PREDIKSI TINGKAT PELANGGAN *CHURN* PADA PERUSAHAAN TELEKOMUNIKASI DENGAN ALGORITMA ADABOOST

Iqbal Muhammad Latief[1], Agus Subekti[2], Windu Gata[3]
[123]Prodi Ilmu Komputer, Sekolah Tinggi Ilmu Manajemen dan Ilmu Komputer Nusa Mandiri
Jl. Jatiwaringin No. 2, Cipinang Melayu, Makasar Jakarta Timur
Telp. (021) 8005722
e-mail : [1]14002323@nusamandiri.ac.id, [2]agus@nusamandiri.ac.id, [3]windu@nusamandiri.ac.id

ABSTRACT

With the rapid advancement of the telecommunications industry, and competition between telecommunications companies is increasing, companies need to predict their customers to determine the level of customer loyalty. One of them is by analyzing customer data by doing a Customer Churn Prediction. Predicting Customer Churn is an important business strategy for the company. To acquire new customers is much higher cost than retaining existing customers. The ease of operator switching is one of the serious challenges that the telecommunications industry must face. By predicting customer churn, companies can take immediate action to retain customers. To retain existing customers, the company must improve customer service, improve product quality, and must know in advance which customers have the possibility to leave the company. Prediction can be done by analyzing customer data using data mining techniques. In line with this, gathering information from the telecommunications business can help predict whether customer relationships will leave the company. The data used in this study are secondary data and amount to 7.403 data customers. The data has 21 variables. This study proposes to use the ensemble method namely adaboost, xgboost and random forest and compare them. Algorithm is validated through training data and testing data with a ratio of 80:20. From the results we got using python tools, it was found that the adaboost algorithm has an accuracy of 80%.

Keywords—*accuracy, adaboost, churn prediction, compare model, data mining.*

ABSTRAK

Dengan pesatnya kemajuan industri telekomunikasi, dan persaingan antar perusahaan telekomunikasi semakin meningkat, perusahaan perlu untuk memprediksi pelanggannya untuk mengetahui tingkat loyalitas pelanggan. Salah satunya adalah dengan menganalisis data pelanggan dengan melakukan prediksi *churn* pelanggan. Prediksi *churn* pelanggan adalah strategi bisnis yang penting bagi perusahaan. Untuk mendapatkan pelanggan baru membutuhkan biaya yang jauh lebih tinggi daripada mempertahankan pelanggan yang sudah ada. Kemudahan operator *switching* merupakan salah satu tantangan serius yang harus dihadapi oleh industri telekomunikasi. Dengan memprediksi pelanggan *churn*, perusahaan dapat secara efektif mengambil keputusan untuk mempertahankannya. Untuk mempertahankan pelanggan yang ada, perusahaan harus meningkatkan layanan pelanggan, meningkatkan kualitas produk, dan harus mengetahui sebelumnya pelanggan mana yang memiliki kemungkinan untuk keluar dari perusahaan. Prediksi dapat dilakukan dengan menganalisis data pelanggan menggunakan teknik *data mining*. Sejalan dengan ini, dengan

mengumpulkan informasi dari bisnis telekomunikasi dapat membantu memprediksi hubungan pelanggan apakah mereka akan meninggalkan perusahaan. Data yang digunakan dalam penelitian ini adalah data sekunder dan berjumlah 7.403 data pelanggan. Data memiliki 21 variabel. Studi ini mengusulkan untuk menggunakan metode *ensemble* yaitu adaboost, xgboost dan random forest dan membandingkannya. Algoritma divalidasi melalui data *training* dan data *testing* dengan rasio 80:20. Dari hasil yang kami dapatkan dengan menggunakan alat bantu python maka ditemukan algoritma adaboost memiliki akurasi tertinggi yaitu 80%.

***Kata Kunci*—**adaboost, akurasi, data mining, komparasi model, prediksi churn.

I. PENDAHULUAN

Bidang telekomunikasi telah menjadi salah satu bisnis fundamental di negara-negara maju. Seiring dengan itu perusahaan baru mulai bermunculan. Munculnya perusahaan baru ini mengakibatkan persaingan yang semakin ketat. Perusahaan mana pun dapat berkembang jika memiliki jumlah pelanggan yang memadai. Memiliki jumlah pelanggan yang besar memungkinkan pendapatan yang optimal untuk kas perusahaan. Berbagai cara akan dilakukan perusahaan untuk menarik pelanggan tersebut agar menggunakan jasanya. Salah satu cara untuk menarik pelanggan adalah dengan menganalisis data pelanggan yang biasanya disimpan di sejumlah besar database perusahaan. Data ini dapat digunakan untuk menganalisis pelanggan mana yang loyal dan *churn*.

Penelitian tentang *churn* pelanggan semakin penting dan menarik banyak perhatian peneliti [1]. Pelanggan dapat dengan mudah menggunakan haknya untuk berpindah penyedia layanan dari satu operator ke operator lainnya. Banyaknya operator seluler mendorong persaingan bisnis yang semakin ketat. Kemudahan operator *switching* merupakan salah satu tantangan serius yang harus dihadapi oleh industri telekomuniksi [2]. Mengingat fakta bahwa industri telekomunikasi mengalami tingkat *churn* tahunan rata-rata 30-35 persen, dan biaya untuk merekrut pelanggan baru 5-10 kali lebih mahal daripada mempertahankan pelanggan yang sudah ada, mempertahankan pelanggan menjadi lebih penting daripada mendapatkan pelanggan [3]. *Churn* pelanggan mengacu pada kehilangan pelanggan berkala dalam sebuah organisasi [4].

Industri telekomunikasi berusaha mengembangkan cara untuk memprediksi pelanggan yang berpotensi *churn* sehingga dapat dilakukan tindakan pencegahan karena pengaruh *direct churn* terhadap penurunan pendapatan perusahaan [5]. Aktivitas *churn* sangat berpengaruh terhadap total profit dan citra bisnis,

sehingga sebaiknya dapat diprediksi dan dicegah [6]. Prediksi *churn* dapat digunakan untuk mengidentifikasi *churnes* lebih awal sebelum mereka pindah, dan dapat membantu departemen CRM (*Customer Relationship Management*) untuk mempertahankannya, sehingga potensi kerugian perusahaan dapat dihindari [7]. Dengan demikian penyedia layanan harus mengoptimalkan kinerja model prediksi *churn* dan menerapkan teknik prediksi *churn* serta menerapkan strategi pemasaran yang tepat untuk mempertahankan pelanggan yang ada [8].

Masalah yang dihadapi adalah bagaimana menganalisis pelanggan mana yang loyal atau *churn*. Dalam studi yang dilakukan oleh Keramati disebutkan bahwa untuk bertahan dalam bisnis telekomunikasi harus mampu membedakan antara pelanggan yang memiliki kemungkinan pindah ke pesaing, dan pelanggan yang enggan pindah [9]. Oleh karena itu, prediksi *churn* pelanggan menjadi isu penting dalam bisnis telekomunikasi. Dalam bisnis yang kompetitif, prediktor pelanggan yang andal dianggap sangat berharga.

Proses menghasilkan informasi dari kumpulan data disebut *data mining*. Contoh penelitian yang telah dilakukan dengan *churn analysis* adalah pada perusahaan televisi berlangganan [10], perusahaan retail [11], dan lain-lain. Diantara metode *data mining* yang telah digunakan untuk menganalisis pelanggan adalah random forest dan decision tree [12] tetapi belum mencapai nilai yang sangat baik. Pada penelitian kali ini akan mencoba untuk mendapatkan akurasi yang lebih baik dari penelitian yang dilakukan sebelumnya yang berjudul *Telecom Customer Churn Prediction* dengan hasil akurasi tertinggi dengan algoritma random forest [12].

Dalam penelitian ini menerapkan teknik klasifikasi *data mining* yang meliputi random forest, adaboost dan xgboost, kemudian membandingkan performanya.

Pada penelitian ini akan mencoba menjawab bagaimana pengklasifikasian data yang paling berpengaruh terhadap tingkat *churn* pada perusahaan telekomunikasi menggunakan metode *ensemble* yaitu random forest, adaboost dan xgboost serta bagaimana akurasi dari ketiga algoritma tersebut.

II. METODE PENELITIAN

Dalam penelitian ini prosedur kerja penelitian ini adalah sebagai berikut:

A) Pemilihan Data

Dataset dalam penelitian ini adalah data sekunder. Dataset tersebut berisi data pelanggan di sebuah perusahaan

telekomunikasi. Dataset ini diunduh dari Kaggle dengan alamat website yaitu https://www.kaggle.com/blastchar/telco-customer-churn. Dataset yang digunakan berukuran 955 KB. Dataset ini memiliki 7.403 data pelanggan dan 21 kolom. 21 kolom tersebut adalah variabel yang akan digunakan sebagai prediksi *churn*. Variabel yang terdapat dalam dataset *churn* adalah sebagai berikut [13]:

A. *Customer ID*: Indeks pelanggan

B. *Gender*: Jenis kelamin pelanggan. Kolom ini memiliki 2 nilai: *male* dan *female.*

C. *SeniorCitizen*: Pelanggan adalah warga negara senior. Kolom ini memiliki 2 nilai yaitu 0 dan 1.

D. *Partners*: Pelanggan memiliki mitra. Kolom ini memiliki 2 nilai, yaitu: *Yes* dan *No.*

E. *Depedents*: Pelanggan memiliki tanggungan. Kolom ini memiliki 2 nilai, yaitu: *Yes* dan *No.*

F. *Tenure*: Jumlah bulan pelanggan menggunakan layanan perusahaan.

G. *PhoneService*: Pelanggan memiliki layanan telepon. Kolom ini memiliki 2 nilai, yaitu *Yes* dan *No.*

H. *MultipleLines*: Pelanggan memiliki layanan multi saluran. Kolom ini memiliki 2 nilai yaitu *Yes* dan *No.*

I. *InternetService*: Penyedia layanan internet pelanggan. Kolom ini mempunyai 3 nilai yaitu DSL, *Fiber Optic* dan *No.*

J. *OnlineSecurity*: Pelanggan memiliki keamanan online. Kolom ini memiliki 3 nilai yaitu *Yes, No* dan *No Internet Service.*

K. *OnlineBackup*: Pelanggan memiliki layanan tampilan *online*. Kolom ini memiliki 3 nilai yaitu *Yes, No* dan *No Internet Service*.

L. *Device Protection*: Pelanggan memiliki layanan perlindungan perangkat. Kolom ini memiliki 3 nilai yaitu *Yes, No* dan *No Internet Service*.

M. *TechSupport*: Pelanggan memiliki dukungan teknis. Kolom ini memiliki 3 nilai yaitu *Yes, No* dan *No Internet Service*.

N. *StreamingTV*: Pelanggan memiliki layanan streaming televisi. Kolom ini memiliki 3 nilai yaitu *Yes, No* dan *No Internet Service*.

O. *StreamingMovies*: Pelanggan memiliki layanan *streaming* film. Kolom ini memiliki 3 nilai yaitu *Yes, No* dan *No Internet Service*.

P. *Contract*: Persyaratan kontrak pelanggan. Kolom ini memiliki 3 nilai yaitu: *Month-to-month, One year, Two year*.

Q. *PaperlessBilling*: Pelanggan memiliki tagihan tanpa kertas. Kolom ini memiliki 2 nilai yaitu: *Yes* dan *No.*

R. *PaymentMethod*: Metode pembayaran pelanggan. Kolom ini memiliki 4 nilai yaitu: *Electronic check, Mailed check, Bank transfer (automatic)* dan *Credit card (automatic).*

S. *MonthlyCharges*: Ini adalah jumlah yang dibebankan kepada pelanggan setiap bulan.

T. *TotalCharges*: Jumlah total layanan yang dibebankan kepada pelanggan.

U. *Churn*: Kategori pelanggan *churn* atau tidak. Kolom ini memiliki 2 nilai yaitu *Yes* dan *No.*

B) Preprocessing

Setelah data terkumpul, proses selanjutnya adalah proses *preprocessing*. *Preprocessing* data merupakan proses yang bertujuan untuk mengubah data menjadi format yang lebih mudah dan efektif bagi pengguna. Beberapa metode *preprocessing* data yang kami gunakan adalah [14]:

A. Pembersihan data, adalah proses pembersihan data yang memiliki *missing value*.

B. Penyesuaian data, adalah proses menyesuaikan jumlah data untuk setiap target.

C. Pemisahan data, merupakan proses pemisahan data menjadi dua kelompok yaitu *train* dan *test*.

C) Transformation

Transformation adalah proses mengubah data yang dipilih, sehingga data tersebut sesuai untuk proses *data mining*. Proses ini merupakan proses kreatif dan sangat bergantung pada jenis atau pola informasi yang akan dicari dalam database. Dalam proses ini, data akan dikelompokkan menggunakan metode *One Hot Encoding*.

D) *Data mining*

Data mining adalah proses mencari pola atau informasi yang menarik dalam data yang dipilih menggunakan teknik atau metode tertentu [15]. Teknik, metode, atau algoritma di dalam *data mining* itu banyak sekali variasinya. Penelitian ini menggunakan algoritma random forest, adaboost dan xgboost.

Algoritma Adaboost

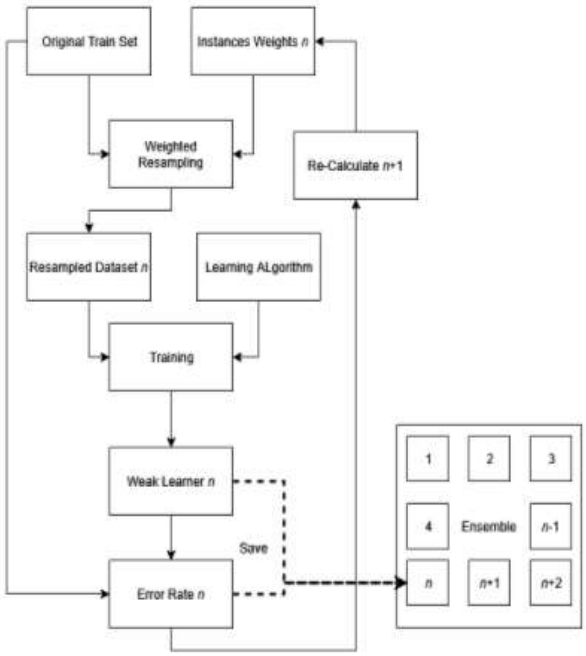
Adaboost adalah salah satu algoritma *boosting* yang paling populer. Mirip dengan *bagging*, ide utama dibalik algoritma adalah membuat sejumlah *weak learners* yang tidak berkolerasi dan kemudian menggabungkan prediksinya. Perbedaan utama dengan *bagging* adalah bahwa alih-alih membuat sejumlah rangkaian-rangkaian *bootsrap* independen, algoritma tersebut secara berurutan melatih setiap *weak learner*, menetapkan bobot ke semua *instance*, dan mengulangi

seluruh proses. Sebagai algoritma *base learner*, biasanya *decision tree* yang terdiri dari satu *node* digunakan. *Decision tree* ini, dengan kedalaman satu tingkat, disebut *decision stumps.*

Algoritma adaboost dapat dideskripsikan secara *high-level* dari langkah dasarnya, langkah-langkahnya:

1) Inisialisasi semua bobot *instance set train* secara merata, sehingga jumlahnya sama dengan 1.
2) Hasilkan set baru dengan pangambilan sampel dengan penggantian, sesuai dengan bobotnya.
3) *Train* sebuah *weak learner* pada set sampel.
4) Hitung *error* pada *train set* asli.
5) Tambahkan *weak learner* ke *ensemble* dan simpan tingkat kesalahannya.
6) Sesuaikan bobot, tambah bobot *instance* yang salah diklasifikasikan, dan kurangi bobot *instance* yang diklasifikasikan dengan benar.
7) Ulangi dari langkah 2.
8) *Weak learners* digabungkan dengan *voting*. *Vote* setiap *learner* diberi bobot sesuai tingkat kesalahannya [16].

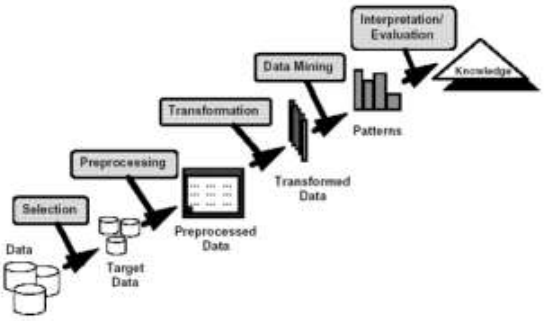
Seluruh proses digambarkan dalam gambar berikut:



Gambar 1. Langkah-Langkah Algoritma Adaboost

E) Interpretasi/Evaluasi

Tahapan mulai dari menyeleksi data yaitu melihat *missing value*, kemudian melakukan transformasi kebentuk *numeric* kemudian mencari mendapatkan model terbaik dan terakhir pola informasi yang dihasilkan dari proses *data mining* perlu ditampilkan dalam bentuk yang mudah dipahami oleh pihak yang berkepentingan. Pada tahap ini, pengetahuan yang dihasilkan dari *data mining* akan dirilis.



Gambar 2. Proses *Data Mining*

III. HASIL

Objek yang digunakan dalam penelitian ini adalah data pelanggan pada perusahaan telekomunikasi. Data pelanggan diperoleh dengan mengunduh di Kaggle pada alamat https://www.kaggle.com/blastchar/telco-customer-churn. Dataset ini berisi 7.043 data pelanggan. Berikut 5 contoh data pelanggan dalam dataset tersebut.

	customerID	gender	SeniorCitizen	Partner	Dependents
0	7590-VHVEG	Female	0	Yes	No
1	5575-GNVDE	Male	0	No	No
2	3668-QPYBK	Male	0	No	No
3	7795-CFOCW	Male	0	No	No
4	9237-HQITU	Female	0	No	No

Gambar 3. Contoh 5 Data Pelanggan dari Dataset

tenure	PhoneService	MultipleLines	InternetService	OnlineSecurity
1	No	No phone service	DSL	No
34	Yes	No	DSL	Yes
2	Yes	No	DSL	Yes
45	No	No phone service	DSL	Yes
2	Yes	No	Fiber optic	No

Gambar 4. Lanjutan Contoh 5 Data Pelanggan dari Dataset

DeviceProtection	TechSupport	StreamingTV	StreamingMovies	Contract
No	No	No	No	Month-to-month
Yes	No	No	No	One year
No	No	No	No	Month-to-month
Yes	Yes	No	No	One year
No	No	No	No	Month-to-month

Gambar 5. Lanjutan Contoh 5 Data Pelanggan dari Dataset

PaperlessBilling	PaymentMethod	MonthlyCharges	TotalCharges	Churn
Yes	Electronic check	29.85	29.85	No
No	Mailed check	56.95	1889.5	No
Yes	Mailed check	53.85	108.15	Yes
No	Bank transfer (automatic)	42.30	1840.75	No
Yes	Electronic check	70.70	151.65	Yes

Gambar 6. Lanjutan Contoh 5 Data Pelanggan dari Dataset

Pertama menemukan *missing values* dari data, kemudian mengisi nya dengan 0 (nol). Terdapat 11 data *missing values* pada variabel *TotalCharges*.

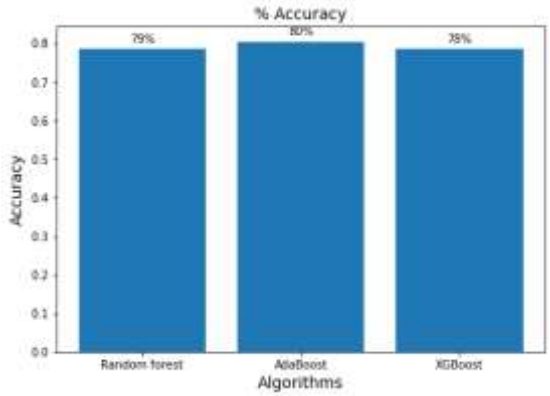
Sasaran pada penelitian ini adalah kolom *Churn*. Jumlah data untuk target "No" adalah 73,4% dan "Yes" adalah 26,6%.

Sebelum membentuk aturan klasifikasi, perlu dilakukan pembagian data menjadi 2 kelompok yaitu *train* dan *test.* Berbagi data ini bertujuan untuk menganalisis apakah aturan klasifikasi yang dihasilkan oleh algoritma adaboost dapat digunakan untuk memprediksi

churn. Distribusi dataset ini menggunakan rasio 80% data *training* dan 20% data *testing* .

Kemudian setelah itu dilakukan pemodelan dengan algoritma random forest, adaboost dan xgboost dengan parameter *default*.

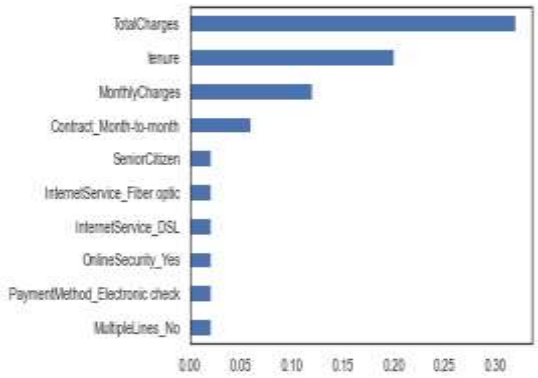
Setalah pemodelan kemudian dilakukan penghitungan akurasi dari masing-masing algoritma, maka didapat hasil seperti gambar 7.



Gambar 7. Perbandingan Ketiga Algoritma

Dari gambar 7 dapat dilihat bahwa adaboost memiliki akurasi tertinggi yaiu 80%.

Dari Algoritma adaboost, *Total Charges, tenure, MonthlyCharges,* adalah variabel prediktor paling penting untuk memprediksi *Churn*.



Gambar 8. Variabel Penting yang Mempengaruhi *Churn* dengan Algoritma Adaboost

Akurasi yang dihasilkan dari pengujian aturan klasifikasi ini dari data pengujiannya adalah 80%.

Dari aturan klasifikasi yang dibentuk menggunakan algoritma adaboost, variabel *TotalCharges* memiliki pengaruh cukup kuat terhadap prediksi *churn* ini.

Dengan mengetahui bahwa variabel *TotalCharges* memiliki pengaruh yang cukup besar dalam memprediksi *churn*, maka perusahaan dapat melakukan promosi yang lebih baik kepada pelanggan dengan adanya *TotalCharges* yang berpotensi *churn*.

IV. KESIMPULAN

Setelah melakukan penelitian ini, dapat ditarik kesimpulan bahwa algoritma adaboost dapat memprediksi masalah *churn* lebih baik dari algoritma random forest dan xgboost serta *TotalCharges* adalah fitur yang paling penting dalam

memprediksi *churn* dengan tingkat akurasi 80% dari pada penelitian sebelumnya dengan algoritma random forest [12].

DAFTAR PUSTAKA

[1] E. B. Lee, J. Kim, and S. G. Lee, "Predicting customer churn in mobile industry using data mining technology," *Ind. Manag. Data Syst.*, vol. 117, pp. 90–109, 2017, doi: 10.1108/IMDS-12-2015-0509.

[2] B. Huang, M. T. Kechadi, and B. Buckley, "Customer churn prediction in telecommunications," *Expert Syst. Appl.*, vol. 39, 2012, doi: 10.1016/j.eswa.2011.08.024.

[3] J. Lu and D. Ph, "Predicting Customer Churn in the Telecommunications Industry — An Application of Survival Analysis Modeling Using SAS," *Techniques*, 2002.

[4] A. Churi, M. Divekar, S. Dashpute, and P. Kamble, "Analysis of customer churn in mobile industry using data mining.," *Int. J. Emerg. Technol. Adv. Eng.*, vol. 5(3), pp. 225–230, 2015.

[5] A. K. Ahmad, A. Jafar, and K. Aljoumaa, "Customer churn prediction in telecom using machine learning in big data platform," *J. Big Data*, 2019, doi: 10.1186/s40537-019-0191-6.

[6] J. Pamina, T. Dhiliphan Rajkumar, S. Kiruthika, T. Suganya, and F. Femila, "Exploring hybrid and ensemble models for customer churn prediction in telecom sector," *Int. J. Recent Technol. Eng.*, 2019, doi: 10.35940/ijrte.A9170.078219.

[7] V. Umayaparvathi and K. Iyakutti, "Applications of Data Mining Techniques in Telecom Churn Prediction," *Int. J. Comput. Appl.*, 2012, doi: 10.5120/5814-8122.

[8] R. Misra, S. Singh, and R. Mahajan, "An empirical study on the cellular subscribers churn, selection factors and satisfaction with the services," 2019, doi: 10.1504/ijpd.2019.10020377.

[9] A. Keramati, R. Jafari-Marandi, M. Aliannejadi, I. Ahmadian, M. Mozaffari, and U. Abbasi, "Improved churn prediction in telecommunication industry using data mining techniques," *Appl. Soft Comput. J.*, 2014, doi: 10.1016/j.asoc.2014.08.041.

[10] N. Suryana, "PREDIKSI CHURN DAN SEGMENTASI PELANGGAN TV BERLANGGANAN (STUDI KASUS TRANSVISION JAWA BARAT)," *J. TEDC*, vol. 11(2), no.

Vol 11 No 2 (2017): Jurnal TEDC, pp. 185–191, 2019, [Online]. Available: http://ejournal.poltektedc.ac.id/index.php/tedc/article/view/77.

[11] N. W. Wardani, G. R. Dantes, and G. Indrawan, "Prediksi Customer Churn dengan Algoritma Decision Tree C4.5 Berdasarkan Segmentasi Pelanggan untuk Mempertahankan Pelanggan pada Perusahaan Retail," *J. Resist. (Rekayasa Sist. Komputer)*, 2018, doi: 10.31598/jurnalresistor.v1i1.219.

[12] M. Manasa, M. N. Reddy, M. L. Sahithi, P. Y. Kumar, and V. Sandhya, "Telecom Customer Churn Prediction," vol. 8, 2020, doi: 10.22214/ijraset.2020.5479.

[13] Kaggle, "Telco Customer Churn: Focused customer retention programs," 2018. https://www.kaggle.com/blastchar/telco-customer-churn (accessed Nov. 06, 2020).

[14] R. R. Rerung, "Penerapan Data Mining dengan Memanfaatkan Metode Association Rule untuk Promosi Produk," *J. Teknol. Rekayasa*, 2018, doi: 10.31544/jtera.v3.i1.2018.89-98.

[15] E. Srikanti, R. F. Yansi, Norvahina, I. Permana, and F. N. Salisah, "Penerapan Data Mining Untuk Menganalisis Penjualan Barang dengan Menggunakan Metode Apriori pada Supermarket Sejahtera Lhoksumawe," *J. Ilm. Rekayasa dan Manaj. Sist. Inf.*, 2018.

[16] G. Kyriakides, *Hands-On Ensemble Learning with Python: Build Highly Optimized Ensemble Machine Learning Models Using Scikit-Learn and Keras*. Birmingham: Packt Publishing Ltd., 2019.

[17] Sabita, H., & Herwanto, R. (2020). Pantauan Prediktif Covid-19 Dengan Menggunakan Metode SIR dan Model Statistik Di Indonesia. *TEKNIKA*, *14*(2), 145-150.