Uji Performa Teknik Klasifikasi untuk Memprediksi *Customer Churn*

Anggito Wicaksono[1], Anita[2], Tesa Nur Padilah[3]
Universitas Singaperbangsa Karawang[1,2,3]
anggito.wicaksono17057@student.unsika.ac.id[1], anita.anita17008@student.unsika.ac.id[2],
tesa.nurpadilah@staff.unsika.ac.id[3]

Abstrak - Perkembangan industri telekomunikasi sangatlah cepat, hal ini dapat dilihat dari perilaku masyarakat yang menggunakan internet dalam berkomunikasi. Perilaku ini menyebabkan banyaknya perusahaan telekomunikasi dan meningkatnya *internet service provider* yang dapat menimbulkan persaingan antar *provider*. Pelanggan memiliki hak dalam memilih *provider* yang sesuai dan dapat beralih dari *provider* sebelumnya yang diartikan sebagai *customer churn*. Peralihan ini dapat menyebabkan berkurangnya pendapatan bagi perusahaan telekomunikasi sehingga penting untuk ditangani. Tujuan dari penelitian ini yaitu untuk mengetahui algoritme klasifikasi terbaik dan sesuai pada permasalahan *customer churn*. Penelitian ini dilakukan berdasarkan metode CRISP-DM sebagai alur penelitian dengan menerapkan tiga algoritme klasifikasi yaitu *Logistic Regression*, *Decision Tree*, dan *Random Forest*, yang dibantu dengan metode *feature selection* yaitu *Backward Elimination* untuk mengurangi variabel yang tidak signifikan. Hasil dari penelitian ini memperoleh bahwa algoritme *Logistic Regression* dengan *Backward Elimination* merupakan algoritme terbaik dengan nilai akurasi sebesar 82,23%, *recall* 57,22%, dan AUC sebesar 0,853 yang termasuk pada pemodelan *good classification*.
Kata Kunci : Industri Telekomunikasi, *Customer Churn*, CRISP-DM, Klasifikasi, *Backward Elimination*

Abstract *- The development of the telecommunications industry is very fast, it can be seen from the behavior of people who use the internet to communicate. This behavior causes the number of telecommunications companies and the increase in internet service provider which can lead to competition between providers. Customers have the right to choose the appropriate provider and can switch from the previous provider which is defined as customer churn. This transition can lead to reduced revenue for telecommunications companies so it is important to handle it. The purpose of this research is to find out the best and suitable classification algorithm for customer churn problems. This research was conducted based on the CRISP-DM method as a research flow by applying three classification algorithms, namely Logistic Regression, Decision Tree, and Random Forest, which was assisted by the feature selection method, namely Backward Elimination to reduce insignificant variables. The results of this study indicate that the Logistic Regression algorithm with Backward Elimination is the best algorithm with an accuracy value of 82.23%, a recall of 57.22%, and an AUC of 0.853 which is included in the good classification modeling.*
Keywords: Telecommunication Industry, Customer Churn, CRISP-DM, Classification, Backward Elimination

I. PENDAHULUAN

Industri telekomunikasi saat ini mengalami perkembangan yang sangat cepat. Hal ini menyebabkan bergesernya perilaku komunikasi masyarakat yang awalnya menggunakan telepon dan *short message service* (SMS) menjadi telekomunikasi berbasiskan data yang didukung oleh internet (Kementerian Komunikasi dan Informatika, 2018). Perkembangan ini mengakibatkan bertambahnya perusahaan telekomunikasi yang dapat memenuhi kebutuhan masyarakat seperti meningkatnya *internet service provider* (ISP). Hal ini tentunya juga dapat menimbulkan persaingan yang ketat antar operator ISP (Wardani & Ariasih, 2019). Persaingan tersebut mengakibatkan setiap operator tidak hanya memperhatikan perkembangan dari produk maupun layanan mereka, akan tetapi juga berfokus terhadap pelanggan, karena dalam hal ini pelanggan dapat memilih operator yang sesuai dengan kebutuhannya yang sewaktu-waktu juga dapat melakukan peralihan (*churn*).

Customer churn merupakan beralihnya pelanggan dari satu *provider* ke *provider* lain yang mengacu pada hilangnya pelanggan secara periodik dalam suatu organisasi sehingga dapat menyebabkan berkurangnya pendapatan secara signifikan bagi perusahaan telekomunikasi (Pamina et al., 2019). Hal ini tentunya merupakan suatu masalah bagi perusahaan tersebut. Untuk mempertahankan pelanggannya, perusahaan telekomunikasi harus meningkatkan produk dan layanannya, serta terlebih dahulu mengetahui pelanggan yang memiliki perilaku yang berkemungkinan akan meninggalkan layanan dari perusahaan (Kavitha, G. Kumar, S. Kumar, & Harish, 2020).

Prediksi *customer churn* merupakan cara untuk mengidentifikasi *churners* sebelum berpindah, yang dapat dilakukan dengan cara menganalisis data dan menemukan pola yang

berguna. Hal ini tentunya dapat dilakukan dengan pendekatan *data mining* (Herawati, Wibowo, & Mukhlash, 2016).

Data mining merupakan rangkaian proses untuk menambang dan mengolah data menjadi pengetahuan. Terdapat beberapa teknik *data mining* yang dapat digunakan untuk mengekstrak data menjadi pengetahuan, salah satunya klasifikasi (Drajana, 2019). Teknik klasifikasi merupakan teknik yang dapat memprediksi *customer churn* (Utami, Shofiana, & Heningtyas, 2020). Dalam melakukan prediksi, pendekatan ini memerlukan data-data masa lalu yang telah dikumpulkan (Yulianti, 2018).

Pada beberapa kasus mengenai prediksi *customer churn*, teknik klasifikasi yang umum digunakan yaitu *decision tree*, *rule-based learning*, dan *neural networks* yang terbukti bahwa ketiga algoritme tersebut dapat melakukan prediksi terhadap permasalahan *customer churn* (Kavitha et al., 2020).

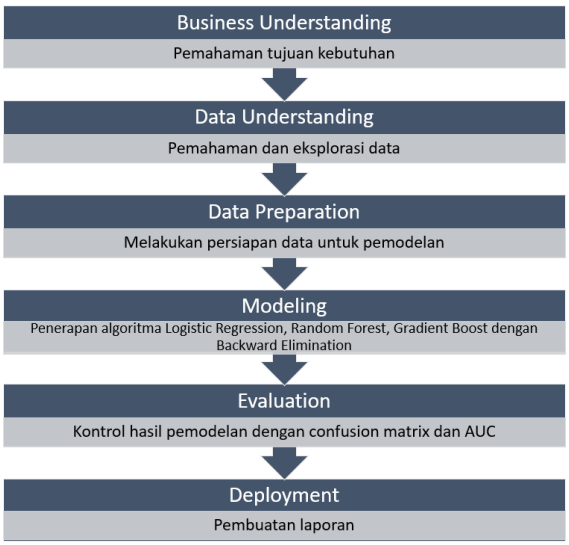
Pada penelitian sebelumnya, melakukan perbandingan terhadap tiga algoritme yaitu *random forest, logistic regression,* dan *XGBoost* untuk memprediksi *customer churn*. Penelitian ini mendapatkan hasil bahwa algoritme *random forest* memiliki akurasi sedikit lebih tinggi dari dua algorime lainnya (Kavitha et al., 2020). Penelitian lain, menggunakan algoritme *decision tree* untuk melakukan prediksi terhadap *customer churn,* selain sering digunakan pada pemodelan klasifikasi, algoritme ini juga mampu dalam memprediksi *customer churn* (Utami et al., 2020).

Pada penelitian ini, teknik *data mining* yang digunakan untuk melakukan prediksi *customer churn* yaitu teknik klasifikasi dengan menerapkan tiga algoritme yaitu *logistic regression* (LR), *decision tree* (DT), dan *random forest* (RF), serta menerapkan metode *feature selection* pada ketiga algoritme tersebut. Ketiga algoritme yang digunakan dipilih dan disesuaikan dengan kondisi dari *dataset* yaitu terdapat salah satu atribut yang berfungsi sebagai label atau kelas target. *Dataset* yang digunakan diambil dari *Public Dataset Kaggle* (*Kaggle*, 2018), yang diproses menggunakan bantuan *tools* RapidMiner.

Berdasarkan hal tersebut, tujuan dari penelitian ini yaitu menerapkan teknik klasifikasi dengan beberapa algoritme yang dipilih, sehingga ditemukan algoritme terbaik yang dapat digunakan untuk memprediksi *customer churn*.

II. METODOLOGI PENELITIAN

Penelitian ini didasarkan pada metode *Cross Industry Standard Process for Data Mining* (CRISP-DM) sebagai alur penelitian dari awal hingga akhir yang terdiri dari enam tahap seperti pada Gambar 1.



Gambar 1. Diagram Alir Penelitian

1. *Business Understanding* (Pemahaman Bisnis)

Tahap ini merupakan tahap awal untuk mengetahui masalah yang akan diselesaikan guna mencapai tujuan yang diinginkan.

2. *Data Understanding* (Pemahaman Data)

Tahap ini merupakan tahap pengumpulan data awal yang dipersiapkan untuk diolah dan dimodelkan, serta mendeteksi bagian data yang menarik sebagai hipotesis awal.

3. *Data Preparation* (Persiapan Data)

Tahap ini meliputi kegiatan yang dilakukan untuk menyiapkan data yaitu pemilihan data, integrasi, serta pembersihan data.

4. *Modelling* (Pemodelan)

Tahap ini merupakan tahap pemilihan teknik *data mining* dan algoritme yang akan digunakan dalam penelitian.

5. *Evaluation* (Evaluasi)

Tahap pengukuran keakuratan hasil dari pemodelan yang dilakukan untuk mengetahui performa dari algoritme yang diterapkan.

6. *Deployment* (Penyebaran)

Tahap penyusunan laporan berdasarkan pengetahuan yang diperoleh (Khumaidi, 2020).

III. HASIL DAN PEMBAHASAN

1. *Business Understanding* (Pemahaman Bisnis)

Pemahaman bisnis merupakan tahap awal dari metode CRISP-DM (Ramadhan & Kurniawati, 2020). Hasil dari pemahaman bisnis diubah menjadi rencana awal dan tujuan dilakukannya penelitian. Tujuan dari penelitian ini yaitu melakukan uji perbandingan algoritme klasifikasi yang diterapkan pada *dataset* pelanggan dari perusahaan telekomunikasi. Perbandingan algoritme ini dilakukan untuk mengetahui algoritme terbaik yang sesuai pada *dataset* pelanggan perusahaan telekomunikasi.

2. *Data Understanding* (Pemahaman Data)
   Data awal merupakan *dataset* yang berisi informasi pelanggan sebuah perusahaan telekomunikasi yang didapat dari situs web *Kaggle* dengan file berformat csv. *Dataset* yang didapat yaitu data sampel *International Business Machines* (IBM) dengan 7043 *record* dan 21 atribut yang dapat dilihat pada Tabel 1.

Setiap baris pada *dataset* mewakili pelanggan dan setiap kolom berisi atribut yang berupa informasi dari pelanggan. Informasi yang terdapat pada *dataset* yaitu sebagai berikut:

a. Pelanggan yang pergi dalam sebulan terakhir yaitu terdapat pada atribut *Churn*.
b. Layanan yang digunakan setiap pelanggan seperti telepon, *multiplelines*, internet, *online security*, *online backup*, *device protection*, *tech support*, serta *streaming* TV dan *movies*.
c. Informasi akun pelanggan seperti lama waktu menjadi pelanggan, kontrak, metode pembayaran, penagihan tanpa kertas, tagihan bulanan, dan biaya total.
d. Info demografis tentang pelanggan seperti usia, jenis kelamin, dan apakah pelanggan memiliki pasangan serta tanggungan.

3. *Data Preparation* (Persiapan Data)

Tahap persiapan data dilakukan untuk membangun *dataset* akhir yang akan digunakan untuk pemodelan (Fadillah, 2015).

a. *Attribute reduction*
   Atribut *customerID* dapat dihapus karena tidak diperlukan dalam proses pemodelan.
b. *Replace name*
   Replace dilakukan terhadap r*ecord* di beberapa atribut dengan isian yang diasumsikan bernilai sama.
   Berdasarkan Tabel terdapat beberapa pilihan yaitu '*Yes*', '*No*', '*No phone service*', dan '*No internet service*'. Untuk pilihan '*No phone service*' dan '*No internet service*' diasumsikan bahwa pelanggan tidak menggunakan layanan tersebut sehingga pilihan tersebut akan di-*rename* menjadi '*No*'.
c. *Data transformation*
   Atribut pada *dataset* yang berbentuk nominal diubah ke dalam bentuk numerik. Sedangkan label *Churn* tetap dalam bentuk nominal karena *dataset* akan diterapkan pada algoritme klasifikasi. Hal ini dapat dilihat pada Tabel .
d. *Handling missing value*
   Terdapat 11 *missing value* pada atribut *TotalCharges* yang merupakan data biaya yang dibebankan kepada pelanggan. Pada kasus ini masalah tersebut ditangani dengan menghapus *record* pada baris yang memiliki *missing value*. Penghapusan dilakukan karena jumlah *record* yang terdapat *missing value* tidak sampai 1% dari jumlah keseluruhan data, sehingga tidak akan memengaruhi hasil pemodelan secara signifikan (Chandranegara, Arifianto, & Wibowo, 2020).
   Dataset akhir setelah dilakukannya tahap *preparation* yaitu menjadi 7032 baris dengan 20 atribut serta seluruh atribut bertipe numerik dan label bertipe nominal yang dapat dilihat pada Tabel .

4. *Modeling* (Pemodelan)

Tahap pemodelan dilakukan dengan membandingkan tiga algoritme klasifikasi yaitu *logistic regression, decision tree, dan random forest* serta ditambahkan dengan metode *backward elimination* (BE). BE merupakan salah satu metode *feature selection* yang dimulai dengan mengambil semua prediktor, kemudian dilanjutkan dengan pengurangan variabel yang tidak signifikan hingga akhirnya sampai pada model yang optimal (Golestan & Hezarkhani, 2018). Dengan demikian, hasil akhir akan menampilkan enam perbandingan dari algoritme yang tidak menggunakan BE dan yang menggunakan BE.

Logistic regression merupakan algoritme yang digunakan untuk masalah klasifikasi biner yang pada labelnya hanya terdapat dua nilai. Setiap fitur dikalikan dengan koefisien regresi, kemudian fungsi sigmoid diperkenalkan, dan terakhir nilai dikeluarkan dalam interval 0 sampai 1. Jika nilainya lebih besar dari 0,5 maka label diklasifikasikan sebagai kelas 1. Jika kurang dari 0,5 maka label diklasifikasikan sebagai kelas 0 (X. Li & Z. Li, 2019).

Decision tree merupakan diagram aturan pengambilan keputusan yang digunakan untuk mengkategorikan subjek menjadi beberapa kelompok (E. Lee, Kim, & S. Lee, 2017). Pembuatan pohon keputusan dilakukan berdasarkan pemilihan atribut yang memiliki nilai *gain* tertinggi berdasarkan nilai *entropy* atribut sebagai poros atribut klasifikasi (Utami et al., 2020).

Random forest merupakan gabungan yang berasal dari teknik *classification and regression tree* (CART) yang berbasiskan pohon keputusan. *Random forest* memprediksi dengan menggabungkan sejumlah pohon. Hasil prediksi *random forest* dipilih berdasarkan kategori atau kelas yang paling sering muncul sebagai hasil prediksi dari sejumlah pohon klasifikasi (Ullah et al., 2019).

Uji perbandingan dari masing-masing model dilakukan menggunakan *split data* dengan rasio sebesar 0,8 sebagai *data training* dan 0,2 sebagai *data testing*. Rasio tersebut dipilih karena semakin besar *data training* maka dapat mewakili kumpulan data secara keseluruhan dengan karakteristik yang berbeda (*Machine*

Learning, 2020). Dengan demikian, dari 7032 baris *dataset* didapatkan sebanyak 5625 baris sebagai *data training* dan 1407 baris sebagai *data testing*.

5. *Evaluation* (Evaluasi)

Tahap evaluasi merupakan pengontrolan terhadap pemodelan yang sudah dilakukan untuk mendapatkan hasil yang sesuai dengan tujuan pada tahap *business understanding*. Pada penelitian ini dilakukan proses evaluasi dengan menggunakan *confusion matrix* dan nilai *area under curve* (AUC). *Confusion matrix* memberikan representasi ringkasan hasil prediksi dengan indeks *true positive* (TP), *true negative* (TN), *false positive* (FP), dan *false negative* (FN) (Zhu, Idemudia, & Feng, 2019).

Tabel 4. *Confusion Matrix*

	True No	*True Yes*
Prediction No	TN	FN
Prediction Yes	FP	TP

Sumber: Bisri & Rachmatika (2019)

Dimana:
TP = *True Positive* (data diprediksi positif dan aktual positif)
FP = *False Positive* (data diprediksi positif dan aktual negatif)
FN = *False Negative* (data diprediksi negatif dan aktual positif)
TN = *True Negative* (data diprediksi negatif dan aktual negatif)

Dari penyajian tabel *confusion matrix* pada Tabel 4 dapat dilakukan perhitungan untuk mengetahui nilai *accuracy* dan *recall* dengan persamaan sebagai berikut.

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

$$Recall = \frac{TP}{TP+FN} \quad \text{..... (2)}$$

Sedangkan nilai AUC diperlukan untuk mengetahui kriteria sebuah pemodelan dari teknik klasifikasi yang telah dilakukan (Bisri & Rachmatika, 2019). Penilaian AUC ditentukan dengan kriteria pernyataan keberhasilan sebagai berikut.

Tabel 5. Kriteria Pemodelan

Nilai	Kriteria
0,90 – 1,00	Sangat baik (*excellent classification*)
0,80 – 0,90	Baik (*good classification*)
0,70 – 0,80	Wajar (*fair classification*)
0,60 – 0,70	Buruk (*poor classification*)
< 0,60	Gagal (*failure*)

Sumber: Bisri & Rachmatika (2019)

a. *Logistic Regression*

Hasil pemodelan dengan *logistic regression* ditunjukkan pada Tabel 6. Berikut merupakan tabel *confusion matrix* hasil pemodelan algoritme *logistic regression.*

Tabel 7. *Confusion Matrix Logistic Regression*

	True No	*True Yes*
Prediction No	940	161
Prediction Yes	93	213

Berdasarkan Tabel 7, diperoleh nilai *accuracy* dan *recall* sebagai berikut.

$$Accuracy = \frac{213+940}{213+93+161+940}$$

$$= \frac{1153}{1407} = 0{,}8195 = 81{,}95\%$$

$$Recall = \frac{213}{213+161} = \frac{213}{374} = 0{,}5695 = 56{,}95\%$$

Nilai AUC pada model *logistic regression* yaitu sebesar 0,851.

b. *Decision Tree*

Hasil pemodelan dengan *decision tree* ditunjukkan pada Gambar 2. Berikut merupakan tabel *confusion matrix* hasil pemodelan algoritme *decision tree.*

Tabel 8. *Confusion Matrix Decision Tree*

	True No	*True Yes*
Prediction No	889	181
Prediction Yes	144	193

Berdasarkan Tabel 8, diperoleh nilai *accuracy* dan *recall* sebagai berikut.

$$Accuracy = \frac{193+889}{193+144+181+889}$$

$$= \frac{1082}{1407} = 0{,}7690 = 76{,}90\%$$

$$Recall = \frac{193}{193+181} = \frac{193}{374} = 0{,}5160 = 51{,}60\%$$

Nilai AUC pada model *decision tree* yaitu sebesar 0,779.

c. *Random Forest*

Hasil pemodelan *random forest* menghasilkan banyak *tree*, salah satunya ditunjukkan pada Gambar 3. Berikut merupakan tabel *confusion matrix* hasil pemodelan algoritme *random forest.*

Tabel 9. *Confusion Matrix Random Forest*

	True No	*True Yes*
Prediction No	984	230
Prediction Yes	49	144

Berdasarkan Tabel 9, diperoleh nilai *accuracy* dan *recall* sebagai berikut.

$$Accuracy = \frac{144+984}{144+49+230+984}$$

$$= \frac{1128}{1407} = 0{,}8017 = 80{,}17\%$$

$$Recall = \frac{144}{144+230} = \frac{144}{374} = 0{,}3850 = 38{,}50\%$$

Nilai AUC pada model *random forest* yaitu sebesar 0,847.

d. *Logistic Regression + BE*
Hasil pemodelan *logistic regression* dengan BE ditunjukkan pada Tabel 10.
Berikut merupakan tabel *confusion matrix* hasil pemodelan algoritme *logistic regression* dengan BE.

Tabel 11. *Confusion Matrix Logistic Regression* dengan BE

	True No	*True Yes*
Prediction No	943	160
Prediction Yes	90	214

Berdasarkan Tabel 11, diperoleh nilai *accuracy* dan *recall* sebagai berikut.

$$Accuracy = \frac{214+943}{214+90+160+943}$$

$$= \frac{1157}{1407} = 0{,}8223 = 82{,}23\%$$

$$Recall = \frac{214}{214+160} = \frac{214}{374} = 0{,}5722 = 57{,}22\%$$

Nilai AUC pada model *logistic regression* dengan BE yaitu sebesar 0,853.

e. *Decision Tree + BE*
Hasil pemodelan *decision tree* dengan BE ditunjukkan pada Tabel 12. Berikut merupakan tabel *confusion matrix* hasil pemodelan algoritme *decision tree* dengan BE.

Tabel 13. *Confusion Matrix Decision Tree* dengan BE

	True No	*True Yes*
Prediction No	972	222
Prediction Yes	61	152

Berdasarkan Tabel 13, diperoleh nilai *accuracy* dan *recall* sebagai berikut.

$$Accuracy = \frac{152+972}{152+61+222+972}$$

$$= \frac{1124}{1407} = 0{,}7989 = 79{,}89\%$$

$$Recall = \frac{152}{152+222} = \frac{152}{374} = 0{,}4064 = 40{,}64\%$$

Nilai AUC pada model *decision tree* dengan BE yaitu sebesar 0,825.

f. *Random Forest + BE*
Hasil pemodelan *random forest* dengan BE ditunjukkan pada Tabel 14. Berikut merupakan tabel *confusion matrix* hasil pemodelan algoritme *random forest* dengan BE.

Tabel 15. *Confusion Matrix Random Forest* dengan BE

	True No	*True Yes*
Prediction No	962	198
Prediction Yes	71	176

Berdasarkan Tabel 15, diperoleh nilai *accuracy* dan *recall* sebagai berikut.

$$Accuracy = \frac{176+962}{176+71+198+962}$$

$$= \frac{1138}{1407} = 0{,}8088 = 80{,}88\%$$

$$Recall = \frac{176}{176+198} = \frac{176}{374} = 0{,}4706 = 47{,}06\%$$

Nilai AUC pada model *random forest* dengan BE yaitu sebesar 0,851.

Berdasarkan enam pengujian yang telah dilakukan, dapat dilihat rekap nilai *accuracy*, *recall*, dan AUC dari masing-masing pemodelan pada gambar 11. Pemodelan menggunakan *logistic regression* dengan BE memperoleh *accuracy*, *recall*, dan AUC dengan nilai tertinggi dari pemodelan lainnya.
Berikut perbandingan hasil dengan penelitian terdahulu.

Tabel 16. Perbandingan dengan Penelitian Terdahulu

	Accuracy	***Recall***	**AUC**
Penelitian terdahulu			
LogReg	79%	56%	-
RF	80%	52%	-
DT	87.03%	96%	-
Penelitian sekarang tanpa BE			
LogReg	81,95%	56,95%	0,851
RF	80,17%	38,50%	0,847
DT	76,90%	51,60%	0,779
Penelitian sekarang + BE			
LogReg	82,23%	57,22%	0,853
RF	80,88%	47,06%	0,851
DT	79,89%	40,64%	0,825

6. *Deployment* (Penyebaran)
Tahap penyebaran merupakan bentuk laporan (*paper*) mengenai informasi yang didapat berdasarkan hasil pengamatan dan analisis perbandingan algoritme teknik klasifikasi untuk prediksi *customer churn*. Selain itu, penyebaran ini juga dapat digunakan sebagai bahan referensi untuk penelitian selanjutnya.

IV. KESIMPULAN

Berdasarkan penelitian yang telah dilakukan, diperoleh bahwa algoritme *logistic regression* dengan *backward elimination* dapat digunakan untuk memprediksi *customer churn.* Nilai akurasi yang diperoleh dengan algoritme ini yaitu sebesar 82,23% yang merupakan nilai akurasi tertinggi dari dua algoritme lainnya. Nilai *recall* yang diperoleh menunjukkan 57,22% bahwa pelanggan terklasifikasi benar melakukan *churn*. Nilai AUC yang didapatkan sebesar 0,853, hal ini berarti bahwa pemodelan teknik klasifikasi yang dilakukan berhasil dengan kriteria *good classification*.

V. REFERENSI

Bisri, A., & Rachmatika, R. (2019). Integrasi Gradient Boosted Trees dengan SMOTE dan Bagging untuk Deteksi Kelulusan

Mahasiswa. *Jurnal Nasional Teknik Elektro Dan Teknologi Informasi (JNTETI)*, *8*(4), 309–314. https://doi.org/10.22146/jnteti.v8i4.529

Chandranegara, D. R., Arifianto, S., & Wibowo, H. (2020). Analisa data pesawat terbang menggunakan metode elimination void data dan smoothing data. *Jurnal POROS TEKNIK*, *12*(1), 1–7.

Drajana, I. C. R. (2019). Prediksi Loyalitas Pelanggan IndiHome dengan Metode K-Nearest Neighbor. *Jurnal Sistem Informasi Dan Teknik Komputer*, *4*(2), 100–103. https://doi.org/10.4249/scholarpedia.1883

Fadillah, A. P. (2015). Penerapan Metode CRISP-DM untuk Prediksi Kelulusan Studi Mahasiswa Menempuh Mata Kuliah (Studi Kasus Universitas XYZ). *Jurnal Teknik Informatika Dan Sistem Informasi*, *1*(3), 260–270. https://doi.org/10.28932/jutisi.v1i3.406

Golestan, F. D., & Hezarkhani, A. (2018). Quadratic investigation of geochemical distribution by backward elimination approach at Glojeh epithermal Au(Ag)-polymetallic mineralization, NW Iran. *Journal of Central South University*, *25*(2), 342–356. https://doi.org/10.1007/s11771-018-3741-8

Herawati, M., Wibowo, I. L., & Mukhlash, I. (2016). Prediksi Customer Churn Menggunakan Algoritma Fuzzy Iterative Dichotomiser 3. *Limits: Journal of Mathematics and Its Applications*, *13*(1), 23–35. https://doi.org/10.12962/j1829605x.v13i1.1913

Kaggle. (2018). *Telco Customer Churn [Data file]*. San Francisco: Author.

Kavitha, V., Kumar, G. H., Kumar, S. M., & Harish, M. (2020). Churn Prediction of Customer in Telecom Industry using Machine Learning Algorithms. *International Journal of Engineering Research and Technology (IJERT)*, *9*(5), 181–184. https://doi.org/10.17577/ijertv9is050022

Kementerian Komunikasi dan Informatika. (2018). *Analisa Industri Telekomunikasi Indonesia untuk Mendukung Efisiensi*. https://balitbangsdm.kominfo.go.id/publikasi_465_3_199

Khumaidi, A. (2020). Data Mining for Predicting the Amount of Coffee Production Using CRISP-DM Method. *Jurnal Techno Nusa Mandiri*, *17*(1), 1–8. https://doi.org/10.33480/techno.v17i1.1240

Lee, E. B., Kim, J., & Lee, S. G. (2017). Predicting customer churn in mobile industry using data mining technology. *Industrial Management and Data Systems*, *117*(1), 90–109. https://doi.org/10.1108/IMDS-12-2015-0509

Li, X., & Li, Z. (2019). A Hybrid Prediction Model for E-Commerce Customer Churn Based on Logistic Regression and Extreme Gradient Boosting Algorithm. *International Information and Engineering Technology Association (IIETA)*, *24*(5), 525–530. https://doi.org/10.18280/isi.240510

Machine Learning. (2020). Https://Developers.Google.Com/Machine-Learning/Crash-Course/Training-and-Test-Sets/Splitting-Data.

Pamina, J., Beschi Raja, J., Sathya Bama, S., Soundarya, S., Sruthi, M. S., Kiruthika, S., Aiswaryadevi, V. J., & Priyanka, G. (2019). An Effective Classifier for Predicting Churn in Telecommunication. *Journal of Advanced Research in Dynamical and Control Systems*, *11*(1), 221–229.

Ullah, I., Raza, B., Malik, A. K., Imran, M., Islam, S. U., & Kim, S. W. (2019). A Churn Prediction Model Using Random Forest: Analysis of Machine Learning Techniques for Churn Prediction and Factor Identification in Telecom Sector. *IEEE Access*, *7*, 60134–60149. https://doi.org/10.1109/ACCESS.2019.2914999

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

Wardani, N. W., & Ariasih, N. K. (2019). Analisa Komparasi Algoritma Decision Tree C4 . 5 dan Naïve Bayes untuk Prediksi Churn Berdasarkan Kelas Pelanggan Retail. *International Journal of Natural Sciences and Engineering*, *3*(3), 103–112.

Yulianti. (2018). Metode Data Mining untuk Prediksi Churn Pelanggan. *Jurnal ICT Akademi Telkom Jakarta*, *17*, 46–52.

Zhu, C., Idemudia, C. U., & Feng, W. (2019). Improved logistic regression model for diabetes prediction by integrating PCA and K-means techniques. *Informatics in Medicine Unlocked*, *17*, 1–7. https://doi.org/10.1016/j.imu.2019.100179

Tabel 1. *Dataset* Pelanggan Telekomunikasi

customerID	gender	SeniorCitizen	Partner	Dependents	tenure	PhoneService	…	Churn
7590-VHVEG	Female	0	Yes	No	1	No	…	No
5575-GNVDE	Male	0	No	No	34	Yes	…	No
3668-QPYBK	Male	0	No	No	2	Yes	…	Yes
7795-CFOCW	Male	0	No	No	45	No	…	No
9237-HQITU	Female	0	No	No	2	Yes	…	Yes
9305-CDSKC	Female	0	No	No	8	Yes	…	Yes
1452-KIOVK	Male	0	No	Yes	22	Yes	…	No
6713-OKOMC	Female	0	No	No	10	No	…	No
…	…	…	…	…	…	…	…	…
3186-AJIEK	Male	0	No	No	66	Yes	…	No

Tabel 2. Atribut Yang Akan Di-*Rename*

MultipleLines	OnlineSecurity	OnlineBackup	DeviceProtection	TechSupport	StreamingTV	StreamingMovies
No	No	No	Yes	Yes	No	No
No	No	Yes	Yes	No	No	Yes
No Phone Service	No	No	Yes	No	No	Yes
No	No Internet Service	No Internet Service	No Internet Service	No Internet Service	No Internet Service	No Internet Service
No	No Internet Service	No Internet Service	No Internet Service	No Internet Service	No Internet Service	No Internet Service
Yes	No	Yes	No	Yes	No	No
No	Yes	Yes	No	Yes	No	No

Tabel 3. *Dataset* Akhir

Churn	Multiple Lines	Online Security	Online Backup	Device Protection	Tech Support	Stream-ingTV	Streaming Movies	gender	…	Total Charges
No	0	0	0	0	0	0	0	0	…	29.85
No	0	1	1	1	0	0	0	1	…	1889.5
Yes	0	1	0	0	0	0	0	1	…	108.15
No	0	1	1	1	1	0	0	1	…	1840.75
Yes	0	0	1	0	0	0	0	1	…	151.65
Yes	1	0	1	1	0	1	1	1	…	820.5
No	1	0	0	0	0	1	0	1	…	1949.4
No	0	1	1	0	0	0	0	1	…	301.9
…	…	…	…	…	…	…	…	…	…	…
No	0	1	1	1	1	1	1	1	…	6844.5

Tabel 6. Hasil Pemodelan *Logistic Regression*

Attribute	**Coefficient**	**Std. Coefficient**	**Std. Error**	**z-Value**	**p-Value**
MultipleLines	0.074	0.037	0.092	0.810	0.418
OnlineSecurity	-0.558	-0.252	0.096	-5.820	0.000
OnlineBackup	0.253	0.121	0.088	2.884	0.004
DeviceProtection	-0.241	-0.115	0.090	-2.667	0.008
TechSupport	-0.468	-0.213	0.095	-4.910	0.000
StreamingTV	-0.130	-0.063	0.097	-1.340	0.180
StreamingMovies	-0.017	-0.008	0.097	-0.171	0.864
gender	-0.026	-0.013	0.072	-0.354	0.723
Partner	-0.005	-0.003	0.086	-0.058	0.953
Dependents	-0.155	-0.071	0.100	-1.548	0.122
PhoneService	-1.239	-0.366	0.166	-7.472	0.000
InternetService	-0.041	-0.030	0.073	-0.562	0.574
Contract	-0.754	-0.627	0.087	-8.648	0
PaperlessBilling	-0.422	-0.208	0.083	-5.109	0.000
PaymentMethod	-0.094	-0.108	0.034	-2.767	0.006
SeniorCitizen	0.192	0.071	0.094	2.038	0.042
tenure	-0.061	-1.496	0.007	-8.769	0
MonthlyCharges	0.030	0.913	0.003	10.478	0
TotalCharges	0.000	0.780	0.000	4.352	0.000
Intercept	-0.300	-1.756	0.216	-1.391	0.164

Tree

```
Contract > 0.500
|   MonthlyCharges > 99.875
|   |   Contract > 1.500
|   |   |   TotalCharges > 7865.500: No {No=81, Yes=0}
|   |   |   TotalCharges ≤ 7865.500
|   |   |   |   tenure > 71.500: No {No=33, Yes=1}
|   |   |   |   tenure ≤ 71.500
|   |   |   |   |   TotalCharges > 7719.175
|   |   |   |   |   |   TotalCharges > 7737.975: No {No=5, Yes=2}
|   |   |   |   |   |   TotalCharges ≤ 7737.975: Yes {No=0, Yes=2}
|   |   |   |   |   TotalCharges ≤ 7719.175
|   |   |   |   |   |   MonthlyCharges > 105.950
|   |   |   |   |   |   |   MonthlyCharges > 108.825: No {No=44, Yes=3}
|   |   |   |   |   |   |   MonthlyCharges ≤ 108.825
|   |   |   |   |   |   |   |   OnlineBackup > 0.500: Yes {No=3, Yes=6}
|   |   |   |   |   |   |   |   OnlineBackup ≤ 0.500: No {No=10, Yes=3}
|   |   |   |   |   |   MonthlyCharges ≤ 105.950: No {No=52, Yes=2}
|   |   Contract ≤ 1.500
|   |   |   TotalCharges > 7961.050
|   |   |   |   Partner > 0.500: No {No=1, Yes=1}
|   |   |   |   Partner ≤ 0.500: Yes {No=0, Yes=3}
|   |   |   TotalCharges ≤ 7961.050
|   |   |   |   TotalCharges > 7044.175
|   |   |   |   |   MonthlyCharges > 108.275
|   |   |   |   |   |   TotalCharges > 7414.100
|   |   |   |   |   |   |   TotalCharges > 7829.150: No {No=4, Yes=0}
|   |   |   |   |   |   |   TotalCharges ≤ 7829.150
|   |   |   |   |   |   |   |   OnlineBackup > 0.500: No {No=2, Yes=0}
|   |   |   |   |   |   |   |   OnlineBackup ≤ 0.500: Yes {No=1, Yes=5}
|   |   |   |   |   |   TotalCharges ≤ 7414.100: No {No=14, Yes=1}
|   |   |   |   |   MonthlyCharges ≤ 108.275: No {No=30, Yes=1}
|   |   |   |   TotalCharges ≤ 7044.175
|   |   |   |   |   PaymentMethod > 0.500
|   |   |   |   |   |   DeviceProtection > 0.500
|   |   |   |   |   |   |   OnlineBackup > 0.500
|   |   |   |   |   |   |   |   MonthlyCharges > 105.850: Yes {No=2, Yes=4}
|   |   |   |   |   |   |   |   MonthlyCharges ≤ 105.850: No {No=18, Yes=4}
|   |   |   |   |   |   |   OnlineBackup ≤ 0.500: No {No=43, Yes=5}
|   |   |   |   |   |   DeviceProtection ≤ 0.500
|   |   |   |   |   |   |   TotalCharges > 6294.700: No {No=6, Yes=1}
|   |   |   |   |   |   |   TotalCharges ≤ 6294.700
|   |   |   |   |   |   |   |   tenure > 36.500: Yes {No=4, Yes=10}
|   |   |   |   |   |   |   |   tenure ≤ 36.500: No {No=2, Yes=0}
|   |   |   |   |   PaymentMethod ≤ 0.500
|   |   |   |   |   |   MonthlyCharges > 108.450
|   |   |   |   |   |   |   MonthlyCharges > 110.425
|   |   |   |   |   |   |   |   MonthlyCharges > 111.375: Yes {No=0, Yes=2}
```

Gambar 2. Hasil Pemodelan *Decision Tree*

Tree

```
MonthlyCharges > 27.550
|   TechSupport > 0.500
|   |   TotalCharges > 8678.625: Yes {No=0, Yes=1}
|   |   TotalCharges ≤ 8678.625
|   |   |   TotalCharges > 46.825
|   |   |   |   tenure > 7.500
|   |   |   |   |   tenure > 70.500
|   |   |   |   |   |   PaymentMethod > 0.500: No {No=230, Yes=0}
|   |   |   |   |   |   PaymentMethod ≤ 0.500
|   |   |   |   |   |   |   TotalCharges > 8046.975
|   |   |   |   |   |   |   |   MonthlyCharges > 112.075: No {No=4, Yes=0}
|   |   |   |   |   |   |   |   MonthlyCharges ≤ 112.075: Yes {No=0, Yes=2}
|   |   |   |   |   |   |   TotalCharges ≤ 8046.975: No {No=19, Yes=0}
|   |   |   |   |   tenure ≤ 70.500
|   |   |   |   |   |   InternetService > 0.500
|   |   |   |   |   |   |   MonthlyCharges > 117.025: Yes {No=0, Yes=4}
|   |   |   |   |   |   |   MonthlyCharges ≤ 117.025
|   |   |   |   |   |   |   |   tenure > 16.500: No {No=376, Yes=110}
|   |   |   |   |   |   |   |   tenure ≤ 16.500: Yes {No=17, Yes=27}
|   |   |   |   |   |   InternetService ≤ 0.500
|   |   |   |   |   |   |   TotalCharges > 295.225
|   |   |   |   |   |   |   |   DeviceProtection > 0.500: No {No=390, Yes=29}
|   |   |   |   |   |   |   |   DeviceProtection ≤ 0.500: No {No=277, Yes=32}
|   |   |   |   |   |   |   TotalCharges ≤ 295.225: Yes {No=0, Yes=3}
|   |   |   |   tenure ≤ 7.500
|   |   |   |   |   MultipleLines > 0.500
|   |   |   |   |   |   StreamingTV > 0.500
|   |   |   |   |   |   |   MonthlyCharges > 64.100
|   |   |   |   |   |   |   |   PaymentMethod > 1.500: Yes {No=0, Yes=1}
|   |   |   |   |   |   |   |   PaymentMethod ≤ 1.500: No {No=2, Yes=0}
|   |   |   |   |   |   |   MonthlyCharges ≤ 64.100: Yes {No=0, Yes=1}
|   |   |   |   |   |   StreamingTV ≤ 0.500
|   |   |   |   |   |   |   OnlineBackup > 0.500: Yes {No=3, Yes=9}
|   |   |   |   |   |   |   OnlineBackup ≤ 0.500: Yes {No=0, Yes=4}
|   |   |   |   |   MultipleLines ≤ 0.500
|   |   |   |   |   |   SeniorCitizen > 0.500: No {No=4, Yes=0}
|   |   |   |   |   |   SeniorCitizen ≤ 0.500
|   |   |   |   |   |   |   TotalCharges > 49.475
|   |   |   |   |   |   |   |   tenure > 6.500: Yes {No=3, Yes=9}
|   |   |   |   |   |   |   |   tenure ≤ 6.500: No {No=50, Yes=50}
|   |   |   |   |   |   |   TotalCharges ≤ 49.475: No {No=2, Yes=0}
|   |   |   TotalCharges ≤ 46.825
|   |   |   |   StreamingTV > 0.500: Yes {No=0, Yes=2}
|   |   |   |   StreamingTV ≤ 0.500
|   |   |   |   |   TotalCharges > 40.275: Yes {No=0, Yes=1}
|   |   |   |   |   TotalCharges ≤ 40.275: No {No=1, Yes=0}
```

Gambar 3. Salah Satu *Tree* Hasil Pemodelan *Random Forest*

Tabel 10. Hasil Pemodelan *Logistic Regression* dengan BE

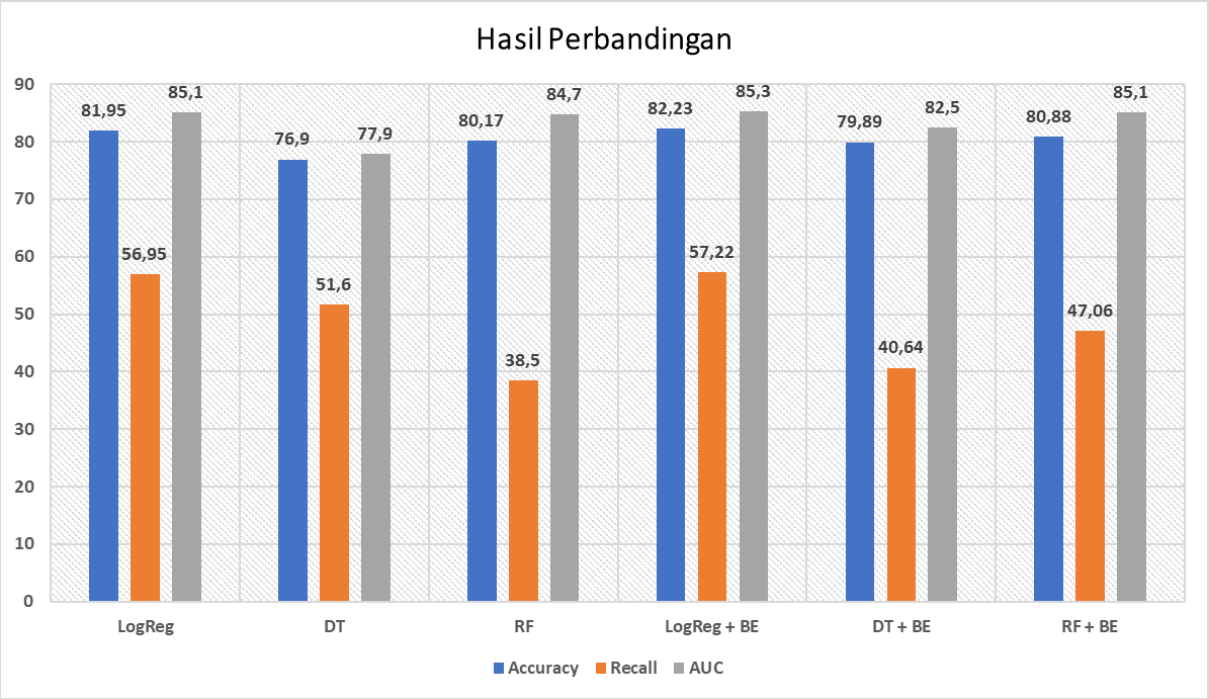
Attribute	Weight
MultipleLines	1
OnlineSecurity	1
OnlineBackup	1
DeviceProtection	1
TechSupport	1
StreamingTV	1
StreamingMovies	1
gender	1
Partner	1
Dependents	0
PhoneService	1
InternetService	1
Contract	1
PaperlessBilling	1
PaymentMethod	1
SeniorCitizen	1
tenure	1
MonthlyCharges	1
TotalCharges	0

Tabel 12. Hasil Pemodelan Decision Tree dengan BE

Attribute	Weight
MultipleLines	1
OnlineSecurity	1
OnlineBackup	1
DeviceProtection	1
TechSupport	1
StreamingTV	1
StreamingMovies	1
gender	1
Partner	1
Dependents	1
PhoneService	1
InternetService	1
Contract	0
PaperlessBilling	1
PaymentMethod	1
SeniorCitizen	1
tenure	1
MonthlyCharges	1
TotalCharges	1

Tabel 14. Hasil Pemodelan *Random Forest* dengan BE

Attribute	Weight
MultipleLines	1
OnlineSecurity	1
OnlineBackup	1
DeviceProtection	1
TechSupport	0
StreamingTV	1
StreamingMovies	1
gender	1
Partner	1
Dependents	0
PhoneService	1
InternetService	1
Contract	1
PaperlessBilling	1
PaymentMethod	1
SeniorCitizen	1
tenure	1
MonthlyCharges	1
TotalCharges	1



Gambar 4. Hasil Perbandingan Pemodelan