2 Kerangka Pemikiran	58
Gambar 3.1 Informasi Dataset Setelah <i>Data Preparation</i>	60
Gambar 3.2 <i>Flowchart</i> Proses <i>One-Hot Encoding</i>	61
Gambar 3.3 <i>Flowchart</i> Proses <i>Label Encoding</i>	64
Gambar 3.4 <i>Flowchart</i> Proses <i>MinMaxScaler</i>	66
Gambar 3.5 Proses <i>SimpleImputer</i>	68
Gambar 3.6 <i>Flowchart</i> Proses <i>SMOTE</i>	71
Gambar 3.7 <i>Flowchart</i> Proses <i>ROS</i>	74
Gambar 3.8 <i>Flowchart</i> Proses <i>Feature Selection</i>	77
Gambar 3.9 <i>Flowchart</i> Pelatihan Model	78
Gambar 3.10 Tampilan Lingkungan Pengembangan Model	79
Gambar 4.1 <i>Confusion Matrix Holdout ANN-SMOTE-LassoReg</i>	89
Gambar 4.2 Kurva ROC-AUC Holdout ANN-SMOTE-LassoReg	89
Gambar 4.3 <i>Confusion Matrix Holdout ANN-SMOTE-mRMR</i>	90
Gambar 4.4 Kurva ROC-AUC Holdout ANN-SMOTE-mRMR	91
Gambar 4.5 <i>Confusion Matrix Holdout ANN-SMOTE-XGBoost</i>	92
Gambar 4.6 Kurva ROC-AUC Holdout ANN-SMOTE-XGBoost	92
Gambar 4.7 <i>Confusion Matrix Holdout ANN-ROS-LassoReg</i>	93
Gambar 4.8 Kurva ROC-AUC Holdout ANN-ROS-LassoReg	94
Gambar 4.9 <i>Confusion Matrix Holdout ANN-ROS-mRMR</i>	95
Gambar 4.10 Kurva ROC-AUC Holdout ANN-ROS-mRMR	95
Gambar 4.11 <i>Confusion Matrix Holdout ANN-ROS-XGBoost</i>	96
Gambar 4.12 Kurva ROC-AUC Holdout ANN-ROS-XGBoost	97
Gambar 4.13 <i>Confusion Matrix 10-Fold ANN-SMOTE-LassoReg</i>	98
Gambar 4.14 Kurva ROC-AUC 10-Fold ANN-SMOTE-LassoReg	98
Gambar 4.15 <i>Confusion Matrix 10-Fold ANN-SMOTE-mRMR</i>	99
Gambar 4.16 Kurva ROC-AUC 10-Fold ANN-SMOTE-mRMR	100
Gambar 4.17 <i>Confusion Matrix 10-Fold ANN-SMOTE-XGBoost</i>	101
Gambar 4.18 Kurva ROC-AUC 10-Fold ANN-SMOTE-XGBoost	101
Gambar 4.19 <i>Confusion Matrix 10-Fold ANN-ROS-LassoReg</i>	102
Gambar 4.20 Kurva ROC-AUC 10-Fold ANN-ROS-LassoReg	103
Gambar 4.21 <i>Confusion Matrix 10-Fold ANN-ROS-mRMR</i>	104

Gambar 4.22 Kurva <i>ROC-AUC 10-Fold ANN-ROS-mRMR</i>	104
Gambar 4.23 <i>Confusion Matrix 10-Fold ANN-ROS-XGBoost</i>	105
Gambar 4.24 Kurva <i>ROC-AUC 10-Fold ANN-ROS-XGBoost</i>	106
Gambar 4.25 Perbandingan Hasil Evaluasi Model dengan Metode <i>Holdout</i>	107
Gambar 4.26 Perbandingan Hasil Evaluasi Model dengan Metode <i>10-Fold CV</i>	107
Gambar 4.27 Tampilan Halaman Utama	113
Gambar 4.28 Tampilan Halaman Utama (Bagian 2)	114
Gambar 4.29 Tampilan Halaman Hasil	114
Gambar 4.30 Tampilan Halaman Hasil (Bagian 2)	115



BAB I

PENDAHULUAN

1.1 Latar Belakang Masalah

Pelanggan merupakan salah satu aset penting dalam dunia bisnis yang dinamis dan pasar yang kompetitif, mereka dapat berpindah ke layanan lain dengan mudah dan hal itu biasanya disebut dengan *Customer Churn* (Amin et al., 2019; Bi, 2019; Xiahou & Harada, 2022). Salah satu alasan utama penyebab terjadinya churn adalah ketidakpuasan pelanggan terhadap layanan yang diberikan (Lalwani et al., 2022). Bagi banyak perusahaan khususnya di sektor telekomunikasi, *Customer Churn* merupakan masalah besar, karena berdampak langsung terhadap pendapatan perusahaan (Abdulsalam et al., 2022; Sina Mirabdolbaghi & Amiri, 2022; Vaudevan et al., 2022)

Untuk mengatasi masalah tersebut, salah satu rencana strategi yang paling menguntungkan adalah mempertahankan pelanggan untuk terus berlangganan (Ahmad et al., 2019). Menjaga loyalitas pelanggan yang sudah ada untuk terus berlangganan merupakan strategi yang cukup murah dibandingkan dengan mendapatkan pelanggan baru, karena mendapatkan pelanggan baru membutuhkan biaya 6 sampai 10 kali lebih banyak daripada mempertahankan pelanggan sehingga menghilangkan banyak keuntungan (Abdulsalam et al., 2022; Khan et al., 2019; Kumar & Kumar, 2019; Lalwani et al., 2022; Wu et al., 2021; Xiahou & Harada, 2022)

Menjaga loyalitas pelanggan bisa lebih efektif jika kita dapat memprediksi pelanggan yang berisiko churn. Hal tersebut menjadi tantangan untuk menentukan metode prediksi churn yang tepat. Salah satu metode yang terbukti cukup baik adalah dengan teknik *Machine Learning* yang sangat efektif khususnya untuk membuat model *Customer Churn prediction* karena dapat memperkirakan informasi berdasarkan data sebelumnya (Amin et al., 2020; Lalwani et al., 2022; Sina Mirabdolbaghi & Amiri, 2022; C. Wang et al., 2019)

Teknik *Machine Learning* telah luas digunakan dalam pengembangan model prediksi churn pelanggan dan menunjukkan hasil yang efektif. Namun, dengan bertambahnya volume data, terjadi pergeseran ke arah penggunaan teknik *deep learning* dalam beberapa tahun terakhir (Fujo et al., 2022). Menurut Fujo, Hal ini disebabkan oleh kemampuan *deep learning* untuk mengelola data dalam skala yang lebih besar dibandingkan dengan teknik *Machine Learning* konvensional. Istilah 'deep' dalam *deep learning* merujuk pada penggunaan jaringan saraf (*neural networks*) yang kompleks. *Deep learning* merupakan bagian dari *Artificial Neural Network* (ANN), yang karakteristiknya mencakup berbagai lapisan dalam Proses pemrosesan data.

Artificial Neural Network sudah banyak digunakan oleh beberapa penelitian sebelumnya untuk membuat model *Customer Churn prediction*, salah satunya penelitian yang membahas pembuatan model *Customer Churn prediction* dengan menggunakan *Artificial Neural Network* mendapatkan akurasi 79% yang dilakukan oleh (Khan et al., 2019). Penelitian lainnya yang membahas pembuatan model *Customer Churn prediction* menggunakan *deep learning* mendapatkan akurasi 88.12% dengan *Artificial Neural Network* yang dilakukan oleh (Fujo et al., 2022). Selain *Artificial Neural Network*, adapun penelitian yang membahas pembuatan model *Customer Churn prediction* menggunakan *K-Nearest Neighbor* (KNN) mendapatkan akurasi 83.97%, *Support Vector Machine* (SVM) mendapatkan akurasi 79.63%, *Decision Tree* (DT) mendapatkan 91.98%, *Random Forest* (RF) mendapatkan akurasi 95.74% yang dilakukan oleh (Rahman & Kumar, 2020).

Dalam pengembangan model *Customer Churn prediction* menggunakan *deep learning* ataupun *Machine Learning* yang telah dilakukan, terdapat masalah dalam Proses pengembangannya. Proses pengembangan model tersebut terdiri dari *Data Preparation*, *Data Preprocessing*, *Training Model*, dan Evaluasi Model. Permasalahan yang terdapat dalam Proses pengembangan model tersebut adalah hasil akurasi yang tidak maksimal

dikarenakan dataset yang kompleks dan bervariasi (Swetha & Dayananda, 2020). Oleh karena itu, pada Proses pengembangan model tersebut dilakukan beberapa Proses tambahan seperti *Feature Selection* untuk mengetahui fitur-fitur apa saja yang sangat berpengaruh terhadap hasil prediksinya dan *Balancing Dataset* untuk mengatasi masalah ketidakseimbangan dataset sehingga mendapatkan hasil yang maksimal (Xiahou & Harada, 2022).

Feature Selection dan teknik *Balancing Dataset* telah banyak digunakan oleh beberapa penelitian sebelumnya untuk meningkatkan akurasi klasifikasi. Salah satu penelitian yang dilakukan oleh (Lalwani et al., 2022) menggunakan Gravitational Search Algorithm (GSA) untuk *Feature Selection* tanpa melakukan *Balancing Dataset* dan hasil akurasinya cukup baik yaitu sekitar 80-81%, tetapi tidak disebutkan perbandingan hasil tanpa menggunakan *Feature Selection* dan yang menggunakan *Feature Selection*. Penelitian lainnya yang dilakukan oleh (Rahman & Kumar, 2020) menggunakan *mRMR* dan *Relief-F* untuk *Feature Selection*, *Oversampling* untuk *Balancing Dataset* dan hasil akurasinya meningkat dari 81.65% (tanpa *Feature Selection*) menjadi 83.97% dengan menggunakan *mRMR* dan 82.15% dengan menggunakan *Relief-F* pada algoritma KNN, sedangkan untuk beberapa algoritma lainnya seperti SVM tidak mengalami perubahan akurasi untuk kedua *Feature Selection* yang digunakan, *Decision Tree* dan *Random Forest* mengalami penurunan akurasi untuk kedua *Feature Selection* yang digunakan. Sementara itu, hasil dari Proses *Balancing Dataset* menggunakan teknik *Oversampling* setelah dilakukannya *Feature Selection* maupun sebelum dilakukannya *Feature Selection* membuat penurunan akurasi untuk semua algoritma yang digunakan. Adapun penelitian yang dilakukan oleh (Fujo et al., 2022) menggunakan *Lasso Regression* untuk *Feature Selection* dan *Random Oversampling* untuk *Balancing Dataset* pada algoritma ANN dan mendapatkan hasil akurasi sekitar 86% - 88% tetapi tidak disebutkan perbandingan hasil tanpa menggunakan *Feature Selection* dan

yang menggunakan *Feature Selection*. Oleh karena itu, pada penelitian ini akan dilakukan perbandingan hasil yang didapatkan oleh beberapa *Feature Selection* dan teknik *Balancing Dataset* yang digunakan, sehingga dapat mengetahui akurasi model terbaik dari beberapa kombinasi metode.

Berdasarkan latar belakang yang telah diuraikan, penelitian ini ditetapkan dengan judul **“OPTIMASI MODEL ARTIFICIAL NEURAL NETWORK DALAM KLASIFIKASI CUSTOMER CHURN BERBASIS WEB”**

1.2 Identifikasi Masalah

Berdasarkan latar belakang masalah di atas, dapat diidentifikasi masalah sebagai berikut:

- a. Membutuhkan solusi untuk mempertahankan pelanggan yang sudah ada karena menarik pelanggan baru membutuhkan biaya yang lebih tinggi.
- b. Terdapat beberapa *Feature Selection* dan teknik *Balancing Dataset* yang dapat digunakan untuk membentuk model *Customer Churn prediction* menggunakan *Artificial Neural Network*

1.3 Ruang Lingkup

Ruang lingkup penelitian ini adalah sebagai berikut:

- a. Program dibuat berbasis *Website*
- b. Dataset yang digunakan adalah dataset sekunder yang didapatkan dari kaggle <https://www.kaggle.com/code/ahmedshahriarsakib/churn-prediction-i-eda-statistical-analysis/input>
- c. Dataset berisi 7043 sampel dan 21 atribut
- d. Informasi yang dihasilkan berupa persentase dari klasifikasi churn atau *non-churn*

1.4 Tujuan dan Manfaat

1.5.1 Tujuan

Tujuan yang ingin dicapai dalam penelitian ini adalah sebagai berikut:

- a. Membuat model *Machine Learning* yang dapat melakukan klasifikasi pelanggan dengan menggunakan algoritma *deep learning*
- b. Membuat aplikasi *Web* yang dapat membantu untuk mengetahui pelanggan mana saja yang berpotensi churn atau tidak
- c. Mengetahui akurasi model dengan kombinasi beberapa *Feature Selection* dan *Balancing Dataset*

1.5.2 Manfaat

Manfaat yang dapat diperoleh dalam penelitian ini adalah sebagai berikut:

- a. Membantu untuk klasifikasi *Customer Churn* secara otomatis
- b. Membantu untuk mengambil keputusan yang lebih baik dan akurat berdasarkan informasi yang diperoleh dari model prediksi churn
- c. Membantu meningkatkan akurasi model untuk klasifikasi *Customer Churn*

1.5 Sistematika Penulisan

Sistematika penulisan bertujuan untuk mempermudah pemahaman dan penelaahan penelitian. Dalam laporan penelitian ini, materi yang tertulis dikelompokkan menjadi beberapa sub bab yang secara garis besar dapat dijelaskan sebagai berikut:

BAB I PENDAHULUAN

Bab ini berisi uraian terkait latar belakang masalah, identifikasi masalah, rumusan masalah, ruang lingkup penelitian, tujuan dan manfaat penelitian, metode penelitian dan sistematika penulisan.

BAB II LANDASAN TEORI

Bab ini berisi uraian terkait teori-teori yang mendasari penelitian ini yang dibagi menjadi tiga bagian yaitu teori umum, teori khusus, dan teori rancangan, tinjauan studi, dan kerangka pemikiran.

BAB III METODOLOGI PENELITIAN

Bab ini berisi uraian terkait metode yang digunakan dalam penelitian.

BAB IV HASIL PEMBAHASAN

Bab ini berisi uraian terkait evaluasi model yang telah dikembangkan, tampilan program, dan pengujian aplikasi.

BAB V SIMPULAN DAN SARAN

Bab ini berisi uraian terkait simpulan penelitian berdasarkan hasil pengujian dan implementasi yang telah dilakukan dan saran-saran yang dapat dikembangkan untuk penelitian selanjutnya

BAB II

LANDASAN TEORI

2.1 Teori Umum

2.1.1 Data

Data merupakan kumpulan informasi, fakta, atau observasi dengan berbagai bentuk seperti nominal, ordinal, interval atau ratio yang dapat digunakan untuk membuat penelitian, mengambil keputusan dan tujuan lainnya (Umar Hussain, 2023). Data dibagi menjadi 3 kategori yaitu data terstruktur, data semi-terstruktur dan data tidak terstruktur. Data terstruktur adalah data yang dapat diakses menggunakan *query* oleh manusia dan mesin, Data semi-terstruktur adalah data seperti JSON dan XML, dan Data tidak terstruktur adalah data yang tidak dapat dicari menggunakan *query* karena tidak mengikuti model yang telah ditentukan (Breitinger & Jotterand, 2023).

2.1.2 Data Mining

Data Mining adalah bagian dari *Artificial intelligence* yang memiliki tujuan untuk menghasilkan informasi dari data dan menyajikannya secara komprehensif dengan tujuan menghasilkan pengetahuan seperti penemuan pola, hubungan, dan tren baru yang dapat berguna bagi pengguna (Schuh et al., 2019). *Data Mining* ahli dalam menggunakan teknologi untuk mencari pola yang kemudian digunakan untuk meramal tren masa depan sehingga memungkinkan pengguna untuk mengambil keputusan (Salem, 2022).

Data Mining juga dapat dikatakan sebagai suatu Proses yang melibatkan pengumpulan dan pemilihan data, pra-pemrosesan data dan analisis data menggunakan metode *Machine Learning*, visualisasi hasil, interpretasi temuan dan implementasi pengetahuan (Schuh et al., 2019).

2.1.3 *Machine Learning*

Machine Learning adalah program komputer yang kinerjanya meningkat seiring waktu tanpa perlu pemrograman eksplisit dan bertujuan untuk otomatisasi pembuatan model analitis yang dicapai dengan menerapkan algoritma secara berulang sehingga memungkinkan komputer menemukan wawasan tersembunyi dan pola yang kompleks (Sharma & Chaudhary, 2023).

Machine Learning juga dapat disebut sebagai bidang penelitian penting dalam *artificial intelligence* yang membantu komputer dalam membuat suatu model berdasarkan pengalaman dan dapat memprediksi kejadian di masa depan secara akurat (Dogan & Birant, 2021).

Machine Learning dibagi menjadi 2 kategori yaitu *Supervised Learning* seperti permasalahan klasifikasi atau menggunakan label dan *Unsupervised Learning* seperti permasalahan klustering atau tanpa menggunakan label. *Supervised Learning* adalah cabang *Machine Learning* yang bertujuan untuk mempelajari pemetaan antara *input* dan *output*. *Unsupervised Learning* adalah *Machine Learning* yang digunakan untuk mengidentifikasi keteraturan dan ketergantungan pada data yang tidak memiliki label (Dogan & Birant, 2021).

2.1.4 *Website*

Website adalah kumpulan dokumen yang berupa halaman *Web* yang berisi teks dalam bentuk format Hyper Text Markup Language (HTML) dan disimpan di *server hosting* yang dapat diakses menggunakan browser dengan jaringan internet melalui Uniform Resource Locator (URL) (Widia & Asriningtias, 2021).

Website dibagi menjadi 2 jenis yaitu *Website* statis dan *Website* dinamis. *Website* statis adalah *Website* yang harus diubah secara manual atau harus mengubah source code jika ingin melakukan perubahan konten. *Website* dinamis adalah *Website*

yang tidak perlu mengubah source codenya dan dapat diperbarui secara berkala karena kontennya disimpan di database dan biasanya ditulis dalam bahasa pemrograman server side seperti PHP, ASP, JSP, dan lainnya (Widia & Asriningtias, 2021).

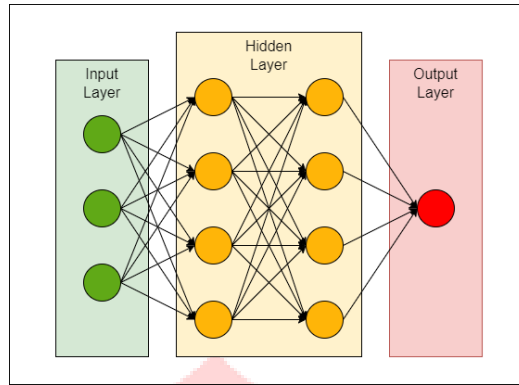
2.2 Teori Khusus

2.2.1 *Customer Churn*

Menurut (Raeisi & Sajedi, 2020), *Customer Churn* adalah customer yang tidak lagi menggunakan layanan yang diberikan oleh perusahaan dalam waktu 1 bulan, 3 bulan atau bahkan 1 tahun sejak terakhir penggunaan layanan. Selain itu, (Kim & Lee, 2021) mendefinisikan *Customer Churn* jika customer tersebut tidak memiliki riwayat mengakses platform seperti *e-commerce* dalam waktu 3 bulan, jika tidak ada riwayat pembelian dalam periode pengumpulan data, atau jika customer tidak membeli lebih dari 1 kali dari suatu influencer.

2.2.2 *Artificial Neural Network*

Artificial Neural Network (ANN) atau sering disebut jaringan saraf tiruan adalah pengklasifikasi yang menggunakan fitur atau *attribute* sebagai input dan membuat sebuah fungsi nonlinier yang bertujuan untuk memprediksi dependent variable (Hussin Adam Khatir & Bee, 2022). Secara umum, ANN bisa direpresentasikan sebagai sistem *neuron* yang saling berhubungan karena terinspirasi dari jaringan saraf biologis yang secara spesifik mengacu pada otak manusia (Chen et al., 2020; Dastres et al., 2021). ANN memiliki struktur yang biasa disebut *Multilayer Perceptron* yang berisi tiga unit layer yaitu *input layer*, *Hidden Layer*, dan *output layer* (Hussin Adam Khatir & Bee, 2022).



Gambar 2.1 *Artificial Neural Network Architecture*

Gambar 2.1 menunjukkan Arsitektur ANN yang digunakan pada penelitian ini yaitu 1 *input layer*, 2 *Hidden Layer*, 1 *output layer*.

2.2.3 **SMOTE**

SMOTE adalah metode oversampling klasik yang diusulkan oleh Chawla et al pada tahun 2002. *SMOTE* meningkatkan jumlah sampel *class* minoritas dengan cara mensintesis sampel *class* minoritas secara acak antara titik sampel dan *k*-neighbor terdekat untuk menghasilkan sampel sintetis baru yang tidak ada dalam dataset dan mencegah *overfitting* (Zhang et al., 2020).

SMOTE adalah singkatan dari Synthetic Minority Oversampling Technique yaitu metode oversampling yang mengambil sampel pada kelas minoritas dengan menciptakan sampel baru secara sintetis, bukan menciptakan sampel baru dengan cara *replacement* secara acak seperti yang dilakukan metode *Random Oversampling*. *SMOTE* (Hussin Adam Khatir & Bee, 2022). *SMOTE* membentuk sampel minoritas baru dengan melakukan interpolasi antara beberapa observasi kelas minoritas yang “berdekatan” satu sama lain (Hussin Adam Khatir & Bee, 2022).

2.2.4 **ROS**

ROS adalah singkatan dari *Random Oversampling* yaitu metode oversampling sebagai solusi paling umum untuk menyelesaikan masalah ketidakseimbangan data dengan cara memproduksi ulang data secara acak yang ada

pada suatu dataset (Fujo et al., 2022). Menurut (Hussin Adam Khatir & Bee, 2022), ROS merupakan salah satu metode paling sederhana untuk meningkatkan *minority class* karena ROS akan menyalin ulang *minority class* yang sudah ada secara acak dan akan diulang sampai dataset seimbang (Hussin Adam Khatir & Bee, 2022). Tetapi hal ini dapat menyebabkan meningkatkan risiko *overfitting* (Fujo et al., 2022; Hussin Adam Khatir & Bee, 2022).

2.2.5 *Lasso Regression*

Lasso Regression adalah metode yang kuat dalam melakukan dua tugas utama yaitu regularisasi dan *Feature Selection* yang dirumuskan oleh Robert Tibshirani pada tahun 1996 (Shafiee et al., 2021). Metode ini memberikan batasan pada nilai absolut dari beberapa parameter model dan menerapkan Proses penyusutan atau regularisasi yang memberikan penalti pada beberapa koefisien variable regresi hingga menjadil nol, sehingga dapat melakukan *Feature Selection* secara efektif dan menghilangkan variable yang kurang relevan (Shafiee et al., 2021).

Lasso Regression memiliki potensi untuk menghilangkan pengaruh variable yang tidak relevan dalam data, yang berarti bahwa dapat sepenuhnya menghapus koefisien suatu variabel menjadi nol dan dengan demikian lebih baik dalam mengurangi varians ketika data terdiri dari beberapa fitur yang tidak signifikan (Fujo et al., 2022).

2.2.6 *mRMR*

mRMR adalah singkatan dari *Minimum Redudancy Maximum Relevance* yang merupakan metode *Feature Selection* tipe filter yang cenderung memilih fitur dengan korelasi tinggi antara fitur terhadap *class* dan korelasi rendah antara fitur dengan fitur lainnya (Billah & Waheed, 2020)(Rahman & Kumar, 2020). *mRMR* adalah teknik *Feature Selection* yang berdasarkan pada kriteria dependensi statistic maksimal dan

meminimalkan ruang fitur melalui peningkatan *dissimilaritas mutual* dengan mempertimbangkan kelas yang diperkenalkan oleh Peng et al (Amin et al., 2020).

2.2.7 *XGBoost*

XGBoost adalah salah satu algoritma boosting tree untuk metode gradient boosting machine (GBM) dan telah banyak digunakan di industry karena kinerjanya yang tinggi dalam memecahkan masalah (Dong et al., 2020). *XGBoost* adalah salah satu implementasi gradient boosting machine (GBM) yang dikenal sebagai salah satu algoritma dengan kinerja terbaik yang digunakan untuk supervised learning (Ibrahim Ahmed Osman et al., 2021).

2.3 Teori Rancangan

2.3.1 *Python*

Python adalah bahasa pemrograman yang terinterpretasi, interaktif, dan berorientasi objek yang menggabungkan *modules*, *exceptions*, tipe data tingkat tinggi dan *class* sehingga *Python* memiliki kekuatan yang luar biasa dengan sintaks yang sangat jelas (*Python Software Foundation*, 2024).

Python juga dapat didefinisikan sebagai bahasa pemrograman tingkat tinggi yang terinterpretasi yang dirancang oleh Guido Van rosum pada akhir tahun 1980-an dengan tipe sistem yang dinamis, mudah dibaca, mudah dipahami, dan mudah diterapkan untuk membuat prototipe secara cepat (Virtanen et al., 2020).

Bahasa pemrograman ini dinamai “*Python*” oleh Guido Van rosum yang terinspirasi dari acara komedi BBC tahun 1970-an yang terkenal yaitu “*Monty Python’s Flying Circus*” karena pada saat itu Guido Van rosum merasa perlu nama yang singkat, unik, dan sedikit misterius untuk bahasa pemrograman yang sedang dikembangkan (*Python Software Foundation*, 2024).

2.3.2 *Pandas*

Pandas adalah sebuah *package* dalam bahasa pemrograman *Python* yang menyediakan struktur data yang cepat, fleksibel, dan ekspresif untuk memudahkan analisis dan manipulasi data “relasional” atau “berlabel” (*Pandas Development Team*, 2024).

Menurut (McKinney, 2022), *Pandas* adalah sebuah *package* atau *library* dalam bahasa pemrograman *Python* yang menyediakan struktur dan fungsi data tingkat tinggi untuk mempermudah pekerjaan dengan data terstruktur atau tabular.

Penggabungan konsep *array* komputasi *NumPy* dengan kemampuan manipulasi data dari *spreadsheet* dan *database* menjadikan *Pandas* sebagai alat yang berguna dalam analisis data dan telah berperan penting dalam menjadikan *Python* sebagai lingkungan yang kuat dan produktif dalam analisis data karena objek utama *Pandas* yaitu *DataFrame* memberikan kemudahan dalam manipulasi data, analisis dan pembersihan data (McKinney, 2022).

2.3.3 *NumPy*

NumPy adalah proyek *open source* yang memungkinkan komputasi numerik dengan menggunakan *Python* yang dibuat pada tahun 2005 dengan dasar dari *Numeric* dan *Numarray library* (*NumPy Developers*, 2024).

NumPy juga dapat disebut sebagai *fundamental package* untuk komputasi ilmiah dalam bahasa pemrograman *Python* yang menyediakan objek *array* multidimensi, matriks, serta berbagai rutinitas untuk operasi cepat pada *array*, termasuk operasi matematika, logika, *shape manipulation*, *sorting*, *selecting* dan banyak lagi (*NumPy Developers*, 2024).

NumPy singkatan dari *Numerical Python* yang telah menjadi dasar utama untuk komputasi numerik dalam lingkungan *Python* dan memiliki fungsi utama yaitu

menyediakan struktur data, algoritma, dan *library* yang diperlukan untuk berbagai aplikasi ilmiah yang melibatkan data numerik dalam bahasa *Python* (McKinney, 2022).

2.3.4 *Matplotlib*

Matplotlib adalah sebuah *library* komprehensif untuk membuat visualisasi statis, animasi, dan interaktif di *Python* yang memudahkan hal-hal mudah dan membuat hal-hal sulit menjadi mungkin dilakukan (*Matplotlib* Development Team, 2024). Hal-hal yang dapat dilakukan oleh *Matplotlib* adalah sebagai berikut (*Matplotlib* Development Team, 2024) :

- a. Membuat grafik berkualitas publikasi
- b. Membuat gambar interaktif yang dapat di-zoom, di-pan (memindahkan gambar), dan di-update
- c. Menyesuaikan *visual style* dan *layout*
- d. Mengekspor ke berbagai format file
- e. Dapat disematkan di *JupyterLab* dan *Graphic User Interface*
- f. Dapat menggunakan banyak *package* pihak ketiga yang telah dibuat pada *Matplotlib*

2.3.5 *Seaborn*

Seaborn adalah *library Python* yang digunakan untuk membuat grafik statistik dengan *interface* tingkat tinggi dan terintegrasi erat dengan *Matplotlib* dan struktur data *Pandas* sehingga mempermudah untuk menerjemahkan pertanyaan tentang data menjadi grafik yang dapat menjawabnya (Waskom, 2021). *Seaborn* juga menawarkan *interface* ke *Matplotlib* yang memungkinkan eksplorasi data yang cepat dan prototyping visualisasi sambil mempertahankan sebagian besar fleksibilitas dan

stabilitas yang diperlukan untuk menghasilkan grafik berkualitas publikasi (Waskom, 2021).

2.3.6 *Scikit-learn*

Scikit-learn adalah sebuah modul atau *library Python* yang dikembangkan menggunakan fungsionalitas dan struktur data yang disediakan oleh *SciPy* yang digunakan untuk *Machine Learning* (*Scikit-learn*, 2024). *Scikit-learn* dibuat oleh David Cournapeau pada tahun 2007, sangat mudah digunakan dan juga dapat mengimplementasikan banyak algoritma *Machine Learning* secara efisien, sehingga bagus menjadi pintu masuk untuk mempelajari *Machine Learning* (Géron, 2022).

2.3.7 *Imbalanced-learn*

Imbalanced-learn adalah sebuah *package Python* yang menyediakan berbagai teknik *re-sampling* dan kompatibel dengan *Scikit-learn* untuk menangani dataset yang menunjukkan ketidakseimbangan antara kelas-kelas (*The Imbalanced-learn Developers*, 2024). *Imbalanced-learn* dibuat oleh Fernando Nogueira pada bulan Agustus 2014 dengan fokus pada implementasi *SMOTE* yang kemudian dibantu oleh Guillaume Lemaitre, Dayvid Victor dan Christos Aridas untuk implementasi metode-metode tambahan untuk *under-sampling* dan *over-sampling* (*The Imbalanced-learn Developers*, 2024).

2.3.8 *Flask*

Flask adalah suatu *framework* aplikasi *Web* ringan yang didesain untuk mempermudah pengembangan awal aplikasi menjadi aplikasi yang kompleks (*Pallets Projects*, 2024). Menurut (Lathkar, 2021), *Flask* adalah sebuah *framework Web* berbasis *Python* yang cukup populer untuk pembuatan aplikasi *Web* dengan skala relatif kecil. *Flask* dikategorikan sebagai *framework* mikro dan proyek

open-source yang dikembangkan oleh Armin Ronacher pada tahun 2004 (Lathkar, 2021).

2.4 Teori Pengujian

2.4.1 *Black Box Testing*

Black Box Testing adalah sebuah teknik pengujian yang digunakan untuk memeriksa aplikasi secara keseluruhan (Yulistiyanti et al., 2022), tanpa melihat atau menguji detail internal dan struktur kode dari aplikasi tersebut (Mahendra & Asmarajaya, 2022) karena *Black Box Testing* hanya berfokus pada fungsional aplikasi (Sholeh et al., 2021).

2.4.2 *Confusion Matrix*

Confusion Matrix adalah sebuah ringkasan yang menjelaskan tentang hasil prediksi pada suatu klasifikasi dengan jumlah prediksi benar dan prediksi salah pada masing-masing kelas. Pada *Confusion Matrix* terdapat istilah TP, FP, TN dan FN yang dapat dijelaskan sebagai berikut: (Dalli, 2022)

- a. *True Positive* (TP) adalah kondisi jika *customer* diprediksi sebagai “Churn” pada data *testing* dan hasil sebenarnya pada data *testing* adalah “Churn”, maka hal itu disebut True Positive (TP)
- b. *False Positive* (FP) adalah kondisi jika *customer* diprediksi sebagai “Churn” pada data *testing* dan hasil sebenarnya pada data *testing* adalah “NotChurn”, maka hal itu disebut False Positive (FP)
- c. *True Negative* (TN) adalah kondisi jika *customer* diprediksi sebagai “NotChurn” pada data *testing* dan hasil sebenarnya pada data *testing* adalah “NotChurn”, maka hal itu disebut True Negative (TN)

- d. *False Negative* (FN) adalah kondisi jika *customer* diprediksi sebagai “NotChurn” pada data *testing* dan hasil sebenarnya pada data *testing* adalah “Churn”, maka hal itu disebut False Negative (FN)

Tabel 2.1 *Confusion Matrix*

		Actual Class	
		Positive	Negative
Predicted Class	Positive	TP	FP
	Negative	FN	TN

Tabel 2.1 menunjukkan ilustrasi *Confusion Matrix* dalam bentuk tabel

Setelah mendapatkan nilai TP, FP, FN dan TN dari *Confusion Matrix*, nilai-nilai tersebut dapat digunakan untuk mengukur nilai beberapa evaluasi matrix yaitu *Accuracy*, *Precision*, *Recall*, dan *AUC* yang dapat dijelaskan sebagai berikut: (Valero-Carreras et al., 2023)

- a. *Accuracy* adalah pengukuran untuk mengetahui seberapa baik model melakukan prediksi pada suatu klasifikasi

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \quad (1)$$

- b. *Precision* adalah pengukuran untuk mengetahui seberapa baik model melakukan prediksi positif pada suatu klasifikasi, yaitu berapa banyak dari semua prediksi positif yang benar-benar positif

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

- c. *Recall* adalah pengukuran untuk mengetahui seberapa baik model melakukan prediksi positif pada data yang benar-benar positif

$$Recall = \frac{TP}{TP + FN} \quad (3)$$

d. *F-1 Score* adalah rata-rata harmonik dari *precision* dan *recall* (S. Wang et al., 2022)

$$F - 1 \text{ Score} = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (4)$$

2.4.3 Kurva ROC-AUC

Kurva ROC adalah grafik yang menggambarkan hubungan antara Sensitivitas atau True Positif Rate (TPR) pada sumbu Y dan Spesifisitas atau False Positif Rate (FPR) pada sumbu X untuk menghasilkan nilai AUC yang dimana AUC itu sendiri merupakan ukuran untuk membandingkan kelas pada klasifikasi sehingga dapat disimpulkan bahwa semakin besar nilai AUC, semakin baik kinerja model dalam membedakan antar kelas pada klasifikasi (Nahm, 2022; Valero-Carreras et al., 2023).

Tabel 2.2 Kategori AUC

Nilai AUC	Kategori
0.90-1.00	Excellent Classification
0.80-0.90	Good Classification
0.70-0.80	Fair Classification
0.60-0.70	Poor Classification
<0.60	Failure

Tabel 2.2 menjelaskan kategori AUC berdasarkan range nilai AUC yang didapatkan dari hasil klasifikasi.

2.5 Tinjauan Studi

2.5.1 *Customer Churn prediction in telecom using Machine Learning in big data platform*

Penelitian berjudul “*Customer Churn prediction in telecom using Machine Learning in big data platform*” yang dilakukan oleh Abdelrahim Kasem Ahmad,

Assef Jafar dan Kadan Aljoumaa pada tahun 2019 bertujuan untuk membuat model *Customer Churn* prediction di perusahaan telecom.

Dataset yang digunakan adalah dataset SyriaTel dengan ukuran lebih dari 70 TeraByte yang berisi data transaksi customer selama 9 bulan terakhir sebelum baseline prediction yang dikumpulkan menggunakan big data platform yaitu Hortonworks Data Platform (HDP).

Klasifikasi dilakukan dengan 4 algoritma yaitu *XGBoost*, *GSM*, *Random Forest*, *Decision Tree* dan dilatih dengan *Statistics Feature*, *SNA Feature* untuk meningkatkan hasil klasifikasi dan teknik balancing yang berbeda yaitu Teknik *Oversampling*, *Undersampling* dan tanpa balancing. Hasil yang didapatkan dari pengujian ini adalah *XGBoost* tanpa menggunakan teknik *Balancing Dataset* dengan akurasi 93.3%. Saat dilakukan Proses teknik *Balancing Dataset*, akurasi *XGBoost* menurun menjadi 93.12% dengan teknik *undersampling* dan 92% dengan teknik *oversampling*.

2.5.2 *Customer Churn prediction in telecommunication industry using data certainty*

Penelitian berjudul “*Customer Churn prediction in telecommunication industry using data certainty*” yang dilakukan oleh Adnan Amin, Feras Al-Obeidat, Babar Shah, Awais Adnan, Jonathan Loo, dan Sajid Anwar pada tahun 2019 bertujuan untuk meningkatkan akurasi prediksi pada model *Customer Churn* prediction dengan pendekatan baru di industri telecommunication menggunakan naïve bayes classifier.

Dataset yang digunakan adalah 4 dataset customer di industri telekomunikasi. Dataset 1 berisi 3,333 *samples*, Dataset 2 berisi 7043 *samples*, Dataset 3 berisi 18,000 *samples*, dan Dataset 4 berisi 100,000 *samples*. Kemudian dilakukan preprocessing

dataset dan *clustering* dengan pendekatan faktor jarak (*Lower Distance* dan *Upper Distance*).

Hasil yang didapatkan dari pengujian ini adalah Akurasi 84.30% untuk UDT dan 82.91% untuk LDT pada Dataset 1, 74.80% untuk UDT dan 70.60% untuk LDT pada Dataset 2, 89.01% untuk UDT dan 70% untuk LDT pada Dataset 3, 56% untuk UDT dan 57% untuk LDT pada Dataset 4.

Kekurangan dari penelitian ini adalah ancaman terhadap validitas karena pemrosesan awal data dan penggunaan classifier yang berbeda.

2.5.3 *Customer Churn prediction system: a Machine Learning approach*

Penelitian ini berjudul “*Customer Churn prediction system: a Machine Learning approach*” yang dilakukan oleh Praveen Lalwani, Manas Kumar Mishra, Jasroop Singh Chadha, dan Pratyush Sethi pada tahun 2022 bertujuan untuk membuat model *Customer Churn prediction* yang efektif dalam industri telekomunikasi menggunakan beberapa algoritma yaitu Logistic Regression, Naïve Bayes, *Support Vector Machine*, *Decision Tree*, *Random Forest*, KNN, AdaBoost, CatBoost, Extra Tree, dan XGBoost.

Dataset yang digunakan adalah dataset customer di industri telekomunikasi yang berisi 7,043 *samples*. Kemudian dilakukan preprocessing dataset dan penggunaan GSA untuk *Feature Selection*.

Hasil yang didapatkan dari pengujian ini adalah CatBoost mendapatkan akurasi tertinggi yaitu 81.8% tetapi dengan nilai AUC 82%, sedangkan AdaBoost dan XGBoost mendapatkan AUC tertinggi yaitu 84% tetapi dengan nilai akurasi 81.71% dan 80.8%.

2.5.4 *Customer Churn Prediction using Artificial Neural Networks (ANN) in Telecom Industry*

Penelitian ini berjudul “*Customer Churn Prediction using Artificial Neural Networks (ANN) in Telecom Industry*” yang dilakukan oleh Yasser Khan, Shahryar Shafiq, Abid Naeem, Sheeraz Ahmed, Nadeem Safwan, dan Sabir Hussain pada tahun 2019 bertujuan untuk membuat model *Customer Churn prediction* dalam industri telekomunikasi menggunakan *Artificial Neural Network*.

Dataset yang digunakan adalah dataset customer di semua perusahaan cellular di Pakistan Telecom Market tahun 2018-2019 yang berisi 20,468 *samples*. Kemudian dilakukan preprocessing dengan menormalisasi semua *attribute* menjadi 0 dan 1.

Model *Customer Churn prediction* didesign dan disetup menggunakan IBM SPSS Statistics 21. *Hidden Layers* dilatih menggunakan back-propagation, mengoptimasi bobot sinaptik menggunakan algoritma gradient descent dan menurunkan kesalahan dengan fungsi aktivasi Hyperbolic Tangent Function. Dan Output Layers menggunakan fungsi aktivasi Softmax Function. Hasil akurasi yang didapatkan dari pengujian ini adalah 79%.

2.5.5 *Improvise XGBoost Machine Learning Algorithm for Customer Churn Prediction*

Penelitian ini berjudul “*Improvise XGBoost Machine Learning Algorithm for Customer Churn Prediction*” yang dilakukan oleh Swetha P, dan Dayananda R B pada tahun 2020 bertujuan untuk mengembangkan model prediksi *XGBoost*.

Dataset yang digunakan adalah dataset South Asia GSM yang berisi 64,107 *samples*. Kemudian dilakukan preprocessing untuk mengubah dan memetakan data

kedalam format yang diperlukan. Hasil yang didapatkan dari pengujian adalah XGBoost dengan akurasi 99.41%, *precision* 99.44% dan *recall* 99.94%.

2.5.6 *Integrated Churn Prediction and Customer Segmentation Framework for Telco Business*

Penelitian ini berjudul “*Integrated Churn Prediction and Customer Segmentation Framework for Telco Business*” yang dilakukan oleh Shuli Wu, Wei-Chuen Yau, Thian-Song Ong, dan Siew-Chin Chong pada tahun 2021 bertujuan untuk mengintegrasikan prediksi churn dan segmentasi pelanggan dalam suatu perangkat kerja menggunakan beberapa algoritma yaitu Logistic Regression, *Decision Tree*, *Random Forest*, Naïve Bayes, AdaBoost dan Multi-Layer Perceptron.

Dataset yang digunakan adalah 3 dataset customer di industri telekomunikasi. Dataset 1 yaitu Dataset IBM, Dataset 2 yaitu Dataset *Telco Customer Churn Prediction Competition 2020*, Dataset 3 yaitu Dataset Cell2Cell. Kemudian dilakukan preprocessing, Exploratory Data Analysis (EDA), dan SMOTE untuk *Balancing Dataset*.

Hasil yang didapatkan dari pengujian ini adalah Logistic Regression mendapatkan akurasi tertinggi yaitu 80.19% pada Dataset 1 tanpa SMOTE, 71.05% pada Dataset 3 tanpa SMOTE, AdaBoost mendapatkan akurasi tertinggi yaitu 77.19% pada Dataset 1 menggunakan SMOTE, *Random Forest* mendapatkan akurasi tertinggi yaitu 95.34% pada Dataset 2 tanpa SMOTE, 93.60% pada Dataset 2 menggunakan SMOTE, dan 63.09% pada Dataset 3 menggunakan SMOTE.

2.5.7 *B2C E-commerce Customer Churn Prediction Based on K-Means and SVM*

Penelitian ini berjudul “*B2C E-commerce Customer Churn Prediction Based on K-Means and SVM*” yang dilakukan oleh Xiancheng Xiahou dan Yoshio Harada

pada tahun 2022 bertujuan untuk mengembangkan model *Customer Churn prediction* menggunakan *Support Vector Machine*, *K-Means* dan *Logistic Regression*.

Dataset yang digunakan adalah dataset aktivitas belanja pengguna pada 23 November 2017 – 4 December 2017 yang dipublish oleh Alibaba Cloud Tianchi Platform untuk Scientific Research dan Big Data Competitions yang berisi 987,994 *samples*. Kemudian dilakukan *preprocessing*, *Segmentation* pelanggan menggunakan *K-Means*, *Feature Selection* menggunakan *Random Forest*, dan *Balancing Dataset* menggunakan *SMOTE*.

Hasil yang didapatkan dari pengujian ini adalah *Support Vector Machine* dengan rata-rata akurasi dari 3 cluster yaitu 91.56% dan *Logistic Regression* dengan rata-rata akurasi dari 3 cluster yaitu 90.66%

2.5.8 *Customer Churn prediction in telecommunication industry using deep learning*

Penelitian ini berjudul “*Customer Churn prediction in telecommunication industry using deep learning*” yang dilakukan oleh Samah Wael Fujo, Suresh Subramanian, dan Moaiad Ahmad Khder pada tahun 2022 bertujuan untuk membuat model *Customer Churn prediction* menggunakan teknik *deep learning*.

Dataset yang digunakan adalah dataset Cell2Cell yang berisi 71,047 *samples* dan dataset IBM Telco yang berisi 7,043 *samples*. Kemudian dilakukan *preprocessing*, menggunakan *Lasso Regression* untuk *Feature Selection*, menggunakan *XGBoost* untuk *Exploratory Data Analysis (EDA)*, dan *Random Oversampling* untuk *Balancing Dataset*.

Hasil yang didapatkan dari pengujian ini adalah *Deep-BP-ANN* mendapatkan akurasi 79.38% menggunakan metode evaluasi *Holdout*, 73.90% menggunakan

metode evaluasi *10-Fold CV* pada dataset Cell2Cell, 88.12% menggunakan metode evaluasi *Holdout* dan 86.57% menggunakan metode evaluasi *10-Fold CV* pada dataset IBM Telco.

2.5.9 *Model Optimization Analysis of Customer Churn Prediction Using Machine Learning Algorithms with Focus on Feature Reduction*

Penelitian ini berjudul “*Model Optimization Analysis of Customer Churn Prediction Using Machine Learning Algorithms with Focus on Feature Reduction*” yang dilakukan oleh Seyed Mohammad Sina Mirabdolbaghi dan Babak Amiri pada tahun 2022 bertujuan untuk mengoptimisasi model *Customer Churn prediction* menggunakan *Feature Reduction*.

Dataset yang digunakan adalah dataset Bank Customer yang berisi 10,000 *samples*, dataset IBM Telco yang berisi 7,043 *samples*, dan dataset Orange 3,333 *samples*. Kemudian dilakukan preprocessing, menggunakan *Principal Component Analysis* (PCA), *Linear Discriminant Analysis* (LDA), *t-distributed Stochastic Neighbor Embedding* (t-SNE), *Autoencoders*, dan *XGBoost* untuk *Feature Reduction*.

Hasil yang didapatkan dari pengujian ini adalah *LightGBM* mendapatkan akurasi terbaik yaitu 86.88% pada dataset Bank Customer, 80.48% pada dataset IBM Telco, dan 95.8% pada dataset Orange menggunakan *Bayesian Optimization*.

2.5.10 *Customer Churn analysis using XGBoosted decision trees*

Penelitian ini berjudul “*Customer Churn analysis using XGBoosted decision trees*” yang dilakukan oleh Muthupriya Vaudevan, Revathi Sathya Narayanan, Sabiyath Fatima Nakeeb, dan Abhisek pada tahun 2022 bertujuan untuk menganalisis *Customer Churn* menggunakan *XGBoost*.

Dataset yang digunakan adalah dataset IBM Telco yang berisi 7,043 *samples*. Kemudian dilakukan *preprocessing* dan menggunakan *GridSearchCV* untuk optimisasi parameter. Hasil yang didapatkan dari pengujian ini adalah *XGBoost* dengan akurasi terbaik yaitu 82.1%.

2.5.11 *Impact of Hyperparameters on Deep learning Model for Customer Churn Prediction in Telecommunication Sector*

Penelitian ini berjudul “*Impact of Hyperparameters on Deep learning Model for Customer Churn Prediction in Telecommunication Sector*” yang dilakukan oleh Anouar Dalli pada tahun 2022 bertujuan untuk mengetahui dampak hyperparameter pada model *Customer Churn prediction*.

Dataset yang digunakan adalah dataset Turkish Telecommunication Company yang berisi 3,333 *samples*. Kemudian dilakukan *preprocessing*, menggunakan *Min-MaxScaler* untuk *standarization*, menggunakan *SMOTE* untuk *Balancing Dataset*, menggunakan 3 Fungsi Aktivasi yaitu *Sigmoid*, *ReLU* dan *Tanh*, menggunakan.

Hasil yang didapatkan dari pengujian ini adalah Kombinasi parameter *ReLU* untuk *Hidden Layers* dan *Sigmoid* untuk *Output Layers* mendapatkan akurasi tertinggi yaitu 86.8%, *Batch Size* dengan jumlah 7 mendapatkan akurasi terbaik yaitu 85.75%, dan *RMSProp* mendapatkan akurasi terbaik yaitu 86.45% sebagai *optimizer*

2.5.12 *Customer Churn Prediction in Telecommunication Industry Using Classification and Regression Trees and Artificial Neural Network Algorithms*

Penelitian ini berjudul “*Customer Churn Prediction in Telecommunication Industry Using Classification and Regression Trees and Artificial Neural Network Algorithms*” yang dilakukan oleh Sulaiman Olaniyi Abdulsalam, Micheal Olaolu

Arowolo, Yakub Kayode Saheed, Jesutofunmi Onaope Afolayan pada tahun 2022 bertujuan untuk membuat model *Customer Churn* prediction menggunakan *Classification and Regression Trees* (CART) dan *Artificial Neural Network* (ANN).

Dataset yang digunakan adalah dataset dari *Francisco Gallery of bigml.com* yang berisi 3,333 *samples*. Kemudian dilakukan *preprocessing*, menggunakan *Relief-F* untuk *Feature Selection*. Hasil yang didapatkan dari pengujian ini adalah ANN mendapatkan akurasi terbaik yaitu 93.88%.

2.5.13 *Just-in-time Customer Churn prediction in the telecommunication sector*

Penelitian ini berjudul “*Just-in-time Customer Churn prediction in the telecommunication sector*” yang dilakukan oleh Adnan Amin, Feras Al-Obeidat, Babar Shah, May Al Tae, Changez Khan, Hamood Ur Rehman Durrani, Sajid Anwar pada tahun 2020 bertujuan untuk model *Customer Churn* prediction menggunakan beberapa algoritma yaitu SVM, *k-NN*, Naïve Bayes, dan Neural Network.

Dataset yang digunakan adalah dataset Telecom yang berisi 3,333 *samples*, dan dataset The UCI KDD Cup Archive yang berisi 5,784 *samples*. Kemudian dilakukan *preprocessing*, menggunakan *mRMR* untuk *Feature Selection*, dan menggunakan Random Under Sampling untuk *Balancing Dataset*. Hasil yang didapatkan dari pengujian ini adalah model Heterogeneous Ensemble (Stacking) mendapatkan akurasi tertinggi yaitu 77.27%.

2.5.14 *Machine Learning Based Customer Churn Prediction in Banking*

Penelitian ini berjudul “*Machine Learning Based Customer Churn Prediction in Banking*” yang dilakukan oleh Manas Rahman dan V Kumar pada tahun 2020 bertujuan untuk membuat model *Customer Churn* prediction menggunakan *k-NN*, SVM, *Decision Tree*, dan *Random Forest*.

Dataset yang digunakan adalah dataset Bank yang berisi 10,000 *samples*. Kemudian dilakukan preprocessing, menggunakan *Relief-F* dan *mRMR* untuk *Feature Selection*, dan menggunakan *Random Oversampling* untuk *Balancing Dataset*.

Hasil yang didapatkan dari pengujian ini adalah *Random Forest* mendapatkan akurasi terbaik yaitu 85.18% tanpa *Oversampling* dan 95.74% menggunakan *Oversampling* pada pengujian yang dilakukan secara langsung (tanpa *Feature Selection*), *k-NN* mendapatkan akurasi terbaik yaitu 83.97% tanpa *Oversampling* dan *Random Forest* mendapatkan akurasi terbaik yaitu 92.95% menggunakan *Oversampling* pada pengujian yang dilakukan menggunakan *mRMR*, *k-NN* mendapatkan akurasi terbaik yaitu 82.15% tanpa *Oversampling* dan *Random Forest* mendapatkan akurasi terbaik yaitu 92.19% menggunakan *Oversampling* pada pengujian yang dilakukan menggunakan *Relief-F*.

2.5.15 Predicting Customer Churn using Artificial Neural Network

Penelitian ini berjudul “*Predicting Customer Churn using Artificial Neural Network*” yang dilakukan oleh Sanjay Kumar dan Manish Kumar pada tahun 2019 bertujuan untuk membuat model *Customer Churn* prediction menggunakan *Artificial Neural Network*.

Dataset yang digunakan adalah dataset *Telecommunication Industry* yang berisi 7,043 *samples*. Kemudian dilakukan preprocessing, menggunakan *ReLU* dan *Sigmoid* untuk Fungsi Aktivasi. Hasil yang didapatkan dari pengujian ini adalah Fungsi Aktivasi *Sigmoid* dengan jumlah *Batch Size* 50, dan 2 *Hidden Layers* dengan *Hidden Unit* (Jumlah *Neuron* pada *Hidden Layers*) 75 dan 35 mendapatkan akurasi terbaik yaitu 93.14%.

2.5.16 Rangkuman Model Penelitian

No	Peneliti	Nama Jurnal dan Tahun	Judul	Latar Belakang	Masalah (Research Question)	Metode	Hasil
1.	Abdelrahim Kasem Ahmad, Assef Jafar, Kadan Aljoumaa.	<i>Journal of Big Data</i> , Vol. 6, No. 28, March 2019.	<i>Customer Churn prediction in telecom using Machine Learning in big data platform.</i>	Untuk bertahan di pasar yang kompetitif memerlukan strategi. Salah satu strateginya adalah mempertahankan pelanggan karena membutuhkan	Bagaimana hasil penggunaan <i>Machine Learning</i> dalam pembuatan model <i>Customer Churn prediction</i> .	<i>Decision Tree</i> , <i>Random Forest</i> , <i>Gradient Boosted Machine Tree (GBM)</i> , <i>Extreme Gradient</i>	<i>XGBoost</i> mendapatkan akurasi tertinggi dengan nilai 93.3% (tanpa teknik <i>Balancing Dataset</i>), dan 93.12% (menggunakan <i>Undersampling</i>),

				biaya yang lebih rendah dibandingkan memperoleh pelanggan baru. Peneliti mengusulkan <i>Machine Learning</i> sebagai metode untuk membuat model <i>Customer Churn prediction</i>.		<i>Boosting (XGBoost)</i>.	92% (menggunakan Oversampling dibandingkan model lainnya.
2.	Adnan Amin, Feras Al-	<i>Journal of Business</i>	<i>Customer Churn prediction in</i>	<i>Customer Churn prediction</i> adalah	Bagaimana dapat dikembangkan	<i>Naïve Bayes</i>	Berdasarkan hasil pengujian,

	Obeidat, Babar Shah, Awais Adnan, Jonathan Loo, Sajid Anwar.	<i>Research, Vol. 94, Pages 290-301, Januari 2019.</i>	<i>telecommunication industry using data certainty.</i>	klasifikasi biner dimana pelanggan dibagi menjadi dua yaitu Churn dan Non-Churn, Selanjutnya perilaku churn diklasifikasi kedalam subkategori menjadi pelanggan yang benar-benar churn dan	pendekatan baru untuk <i>Customer Churn Prediction (CCP)</i> dalam industri telekomunikasi dengan mengklasifikasi tingkat kepastian churn dengan faktor jarak, dan Bagaimana pengaruhnya terhadap tingkat akurasi prediksi?		klasifikasi menggunakan Faktor Jarak mendapatkan akurasi tertinggi pada Zona Atas (UDT) dengan nilai 84.3% untuk Dataset 1, 74.8% untuk Dataset 2, 89.01% untuk Dataset 3, dan 56% untuk Dataset 4.
--	--	--	---	--	---	--	---

				pelanggan yang tidak sengaja churn. Peneliti mengusulkan Naïve Bayes dan Pendekatan Faktor Jarak untuk klasifikasi data yang akan divalidasi kepastian datanya.			
3.	Praveen Lalwani, Manas Kumar	<i>Computing. 104, 271-294, (2022)</i>	<i>Customer Churn prediction system: a</i>	Di era persaingan yang semakin	Bagaimana dapat dikembangkan metodologi	<i>Logistic Regression, Naïve Bayes,</i>	<i>CatBoost</i> mendapatkan akurasi tertinggi

	Mishra, Jasoop Singh Chadha, Pratyush Sethi.		<i>Machine Learning approach</i>	kompetitif, memaksimalkan keuntungan sudah menjadi hal wajib. Diantara semua strategi mempertahankan pelanggan adalah strategi yang paling murah dibandingkan strategi lain. Peneliti	<i>Customer Churn prediction (CCP)</i> yang efektif dalam industri telekomunikasi?	<i>Support Vector Machine, Decision Trees, Random Forest Classifier, Extra Tree Classifier, AdaBoost, XGBoost, CatBoost.</i>	yaitu 81.8% tetapi dengan nilai AUC 82%, sedangkan AdaBoost dan XGBoost mendapatkan AUC tertinggi yaitu 84% tetapi dengan nilai akurasi 81.71% dan 80.8%.
--	--	--	----------------------------------	---	---	--	--

				mengusulkan <i>Machine Learning</i> untuk membuat model <i>Customer Churn prediction</i>			
4.	Yasser Khan, Shahryar Shafiq, Abid Naeem, Sheeraz Ahmed, Nadeem Safwan, Sabir Hussain.	<i>International Journal of Advanced Computer Science and Application, Vol. 10, No. 9, 2019</i>	<i>Customer Churn Prediction using Artificial Neural Networks (ANN) in Telecom Industry</i>	Akuisisi pelanggan baru membutuhkan biaya 6 sampai 7 kali lebih banyak daripada mempertahankan pelanggan yang sudah ada,	Bagaimana akurasi prediksi churn dalam industri telekomunikasi dapat ditingkatkan melalui pendekatan jaringan saraf tiruan dengan	<i>Artificial Neural Network</i>	Berdasarkan hasil pengujian, akurasi yang didapatkan adalah 79%

				sehingga sebagian besar perusahaan di dunia saat ini sangat menderita karena berpindahnya pelanggan yang tidak puas ke pesaing baru.	teknik prediksi lainnya?		
5.	Swetha P, Dayananda R B	<i>EAI Endorsed Transactions on Energy Web, Vol 7, Issue 30, 2020.</i>	<i>Improvised_XGBoost Machine Learning Algorithm for Customer Churn Prediction.</i>	Retensi pelanggan telah menjadi salah satu isu utama bagi perusahaan	Bagaimana cara mengatasi ketidakseimbangan dan ukuran dataset yang besar dalam	<i>XGBoost</i>	Berdasarkan hasil pengujian, <i>XGBoost</i> mendapatkan akurasi sangat

			berbasis layanan. Secara umum, alasan berpindahnya pelanggan terjadi karena masalah layanan pelanggan, biaya data, masalah jaringan, dan lainnya. Mendapatkan pelanggan baru membutuhkan biaya 5 sampai 6	membangun model prediksi churn pelanggan yang efisien di industry telekomunikasi?		tinggi dengan nilai 99.41% dibandingkan dengan beberapa algoritma lainnya.
--	--	--	--	--	--	---

				kali lebih mahal dibandingkan mempertahankan pelanggan baru. Oleh karena itu, tujuan utama penelitian ini untuk membuat model yang dapat memprediksi tingkat churn pelanggan.			
6.	Shuli Wu, Wei Chuen Yau,	<i>IEEE Access</i> , Vol.9, 2021	<i>Integrated Churn Prediction and</i>	Dalam industry telekomunikasi,	Bagaimana mengintegrasikan	<i>Logistic Regression</i> ,	Pada pengujian Dataset Ke-1,

	Thian-Song Ong, Siew- Chin Cong		<i>Customer Segmentation Framework for Telco Business</i>	menarik pelanggan baru bukan lagi merupakan strategi yang baik karena biaya untuk mempertahankan pelanggan yang sudah ada jauh lebih rendah dan manajemen churn menjadi sangat penting. Oleh karena itu,	prediksi churn dan segmentasi pelanggan dalam suatu kerangka kerja analitik untuk manajemen churn di industry telekomunikasi?	<i>Decision Tree, Random Forest, Naïve Bayes, Ada Boost, Multi- Layer Perceptron.</i>	<i>Logistic Regression</i> mendapatkan akurasi tertinggi dengan nilai 80.19% tanpa <i>SMOTE</i> , dan <i>AdaBoost</i> mendapatkan akurasi tertinggi dengan nilai 77.19% menggunakan <i>SMOTE</i> . Kemudian pada
--	---------------------------------------	--	---	--	--	---	---

			 penelitian ini membuat model yang menggabungkan prediksi churn dan segmentasi pelanggan untuk manajemen churn.		pengujian Dataset Ke-2, <i>Random Forest</i> mendapatkan nilai akurasi tertinggi dengan nilai 95.34% tanpa <i>SMOTE</i> , dan 93.60% menggunakan <i>SMOTE</i> . Sedangkan pada pengujian Dataset Ke-3, <i>Logistic</i>
--	--	--	---	--	--

							<i>Regression</i> mendapatkan akurasi tertinggi dengan nilai 71.05% tanpa <i>SMOTE</i>, dan <i>Random Forest</i> mendapatkan akurasi tertinggi dengan nilai 63.09% menggunakan <i>SMOTE</i>
--	--	--	--	--	--	--	---

7.	Xiancheng Xiahou, Yoshio Harada	<i>Journal of Theoretical and Applied Electronic Commerce Research</i>	<i>B2C E-commerce Customer Churn Prediction Based on K-Means and SVM</i>	Memanfaatkan sumber daya pelanggan yang ada dan menghindari hilangnya pelanggan tersebut sangat penting untuk mempertahankan keunggulan pasar karena pelanggan adalah salah satu aset terpenting	Bagaimana mengembangkan model prediksi churn pelanggan yang efektif dan akurat untuk industri E-commerce B2C menggunakan algoritma <i>K-Means</i> dan SVM	<i>K-Means, Support Vector Machine, Logistic Regression</i>	Pada pengujian sebelum segmentasi SVM mendapatkan akurasi dengan nilai 90.81%. Setelah dilakukan segmentasi, SVM mendapatkan akurasi tertinggi untuk semua cluster (Cluster
----	---------------------------------------	--	--	--	---	---	---

			suatu perusahaan. Biaya untuk mengembangkan pelanggan baru juga seringkali lebih tinggi dibanding mempertahankan pelanggan yang sudah ada. Oleh karena itu, penelitian ini mengusulkan model <i>Customer</i>		1-3) dengan nilai rata-rata 91.56% sedangkan LR mendapatkan akurasi dengan nilai rata-rata 90.66% untuk semua cluster (Cluster 1-3).
--	--	--	--	--	---

				Churn prediction berdasarkan kombinasi segmentasi pelanggan dan prediksi churn.			
8.	Samah Wael Fujo, Suresh Subramanian, Moaiad Ahmad Khder	<i>Information Sciences Letter</i>	<i>Customer Churn prediction in telecommunication industry using deep learning</i>	Industri telekomunikasi menghadapi masalah besar terkait dengan pergantian pelanggan. Permasalahan ini tidak hanya	Bagaimana kinerja model Deep-BP-ANN dibandingkan dengan metode <i>Machine Learning</i> seperti <i>XGBoost</i> , Logistic Regression, Naïve Bayes, dan KNN	<i>Logistic Regression, Naïve Bayes, k-NN, XGBoost, Artificial Neural Network (ANN)</i>	Deep-BP-ANN mendapatkan akurasi 79.38% menggunakan metode evaluasi <i>Holdout</i> , 73.90% menggunakan metode evaluasi

			berdampak pada pertumbuhan perusahaan yang pesat, namun juga dapat berdampak pada pendapatan. Oleh karena itu, pada penelitian ini mengusulkan model <i>Customer Churn prediction</i> untuk mengatasi masalah tersebut.	dalam memprediksi <i>Customer Churn</i>	<i>10-Fold CV</i> pada dataset Cell2Cell, 88.12% menggunakan metode evaluasi <i>Holdout</i> dan 86.57% menggunakan metode evaluasi <i>10-Fold CV</i> pada dataset IBM Telco
--	--	--	---	--	--

9.	Seyed Mohammad Sina Mirabdolbaghi, Babak Amiri	<i>Discrete Dynamics in Nature and Society</i>	<i>Model Optimization Analysis of Customer Churn Prediction Using Machine Learning Algorithms</i>	Permasalahan <i>Customer Churn</i> merupakan tantangan besar dalam dunia industri khususnya industri telekomunikasi. Untuk mengatasi masalah ini, salah satu cara yang paling efektif untuk adalah	Seberapa efektif PCA, Autoencoders, LDA, T-SNE, dan XGBoost dalam mengurangi fitur untuk prediksi churn dalam industri yang diberikan?	<i>LightGBM, Decision Tree, Random Forest, Support Vector Machine, Logistic Regression, XGBoost, AdaBoost</i>	LightGBM mendapatkan akurasi terbaik yaitu 86.88% pada dataset Bank Customer, 80.48% pada dataset IBM Telco, dan 95.8% pada dataset Orange menggunakan Bayesian Optimization
----	--	--	---	---	---	--	---

				mengevaluasi perilaku pelanggan untuk mendeteksi kemungkinan churn. Oleh karena itu, sangat penting untuk memiliki sistem prediksi churn yang efisien.			
10.	Muthupriya Vaudevan, Revathi Sathya	Indonesian Journal of Electrical	Customer Churn analysis using	Tingkat churn yang tinggi menyebabkan	Bagaimana performa algoritma XGBoost dalam	XGBoost	Berdasarkan hasil pengujian, XGBoost

	Narayanan, Sabiyath Fatima Nakeeb, Abhishek	<i>Engineering and Computer Science, Vol. 25, Issue 1, Pages 488-495, 2022</i>	<i>XGBoosted decision trees</i>	kerugian yang besar. Meskipun kerugian dalam bisnis tidak dapat dihindari, kerugian akibat churn dapat dikelola. Oleh karena itu, perlu dikembangkan metode yang baik dan meningkatkan metode yang sudah ada.	menganalisis churn pelanggan dibandingkan dengan model <i>Machine Learning</i> lainnya seperti <i>Random Forest</i> atau Logistic Regression?	mendapatkan akurasi tertinggi dengan nilai 82.1%
--	---	--	-------------------------------------	---	---	---

11.	Anouar Dalli	<i>Mathematical Problems in Engineering, 2020</i>	<i>Impact of Hyperparameter on Deep learning Model for Customer Churn Prediction in Telecommunication Sector</i>	Dalam beberapa tahun terakhir, industri telekomunikasi mengalami perubahan besar seperti pelayanan baru, peningkatan teknologi, dan meningkatnya persaingan. Oleh karena itu, memprediksi <i>Customer Churn</i>	Bagaimana dampak dari beberapa kombinasi hyperparameter terhadap performance model?	<i>Neural Network</i>	Berdasarkan hasil pengujian, kombinasi parameter <i>ReLU</i> untuk <i>Hidden Layers</i> dan <i>Sigmoid</i> untuk <i>Output Layers</i> mendapatkan akurasi tertinggi yaitu 86.8%, <i>Batch Size</i> dengan jumlah 7 mendapatkan akurasi terbaik
-----	--------------	---	--	---	---	-----------------------	--

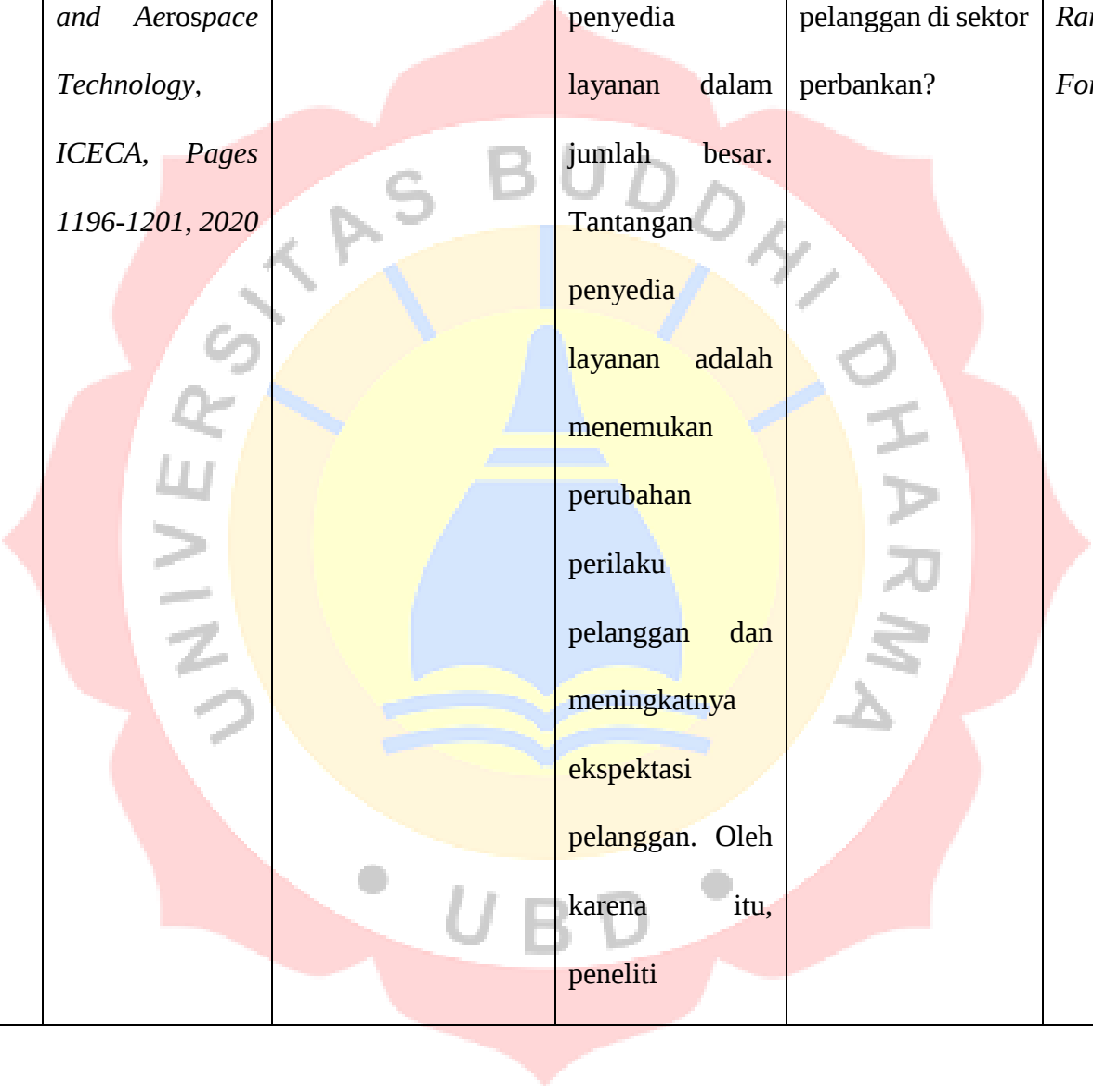
				di sector telekomunikasi menjadi parameter penting. Namun, hal itu terlihat sangat kurangnya informasi heuristic yang diperoleh secara empiris sehingga harus mempengaruhi hyperparameter		yaitu 85.75%, dan RMSProp mendapatkan akurasi terbaik yaitu 86.45% sebagai <i>optimizer</i>
--	--	--	---	--	--	--

				sebagai konsekuensinya			
12.	Sulaiman Olaniyi Abdulsalam, Micheal Olaolu Arowolo, Yakub Kayode Saheed, Jesutofunmi Onaope Afolayan	<i>Indonesian Journal of Electrical Engineering and Informatics, Vol. 10, Issue 2, Pages 431-440, 2022</i>	<i>Customer Churn Prediction in Telecommunication Industry Using Classification and Regression Trees and Artificial Neural Network Algorithm</i>	Perpindahan pelanggan adalah masalah serius, karena dampaknya langsung terhadap pendapatan perusahaan. Oleh karena itu, sangat penting untuk menyelidiki	Bagaimana kinerja model prediksi churn pelanggan menggunakan metode klasifikasi CART dan ANN?	<i>Classification and Regression Tree (CART), Artificial Neural Networks (ANN)</i>	Berdasarkan hasil pengujian, ANN mendapatkan akurasi terbaik yaitu 93.88%

				masalah yang mempengaruhi <i>Customer Churn</i> agar dapat mengambil tindakan yang tepat untuk mengurangi churn			
13.	Adnan Amin, Feras Al-Obeidat, Babar Shah, May Al Tae, Changez Khan, Hamood	<i>Journal of Supercomputing</i> , Vol. 76, Issue. 6, Pages 3924-3948, 2022	<i>Just-in-time Customer Churn prediction in the telecommunication sector</i>	Terdapat persaingan yang ketat antar perusahaan dalam sektor telekomunikasi	Bagaimana kinerja model JIT-Homogeneous dan JIT-Heterogeneous ensemble model untuk CCP di	<i>Support Vector Machine, k-NN, Naïve Bayes,</i>	Berdasarkan hasil pengujian, model Heterogeneous Ensemble (Stacking)

	Ur Rehman Durrani, Sajid Anwar			untuk memaksimalkan pendapatan dan meningkatkan pangsa pasar melalui akuisisi lebih banyak pelangga. Namun, perolehan pelanggan baru dari jauh lebih mahal, sekitar sepuluh kali lipat dibandingkan	sektor telekomunikasi?	<i>Neural Network</i>	mendapatkan akurasi tertinggi yaitu 77.27%.
--	--------------------------------------	--	--	--	---------------------------	---------------------------	---

				mempertahan pelanggan yang sudah ada. Oleh karena itu, peneliti mengusulkan model Just-in- time (JIT) untuk prediksi churn			
14.	Manas Rahman, V Kumar	<i>Proceedings of the 4th international Conference on Electronics, Communication</i>	<i>Machine Learning Based Customer Churn Prediction in Banking</i>	Pasar saat ini sangat dinamis dan sangat kompetitif. Hal ini karena banyaknya	Bagaimana penggunaan teknik <i>Machine Learning</i> dapat membantu dalam memprediksi churn	<i>k-NN, Support Vector Machine, Decision Tree,</i>	<i>Random Forest</i> mendapatkan akurasi terbaik yaitu 85.18% tanpa Oversampling

		and Aerospace Technology, ICECA, Pages 1196-1201, 2020		penyedia layanan dalam jumlah besar. Tantangan penyedia layanan adalah menemukan perubahan perilaku pelanggan dan meningkatnya ekspektasi pelanggan. Oleh karena itu, peneliti	pelanggan di sektor perbankan?	Random Forest	dan 95.74% menggunakan Oversampling pada pengujian yang dilakukan secara langsung (tanpa Feature Selection), k-NN mendapatkan akurasi terbaik yaitu 83.97% tanpa Oversampling dan Random Forest
--	--	--	---	---	---------------------------------------	----------------------	--

			 mengusulkan model <i>Customer Churn prediction</i> untuk mengatasi tantangan tersebut		mendapatkan akurasi terbaik yaitu 92.95% menggunakan Oversampling pada pengujian yang dilakukan menggunakan <i>mRMR</i>, <i>k-NN</i> mendapatkan akurasi terbaik yaitu 82.15% tanpa Oversampling dan <i>Random</i>
--	--	--	--	--	---

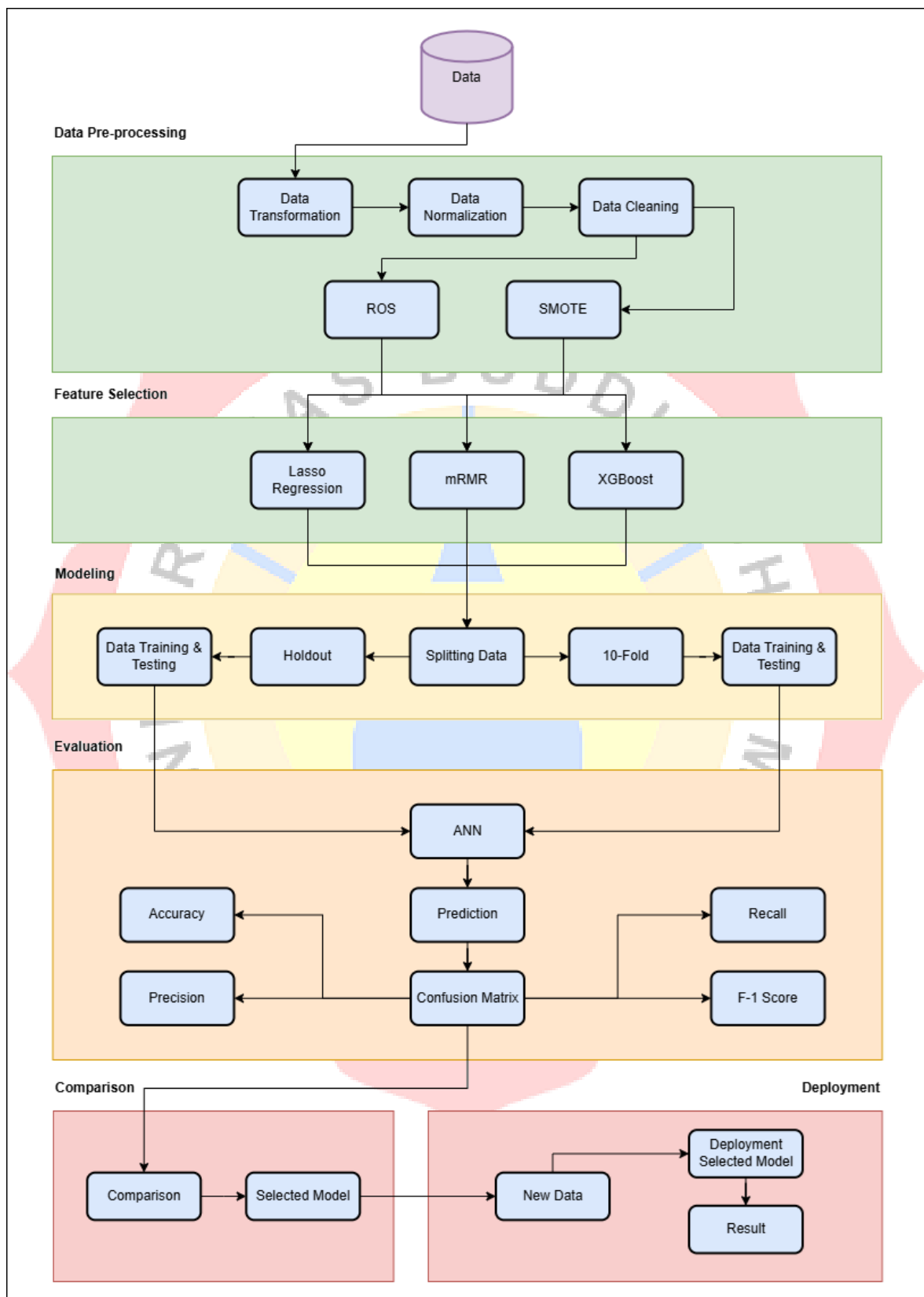
							<i>Forest</i> mendapatkan akurasi terbaik yaitu 92.19% menggunakan Oversampling pada pengujian yang dilakukan menggunakan <i>Relief-F</i>
15.	Sanjay Kumar, Manish Kumar	<i>Communications in Computer and Information Science, Vol.</i>	<i>Predicting Customer Churn using Artificial Neural Network</i>	<i>Customer Churn</i> adalah salah satu faktor yang paling berpengaruh	Bagaimana penggunaan berbagai fungsi aktivasi dalam ANN	<i>Artificial Neural Network (ANN)</i>	Berdasarkan hasil pengujian, Fungsi Aktivasi <i>Sigmoid</i> dengan jumlah <i>Batch</i>

		1000, Pages 299-306, 2019	bagi berbagai industry berbasis layanan terutama di industri telekomunikasi. Membiarkan <i>Customer Churn</i> tidak terkendali dapat sangat menurunkan bisnis perusahaan. Oleh karena itu, peneliti mengusulkan	mempengaruhi performa model?		Size 50, dan 2 <i>Hidden Layers</i> dengan <i>Hidden Unit</i> (Jumlah <i>Neuron</i> pada <i>Hidden Layers</i>) 75 dan 35 mendapatkan akurasi terbaik yaitu 93.14%.
--	--	------------------------------	---	------------------------------	--	---

				model <i>Customer Churn prediction</i> untuk menyelesaikan masalah <i>Customer Churn</i>			
--	--	--	--	--	--	--	--

Berdasarkan hasil tinjauan studi dan rangkuman penelitian, peneliti sebelumnya telah melakukan penelitian untuk membuat *Customer Churn model prediction* menggunakan beberapa algoritma seperti ANN, XGBoost, SVM dan lainnya dan beberapa *Feature Selection* seperti *Lasso Regression*, *mRMR*, *Relief-F*, *XGBoost* dan lainnya. Oleh karena itu, peneliti akan melakukan perbandingan model dengan menggunakan algoritma ANN dan beberapa *Feature Selection* yaitu *Lasso Regression*, *mRMR*, dan *XGBoost* yang kemudian akan digunakan untuk *Customer Churn prediction*. Penelitian bertujuan untuk mengoptimalkan model *Customer Churn prediction* yang sudah ada sebelumnya dengan melakukan penambahan metode *Balancing Dataset* dan *Feature Selection* yang berbeda sehingga penelitian ini diharapkan dapat meningkatkan akurasi menjadi lebih baik.

2.6 Kerangka Pemikiran



Gambar 2.2 Kerangka Pemikiran

BAB III

METODOLOGI PENELITIAN

3.1 Pengumpulan Data

Peneliti menggunakan dataset IBM Telco yang berisi 7043 sampel informasi customer dan 21 *attribute* yaitu:

- | | |
|---------------------------|--|
| a. <i>customerID</i> | l. <i>DeviceProtection</i> |
| b. <i>gender</i> | m. <i>TechSupport</i> |
| c. <i>SeniorCitizen</i> | n. <i>StreamingTV</i> |
| d. <i>Partner</i> | o. <i>StreamingMovies</i> |
| e. <i>Dependents</i> | p. <i>Contract</i> |
| f. <i>tenure</i> | q. <i>PaperlessBilling</i> |
| g. <i>PhoneService</i> | r. <i>PaymentMethod</i> |
| h. <i>MultipleLines</i> | s. <i>MonthlyCharges</i> |
| i. <i>InternetService</i> | t. <i>TotalCharges</i> |
| j. <i>OnlineSecurity</i> | u. <i>Churn (Dependent Variable/Class)</i> |
| k. <i>OnlineBackup</i> | |

Dataset yang digunakan adalah dataset sekunder dan telah banyak digunakan di beberapa penelitian sebelumnya (Fujo et al., 2022). Dataset ini tersedia di *Kaggle* :

<https://www.kaggle.com/code/ahmedshahriarsakib/churn-prediction-i-eda-statistical-analysis/input>

3.2 Pengolahan Data

Dataset IBM Telco berisi *categorical attribute* dengan 2 sampai 3 *unique value*. Oleh karena itu, dataset diolah agar bisa digunakan untuk pembuatan model yang membutuhkan nilai *attribute* dalam bentuk numerik.

3.2.1 Data Preparation

Melakukan penghapusan *attribute* “customerID” karena tidak memiliki pengaruh penting dalam pembuatan model dan mengubah type data “TotalCharges” dari object menjadi float karena *attribute* tersebut memiliki nilai numerik.

```
def DataPreparation():  
    dataset = pd.readcsv('dataset/IBM Telco.csv')  
    del dataset['customerID']  
    dataset['TotalCharges']=pd.to_numeric(dataset['TotalCharges'],  
    errors='coerce')  
    return dataset  
    print("\n")  
    DataPreparation().info()  
    print("\n")
```

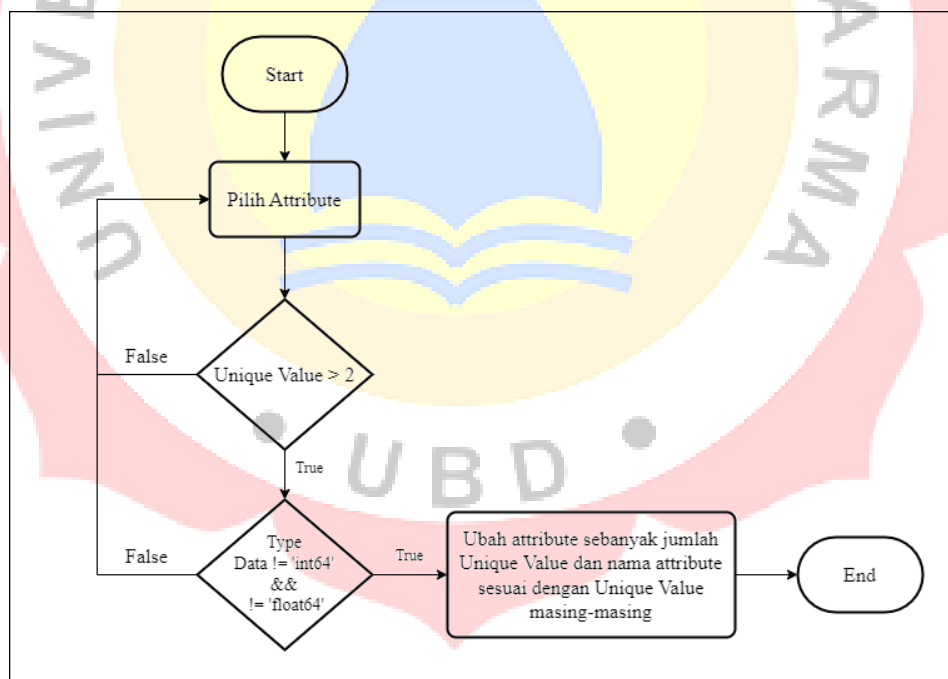
```
<class 'pandas.core.frame.DataFrame'>  
RangeIndex: 7043 entries, 0 to 7042  
Data columns (total 20 columns):  
#   Column                Non-Null Count  Dtype  
---  ---  
0   gender                7043 non-null   object  
1   SeniorCitizen         7043 non-null   int64  
2   Partner               7043 non-null   object  
3   Dependents            7043 non-null   object  
4   tenure                7043 non-null   int64  
5   PhoneService          7043 non-null   object  
6   MultipleLines         7043 non-null   object  
7   InternetService       7043 non-null   object  
8   OnlineSecurity        7043 non-null   object  
9   OnlineBackup          7043 non-null   object  
10  DeviceProtection      7043 non-null   object  
11  TechSupport           7043 non-null   object  
12  StreamingTV           7043 non-null   object  
13  StreamingMovies       7043 non-null   object  
14  Contract              7043 non-null   object  
15  PaperlessBilling      7043 non-null   object  
16  PaymentMethod         7043 non-null   object  
17  MonthlyCharges        7043 non-null   float64  
18  TotalCharges          7032 non-null   float64  
19  Churn                 7043 non-null   object  
dtypes: float64(2), int64(2), object(16)  
memory usage: 1.1+ MB
```

Gambar 3.1 Informasi Dataset Setelah *Data Preparation*

Gambar 3.1 menunjukkan informasi dataset yang berisi 20 *attribute* diantaranya 19 (*independent*), 1 (*dependent*) dengan jumlah data 7043 dan tipe datanya.

3.2.2 Data Transformation

Pembuatan model *Customer Churn Prediction* menggunakan ANN memerlukan dataset yang berisi nilai numerik, maka diperlukan Proses transformasi data. Pada Proses transformasi data, ada 2 metode yang akan digunakan, yaitu: *One-Hot Encoding*, *Label Encoding*. Tahap Pertama, metode *One-Hot Encoding* dari library Python yaitu *Pandas* menggunakan `pd.get_dummies` untuk mengubah *attribute* kategorikal menjadi representasi nilai numerik. Tetapi dalam eksperimen ini hasilnya dalam bentuk Boolean. Oleh karena itu perubahan ke bentuk numerik akan dilakukan pada tahap selanjutnya.



Gambar 3.2 Flowchart Proses *One-Hot Encoding*

Gambar 3.2 menunjukkan alur Proses metode *One-Hot Encoding* mulai dari pemilihan *attribute* dengan syarat *attribute* hanya memiliki unique value lebih dari 2 dan tipe data tidak sama dengan int64 dan tidak sama dengan float64 sampai

mengubah *attribute* sesuai dengan jumlah unique valuenya dan namanya masing-masing. Attribute yang terpilih adalah sebagai berikut: MultipleLines, InternetService, OnlineSecurity, OnlineBackup, DeviceProtection, TechSupport, StreamingTV, StreamingMovies, Contract, PaymentMethod.

Contoh Sebelum Proses *One-Hot Encoding*:

Warna
Biru
Merah
Hijau

Hasil Sesudah Proses *One-Hot Encoding* :

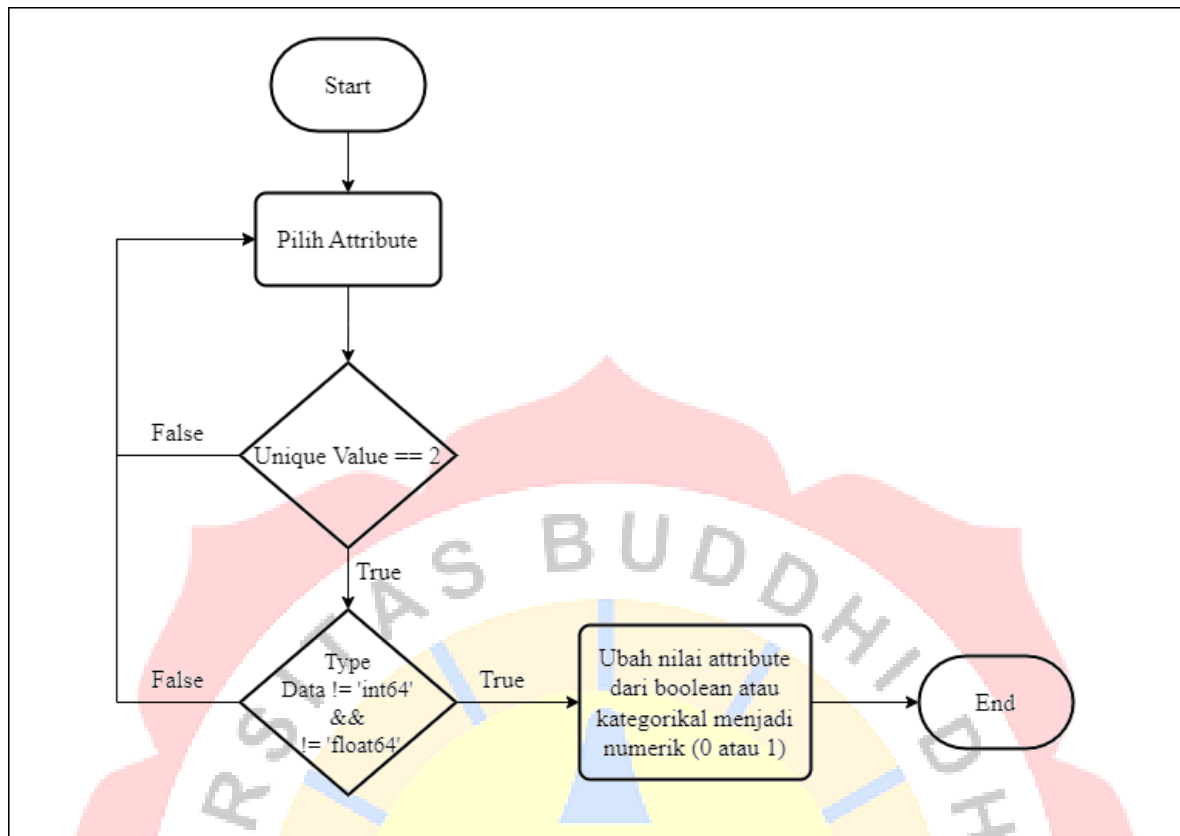
Warna_Biru	Warna_Merah	Warna_Hijau
True	False	False
False	True	False
Flase	False	True

Tabel 3.1 Hasil Proses *One-Hot Encoding*

No	gender	SeniorCitizen	Partner	...	PaymentMethod_Credit card (automatic)	PaymentMethod_Electronic check	PaymentMethod_Mailed check
0	Female	0	Yes	...	False	True	False
1	Male	0	No	...	False	False	True
2	Male	0	No	...	False	False	True
3	Male	0	No	...	False	False	False
4	Female	0	No	...	False	True	False
...	...	...	...	...	...	...	...
7038	Male	0	Yes	...	False	False	True
7039	Female	0	Yes	...	True	False	False
7040	Female	0	Yes	...	False	True	False
7041	Male	1	Yes	...	False	False	True
7042	Male	0	No	...	False	False	False

Tabel 3.1 menunjukkan hasil Proses metode *One-Hot Encoding* sehingga total atributnya menjadi 40 (*independent*)

Tahap kedua, metode *Label Encoding* dari library Python yaitu *Scikit-learn* yang digunakan untuk mengubah nilai non-numerik atau boolean menjadi nilai numerik. Dalam kasus ini, nilai (Male = 1 dan Female = 0), (Yes = 1 dan No = 0), (True = 1 dan False = 0).



Gambar 3.3 Flowchart Proses Label Encoding

Gambar 3.3 menunjukkan alur Proses metode *Label Encoding* mulai dari pemilihan *attribute* dengan syarat *attribute* hanya memiliki unique value sama dengan 2 dan tipe data tidak sama dengan int64 dan tidak sama dengan float64 sampai mengubah nilai *attribute* dari boolean atau kategorikal menjadi numerik (0 atau 1). Semua *attribute* terpilih kecuali 4 *attribute* berikut : SeniorCitizen, tenure, MonthlyCharges, TotalCharges.

Sebelum :

Warna_	Warna_	Warna_
Biru	Merah	Hijau
True	False	False
False	True	False
False	False	True

Sesudah :

Warna_	Warna_	Warna_
Biru	Merah	Hijau
1	0	0
0	1	0
0	0	1

Tabel 3.2 Hasil Proses *Label Encoding*

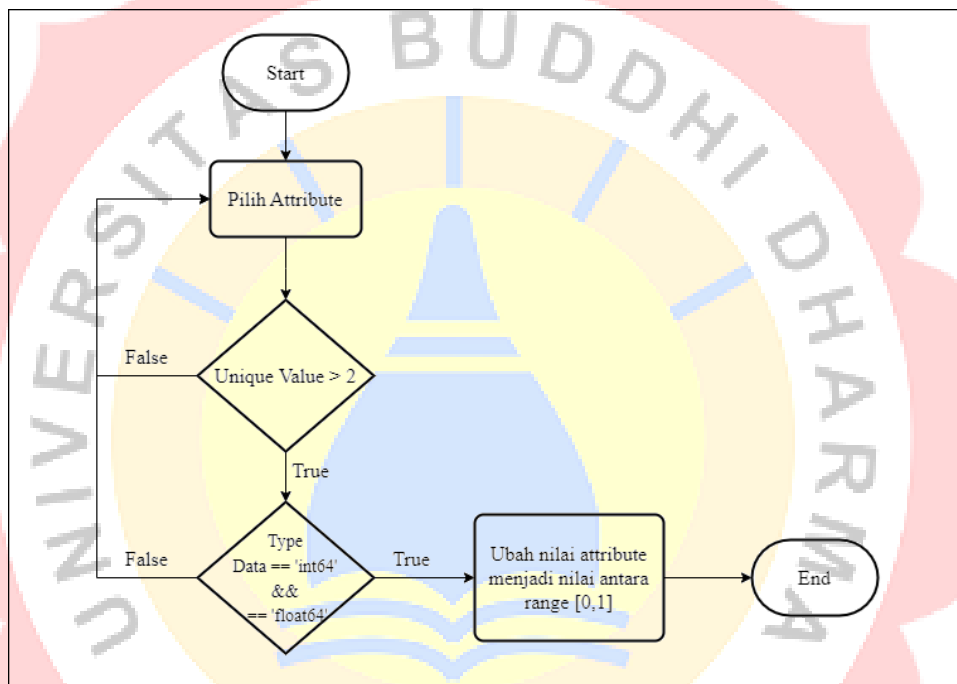
No	gender	SeniorCitizen	Partner	...	PaymentMethod_Credit card (automatic)	PaymentMethod_Electronic check	PaymentMethod_Mailed check
0	0	0	1	...	0	1	0
1	1	0	0	...	0	0	1
2	1	0	0	...	0	0	1
3	1	0	0	...	0	0	0
4	0	0	0	...	0	1	0
...	...	...	...	...	...	...	...
7038	1	0	1	...	0	0	1
7039	0	0	1	...	1	0	0
7040	0	0	1	...	0	1	0
7041	1	1	1	...	0	0	1
7042	1	0	0	...	0	0	0

Tabel 3.2 menunjukkan hasil Proses metode *Label Encoding* sehingga semua nilai *attribute* yang dengan tipe data *boolean* dan kategorikal menjadi numerik (0 atau 1).

3.2.3 Data Normalization

Berdasarkan informasi dari penelitian Fujo et al, dataset IBM Telco memiliki beberapa *attribute* dengan rentang nilai yang cukup tinggi. Oleh karena itu, dilakukan *Normalization* menggunakan metode *MinMaxScaler* dari *library Python* yaitu *Scikit-learn* yang digunakan untuk mengubah nilai *attribute* menjadi nilai dalam *range* [0,1] yang hanya diterapkan pada *attribute* yang memiliki rentang nilai yang tinggi.

Gambar 3.4 Flowchart Proses *MinMaxScaler*



Gambar 3.4 menunjukkan alur Proses metode *MinMaxScaler* mulai dari pemilihan *attribute* dengan syarat *attribute* hanya memiliki unique value lebih dari 2 dan tipe data sama dengan int64 dan sama dengan float64 sampai mengubah nilai *attribute* menjadi nilai antara range [0, 1]. Attribute yang terpilih adalah sebagai berikut : tenure, MonthlyCharges, TotalCharges.

Tabel 3.3 Data Sebelum Proses *MinMaxScaler*

No	tenure	MonthlyCharges	TotalCharges
0	1	29.85	29.85
1	34	56.95	1889.5
2	2	53.85	108.15
3	45	42.3	1840.75
4	2	70.7	151.65
...	...	...	...
7038	24	84.8	1990.5
7039	72	103.2	7362.9
7040	11	29.6	346.45
7041	4	74.4	306.6
7042	66	105.65	6844.5

Tabel 3.3 menunjukkan data sebelum diProses menggunakan metode *MinMaxScaler*.

Tabel 3.4 Data Setelah Proses *MinMaxScaler*

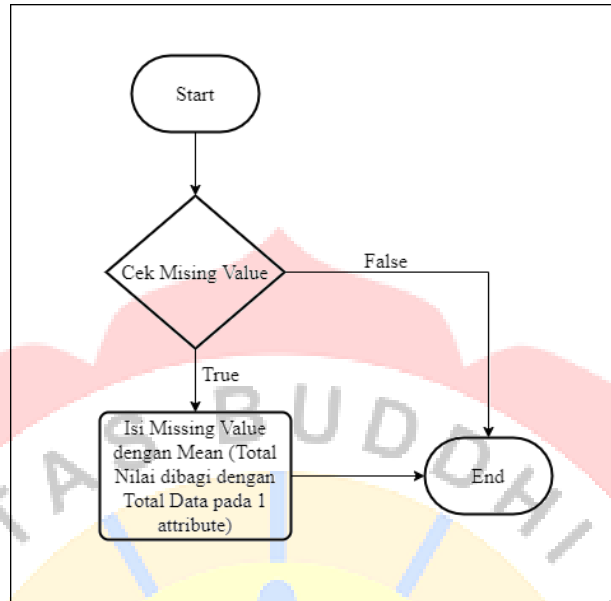
No	tenure	MonthlyCharges	TotalCharges
0	0.013889	0.115423	0.001275
1	0.472222	0.385075	0.215867
2	0.027778	0.354229	0.01031
3	0.625	0.239303	0.210241
4	0.027778	0.521891	0.01533
...	...	...	...
7038	0.333333	0.662189	0.227521
7039	1	0.845274	0.847461
7040	0.152778	0.112935	0.037809
7041	0.055556	0.558706	0.03321
7042	0.916667	0.869652	0.787641

Tabel 3.4 menunjukkan data sesudah diProses menggunakan metode *MinMaxScaler*.

3.2.4 Missing Value

Berdasarkan informasi dari penelitian Fujio et al, di dalam dataset IBM Telco terdapat *Missing Value*. Oleh karena itu, dilakukan pengecekan dan menggunakan

metode *SimpleImputer* dari *library Python* yaitu *Scikit-learn* untuk di-replace dengan nilai mean.



Gambar 3.5 Proses *SimpleImputer*

Gambar 3.5 menunjukkan alur Proses metode *SimpleImputer* mulai dari pengecekan *Missing Value* kemudian diisi dengan nilai rata-rata (*mean*). Attribute yang memiliki *Missing Value* adalah *TotalCharges* dengan total 11 *Missing Value*.

Tabel 3.5 Data Sebelum Proses *SimpleImputer*

No	gender	SeniorCitizen	Partner	Dependent	tenure	PhoneService	PaperlessBilling	MonthlyCharges	TotalCharges	Churn
488	0	0	1	1	0.0	0	1	0.341294	NaN	0
753	1	0	0	1	0.0	1	0	0.019900	NaN	0
936	0	0	1	1	0.0	1	0	0.622886	NaN	0
1082	1	0	1	1	0.0	1	0	0.074627	NaN	0
1340	0	0	1	1	0.0	0	0	0.376119	NaN	0
3331	1	0	1	1	0.0	1	0	0.015920	NaN	0
3826	1	0	1	1	0.0	1	0	0.070647	NaN	0
4380	0	0	1	1	0.0	1	0	0.017413	NaN	0
5218	1	0	1	1	0.0	1	1	0.014428	NaN	0
6670	0	0	1	1	0.0	1	0	0.548259	NaN	0
6754	1	0	0	1	0.0	1	1	0.434328	NaN	0

Tabel 3.5 menunjukkan data sebelum diProses menggunakan metode *SimpleImputer*.

Tabel 3.6 Data Setelah Proses *SimpleImputer*

No	gender	SeniorCitizen	Partner	Dependent	tenure	PhoneService	PaperlessBilling	MonthlyCharges	TotalCharges	Churn
488	0	0	1	1	0.0	0	1	0.341294	0.261309	0
753	1	0	0	1	0.0	1	0	0.019900	0.261309	0
936	0	0	1	1	0.0	1	0	0.622886	0.261309	0
1082	1	0	1	1	0.0	1	0	0.074627	0.261309	0
1340	0	0	1	1	0.0	0	0	0.376119	0.261309	0
3331	1	0	1	1	0.0	1	0	0.015920	0.261309	0
3826	1	0	1	1	0.0	1	0	0.070647	0.261309	0
4380	0	0	1	1	0.0	1	0	0.017413	0.261309	0
5218	1	0	1	1	0.0	1	1	0.014428	0.261309	0
6670	0	0	1	1	0.0	1	0	0.548259	0.261309	0
6754	1	0	0	1	0.0	1	1	0.434328	0.261309	0

Tabel 3.6 menunjukkan data setelah diProses menggunakan metode *SimpleImputer*.

3.2.5 *Balancing Dataset*

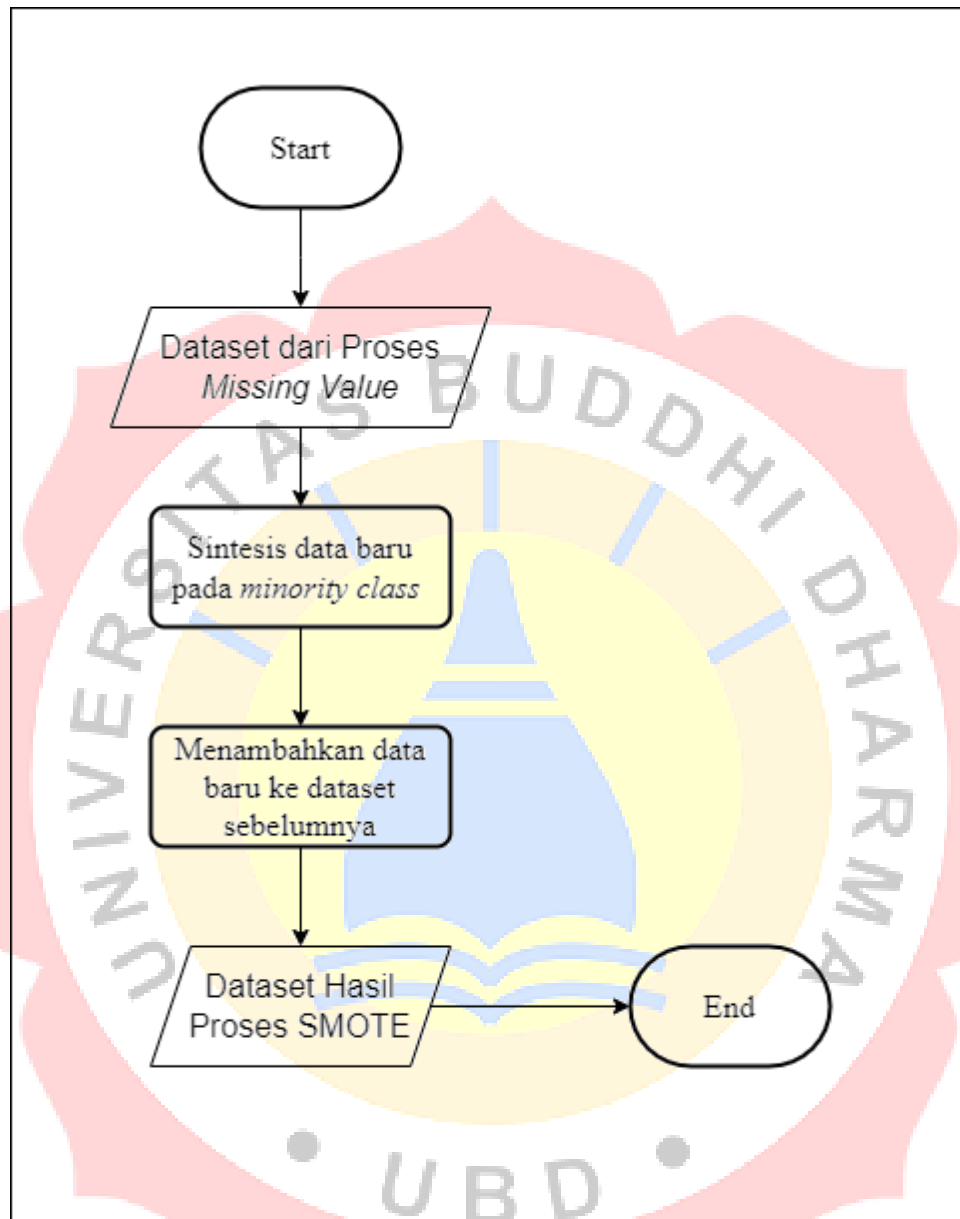
Dataset IBM Telco memiliki perbandingan jumlah positif dan negatif class yang cukup tinggi yaitu 26.54% (Churn = Yes) dan 73.46% (Churn = No). Jumlah tersebut didapat dari perhitungan menggunakan persamaan dibawah ini.

$$\text{Persentase Positif Class} = \frac{\text{Jumlah Class (Yes)}}{\text{Jumlah Data}} \quad (5)$$

$$\text{Persentase Negatif Class} = \frac{\text{Jumlah Class (No)}}{\text{Jumlah Data}} \quad (6)$$

Oleh karena itu, dilakukan *Balancing* data menggunakan metode *SMOTE* dari library *Python* yaitu *Imbalanced-learn* dan metode *ROS* dari library *Python* yaitu

Imbalanced-learn untuk mengatasi ketidakseimbangan tersebut dan mencegah bias terhadap kelas mayoritas yang akan mengakibatkan hasil akurasi yang kurang akurat.



Gambar 3.6 Flowchart Proses SMOTE

Gambar 3.6 menunjukkan alur Proses metode *SMOTE* mulai dari pengambilan data dari hasil Proses sebelumnya kemudian melakukan sintesis data untuk membuat data baru pada *minority class*

Tabel 3.7 Data Sebelum Proses *SMOTE*

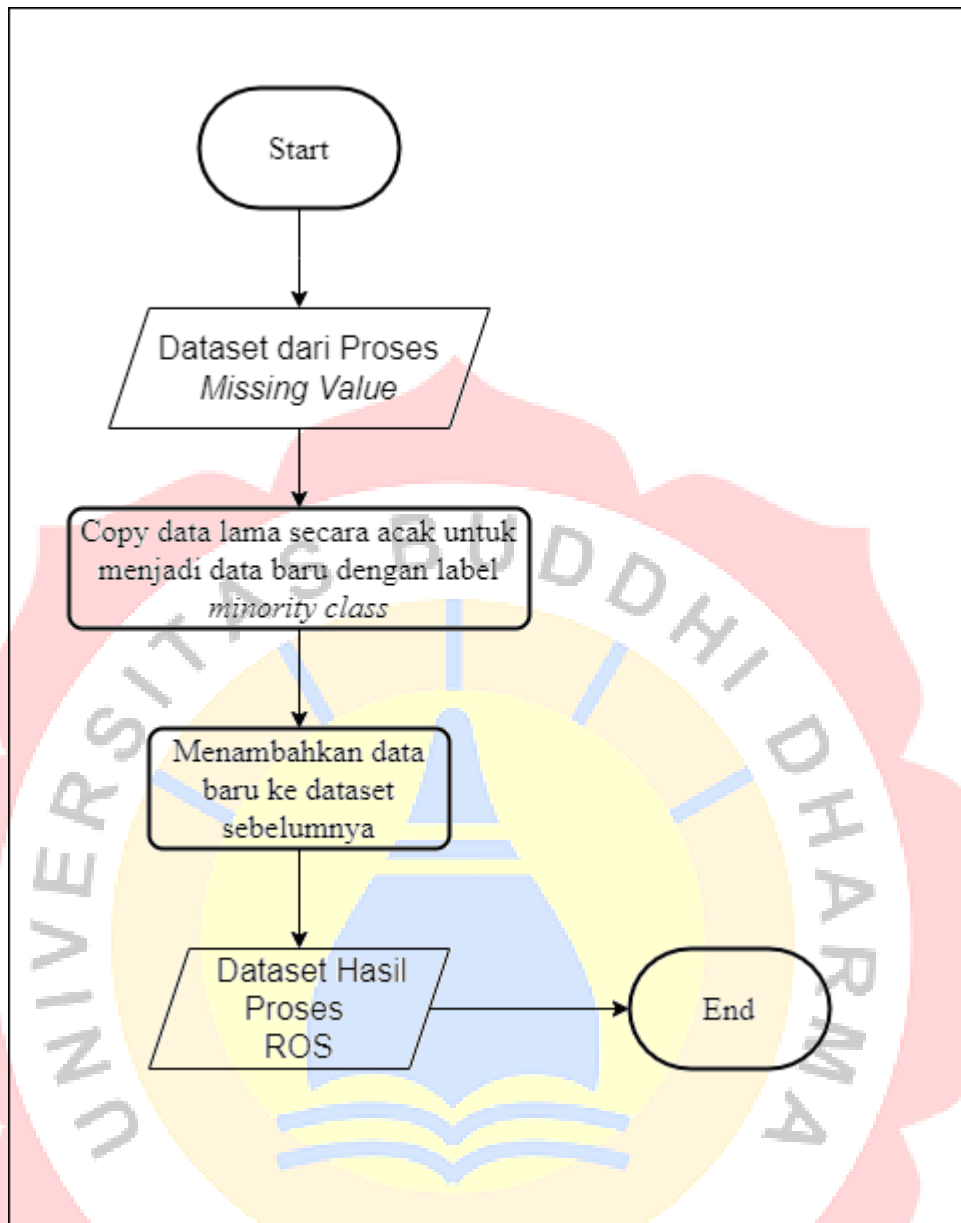
No	gender	SeniorCitizen	Partner	...	PaymentMethod_Credit card (automatic)	PaymentMethod_Electronic check	PaymentMethod_Mailed check
0	0	0	1	...	0.0	1	0
1	1	0	0	...	0.0	0	1
2	1	0	0	...	0.0	0	1
3	1	0	0	...	0.0	0	0
4	0	0	0	...	0.0	1	0
...	...	...	...	...	...	...	...
7038	1	0	1	...	0.0	0	1
7039	0	0	1	...	1.0	0	0
7040	0	0	1	...	0.0	1	0
7041	1	1	1	...	0.0	0	1
7042	1	0	0	...	0.0	0	0

Tabel 3.7 menunjukkan data sebelum diProses menggunakan *SMOTE*

Tabel 3.8 Data Setelah Proses *SMOTE*

No	gender	SeniorCitizen	Partner	...	PaymentMethod_Credit card (automatic)	PaymentMethod_Electronic check	PaymentMethod_Mailed check
0	0	0	1	...	0	1	0
1	1	0	0	...	0	0	1
2	1	0	0	...	0	0	1
3	1	0	0	...	0	0	0
4	0	0	0	...	0	1	0
...	...	...	...	...	...	...	...
10343	0	0	0	...	0	1	0
10344	1	0.767279	0.767279	...	0	0	1
10345	0	0	0	...	0	1	0
10346	0.483691	0.516309	0	...	0	1	0
10347	1	0.684447	0.315553	...	0	0	1

Tabel 3.8 menunjukkan jumlah data meningkat menjadi 10,348 setelah diProses menggunakan *SMOTE*



Gambar 3.7 Flowchart Proses ROS

Gambar 3.7 menunjukkan alur Proses metode ROS mulai dari pengambilan data dari hasil Proses sebelumnya kemudian menyalin data lama secara acak untuk membuat data baru pada *minority class*

Tabel 3.9 Data Sebelum Proses ROS

No	gender	SeniorCitizen	Partner	...	PaymentMethod_Credit card (automatic)	PaymentMethod_Electronic check	PaymentMethod_Mailed check
0	0	0	1	...	0.0	1	0
1	1	0	0	...	0.0	0	1
2	1	0	0	...	0.0	0	1
3	1	0	0	...	0.0	0	0
4	0	0	0	...	0.0	1	0
...	...	...	...	...	...	...	...
7038	1	0	1	...	0.0	0	1
7039	0	0	1	...	1.0	0	0
7040	0	0	1	...	0.0	1	0
7041	1	1	1	...	0.0	0	1
7042	1	0	0	...	0.0	0	0

Tabel 3.9 menunjukkan data sebelum diProses menggunakan ROS

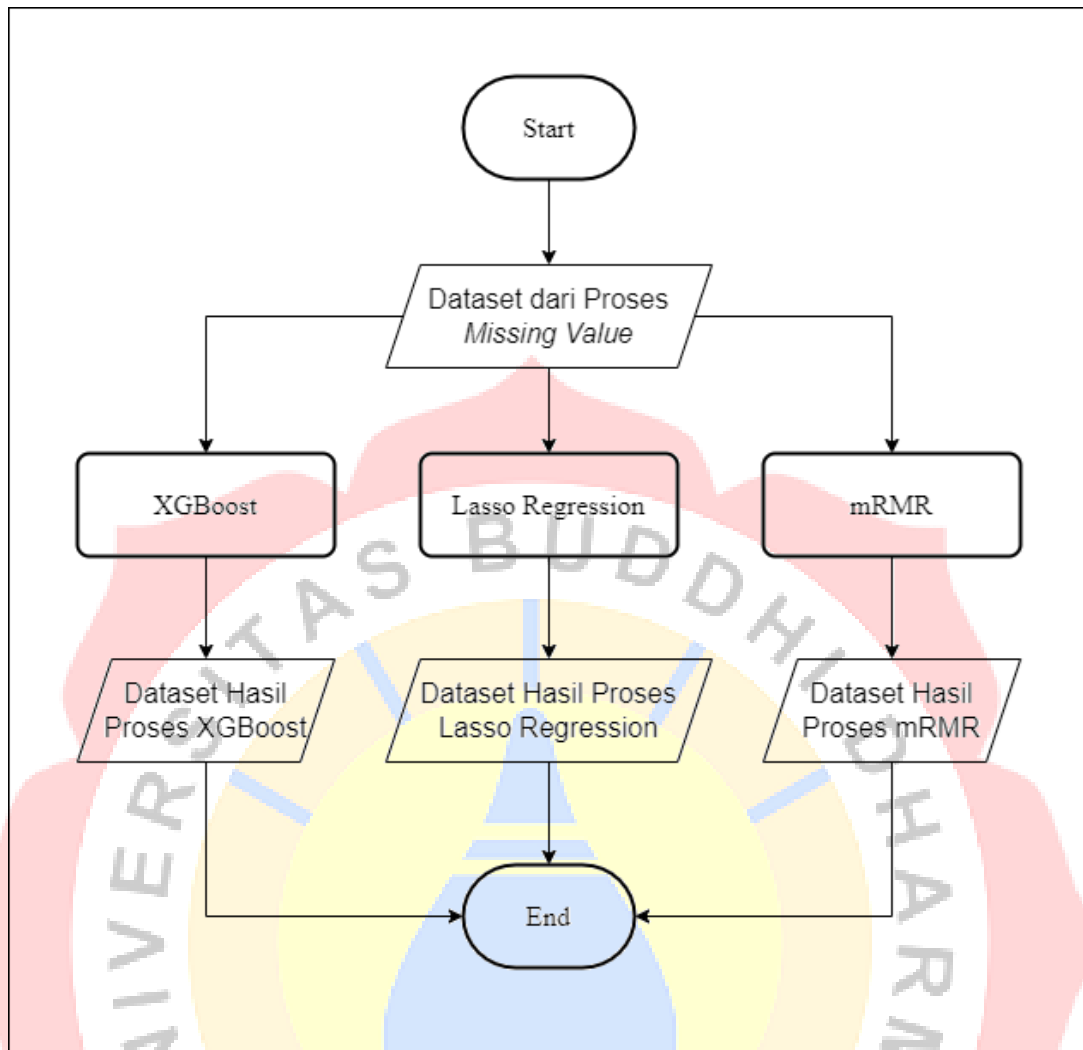
Tabel 3.10 Data Setelah Proses ROS

No	gender	SeniorCitizen	Partner	..	PaymentMethod_Credit card (automatic)	PaymentMethod_Electronic check	PaymentMethod_Mailed check
0	0	0	1	..	0	1	0
1	1	0	0	..	0	0	1
2	1	0	0	..	0	0	1
3	1	0	0	..	0	0	0
4	0	0	0	..	0	1	0
...	...	...	...	..	...	...	...
10343	1	0	0	..	0	0	1
10344	1	1	0	..	0	1	0
10345	0	1	0	..	0	0	1
10346	1	0	0	..	0	1	0
10347	1	0	1	..	0	0	1

Tabel 3.10 menunjukkan jumlah data meningkat menjadi 10348 setelah diProses menggunakan ROS

3.2.6 Feature Selection

Untuk mengurangi dimensi data dan kompleksitas model, serta meningkatkan kinerja model, dilakukan Proses pemilihan *Feature* menggunakan 3 metode yaitu *XGBoost*, *LassoRegression*, dan *mRMR* dari *library Python* yaitu *XGBoost*, *Scikit-learn*, dan *mRMR* untuk memilih *Feature* yang relevan atau berpengaruh terhadap nilai *class* pada dataset.



Gambar 3.8 *Flowchart Proses Feature Selection*

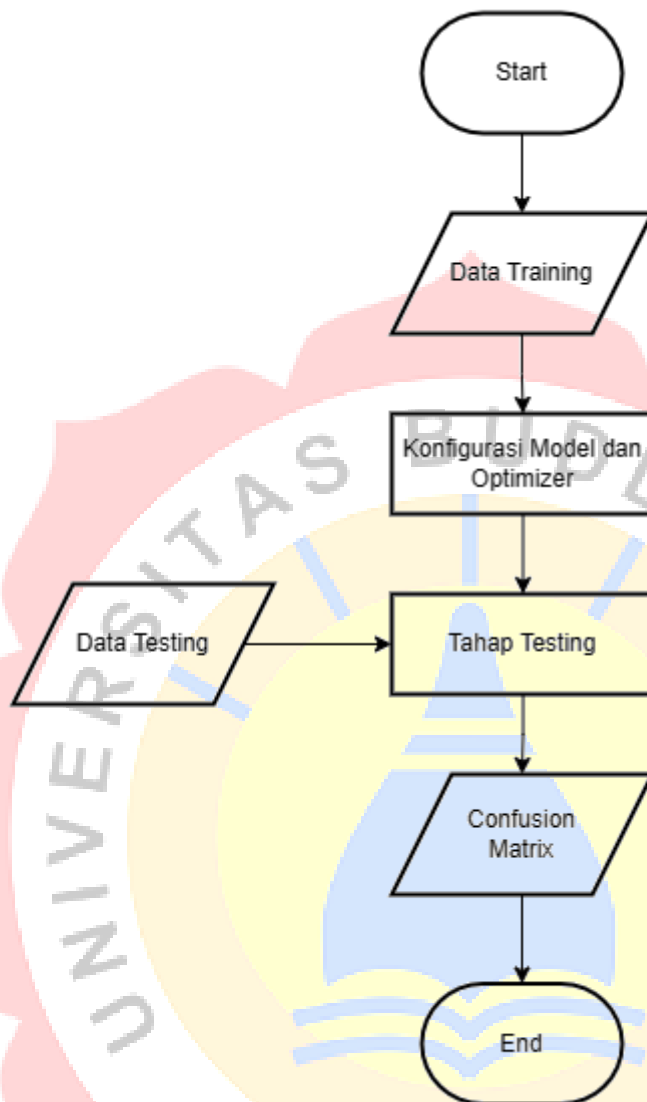
Gambar 3.8 menunjukkan alur Proses metode *SMOTE* mulai dari pengambilan data dari hasil Proses sebelumnya kemudian melakukan Proses pemilihan *Feature* menggunakan beberapa *Feature Selection*

3.3 Pelatihan Model

3.3.1 *Splitting Data*

Pada tahap ini, dataset dibagi menjadi 2 jenis yaitu Data *Training* dan Data *Testing*. Metode yang digunakan adalah *Holdout* (90% untuk *Training* dan 10% untuk *Testing*) dan *10-Fold Cross Validation* berdasarkan dari penelitian sebelumnya yaitu Fujo et al.

3.3.2 Flowchart Pelatihan Model



Gambar 3.9 Flowchart Pelatihan Model

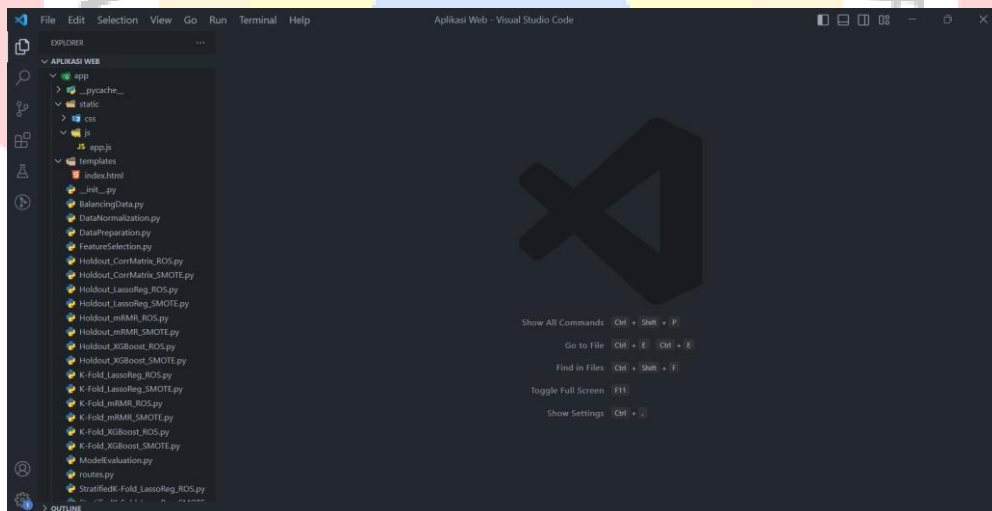
3.3.3 Perangkat Keras

Untuk melatih model *Artificial Neural Network* tentu saja dibutuhkan perangkat keras (Hardware). Kecepatan Proses pelatihan model dipengaruhi oleh spesifikasi perangkat keras. Oleh karena itu, semakin tinggi spesifikasi perangkat keras yang digunakan akan semakin cepat juga Proses pelatihan model. Perangkat keras yang digunakan untuk melatih model pada penelitian ini adalah sebagai berikut:

Jenis Perangkat Keras	Nama Perangkat Keras
Processor	Intel core i7 – 10510U
RAM	8 GB DDR3 2133 MHz
Kartu Grafis	NVIDIA GeForce MX250
Penyimpanan	SSD 512 GB
Peripheral	Keyboard Mouse Monitor

3.3.4 Lingkungan Pengembangan

Pelatihan model *Artificial Neural Network* dilakukan menggunakan *Visual Studio Code*, *Python Version 3.11.5*, *Scikit-learn Version 1.3.1*, dan model *Artificial Neural Network* didapatkan dari *Scikit-learn*.



Gambar 3.10 Tampilan Lingkungan Pengembangan Model

3.3.5 Konfigurasi Model

Konfigurasi model dilakukan pada masing-masing metode pembagian data. Setiap model dari masing-masing *Feature Selection* dan masing-masing metode

Splitting Data dibuat menjadi 1 file terpisah untuk mempermudah dalam melakukan eksperimen pelatihan model.

Tabel 3.11 Konfigurasi Parameter Model *Artificial Neural Network*

Parameter Model	
Parameter	Nilai
<i>Hidden Layer</i>	200, 250
<i>Activation Function</i>	<i>ReLu</i>
<i>Optimizer</i>	<i>Adam</i>

3.3.6 Kompilasi Model

Proses *Training* dilakukan sesuai dengan konfigurasi masing-masing model. Berdasarkan perangkat keras yang digunakan dan telah dijelaskan spesifikasi sebelumnya, pelatihan model membutuhkan waktu dari 55 detik hingga 901 detik tergantung arsitektur *Artificial Neural Network* yang dilatih. Berikut merupakan *Feature* yang terpilih dari ketiga model *Artificial Neural Network* dengan masing-masing metode *balancing* data dan *Splitting Data*:

Tabel 3.12 Top 10 Feature (XGBoost + SMOTE + Holdout (90%, 10%))

XGBoost + SMOTE + Holdout (90%, 10%)		
No.	Feature	Importance Score
1	Contract_Month-to-month	0.343273
2	InternetService_Fiber optic	0.139904
3	OnlineSecurity_No	0.0417598
4	TechSupport_No	0.0356718
5	StreamingMovies_Yes	0.0302177
6	InternetService_DSL	0.0290379
7	PaymentMethod_Electronic check	0.0270609
8	Dependents	0.0250466
9	PaperlessBilling	0.0220574
10	Partner	0.0214875

Tabel 3.13 Top 10 Feature (XGBoost + ROS + Holdout (90%, 10%))

XGBoost + ROS + Holdout (90%, 10%)		
No.	Feature	Importance Score
1	Contract_Month-to-month	0.377104
2	InternetService_Fiber	0.218794
3	StreamingMovies_Yes	0.0413966
4	OnlineSecurity_No	0.0320359
5	TechSupport_No	0.0240612
6	Contract_Two year	0.0193264
7	Contract_One year	0.019087
8	InternetService_DSL	0.0172029
9	MultipleLines_Yes	0.017006
10	PhoneService	0.0162392

Tabel 3.14 *Feature* yang terpilih (*Lasso Regression + SMOTE + Holdout (90%, 10%)*)

<i>Lasso Regression + SMOTE + Holdout (90%, 10%)</i>		
No.	<i>Feature</i>	Coefficient
1	Dependents	0.205479
2	tenure	0.144516
3	PaperlessBilling	0.0905163
4	InternetService_Fiber optic	0.0768326
5	OnlineSecurity_No	0.0764348
6	OnlineBackup_No	0.0442784
7	TechSupport_No	0.018653
8	StreamingTV_Yes	0.010041
9	StreamingMovies_Yes	0.00740685
10	Contract_Month-to-month	-0.00964155
11	PaymentMethod_Electronic check	-0.243856

Tabel 3.15 *Feature* yang terpilih (*Lasso Regression + ROS + Holdout (90%, 10%)*)

<i>Lasso Regression + ROS + Holdout (90%, 10%)</i>		
No.	<i>Feature</i>	Coefficient
1	Dependents	0.205378
2	tenure	0.143848
3	PaperlessBilling	0.0878352
4	MultipleLines_No	0.0764764
5	InternetService_Fiber optic	0.0717387
6	OnlineSecurity_No	0.0321194
7	OnlineBackup_No	0.0217853
8	TechSupport_No	0.00554828
9	StreamingMovies_Yes	-0.00590237
10	Contract_Month-to-month	-0.00613283
11	Contract_Two year	-0.00651621
12	PaymentMethod_Electronic check	-0.222267

Tabel 3.16 *Top 10 Feature (mRMR + SMOTE + Holdout (90%, 10%))*

<i>mRMR + SMOTE + Holdout (90%, 10%)</i>		
No.	Feature	Relevance Score
1	Contract_Month-to-month	3404.56
2	MonthlyCharges	602.947
3	OnlineSecurity_No	2131.58
4	tenure	2305.76
5	TechSupport_No	2094.07
6	Contract_Two year	1927.85
7	PaymentMethod_Electronic check	1486.76
8	InternetService_Fiber optic	1543.86
9	OnlineBackup_No	1121.81
10	DeviceProtection_No	974.864

Tabel 3.17 *Top 10 Feature (mRMR + ROS + Holdout (90%, 10%))*

<i>mRMR + ROS + Holdout (90%, 10%)</i>		
No.	Feature	Relevance Score
1	Contract_Month-to-month	3169.35
2	MonthlyCharges	563.467
3	Contract_Two year	1849.15
4	OnlineSecurity_No	1872.61
5	tenure	2121.35
6	TechSupport_No	1919.76
7	PaymentMethod_Electronic check	1307.59
8	InternetService_Fiber optic	1442.35
9	OnlineBackup_No	1054.16
10	DeviceProtection_No	996.311

Tabel 3.18 *Top 10 Feature (XGBoost + SMOTE + 10-Fold)*

<i>XGBoost + SMOTE + K-Fold (K = 10)</i>		
No.	<i>Feature</i>	<i>Importance Score</i>
1	Contract_Month-to-month	0.343273
2	InternetService_Fiber optic	0.139904
3	OnlineSecurity_No	0.0417598
4	TechSupport_No	0.0356718
5	StreamingMovies_Yes	0.0302177
6	InternetService_DSL	0.0290379
7	PaymentMethod_Electronic check	0.0270609
8	Dependents	0.0250466
9	PaperlessBilling	0.0220574
10	Partner	0.0214875

Tabel 3.19 *Top 10 Feature (XGBoost + ROS + 10-Fold)*

<i>XGBoost + ROS + K-Fold (K = 10)</i>		
No.	<i>Feature</i>	<i>Importance Score</i>
1	Contract_Month-to-month	0.377104
2	InternetService_Fiber optic	0.218794
3	StreamingMovies_Yes	0.0413966
4	OnlineSecurity_No	0.0320359
5	TechSupport_No	0.0240612
6	Contract_Two year	0.0193264
7	Contract_One year	0.019087
8	InternetService_DSL	0.0172029
9	MultipleLines_Yes	0.017006
10	PhoneService	0.0162392

Tabel 3.20 *Feature* yang terpilih (*Lasso Regression + SMOTE + 10-Fold*)

<i>Lasso Regression + SMOTE + 10-Fold</i>		
No.	<i>Feature</i>	Coefficient
1	Dependents	0.205479
2	tenure	0.144516
3	PaperlessBilling	0.0905163
4	InternetService_Fiber optic	0.0768326
5	OnlineSecurity_No	0.0764348
6	OnlineBackup_No	0.0442784
7	TechSupport_No	0.018653
8	StreamingTV_Yes	0.010041
9	StreamingMovies_Yes	0.00740685
10	Contract_Month-to-month	-0.00964155
11	PaymentMethod_Electronic check	-0.243856

Tabel 3.21 *Feature* yang terpilih (*Lasso Regression + ROS + 10-Fold*)

<i>Lasso Regression + ROS + 10-Fold</i>		
No.	<i>Feature</i>	Coefficient
1	Dependents	0.205378
2	tenure	0.143848
3	PaperlessBilling	0.0878352
4	MultipleLines_No	0.0764764
5	InternetService_Fiber optic	0.0717387
6	OnlineSecurity_No	0.0321194
7	OnlineBackup_No	0.0217853
8	TechSupport_No	0.0055483
9	StreamingMovies_Yes	-0.0059024
10	Contract_Month-to-month	-0.0061328
11	Contract_Two year	-0.0065162
12	PaymentMethod_Electronic check	-0.222267

Tabel 3.22 Top 10 Feature (mRMR + SMOTE + 10-Fold)

<i>mRMR + SMOTE + 10-Fold</i>		
No.	Feature	Relevance Score
1	Contract_Month-to-month	3404.56
2	MonthlyCharges	602.947
3	OnlineSecurity_No	2131.58
4	tenure	2305.76
5	TechSupport_No	2094.07
6	Contract_Two year	1927.85
7	PaymentMethod_Electronic check	1486.76
8	InternetService_Fiber optic	1543.86
9	OnlineBackup_No	1121.81
10	DeviceProtection_No	974.864

Tabel 3.23 Top 10 Feature (mRMR + ROS + 10-Fold)

<i>mRMR + ROS + 10-Fold</i>		
No.	Feature	Relevance Score
1	Contract_Month-to-month	3169.35
2	MonthlyCharges	563.467
3	Contract_Two year	1849.15
4	OnlineSecurity_No	1872.61
5	tenure	2121.35
6	TechSupport_No	1919.76
7	PaymentMethod_Electronic check	1307.59
8	InternetService_Fiber optic	1442.35
9	OnlineBackup_No	1054.16
10	DeviceProtection_No	996.311

3.4 Metode Evaluasi

Evaluasi model merupakan hal penting dalam melatih sebuah model *Machine Learning* untuk mengukur performa dari model tersebut. Pada penelitian ini metode yang digunakan adalah *Confusion Matrix* dan Evaluasi Matrix.

3.4.1 *Confusion Matrix*

Confusion Matrix adalah tabel yang digunakan untuk mengevaluasi kinerja suatu model dengan menggambarkan jumlah prediksi yang benar dan salah pada data testing setelah pelatihan model. *Confusion Matrix* terdiri dari:

- a. *True Positive* (TP): Jumlah kasus positif yang diprediksi benar
- b. *True Negative* (TN): Jumlah kasus negatif yang diprediksi benar
- c. *False Positive* (FP): Jumlah kasus positif yang diprediksi salah
- d. *False Negative* (FN): Jumlah kasus negatif yang diprediksi salah

3.4.2 Evaluasi Matrix

Evaluasi Matrix adalah kumpulan matrix atau ukuran yang digunakan untuk mengevaluasi model. Evaluasi matrix yang digunakan pada penelitian ini adalah sebagai berikut:

- a. Akurasi: Mengukur total prediksi benar dari semua prediksi yang dibuat oleh model
- b. Presisi: Mengukur kasus positif yang diprediksi benar dari semua kasus yang diprediksi positif
- c. *Recall*: Mengukur seberapa baik model dalam menemukan semua kasus positif yang sebenarnya
- d. F-1 Score: Rata-rata harmonik dari presisi dan *recall*