IJIEB: Indonesian Journal of Islamic Economics and Business
Volume 8, Nomor 1, June 2023, 01-20
E_ISSN: 2540-9506 P_ISSN: 2540-9514
http://e-journal.lp2m.uinjambi.ac.id/ojp/index.php/ijoieb

Prediksi *Customer Churn* Menggunakan Metode CRISP-DM Pada Industri Telekomunikasi Sebagai Implementasi Mempertahankan Pelanggan

Yudi Yudiana[1], Asiroch Yulia Agustina[2] dan Nur Khofifah[3]
[1]yudi@unusia.ac.id, Universitas Nahdlatul Ulama Indonesia
[2]asirochyulia@unusia.ac.id, Universitas Nahdlatul Ulama Indonesia
[3]nur.khofifah1613@gmail.com, Universitas Nahdlatul Ulama Indonesia

Abstract

Fulfilling the needs and desires of customers is a strategy in retaining customers. Among these strategies is a quality product and quality service will also fulfill customer expectations. The purpose of this research is to predict whether customers will become loyal customer or leave the company's services. The research method uses the CRISP-DM (Cross Industry Standard Process for Data Mining) technique and the hypermeter tuning method with the ridge classifier algorithm and the confusion matrix. CRISP-DM is a process model that serves as the base for a data science process with six sequential phases. The data used is 7,043 records. divided into 70:30 data train and data test. From the selection of features, most unsubscribed customers are customers who do not use several services such as VPN, Data Backup, Device Protection, Technician Assistance, TV Streaming, Movie Streaming monthly contracts and use the E-Wallets payment method. Then the results of the accuracy research using the confusion matrix show quite good results with an accuracy of 80.5%, Precision of 85.7% and Recall of 89.8%. As many as 73% or 5,174 continue to use the service and 1,869 or around 27% of customers stop using Telkom Indonesia services.

Keywords: CRISP-DM, Loyal Customer, Quality Services, Customer Retention

PENDAHULUAN

Dalam dunia usaha, perusahaan harus berupaya untuk mampu memahami kebutuhan dan keinginan pelanggan dengan tujuan agar terjalin hubungan baik antara kedua belah pihak. Apabila konsumen terpenuhi kebutuhan dan keinginannya, maka konsumen akan merasa puas dan akan membangun loyalitas pelanggan pada perusahaan sehingga terjalin ikatan yang erat dan saling menguntungkan antara kedua belah pihak (Rabiqy, 2019). Pada dasarnya kepuasan pelanggan secara keseluruhan tidak akan tercapai sekalipun dalam kurun waktu sementara, karena kepuasan pada manusia bersifat kompleks dan tidak dapat diukur dan berbeda setiap individu. Upaya perbaikan dan pembaharuan selalu dilakukan dengan berbagai cara dan strategi. Mempertahankan pelanggan sangat berkaitan dengan kepuasan pelanggan.

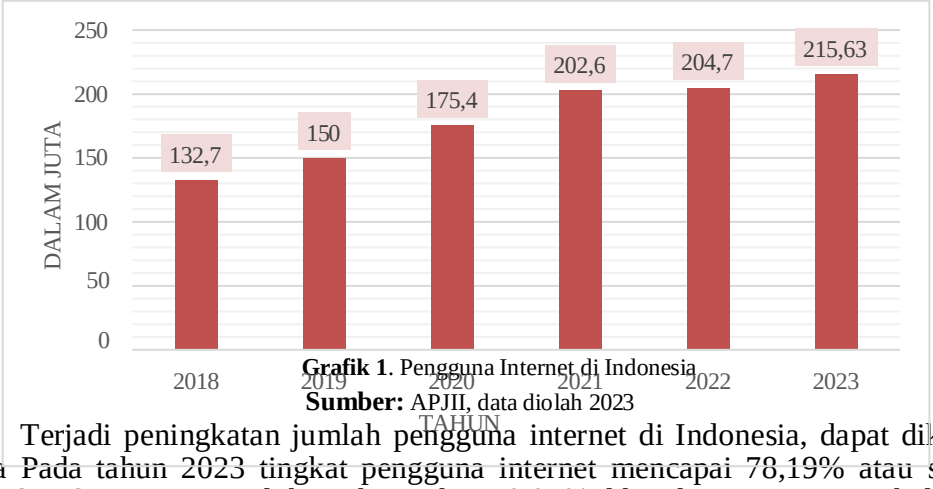
Kepuasan pelanggan yang diperoleh dari barang dan jasa dapat berdampak pada pelanggan yang akan membeli barang atau jasa lebih sering atau berulang, sehingga menjadi pelanggan setia, atau disebut sebagai Customer Retention (T. Mulyana, 2019). Pelanggan merupakan publik eksternal perusahaan yang memiliki peran penting dalam kemajuan bisnis suatu perusahaan. Keberhasilan suatu perusahaan dalam memberikan kepuasan terhadap para pelanggan menjadi tolak ukur keberhasilan perusahaan. Perusahaan menyadari bahwa strategi komunikasi pemasaran satu arah bagi semua orang sudah tidak mencapai target yang diinginkan perusahaan. Ekspektasi konsumen yang tinggi membuat perusahaan harus merencanakan strategi baru. Efektivitas kepuasan pelanggan sangat dibutuhkan untuk memenangkan persaingan bisnis yang semakin kompetitif (Rohana, 2020). Pelanggan ingin diakui sebagai individu yang unik, mendapatkan perhatian dan perlakuan khusus dari perusahaan serta tersedianya varian produk yang dibutuhkan menjadi faktor penting dalam pemenuhan ekspektasi konsumen. Salah satu strategi untuk mengetahui apakah konsumen tersebut akan melakukan pembelian ulang atau tidak, maka perusahaan perlu menganalisis kebiasaan dari konsumen. Kendati demikian, menganalisis kebiasaan konsumen bukan perkara yang mudah serta seringkali membutuhkan waktu yang lama dan biaya yang tidak sedikit. Sehingga, diperlukan suatu alat bantu (tools) sebagai media sebagai usaha dalam mengatasi masalah tersebut.

Banyak perusahaan sering memberikan penawaran produk berupa diskon atau promo lainnya kepada konsumen. Hal ini dimaksudkan agar perusahaan dapat menarik calon pembeli baru. Namun, tentu saja perusahaan tidak dapat secara acak dalam memberikan penawaran kepada pembeli. Untuk menjaga faktor efektifitas dan efisien pemilihan calon konsumen, perusahaan harus memprediksi dan memperhitungkan calon pembeli mana yang berpotensi dapat memberikan feedback keuntungan bagi perusahaan. Salah satu hal yang dapat dikatakan menjadi penentu utama profitabilitas perusahaan adalah loyalitas dari para pembelinya. Loyalitas konsumen memegang peranan yang sangat penting bagi keberlangsungan suatu bisnis. Menurut (Rowley, 2005) kunci dari pengembangan dan profitabilitas bisnis yaitu loyalitas pelanggan. Pelanggan adalah aset perusahaan yang memiliki peran terbesar dalam profitabilitas, image dan citra perusahaan. Perusahaan diharapkan membuat satu divisi untuk mempertahankan pelanggan, ini merupakan hal yang penting, sebab mempertahankan pelanggan jauh lebih baik dibandingkan mencari pelanggan baru. dalam penelitian ini, diharapkan dapat membantu perusahaan dalam menentukan pelanggan yang dapat dipertahankan dan tidak dapat dipertahankan, sehingga perusahaan dapat lebih fokus mempertahankan dan memperbaiki kualitas layanan maupun produk yang ditawarkan.

Dalam perkembangan teknologi dan informasi yang berkembang sangat pesat, kebutuhan manusia semakin beragam(Yaqin, 2022). Salah

satunya adalah kebutuhan dalam mendapatkan informasi secara cepat, kemudahan dalam melakukan aktivitas dalam jaringan seperti *meeting*, melakukan pembelajaran dalam internet atau dapat memudahkan manusia untuk berkomunikasi dibelahan dunia lain.

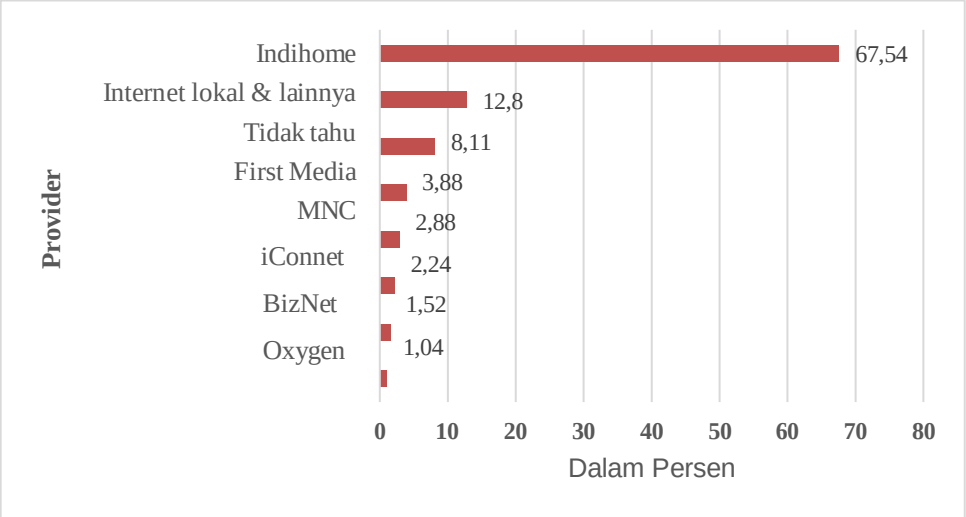
Indonesia merupakan masyakarat yang memiliki kebutuhan informasi internet, khususnya pada saat pandemi covid-19 karena pemerintah menerapkan *social distancing* sehingga masyakarat memiliki keterbatasan aktivitas diluar rumah. Menurut Asosiasi Penyelenggara Jasa Internet Indonesia (APJII), dari tahun ke tahun jumlah pengguna internet semakin meningkat. Berikut adalah grafik pengguna internet di Indonesia:



Grafik 1. Pengguna Internet di Indonesia
Sumber: APJII, data diolah 2023

Terjadi peningkatan jumlah pengguna internet di Indonesia, dapat diketahui bahwa Pada tahun 2023 tingkat pengguna internet mencapai 78,19% atau sebesar 215.626.156 juta, ini terjadi kenaikan sekitar 0,05% dibanding pengguna sebelumnya ditahun 2022 yaitu 204,7 juta. Angka ini mulai mendekati angka jumlah penduduk Indonesia yaitu 275.773.901 jiwa (Kusnandar, 2022) Pada tahun 2018 – 2023 terjadi kenaikan jumlah pengguna internet sebesar 0,46% dalam 5 tahun terakhir.

Dalam penelitian ini lokasi penelitian dilakukan di PT Telkom Indonesia, karena PT Telkom Indonesia merupakan perusahaan jasa jaringan telekomunikasi, jasa layanan informasi dan komunikasi yang telah berdiri sejak 1965 di Indonesia. Sebagai perusahaan telekomunikasi telah berdiri lama di Indonesia, PT Telkom Indonesia selalu memberikan pelayanan yang terbaik untuk pelanggan. Indihome merupakan layanan milik PT Telkom Indonesia yang berdiri pada tahun 2015. Indihome mencakup akses internet, tayangan televisi berbayar dan saluran telpon. Berikut adalah jumlah pengguna provider di Indonesia pada Februari tahun 2022.



Grafik 2 : Jumlah Pengguna Provider di Indonesia

Sumber: Data diolah, 2023

Survei yang dilakukan oleh APJII melibatkan 7.568 responden seluruh Indonesia. Survei ini dilakukan pada tanggal 11 Januari – 24 Februari 2022. Responden berusia 13 – 55 tahun. Mayoritas kelompok pengguna internet berupa 19 – 54 tahun. Pengguna provider terbanyak yaitu pada produk telkom, indihome. Produk indihome ini digunakan oleh 67,54% responden. Menurut survei Speedtest, indihome merupakan produk telkom yang memiliki kecepatan internet terendah kedua setelah MNC. Kemudian sebanyak 3,88% menggunakan First Media, 2,88% menggunakan MNC, iConnet digunakan sebanyak 2,24%, sebanyak 1,52% menggunakan Biznet. Menurut survei Speedtest Biznet memiliki kecepatan internet tercepat dibandingkan provider lain, dan Oxygen memiliki jumlah pengguna terendah sebanyak 1,04%.

Tujuan Penelitian

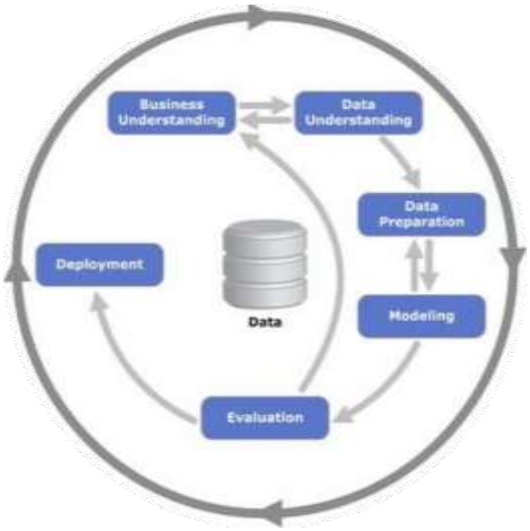
Penelitian ini dibuat untuk melakukan prediksi retention customer dengan menggunakan metode Ridge Classifier. Data yang digunakan adalah data Telkom Indonesia Customer Churn. Dataset ini berisi data behaviour dari Telkom Indonesia customer churn. Perusahaan Telekomunikasi yaitu Telkom Indonesia merupakan salah satu perusahaan yang bergerak dibidang jasa layanan informasi dan komunikasi serta jaringan telekomunikasi. Dataset akan dilakukan pemetaan dimana setiap baris data merepresentasikan data pelanggan dan setiap kolom merepresentasikan informasi mengenai kebiasaan dari masing-masing pelanggan. Penelitian ini diharapkan dapat membantu perusahaan dalam menentukan pelanggan yang dapat dipertahankan dan yang tidak dapat dipertahankan.

TINJAUAN PUSTAKA

Landasan Teori

Metode CRIPS-DM atau *Cross Industry Standard Process for Data Mining* biasanya digunakan dalam bisnis. Pada penelitian (Hasanah, 2021) sejak tahun 1996 standar CRISP-DM sudah dikembangkan guna menjadikan proses industri bisnis dalam sebuah penelitian. Pada awal tahun 1996, 5 perusahaan ternama yaitu Perusahaan asuransi bernama OHRA, Integral Solutions Ltd (ISL) perusahaan penyedia solusi data mining, perusahaan penyedia database bernama NCR Corporation, Terdata dan Daimler AG perusahaan penyedia produsen mobil (Pambudi et al., 2023). 5 Perusahaan membuat Framework kemudian dikembangkan oleh ratusan perusahaan dan organisasi di Eropa, , insinyur berpengalaman mengembangkan metodologi CRISP DM versi pertama yang dipresentasikan pada CRISP-DM SIG Workshop pada bulan Maret 1999 di Brussels (Martinez-Plumed et al., 2021).

Pada tahun 1999 ini diperkenalkan *Cross-Industry Standard Process for Data Mining* atau biasa disingkat CRISP-DM. Pada tahun 2020 di Indonesia, CRISP-DM menjadi acuan Standar Kompetensi Kerja Nasional berdasarkan Keputusan Menteri Ketenagakerjaan No. 299 tahun 2020 (Fauziyah, 2020). William Vorheis, salah satu pencetus CRISP-DM (dari *Data Science Central*). Terdapat 6 proses tahapan dalam metode CRISP DM yaitu, Business Understanding, Data Understanding, Data Preparation, Modelling, Evaluation, dan Deployment.



Gambar 1. Model CRISP - DM

1. Business Understanding

Bussines understanding yaitu tahap pemahaman tujuan, kebutuhan, batasan dan sudut pandang bisnis, kemudian memasukkan pemahaman tersebut kedalam definisi masalah, menentukan strategi yang akan di capai pada data mining(Wijaya & Girsang, 2016).

2. Data Understanding

Data understanding yaitu tahapan mengumpulkan data, pemahaman data, mendeskripsikan data, mengidentifikasi data, dan mengevaluasi kualitas data yang akan di gunakan untuk proses selanjutnya (Aditra Pradyana et al., 2020:5).

3. Data Preparation

Proses data preparation yaitu membangun data mentah menjadi dataset akhir. Ada beberapa hal yang dilakukan seperti pembersihan data, pemilihan data, memeriksa atribut-atribut, memilih variabel yang sesuai untuk dianalisis, memeriksa distribusi data serta melakukan transformasi terhadap data(Joko Suntoro, 2019, p. 103).

4. Modelling

Pada tahap modelling, yaitu mengaplikasikan teknik pemodelan yang sesusai, mengidentifikasi dan menampilkan pola. Pada tahap modelling yang telah menggunakan *machine learning*, yaitu membagi data dan menentukan algoritma yang cocok untuk data serta mengkalibrasi aturan model agar mendapatkan hasil yang optimal (Aditra Pradyana et al., 2020). Penelitian ini menggunakan algoritma ridge classifier yang paling baik skornya. Parameter yang akan dicari untuk Ridge Classifier yaitu sebagai berikut:

1) 'solver' = Solver digunakan untuk penyelesaian masalah. Opsi parameternya auto, svd, cholesky, lsqr , sparse_cg, ibfgs, saga dan sag.
2) 'alpha' = Parameter kekuatan regularisasi, untuk meningkatkan pengkondisian masalah dan mengurangi varians dari perkiraan. Semakin tinggi nilai alpha maka samakin kuat regulasi. Opsi nilai parameter yang dipilih yaitu 1e-5 1e-5, 1e-4, 1e-3, 1e-2, 1e-1, 1, 10, 100.
3) 'fit_intercept' = Penggunaan intersep untuk model. Opsi parameternya yaitu True, False.

5. Evaluation

Tahap proses ini yaitu mengevaluasi satu atau lebih dari satu model yang akan digunakan dalam penelitian, memastikan tidak ada permasalahan penting yang tidak tertangani dengan baik, memastikan dan meng melakukan perhitungan model sesuai dengan tahap awal, melihat pola yang di hasilkan dari algoritma, melihat parameter yang akan digunakan untuk evaluasi komparasi algoritma(Aditra Pradyana et al., 2020) evaluasi algoritma yang digunakan adalah Confusion Matrix dengan melihat nilai akurasi, presisi dan recall. Dalam dunia statistik, akurasi didefinisikan sebagai tingkat ketepatan antara nilai prediksi dengan nilai fakta dan aktual dari semua pengamatan. Presisi merupakan tingkat ketepatan antara informasi yang diprediksi pengguna dengan jawaban yang diberikan oleh sistem, sedangkan recall adalah tingkat keberhasilan sistem mengukur kasus dalam menemukan kembali identifikasi informasi(Mulaab, 2017). Nilai dapat dilihat dalam perhitungan sebagai berikut:

$$Accuracy = \frac{\text{Prediksi benar}}{\text{Semua kasus}} \times 100\% = \frac{TP + TN}{TP + FP + TN + FN} \times 100\%$$

$$Precision = \frac{\text{True Positive}}{\text{Total Prediksi Positif}} \times 100\% = \frac{TP}{TP + FP} \times 100\%$$

$$Recall = \frac{\text{True Positif}}{\text{Total aktual Positif}} \times 100\% = \frac{TP}{TP + FN} \times 100\%$$

6. *Deployment*

Tahap *deployment* yakni membuat laporan hasil dari seluruh data yang diolah, mengembangkan dan memvisualisasikan data.

Penelitian Terdahulu

Penelitian yang telah dilakukan oleh Bimo Adiprawa Dwijaya dan Setia Wirawan (Dwijaya & Wirawan, 2021) dengan menggunakan metode CRIPS- DM dengan memanfaatkan machine learning seperti Random Forest, XGBoost, dan K-Nearest Neighbor (KNN). Data yang di gunakan yaitu 10.000 data pelanggan yang meninggalkan layanan bank. Di bagi menjadi 80:20 data train dan data test. Hasil menunjukkan dari seleksi fitur terdapat 8 variabel yang mempengaruhi customer churn diantaranya umur, estimasi gaji, skor kredit, keadilan, produk, tenure, aktivasi pelanggan dan geografi. Kemudian hasil akurasi menggunakan confusion matrix yaitu Random Forest model 87,05%, Xgboost 90,75% dan KNN 100%. Hasil KNN mengungguli semua model dengan akurasi 100% dan score AUC yaitu 1. Penelitian ini dapat dikembangkan dalam dunia perbankan untuk mengurangi pelanggan yang berhenti menggunakan layanan.

Penelitian lain Msy Aulia Hasanah, Sopian Soim dan Ade Silvia Handayani. Penelitian ini bertujuan mengklasifikasi hujan lebat dan hujan sangat lebat menggunakan tehnik data mining dengan metodologi CRIPS- DM, algoritma yang di gunakan dalam klasifikasi ini adalah CART (Classification and regression Tree). Menggunakan dataset sebanyak 3.653 curah hujan yang diambil dari data BMKG. terdapat 123 record yang di kotomi curah hujan dengan intensitas hujan lebat dan hujan sangat lebat. Berdasarkan uji parameter confusion matrix dihasilkan akurasi cukup baik sebesar 89,4% dengan jumlah data testing 123 data jumlah prediksi data 110 benar(Hasanah et al., 2021).

Penelitian yang dilakukan Hardian Kokoh Pambudi, Putu Giri Artha Kusuma, Femi Yulianti, dan Kevin Ahessa menggunakan metodologi (CRISP- DM). data yang digunakan dalam penelitian yaitu data ekspedisi pesawat udara yang dirilis oleh the International Air Transport Association (IATA) yakni perusahaan logistik kargo. Penelitian menggunakan 3 metode yaitu regresi logistik, random forest dan Artificial Neural Network (ANN). Hasil penelitian menjelaskan bahwa metode regresi logistik memiliki nilai kurasi sebesar 72,84%, random forest 76,6%, dan metode ANN sebesar 73,81%. Artinya, metode yang memiliki nilai akurasi lebih baik yaitu metode random forest dibandingkan dengan metode lainnya (Pambudi et al., 2020).

Jairo Acosta Solano, Diana Janeth Lancheros Cuesta, Samir F. Umaña Ibáñez, Jairo R. Coronado-Hernández (Solano et al., 2021) melakukan penelitian evaluasi pembelajaran machine learning dengan metode CRISP- DM. Tujuan penelitian tersebut yaitu untuk memprediksi tingkat kinerja sekolah SMP di Karibia Kolombia dengan menggunakan Saber 11 test. Dengan mengusulkan metodologi baru untuk mengevaluasi hasil ujian berdasarkan wilayah tersebut dengan mempertimbangkan sosial ekonomi masing-masing siswa tersebut. Data yang digunakan yaitu sistem database ICFES (Colombian Institute for the Evaluation of Education) tahun 2017 sampai dengan 2019. Penelitian ini mengevaluasi 4 model yaitu (C4.5, LMT, PART dan MLP. Hasil dari penelitian tersebut adalah model pembelajaran (LMT) diidentifikasi sebagai salah satu yang menawarkan metrik terbaik di antara semua model, model ini dilatih dan dievaluasi dengan parameter konservatif yang datang secara default di WEKA untuk membuat prediksi kinerja siswa di kawasan Karibia, penerapan model ini akan memungkinkan para pelaku sistem pendidikan di wilayah Karibia untuk melakukan tindakan intervensi dengan para siswa yang memiliki prediksi kinerja minimal atau tidak memadai. Struktur utama pohon LMT menentukan bahwa atribut dengan perolehan hasil tertinggi adalah sifat sekolah yang dievaluasi, memiliki koneksi Internet, dan hari sekolah ini. Demikian juga, untuk menunjukkan bahwa aspek lain mungkin berdampak pada kinerja, serangkaian aturan asosiasi dikembangkan dengan algoritma apriori, dengan tingkat kepercayaan minimal 60%, 20 aturan ini memiliki konsekuensi kelas kinerja global di 11° Saber, aspek-aspek tersebut meliputi penggunaan Internet, ketersediaan perangkat komputer, luas lokasi dan sifat sekolah, dan usia, sangat menentukan kinerja siswa.

Penelitian yang lakukan oleh Arwa A. Jamjoom (2021) menggunakan 3 tehnik yaitu K-means, regresi logistik dan jaringan saraf. Metode Descision Tree digunakan pada fase model CRISP-DM untuk mengidentifikasi atribut pelanggan yang keluar dan tidak dapat dipertahankan. Data yang digunakan yaitu data perusahaan asuransi pada tahun 2018, tujuan penelitian ini yaitu untuk menyelidiki penggunaan ekstraksi pengetahuan dalam memprediksi customer churn. Hasil penelitian menunjukkan bahwa prosedur penambangan data bisa sangat berhasil dalam mengekstraksi informasi

tersembunyi dan mengetahui informasi pelanggan. Distribusi data training 50:50 menghasilkan hasil yang efektif ketika teknik regresi logistik digunakan selama penelitian ini. Distribusi 70:30 bekerja secara efektif untuk teknik jaringan saraf. Dalam hal ini, disimpulkan bahwa setiap teknik bekerja secara efektif dengan distribusi training set yang berbeda. Penelitian ini mengusulkan metode penambangan data untuk membantu membangun kerangka kerja yang berisi kumpulan dan klasifikasi pelanggan. Algoritma K- means digunakan untuk segmentasi pelanggan, prediksi churn dilakukan dengan menggunakan customer attractiveness dan customer churn. Sistem yang disarankan dapat memungkinkan perusahaan asuransi kesehatan memprediksi kemungkinan menyelidiki perilaku masa lalu, klien saat ini, dan masa depan. Dengan pemahaman yang lebih baik tentang fitur churn, manajer asuransi kesehatan dapat mencegah terjadinya customer churn dengan mempertimbangkan beberapa strategi. Strategi tersebut antara lain menyediakan fasilitas yang dibutuhkan, meningkatkan kualitas layanan dan produk, mengidentifikasi perbedaan kebutuhan konsumen, dan meningkatkan daya tanggap pelanggan (Jamjoom, 2021).

METODE PENELITIAN

Objek penelitian yang menjadi fokus peneliti yaitu *Data Science*. Data Science merupakan cabang dari domain AI (*Artificial Intelligence*) yang digunakan untuk mengolah data menjadi informasi agar dapat dipahami dalam merancang strategi bisnis seperti halnya memprediksi pelanggan mana yang akan berhenti menggunakan layanan atau produk sebagai implementasi customer retention. Penelitian ini dibuat untuk melakukan prediksi retention customer dengan menggunakan *metode Ridge Classifier.*

Data

Data yang digunakan adalah data Telkom Indonesia Customer Churn.
Dataset ini berisi data behaviour dari Telkom Indonesia customer churn. Perusahaan Telekomunikasi yaitu Telkom Indonesia merupakan salah satu perusahaan yang bergerak dibidang jasa layanan informasi dan komunikasi serta jaringan telekomunikasi. Dataset akan dilakukan pemetaan dimana setiap baris data merepresentasikan data pelanggan dan setiap kolom merepresentasikan informasi mengenai kebiasaan dari masing-masing pelanggan. Penelitian ini diharapkan dapat membantu perusahaan dalam menentukan pelanggan yang dapat dipertahankan dan yang tidak dapat dipertahankan.

Metode Analisis Data

Metode pengelolaan data dan proses dalam penelitian ini menggunakan metode CRIPS-DM. CRISP-DM adalah metode yang menyediakan proses standar dalam data mining untuk memecahkan masalah dalam bisnis. CRISP-DM lebih mudah diterapkan karena setiap tahapan atau

fase didefinisikan dan terstruktur dengan jelas serta memiliki metodologi data mining yang lengkap dan terdokumentasi dengan baik (Khumaidi, 2020).

HASIL DAN PEMBAHASAN

1. Bussiness Understanding

Data penelitian yang di kumpulkan yaitu berasal dari penelitian ini dibuat untuk melakukan prediksi retention customer. Data yang digunakan adalah data set Telkom Indonesia *customer churn*, Perusahaan tersebut merupakan salah satu Perusahaan Telekomunikasi yaitu Telkom Indonesia yang bergerak dibidang jasa layanan informasi dan komunikasi serta jaringan telekomunikasi. Tujuan penelitian ini adalah membantu perusahaan dalam menentukan pelanggan yang dapat dipertahankan dan yang tidak dapat dipertahankan., sehingga perusahaan dapat mempertahankan, mengembangkan dan memperbaiki kualitas layanan maupun produk yang ditawarkan.

2. Data Understanding

Data yang digunakan adalah data Telkom Indonesia customer churn. Data yang didapatkan berupa data mentah sebanyak 7.043 record, dataset yang digunakan dalam penelitian ini berisi tentang informasi pelanggan dan layanan yang digunakan oleh pelanggan. dimana setiap baris data merepresentasikan data pelanggan dan setiap kolomnya terdapat informasi mengenai behaviour dari masing-masing pelanggan. Beberapa variabel yang terdapat dalam dataset tersebut yaitu:

Tabel 1. Dataset

No.	Nama Kolom	Deskripsi	Tipe Data	Isi Data
1.	ID Pelanggan	ID unik pelanggan	Kategorikal	-
2.	Jenis Kelamin	Jenis kelamin pelanggan	Kategorikal	Laki-laki, Perempuan
3.	Bekerja	Pelanggan bekerja atau tidak	Kategorikal	Ya, Tidak
4.	Status Pernikahan	Apakah pelanggan sudah menikah atau tidak	Kategorikal	Ya, Tidak
5.	Tanggungan	Apakah pelanggan saat ini memiliki tanggungan atau tidak	Kategorikal	Ya, Tidak
6.	Lama Berlangganan	Berapa lama pelanggan berlangganan	Numerik	(Jumlah bulan pelanggan telah berlangganan)
7.	Layanan Telepon	Apakah pelanggan berlangganan jasa layanan telepon atau tidak	Kategorikal	Ya, Tidak
8.	Multi Jaringan	Apakah pelanggan menggunakan layanan multi jaringan atau tidak	Kategorikal	Tidak berlangganan telepon, Ya, Tidak

9.	Layanan Internet	Layanan jenis Internet yang digunakan pelanggan	Kategorikal	FUP, Unlimited, Tidak
10.	VPN	Apakah pelanggan menggunakan layanan VPN atau tidak	Kategorikal	Tidak berlangganan internet, Tidak, Ya
11.	Backup Data	Apakah pelanggan menggunakan layanan backup data atau tidak	Kategorikal	Tidak berlangganan internet, Tidak, Ya
12.	Perlindungan Device	Apakah pelanggan menggunakan layanan perlindungan device	Kategorikal	Tidak berlangganan internet, Tidak, Ya
13.	Layanan Teknisi	Apakah pelanggan menggunakan layanan jasa teknisi	Kategorikal	Tidak berlangganan internet, Tidak, Ya
14.	Streaming TV	Apakah pelanggan menggunakan layanan streaming TV atau tidak	Kategorikal	Tidak berlangganan internet, Tidak, Ya
15.	Streaming Film	Apakah pelanggan menggunakan layanan streaming film atau tidak	Kategorikal	Tidak berlangganan internet, Tidak, Ya
16.	Kontrak	Jangka waktu kontrak pelanggan	Numerik	Perbulan, 1 Tahun, 2 Tahun
17.	Promo	Apakah pelanggan menggunakan promo atau tidak	Kategorikal	Ya, Tidak
18.	Metode Pembayaran	Metode pembayaran yang digunakan oleh pelanggan	Kategorikal	Transfer Bank, Kartu kredit, E-Wallets, Lainnya
19.	Tagihan Bulanan	Berapa tagihan bulanan pelanggan tersebut	Numerik	(Jumlah tagihan perbulan)
20.	Total tagihan	Berapa total tagihan pelanggan tersebut	Numerik	(Lama berlangganan x Tagihan bulanan)
21.	Berhenti (*Churn*)	Status pelanggan, apakah pelanggan berhenti menggunakan layanan atau tidak	Kategorikal	Ya, Tidak

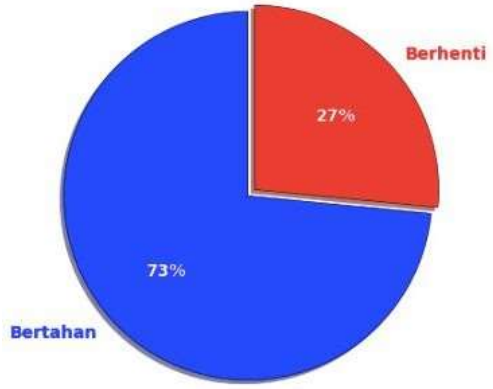
3. Tahap *Preparation*

Dalam tahap ini penulis memeriksa distribusi data menggunakan bantuan kombinasi matplotlib, plotly, dan seaborn untuk visualisasi.

Seaborn merupakan kumpulan data gambar, animasi, diagram, grafik dan lain-lain yang dapat memudahkan untuk merepresentasikan informasi dalam cakupan Python. Seaborn bersifat terbuka dari sumber yang berlisensi *Berkeley Software Distribution* (BSD). Sumber lisensi BSD merupakan perangkat lunak yang memungkinkan akses terbuka dan dapat digunakan, dimodifikasi secara gratis. Seaborn dibangun di atas pustaka matplotlib. Seaborn akan menganalis data untuk membuat visualisasi tanpa kostumisasi yang rumit seperti pada matplotlib. Seaborn lebih sering digunakan data science dalam tahap Explore dan Presentations. Matplotlib adalah pustaka visualisasi yang fleksibel. Perbedaan antara seaborn dan matplotlib adalah, seaborn menyediakan berbagai pola visualisasi tingkat lanjut dengan sintaks yang tidak terlalu rumit sedangkan matplotlib digunakan untuk plotting dasar, batang, lingkaran, garis, plot pencar, dan jenis data lainnya. Hasil pemeriksaan distribusi data adalah sebagai berikut:

Dapat dilihat dari data diatas bahwa sebanyak 1.869 atau sekitar 27% pelanggan berhenti menggunakan layanan Telkom Indonesia sedangkan sebanyak 5.174 atau sekitar 73% pelanggan tetap menggunakan layanan dari perusahaan.

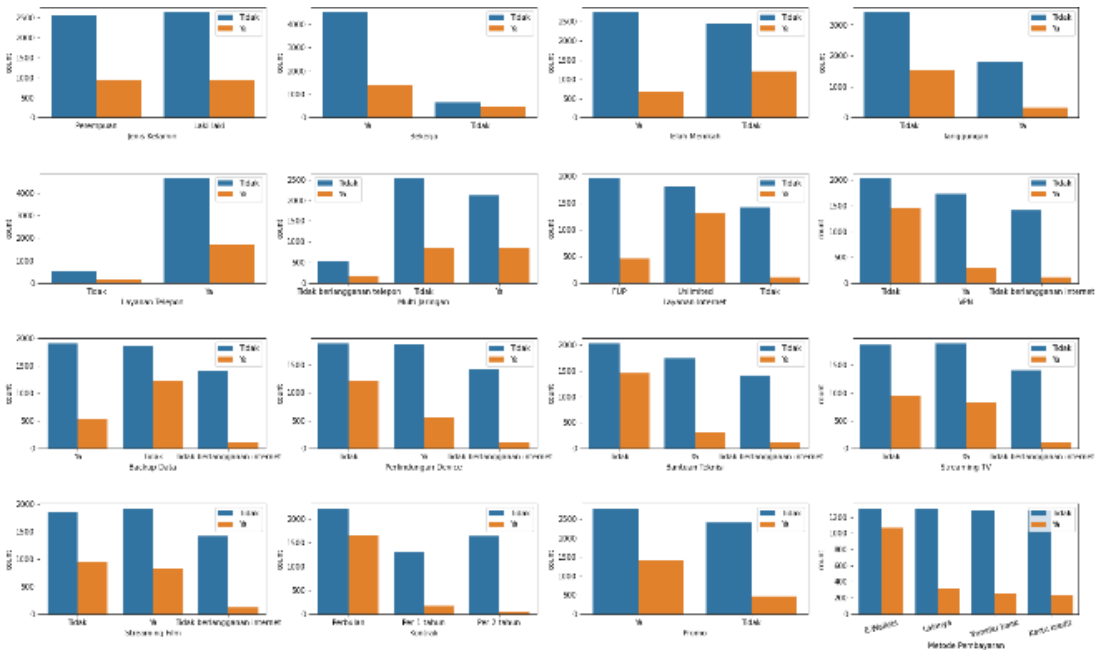
Gambar 2. Distribusi variabel target Sumber:



Data penelitian diolah

Dapat dilihat dari data diatas bahwa sebanyak 1.869 atau sekitar 27% pelanggan berhenti menggunakan layanan Telkom Indonesia sedangkan sebanyak 5.174 atau sekitar 73% pelanggan tetap menggunakan layanan dari perusahaan

.



Gambar 3. Distribusi variabel kategorikal terhadap variabel target
Sumber: Data diolah

Dari gambar grafik diatas dapat dilihat bahwa sebagian besar pelanggan yang berhenti berlangganan adalah pelanggan yang tidak menggunakan beberapa layanan seperti VPN, Backup Data, Perlindungan Device, Bantuan teknisi, Streaming TV, Streaming Film dan sebagian besar kontrak bulanan/pengguna bulanan serta menggunakan metode pembayaran E-Wallets.

Tabel 2. Tipe Data

No	Atribut	Non-Null Count	Dtype
0	jenis kelamin	7043 non-null	object
1	bekerja	7043 non-null	object
2	telah menikah	7043 non-null	object
3	tanggungan	7043 non-null	object
4	lama berlangganan	7043 non-null	integral
5	layanan telepon	7043 non-null	object
6	multi jaringan	7043 non-null	object
7	layanan internet	7043 non-null	object
8	vpn	7043 non-null	object
9	backup data	7043 non-null	object
10	perlingdungan device	7043 non-null	object
11	bantuan teknisi	7043 non-null	object
12	streaming tv	7043 non-null	object

13	streaming film	7043 non-null	object
14	kontrak	7043 non-null	object
15	promo	7043 non-null	object
16	metode pembayaran	7043 non-null	object
17	tagihan bulanan	7043 non-null	integral
18	total tagihan	7043 non-null	integral
19	berhenti	7043 non-null	object

Berdasarkan tabel 2 diatas, dapat diketahui bahwa Lama Berlangganan, Tagihan Bulanan, dan Total Tagihan merupakan data numerik dan sisanya adalah data ketagorikal. ID Pelanggan perlu dikeluarkan dari data karena bersifat unik untuk setiap pelanggan dan tidak diperlukan untuk pemodelan. Total tagihan juga perlu dikeluarkan karena bersifat redundant yang merupakan perkalian dari Tagihan bulanan x Lama berlangganan. Selanjutnya terdapat 11 missing values yang akan diisi dengan nilai median.

Tabel 3. Data sebelum Binary

1	Kelamin	Perempuan	Laki-laki	Laki-laki	Laki-laki	Perempuan
2	Bekerja	Ya	Ya	Ya	Ya	Ya
3	Telah Menikah	Ya	Tidak	Tidak	Tidak	Tidak
4	Tanggungan	Tidak	Tidak	Tidak	Tidak	Tidak
5	Lama Berlangganan	1	34	2	45	2
6	Layanan Telepon	Tidak	Ya	Ya	Tidak	Ya
7	Multi Jaringan	Tidak berlangganan telepon	Tidak	Tidak	Tidak berlangganan telepon	Tidak
8	Layanan Internet	FUP	FUP	FUP	FUP	Unlimited
9	VPN	Tidak	Ya	Ya	Ya	Tidak
10	Backup Data	Ya	Tidak	Ya	Tidak	Tidak
11	Perlindungan Device	Tidak	Ya	Tidak	Ya	Tidak
12	Bantuan Teknisi	Tidak	Tidak	Tidak	Ya	Tidak
13	Streaming TV	Tidak	Tidak	Tidak	Tidak	Tidak
14	Streaming Film	Tidak	Tidak	Tidak	Tidak	Tidak
15	Kontrak	Perbulan	Per 1 tahun	Perbulan	Per 1 tahun	Perbulan
16	Promo	Ya	Tidak	Ya	Tidak	Ya

17	Metode Pembayaran	E-Wallets	Lainnya	Lainnya	Transfer bank	E-Wallets
18	Berhenti	Tidak	Tidak	Ya	Tidak	Ya

Tabel 3 merupakan kolom yang berisi data sebelum di ubah menjadi binary. Terlebih dahulu penulis melakukan perubahan penulisan baik itu pada nama kolom maupun isi datanya seperti menghilangkan spasi dan simbol dengan fungsi replace serta mengecilkan huruf dengan fungsi lower. Untuk atribut yang memiliki kategori lebih dari 2 kami menggunakan metode pd.get_dummies dengan tujuan membuatkan kolom baru berdasarkan kategorinya masing-masing sehingga total fitur menjadi 29. Selanjutnya, melakukan proses normalisasi data untuk data numerik dikarenakan distribusi data yang tidak normal.

Tabel 4. Data setelah *Binary*

1	**Jenis Kelamin**	**0**	**1**	**1**	**1**	**0**
2	Bekerja	1	1	1	1	1
3	Telah Menikah	1	0	0	0	0
4	Tanggungan	0	0	0	0	0
5	Lama Berlangganan	1	34	2	45	2
6	Layanan Telepon	0	1	1	0	1
7	Multi Jaringan	1	0	0	1	0
8	Layanan Internet	0	0	0	0	1
9	VPN	0	1	1	1	0
10	Backup Data	1	0	1	0	0
11	Perlindungan Device	0	1	0	1	0
12	Bantuan Teknisi	0	0	0	1	0
13	Streaming TV	0	0	0	0	0
14	Streaming Film	0	0	0	0	0
15	Kontrak	0	1	0	1	0
16	Promo	1	0	1	0	1
17	Metode Pembayaran	0	1	1	1	0
18	Berhenti	0	0	1	0	1

Tabel 4 merupakan Kemudian mengubah setiap data ketegorikal ke dalam bentuk array berisi 1 dan 0 dengan mendefinisikan fungsi binary map untuk setiap atribut yang memiliki kategori ya = 1, dan tidak = 0. Dan juga untuk atribut jenis kelamin dengan kategori laki-laki = 1, dan perempuan = 0.

1. Modelling

Pada tahap *modelling*, hal pertama yang dilakukan adalah membagi data menjadi 2, data training dan data testing. Data training adalah data yang membentuk dan melatih suatu model, sedangkan data testing digunakan

untuk mengukur evaluasi algoritma (Joko Suntoro, 2019) masing-masing data terbagi menjadi 70:30. Data training 70% dan data testing 30%. Selanjutnya mendefinisikan evaluasi keseluruhan model algoritma yang akan dites dan akan menampilkan tabel nilai skor *accuracy*, *precision*, *recall* dan *f1 score* untuk masing-masing model algoritma.

Setelah itu, variabel membuat daftar model algoritma yang akan digunakan dengan menggunakan parameter defaultnya masing-masing, yaitu: *Logistic Regression*, *Ridge Classifier*, *KNN*, *Gaussian*, *SVC*, *MLP Classifier*, *Decision Tree*, *Random Forest*, *Gradient Boosting*, *AdaBoost*, *CatBoost*, *XGBoost*, *dan LightGBM*. Dasar alasan pemilihan model-model algoritma tersebut adalah karena model-model tersebut biasa digunakan untuk pemodelan machine learning klasifikasi atau prediksi dan nantinya akan dipilih model mana yang memiliki skor terbaik. Selain berfokus pada *Accuracy*, nilai f1-score merupakan nilai yang menjadi acuan untuk mencari skor model terbaik karena, f1-score menggambarkan perbandingan rata-rata precision dan recall, nilai f1-score yang baik mengindikasikan bahwa model memiliki precision dan recall yang baik. Selanjutnya, melihat hasil skor tiap model berdasarkan parameter defaultnya dengan memasukkan list model ke data train sekaligus prediksi terhadap data test.

Tabel 5. Model *Algoritma*

MODEL	ACCURACY	PRECISION	RECALL	F1 SCORE
Logistic Regression	0.797918	0.628458	0.570916	0.794699
Ridge Classifier	0.800757	0.646552	0.538600	0.794574
KNN	0.752958	0.531306	0.533214	0.753100
Gaussian	0.630857	0.409128	0.901257	0.649534
SVC	0.795551	0.637363	0.520646	0.788492
Neural Network	0.777094	0.583658	0.538600	0.774138
Decision Tree	0.721249	0.473684	0.517056	0.725029
Random Forest	0.779460	0.600887	0.486535	0.771497
Gradient Boosting Classifier	0.796025	0.633475	0.536804	0.790312
AdaBoost Classifier	0.798391	0.632860	0.560144	0.794278
CatBoost Classifier	0.793658	0.630108	0.526032	0.787333
Hist Gradient Boosting	0.781354	0.595573	0.531418	0.777198
XGBoost	0.797444	0.638710	0.533214	0.791236
LightGBM	087506	0.613924	0.522442	0.781713

Berdasarkan parameter default, masing-masing model algoritma didapat hasil skor sementara yaitu model Logistic Regression memiliki skor tertinggi. Selanjutnya untuk meningkatkan skor, menurunkan kompleksitas, dan mempercepat proses model, dilakukan tahap seleksi fitur. Seleksi fitur adalah sebuah prosedur memilih atau melakukan filtrasi berbasis atribut pada data dalam jumlah yang besar kemudian direduksi ke tingkat yang lebih musah dikelola dengan tujuan menghilangkan atribut atau fitur yang tidak

mempunyai relevansi pengaruh terhadap data (Pratama et al., 2022). Setelah dilakukan seleksi fitur, didapat jumlah fitur atau atribut yang akan digunakan, sebelumnya berjumlah 29 menjadi 21. Setelah itu peneliti memperbaharui data training dan testing menggunakan atribut yang telah diseleksi. Seleksi fitur menghasilkan skor sebagai berikut:

Tabel 6. Hasil Score Seleksi Fitur

MODEL	ACCURACY	PRECISION	RECALL	F1 SCORE
Logistic Regression	0.800284	0.634731	0.570916	0.796764
Ridge Classifier	0.805017	0.657267	0.543986	0.798740
KNN	0.769049	0.566990	0.524237	0.766062
Gaussian	0.672977	0.439531	0.874327	0.691908
SVC	0.796498	0.639560	0.522442	0.789472
Neural Network	0.787033	0.603883	0.558348	0.784279
Decision Tree	0.733554	0.494828	0.515260	0.735257
Random Forest	0.773781	0.577909	0.526032	0.770255
Gradient Boosting Classifier	0.796971	0.631687	0.551167	0.792325
AdaBoost Classifier	0.801704	0.644958	0.551167	0.796444
CatBoost Classifier	0.791765	0.621622	0.536804	0.786624
Hist Gradient Boosting	0.799337	0.634343	0.563734	0.795384
XGBoost	0.800757	0.643460	0.547576	0.795325
LightGBM	0.799337	0.634343	0.563734	0.795384

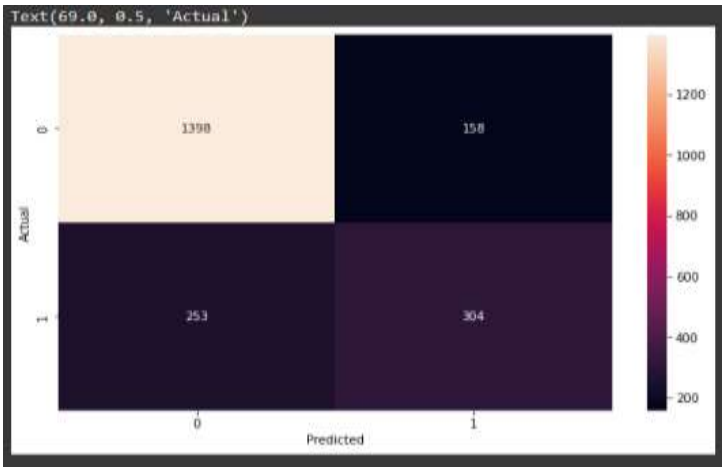
Dengan seleksi fitur, rata-rata setiap model algoritma mendapat peningkatan skor. Dan didapat model algoritma dengan accuracy dan f1 score tertinggi adalah Ridge Classifier maka model ini yang akan dipilih untuk pemodelan penelitian ini. Selanjutnya, kami melakukan tuning terakhir pada model yang dipilih berupa hyperparameter tuning agar skor diharapkan dapat meningkat lagi. Parameter yang akan dicari untuk *Ridge Classifier* yaitu sebagai berikut:

1. ‘solver’ = Solver yang digunakan untuk penyelesaian masalah. Opsi parameternya yaitu auto, svd, cholesky, lsqr , sparse_cg, ibfgs, saga dan sag.
2. ‘alpha’ = Parameter kekuatan regularisasi, untuk meningkatkan pengkondisian masalah dan mengurangi varians dari perkiraan. Semakin tinggi nilai alpha maka samakin kuat regulasi. Opsi nilai parameter yang dipilih yaitu 1e-5 1e-5, 1e-4, 1e-3, 1e-2, 1e-1, 1, 10, 100.
3. ‘fit_intercept’ = Penggunaan intersep untuk model. Opsi parameternya yaitu True, False.

Didapat kombinasi *hypmer parameter tuning* terbaik yaitu ‘solver’ = lsqr, ‘alpha’ = 1e-5, dan ‘fit_intercept’ = True.

1. Evaluation

Setelah pola klasifikasi didapatkan pada algoritma ridge classifier, Kemudian dilakukan pemeriksaan confussion matrix untuk melihat hasil evaluasi klasifikasi sebuah model machine learning dengan membandingkan hasil klasifikasi sebenarnya. Yaitu, melihat nilai akurasi, presisi dan recall. Berikut hasil evaluasi pada model yang telah dibuat dengan algoritma ini. Terdapat peningkatan skor dari sebelum dilakukan *hyperparameter tuning*



Gambar 4. *Confussion matrix*

Sumber: Data diolah

Dari 1.398 data pelanggan di prediksi tetap berlangganan dan ternyata memang tetap berlangganan (*True Positif*), lalu sebanyak 158 data customer diprediksi model berhenti berlangganan tetapi kenyataan tetap berlangganan *(False Negatif)*. Sedangkan sebanyak 253 data pelanggan diprediksi tetap berlangganan tetapi kenyataannya berhenti berlangganan *(False Positif)*, dan sebanyak 304 data baik prediksi maupun kenyataan berhenti berlangganan *(True Negatif)*. Berdasarkan gambar *confussion matrix* hasil *accuracy* 0.805 , *precision* 0.857 dan *recall* 0.898.

$$\text{Accuracy} = \frac{1.702}{2.113} \times 100\% = 80{,}5\%$$

$$\text{Precision} = \frac{1.398}{1.651} \times 100\% = 85{,}7\%$$

$$\text{Recall} = \frac{1.398}{1.756} \times 100\% = 89{,}8\%$$

Sehingga di dapatkan tingkat akurasi algoritma *Ridge Classifier* sebesar 80,5%, *presisi* dan *recall* masing-masing 85,7% dan 89,8%. Setelah model cukup memuaskan tahap selanjutnya adalah menyimpan model tersebut agar dapat di lanjut ke tahap *deployement*.

2. *Deployment*

Yaitu tahap evaluasi implementasi seluruh model dengan detail, menyesuaikan tahapan model hingga menghasilkan suatu capaian yang sesuai dengan target CRIPS-DM tersebut.

SIMPULAN

Penelitian ini berfokus untuk memprediksi pelanggan yang bertahan (loyal) atau pelanggan yang berhenti menggunakan layanan. Data yang digunakan berdasarkan data pelanggan Perusahaan Telekomunikasi (Telkom Indonesia). Sample data yang digunakan sebanyak 7.043 data record dan 21 atribut yang diambil dari pengguna layanan Telkom Indonesia Februari 2022 dengan metode purposive sampling. Hasil penelitian menunjukan sebanyak 73% atau 5.174 tetap menggunakan layanan dan 1.869 atau sekitar 27% pelanggan berhenti menggunakan layanan Telkom Indonesia. Metode yang digunakan CRISP-DM, dengan algortima ridge classifier dengan menggunakan paramenter confusion matrix, menghasilkan algoritma yang memiliki hasil cukup baik dengan akurasi 80,5%, precision 85,7% dan recall 89,8%. Penelitian ini diharapkan dapat membantu perusahaan dalam menentukan pelanggan yang dapat dipertahankan dan yang tidak dapat dipertahankan.

DAFTAR PUSTAKA

(a) BUKU

Aditra Pradyana, G., Mahendra Darmawiguna, I. gede, & Saputra Wahyu Wijaya, I. N. 2020. *Data Mining: Menemukan Pengetahuan Dalam Data* (1st ed.). PT Rajagrafindo Persada.

Dwijaya, B. A., & Wirawan, S. 2021. *Customer Churn Prediction in The Banking Industry Using CRISP-DM Utilizing Machine Learning Techniques*. 6(2), 92–106.

Hasanah, M. A., Soim, S., & Handayani, A. S. 2021. *Implementasi CRISP-DM Model Menggunakan Metode Decision Tree dengan Algoritma CART untuk Prediksi Curah Hujan Berpotensi Banjir*. 5(2).

Joko Suntoro. 2019. *Data Mining: Algortima dan Implementasi dengan Pemrograman PHP* (1st ed.). PT Elex Media Komputindo.

Mulaab. 2017. *DATA MINING: Konsep dan Aplikas*i (1st ed.). Media Nusa Creative.

Wijaya, A., & Girsang, A. S. 2016. *The Use of Data Mining For Prediction of Customer Loyalty*. 10(1), 41–47.

(b) JURNAL

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

Pambudi, H. K., Kusuma, P. G. A., Yulianti, F., & Julian, K. A. 2020. Prediksi Status Pengiriman Barang Menggunakan Metode Machine Learning. *Jurnal Ilmiah Teknologi Infomasi Terapan*, 6(2), 100−109. https://doi.org/10.33197/jitter.vol6.iss2.2020.396

Pratama, I., Chandra, A. Y., & Presetyaningrum, P. T. 2022. Seleksi Fitur dan Penanganan Imbalanced Data menggunakan RFECV dan ADASYN. *Jurnal Eksplora Informatika*, 11(1), 38−49. https://doi.org/10.30864/eksplora.v11i1.578

Rabiqy, Y. 2019. Pengaruh Ekspektasi Pelanggan, Kualitas Produk Dan Kepuasan Pelanggan Terhadap Loyalitas Pelanggan Internet Telkomsel Di Banda Aceh. *Jurnal Bisnis Dan Kajian Strategi Manajemen*, 1(1), 53−63. https://doi.org/10.35308/jbkan.v1i1.912

Rohana, T. 2020. Pengaruh Kepuasan Terhadap Loyalitas Pelanggan. *Jurnal Ilmu Manajemen*, 8(1), 28−32.

Rowley, J. (2005). The four Cs of customer loyalty. *Marketing Intelligence and Planning*, 23(6), 574−581. https://doi.org/10.1108/02634500510624138

Am, S., & Harun, H. (2023). Determining Qibla Direction of Mosques in Jambi Province : Method , Conflict , and Resolution. 01(01), 166–186.

Arrahman, A., & Yanti, I. (2022). Halal Industry in Javanese Culture; Yogyakarta Regional Government Policy in obtaining its economic values. INFERENSI: Jurnal Penelitian Sosial Keagamaan, 16(1), 151–174. https://doi.org/10.18326/infsl3.v16i1.151-174

Nengsih, T. A., Abduh, M., Ladini, U., & Mubarak, F. (2023). The Impact of Islamic Financial Development, GDP, and Population on Environmental Quality in Indonesia. International Journal of Energy Economics and Policy, 13(1), 7–13. https://doi.org/10.32479/ijeep.13727

Putra, D. . A., & Addiarrahman, A. (2023). Quranic Exegesis Journalism in Islamic Magazines in Indonesia Between 1970-1980. Journal of Indonesian Islam, 17(2), 483. https://doi.org/10.15642/jiis.2023.17.2.483-509

Rafidah, R. (2023). Indonesian islamic bank return on assets analysis: Moderating effect of musyarakah financing. Al-Uqud: Journal of Islamic Economics, 7(2), 200–216. https://journal.unesa.ac.id/index.php/jie/article/view/20310%0Ahttps://journal.unesa.ac.id/index.php/jie/article/download/20310/10813