Current Issues & Research in Social Sciences, Education and Management (CIRSSEM) Vol. 4 No. 1, pp. 16-26, January, 2026
E-ISSN: 3032-324X
https://www.ejournal.aibpmjournals.com/index.php/ssem/index

Ramalan Keciciran Pelanggan Telco Menggunakan Algoritma Pembelajaran Mesin

Xinying Chew*
School of Computer Sciences, Universiti Sains Malaysia, 11800 Pulau Pinang, Malaysia
*Corresponding Email: xinying@usm.my

ARTICLE INFORMATION

Publication information

Research article

HOW TO CITE
Chew, X. (2025). *Ramalan keciciran pelanggan telco menggunakan algoritma pembelajaran mesin*. Current Issues & Research in Social Sciences, Education and Management (CIR-SSEM), *4*(1), 16–26.

Received: 22 November 2025
Accepted: 23 December 2025
Published: 24 January 2026

ABSTRACT

Kertas kerja ini membentangkan satu kajian mengenai peramalan keciciran pelanggan dalam industri telekomunikasi menggunakan algoritma pembelajaran mesin. Satu koleksi data pelanggan sejarah daripada sebuah firma telekomunikasi digunakan dalam kajian ini, yang merangkumi maklumat demografi, butiran akaun pengguna dan trend penggunaan. Lapan algoritma pembelajaran mesin seperti Regresi Logistik (LR), Hutan Rawak (RF), Pengelas Pokok Keputusan (DTC), dan Mesin Vektor Sokongan (SVM), Naif Bayes (NB), K-Jiran Terdekat (KNN), AdaBoost, XGBoost telah diaplikasikan ke atas set data untuk meramalkan kebarangkalian keciciran pelanggan. Pembersihan data, analisis data penerokaan, pemilihan ciri dan, Pengekodan Integer, Pengekodan Satu Panas juga dilakukan sebelum membina model-model tersebut. Kajian menilai prestasi model berdasarkan pelbagai metrik penilaian seperti ketepatan, kejituan, dapatan semula dan skor F1. Keputusan menunjukkan model XGBoost mengatasi model lain dalam meramalkan keciciran pelanggan. Kajian ini memberikan wawasan berharga mengenai aplikasi algoritma pembelajaran mesin untuk meramalkan keciciran pelanggan dalam industri telekomunikasi, yang boleh membantu syarikat mengambil langkah proaktif untuk mengekalkan pelanggan dan mengurangkan kadar keciciran pelanggan.

Keywords: Pembelajaran Mesin, Peramalan Keciciran, K-Jiran Terdekat, Pokok Keputusan, Regresi Logistik, Hutan Rawak, Naif Bayes, AdaBoost, XGBoost

PENGENALAN

Keciciran pelanggan ialah peratusan pelanggan yang berhenti membeli produk atau perkhidmatan syarikat anda dalam tempoh masa tertentu. Sepanjang beberapa dekad yang lalu, industri telekomunikasi telah mengalami pertumbuhan yang ketara disebabkan oleh penyebaran peranti mudah alih dan internet. Pengekalan pelanggan telah menjadi aspek yang penting bagi syarikat telekomunikasi untuk terus menguntungkan memandangkan persaingan dengan pesaing yang semakin meningkat dan ketersediaan pelbagai pembekal perkhidmatan.

Kajian ini bertujuan untuk meramalkan keciciran pelanggan syarikat telekomunikasi menggunakan kaedah pembelajaran mesin terselia. Model-model dibina menggunakan data keciciran pelanggan telekomunikasi. Penemuan kajian ini akan memberikan wawasan berharga tentang penggunaan algoritma pembelajaran mesin untuk meramalkan keciciran pelanggan dalam industri telekomunikasi, dan syarikat dapat mengambil langkah proaktif supaya kadar keciciran pelanggan dapat diminimumkan.

Objektif kajian ini adalah untuk mengetahui ciri-ciri mana yang lebih berkaitan dengan keciciran pelanggan telekomunikasi, untuk membangunkan model pembelajaran mesin terselia untuk meramalkan keciciran pelanggan telekomunikasi pada masa depan, dan untuk memilih dan mencadangkan model yang paling sesuai dan tepat untuk meramalkan keciciran pelanggan telekomunikasi.

Untuk memenuhi objektif yang disenaraikan di atas, soalan penyelidikan berikut telah dirangka.

RQ1: Apakah teknik pembelajaran mesin terselia yang boleh digunakan untuk meramalkan keciciran pelanggan telekomunikasi?

RQ2: Ciri-ciri mana yang sangat berkorelasi dengan keciciran pelanggan telekomunikasi.

RQ3: Model pembelajaran mesin terselia mana yang memberikan prestasi terbaik dalam meramalkan keciciran pelanggan telekomunikasi?

RQ4: Apakah strategi dan cadangan yang boleh dicadangkan berdasarkan model pembelajaran mesin yang dipilih.

Kajian ini memberi tumpuan kepada pembinaan model pembelajaran mesin untuk meramalkan keciciran pelanggan telekomunikasi menggunakan set data yang sedia ada dan memilih model terbaik menggunakan metrik prestasi yang dipilih untuk model pembelajaran mesin. Kajian ini juga menemui korelasi antara ciri-ciri dan lajur sasaran iaitu keciciran pelanggan. Oleh itu, kajian ini tidak mempertimbangkan dan memasukkan pemboleh ubah lain yang tidak terdapat dalam set data yang digunakan yang mungkin mempengaruhi keputusan. Pelaksanaan dan penyebaran hasil daripada kajian lain juga tidak termasuk dalam skop kajian ini.

SOROTAN KAJIAN

Keciciran pelanggan berlaku apabila seorang pelanggan menghentikan hubungan dengan sebuah syarikat dan beralih kepada pesaing. Ini merupakan kebimbangan utama bagi perniagaan kerana ia menjejaskan pendapatan dan bahagian pasaran mereka. Keciciran boleh berlaku dalam pelbagai industri dan boleh disebabkan oleh

perkara seperti ketidakpuasan hati, harga yang tinggi, perkhidmatan pelanggan yang lemah, atau ketersediaan alternatif yang lebih baik. Memahami dan meramalkan keciciran pelanggan adalah penting bagi perniagaan untuk mengambil langkah proaktif bagi mengekalkan pelanggan dan mengurangkan kerugian pendapatan.

Peramalan keciciran pelanggan adalah tugas penting bagi syarikat dalam pelbagai industri, termasuk telekomunikasi, perbankan, e-dagang dan lain-lain. Ia melibatkan mengenal pasti pelanggan yang berkemungkinan akan berhenti atau beralih kepada perkhidmatan pesaing. Dengan meramalkan keciciran secara tepat, perniagaan boleh mengambil langkah proaktif untuk mengekalkan pelanggan, meningkatkan kepuasan pelanggan dan menambah keuntungan. Sorotan literatur ini memberikan gambaran keseluruhan penyelidikan dan pendekatan sedia ada untuk meramalkan keciciran pelanggan.

Regresi logistik adalah pendekatan pembelajaran mesin paling asas dan salah satu model linear teritlak (GLM). LR adalah teknik pengelasan binari yang terkenal dalam pembelajaran mesin terselia (Osisanwo et al., 2017). Ini adalah fungsi pengelasan yang dibina atas konsep kelas dan menggunakan model regresi logistik multinomial tunggal dengan satu penganggar. Ia boleh digunakan pada pemboleh ubah bersandar kategori, dengan hasilnya adalah pemboleh ubah kategori binari atau diskret 0 atau 1. Model regresi logistik lebih teguh berbanding regresi linear piawai (Nibbering & Hastie, 2022). Regresi logistik mempunyai ketepatan 85.2385% berdasarkan satu kajian juga mengenai ramalan keciciran menurut kajian yang dilakukan oleh Hemlata et al. pada tahun 2022.

Pokok Keputusan adalah sejenis pembelajaran mesin terselia yang boleh digunakan untuk pengelasan atau regresi. Kaedah ini bermula dari akar pokok, bergerak ke bawah ke dahan pokok yang sepadan dengan nilai atribut dalam contoh yang diberikan, dan kemudian diulang. Pendekatan ini diulang pada setiap nod sehingga daun ditemui, di mana kelas ramalan kes yang diberikan menjadi label (Alpaydin, 2014). Dalvi et al. (2016) menekankan kepentingan meramalkan keciciran pelanggan berdasarkan pokok keputusan sebagai teknik pembelajaran mesin utama untuk menilai keberkesanannya dalam menghasilkan keputusan.

Pengelas hutan rawak (RF) (Breiman, 2001) telah mendapat perhatian yang semakin meningkat sejak dua dekad yang lalu kerana keputusan pengelasan yang cemerlang yang diperoleh dan kelajuan pemprosesan (Du et al., 2015, Pal, 2005, Rodriguez-Galiano et al., 2012). Dengan menggunakan ramalan yang diperoleh daripada satu ensemble pokok keputusan, pengelas RF menghasilkan klasifikasi yang boleh dipercayai (Breiman, 2001). Tambahan pula, pengelas ini boleh digunakan untuk memilih dan mengkedudukan pemboleh ubah yang mempunyai keupayaan terbesar untuk membezakan antara kelas sasaran. Menurut kajian oleh Xie et al. (2009), penyelidik mencadangkan versi diperbaiki hutan rawak yang dikenali sebagai Improved Balanced Random Forest (IBRF) untuk meramalkan keciciran dalam industri perbankan. IBRF mempunyai kelebihan mengubah taburan kelas dan memberi penalti yang lebih teruk terhadap pengelasan salah kelas minoriti dengan menggabungkan teknik pensampelan dengan pembelajaran sensitif kos.

Mesin vektor sokongan adalah pengelas linear yang mengelaskan data dengan menyelesaikan masalah pengoptimuman kuadratik untuk menentukan satah pemisahan terbaik antara set data. Melalui kaedah pendarab Lagrange, SVM menukar masalah asal kepada masalah dwi yang lebih boleh diselesaikan (Scholkopf & Smolla, 2002). Dalam satu penyelidikan keciciran, untuk meningkatkan prestasi, parameter SVM dioptimumkan menggunakan carian grid dengan metrik penilaian tersuai yang boleh

diselaraskan berdasarkan kos strategi kempen pengekalan. Model yang dicadangkan menunjukkan keputusan yang memberangsangkan dan kuasa ramalan yang tinggi (Rodan et al., 2014).

Salah satu pendekatan pembelajaran mesin paling asas dan sering digunakan untuk mengkategorikan model data berdasarkan pengelasan jiran mereka. Sebagai contoh, apabila data dikategorikan sebagai bebas bunyi dan kecil. Set data harus disediakan sebelum menggunakan alat KNN (). Selepas menggunakan kaedah kNN untuk meramalkan hasil, prestasi diagnostik model harus dinilai. Nilai k, pengiraan jarak, dan pemilihan peramal semuanya mempunyai pengaruh besar terhadap prestasi model (Zhang, 2016).

Menurut kajian yang dilakukan oleh Sjarif et al. pada 2019, SVM meramalkan keciciran pada 83.67%, KNN pada 80.45% dan Hutan Rawak pada 76.36%. Semasa fasa pengujian, algoritma KNN mencapai ketepatan tertinggi iaitu 97.78%, manakala dua algoritma lain mencapai kurang daripada 80%, dengan Hutan Rawak mencapai 76.85% dan SVM mencapai 79.41%.

Salah satu kaedah pembelajaran mesin terselia yang paling popular adalah Pengelas Naif Bayes, pengelas kebarangkalian mudah berdasarkan teorem Bayes. Apabila teorem Bayes dan "Naif" digabungkan, setiap atribut adalah bebas (Hartatik et al., 2020). Sebaliknya, kaedah konvensional pemilihan atribut dalam Naif Bayes mempunyai kos pengiraan yang tinggi (Chen et al., 2020). Menurut kajian oleh Langarizadeh dan Moghbeli pada 2016, ulasan sistematik (23 kajian, 53,725 pesakit) menunjukkan bahawa menggunakan Rangkaian Naif Bayes untuk meramalkan penyakit mengatasi algoritma lain untuk majoriti penyakit. Berdasarkan ketepatan yang dilaporkan, Naif Bayes menunjukkan prestasi lebih baik daripada algoritma lain dalam kebanyakan situasi.

Untuk meningkatkan ketepatan pengkategorian, kaedah pembelajaran ensemble sering dicari dan digunakan. Dari sudut pandangan ini, penambahan pendekatan ensemble telah meningkatkan pengelas dengan ketara. Kaedah ini menggabungkan beberapa pelajar lemah atau pengelas untuk mengkategorikan set data secara berulang dan menawarkan klasifikasi yang boleh dipercayai. Teknik ensemble yang paling banyak digunakan adalah algoritma peningkatan (Schapire dan Singer, 1999). Adaptive Boosting, varian algoritma Boosting (AdaBoost), adalah salah satu teknik ini dalam penyelidikan Schapire dan Singer (1997). Menurut kajian keciciran pelanggan e-dagang oleh Wu et al. (2022) yang membandingkan model PCA-AdaBoost yang dicipta dalam kajian ini dengan beberapa model daripada literatur mereka. Penemuan menyokong keupayaan model PCA-AdaBoost untuk meramalkan keciciran pelanggan dengan ketepatan yang lebih tinggi dan kestabilan keseluruhan yang lebih baik.

Dalam pelbagai kajian peramalan keciciran pelanggan, XGBoost telah terbukti mengatasi algoritma pembelajaran mesin popular lain. Sebagai contoh, XGBoost didapati mempunyai ketepatan dan skor F1 tertinggi antara model yang diuji dalam satu kajian membandingkan algoritma pembelajaran mesin berbeza untuk peramalan keciciran pelanggan dalam industri telekomunikasi (Malik et al., 2023). Kajian lain menemui bahawa XGBoost mengatasi algoritma pembelajaran mesin lain dalam industri perbankan dalam meramalkan keciciran pelanggan.

METODOLOGI

Pemulihan Data

Set data yang digunakan untuk kajian ini ialah Telco Customer Churn, yang menawarkan butiran pengguna telekomunikasi dan perkhidmatan yang dipilih. Set data ini bebas diakses dan sedia untuk dimuat turun dari Kaggle. Selepas memuat turun set data dalam format fail CSV, fungsi read.csv() menggunakan pustaka Pandas telah digunakan untuk mengimport data daripada set data ke dalam Python menggunakan Jupyter Notebook. Fail CSV boleh ditukar menjadi bingkai data atau struktur data berlabel 2-dimensi dengan lajur pelbagai jenis menggunakan fungsi read_csv().

Setiap baris mewakili seorang pelanggan, setiap lajur mengandungi atribut pelanggan yang diterangkan pada Metadata lajur. Set data termasuk maklumat tentang:

- Pelanggan yang telah meninggalkan syarikat dalam bulan lepas – lajur ini dipanggil **Churn**
- Perkhidmatan yang telah didaftarkan oleh setiap pelanggan – telefon, pelbagai talian, internet, keselamatan dalam talian, sandaran dalam talian, perlindungan peranti, sokongan teknikal, dan strim TV serta filem
- Maklumat akaun pelanggan – tempoh mereka menjadi pelanggan, kontrak, kaedah pembayaran, pengebilan tanpa kertas, caj bulanan, dan jumlah caj
- Maklumat demografi tentang pelanggan – jantina, julat umur, dan sama ada mereka mempunyai pasangan serta tanggungan

Analisis Data Eksploratori (EDA)

Statistik deskriptif digunakan untuk merumuskan ciri-ciri fitur dalam set data dari segi taburan, kecenderungan memusat, dan kebolehubahan untuk pemboleh ubah selanjar serta perkadaran untuk pemboleh ubah kategori. Dengan menggunakan fungsi describe() dari Pandas, analisis deskriptif asas untuk pemboleh ubah selanjar telah dilakukan.

Peta haba korelasi hanya mendedahkan korelasi linear antara dua pemboleh ubah. Ini menunjukkan kemungkinan adanya hubungan langsung antara mereka. Tambahan pula, plot ini memudahkan analisis set data dan memahami cara pemboleh ubah berinteraksi antara satu sama lain. Matriks korelasi dipaparkan menggunakan fungsi peta haba Seaborn.

Pra-Pemprosesan Data

Set data asal mengandungi lajur ID pelanggan. Lajur ini dikeluarkan menggunakan fungsi dataframe.drop() dari pustaka pandas kerana ia hanya bertindak sebagai nombor pengenalan unik untuk merekodkan pelanggan dan tidak akan memberikan sebarang nilai yang bermakna serta berguna untuk meramalkan kecicirian.

Jumlah caj mula-mula ditukar kepada jenis data berangka menggunakan fungsi pd.to_numeric(). Kemudian, fungsi dataframe.isnull() dan sum() digunakan untuk mencari dan menunjukkan nilai yang hilang. Baris-baris yang mengandungi nilai yang hilang dikeluarkan menggunakan fungsi drop().

Pemilihan fitur digunakan untuk memilih fitur-fitur yang penting kepada pemboleh ubah sasaran. Untuk pemilihan fitur bagi fitur kategori, Ujian Chi-Square digunakan untuk melakukan pemilihan fitur. Pemilihan fitur bagi fitur berangka dilakukan menggunakan ujian ANOVA. Selepas ujian-ujian ini, fitur-fitur PhoneService, gender, StreamingTV, StreamingMovies, MultipleLines, InternetService dikeluarkan.

Penskalaan fitur dilakukan menggunakan fungsi StandardScaler() dan fungsi Scaler.fit_transform(). Menskala fitur dalam model pembelajaran mesin dapat mempercepatkan peminimuman fungsi kos dan membuat aliran kecerunan turunan lebih lancar, yang kedua-duanya membantu proses pengoptimuman. Tanpa menskala fitur, algoritma mungkin cenderung kepada fitur dengan nilai magnitud yang lebih besar.

Pengekodan integer dilakukan untuk menukar "Ya" dan "Tidak" kepada 1 dan 0. Lajur-lajur lain yang mengandungi "ya" dan "tidak" termasuk Partner, Dependents, OnlineSecurity, OnlineBackup, DeviceProtection, TechSupport, PaperlessBilling dan fitur sasaran, Churn. Semua lajur-lajur ini dengan 'ya' digantikan dengan 1 dan 'tidak' digantikan dengan 0.

Sebelum boleh digunakan untuk melatih dan menguji model, data teks atau kategori mesti dikodkan sebagai nilai integer. Dalam kajian ini, pemboleh ubah kategori ditukar kepada pemboleh ubah berangka menggunakan Pengekod Satu Panas dari pustaka pra-pemprosesan Scikit-learn. Pengekodan satu panas dilakukan untuk lajur kategori khususnya untuk Contract dan PaymentMethod. Contract dan PaymentMethod dibahagikan kepada lajur berbeza untuk setiap atribut. Sebagai contoh untuk kontrak, ia dibahagikan kepada Contract_Month-to-month, Contract_Oneyear, Contract_Twoyear, dan begitu juga untuk PaymentMethod.

Pembahagian Data
Kajian ini menunjukkan bahawa pembahagian data adalah teknik yang digunakan secara meluas untuk pengesahan model, di mana set data yang diberikan dibahagikan kepada dua set berasingan untuk latihan dan pengujian. Setelah disesuaikan dengan set latihan, model pembelajaran mesin kemudiannya diuji pada set ujian. Kita boleh menilai keberkesanan pelbagai model tanpa mengambil kira sebarang bias yang diperkenalkan semasa latihan mereka dengan menyediakan satu set data untuk pengesahan tambahan kepada latihan mereka (Joseph, 2022).

Untuk kajian ini, set data dibahagikan kepada 80% set latihan dan 20% set ujian. Ini dicapai menggunakan fungsi train_test_split dari pustaka pemilihan model Scikit-learn. Apabila parameter ditetapkan, fungsi ini mengekstrak 80% sampel bersama-sama dengan label yang sepadan untuk dijadikan set latihan dan baki 20% untuk dijadikan set ujian. Selain itu, parameter random_state ditetapkan untuk menghasilkan keputusan yang sama setiap kali fungsi dipanggil.

Pemodelan Data
Terdapat lapan model pengelasan iaitu LR, DT, RF, SVM, KNN, NB, AdaBoost, XGBoost dan semua model ini akan dibina dalam peringkat ini.

Untuk setiap model, algoritma dan parameter yang dibekalkan pada mulanya disimpan dalam suatu objek. Objek ini juga mengandungi data yang diekstrak algoritma dari set latihan. Model pengelasan kemudian dibangunkan pada data latihan dengan mengaplikasikan kaedah fit kepada objek yang dicipta sebelumnya untuk mengambil argumen sebagai tatasusunan NumPy, x_train dan y_train, di mana x adalah ciri-ciri dan y adalah sasaran. Melaksanakan kaedah predict pada objek yang mengandungi ciri-ciri set ujian (x_test) kemudiannya akan menghasilkan ramalan.

Penilaian Model
Skor pengesahan silang, ketepatan, skor ROC-AUC, dan skor F1 digunakan sebagai empat metrik prestasi untuk menilai prestasi pengelas. Matriks kekeliruan berfungsi sebagai asas untuk metrik prestasi. Matriks kekeliruan adalah sejenis jadual khusus yang membolehkan visualisasi prestasi sesuatu pengelas. Pengelas yang baik

mempunyai nombor Positif Benar dan Negatif Benar yang besar serta nilai Positif Palsu dan Negatif Palsu yang kecil.

Pengesahan silang ialah teknik yang digunakan untuk menilai prestasi model pada data yang belum diuji. Keupayaan model untuk mengitlur kepada data baharu ditunjukkan oleh skor pengesahan silang, juga dikenali sebagai ketepatan pengesahan silang atau ketepatan disahkan silang. Set data yang tersedia dibahagikan kepada beberapa lipatan atau subset untuk pengesahan silang. Sebahagian data (set latihan) digunakan untuk melatih model, dan baki data (set pengesahan) digunakan untuk menilainya. Prosedur ini diulang beberapa kali, menggunakan kombinasi set latihan dan pengesahan yang berbeza setiap kali. Prestasi model merentasi semua set pengesahan kemudiannya diratakan untuk menentukan skor pengesahan silang.

Ketepatan atau kadar kejayaan keseluruhan model adalah ukuran sejauh mana rekod telah dikelaskan dengan betul. Ia ditentukan dengan mendarabkan jumlah keseluruhan kejadian dengan jumlah TP dan TN. Ia mengira perkadaran peristiwa yang diramalkan dengan betul. Perkadaran hasil yang diramalkan secara salah diukur oleh kadar pengelasan salah, yang sering wujud. Model yang mantap juga sepatutnya mempunyai kadar pengelasan salah yang lebih rendah; skor ketepatan tinggi sahaja tidak menunjukkan ini.

$$Ketepatan = \frac{TP + TN}{TP + FP + TN + FN}$$

Memilih model berprestasi terbaik semata-mata berdasarkan ketepatan dan kadar pengelasan salah boleh mengelirukan kerana set data yang tidak seimbang. Dalam kes ini, ketepatan tinggi dan kadar pengelasan salah yang rendah boleh dicapai dengan meramalkan kelas majoriti atau kelas positif secara tepat. Oleh itu, ukuran lain digunakan untuk mendapatkan model terunggul, yang diterangkan dalam subseksyen berikut.

Memplot Kadar Positif Benar (TPR) berbanding Kadar Positif Palsu (FPR) pada pelbagai ambang pengelasan menghasilkan lengkung ROC. TPR mengukur perkadaran ramalan positif benar (kes positif yang dikenal pasti dengan betul) antara semua kes positif sebenar, manakala FPR mengukur perkadaran ramalan positif palsu (kes negatif yang dikenal pasti secara salah) antara semua kes negatif sebenar. Skor ROC AUC, yang berada dalam julat 0 hingga 1, adalah luas di bawah lengkung ROC. Pengelas sempurna yang boleh membezakan antara kelas positif dan negatif dengan sempurna mempunyai skor ROC AUC 1. Pengelas yang berprestasi tidak lebih baik daripada tekaan rawak mempunyai skor ROC AUC 0.5, manakala model yang berprestasi lebih teruk daripada rawak mempunyai skor di bawah 0.5.

$$Skor\ F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

KEPUTUSAN DAN PERBINCANGAN

Pemilihan fitur telah dilakukan dalam bahagian sebelumnya. Selepas pemilihan fitur dan pemeriksaan korelasi, hanya 14 fitur yang tinggal akan dipilih iaitu SeniorCitizen, Partner, Dependents, tenure, OnlineSecurity, OnlineBackup, DeviceProtection, TechSupport, PaperlessBilling, MonthlyCharges, TotalCharges. Dua lagi fitur yang telah melalui pengekodan satu panas ialah Contract yang dibahagikan kepada 3 lajur lagi dan PaymentMethod kepada 4 lajur.

Dalam kajian ini, lapan algoritma pembelajaran mesin terselia terpilih digunakan untuk meramalkan keciciran. Skor pengesahan silang untuk lapan model ini hanya mempunyai julat kecil dari 79.98% (KNN) hingga 84.85% (XGBoost). Ini mencadangkan bahawa, secara purata, XGBoost menunjukkan prestasi baik merentasi pelbagai subset data. Ini menunjukkan model telah mengenal pasti corak signifikan dan boleh secara konsisten membuat ramalan tentang data yang tidak diketahui.

Ketepatan adalah istilah yang digunakan untuk menggambarkan sejauh mana model meramalkan atau mengkategorikan data. Dengan menggunakan Python untuk melaksanakan proses pembelajaran mesin, ketepatan lapan algoritma berjulat dari 69.72% hingga 79.39%. Model ketepatan terendah adalah Naif Bayes manakala XGBoost mempunyai ketepatan tertinggi. Oleh itu, XGBoost boleh dikatakan mempunyai keupayaan terbaik untuk menunjukkan perkadaran kes yang dikelaskan dengan betul daripada jumlah kes dalam set data.

Dalam pembelajaran mesin untuk masalah pengelasan binari, skor ROC AUC (Receiver Operating Characteristic Area Under the Curve) adalah metrik penilaian yang kerap digunakan. Semakin tinggi skor ROC AUC, semakin baik keupayaan model untuk membezakan antara kelas positif dan negatif. Skor berjulat dari 67.95% (SVM) hingga 73.24% (NB) bermaksud Naif Bayes mempunyai keupayaan terbaik untuk membezakan antara kelas positif dan negatif.

Semua model mempunyai skor F1 yang rendah yang hanya berjulat dari 53% hingga 59%. Model dengan skor F1 terendah adalah SVM dan RF manakala tertinggi ialah NB. Skor F1 yang lebih rendah menunjukkan bahawa kejituan dan dapatan semula model adalah tidak seimbang secara signifikan. Sebab terperinci mengapa skor F1 rendah akan dijelaskan lanjut dalam bahagian perbincangan.

XGBoost mendahului dan cemerlang kerana ia mempunyai skor pengesahan silang dan ketepatan tertinggi sambil mempunyai skor ROC-AUC dan skor F1 yang sederhana. Oleh itu, daripada kajian ini dapat disimpulkan bahawa XGBoost adalah model yang paling sesuai untuk meramalkan keciciran pelanggan telekomunikasi.

Jadual 1. Keputusan prestasi untuk model pembelajaran mesin

	Skor Cross-validation	ROC-AUC	Ketepatan	Skor F1
LR	84.33%	70.08%	78.61%	56%
DT	82.56%	69.04%	78.46%	55%
RF	84.48%	68.08%	79.18%	53%
SVM	79.98%	67.95%	78.61%	53%
KNN	81.34%	69.46%	77.83%	55%
NB	82.04%	73.24%	69.72%	59%
AdaBoost	84.51%	69.25%	77.90%	55%
XGBoost	84.85%	69.93%	79.39%	56%

EDA menunjukkan bilangan pengguna lelaki dan perempuan hampir sama, dan jantina tidak dipilih sebagai fitur kerana ia tidak mempengaruhi keputusan keciciran. Kontrak Berasaskan Bulan-ke-Bulan menunjukkan kadar keciciran pelanggan yang tinggi. Ini mungkin kerana pelanggan sedang mencuba pelbagai perkhidmatan yang tersedia

untuk mereka, jadi untuk menjimatkan wang, mereka hanya menguji perkhidmatan satu bulan. Faktor lain mungkin ialah kualiti keseluruhan perkhidmatan telefon, strim dan internet berbeza-beza. Setiap pelanggan mempunyai keutamaan yang berbeza, jadi jika mana-mana satu daripada tiga memenuhi jangkaan, keseluruhan perkhidmatan digantung. PaperlessBilling menunjukkan kadar keciciran pelanggan yang tinggi. Ini kemungkinan besar adalah hasil daripada isu pembayaran atau resit. Pelanggan yang tidak keciciran pada bulan-bulan awal juga cenderung untuk kekal lebih lama dan terus menggunakan perkhidmatan telekomunikasi.

Dari segi metrik prestasi, seperti yang dapat dilihat daripada keputusan, skor F1 adalah agak rendah untuk semua model. Dengan menggunakan set data ini, ia mengakibatkan dapatan semula yang rendah kerana ia kehilangan sejumlah besar kes positif sebenar. Ini adalah kerana bilangan negatif yang mewakili tiada keciciran adalah lebih tinggi dalam set data, menyebabkan set data tidak seimbang kerana Negatif Benar akan jauh lebih banyak daripada Positif Benar. Daripada Lampiran, dapat dilihat bahawa kes positif (keciciran) adalah jauh lebih rendah berbanding kes negatif (tiada keciciran).

KESIMPULAN

Kesimpulannya, kajian ini bertujuan untuk meramalkan keciciran pelanggan syarikat telekomunikasi menggunakan set data dan untuk mengetahui ciri-ciri yang lebih penting dan mempunyai korelasi lebih tinggi dengan pemboleh ubah sasaran iaitu keciciran. Satu sorotan literatur juga dijalankan untuk mengetahui lebih lanjut tentang penyelidikan dan kajian lepas mengenai algoritma pembelajaran mesin yang digunakan untuk kajian ini dan bidang berkaitan lain. Data dibahagikan kepada nisbah latihan:ujian 80:20 dan dinilai menggunakan metrik prestasi seperti yang ditunjukkan dalam Jadual 4.1. Model yang menunjukkan prestasi terbaik dari kajian ini adalah XGBoost berdasarkan metrik prestasi. Oleh itu, daripada kajian ini dapat disimpulkan bahawa XGBoost adalah algoritma pengelasan yang paling sesuai untuk menyelesaikan ramalan keciciran pelanggan telekomunikasi.

Kajian ini memberikan wawasan apabila ciri-ciri yang mempengaruhi pemboleh ubah sasaran keciciran ditemui. Pembelajaran mesin terselia terbaik untuk pengelasan juga telah dipilih. XGBoost boleh digunakan untuk meramalkan keciciran menggunakan ciri-ciri yang dipilih dan kajian ini boleh menjadi rujukan untuk kajian pembelajaran mesin masa depan untuk diaplikasikan dalam bidang berkaitan.

Batasan pertama kajian ini ialah tidak semua pemboleh ubah relevan yang mungkin mempunyai kesan ke atas keciciran pelanggan diambil kira semasa mencipta set data. Set data semasa mungkin tidak merangkumi semua ciri relevan, dan akibatnya, pengelas tidak dilatih pada ciri-ciri tersebut.

Selain itu, hanya satu set data digunakan untuk penemuan kajian ini, jadi penemuannya terhad kepada maklumat yang dikumpul daripada data yang dikandungnya. Penemuan tidak boleh digunakan untuk membuat kesimpulan tentang set data lain dalam konteks atau bidang yang berbeza.

Skop kajian ini boleh diperluaskan dengan menilai kepentingan setiap ciri dalam set data dan dengan menggabungkan ciri tambahan yang penting tetapi tidak dimasukkan dalam set data semasa. Untuk membangunkan dan memperhalusi hasil kajian, penyelidikan

masa depan boleh bereksperimen dengan atribut kepentingan ciri berbeza untuk setiap pengelas untuk melihat sejauh mana kegunaan setiap ciri dalam meramalkan hasil dalam senario pertama. Dalam kes kedua, lebih banyak penyiasatan diperlukan untuk mengenal pasti faktor bebas tambahan yang memberi kesan signifikan ke atas keciciran pelanggan. Ini boleh dianalisis secara statistik dengan merumuskan beberapa hipotesis dan menentukan sama ada ia signifikan secara statistik atau tidak.

Akhir sekali, untuk menangani set data tidak seimbang yang mempunyai terlalu banyak 'tidak' (0) untuk keciciran berbanding keciciran (1), SMOTE boleh digunakan. Terdapat dua pilihan untuk melakukan ini. Pertama, pensampelan bawah iaitu mengurangkan sampel majoriti pemboleh ubah sasaran. Kedua, pensampelan atas iaitu menambah sampel minoriti pemboleh ubah sasaran kepada sampel majoriti. Percubaan dan kesilapan boleh dilakukan untuk memilih pilihan mana yang memberikan hasil terbaik.

RUJUKAN

Alpaydin, E. (2014). *Introduction to machine learning*.

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32.

Chen, S., Webb, G. I., Liu, L., & Ma, X. (2020). A novel selective naïve Bayes algorithm. *Knowledge-Based Systems, 192*, 105361. https://doi.org/10.1016/j.knosys.2019.105361

Dalvi, P. K., Khandge, S. K., Deomore, A., Bankar, A., & Kanade, V. (2016). Analysis of customer churn prediction in telecom industry using decision trees and logistic regression. *2016 Symposium on Colossal Data Analysis and Networking (CDAN)*, 1–4. IEEE.

Du, P., Samat, A., Waske, B., Liu, S., & Li, Z. (2015). Random forest and rotation forest for fully polarized SAR image classification using polarimetric and spatial features. *ISPRS Journal of Photogrammetry and Remote Sensing, 105*, 38–53.

Hartatik, Kusrini, K., & Prasetio, A. B. (2020). Prediction of student graduation with Naive Bayes algorithm. *2020 Fifth International Conference on Informatics and Computing (ICIC)*. https://doi.org/10.1109/ICIC50835.2020.928862

Hemlata, J., Ajay, K., & Sumit, S. (2019). Churn prediction in telecommunication using logistic regression and LogitBoost. *International Conference on Computational Intelligence and Data Science*, 101–102, 10–12.

Joseph, V. R. (2022). Optimal ratio for data splitting. *Statistical Analysis and Data Mining: The ASA Data Science Journal, 15*(4), 531–538. https://doi.org/10.1002/sam.11583

Langarizadeh, M., & Moghbeli, F. (2016). Applying naive Bayesian networks to disease prediction: A systematic review. *Acta Informatica Medica, 24*(5), 364–369. https://doi.org/10.5455/aim.2016.24.364-369

Malik, S., & Runwal, S. (2023). A study on customer churn prediction. *International Research Journal of Modernization in Engineering Technology and Science, 5*(4), 3600–3604. https://doi.org/10.56726/IRJMETS36156

Nibbering, D., & Hastie, T. J. (2022). Multiclass-penalized logistic regression. *Computational Statistics & Data Analysis, 169*, 107414.

Osisanwo, F. Y., Akinsola, J. E. T., Awodele, O., Hinmikaiye, J. O., Olakanmi, O., & Akinjobi, J. (2017). Supervised machine learning algorithms: Classification and comparison. *International Journal of Computer Trends and Technology, 48*(3), 120–126.

Rodan, A., Faris, H., Alsakran, J., & Al-Kadi, O. (2014). A support vector machine approach for churn prediction in telecom industry. *Information – An International Interdisciplinary Journal, 17*(8).

Schapire, R. E., & Singer, Y. (1999). Improved boosting algorithms using confidence-rated predictions. *Machine Learning, 37*(3), 297–336. https://doi.org/10.1023/A:1007614523901

Schölkopf, B., & Smola, A. (2002). *Learning with kernels: Support vector machines, regularization, optimization, and beyond*. MIT Press.

Sjarif, N. N. A., Yusof, M. R. M., Wong, D. H., Ya'akob, S., Ibrahim, R., & Osman, M. Z. (2019). Customer churn prediction using Pearson correlation function and K-nearest neighbor algorithm for telecommunication industry. *International Journal of Advance Soft Computing and Applications, 11*(2).

Wu, Z., Jing, L., & Wu, B. (2022). A PCA-AdaBoost model for e-commerce customer churn prediction. *Annals of Operations Research*. https://doi.org/10.1007/s10479-022-04526-5

Xie, Y., Li, X., Ngai, E. W. T., & Ying, W. (2009). Customer churn prediction using improved balanced random forests. *Expert Systems with Applications, 36*(3), 1–? https://doi.org/10.1016/j.eswa.2008.06.121

Zhang, Z. (2016). Introduction to machine learning: k-nearest neighbors. *Annals of Translational Medicine, 4*(11).

MENGENAI PENGARANG

XINYING CHEW menerima ijazah Ph.D. dalam kawalan kualiti statistik dari Universiti Sains Malaysia. Beliau kini merupakan Profesor Madya di Pusat Pengajian Sains Komputer, Universiti Sains Malaysia. Beliau juga merupakan Jurulatih Bertauliah dengan Tabung Pembangunan Sumber Manusia (HRDF) dan Teknologis Profesional dengan Lembaga Teknologis Malaysia (MBOT). Bidang penyelidikan beliau merangkumi analisis data lanjutan dan kawalan kualiti/proses statistik.