Prediksi Customer Churn Menggunakan Logistic Regression dan Decission Tree

Adhy Mauludin Nur Aziz [1], Adni Mauliddin[2], Virgiantara Armanda Sintalana,[3] Rafi Daryl Hafiz[4], Ali Akbar Rismayadi[5]
[1,4,5]Program Studi Sistem Informasi, Universitas Adhirajasa Reswara Sanjaya, Bandung
[2,3]Program Studi Teknik Informatika, Universitas Adhirajasa Reswara Sanjaya, Bandung
e-mail: [1] adhyadhyan31@gmail.com, [2] adni.unin@gmail.com , [3] vasintalana@gmail.com, [4]rafidaryl@gmail.com, [5]ali@ars.ac.id

Abstrak

Pesatnya perkembangan teknologi mendorong setiap orang untuk menggunakan internet sebagai media telekomunikasi, hal ini menjadi peluang bagi perusahaan telekomunikasi untuk bersaing menarik perhatian pelanggan, dan pelanggan pun berhak menentukan perusahaan atau provider mana yang mereka pilih untuk digunakan serta pelanggan dengan bebas beralih ke perusahaan telekomunikasi atau provider lain. Peralihan keputusan pelanggan *(Customer Churn)* salah satunya dapat terjadi dikarenakan ketidakpuasan satu dan lain hal terkait kebijakan yang diberikan pihak perusahaan. Oleh karena itu perusahaan telekomunikasi harus mengantisipasi agar tidak kehilangan pelangganya. Tujuan dari penelititan ini untuk mengetahui model klasifikasi yang lebih baik dari segi tingkat akurasinya. Dalam penelitian ini penulis menggunakan data dari Kaggle yang terdiri dari 7043 baris data dan dengan kolom *Churn* sebagai variable target yang dikalsifikasikan menggunakan *Logistic Regression* dan *Decission Tree*. Hasil dari penelitian ini memperoleh bahwa akurasi dari model klasifikasi Logistic Regression sebesar 80% dan akurasi model klasifikasi *Decission Tree* sebesar 72%.

Kata kunci *Customer Churn, Logistic Regression, Decission Tree*

Abstract

Technological developments encourage everyone to use the internet as a means of telecommunication, this is an opportunity for telecommunications companies to compete to recruit customers, and customers has the right to determine which company or provider they choose to use and customers are free to switch to other telecommunications companies or providers. One of the changes in customer decisions (Customer Churn) can occur due to dissatisfaction with one thing or another regarding the policies provided by the company. Therefore, telecommunications companies must anticipate so as not to lose their customers. The purpose of this research is to find out a better classification model in terms of the level of accuracy. In this study, the author uses data from Kaggle which consists of 7043 rows of data and with the Churn column as the target variable which is classified using Logistic Regression and Decission Tree. The results of this study obtained that the accuracy of the Logistic Regression classification model is 80% and the accuracy of the Decission Tree classification model is 72%.

Keywords *Customer Churn, Logistic Regression, Decission Tree*

Corresponding Author:
Ali Akbar Rismayadi
Email: ali@ars.ac.id

1. PENDAHULUAN

Telekomunikasi sudah menjadi bagian pokok bagi kehidupan baik untuk komunikasi, memperoleh informasi, bekerja, bahkan untuk sarana hiburan. Hal demikian membuat perusahaan telekomunikasi bersaing ketat untuk mendapatkan pelanggan, sehingga perusahaan mencari cara agar sebisa mungkin memprtahanka pelanggan mereka bukan hanya memperhatikan perkembangan produk atau layanan mereka [1]. Pelanggan sangat penting bagi perusahaan karena pelangganlah yang menggunakan jasa atau produk yang dihasilkan oleh perusahaan sehingga tanpa adanya perusahaan tidak akan mendapatkan keuntungan [2].

Customer Churn menyebabkan perusahaan mengalami kerugian. Dengan adanya model yang dapat memprediksi kapan dan mengapa pelanggan berhenti berlanggan maka perusahaan mencegah itu sehigga perusahaan dapat terhidar dari dari kerugian [3].

Dalam studi yang dilakukan oleh Keramati dalam jurnal Iqbal Muhammad Latief dkk [4] disebutkan bahwa untuk bertahan dalam bisnis telekomunikasi harus mampu membedakan antara pelanggan yang memiliki kemungkinan pindah ke pesaing, dan pelanggan yang enggan pindah. Dengan demikian perusahaan harus menganalisis apakah pelanggan *churn* atau tidak.

Tingkat *churn* yang tinggi akan berdampak buruk pada laba perusahaan dan menghambat pertumbuhan. Prediksi *churn* kami akan dapat memberikan kejelasan kepada perusahaan telekomunikasi tentang seberapa baik mempertahankan pelanggan yang ada dan memahami apa alasan mendasar yang menyebabkan pelanggan lama memutuskan kontrak mereka (tingkat *churn* tinggi).

Salah satu penelitian sebelumnya yang berjudul Analisis *Churn Prediction* Menggunakan Metode *Logistic Regression* Dan Smote *(Synthetic Minority Over-Sampling Technique)* Pada Perusahaan Telekomunikasi dengan menggunakn data dari WITEL PT.Telekomunikasi Regional [5]. Penelitian yang dilakukan menggunakan metode logistic regression dan penanganan *imbalance* data dengan SMOTE memiliki hasil performan dengan tingkat akurasi sebesar 92,4% dan *f1-measure* sebesar 31,27% [6].

Penelitian selanjutnya yang dilakukan oleh Iqbal Muhammad Latief dkk [7] yang berjudul Prediksi Tingkat Pelanggan *Churn* Pada Perusahaan Telekomunikasi Dengan Algoritma *adaboost*. Kesimpulan yang didapat dari penelitiannya yaitu 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%.

Machine Learning merupakan teknik untuk mengolah data menghasilkan pola-pola tertentu agar mudah dianalisis dan diterapkan dalam system. *Machine learning* dapat digunakan untuk mengklasifikasikan jenis data. Klasifikasi data bertujuan untuk memprediksi label pada kumpulan data [8]. Dalam klasifikasi algortma decision tree lebih baik dari algortma Naïve bayes dan K-NN [9].

Dalam penelitian ini, penulis menggunakan model *Logistic Regression* dan *Decission Tree* kemudian membandingkan hasil dari kedua model tersebut untuk mengetahui model mana yang paling baik dari tingkat akurasinya.

2. METODE PENELITIAN

2.1. Dataset

Pada penelitian ini, penulis mengambil satu dataset yang bersumber dari Kaggle. Dataset tersebut berisi tentang data pelanggan disebuah perusahaan telekomunikasi yang memiliki 7.403 baris data dan 21 kolom (variabel). Untuk pemilihan variabel target, penulis menggunakan kolom atau *variable churn* sebagai variabel targetnya

2.2. Customer Churn

Customer Churn adalah dimana *customer* lebih memilih untuk berhenti berlangganan atau pindah berlanggan ke perusahaan lain [10] *customer churn* ini tentunya ditentukan oleh beberapa *variable*. Dalam penelitian Iqbal dkk disebutkan bahwa [11] prediksi *cusromer churn* adalah *factor* yang sangat penting dalam persaingan bisnis oleh karena itu penelitian ini dinilai sangat penting.

Customer churn adalah salah satu masalah terbesar dari setiapp perusahaan manapun. Jika kita dapat mengetahui mengapa pelanggan pergi dan kapan mereka pergi dengan akurasi yang tepat itu akan sangat membantu perusahaan untuk menyusun strategi pencegahan mereka.

Customer Churn menyebabkan perusahaan mengalami kerugian. Dengan adanya model yang dapat memprediksi kapan dan mengapa pelanggan berhenti berlanggan maka perusahaan mencegah itu sehigga perusahaan dapat terhidar dari dari kerugian [12].

Perusahaan telekomunikasi dapat menggunakan analisis kami untuk mengukur apakah ia menyediakan produk yang bermanfaat dibandingkan dengan produk yang disediakan oleh pesaingnya. Karena biaya untuk mendapatkan pelanggan baru jauh lebih tinggi daripada mempertahankan pelanggan yang sudah ada, perusahaan dapat menggunakan analisis tingkat *churn* untuk memberikan diskon, penawaran khusus, dan produk unggulan untuk mempertahankan pelanggan saat ini.

2.3. Data Preparation

Untuk dataset kita menggunakan dataset *Telco Customer Churn* dari Kaggle. Berikut adalah sample 5 data teratas dari dateset *TelCo Customer Churn.*

	customerID	gender	SeniorCitizen	Partner	Dependents	tenure	PhoneService	MultipleLines	InternetService	OnlineSecurity	OnlineBackup	DeviceProtection	TechSupport	StreamingTV	StreamingMovies	Contract	PaperlessBilling	PaymentMethod	MonthlyCharges	TotalCharges	Churn
0	7590-VHVEG	Female	0	Yes	No	1	No	No phone service	DSL	No	Yes	No	No	No	No	Month-to-month	Yes	Electronic check	29.85	29.85	No
1	5575-GNVDE	Male	0	No	No	34	Yes	No	DSL	Yes	No	Yes	No	No	No	One year	No	Mailed check	56.95	1889.5	No
2	3668-QPYBK	Male	0	No	No	2	Yes	No	DSL	Yes	Yes	No	No	No	No	Month-to-month	Yes	Mailed check	53.85	108.15	Yes
3	7795-CFOCW	Male	0	No	No	45	No	No phone service	DSL	Yes	No	Yes	Yes	No	No	One year	No	Bank transfer (automatic)	42.30	1840.75	No
4	9237-HQITU	Female	0	No	No	2	Yes	No	Fiber optic	No	No	No	No	No	No	Month-to-month	Yes	Electronic check	70.70	151.65	Yes

Gambar 1. Dataset *Telco Customer Churn*

Data set ini memiliki 7043 baris dan 21 kolom yang terdiri dari

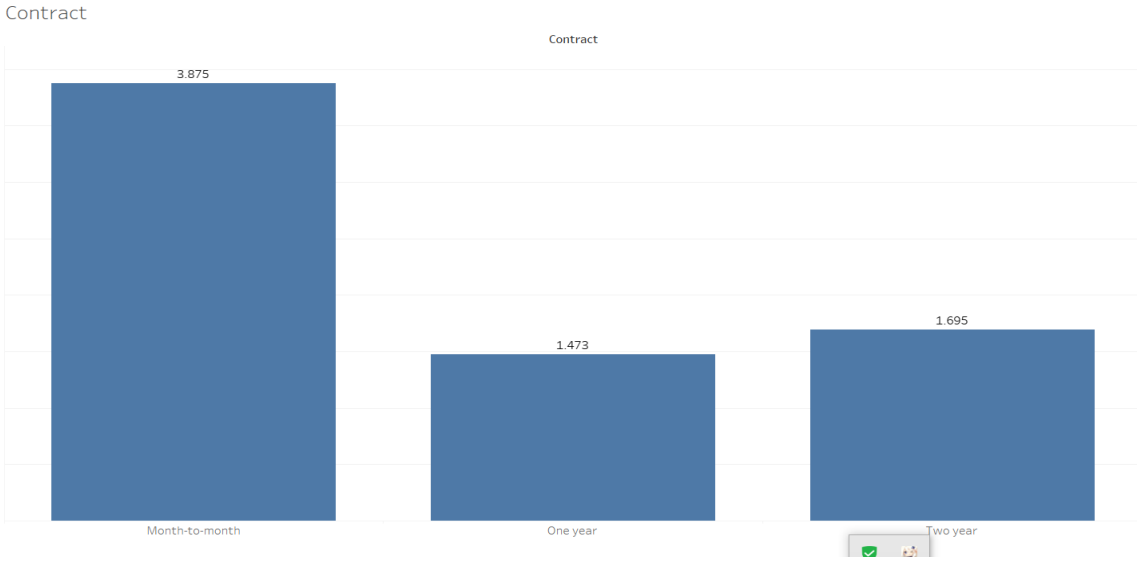
1. Customer ID
2. Gender
3. Senior Citizen
4. Partner
5. Dependents
6. Tenure
7. Phone Service
8. Multiple Lines
9. Internet Service
10. Online Security
11. Online Backup
12. Device Protection
13. Tech Support
14. Streaming TV
15. Streaming Movies
16. Contract
17. Paperless Billing
18. Payment method

19. Monthly Charges
20. Total Charges
21. Churn (Variable Target)

2.4. Data Eksploration

2.4.1. Analisis Variable Contract

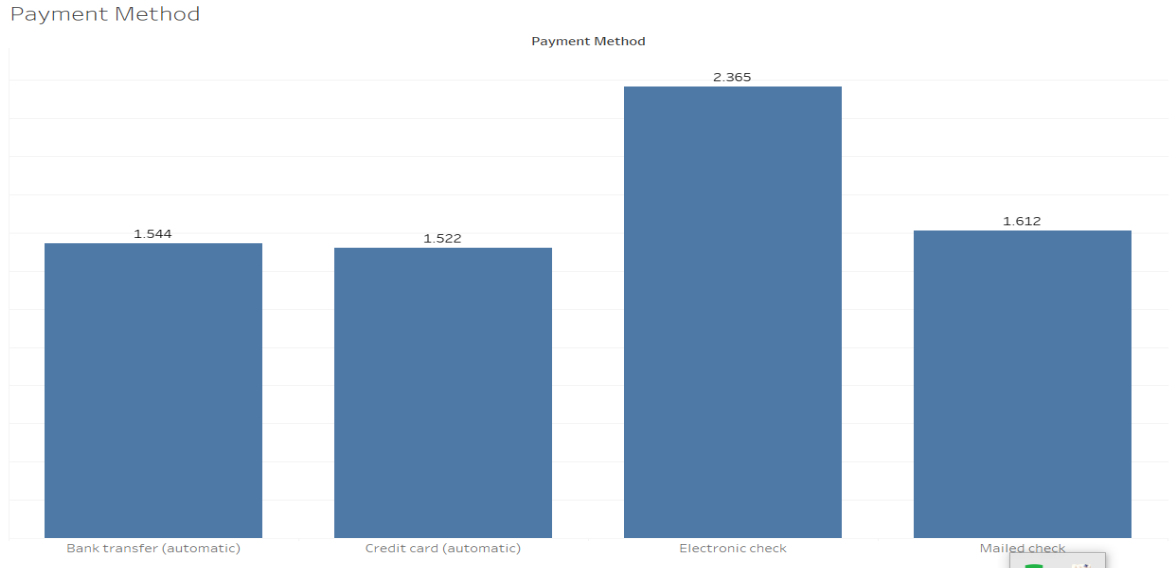
Sebagian besar pelanggan memilih kontrak perbulan dan juga jumlah pelanggan yang memilih kontak 1 tahun dan 2 tahun tidak jauh berbeda. Berikut diagram yang menunjukan jumlah pelanggan dalam memilih kontrak.



Gambar 2. Diagram *Variable Contract*

2.4.2. Analisis Variable Payment Method

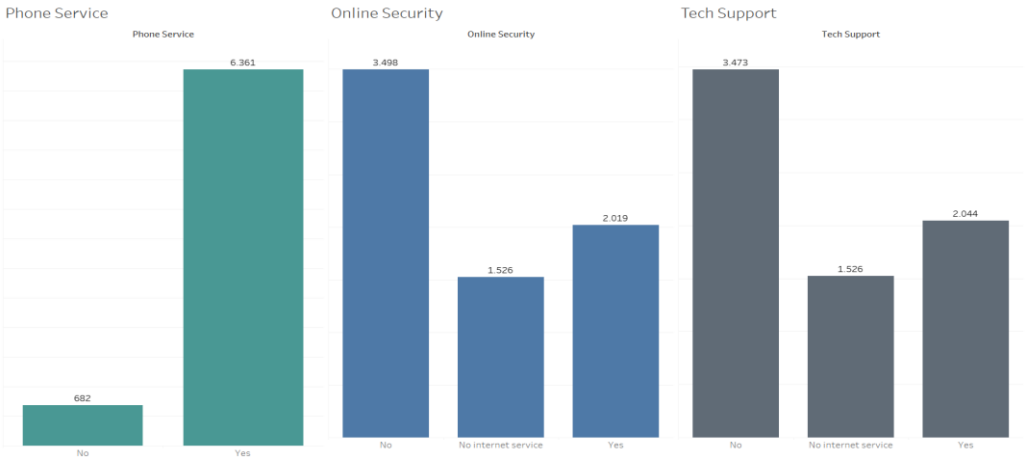
Dari dataset dapat dilihat pelanggan lebih memilih pembayaran elektronik dan diikuti oleh *mailed check, bank tranfer, creditcard*. Berikut diagram yang menggambarkan jumlah pelanggan dalam memilih jenis pembayaran.



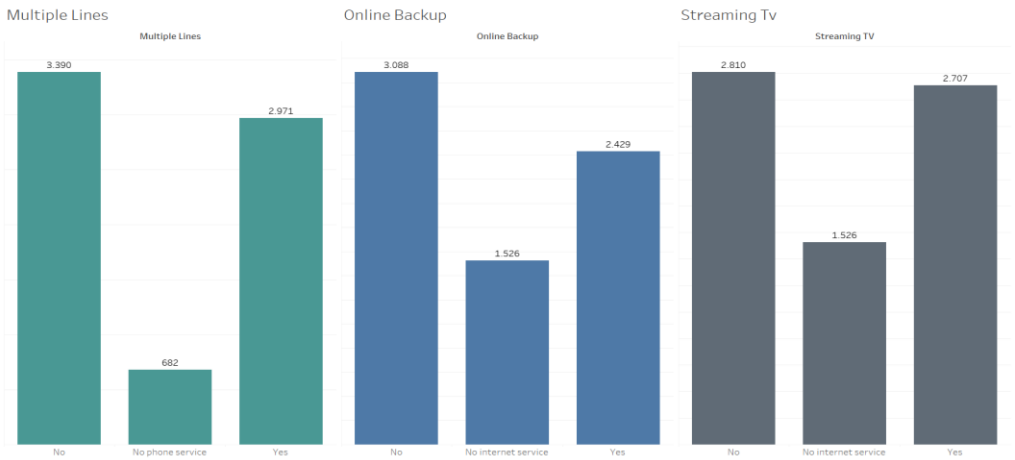
Gambar 3. Diagram *Variable Payment Method*

2.4.3. Analisis Encoding Varible

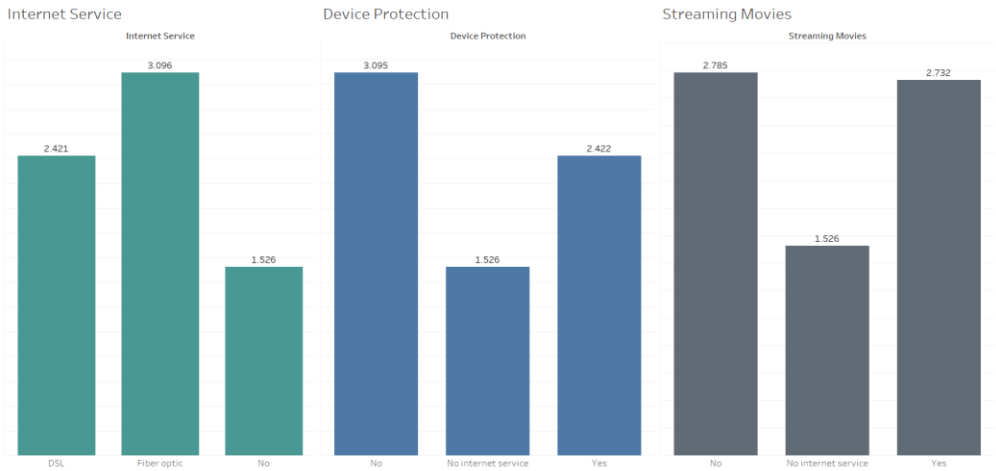
Berikut adalah diagram hasil analisis data dari *encoding variable*:



Gambar 4. *Diagram Variable Encoding* 1



Gambar 5. *Diagram Variable Encoding* 2



Gambar 6. *Diagram Variable Encoding* 3

Dari gambar diatas dapat diartikan

1. Sebagian besar pelanggan memiliki layanan telepon dan hampir setangahnya memiliki banyak saluran.
2. ¾ pelanggan telah memiliki layanan internet berupa *fiber optic* maupun *DSL*, dan hampir dari setengah berlangganan *streaming tv* dan film.
3. Pelanggan yang telah memanfaatkan fitur *online security*, *online backup* dan *device protection* terlihat lebih sedikit dibandingkan dengan pelanggan yang tidak berlanggan fitur-fitur tersebut

Data eksplorasi ini menunjukan fitur-fitur apa saja yang sesuai dengan keinginan pelanggan jika fitur-fitur sesuai dengan keinginan pelanggan maka tinkat kepuasan yang dirasakan akan semakin tinggi [13] dan akan mengurangi tangka *churn*.

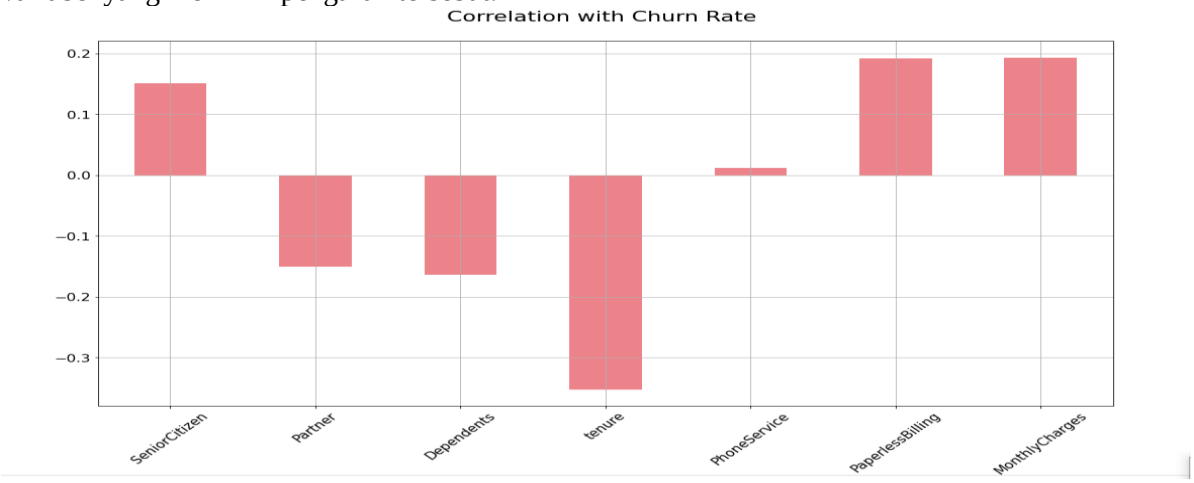
2.5. Variable Target

Variable Churn adalah variabel target dalam penelitian ini dimana variabel ini menentukan dimana pelanggan akan berhenti berlaggan atau tidak. Data yang digunakan dalam penelitian ini adalah data yang bersumber dari Kaggle dan berisi tentang data pelanggan disebuah perusahaan telekomunikasi yang memiliki 7.403 baris data dan 21 kolom (variabel).

Dari dataset tersebut terdapat kolom atau *variable churn* yang kemudian dijadikan sebagai variabel target. Diketahui sebanyak 73,4% dari 7.403 pelanggan yang memutuskan untuk loyal dengan *provider* yang sedang digunakannya, dan sebanyak 26,6% dari 7.403 pelanggan yang memutuskan berhenti berlangganan (*churn*).

2.5. Korelasi

Dari dataset diatas terdapat beberapa variabel yang memiliki korelasi yang cukup kuat dan memiliki variasi yang tinggi terhadap variabel target. Pada gambar dibawah ditunjukkan variabel yang memiliki pengaruh tersebut.



Gambar 8. Korelasi

```
Most Positive Correlations:
 MonthlyCharges      0.193356
PaperlessBilling    0.191825
SeniorCitizen       0.150889
PhoneService        0.011942
dtype: float64

Most Negative Correlations:
 Partner         -0.150448
Dependents      -0.164221
TotalCharges    -0.199426
tenure          -0.352229
dtype: float64
```

Gambar 9. Korelasi *Positif* dan *Negatif*

Ada korelasi *positif* yang cukup besar dari payment method paperless billing dan variable *monthly charges* terhadap variabel tatget (*churn*). Dan ada jugan korelasi *negative* yang tinggi dari *variable tenure* terhadap variabel target

3. HASIL DAN PEMBAHASAN

3.1. Logistic Regression

Logistic Regression ini berguna untuk memprediksi varibel bebas berdasarkan variable target [14]. Dibawah ini adalah model hasil dari klasifikasi menggunakan *Logistic Regression* dengan pendistribusian *data train* dan *data testing* dengan rasio yang sudah ditentukan sebelumnya yaitu 70:30.

	Model	Accuracy	Precision	Recall	F1 Score
0	Logistic Regression	0.818311	0.690323	0.572193	0.625731

Gambar 11. Klasifikasi Logistic Regression

Berdasarkan gambar diatas, didapat hasil klasifikasi terhadap variabel target menggunakan model *Logistic Regression* yakni 80% nilai akurasi, 69% nilai *precission*, 57% nilai *recall* dan 62% nilai *F1 Score*.

3.2. Decission Tree

Dibawah ini adalah model hasil dari klasifikasi menggunakan *Decission Tree* dengan pendistribusian *data train* dan *data testing* dengan rasio yang sudah ditentukan sebelumnya yaitu 70:30.

	Model	Accuracy	Precision	Recall	F1 Score
0	Decision Tree	0.729595	0.490566	0.486631	0.488591

Gambar 12. Klasifikasi Decission Tree

Berdasarkan gambar 12 diatas, didapat hasil klasifikasi terhadap variabel target menggunakan model *Decission Tree* yakni nilai akurasi sebesar 72%, nilai *precission* 49%, nilai *recall* 48% dan nilai *F1 Score* 48%.

4. KESIMPULAN

Dari pembahasan dia atas dapat disimpulkan bahwa dalam memprediksi *customer churn* pada sebuah perusahaan telekomunikasi model klasifikasi *Logistic Regression* lebih berpeluang untuk mencapai tingkat akurasi terbaik yaitu sebesar 80% dibandingkan dengan model klasifikasi *Decission Tree* yang hanya mencapai tingkat akurasi 72%. Diharap kedepannya model ini dapat ditinggkatkan lagi agar dapat membantu banyak perusahaan dalam menagani *customer churn.*

UCAPAN TERIMA KASIH

Penulis mengucapka terimaksih banyak kepada orang tua, dosen pembimbing, dosen penguji serta teman-teman yang telah mendukung penulis untuk menyelasaikan penulisan jurnal ini.

DAFTAR PUSTAKA

[1] A.Wicaksono, A.Anita, dan T.N.Padilah, 2021, Uji Performa Teknik Klasifikasi untuk Memprediksi Customer Churn, *Bianglala Informatika*, vol. 9, no.1.

[2] Koharudin, M.Galih Pradana, dan Kusrini, 2019, Prediksi Customer Churn Perusahaan Telekomunikasi Menggunakan Naïve Bayes dan K-Nearest Neighbor, *Jurnal informasi Interaktif*, vol.4, no.3.

[3] D. H. Tisantri, R. C. Wihandika, dan S. Adinugroho, 2019, Prediksi Keputusan Pelanggan Menggunakan Extreme Learning Machine Pada Telco Customer Churn, *Jurnal Pegembangan Teknologi Informasi dan Ilmu Komputer*, vol.3 No.11.

[4] I. M. Latief, A. Subekti, dan W. Gata, 2021, Prediksi Tingkat Pelanggan Churn Pada Perusahaan, *Informatics Bus. Inst Darmajaya*, vol. 21, no. 1.

[5] V. Kavitha, G.H. Kumar, V.M Kumar, dan M. Harish, 2020, Churn Prediction of Customer in Telecom Industry Using Machine Learnig Algorithms, *int. J. Eng. Res. Technol*, vol. 9 no. 05.

[6] M. F. Mujaddid, Adiwijaya, dan S. Al-faraby. 2017, Analisis Churn Prediction menggunakan metode Logisitic Regression dan SMOTE (Synthetic Minority Over Sampling Technique) Pada Perusahaan Telekomunikasi, *e-proceeding of engineering*, vol.4 no. 3.

[7] I. M. Latief, A. Subekti, dan W. Gata, 2021, Prediksi Tingkat Pelanggan Churn Pada Perusahaan, *Informatics Bus. Inst Darmajaya*, vol. 21, no. 1.

[8] Toni Arifin, Rizal Rachman, 2020, Analisis Decision Tree menggunakan Particle Swarm Optimization Untuk Klasifikasi Sel Pap Smear, Jurnal Teknik Informatika dan Sistem Informasi, vol. 7, no 3.

[9] Toni Arifin, Rizal Rachman, 2020, Analisis Decision Tree menggunakan Particle Swarm Optimization Untuk Klasifikasi Sel Pap Smear, Jurnal Teknik Informatika dan Sistem Informasi, vol. 7, no 3.

[10] A. Alamsyah, N. Salma, A Comparative Study of Employee Churn Prediction Model, in 2018 4th Internation Conference on Science and Technology (ICST).

[11] I. M. Latief, A. Subekti, dan W. Gata, 2021, Prediksi Tingkat Pelanggan Churn Pada Perusahaan, *Informatics Bus. Inst Darmajaya*, vol. 21, no. 1.

[12] D. H. Tisantri, R. C. Wihandika, dan S. Adinugroho, 2019, Prediksi Keputusan Pelanggan Menggunakan Extreme Learning Machine Pada Telco Customer Churn, *Jurnal Pegembangan Teknologi Informasi dan Ilmu Komputer*, vol.3 No.11.

[13] Yuda Yuliana, Rangga Sanjaya, Mayya Nurbayanti Shobary, 2006, Analisis Kepuasan Pegawai Terhadap Layanan Unit Sistem Informasi Menggunakan Technology Acceptance Model Di PT KAI (Persero), *jurnal Informatika*, vol. 3, hal 290-298.

[14] J. Kinoto, K.L. Damanik, E. Tri, S. Situmorang, J, Siregar, dan M. Harahap, 2020, Prediksi Employee Churn Dengan Uplift Modeling Menggunakan Algoritma Logistic Regression, *J. Penelit. Tek. Inform. Univ Prima Indones*, vol. 3 no. 2.