Analisis Perbandingan Algoritma Machine Learning untuk Prediksi Potensi Hilangnya Nasabah Bank

Application of Machine Learning to Predict Potential Loss of Bank Customer

Mohammad Farid Naufal[1], Subrata[2], Alvin Fernando Susanto[3], Christian Nathaneil Kansil[4]
Solichul Huda[5]
[1,2,3,4] Program Studi Teknik Informatika, Universitas Surabaya, Surabaya, Jawa Timur
[5]Program Studi Teknik Informatika, Universitas Dian Nuswantoro
E-mail: [1]faridnaufal@staff.ubaya.ac.id, [2]s160420002@student.ubaya.ac.id,
[3]s160420013@student.ubaya.ac.id, [4]s160420069@student.ubaya.ac.id,
[5]solichul.huda@dsn.dinus.ac.id

Abstrak

Nasabah adalah salah satu aset paling berharga dari sebuah bisnis perbankan. Mereka adalah ujung tombak pengguna produk yang nantinya memberikan keuntungan bagi bank, terutama pada produk kartu kredit. Penelitian ini bertujuan untuk mengetahui nasabah mana sajakah yang berpotensi untuk meninggalkan layanan kartu kredit dari sebuah bank. Pada penelitian sebelumnya belum ada yang melakukan analisis perbandingan algoritma *machine learning* dengan berbagai macam tahapan *preprocessing* untuk memprediksi potensi hilangnya nasabah bank. Penelitian ini melakukan analisis perbandingan algoritma *machine learning* dengan kombinasi tahapan *preprocessing* untuk memprediksi potensi hilangnya nasabah bank. Analisis ini penting untuk pemilihan algoritma yang paling cocok untuk prediksi potensi hilangnya nasabah bank. Pada tahapan *preprocessing* diterapkan *dimensionality reduction* dan *feature selection* menggunakan metode *Variance threshold* dan *Correlation coefficient*. Metode klasifikasi yang digunakan adalah *Logistic regression* (LR), *Decision tree* (DT), dan *Naïve Bayes* (NB). Hasil tertinggi dari ketiga metode tersebut adalah *Decision tree* yang mampu memiliki nilai *F1 Score* sebesar 96% dan nilai akurasi mencapai 93%. *Logistic regression* dan *Naïve Bayes* berada pada urutan kedua dan ketiga setelah *decision tree*. Tahapan *data preprocessing* tidak memberikan pengaruh yang signifikan pada nilai *F1 Score* dan akurasi.

Kata kunci: Klasifikasi, Bank, Nasabah, Hilang

Abstract

Customers are one of the most valuable assets of a banking business. They are the spearhead of product users who will provide benefits for banks, especially in credit card products. This study aims to find out which customers have the potential to leave credit card services from a bank. In previous studies, no one has conducted a comparative analysis of machine learning algorithms with various preprocessing stages to predict the potential loss of bank customers. This study performs a comparative analysis of machine learning algorithms with a combination of preprocessing stages to predict the potential loss of bank customers. This analysis is important for selecting the most suitable algorithm for predicting potential loss of bank customers. At the preprocessing stage, dimensionality reduction and feature selection are applied using the Variance threshold and Correlation coefficient methods. The classification methods used are Logistic regression (LR), Decision tree (DT), and Naïve Bayes (NB) algorithms. The highest result of the three methods is the Decision Tree which is able to have an F1 score of 96% and an accuracy value of 93%. Logistic regression and Naïve Bayes are second and third after the decision tree. It was also found that the presence or absence of data preprocessing stages did not have a significant effect on the F1 score and accuracy.

Keywords: Classification, Banks, Customers, Loss

1. PENDAHULUAN

Berkembangnya zaman yang begitu cepat menjadikan manusia memiliki kebutuhan hidup yang semakin tinggi. Manusia juga membutuhkan kemudahan untuk memenuhi kebutuhan hidup tersebut. Bank hadir untuk memenuhi kedua kebutuhan tersebut. Melalui produk kartu kreditnya, bank mampu membantu setiap orang yang memenuhi persyaratan tertentu untuk melakukan pinjam bayar, yakni bank membayar terlebih dahulu biaya belanja nasabah dan nasabah dapat membayar total biaya yang dikeluarkan tersebut secara mencicil. Bisnis kartu kredit ini mendatangkan keuntungan bagi bank melalui bunga yang dikenakan kepada nasabah. Mayoritas atau hampir seluruh bank memberikan layanan kartu kepada nasabah yang memenuhi kriteria yang telah ditetapkan. Akan tetapi, hal ini memunculkan persaingan antara satu bank dengan bank lainnya. Persaingan tersebut menjadikan bank membutuhkan proses analisis terhadap nasabah yang berpotensi untuk meninggalkan layanan kartu kredit milik mereka dan berpindah ke layanan kartu kredit dari bank lainnya. Lancar atau tidaknya sebuah bisnis bank bergantung dari bagaimana bank tersebut mampu memberikan pelayanan terbaik pada setiap nasabah yang mereka miliki. Kehilangan nasabah menjadi permasalahan sangat serius bagi bank, karena nasabah adalah tombak utama mereka untuk menghasilkan keuntungan. Umumnya, salah satu penghasilan terbesar bank adalah melalui penawaran kartu kredit yang mampu mereka berikan kepada nasabah.

Beralihnya nasabah dari suatu bank ke bank lain umumnya dikenal dengan istilah *Churn*. *Customer churn* harus ditangani dengan segera agar menghindari dampak negatif untuk bank. Bagi bank, terjadinya *customer churn* layaknya kehilangan ujung tombak dari bisnis itu sendiri. Dengan adanya *Machine Learning*, maka bank dapat terbantu dalam menganalisis nasabah mana saja yang berpotensi untuk melakukan *churn*. Jenis *Machine Learning* yang sesuai untuk permasalahan tersebut adalah *metode klasifikasi* dengan *output* berupa label. Sementara itu, beberapa metode *classification* yang mampu menghasilkan akurasi dengan nilai tinggi pada dataset *customer churn* adalah *Logistic regression* (LR), *Decision tree* (DT), dan *Naïve Bayes* (NB). Semakin tinggi nilai akurasi dari metode-metode yang digunakan, maka semakin tinggi pula kemungkinan penelitian ini dapat menghasilkan *output* prediksi yang sesuai.

Berdasarkan penelitian terdahulu yang pernah dilakukan oleh Ahmad et al. [1] ditemukan bahwa permasalahan klasifikasi pada dataset sejenis menggunakan metode *Decision tree* pada perusahaan telecom dan menghasilkan akurasi 93,3%. Namun penelitinan ini tidak menyajikan perhitungan F1 Score. Namun demikian, hasil penelitian lain menemukan bahwa permasalahan hilangnya nasabah kartu kredit dengan metode *Decision tree* mampu mencapai nilai 90,50% dengan metode *Shuffled Sampling*. Hasil ini mampu mencapai nilai yang lebih baik ketika dilakukan *Feature selection* dengan metode backward Shuffled Sampling [2]. Gavril et al. [3] menggunakan data mining untuk memprediksi hilangnya nasabah pada dataset yang terdiri 3333 pelanggan dengan 21 fitur. Beberapa fitur meliputi informasi tentang jumlah pesan masuk dan keluar serta pesan suara untuk setiap pelanggan. Penelitian ini menerapkan *Principal Component Analysis* (PCA) untuk mengurangi dimensi data. Tiga algoritma machine learning yang digunakan adalah Neural Networks, Support Vector Machine, dan Naïve bayes. Nilai AUC masing-masing adalah 99,10%, 99,55% dan 99,70% untuk Naïve bayes, Neural network dan Support Vector Machine. Namun dataset yang digunakan dalam penelitian ini kecil dan tidak terdapat *missing values*. Penelitian ini hanya menggunakan *dimensionality reduction* dan tidak menambahkan metode *feature selection*. Dia et al. [4] mengusulkan model klasifikasi menggunakan Neural Network untuk memprediksi potensi hilangnya nasabah di sebuah perusahaan telekomunikasi besar China yang memiliki sekitar 5,23 juta pelanggan. Akurasi yang dihasilkan mencapai 91,1%. Idris et al. [5] menggunakan *genetic programming* dengan *AdaBoost* untuk memprediksi potensi hilangnya nasabah di bidang telekomunikasi. Model diuji pada dua dataset yaitu Orange Telecom dan cell2cell. Akurasi yang dihasilkan adalah 63% untuk dataset Orange Telecom dan 89% untuk cell2cell. Makhtar et al. [6] mengusulkan algoritma *rough set theory* di sebuah perusahaan telekomunikasi untuk memprediksi hilangnya pelanggan. Seperti disebutkan dalam makalah ini algoritma klasifikasi *rough set* mengungguli algoritma lain seperti *Linear Regression, Decision*

Tree, dan Voted Perception Neural Network. Berbagai penelitian mempelajari masalah *unbalanced dataset* dimana distribusi data kelas pelanggan yang hilang lebih sedikit dibandingkan jumlah pelanggan yang tidak hilang. Amin et al. [7] membandingkan enam teknik pengambilan sampel yang berbeda untuk *oversampling* mengenai masalah prediksi churn telekomunikasi. Hasil penelitian menunjukkan bahwa *rules-generation based on genetic algorithms* lebih baik jika dibandingkan dengan metode oversampling yang lain. Burez et al. [8] membandingkan teknik oversampling untuk model prediksi potensi hilangnya nasabah menggunakan *Random Sampling, Advanced Under-Sampling, Gradient Boosting Model*, dan *Weighted Random Forests* [9]. Metode evaluasi yang digunakan adalah AUC. Hasilnya menunjukkan bahwa teknik *undersampling* mengungguli teknik lain yang diuji. Aksama et al. [10] melakukan prediksi potensi hilangnya nasabah menggunakan Naïve Bayes dan Decision Tree. Akurasi yang dihasilkan oleh Naïve Bayes adalah 85,17% sedangkan oleh Decision Tree adalah 79,17%. Tahapan feature selection pada penelitian ini dilakukan dengan dengan melihat secara manual atribut mana yang tidak berpengaruh terhadap output. Hanif et al. [11] menggunakan algoritma *Logistic regression* dan menghasilkan akurasi tertinggi 55,66%. Akan tetapi, nilai tersebut meningkat signifikan menjadi 85,53% apabila digabungkan dengan menambahkan metode untuk menyeimbangkan dataset yaitu *underbagging 15*.

Dari beberapa penelitian sebelumnya belum ada yang melakukan perbandingan performa algoritma *machine learning* secara mendetail dengan mengkombinasikan beberapa tahapan *feature selection* dan *preprocessing* untuk memprediksi potensi hilangnya nasabah bank. Penelitian ini menyajikan hasil performa beberapa algoritma *machine learning* dengan dan tanpa adanya tahapan *preprocessing* untuk mengklasifikasikan potensi hilangnya nasabah bank. Mengacu pada permasalahan tersebut, penelitian ini menggunakan dataset *churn* dari kaggle [12] yang selanjutnya akan dianalisis untuk menghasilkan *output* berupa hilang atau tidaknya nasabah pada bank. Hasil dari analisis ini akan membantu pihak bank untuk secara proaktif memberikan pelayanan yang semakin baik serta membantu mereka untuk mengambil keputusan paling tepat.

2. METODE PENELITIAN

Penelitian ini menggunakan beberapa tahapan dimulai dari menentukan dataset, melakukan reduksi dimensi, *preprocessing* dataset, *training* dan *testing*, membuat model klasifikasi, dan evaluasi model.

a. Pemilihan Dataset

Dataset yang digunakan memiliki 19 *features* dan 1 target dari dataset *churn* kaggle [12]. *Feature* ini akan digunakan untuk memprediksi apakah nasabah meninggalkan layanan kartu kreditnya atau tidak. Berikut jenis atribut dan tipe data beserta keterangannya pada Tabel 1.

b. Feature Selection

Feature selection adalah pengurangan jumlah atribut dalam suatu dataset berdasarkan pertimbangan tertentu guna meningkatkan kinerja algoritma klasifikasi. Algoritma reduksi dimensi yang digunakan dalam penelitian ini yaitu *feature selection variance threshold* dan *correlation coefficient.* Dalam *variance threshold* akan dipilih atribut yang memiliki varians lebih tinggi daripada nilai *threshold*. Jika nilai atribut tidak lebih tinggi dari *threshold* yang ditentukan maka atribut tersebut dihapus dari dataset. Persamaan (1) menunjukkan formula untuk menentukan variance. x_i adalah nilai asli data ke-I, $\bar{x}$ adalah nilai rata-rata, dan n adalah banyak data.

$$Var(x) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \, , \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{1}$$

Tabel 1. Keterangan Atribut

Atribut	**Tipe Data**	**Keterangan**
Attrition_Flag	Label	Apakah nasabah tutup akun
Customer_Age	Angka	Usia nasabah (Tahun)
Gender	Label	Jenis kelamin
Dependent_count	Angka	Jumlah tanggungan
Education_Level	Label	Kualifikasi pendidikan
Martial_Status	Label	Status pernikahan
Income_Category	Label	Kategori pendapatan
Card_Category	Label	Jenis kartu kredit
Months_on_book	Angka	Jangka waktu hubungan dengan bank (Bulan)
Total_Relationship_Count	Angka	Banyak produk yang dimiliki nasabah
Months_Inactive_12_mon	Angka	Jumlah bulan nasabah tidak aktif (12 bulan terakhir)
Contacts_Count_12_mon	Angka	Jumlah kontak (12 bulan terakhir)
Credit_Limit	Angka	Jumlah limit pada kartu kredit
Total_Revolving_Bal	Angka	Total saldo bergulir pada kartu kredit
Avg_Open_To_Buy	Angka	Rata-rata transaksi kartu kredit
Total_Amt_Chng_Q4_Q1	Angka	Perubahan jumlah nominal transaksi Q4 sampai Q1
Total_Trans_Amt	Angka	Jumlah nominal transaksi (12 bulan terakhir)
Total_Trans_Ct	Angka	Jumlah transaksi (12 bulan terakhir)
Total_Ct_Chng_Q4_Q1	Angka	Perubahan jumlah transaksi Q4 sampai Q1
Avg_Utilization_Ratio	Angka	Rasio penggunaan kartu rata-rata

c. Data Preprocessing

Preprocessing adalah sekumpulan teknik yang dilakukan terhadap dataset agar data tidak mengandung *noise, missing value* dan *error* lainnya sehingga data siap untuk diproses lebih lanjut. Adapun teknik *preprocessing* yang digunakan adalah scaling (*standard scaler, minmax scaling*, dan *normalization*).

Correlation coefficient adalah algoritma untuk menyeleksi *feature* yang digunakan dengan menghitung korelasi antar atribut. Persamaan (2) menunjukkan formula untuk mencari I*correlation* antar fitur. x_i adalah nilai asli data ke-*i*, $\bar{x}$ adalah Nilai rata-rata, n adalah banyak data.

$$corr(x_1, x_2) = \frac{\sum_{i=1}^{n}(x_{1i} - \bar{x}_1)(x_{2i} - \bar{x}_2)}{\sqrt{\sum_{i=1}^{n}(x_{1i} - \bar{x}_1)^2}\sqrt{\sum_{i=1}^{n}(x_{2i} - \bar{x}_2)^2}} \quad (2)$$

Standard Scaler adalah proses membuat skala ulang pada data *feature train* dengan rata-rata 0 dan varians 1, kemudian di transformasikan ke *feature testing*. Persamaan (3) menunjukkan formula untuk *standardization*. x_i' adalah hasil standarisasi data ke-*i*, x_i adalah nilai asli data ke *i*, μ adalah nilai rata-rata, σ adalah standar deviasi.

$$x_i' = \frac{x_i - \mu x}{\sigma x} \quad (3)$$

Min Max Scaling adalah proses mengubah skala nilai data dari rentang nilai sesungguhnya menjadi dalam rentang 0 hingga 1 atau -1 hingga 1. Persamaan (4) menunjukkan formula untuk scaling. x_i' adalah hasil *scaling* data ke-*i*, x_i adalah nilai asli data ke-*i*, a adalah batas skala minimal, b adalah batas skala maksimal, $\min(X)$ adalah nilai minimal dari *X*, $\max(X)$ adalah nilai maksimal dari *X*.

$$x_i' = a + \frac{(x_i - min(X))(b - a)}{max(X) - min(X)} \quad (4)$$

Normalizing adalah proses mengubah nilai data dalam skala yang sama. *Normalizing* yang digunakan adalah Manhattan atau L1 dengan rumus sebagai seperti di persamaan (5). Sedangkan untuk perhitungan normalisasi Manhattan ada di persamaan (6). x_i' adalah Hasil standarisasi data ke-i, x adalah nilai asli data x, $\|x\|$ adalah hasil manhattan atau L1 norm, $|x_1|$ adalah hasil mutlak data x ke-1, $|x_2|$ adalah hasil mutlak data x ke-2, $|x_n|$ adalah hasil mutlak data x ke-n

$$x' = \frac{1}{\|x\|} x \tag{5}$$

$$\|x\| = |x_1| + |x_2| + \cdots + |x_n| \tag{6}$$

d. Training dan Testing

Training merupakan tahapan untuk melatih model *machine learning* dengan dataset yang diberikan, setelah mesin selesai dilatih maka dilakukan pengujian terhadap mesin atau biasa disebut *testing.* Sebelum melakukan *training* dan *testing* kami membagi dataset dengan proporsi 75% digunakan untuk *training* dan 25% untuk *testing*.

e. Model Klasifikasi

Logistic regression adalah sebuah metode klasifikasi yang digunakan untuk memprediksi probabilitas hubungan antar variabel dependen dan independent [13]. Dalam kasus ini digunakan untuk memprediksi kemungkinan nasabah untuk meninggalkan layanan kartu kredit pada suatu bank. Formula dari *logistic regression* dapat dilihat pada persamaan (7). ln adalah logaritma natural, p adalah probabilitas logistik

$$ln\left(\frac{p}{1-p}\right) = \boldsymbol{w}^T \boldsymbol{X}, p = \frac{1}{1+e^{-\mathbf{w}^T\mathbf{x}}} \tag{7}$$

Decision tree adalah algoritma klasifikasi yang menggunakan keputusan berdasarkan struktur yang dimodelkan seperti pohon [14]. Untuk memilih atribut sebagai akar berdasarkan nilai gain tertinggi dari atribut yang digunakan. Persamaan (8) menunjukkan formula dari *decision tree*. S = Himpunan kasus, A adalah atribut, n adalah Jumlah partisi atribut A, |Si| adalah proporsi Si terhadap S, $|S|$ adalah Jumlah kasus dalam S

$$Gain(S, A) = Entropy(S) - \sum_{i=1}^{n} \frac{|Si|}{|S|} * Entropy\ (Si) \tag{8}$$

Naïve Bayes adalah metode klasifikasi yang menggunakan nilai dari probabilitas untuk memprediksi peluang yang akan terjadi di masa mendatang berdasarkan data yang ada sebelumnya [15]. Formula (9) menunjukkan perhitungan *naïve bayes* [16]. $P(A|B)$ adalah Probabilitas A terjadi dengan syarat B telah terjadi (probabilitas superior), $P(B|A)$ adalah Probabilitas B terjadi dengan syarat A telah terjadi, $P(A)$ adalah peluang terjadinya A, $P(B)$ adalah peluang terjadinya B.

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \tag{9}$$

f. Evaluasi Model

Akurasi merupakan metode pengujian untuk melihat kesesuaian nilai prediksi dengan nilai aktual. *Precision* adalah metode pengujian dengan membandingkan jumlah total data positif yang diklasifikasikan benar dengan jumlah total data positif yang diprediksi. Semakin tinggi *precision* maka semakin banyak contoh data berlabel positif yang benar (FP rendah). *Recall* adalah metode pengujian dengan membandingkan jumlah total data positif yang diklasifikasikan benar dengan jumlah total data positif. *F1 Score* adalah bobot rata-rata dari perbandingan *precision* dan *precision.* Persamaan (10-11) masing-masing menunjukkan perhitungan akurasi dan *F1 Score*.

$$Akurasi = \frac{TP + TN}{TP + FN + FP + FN} * 100\% \quad (10)$$

$$F1\ score = \frac{2 * Recall * Precision}{Recall + Precision} \quad (11)$$

g. Metode Perbandingan Performa Algoritma

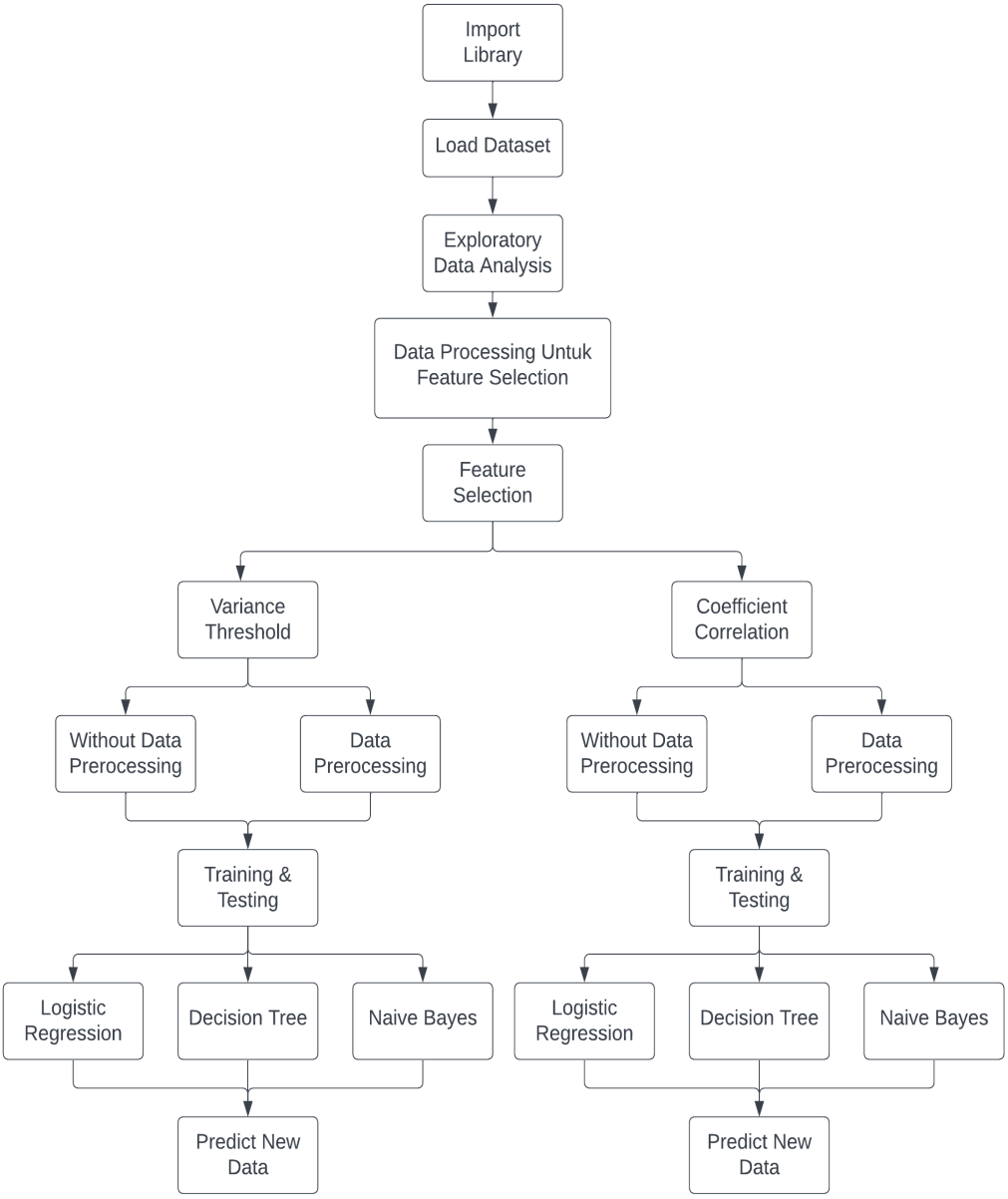
Perbandingan algoritma machine learning yang dilakukan adalah berdasarkan nilai metrik Akurasi dan *F1 Score*. Skenario uji coba yang dilakukan adalah menggunakan dan tanpa adanya *preprocessing*. Metode *preprocessing* yang digunakan adalah *Standarization (St), Scaling (Sc),* dan *Normalizing (N)*. Setiap model diuji coba dengan kombinasi metode preprocessing. Terdapat 15 kombinasi *preprocessing* + 1 tanpa *preprocessing*. Detail dari kombinasi *preprocessing* dijelaskan pada bab hasil dan pembahasan.

3. HASIL DAN PEMBAHASAN

Penelitian ini menggunakan 2 skenario *feature selection* terhadap dataset untuk memilih beberapa fitur tertentu yang memang memiliki keterkaitan dengan permasalahan yang diuji-cobakan. Skenario pertama menggunakan *variance threshold*, sedangkan skenario kedua menggunakan *correlation coefficient*. Pada masing-masing skenario, penelitian ini melakukan berbagai kemungkinan kombinasi metode *data preprocessing*, seperti *Standarization* (St), *Scaling* (Sc), dan *Normalizing* (N). Hasil dari masing-masing kombinasi *Data preprocessing* tersebut selanjutnya diuji menggunakan algoritma *Logistic regression* (LR), *Decision tree* (DT), dan *Naïve Bayes* (NB). Hasil dari percobaan tersebut berupa *F1 Score* dan nilai akurasi. Akan dilakukan pula uji coba *predict new data* dengan menggunakan kolom sesuai dengan hasil *Variance threshold* dan *Correlation coefficient*. Penelitian ini tidak menggunakan deteksi *outlier* dengan alasan algoritma LR dan NB baik dalam mengatasi *outlier*. Akan tetapi, algoritma DT sebenarnya tidak sebaik itu dalam mengatasi *outlier*. Namun adanya *Data preprocessing* membantu DT untuk tetap dapat memiliki performa yang baik walaupun masih terdapat *outlier* pada kumpulan data. Skema dari percobaan ini dapat dilihat pada gambar diagram berikut.

a. Skenario Uji Coba menggunakan *Feature selection Variance threshold*

Berdasarkan hasil uji coba algoritma *Feature selection* dengan *Variance threshold* terhadap 19 fitur dalam dataset yang digunakan, didapatkan bahwa algoritma ini menyederhanakan fitur-fitur tersebut menjadi 14 fitur. Penelitian ini membagi dataset menjadi dataset training dan dataset testing dengan proposi 25% data testing dan 75% data training dari keseluruhan total data. Pemilihan data tersebut diacak dengan inisialisasi random_state 42 pada *library* sklearn. Selanjutnya, dihitung varians dari threshold dengan nilai 0.5.



Gambar 1. Diagram skenario iji coba penelitian

Selanjutnya, data testing dan data training yang sudah dihasilkan dari *splitting* dataset dan *feature selection* akan dilakukan *data preprocessing*. *Data preprocessing* ditujukkan untuk memperbaiki kesalahan maupun kesenjangan pada *raw data* yang kerap tidak lengkap atau tidak menggunakan format yang tidak teratur. Pada langkah ini, dilakukan berbagai kombinasi algoritma *Data preprocessing* guna memastikan bahwa semua kemungkinan telah dicoba pada data training untuk membentuk hasil terbaik.

Berdasarkan berbagai hasil percobaan yang dilakukan dan hasil yang disajikan pada Tabel 3, ditemukan bahwa algoritma LR, DT, dan NB mampu menghasilkan nilai *F1 Score* dan akurasi yang tinggi dengan data hasil *preprocessing*. Akan tetapi, hasil ini tidak jauh berbeda dengan data tanpa *preprocessing*. Ketiga hasil ini membuktikan bahwa ketiga algoritma *classification* tersebut dapat menghasilkan performa yang baik meskipun terdapat data *outlier* dan perbedaan besaran data pada dataset.

Tabel 3. Hasil *F1 Score* dan akurasi skenario 1 dengan 3 macam kombinasi *data preprocessing*

		LR	DT	NB
Tanpa Preprocessing	***F1 Score***	0.93	**0.95**	**0.93**
	Akurasi	0.88	**0.92**	**0.88**
St	***F1 Score***	**0.93**	**0.95**	0.92
	Akurasi	**0.89**	**0.92**	0.87
St-Sc	***F1 Score***	0.93	**0.95**	0.92
	Akurasi	0.88	**0.92**	0.87
St-N	***F1 Score***	0.93	0.94	0.92
	Akurasi	0.87	0.90	0.87
St-Sc-N	***F1 Score***	0.92	0.91	0.92
	Akurasi	0.86	0.83	0.87
St-N-Sc	***F1 Score***	0.93	0.91	0.92
	Akurasi	0.87	0.83	0.87
Sc	***F1 Score***	0.93	**0.95**	0.92
	Akurasi	0.88	**0.92**	0.87
Sc-St	***F1 Score***	0.93	**0.95**	0.92
	Akurasi	0.89	**0.92**	0.87
Sc-N	***F1 Score***	0.92	0.95	0.92
	Akurasi	0.86	0.91	0.87
Sc-St-N	***F1 Score***	0.93	0.91	0.92
	Akurasi	0.87	0.83	0.87
Sc-N-St	***F1 Score***	**0.93**	0.79	0.92
	Akurasi	**0.89**	0.69	0.87
N	***F1 Score***	0.91	0.92	0.89
	Akurasi	0.83	0.87	0.91
N-St	***F1 Score***	0.93	0.93	0.89
	Akurasi	0.87	0.88	0.81
N-Sc	***F1 Score***	0.92	0.93	0.89
	Akurasi	0.87	0.88	0.81
N-St-Sc	***F1 Score***	0.92	0.91	0.89
	Akurasi	0.87	0.83	0.81
N-Sc-St	***F1 Score***	0.93	0.93	0.89
	Akurasi	0.87	0.88	0.81

Dengan nilai *F1 Score* dan akurasi yang tinggi dari ketiga algoritma, maka ketiga algoritma tersebut mampu memberikan hasil klasifikasi yang akurat terdahap data baru yang diujikan. Data *input* yang kami gunakan disesuaikan dengan Tabel 1 dataset yang selanjutnya digunakan untuk memvalidasi hasil. Nilai fitur dari data input yang urut sesuai dengan Tabel 1 adalah 46, 4, 4, 2, 5, 2374.0, 36, 5, 2, 1, 1332, 1042.0, 4253, dan 81. Hasil dari percobaan tersebut dapat dilihat pada Gambar 2. Hasil ini telah sesuai dengan data sumber yang digunakan, dimana serangkaian data tersebut benar memiliki *output* berupa *Existing customer*

```
Decision Tree Result:  ['Existing Customer']
Logistic Regression Result:  ['Existing Customer']
Gaussian Naive Bayes Result:  ['Existing Customer']
```

Gambar 2. Hasil Prediksi Data Baru dengan Ketiga Algoritma *Classification*

b. Skenario Uji Coba menggunakan *Feature selection Correlation coefficient*

Berdasarkan hasil uji coba *Feature selection* dengan metode *Correlation coefficient* pada ke-19 atribut dalam dataset yang digunakan, ditemukan bahwa hanya terdapat 1 kolom yang hilang, yaitu fitur *Avg Open to Buy*. Skenario ini dilakukan dengan cara menemukan terlebih dahulu korelasi antar variabel/atribut yang terdapat dalam dataset. Kemudian, dilakukan

pencarian atribut dengan korelasi di atas 90%. Selanjutnya, dilakukan perulangan untuk memperoleh kolom apa saja yang memiliki korelasi yang meniliki nilai di atas threshold. Dataset yang sudah diseleksi fiturnya d di-*splitting* menjadi data training dan testing dengan ukuran 25% data testing dan 75% data training seperti yang sudah dijelaskan pada bagian sebelumnya. Hasil dari dataset yang sudah diseleksi juga akan dilakukan *data preprocessing*, serta dilakukan evaluasi hasil dengan ketiga algoritma *classification*.

Tabel 4. Hasil *F1 Score* dan akurasi skenario 2 dengan 3 macam kombinasi *data preprocessing*

		LR	DT	NB
Tanpa Preprocessing	***F1 Score***	0.92	**0.96**	**0.93**
	Akurasi	0.86	**0.93**	**0.88**
St	***F1 Score***	**0.94**	**0.96**	0.92
	Akurasi	**0.90**	**0.93**	0.87
St-Sc	***F1 Score***	0.94	**0.96**	0.92
	Akurasi	0.89	**0.93**	0.87
St-N	***F1 Score***	0.93	0.95	0.92
	Akurasi	0.88	0.91	0.87
St-Sc-N	***F1 Score***	0.92	0.95	0.92
	Akurasi	0.86	0.92	0.87
St-N-Sc	***F1 Score***	0.93	0.95	0.92
	Akurasi	0.88	0.92	0.87
Sc	***F1 Score***	0.94	**0.96**	0.92
	Akurasi	0.89	**0.93**	0.87
Sc-St	***F1 Score***	0.94	**0.96**	0.92
	Akurasi	0.90	**0.93**	0.87
Sc-N	***F1 Score***	0.91	0.95	0.92
	Akurasi	0.83	0.92	0.87
Sc-St-N	***F1 Score***	0.93	0.95	0.92
	Akurasi	0.88	0.91	0.87
Sc-N-St	***F1 Score***	**0.94**	0.95	0.92
	Akurasi	**0.90**	0.92	0.87
N	***F1 Score***	0.91	0.93	0.92
	Akurasi	0.83	0.88	0.87
N-St	***F1 Score***	0.93	0.92	0.88
	Akurasi	0.89	0.87	0.81
N-Sc	***F1 Score***	0.91	0.92	0.88
	Akurasi	0.83	0.87	0.81
N-St-Sc	***F1 Score***	0.93	0.93	0.88
	Akurasi	0.88	0.88	0.81
N-Sc-St	***F1 Score***	0.93	0.92	0.88
	Akurasi	0.89	0.87	0.81

Layaknya hasil uji coba pada bagian sebelumnya, algoritma *classification* yang digunakan dapat menghasilkan nilai *F1 score* dan akurasi yang tinggi pada dataset hasil *preprocessing*. Namun, algoritma LR menghasilkan nilai yang lebih rendah apabila memproses data tanpa dilakukan *preprocessing* terlebih dahulu meski dengan hasil yang masih terbilang tinggi.

Dengan nilai *F1 Score* dan akurasi yang tinggi dari ketiga algoritma, maka ketiga algoritma tersebut mampu memberikan hasil prediksi yang akurat terdahap data baru yang diujikan. Data *input* yang kami gunakan disesuaikan dengan tabel dataset yang selanjutnya digunakan untuk memvalidasi hasil. Data *input* tersebut adalah 45, 1, 3, 4, 2, 3, 1, 12691.0, 39, 5, 1, 3, 777, 1.335, 1144, 1.625, 42, dan 0.061. Hasil dari percobaan tersebut dapat dilihat pada

Gambar 3. Hasil ini telah sesuai dengan data sumber yang digunakan, dimana serangkaian data tersebut benar memiliki *output* berupa *Existing customer*.

```
Decision Tree Result:  ['Existing Customer']
Logistic Regression Result:  ['Existing Customer']
Gaussian Naive Bayes Result:  ['Existing Customer']
```

Gambar 3. Hasil Prediksi Data Baru dengan Ketiga Algoritma *Classification*

Apabila hasil dari *F1 Score* dan akurasi dari masing-masing algoritma dibandingkan dari hasil *feature selection* yang berbeda, maka algortima LR dan DT memiliki performa yang lebih baik pada dataset yang fiturnya diseleksi menggunakan *correlation coefficient*. Sementara itu, algoritma NB stabil pada kedua hasil reduksi dataset. Jika dilihat lebih detail lagi, algoritma NB memiliki performa stabil pada mayoritas hasil *data preprocessing* di kedua hasil reduksi dataset yang digunakan.

4. KESIMPULAN DAN SARAN

Perbedaan metode *feature selection* dengan menggunakan *variance threshold* dan *Correlation coefficient* pada untuk mengklasifikasikan potensi hilangnya nasabah bank tidak memiliki pengaruh yang signifikan terhadap performa dari ketiga algoritma *classification* yang digunakan. Algoritma *Logistic regression* (LR) mampu menghasilkan performa yang sedikit lebih baik dengan menggunakan data hasil *preprocessing*. Sementara itu, algoritma *Decision tree* (DT) dan *Naïve Bayes* (NB) mampu menghasilkan performa yang lebih baik pada data tanpa *preprocessing*. Algoritma *Naïve Bayes* juga memiliki performa yang stabil pada berbagai macam jenis data (baik dilakukan *data preprocessing* maupun tidak). Dari ketiga algoritma tersebut, hasil *F1 Score* dan akurasi yang paling tinggi ditunjukkan oleh algoritma *Decision tree* dengan tahapan *feature selection correlation coefficient* dengan 0.96 dan 0.93.

Dataset yang digunakan tidak dilakukan *detecting outlier* karena algoritma LR dan NB memiliki kemampuan baik dalam mengatasi data *outlier*. Namun demikian, algoritma DT juga terbukti masih memiliki performa yang stabil walaupun memiliki sifat asli yang rentan terhadap data *outlier*. Hasil prediksi data baru memvalidasi bahwa ketiga algoritma mampu memprediksi data dengan performa baik.

DAFTAR PUSTAKA

[1] A. K. Ahmad, A. Jafar, and K. Aljoumaa, "Customer churn prediction in telecom using machine learning in big data platform," *J. Big Data*, vol. 6, no. 1, pp. 1–24, Dec. 2019, doi: 10.1186/S40537-019-0191-6/TABLES/4.

[2] H. ANNUR, "PREDIKSI BERHENTINYA NASABAH KARTU KREDIT MENGGUNAKAN ALGORITMA DECISION TREE BERBASIS BACKWARD ELIMINATION".

[3] I. Brandusoiu, G. Toderean, and H. Beleiu, "Methods for churn prediction in the pre-paid mobile telecommunications industry," *IEEE Int. Conf. Commun.*, vol. 2016-August, pp. 97–100, Aug. 2016, doi: 10.1109/ICCOMM.2016.7528311.

[4] Y. He, Z. He, and D. Zhang, "A study on prediction of customer churn in fixed communication network based on data mining," *6th Int. Conf. Fuzzy Syst. Knowl. Discov. FSKD 2009*, vol. 1, pp. 92–94, 2009, doi: 10.1109/FSKD.2009.767.

[5] A. Idris, A. Khan, and Y. S. Lee, "Genetic programming and Adaboosting based churn prediction for telecom," *Conf. Proc. - IEEE Int. Conf. Syst. Man Cybern.*, pp. 1328–1332, 2012, doi: 10.1109/ICSMC.2012.6377917.

[6] M. Makhtar, S. Nafis, M. A. Mohamed, M. K. Awang, M. N. A. Rahman, and M. M. Deris, "Churn classification model for local telecommunication company based on rough set theory," *J. Fundam. Appl. Sci.*, vol. 9, no. 6S, p. 854, Feb. 2018, doi:

10.4314/JFAS.V9I6S.64.

[7] A. Amin *et al.*, "Customer churn prediction in the telecommunication sector using a rough set approach," *Neurocomputing*, vol. 237, pp. 242–254, May 2017, doi: 10.1016/J.NEUCOM.2016.12.009.

[8] J. Burez and D. Van den Poel, "Handling class imbalance in customer churn prediction," *Expert Syst. Appl.*, vol. 36, no. 3, pp. 4626–4636, Apr. 2009, doi: 10.1016/J.ESWA.2008.05.027.

[9] V. Jain, J. Sharma, K. Singhal, and A. Phophalia, "Exponentially Weighted Random Forest," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 11941 LNCS, pp. 170–178, 2019, doi: 10.1007/978-3-030-34869-4_19.

[10] M. C. Aksama and A. Wahyuniati, "Prediksi Churn Nasabah Bank Menggunakan Klasifikasi Naïve Bayes dan ID3".

[11] T. T. Hanif, A. Adiwijaya, and S. Al-Faraby, "Analisis Churn Prediction Pada Data Pelanggan Pt. Telekomunikasi Menggunakan Underbagging Dan Logistic Regression," *eProceedings Eng.*, vol. 4, no. 2, 2017.

[12] A. Chaunan, "Credit Card customers | Kaggle," *kaggle*, 2021. https://www.kaggle.com/datasets/whenamancodes/credit-card-customers-prediction (accessed Dec. 23, 2022).

[13] C. Y. J. Peng, K. L. Lee, and G. M. Ingersoll, "An introduction to logistic regression analysis and reporting," *J. Educ. Res.*, vol. 96, no. 1, pp. 3–14, 2002, doi: 10.1080/00220670209598786.

[14] L. Rokach and O. Maimon, "Decision Trees," *Data Min. Knowl. Discov. Handb.*, pp. 165–192, May 2006, doi: 10.1007/0-387-25465-X_9.

[15] D. A. Setiawan, R. Halilintar, and L. S. Wahyuniar, "Penerapan Metode Naive Bayes Untuk Klasifikasi Penentuan Penerima Bantuan PKH," in *Prosiding SEMNAS INOTEK (Seminar Nasional Inovasi Teknologi)*, 2021, vol. 5, no. 2, pp. 249–254.

[16] Vikramkumar, V. B, and Trilochan, "Bayes and Naive Bayes Classifier," Apr. 2014, doi: 10.48550/arxiv.1404.0933.