PROKASDADIK, Volume 1, Desember (2023) E-ISSN: 3063-5845

Analisis dan Prediksi Kehilangan Nasabah pada Layanan Perbankan dengan *Machine Learning*

Agus Susilo
Teknik Informatika, Program Pascasarjana, Universitas Pamulang
e-mail: susilo.agus17@email.com

Abstrak*—Dalam era digital yang terus berkembang, industri perbankan Indonesia menghadapi tantangan yang signifikan dalam mempertahankan dan memperluas basis pelanggannya. Kehilangan nasabah (customer churn) menjadi fokus utama, mengingat dampaknya yang dapat merugikan secara finansial dan reputasi bagi bank. Penelitian ini bertujuan untuk menggali lebih dalam pola-pola dan faktor-faktor yang mempengaruhi keputusan pelanggan dalam meninggalkan sebuah bank. Kami memanfaatkan analisis data mendalam dan teknik pemodelan prediktif untuk meramalkan perilaku churn dan memberikan wawasan strategis bagi industri perbankan. Dalam penelitian ini, kami menyelidiki alasan di balik perpindahan nasabah dan penghentian penggunaan layanan perbankan. Data historis pelanggan dari sejumlah bank di Indonesia dianalisis secara menyeluruh. Kami menggunakan metode-metode statistik dan teknik machine learning untuk mengidentifikasi pola-pola perilaku nasabah sebelum mereka melakukan churn. Hasil analisis ini memberikan pemahaman mendalam tentang faktor-faktor yang menyebabkan kehilangan nasabah, termasuk tingkat kepuasan pelanggan, kualitas layanan, dan penawaran produk. Selain itu, kami mengembangkan model prediktif yang memanfaatkan variabel-variabel yang ditemukan melalui analisis data. Model ini memungkinkan prediksi kecenderungan churn berdasarkan pola perilaku pelanggan yang teramati. Dengan memahami sifat-sifat perilaku ini, bank dapat mengambil tindakan preventif dan proaktif untuk mempertahankan pelanggan. Selain itu, model prediktif ini dapat menjadi landasan untuk pengembangan strategi retensi yang disesuaikan, meningkatkan interaksi dengan pelanggan, dan memperbaiki layanan. Dengan memahami motivasi di balik keputusan pelanggan untuk berpindah atau berhenti menggunakan layanan, bank dapat mengurangi churn melalui peningkatan layanan pelanggan, personalisasi produk, dan pengelolaan proaktif hubungan pelanggan. Dengan demikian, penelitian ini bukan hanya memperluas pemahaman akademik tentang customer churn, tetapi juga memberikan pandangan praktis dan aplikatif bagi industri perbankan Indonesia dalam menghadapi tantangan retensi pelanggan di era digital ini.

***Kata Kunci—Churn, Machine Learning,* Prediksi.**

I. PENDAHULUAN

Dalam era digital yang berkembang pesat, industri perbankan menjadi salah satu sektor yang paling dipengaruhi oleh perubahan teknologi dan dinamika pasar yang cepat. Dalam konteks ini, retensi pelanggan merupakan faktor kunci yang menentukan keberhasilan suatu bank. Menjaga nasabah yang ada tidak hanya mengurangi biaya pemasaran dan akuisisi baru, tetapi juga meningkatkan reputasi dan kepercayaan pelanggan, yang menjadi modal penting dalam persaingan yang semakin ketat.

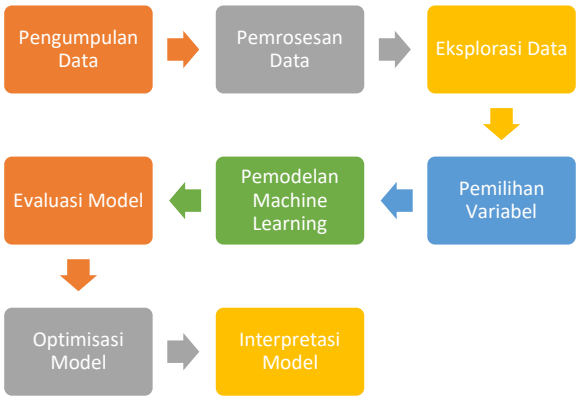
Salah satu aspek yang paling menantang dalam retensi pelanggan adalah customer churn, yaitu fenomena di mana nasabah memutuskan hubungan dengan bank dan beralih ke penyedia layanan finansial lainnya. Customer churn bukan sekadar kehilangan pelanggan; hal ini mencerminkan ketidakpuasan atau perubahan kebutuhan yang tidak terpenuhi. Oleh karena itu, pemahaman mendalam tentang faktor-faktor yang mempengaruhi customer churn dan kemampuan untuk meramalkannya menjadi sangat penting dalam strategi retensi pelanggan bank.

Penelitian ini bertujuan untuk menyelidiki dan memahami perilaku customer churn pada layanan perbankan dengan memanfaatkan kekuatan teknologi machine learning. Dengan menggunakan pendekatan data-driven yang canggih, penelitian ini akan menggali data historis pelanggan untuk mengidentifikasi pola-pola perilaku dan faktor-faktor yang memiliki dampak signifikan pada keputusan nasabah untuk meninggalkan layanan perbankan. Penerapan metode machine learning, seperti regresi logistik, decision trees, dan algoritma neural networks, akan memungkinkan kami untuk memodelkan hubungan kompleks antara variabel-variabel yang mempengaruhi customer churn [1].

Kecerdasan Buatan merupakan salah satu bidang dalam ilmu komputer yang ditujukan pada pembuatan software dan hardware yang dapat berfungsi sebagai sesuatu yang dapat berpikir seperti manusia. [2] Machine learning dapat didefinisikan sebagai aplikasi komputer dan algoritma matematika yang diadopsi dengan cara pembelajaran yang berasal dari data dan menghasilkan prediksi di masa yang akan datang. [3],[4] Penelitian terkini mengungkapkan bahwa machine learning terbagi menjadi tiga kategori: Supervised Learning, Unsupervised Learning, Reinforcement Learning. Teknik yang digunakan oleh Supervised Learning adalah metode klasifikasi di mana kumpulan data sepenuhnya diberikan label untuk mengklasifikasikan kelas yang tidak dikenal. Sedangkan teknik Unsupervised Learning sering disebut cluster dikarenakan tidak ada kebutuhan untuk pemberian label dalam

kumpulan data dan hasilnya tidak mengidentifikasi contoh di kelas yang telah ditentukan. [5],[6][7] Supervised learning memiliki beberapa algoritma populer seperti Back-propagation Linear regression, Random Forest, Support Vector Machines, Naive Bayesian, Metode Rocchio, Decision Tree, k-Nearest Neighbor, Neural Network, Logistic Regression dan Neural Network.

II. METODE PENELITIAN



Gambar 1.
Langkah-Langkah Metodologi Penelitian

A. Pengumpulan Data

Data pelanggan akan dikumpulkan dari satu bank yang relevan dengan tujuan penelitian. Data ini mencakup beragam informasi seperti riwayat transaksi, jumlah transaksi bulanan, jenis produk atau layanan yang digunakan, lamanya keanggotaan nasabah, informasi demografis (seperti usia, jenis kelamin, dan alamat), serta umpan balik pelanggan jika tersedia. Data juga dapat mencakup informasi lain yang dianggap relevan dengan keputusan customer churn, seperti tingkat kepuasan pelanggan atau interaksi pelanggan dengan layanan pelanggan.

B. Pemrosesan Data

Data yang terkumpul akan melalui tahap pemrosesan. Proses ini mencakup pembersihan data untuk mengatasi nilai yang hilang atau tidak valid, pengkodean variabel kategorikal seperti gender, status keanggotaan, dan normalisasi data untuk memastikan konsistensi dalam skala variabel.

C. Eksplorasi Data

Analisis eksplorasi data akan dilakukan untuk mengidentifikasi pola-pola awal dan hubungan antar variabel. Visualisasi grafis, statistik deskriptif, dan teknik eksplorasi data lainnya akan digunakan untuk memahami karakteristik data. Melalui eksplorasi data yang mendalam ini, pola-pola yang mungkin tidak terlihat pada pandangan pertama dapat diidentifikasi, membimbing pilihan variabel yang akan dimasukkan dalam model machine learning, dan memahami konteks yang lebih mendalam tentang perilaku pelanggan dalam konteks customer churn di layanan perbankan.

D. Pemilihan Variabel

Variabel-variabel yang memiliki dampak signifikan terhadap customer churn akan dipilih. Analisis korelasi dan teknik pemilihan fitur (*feature selection*) seperti Recursive Feature Elimination atau analisis keterbatasan varians (VIF) akan digunakan untuk memilih variabel-variabel yang paling informatif.

E. Pemodelan Machine Learning

Beberapa model machine learning akan dikembangkan untuk meramalkan customer churn. Ini termasuk regresi logistik, decision trees, random forests, dan neural networks. Data akan dibagi menjadi set pelatihan (*training set*) dan set pengujian (*testing set*) untuk melatih dan menguji kinerja model-model tersebut.

F. Evaluasi Model

Model-model yang dikembangkan akan dievaluasi menggunakan metrik-metrik seperti akurasi, presisi, recall, dan F1-score. Selain itu, teknik validasi silang (*cross-validation*) akan digunakan untuk memastikan keandalan model terhadap data yang tidak terlihat sebelumnya.

G. Optimisasi Model

Model-model yang telah dievaluasi akan dioptimalkan untuk meningkatkan kinerja mereka. Proses ini melibatkan penyesuaian parameter-model, teknik pengaturan ambang (*threshold tuning*), dan pemrosesan lebih lanjut pada variabel untuk meningkatkan prediksi customer churn.

H. Interpretasi Hasil

Hasil dari model-model yang telah dioptimalkan akan diinterpretasikan untuk mengidentifikasi faktor-faktor yang paling mempengaruhi keputusan customer churn. Analisis kontribusi variabel-variabel terhadap prediksi churn akan membantu bank dalam mengambil tindakan yang relevan untuk meningkatkan retensi pelanggan.

I. Pelaporan dan Presentasi

Temuan dari penelitian ini akan disajikan dalam laporan yang komprehensif, termasuk analisis data, metodologi, hasil evaluasi model, dan rekomendasi. Selain itu, presentasi visual, seperti grafik dan diagram, akan digunakan untuk menyajikan temuan secara jelas dan mudah dipahami.

Melalui langkah-langkah metodologi ini, penelitian ini akan memberikan pemahaman mendalam tentang pola-pola perilaku nasabah yang mengarah pada churn, memungkinkan bank-bank untuk mengoptimalkan strategi retensi pelanggan mereka dengan menggunakan pendekatan data-driven dan teknologi *machine learning*.

III. HASIL DAN PEMBAHASAN

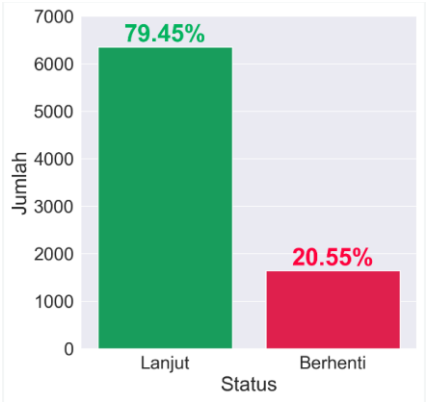
A. Pemrosesan Data

Dataset yang kami miliki memiliki 14 fitur atau atribut dan 10 ribu pelanggan. Fitur terakhir yaitu Exited adalah sebagai variable target yang menunjukan apakah pelanggan telah berpindah atau masih lanjut (0 = tidak dan 1 = ya). Fitur yang sifatnya spesifik akan kami hapus karena tidak diperlukan untuk dilanjutkan untuk proses Analisa selanjutnya seperti RowID, CustomerID dan FullName.

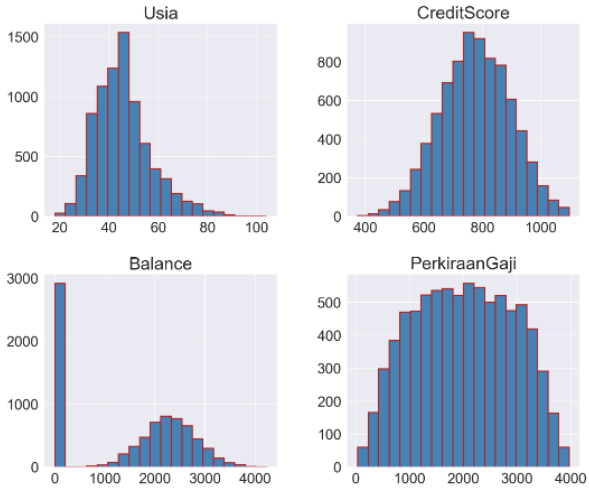
B. Eksplorasi Data Analisis

Target variable kami yang digunakan yaitu Exited. Target variable sudah diubah menjadi 2 kemungkinan, yaitu

- Nol (0) untuk customer yang tidak melanjutkan
- Satu (1) untuk customer yang melanjutkan



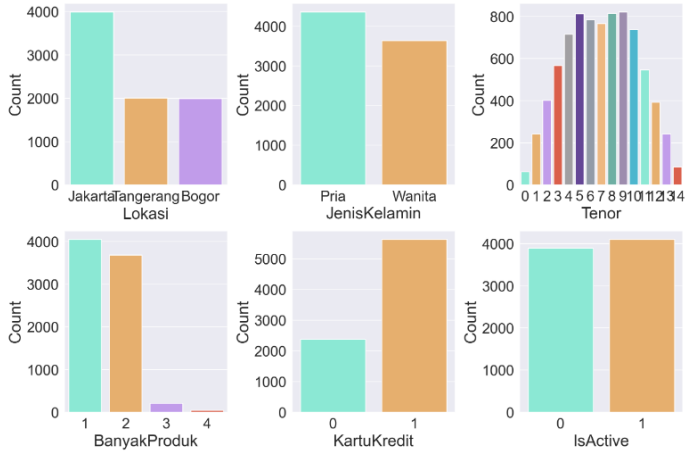
Gambar 2.
EDA dengan target Exited



Gambar 3.
Visualisasi dengan variable kontinue

Sebanyak 80% yang tetap bertahan sebagai client pada bank tersebut. Jika diperhatikan, dataset tersebut tidak seimbang dikarenakan jumlah instance dalam kelas Lanjut lebih banyak dibandingkan dengan jumlah instance dalam kelas Berhenti. Oleh karena itu, akurasi mungkin bukan metrik terbaik untuk kinerja model. Teknik visualisasi yang berbeda berlaku untuk jenis variabel yang berbeda, oleh karena itu, berguna untuk membedakan antara variabel kontinu dan kategorikal serta memeriksanya secara terpisah.

'Usia' memiliki ekor yang sedikit lebih panjang, yaitu, lebih banyak nilainya berada di sebelah kanan median daripada di sebelah kiri, Sebagian besar nilai untuk 'CreditScore' berada di atas 600, Jika kami abaikan kotak pertama, 'Balance' mengikuti distribusi yang cukup normal, dan Distribusi 'PerkiraanGaji' lebih atau kurang seragam dan memberikan sedikit informasi.



Gambar 4.
Visualisasi Variable kategorikal

Kali ini kami melakukan visualisasi dengan menggunakan variable kategorikal. Ada beberapa point yang dapat diambil dari visualisasi tersebut. Bank memiliki nasabah di 3 lokasi dan paling banyak berada di Jakarta. Wanita lebih banyak menjadi nasabah pada bank tersebut. Tenor antara 4 sampai 10 memiliki jumlah nasabah yang hampir sama. Kebanyakan nasabah membeli program 1 dan 2. Kebanyakan nasabah pada bank ini memiliki kartu kredit. Hampir 50 persen nasabah yang sudah tidak aktif.

C. Data Preprocessing

Tenor dan Kartu Kredit memiliki *chi square* yang rendah dan hipotesis awal ini bahwa kedua fitur ini tidak menyampaikan informasi yang berguna. Sehingga menggunakan metode *drop()* untuk menghapus tiga fitur ini dari *set data train*.

Table 1.
Chi Square Selection

	Variable	Chi-square	p-value
3	BanyakProduk	1233.595482	3.767234e-267
0	Lokasi	230.747892	7.829463e-51
5	IsActive	195.314895	2.199308e-44
1	JenisKelamin	90.172536	2.182709e-21
2	Tenor	11.998323	6.064375e-01
4	KartuKredit	0.300882	5.833299e-01

D. Data Modelling

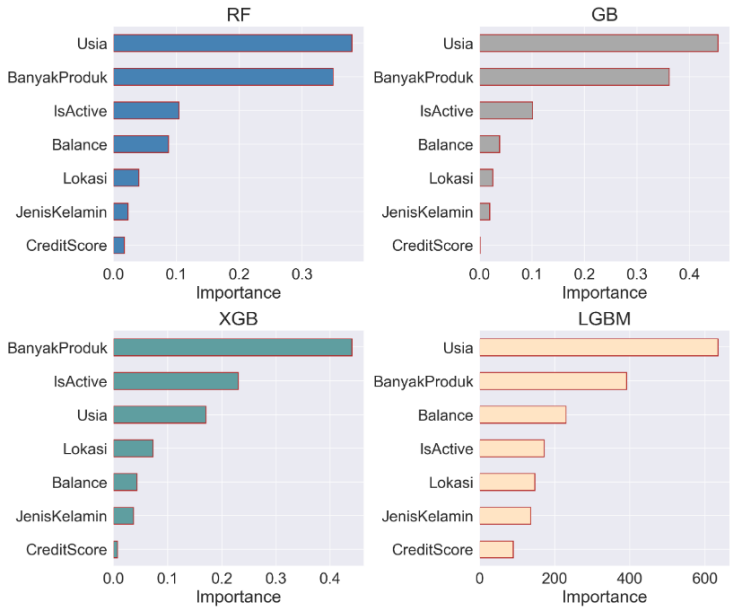
Dimulai dengan membuat dua model sederhana untuk memperkirakan kinerja dasar pada *training set*. Secara khusus, model yang akan digunakan yaitu *Gaussian Naïve Bayes* dan *Regresi Logistik*. Dengan menggunakan parameter default dan mengevaluasi recall (rata-rata) dengan melakukan validasi silang k-fold. Ide di balik validasi silang k-fold, yang diilustrasikan dalam gambar ini, adalah sederhana. ini membagi set pelatihan menjadi k subset/lipatan, melatih model menggunakan k-1 lipatan, dan mengevaluasi model pada satu lipatan yang tersisa. Proses ini diulang sampai setiap lipatan diuji satu kali.

E. Optimasi Model

Pada tahapan ini akan menggunakan 6 teknik klasifikasi yaitu *Logistic Regression, Support Vector Classifier, Random Forest Classifier, Gradient Boosting Classifier, Xtreme Gradient Boosting Classifier* dan *Light Gradient Boosting Machine*. Kemudian prediksi dari semua klasifikasi digabungkan untuk menentukan apakah akan didapatkan kinerja prediksi yang lebih baik.

F. Hasil

Untuk semua model, ada selisih kecil antara dua kurva di akhir pelatihan. Pengamatan ini menunjukkan bahwa tidak memperoverfit set pelatihan. Fitur Usia dan BanyakProduk terlihat lebih berguna untuk semua klasifikasi, diikuti dengan fitur IsActive dan Balance. Di sisi lain, CreditScore adalah fitur yang paling tidak penting dengan nilai kecil yang mendekati nol untuk semua estimator kecuali LGBM.



Gambar 5.
Fitur Penting

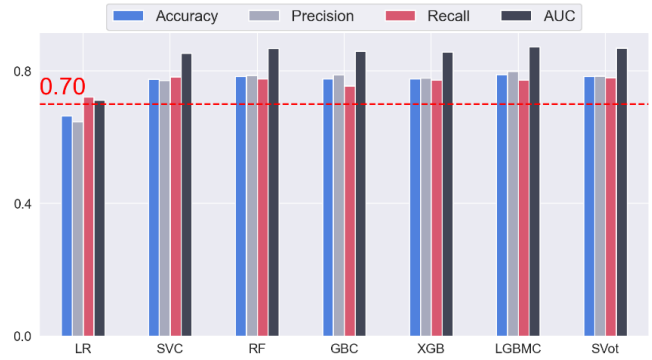
G. Evaluasi Hasil

Kita dapat membandingkan kinerja klasifikasi dalam hal empat metrik individual (akurasi, presisi, recall, dan luas area di bawah kurva ROC atau disebut AUC).

	Accuracy	Precision	Recall	AUC
LR	0.664176	0.647117	0.722152	0.712269
SVC	0.775330	0.771486	0.782410	0.853687
RF	0.783512	0.787492	0.776589	0.867705
GBC	0.776668	0.788894	0.755507	0.859650
XGB	0.776746	0.779118	0.772498	0.858127
LGBMC	0.789254	0.799089	0.772813	0.872784
SVot	0.783197	0.784630	0.780680	0.869363

Gambar 6.
Perbandingan Performa

Semua klasifikasi lain memiliki recall lebih tinggi dari 70% (kinerja dasar). XGB adalah model dengan recall tertinggi (78,5%). Namun, klasifikasi LGBM memiliki kinerja keseluruhan terbaik dengan akurasi, presisi, dan AUC tertinggi. Menggunakan satu metrik bukanlah satu-satunya cara untuk membandingkan kinerja prediksi model klasifikasi. Kurva ROC (Receiver Operating Characteristic) adalah grafik yang menunjukkan kinerja suatu klasifikasi pada berbagai ambang klasifikasi. Ini menggambarkan tingkat positif benar (nama lain dari recall) melawan tingkat positif salah.



Gambar 7.
Perbandingan Klasifikasi

Kinerja pada set uji untuk semua model cukup mirip dengan training set, hal ini membuktikan bahwa pengujian tidak memperoverfit *training set*. Oleh karena itu, prediksi kehilangan nasabah dengan recall sekitar 78%.

IV. KESIMPULAN

EDA dapat membantu dalam mengidentifikasi fitur-fitur yang berkontribusi pada pengunduran pelanggan. Selain itu, analisis pentingnya fitur dapat mengukur pentingnya setiap fitur dalam memprediksi kemungkinan pengunduran pelanggan. Hasil kami menunjukkan bahwa fitur yang paling signifikan adalah usia (pelanggan yang lebih tua lebih mungkin mengundurkan diri), diikuti oleh jumlah produk (memiliki lebih banyak produk meningkatkan kemungkinan pengunduran pelanggan). Bank dapat menggunakan temuan kami untuk menyesuaikan dan meningkatkan layanannya dengan cara yang meningkatkan kepuasan bagi pelanggan yang lebih mungkin mengundurkan diri.

DAFTAR PUSTAKA

[1] Roihan, A., Sunarya, P. A., & Wijaya, C. (2019) Auto Tee Prototype as Tee Golf Automation in Golf Simulator Studio. 2018 6th International Conference on Cyber and IT Service Management, CITSM 2018
[2] Goldberg, D. E., & Holland, J. H. (1988). Genetic algorithms and machine learning. Machine Learning, 3(2), 95–99.
[3] Somvanshi, M., & Chavan, P. (2016). A review of machine learning techniques using decision tree and support vector machine. 2016 International Conference on Computing Communication Control and Automation (ICCUBEA), 1–7. https://doi.org/10.1109/ICCUBEA.2016.7860040.
[4] Thupae, R., Isong, B., Gasela, N., & AbuMahfouz, A. M. (2018). Machine Learning Techniques for Traffic Identification and Classifiacation in SDWSN: A Survey. IECON 2018 - 44th Annual Conference of the IEEE Industrial Electronics Society, 4645–4650. https://doi.org/10.1109/IECON.2018.8591178.
[5] Negnevitsky, M. (2005). Artificial intelligence: a guide to intelligent systems. Pearson education.
[6] Brownlee, J. (2016). Master Machine Learning Algorithms: discover how they work and implement them from scratch.
[7] Lakshmi, J. V. N., & Sheshasaayee, A. (2015). Machine learning approaches on map reduce for Big Data analytics. 2015 International Conference on Green Computing and Internet of Things (ICGCIoT), 480–484.