JASIE "Jurnal Aplikasi Sistem Informasi dan Elektronika"
Vol. 5, No. 1, Juli 2023, Hal: 28-37
p-ISSN: 2714-612X, e-ISSN: 2798-1509

Perbandingan Optimasi Algoritma *Random Forest* Menggunakan Teknik *Boosting* Terhadap Kasus Klasifikasi *Churn* Pelanggan Di Industri Telekomunikasi

Noeril Agian Septa Dinata[1*], Ginanjar Abdurrahman[2], Nur Qodariyah Fitriyah[3]
Teknik Informatika, Fakultas Teknik, Universitas Muhammadiyah Jember[123]
Email: agianasd@gmail.com[1*], abdurrahmanginanjar@unmuhjember.ac.id[2], nurfitriyah@unmuhjember.ac.id[3]

ABSTRAK

Industri telekomunikasi berkembang sangat pesat dan perusahaan telekomunikasi terus melakukan berbagai inovasi untuk mendukung persaingan bisnis yang benar-benar sengit dan semakin sulit mendapatkan pelanggan. Persaingan ini menghasilkan *churn* pelanggan. *Churn* pelanggan yang tinggi adalah salah satu tingkat kegagalan perusahaan, oleh karena itu *churn* harus dikurangi. *Algoritma Random Forest* dipilih karena memiliki kemampuan untuk mengklasifikasikan data yang tidak lengkap dan dapat menangani data sampel yang besar. Tujuan khusus dari penelitian ini adalah untuk membandingkan kinerja algoritma *Random Forest* dengan optimasi menggunakan teknik *boosting* (*XGBoost* dan *AdaBoost*). Pada penelitian menggunakan *Upsampled* untuk mengatasi data yang tidak seimbang dan metode *interquartile range* dalam mengatasi pencilan. Hasil penelitian ini menunjukkan bahwa optimasi algoritma *Random Forest* menggunakan *boosting AdaBoost* menghasilkan kinerja yang paling optimal dengan hasil akurasi (0,9913), presisi (0,9831), *recall* (1,0) dan *F1-Score* (0,9915).

Kata Kunci: Telekomunikasi, *Churn*, Optimasi, *Random Forest*, *Boosting*, *XGBoost*, *AdaBoost*

ABSTRACT

The telecommunications industry is growing very fast and telecommunications companies continue to carry out various innovations to support business competition that is really fierce and increasingly difficult to get customers. This competition results in customer churn. High customer churn is one of the failure rates of the company, therefore churn must be reduced. The Random Forest algorithm was chosen because it has the ability to classify incomplete data and can handle large sample data. The general goal of the study is to predict which customers will switch services. While the specific purpose of this study is to compare the performance of the Random Forest algorithm with optimization using boosting techniques (XGBoost and AdaBoost). The results of this study optimized algorithms using AdaBoost boosting produced the most optimal performance with results of accuracy (0.9913), precision (0.9831), recall (1.0) and F1-Score (0.9915).

Keywords: *Telecomunication, Churn, Optimization, Random Forest, Boosting, XGBoost, Adaboost*

1. PENDAHULUAN

Dengan kemajuan teknologi informasi banyak perusahaan yang bergerak di bidang industri telekomunikasi. Industri telekomunikasi merupakan salah satu industri dengan pengguna terbanyak dan menempati urutan teratas dalam daftar industri dengan pertumbuhan tercepat, menempati 90% dari populasi (Hashmi dkk., 2013). Hal ini membuat perusahaan telekomunikasi terus menjalankan bermacam-macam inovasi untuk menunjang persaingan usaha yang sungguh ketat dan semakin sulit mendapatkan pelanggan. *Churn* pelanggan yang tinggi merupakan salah satu tingkat kegagalan perusahaan, oleh karena itu *churn* harus dikurangi. Perusahaan lebih memilih mempertahankan pelanggan untuk menghindari resiko *churn* karena biayanya lebih murah daripada menambah pelanggan baru (Arina & Ulfah, 2022). Pelanggan yang menghentikan layanan dan pelanggan beralih ke layanan perusahaan telekomunikasi lainnya. Perilaku pelanggan ini dikenal sebagai *churn*. *Churn* merupakan penghentian layanan telekomunikasi oleh pelanggan atau perusahaan (Saputra, 2021). Dalam mempertahankan pelanggan, perusahaan telekomunikasi memerlukan cara untuk memprediksi *churn* pelanggan. Klasifikasi pelanggan dapat dilakukan dengan menggunakan *machine learning*.

Sejalan dengan perkembangan teknologi kehadiran *machine learning* di bidang komputer telah menarik banyak perhatian. *Machine Learning* menjadi sebuah pengembangan algoritma dan model data itu sendiri. *Machine learning* memiliki peran penting dalam pembangunan, terutama di pembangunan analisis data (Mutmainnah dkk., 2018). Salah satu algoritma yang ada dalam *machine learning* yaitu klasifikasi.

Beberapa algoritma klasifikasi seperti *K-Nearest Neighbors* dan *Random Forest* telah diterapkan oleh penelitian terdahulu dengan judul "Pendekatan *Data Science* untuk Menemukan *Churn* Pelanggan pada Sector Perbankan dengan *Machine Learning*". Penelitian ini menggunakan *Random Forest* dasar dan menghasilkan akurasi lebih baik sebesar 86% dan algoritma *Random Forest* termasuk dalam algoritma populer, efektif dan memberikan prediksi yang relatif baik (Husein & Harahap, 2021). Pada penelitian selanjutnya dengan judul "Perbandingan Metode Klasifikasi *Supervised Learning* pada Data Bank *Customers* Menggunakan Python". Penelitian ini menggunakan klasifikasi *Supervised Learning* yaitu *Regresi Logistik, K-Nearst Neighboor, Super Vector Machine, Naïve bayes, Decision Tree*, dan *Random Forest.* Dari beberapa metode klasifikasi *Supervised* ini *Random Forest* menghasilkan nilai terbaik 0,862 (Pamungkas dkk., 2020).

Pada penelitian berikutnya dengan optimasi *boosting* yang berjudul "COVID-19 *Patient Health Prediction Using Boosted Random Forest Algorithm*" algoritma yang digunakan menggunakan *Random Forest* dan disempurnakan oleh teknik *boosting* yaitu AdaBoost menghasilkan akurasi 94% dan *F1-Score* 0,86 pada dataset yang digunakan (Iwendi dkk., 2020). Peneliti selanjutnya dengan judul "*Glioma Segmentation and Classification System Based on Proposed Textures Features Extraction Method and Hybrid Esemble Learning*". Metode yang digunakan SVM, *Naïve Bayes*, RF, KNN, ADBRF dan XGBRF. Hasil dari metode kombinasi XGB dengan RF (XGBRF) mendapatkan akurasi yang memuaskan sebesar 99,25% pada paduan urutan MRI T1C+T2+Flair dan akurasi 96.75% pada perpaduan urutan MRI T1+T1C+T2+Flair (Bhatele & Bhadauria, 2020).

Permasalahan diatas dapat disimpulkan bahwa penelitian ini akan melakukan klasifikasi *churn* pelanggan menggunakan algoritma *Random Forest*. Algoritma *Random Forest* dipilih karena memiliki kemampuan dalam mengklasifikasi data yang tidak lengkap dan serta dapat menangani data sampel yang banyak. *Random Forest* merupakan algoritma yang terdiri dari banyak *tree* dan memprediksi dengan cara *voting class* dari masing - masing *tree*, *class* dengan jumlah *vote* terbanyak akan menjadi final class (Rachmi, 2020). Teknik optimasi yang akan digunakan dalam penelitian ini adalah teknik *Boosting* menggunakan *XGBoost* dan *AdaBoost*. *XGBoost* dipilih karena memiliki kemampuan beradaptasi di berbagai situasi, fitur yang berguna mempercepat sistem perhitungan dan mencegah *overfitting* (Yulianti dkk., 2022). *Adaboost* digunakan karena sangat mudah diterapkan, tidak perlu mengatur parameter dan fleksibel sehingga dapat digabungkan dengan berbagai algoritma (Prasetio & Susanti, 2019).

2. KAJIAN PUSTAKA

A. *Data Science*

Menurut Fadli (2020), *data science* merupakan suatu proses yang dilakukan untuk menghasilkan pengetahuan data (*data insight*). Pengetahuan data tersebut merupakan sebuah kesimpulan yang dapat memberikan rekomendasi atau prediksi untuk kebutuhan tertentu. *Data scientist* adalah seseorang yang harus mampu melakukan penambangan data dengan mengekstraknya hingga menemukan data yang akurat yang dapat digunakan oleh para pemangku kebijakan. Sehingga dengan demikian seorang *data scientist* harus mampu mengidentifikasi permasalahan, mengumpulkan data dari berbagai sumber yang berbeda, mengatur informasi dan menerjemahkan hasil menjadi solusi.

B. *Churn* Pelanggan

Churn pelanggan adalah kecenderungan seorang pelanggan untuk meninggalkan satu penyedia layanan atau berpindah dari satu penyedia layanan ke penyedia layanan lainnya (Atthariq, 2020). Pelanggan *churn* menjadi masalah bagi sebagian besar perusahaan karena berdampak pada pendapatan perusahaan karena pelanggan beralih dari perusahaan penyedia layanan ke perusahaan lain di sektor telekomunikasi (El Kassem dkk., 2020).

C. Klasifikasi *Random Forest*

Klasifikasi merupakan proses untuk menemukan pola atau fitur menjelaskan atau membedakan konsep atau data kelas untuk tujuan estimasi objek kelas yang namanya tidak diketahui. Klasifikasi banyak digunakan dalam berbagai aplikasi termasuk deteksi penipuan, manajemen pelanggan, diagnosis medis, peramalan penjualan (Elfaladonna & Rahmadani, 2019).

Random Forest (RF) merupakan metode pembelajaran untuk klasifikasi dan regresi. Metode ini membuat serangkaian pohon keputusan pada saat yang sama dengan pelatihan. Mengklasifikasikan kasus baru dengan menugaskan kasus baru ke setiap pohon. Setiap pohon melakukan klasifikasi dan menghasilkan kelas. Kelas keluaran dipilih berdasarkan suara terbanyak, yaitu jumlah maksimum kelas serupa yang dihasilkan oleh berbagai pohon dianggap sebagai keluaran dari *Random Forest* (Andreyestha & Subekti, 2020).

D. *Boosting*

Seluruh konsep dengan *boosting* bekerja dengan menyusun kumpulan model secara berurutan dan kemudian menggabungkan seluruh model untuk menciptakan ekspektasi, model berikutnya memanfaatkan kesalahan model sebelumnya (Yulianti dkk., 2022). Berikut tipe - tipe *boosting*:

1. *XGBoost*

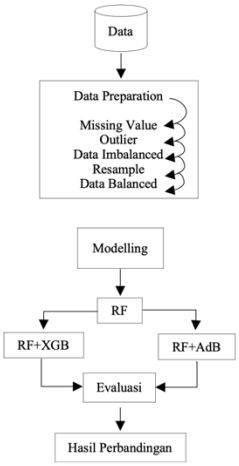
XGBoost merupakan pengembangan dari *Gradient Boosting*, metode ini dapat membantu mengoptimalkan algoritma dengan fleksibel terutama dalam regresi, klasifikasi dan perangkingan. *Extreme* yang tinggi dan fleksibilitas. *XGBoost* membantu kelancaran bobot terakhir yang dipelajari untuk menghindari *overfitting*. Selain mencegah *overfitting*, *XGBoost* juga mendukung pengambilan sampel baris dan kolom untuk memecahkan masalah. Eksplorasi model yang lebih cepat dimungkinkan sebagai paralel dan komputasi terdistribusi memastikan pembelajaran yang lebih cepat (Zhang, dkk., 2021).

2. *AdaBoost*

Adaptive Boosting (*Adaboost*) merupakan salah satunya metode dari algoritma *boosting*. Algoritma *AdaBoost* merupakan algoritma pertama yang ada pada teknik *boosting* yang masih digunakan dan dikembangkan. *Boosting* ini dapat gabungkan dengan algoritma klasifikasi lainnya untuk mengoptimalkan performa klasifikasi (Pristyanto, 2019).

3. METODE PENELITIAN

Tahapan pertama yang dilakukan sebelum pemrosesan data yang akan digunakan dalam penelitian yaitu pencarian data. Pencarian data dilakukan secara daring dan data pada penelitian ini diperoleh dari www.kaggle.com. Dataset berupa data *telecom churn* pada salah satu perusahaan telekomunikasi. *Churn* pelanggan adalah kecenderungan seorang pelanggan untuk meninggalkan satu penyedia layanan atau berpindah dari satu penyedia layanan ke penyedia layanan lainnya (Atthariq, 2020). Pelanggan *churn* menjadi masalah bagi sebagian besar perusahaan karena berdampak pada pendapatan perusahaan karena pelanggan beralih dari perusahaan penyedia layanan ke perusahaan lain di sektor telekomunikasi (EL Kassem dkk., 2020). Langkah-langkah dalam penelitian ini dapat dilihat pada Gambar 1.



Gambar 1. Kerangka Penelitian

Dataset ini terdiri dari 3333 baris dan 11 kolom. *Data Preparation* tahapan penyiapan data sebelum diproses untuk menghasilkan data dengan kualitas yang baik. Tahapan ini juga melakukan pemeriksaan data yaitu *missing value*, tipe data, duplikasi data, memeriksa *outlier* dan data *imbalanced*. Pada dataset *telecom churn* ini terdapat *outlier* dan data imbalanced. Munculnya nilai parameter ekstrim dalam kumpulan data dikenal sebagai keadaan *outlier*, masalah tersebut harus di atasi karena mengakibatkan kesalahan klasifikasi model, bias dalam estimasi parameter, hasil yang tidak akurat data dibawah standar semuanya dapat dipengaruhi oleh pencilan (*outlier*) dan *outlier* dapat diatasi dengan metode *interquartile range* (Siringoringo dkk., 2022). Ketidakseimbangan data berdampak pada temuan klasifikasi yang sering salah mengklasifikasikan kelompok minoritas sebagai kelas mayoritas (Siringoringo, 2018). Masalah ketidakseimbangan kelas dapat diselesaikan dengan SMOTE *Upsampling*. SMOTE *Upsampling* sangat efektif dalam meningkatkan kinerja sistem klasifikasi dengan memodifikasi kumpulan data yang tidak seimbang dengan menghasilkan kelas minoritas baru (Singgalen, 2022). *Feature scaling* merupakan tahapan *pre-processing* yang dikenal dengan standardisasi dapat membantu mempercepat perhitungan algoritma (Thara dkk., 2019). Standardisasi yang digunakan menggunakan *MinMaxScaller* karena *MinMaxScaler* berkinerja terbaik dibandingkan dengan *StandardScaler,Scale, RobustScaler, QuantileTransform, PowerTransform*, dan *MaxAbsScaler* (Raju dkk., 2020).

	Churn	AccountWeeks	ContractRenewal	DataPlan	DataUsage	CustServCalls	DayMins	DayCalls	MonthlyCharge	OverageFee	RoamMins
0	0	128	1	1	2.70	1	265.1	110	89.0	9.87	10.0
1	0	107	1	1	3.70	1	161.6	123	82.0	9.78	13.7
2	0	137	1	0	0.00	0	243.4	114	52.0	6.06	12.2
3	0	84	0	0	0.00	2	299.4	71	57.0	3.10	6.6
4	0	75	0	0	0.00	3	166.7	113	41.0	7.42	10.1
...	...	...	...	...	...	...	...	...	...	...	...
3328	0	192	1	1	2.67	2	156.2	77	71.7	10.78	9.9
3329	0	68	1	0	0.34	3	231.1	57	56.4	7.67	9.6
3330	0	28	1	0	0.00	2	180.8	109	56.0	14.44	14.1
3331	0	184	0	0	0.00	2	213.8	105	50.0	7.98	5.0
3332	0	74	1	1	3.70	0	234.4	113	100.0	13.30	13.7

3333 rows × 11 columns

Gambar 2. Data Sampel

Algoritma yang akan digunakan dalam penelitian ini yaitu *Random Forest* dan *Random Forest* menggunakan *boosting* (*XGBoost* dan *AdaBoost*). Data yang digunakan dibagi menjadi dua yaitu 80% data *training* dan 20% data *testing*. Rasio ini dipilih karena kumpulan data pelatihan yang lebih besar dapat mewakili seluruh kumpulan data dengan lebih baik (Wicaksono dkk., 2021). Dan penelitian ini juga menetapkan *random state* = 42, agar hasil tetap sama jika dites ulang.

Tahap evaluasi bertujuan untuk menilai efektivitas algoritma yang digunakan, dilakukan tahapan penilaian kebenaran hasil pemodelan menggunakan *confusion matrix*, akurasi, presisi, *recall* dan *F1-Score*. Selain itu, tahapan penelitian ini juga mencoba 2 percobaan yaitu penelitian dengan mengatasi masalah data dan tidak mengatasi masalah data. *Confusion matrix* merupakan salah satu cara untuk mengukur kinerja pembelajar. Metode ini dapat diterapkan pada masalah prediksi biner atau multikelas (Lukas dkk., 2022). Menghitung performa kinerja klasifikasi dapat dihitung menggunakan rumus berikut (Sitorus dkk., 2021):

Akurasi merupakan perbandingan antara data yang diklasifikasi dengan benar dan semua data, nilai akurasi menunjukkan seberapa baik sistem dapat mengklasifikasikan data. Perhitungan akurasi ditunjukkan pada rumus 1.

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

Presisi merupakan rasio jumlah data kategori positif yang dikategorikan dengan benar ke semua data kategori positif. Perhitungan nilai presisi ditunjukkan pada rumus 2.

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

Recall merupakan nilai untuk menampilkan proporsi data kelas positif yang berhasil diklasifikasi oleh algoritma. Perhitungan nilai *recall* ditunjukkan pada rumus 3.

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

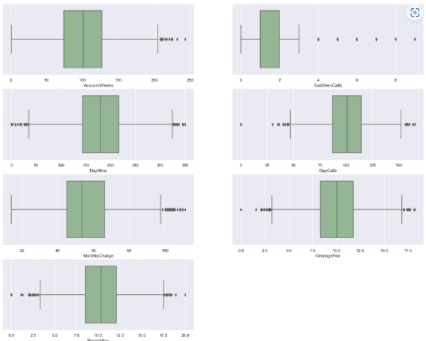
Rumus 4 adalah rumus perhitungan nilai *F1-Score*. *F1-Score* merupakan nilai rata-rata akurasi dan *recall*. Hasil terkecil pada *F1-Score* adalah 0, sedangkan nilai terbesar adalah 1,0. Jika nilai *F1-Score* tinggi, maka sistem kategorisasi dibuat dengan mempertimbangkan presisi dan *recall*.

$$F1\text{-}Score = \frac{2TP}{2TP+FP+FN} \quad (4)$$

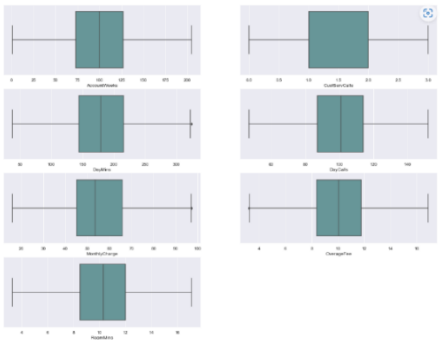
4. HASIL DAN PEMBAHASAN

A. Mengatasi *Outlier*

Pada data pelanggan *churn* ini terdapat *outlier* yang dapat dilihat dengan *BloxPot* dan untuk mengatasi *outlier* pada penelitian ini menggunakan metode *Interquartile Range* (IQR) dalam mendeteksi *outlier* dapat dilihat pada gambar 3 dan pembersihan data *outlier* dapat dilihat pada gambar 4 sehingga jumlah data menjadi (2906, 11).



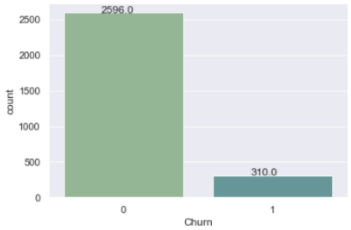
Gambar 3. Data *Outlier*



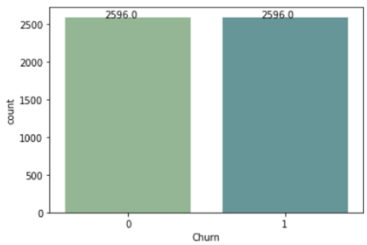
Gambar 4. Data Tanpa *Outlier*

B. Mengatasi *Data Imbalanced*

Pada data yang mengalami *imbalanced* dapat diatasi menggunakan metode *resample* dengan mengambil data *upsampled*. Grafik data terdapat *downsampled* dan *upsampled* yang dapat dilihat pada Gambar 4 dan data yang sudah diatasi dengan *Upsampled* dapat dilihat pada Gambar 5.



Gambar 5. Grafik *Data Imbalanced*



Gambar 6. Grafik *Data Balanced*

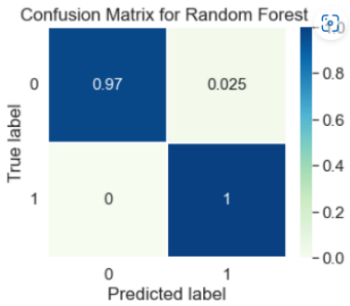
C. Klasifikasi Menggunakan *Random Forest*

Klasifikasi adalah proses dimana *classifier* mempelajari informasi berlabel dari sampel. Kemudian akurasi klasifikasi pengklasifikasi diuji dengan menggunakannya untuk menguji sampel data. Setiap sampel dalam pelatihan memiliki nilai target dan beberapa atribut (Silva dkk., 2021). *Random Forest* merupakan algoritma yang terdiri dari banyak *tree* dan memprediksi dengan cara *voting class* dari masing-masing *tree*, *class* dengan jumlah *vote* terbanyak akan menjadi final *class* (Rachmi, 2020). Kelas keluaran dipilih berdasarkan suara terbanyak, yaitu jumlah maksimum kelas serupa yang dihasilkan oleh berbagai pohon dianggap sebagai keluaran dari *Random Forest* (Andreyestha & Subekti, 2020). Pada pengujian yang dilakukan dengan algoritma *Random Forest* mendapatkan hasil seperti pada Tabel 1.

Tabel 1. Hasil *Random Forest*

Akurasi	0.9874
Presisi	0.9758
Recall	1.0
F1-Score	0.9877

Berdasarkan hasil pengujian yang telah dilakukan dengan algoritma *Random Forest* menunjukkan hasil akurasi (0.9874), presisi (0.9758), *recall* (1.0) dan *F1-Score* (0.9877).



Gambar 6. *Confusion Matrix Random Forest*

D. Klasifikasi Menggunakan *Random Forest* dengan *Extreme Gradien Boosting*

Extreme Gradient Boosting (*XGBoost*) memiliki keunggulan efisiensinya yang tinggi dan fleksibilitas. *XGBoost* membantu kelancaran bobot terakhir yang dipelajari untuk menghindari *overfitting*. Selain mencegah *overfitting*, *XGBoost* juga mendukung pengambilan sampel baris dan kolom untuk memecahkan masalah. Eksplorasi model yang lebih cepat dimungkinkan sebagai paralel dan komputasi terdistribusi memastikan pembelajaran yang lebih cepat (Zhang dkk., 2021). Pada pengujian yang dilakukan dengan algoritma *Random Forest* menggunakan teknik *boosting XGBoost* mendapatkan hasil seperti pada Tabel 2.

Tabel 2. Hasil *Random Forest* dengan *XGBoost.*

Akurasi	0.8873
Presisi	0.9232
Recall	0.8476
F1-Score	0.8838

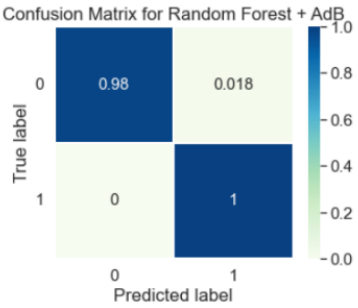
E. Klasifikasi Menggunakan *Random Forest* dengan *AdaBoost*

Algoritma *AdaBoost* merupakan algoritma pertama yang ada pada teknik *boosting* yang masih digunakan dan dikembangkan. *Boosting* ini dapat gabungkan dengan algoritma klasifikasi lainnya untuk mengoptimalkan performa klasifikasi (Pristyanto, 2019). *Adaptive boosting* (*AdaBoost*) adalah teknik yang membantu meningkatkan pengklasifikasi lemah menjadi satu pengklasifikasi yang kuat. Pembelajar yang lemah adalah pengklasifikasi yang menghasilkan prediksi yang lebih baik daripada menebak acak. Struktur *AdaBoost* bekerja dengan menerapkan algoritma klasifikasi secara berurutan ke bentuk kumpulan data pelatihan yang diberi bobot ulang. Selanjutnya, algoritma ini akan mengambil suara mayoritas tertimbang dari semua pengklasifikasi (Mirabdolbaghi dkk., 2022). Pada pengujian yang dilakukan dengan algoritma *Random Forest* menggunakan teknik *boosting* AdaBoost mendapatkan hasil seperti pada Tabel 3.

Tabel 3. Hasil *Random Forest* dengan *AdaBoost*

Akurasi	0.9913
Presisi	0.9831
Recall	1.0
F1-Score	0.9915

Berdasarkan hasil pengujian yang telah dilakukan dengan algoritma *Random Forest* menggunakan *boosting AdaBoost* menunjukkan hasil akurasi (0.9913), presisi (0.9831), *recall* (1.0) dan *F1-Score* (0.9915).



Gambar 7. *Confusion Matrix Random Forest* dengan *AdaBoost*

Setelah dilakukkan pengujian dari algoritma *Random Forest* menggunakan teknik *boosting* kemudian hasil yang didapatkan untuk menentukan ke dalam *good classification* dan paling baik dalam klasifikasi *churn* pelanggan, yang mana dari ketiganya menggunakan ukuran *split* data yang sama yaitu 80/20 dan parameter *random state* dengan jumlah *seed* 42 disetiap model agar saat pengujian hasil tidak berubah-ubah dan juga mengutamakan hasil optimasi original dari (*XGBoost* dan *AdaBoost*). Berikut adalah hasil perbandingannya:

Tabel 7. Tanpa Mengatasi *Outlier* dan *Data Imbalanced*

Model	Akurasi	Presisi	Recall	*F1-Score*
Random Forest	0.9265	0.8611	0.6138	0.7167
RF + XGBoost	0.9175	0.8484	0.5544	0.6706
RF + AdaBoost	0.9280	0.8732	0.6138	0.7209

Berdasarkan Tabel 7. dapat diliihat algoritma *Random Forest* menghasilkan performa yang cukup bagus, tetapi hasil klasifikasi tidak akurat jika dihitung manual menggunakan rumus *confusion matrix*, akurasi, presisi, *recall* dan *F1-Score* hal ini diakibatkan adanya *outlier* dan data *imbalanced* untuk hasil optimasi menggunakan teknik *boosting AdaBoost* dapat meningkatkan performa pada algoritma *Random Forest*.

Tabel 8. Mengatasi *Outlier* dan *Data Balanced*

Model	Akurasi	Presisi	*Recall*	*F1-Score*
Random Forest	0.9874	0.9758	1.0	0.9877
RF + XGBoost	0.8873	0.9232	0.8476	0.8838
RF + AdaBoost	0.9913	0.9831	1.0	0.9915

Berdasarkan Tabel 8. Dapat dilihat algoritma *Random Forest* dengan mengatasi *outlier* dan mengatasi data yang tidak seimbang menghasilkan akurasi 0.9874 dan hasil optimasi *Random Forest* oleh *boosting AdaBoost* menghasilkan performa yang sangat baik dengan akurasi 0.9913, presisi 0.9831, *recall* 1.0, *F1-Score* 0.9915 dan hasil klasifikasi yang baik. Pada Tabel ini dapat disimpulkan bahwa *Random Forest* menggunakan teknik *boosting Adaboost* dapat meningkatkan kinerja algoritma dalam klasifikasi dan mengatasi *outlier* dan data *imbalanced* ini dapat mencegah terjadinya kesalahan klasifikasi. Sedangkan saat *outlier* ditangani pada optimasi *XGBoost* mengalami penurunan setelah berkurangnya data latih dan data uji.

5. KESIMPULAN

Penelitian ini melakukan optimasi algoritma *Random Forest* menggunakan teknik *boosting* (*XGBoost* dan *AdaBoost*) untuk klasifikasi *churn* pelanggan pada industri telekomunikasi. Berdasarkan hasil percobaan pada penelitian ini, maka dapat ditarik kesimpulan bahwa algoritma *boosting AdaBoost* dapat meningkatkan performa paling optimal algoritma *Random Forest* dalam klasifikasi *churn* pelanggan dengan *train test* 80/20 dan jumlah *seed* 42 menghasilkan akurasi (0.9874), presisi (0.9758), *recall* (1.0) dan *F1-Score* (0.9877) menjadi akurasi (0.9913), presisi

(0.9831), *recall* (1.0) dan *F1-Score* (0.9915) pada data yang *balanced* dan *outlier* yang sudah diatas. Sedangkan hasil klasifikasi tidak mengatasi *outlier* dan data *imbalanced* menghasilkan klasifikasi yang tidak baik. Berdasarkan hasil kesimpulan penelitian ini dapat dikembangkan penelitian lebih lanjut menggunakan algoritma *Classification* yang lain seperti *Decision tree*, SVM, KNN, *Naïve bayes* dengan optimasi teknik *boosting Adaboost* dan *XGBoost*. Dan dalam menangani data yang *outlier* dan data *imbalanced*, dapat dilakukan percobaan normalisasi dengan *StandardScaler, Scale, RobustScaler* pada data yang mengalami *outlier* dan pada data *imbalanced* dapat melakukan *Downsampling*

6. DAFTAR PUSTAKA

Andreyestha & Subekti, A. (2020). Analisa Sentiment Pada Ulasan Film Dengan Optimasi Ensemble Learning. *Jurnal Informatika*, Vol.7 No.1, pp. 15–23.

Arina, F. & Ulfah, M. (2022). Analisa Survival Untuk Mengurangi Customer Churn Pada Perusahaan Telekomunikasi. *Journal Industrial Servicess*, vol. 8, no. 1, Juni 2022.

Atthariq, A. S. (2020). *Klasifikasi Customer Churn Berdasarkan Segmentasi Pelanggan Menggunakan Algoritma Naïve Bayes (Studi Kasus : Esl Express Tasikmalaya)*. Sarjana thesis, Universitas Siliwangi.

Bhatele, K. R. & Bhadauria, S. S. (2020). Glioma segmentation and classification system based on proposed texture features extraction method and hybrid ensemble learning. *Traitement du Signal*. Vol. 37, No. 6, December, 2020, pp. 989-1001.

Elfaladonna, F. & Rahmadani, A. (2019). Analisa Metode Classification-Decission Tree Dan Algoritma C.45 Untuk Memprediksi Penyakit Diabetes Dengan Menggunakan Aplikasi Rapid Miner. *SINTECH Journal*, vol. 2, no. 1, pp. 10–17.

El Kassem, E. A., Hussein, S. A., Abdelrahman, A. M., & Alsheref, F. K. (2020). Customer Churn Prediction Model and Identifying Features to Increase Customer Retention based on User Generated Content. *(IJACSA) International Journal of Advanced Computer Science and Applications*, Vol. 11, No. 5, pp. 522–531.

Fadli, A. (2020). Konsep Dasar Data Science. Komunitas eLearning IlmuKomputer, pp. 1–7.

Hashmi, N., Butt, N.A., & Iqbal, M. (2013). Customer Churn Prediction in Telecommunication: A Decade Review and Classification. *International Journal of Computer Science Issues*, Vol. 10, Issue 5, No 2, September 2013, pp. 271–282, 2013.

Husein, A.M., Harahap, M., & Fernandito, P. (2021). Pendekatan Data Science untuk Menemukan Churn Pelanggan Pada Sector Perbankan dengan Machine Learning. *Jurnal Data Science Indonesia* (DSI), vol. 1, no. 1, pp. 8–13, 2021.

Iwendi, C., Bashir, A.K., Peshkar, A., Sujatha, R., Chatterjee, J.M., Pasupuleti, S., Mishra, R., Pillai, S. & Jo, O. (2020). COVID-19 Patient Health Prediction Using Boosted Random Forest Algorithm. *Front. Public Health* 8:357

Lukas, S., Vigo, O., Krisnadi, D. & Widjaja, P. (2022). Perbandingan Performa Bagging Dan Adaboost Untuk Klasifikasi Data Multi-Class. *Journal Information System Development (ISD)*, vol 7, no 2, p. 7 - 12.

Mirabdolbaghi, S. M. S. & Amiri, B. (2022). Model Optimization Analysis of Customer Churn Prediction Using Machine Learning Algorithms with Focus on Feature Reductions. *Discrete Dynamics in Nature and Society*, vol. 2022.

Mutmainnah, S. (2020). *Optimasi Algoritma C4. 5 Menggunakan Teknik Bagging Pada Data Kadar Karat Emas*. Universitas Muhammadiyah Jember.

Pamungkas, F. S., Prasetya, B. D., & Kharisudin, I. (2020). Perbandingan Metode Klasifikasi Supervised Learning pada Data Bank Customers Menggunakan Python. *PRISMA, Prosiding Seminar Nasional Matematika*, 3, 692-697.

Prasetio, R. T. & Susanti, S. (2019). Prediksi Harapan Hidup Pasien Kanker Paru Pasca Operasi Bedah Toraks Menggunakan Boosted k-Nearest Neighbor. *Jurnal Responsif*, vol. 1, no. 1, pp.

64–69.

Pristyanto, Y. (2019). Penerapan Metode Ensemble Untuk Meningkatkan Kinerja Algoritme Klasifikasi Pada Imbalanced Dataset. *Jurnal Teknoinfo*, vol. 13, no. 1, p. 11.

Rachmi, A.N. (2020). *Implementasi Metode Random Forest Dan Xgboost Pada Klasifikasi Customer Churn.* Universitas Islam Indonesia.

Raju, V. N. G., Lakshmi, K. P., Jain, V. M., Kalidindi, A., & Padma, V. (2020). Study the Influence of Normalization/Transformation process on the Accuracy of Supervised Classification. *Third International Conference on Smart Systems and Inventive Technology (ICSSIT)*, Tirunelveli, India, pp. 729-735.

Saputra, F.D. (2021). *Prediksi Churn Dan Strategi Retensi Pada Kasus Perusahaan Telekomunikasi*. Tesis, pp. 1–120, 2021.

Silva, A., Kocayusufoglu, F., Jafarpour, S., Bullo, F., Swami, A. & Singh, A. (2021). Combining Physics and Machine Learning for Network Flow Estimation. *International Conference on Learning Representations*, vol. 1, pp. 1–18.

Singgalen, Y. (2022). Analisis Sentimen Wisatawan Melalui Data Ulasan Candi Borobudur di Tripadvisor Menggunakan Algoritma Naïve Bayes Classifier. *Building of Informatics, Technology and Science (BITS)*, 4(3), 1343−1352.

Siringoringo, R. (2018). Klasifikasi Data Tidak Seimbang Menggunakan Algoritma Smote Dan K-Nearest Neighbor. *Journal Information System Development* (ISD), vol. 3, no. 1, pp. 44–49.

Siringoringo, R., Angin, R. P., & Rumahorbo, B. (2022). Model Klasifikasi Genetic-XGBoost Dengan T-Distributed Stochastic Neighbor Embedding Pada Peramalan Pasar. *Jurnal Times*, vol. XI, no. 1, pp. 30–36.

Sitorus, Y. W., Sukarno, P., & Mandala, S. (2021). Analisis Deteksi Malware Android menggunakan metode Support Vector Machine & Random Forest. *e-Proceeding of Engineering*: Vol.8, No.6. pp. 12500–12518.

Thara D.K., PremaSudha, B. G., & Xiong, F. (2019). Auto-detection of epileptic seizure events using deep neural network with different feature scaling techniques. *Pattern Recognition Letters*, volume 128, p 544-550.

Wicaksono, A., Anita, & Padilah, T. N. (2021). Uji Performa Teknik Klasifikasi untuk Memprediksi Customer Churn. *Bianglala Inform*, vol. 9, no. 1, pp. 37–45.

Yulianti, S. E. H., Soesanto, O., & Sukmawaty, Y. (2022). Penerapan Metode Extreme Gradient Boosting (XGBOOST) pada Klasifikasi Nasabah Kartu Kredit. *Journal of Mathematics: Theory and Applications*, vol. 4, no. 1, pp. 21–26.

Zhang, W., Wu, C., Zhong, H., Li, Y., & Wang, L. (2021). Prediction of undrained shear strength using extreme gradient boosting and random forest based on Bayesian optimization. *Geoscience Frontiers*, vol. 12, no. 1, pp. 469–477.