ISBN 978-623-93343-5-2
Seminar Nasional Informatika, Sistem Informasi dan Keamanan Siber
(SEINASI-KESI)
Jakarta-Indonesia, Januari 2022

IMPLEMENTASI CHURN PREDICTION DI INDUSTRI TELEKOMUNIKASI DENGAN METODE LOGISTIC *REGRESSION* DAN *CORRELATION-BASED FEATURE SELECTION*

Dhea Laksmi Prianto[1], Iin Ernawati[2], Nurul Chamidah[3].

Fakultas Ilmu Komputer
Universitas Pembangunan Nasional Veteran Jakarta
Email : dhealaksmi@gmail.com[1], iinerti@gmail.com[2] ,nurul.chamidah@upnvj.ac.id[3]
Jl. Rs. Fatmawati, Pondok Labu, Jakarta Selatan, DKI Jakarta, 12450, Indonesia

Abstrak

Industri telekomunikasi menjadi salah satu industri yang semakin kompetitif dan telah menumbuhkan ketertarikan akan prediksi churn untuk pelanggan. Prediksi churn tentang bagaimana mendeteksi pelanggan yang memiliki kecenderungan untuk meninggalkan layanan. Prediksi ini merupakan salah satu upaya perusahaan untuk mempertahankan pelanggan didalam Customer Relationship Management (CRM). Beberapa penulis mengemukakan bahwa metode logistic regression memiliki pemodelan dan hasil performa yang bagus untuk diaplikasikan untuk data prediktif. Dataset diambil dari salah satu perusahaan telekomunikasi di Amerika bernama Orange yang tersedia di situs Kaggle yang kemudian diolah untuk menganalisis performa prediksi churn menggunakan data mining dengan teknik pemilihan correlation-based feature selection forward selection serta algoritma machine learning logistic regression.

Kata kunci: Prediksi churn, logistic regression, data mining, correlation-based feature selection, forward selection

1 PENDAHULUAN

Perkembangan teknologi berdampak besar dalam persebaran informasi. Yang menjadikan industri telekomunikasi sebagai aspek yang paling penting dalam persebaran informasi tersebut. Perusahaan telekomunikasi bersaing dalam meningkatkan kualitas layanan untuk menjaga kepuasan pengguna dan menjaga konsumen yang telah ada agar tidak beralih ke layanan yang disediakan oleh pesaing. Di dalam bisnis ekonomi, konsep ini dikenal dengan sebutan Customer Relationship Management (CRM). CRM adalah strategi bisnis yang menargetkan untuk meyakinkan kepuasan pelanggan. Perusahaan yang berhasil menerapkan CRM kedalam bisnis mereka hampir selalu berhasil dalam peningkatan dalam menjaga pelanggan untuk tidak berpaling.

Industri telekomunikasi memproses dan menghasilkan jumlah data yang sangat besar dan cocok untuk penerapan data mining. Pada dasarnya, pertumbuhan pasar telekomunikasi sangat besar dan tumbuh secara eksponensial.

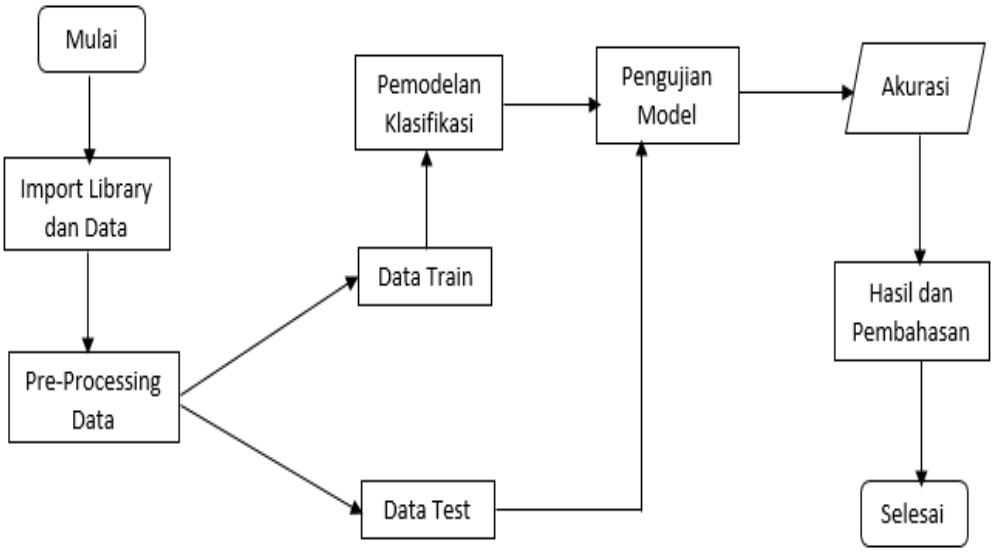
Churn di dalam industri telekomunikasi digunakan untuk menganalisis kemungkinan pelanggan untuk merubah layanan atau layanan provider selama jangka waktu tertentu. Secara sukarela maupun tidak sukarela (Mahajan, Misra, Mahajan, 2015). Faktor yang mempengaruhi perpindahan pelanggan bisa terjadi karena beberapa hal seperti pelanggan yang memutuskan untuk berhenti berlangganan, pelanggan yang terpaksa berhenti berlangganan bisa karena masalah finansial atau pindah lokasi yang tidak menyediakan layanan tersebut atau perusahaan memutuskan layanan pelanggan secara sepihak (Lazarov dan Capota, 2007). Mengurangi pengurangan pelanggan memiliki pengaruh yang signifikan tidak hanya untuk meningkatkan profit untuk komersial bank, tetapi juga meningkatkan inti kompetitif mereka (He, Shi, Wan dan Zhao, 2014).

Dengan berkembangnya teknologi dan ilmu pengetahuan, maka dibutuhkan teknik untuk mengatasi hal tersebut. Salah satu caranya ialah menggunakan algoritma machine learning. Algoritma logistic regression terbukti menghasilkan nilai performa yang tinggi untuk pengaplikasian prediksi churn, yaitu sebesar 95% tertinggi diantara algoritma lain (Petkovski, Aleksandar, Risteska, Biljana, Trivodaliev, 2016). Logistic regression berhasil di aplikasikan didalam perusahaan telekomunikasi untuk mendeteksi pelanggan yang berpeluang untuk meninggalkan perusahaan dan penemuan mayoritas faktor-faktor penting penyebab churn (Mandák dan Hančlová, 2019) [5] .Dengan banyak dan besar nya data yang ada, pendekatan menggunakan correlation-based feature selection menghasilkan hasil yang lebih efektif untuk bekerja dengan data yang lebih besar(Yu dan Liu, 2003) [6]. Penelitian Yu dan Liu menerapkan metode relevansi dan analisis redundansi untuk seleksi fitur yang menargetkan dataset untuk data yang besar

Diharapkan peneliti dapat mengetahui hasil performa akan penggunaan correlation-based feature selection dan algoritma logistic regression yang dipakai untuk memprediksi churn terhadap layanan telekomunikasi dan dapat berkontribusi untuk perkembangan akan penelitian selanjutnya.

2 METODOLOGI PENELITIAN

Proses pada penelitian ini untuk memprediksi mengenai *income* serta proses dalam mencari akurasi dari model-model yang digunakan. Dibawah ini adalah gambar dari tahapan yang akan digunakan.



Gambar 1: Metode Penelitian

Pada gambar 1 terdapat metode penelitian dimulai dari import library dan data lalu melakukan tahapan *pre-processing* yang meliputi, data yang telah melewati tahapan *pre-processing* akan dibagi menjadi dua bagian mencakup *data train* dan *data test*, kemudian dmelakukan pemodelan klasifikasi menggunakan *data train* yang selanjutnya ialah pengujian model untuk data yang telah melewati tahap pemodelan klasifikasi maupun *data test* untuk melihat kedua akurasi dari data tersebut.

2.1 Data

Diharapkan peneliti dapat mengetahui hasil performa akan penggunaan correlation-based feature selection dan algoritma logistic regression yang dipakai untuk memprediksi churn terhadap layanan telekomunikasi dan dapat berkontribusi untuk perkembangan akan penelitian selanjutnya.

Atribut yang ada pada dataset adalah *State,Account length, Area code,,International plan, Voice mail plan,Number vmail messages, Total day minutes, Total day calls, Total day charge, Total eve minutes, Total eve calls, Total eve charge, Total night minutes, Total night calls, Total night charge, Total intl minutes, Total intl calls, Total intl charge, Customer service calls*
Churn.

2.2 Pre-processing data

Pre-processing data meliputi *wrangling*, yaitu proses mengubah data yang berbentuk kategorik menjadi bentuk numerik.

Berikut data kategori yang diubah menjadi numerik:

- *State*
- *International plan*
- *Voice mail plan*

2.3 Split Data

Penentuan split data atau pembagian data yang dilakukan dengan rasio 80% data *training* dan 20% data *testing*.

2.4 Forward Selection

Menurut (Mark A.Hall, 1999) Forward selection merupakan sebuah correlation-based feature selector dimana forward selection dimulai dengan tanpa fitur dan menambahkan satu fitur dalam satu waktu sampai tidak adanya kemungkinan penambahan fitur tersebut menghasilkan nilai evaluasi yang lebih tinggi..

Metode ini menerapkan pencarian secara berurut dengan menambahkan fitur kedalam subset kosong dan menambahkan fitur satu persatu dengan menggunakan hasil evaluasi R2 dari setiap penambahan fitur tersebut. Subset tersebut dipilih dengan menambahkan satu fitur dalam satu waktu untuk setiap iterasi dengan menghasilkan hasil yang berbeda untuk setiap kali penambahan fitur (Marcano-Cedeño, 2010).

2.4.1. Sequential Forward Selection

Sebuah prosedur yang diaplikasikan setelah setiap satu tahapan forward selection dan menerapkan metode backward selama hasil dari subset lebih baik dari subset sebelumnya. (Somol P., 2010).

Meskipun pengaplikasian Sequential Forward Selection dikatakan menggunakan metode backward, metode ini tidak menggunakan metode backward sama sekali jika hasil tingkat

aktual dari dimensi yang sesuai tidak meningkat. Aturan ini berlaku juga untuk metode backward. Kedua algoritma tersebut menggunakan sebuah metode yang disebut "self-controlled backtracking" dimana mereka bisa mencari solusi yang baik dengan menyesuaikan kualitas tahap diantara forward dan backward secara dinamis.

Penerapan sequential forward selection memiliki prinsip yang sama dengan forward selection dengan penambahan fitur K. K berfungsi sebagai penentuan fitur yang akan dicari.

2.5. Logistic Regression

Logistic Regression mempelajari asosiasi diantara variabel yang bergantung dan sebuah set dari variabel yang tidak saling bergantung. Logistic Regression bersaing dengan analisis diskriminan sebagai sebuah metode analisis variabel respon kategoris. Banyak ahli statistic merasa bahwa Logistic Regression lebih fleksibel dan lebih cocok untuk sebagian besar situasi pemodelan. (NCSS, 2020).

Logistic regression digunakan saat variabel yang saling bergantung memiliki 2 nilai seperti 0 dan 1 atau Yes dan No maupun multinomial dan ordinal. Logistic regression mengambil input nilai asli dan membuat seperti sebuah prediksi yang akan menentukan tempat nilai tersebut, apakah 0 atau 1 yang menunjukan Churn atau Non-churn.

Menurut Hemlata, Ajay dan Sumit menyebutkan Logistic regression didapatkan dengan mengambil dari P_i/(1-P_i) dimana P adalah kemungkinan untuk churn dan non-churn. P selalu diantara nilai 0 sampai 1. Perhitungan matematika untuk pemodelan logistic regression adalah sebagai berikut

$$Z_i = In\left({}^{P_i}/_{1-P_i}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_n X_n \quad (1)$$

β merupakan koefisian untuk dipelajari dan X1 … Xn adalah variabel bergantung. Disini, churn adalah variabel bergantung dan fitur-fitur yang lain adalah variabel tidak bergantung

2.6. Evaluasi

Data *training* yang telah dibangun akan dilakukan tahap pengujian yang meliputi nilai akurasi dan nilai *F1 Score*. Nilai akurasi didapatkan dari prediksi benar untuk data positif dan negatif dari keseluruhan data. *F1 Score* adalah nilai yang menandakan jika model yang dibangun memiliki nilai presisi dan *recall* yang baik.

Data *training* yang telah dibangun akan dilakukan tahap pengujian yang meliputi nilai akurasi dan nilai *F1 Score*. Nilai akurasi didapatkan dari prediksi benar untuk data positif dan negatif dari keseluruhan data. *F1 Score* adalah nilai yang menandakan jika model yang dibangun memiliki nilai presisi dan *recall* yang baik. Untuk memperoleh nilai akurasi, nilai *precision*, nilai *recall*, serta *f1 score* dapat menggunakan *Confusion Matrix*.

		Nilai Aktual	
		Positive	Negative
Nilai Prediksi	Positive	TP (True Positive)	FP (False Positive)

Negative	FN (False Negative)	TN (True Negative)

2.6.1. *Accuracy*

Accuracy adalah rasio yang memiliki prediksi nilai benar (nilai positif dan nilai negatif) berdasarkan keseluruhan data. Akurasi dapat menggambarkan keakuratan model klasifikasi yang digunakan. Nilai akurasi dapat diperoleh menggunakan persamaan berikut ini

$$accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (2)$$

2.6.2. *Precision*

Precision adalah rasio yang memiliki prediksi nilai benar positif jika dibandingkan dengan keseluruhan hasil yang diprediksi positif. *Precision* dapat menggambarkan keakuratan data yang diinginkan dengan hasil prediksi yang diperoleh model klasifikasi. Nilai *Precision* dapat diperoleh menggunakan persamaan berikut ini

$$precision = \frac{TP}{TP+FP} \quad (3)$$

2.6.3. *Specitifity*

Specitifity merupakan kebenaran dalam memprediksi negatif dengan keseluruhan data yang positif. Nilai *Recall* dapat diperoleh menggunakan persamaan berikut ini

$$Specificity = TP/(TP + FP) \quad (4)$$

3 HASIL DAN PEMBAHASAN

Hasil eksperimen dengan menerapkan seleksi fitur pada dataset dengan menggunakan algoritma klasifikasi untuk mengetahui efektifitas reduksi data. Perbandingan korelasi untuk keseluruhan fitur meliputi :

No.	State	Account length	Area code	International plan	Voice mail plan	...	Churn
1.	State	1.000.000	0.003678	0.015814	-0.004597	...	0.007780
2.	Account length	0.003678	1.000.000	-0.012463	0.024735	...	0.016541
3.	Area code	0.015814	-0.012463	1.000.000	0.048551	...	0.006174
4.	International plan	-0.004597	0.024735	0.048551	1.000.000	...	0.259852
5.	Voice mail plan	-0.031664	0.002918	-0.000747	0.006006	...	-0.102148
...	...	...	...	...	...	...	...
3333.	Churn	0.007780	0.016541	0.006174	0.259852	...	1.000.000

Gambar 2: Nilai Korelasi Pearson

Yang selanjutnya akan dilakukan pemodelan berdasarkan 5 fitur terbaik menggunakan sequential forward selection yang dihasilkan seperti di table 2

Tabel 1. Score seleksi fitur

No	Variabel	Score
1.	Customer service calls	-0.16234406614124422
2.	International plan, Customer service calls	-0.13831843019637002

3.	International plan, Total intl calls, Customer service calls	-0.12029587177000296
4.	International plan, Total intl minutes, Total intl calls, Customer service calls	-0.09026194383542108
5.	State, International plan, Total intl minutes, Total intl calls, Customer service calls	-0.08725808753341176

Tabel diatas merupakan tabel iterasi yang dilakukan untuk mendapatkan 5 fitur terbaik setelah mengaplikasikan metode sequential feature selection. 5 fitur terpilih dengan kombinasi seperti di tabel 2 menunjukan bahwa kombinasi tersebut merupakan kombinasi terbaik dengan score akhir -0.08725808753341176

Selanjutnya dilakukan tahapan pemodelan. Pemodelan dengan menerapkan Logistic Regression pada keseluruhan data dengan seluruh variabel dan dengan variabel yang sudah diseleksi dengan Forward Selection sebelumnya.

Keseluruhan Variabel menghasilkan confusion matrix seperti :

Tabel 2 Confusion Matrix Seluruh Data

	False	True
False	566	4
True	93	4

Dari tabel diatas terdapat 93 data yang Churn dan terprediksi salah, sedangkan terdapat 4 data yang tidak churn tetapi diprediksi churn dan 4 data yang churn tetapi diprediksi tidak churn, model Regresi Logistik yang diterapkan lebih memprediksi data tersebut churn daripada tidak churn dengan perbandingan 566:4

Akurasi untuk keseluruhan data atau prediksi benar untuk data positif dan negatif ialah:

$$AC = \frac{4 + 566}{4 + 566 + 93 + 4} = 0{,}8545$$

Presisi atau rasio prediksi benar positif dibandingkan dengan keseluruhan hasil positif dari analisis menggunakan keseluruhan fitur ialah :

$$Precision = \frac{4}{4+4} = 0{,}5$$

Kebenaran algoritma dalam memprediksi negatif dengan keseluruhan data positif nya atau yang lebih dikenal dengan nilai specificity dari keseluruhan data ialah :

$$Specificity = \frac{566}{566+4} = 0{,}9929$$

Kemudian data dengan variabel terpilih akan dimasukkan kedalam model regresi logistik untuk memprediksi kemungkinan persentase akan prediksi yang diharapkan.

Tabel 3 Confusion Matrix Variabel Terpilih

	False	True
False	565	9
True	86	7

Dengan implementasi seleksi fitur didapatkan data seperti tabel diatas yang di artikulasikan menjadi 86 data yang Churn dan terprediksi salah, sedangkan terdapat 9 data yang tidak churn tetapi diprediksi churn dan 86 data yang churn tetapi diprediksi tidak churn, model Regresi Logistik yang diterapkan lebih memprediksi data tersebut churn daripada tidak churn dengan perbandingan 565:9.

Akurasi untuk keseluruhan data atau prediksi benar untuk data positif dan negatif yang diterapkan pada fitur pilihan ialah :

$$AC = \frac{7+565}{7+565+86+9} = 0{,}8575$$

Presisi atau rasio prediksi benar positif dibandingkan dengan keseluruhan hasil positif dari analisis menggunakan fitur terpilih ialah :

$$Precision = \frac{7}{9+7} = 0{,}4375$$

Kebenaran algoritma dalam memprediksi negatif dengan keseluruhan data positif nya atau yang lebih dikenal dengan nilai specificity dari data fitur terpilih ialah :

$$Specificity = \frac{565}{565+9} = 0{,}9843$$

Untuk perhitungan akurasi regresi logistic dengan dataset churn di industry telekomunikasi dengan menggunakan keseluruhan variabel sebesar 85,45% sedangkan hasil akurasi dengan menerapkan correlation based feature selection – forward selection menghasilkan nilai akurasi algoritma sebesar 85,75%.

Dengan klasifikasi report untuk kedua percobaan ialah:

Tabel 4 Laporan Klasifikasi

		Precision	Recall	F-1 Score
Seluruh Variabel	False	0.858877	0.992982	0.921074
	True	0.500000	0.041237	0.076190
	Akurasi		0.854573	
		Precision	Recall	F-1 Score
5 Variabel	False	0.867896	0.984321	0.922449
	True	0.437500	0.075269	0.128440
	Akurasi		0.857571	

Dari laporan diatas menghasilkan untuk nilai dari data yang diprediksi benar memiliki peningkatan dari 0.8545 menjadi 0.8575 dengan menerapkan Correlation Based Feature Selection-Forward Selection

4 KESIMPULAN

Kesimpulan yang diperoleh dari pembahasan ialah performa penerapan algoritma Logistic Regression untuk memprediksi Churn dengan Correlation Based Feature Selection-Forward Selection menghasilkan tingkat akurasi prediksi churn pelanggan telekomunikasi perusahaan Orange sebesar 85,75% terbukti lebih baik daripada dengan menerapkan Logistic Regression tanpa menerapkan Forward Selection dengan akurasi prediksi churn sebesar 85,45%. Kesimpulan lain ialah fitur yang sangat berpengaruh dalam prediksi churn khususnya untuk perusahaan telekomunikasi Orange ialah fitur State, International Plan, Total intl minute, Total intl call, Customer Service call karena fitur tersebut yang terseleksi menggunakan

Referensi

Hall, M. A. (1999). *Correlation-based Feature Selection for Machine Learning*.
Marcano-Cedeño, A., Quintanilla-Domínguez, J., Cortina-Januchs, M. G., & Andina, D. (2010). Feature selection using Sequential Forward Selection and classification applying Artificial Metaplasticity Neural Network.

He, B., Shi, Y., Wan, Q. and Zhao, X., 2014. Prediction of Customer Attrition of Commercial Banks Based on SVM Model. 2nd International Conference on Information Technology and Quantitative Management, ITQM,.

Lazarov, V., dan Capota, M. 2007. Churn Prediction

Mahajan, Vishal dan Misra, Richa dan Mahajan, Renuka. 2015. Review of Data Mining Techniques for Churn Prediction in Telecom. Journal of Information and Organizational Sciences. 39. 183-197

Mandák, J. and Hančlová, J., 2019. Use of Logistic Regression for Understanding and Prediction of Customer Churn in Telecommunications.

Petkovski, Aleksandar dan Risteska Stojkoska, Biljana dan Trivodaliev, Kire dan Kalajdziski, Slobodan. 2016. Analysis of churn prediction: A case study on telecommunication services in Macedonia. 1-4.

Santhalingam, Babu dan Ananthanarayanan, N.R.. (2018). Enhanced Prediction Model for Customer Churn in Telecommunication Using EMOTE. NCSS 2020 Statistical Software. 2020. NCSS, LLC.

Yu, L., & Liu, H. (2003). Feature Selection for High-Dimensional Data: A Fast Correlation-Based Filter Solution. Proceedings, Twentieth International Conference on Machine Learning, 2