P-ISSN : 2549-7480
E-ISSN : 2549-6301

JURNAL ILMIAH INFORMATIKA
E-ISSN : 2549-6301
ISSN : 2549-7480
Faculty of Sains and Technology, Ibrahimy University
S4-Accredited : 85/M/KPT/2020
Journal page is available to : https://journal.ibrahimy.ac.id/index.php/JIMI

PENGARUH JUMLAH *RECORD DATASET* TERHADAP ALGORITMA KLASIFIKASI BERDASARKAN DATA *CUSTOMER CHURN*

Anita[1)], Anggito Wicaksono[2)], Tesa Nur Padilah[3)]

[1, 2, 3] *Program Studi Teknik Infomatika, Universitas Singaperbangsa Karawang*

email: [1] *anita.anita17008@student.unsika.ac.id,* [2] *anggito.wicaksono17057@student.unsika.ac.id,* [3] *tesa.nurpadilah@staff.unsika.ac.id*

ARTICLE INFO

Article History:
Recieved : 20 April 2021
Revised : 11 June 2021
Accepted : 28 June 2021
Published : 30 June 2021

Keywords:
Dataset
Customer Churn
Classification
Data Mining
CRISP-DM

IEEE style in citing this article:
Anita, A. Wicaksono, and T. N. Padilah, "Pengaruh Jumlah Record Dataset Terhadap Algoritma Klasifikasi Berdasarkan Data Customer Churn", Jurnal.ilmiah.informatika, vol. 6, no. 1, pp. 1-10, Jun. 2021.

ABSTRACT

Telecommunication is one of the fastest growing industrial sectors so that there are more telecommunication companies. This can create various threats if the company does not use the strategy properly. Customer churn refers to the level of customer reduction which is one of the threats to reducing the company's revenue. This is an important issue for developing companies to evaluate in order to reduce the potential for churn that occurs. The initial stage that needs to be done is to predict customers who have the potential to switch from the company, one of which is the data mining approach. Classification is a data mining technique that can predict the class of datasets with various existing classification algorithms. The purpose of this study is to identify the effect of the number of dataset records on several classification algorithms. This research was conducted based on the CRISP-DM method by applying three classification algorithms, namely Logistic Regression, Naïve Bayes, and Decision Tree C4.5. The results showed that the greater the number of records in the dataset, the higher the accuracy value will be obtained. In dataset-1, logistic regression is a better algorithm based on an accuracy value of 80.09%, while naïve Bayes is superior based on an AUC value of 0.733 and an execution time of 0.00798 seconds. In dataset-2, it is found that decision tree is an algorithm that is more suitable than logistic regression and naïve Bayes algorithms, with an accuracy of 91.9% and an AUC value of 0.846 which is included in the good classification criteria. However, in execution time, the naïve Bayes algorithm only takes a processing time of 0.00403 seconds.

1. PENDAHULUAN

Telekomunikasi telah menjadi salah satu sektor industri yang diminati dan berkembang pesat di seluruh penjuru dunia. Perkembangan yang terjadi dapat mengakibatkan semakin banyaknya

DOI : https://doi.org/10.35316/jimi.v6i1.1223

P-ISSN : 2549-7480
Anita dkk / JIMI 6 (1) pp. 1-10 E-ISSN : 2549-6301

perusahaan industri telekomunikasi, sehingga berbagai ancaman dapat terjadi jika perusahaan tidak menggunakan strategi yang baik dalam mempertahankan produk dan meningkatkan pendapatan. Adapun strategi utama dalam meningkatkan pendapatan yaitu dengan memperoleh pelanggan baru, meningkatkan penjualan, dan meningkatkan periode retensi pelanggan [1]. Namun pada saat membandingkan strategi yang dilakukan dengan *return on investment* (RoI) menunjukkan bahwa pendekatan ketiga adalah strategi yang sesuai dan lebih menguntungkan karena biaya dalam mempertahankan pelanggan lama jauh lebih rendah dibandingkan dengan biaya dalam mendapatkan pelanggan baru [2]. Pendekatan ketiga dapat diterapkan dengan cara mengurangi potensi *churn* pada pelanggan yaitu peralihan yang dilakukan oleh pelanggan dari satu perusahaan ke perusahaan lain dalam waktu tertentu [3]. *Customer churn* mengacu pada tingkat pengurangan pelanggan di perusahaan. Hal ini menjadi masalah penting untuk dievaluasi oleh perusahaan bisnis yang sedang berkembang [4]. Tahap awal yang dilakukan untuk mengurangi potensi *churn* yaitu dengan memprediksi pelanggan yang berpotensi beralih maupun keluar dari perusahaan sehingga dapat meningkatkan keuntungan perusahaan. Prediksi *customer churn* dapat dilakukan dengan pendekatan *data mining*.

Data mining telah banyak digunakan di berbagai bidang, seperti pada bidang bisnis dan teknologi informasi. Adapun tugas utama dari *data mining* yaitu proses penggalian informasi hingga mengekstraksi pengetahuan dari data yang berjumlah besar [5]. Proses utama *data mining* terdiri dari eksplorasi seperti *preprocessing* data, transformasi data, seleksi atribut, dan lain sebagainya. Selanjutnya proses implementasi model, artinya memilih dan menerapkan metode yang akan digunakan berdasarkan tujuan dari proses *mining*. Tahap terakhir yaitu pengembangan [6]. *Dataset* yang didapatkan diolah dengan berbagai teknik yang ada pada *data mining*, salah satunya teknik klasifikasi. Teknik ini dapat digunakan untuk memprediksi kelas *dataset* berdasarkan data latih dan data uji untuk menghitung akurasi model yang digunakan [7].

Pada implementasinya, setiap teknik yang digunakan memiliki beberapa algoritma tertentu untuk memberikan solusi terhadap masalah yang didapatkan. Namun, setiap algoritma memberikan hasil dengan tingkat akurasi yang berbeda terhadap *dataset* yang digunakan sehingga diperlukan pemahaman awal mengenai *dataset* yang nantinya dapat digunakan untuk memilih algoritma yang sesuai dengan kondisi *dataset* tersebut. Salah satu kondisi yang berpengaruh yaitu jumlah *record* pada *dataset* terhadap algoritma yang diterapkan.

Pada penelitian sebelumnya, algoritma *naive bayes* memiliki kinerja yang lebih baik daripada *decision tree* c4.5 dalam melakukan prediksi terhadap *customer churn* [8]. Pada kasus lainnya, dalam melakukan prediksi, algoritma *decision tree* c4.5 memiliki tingkat akurasi dan kesalahan yang lebih baik dibandingkan algoritma *naive bayes* [9]. Penelitian lain dengan menerapkan algoritma *logistic regression* dan *logitboost*, didapatkan bahwa algoritma *logistic regression* memiliki nilai akurasi sedikit lebih tinggi dibandingkan *logitboost* yaitu 85.2385% pada *logistic regression* dan 85.1785% pada *logitboost*.

Berdasarkan hal tersebut, penelitian ini menggunakan algoritma *logistic regression*, *naive bayes*, dan *decision tree* c4.5 yang dipilih berdasarkan hasil akurasi yang baik dari penelitian sebelumnya. Maka penelitian ini akan mengidentifikasi pengaruh jumlah *record* dan

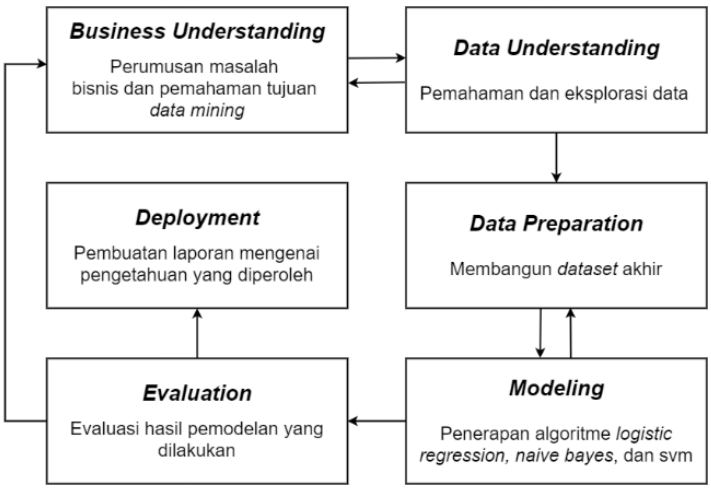
DOI : https://doi.org/10.35316/jimi.v6i1.1223

P-ISSN: 2549-7480
E-ISSN: 2549-6301

membandingkan ketiga algoritma yang digunakan untuk mengetahui algoritma yang sesuai berdasarkan jumlah *record*. Objek yang digunakan merupakan dua *dataset customer churn* yang diambil dari situs *public dataset Kaggle* yaitu *Telco Customer Churn* sebagai *dataset*-1 dan *Telco Churn Prediction* sebagai *dataset*-2 dengan jumlah masing-masing *record* yaitu 7043 dan 3333 *record*. Proses *data mining* dilakukan dengan bantuan *tools* Jupyter Notebook dan bahasa pemrograman Python.

2. METODE PENELITIAN

Penelitian dilakukan berdasarkan metodologi *Cross Industry Standard Process for Data Mining* (CRISP-DM) yang merupakan strategi pemecahan masalah secara umum dari bisnis sebagai alur penelitian yang terdiri dari enam tahap sebagai berikut [10].



Gambar 1. Alur Penelitian

1. *Business Understanding*

   Tahap ini merupakan tahap untuk mengenal proses bisnis dengan merumuskan beberapa masalah yang akan diselesaikan yang dapat dilihat dari situasi bisnis suatu perusahaan guna untuk mencapai tujuan pada *data mining*.

2. *Data Understanding*

   Suatu tahap mengumpulkan, mengeksplorasi data untuk memahami data yang didapatkan, dan mendeskripsikan data yang akan digunakan.

3. *Data Preparation*

   Suatu proses pemilihan data yang dilakukan untuk mempersiapkan data yang sesuai dengan kebutuhan proses *mining*. Tahap ini dapat dilakukan dengan memilih data yang sesuai, membersihkan data, dan melakukan integrasi terhadap data sehingga data termasuk pada *dataset* akhir yang siap untuk digunakan pada tahap selanjutnya.

4. *Modeling*

   Modeling merupakan tahap dalam mengolah data dengan menerapkan algoritma yang dipilih untuk mendapatkan informasi yang dibutuhkan.

5. *Evaluation*

   Tahap evaluasi keakuratan hasil yang telah didapatkan pada proses *modeling* guna mengetahui kinerja dari algoritma yang digunakan.

6. *Deployment*

   Menyusun laporan yang berdasar pada proses serta pengetahuan yang didapatkan [11]–[14].

DOI : https://doi.org/10.35316/jimi.v6i1.idpublikasi

P-ISSN : 2549-7480
Anita dkk / JIMI 6 (1) pp. 1-10 E-ISSN : 2549-6301

3. HASIL DAN PEMBAHASAN

Berikut merupakan hasil yang diperoleh yang disesuaikan dengan metodologi CRISP-DM.

3.1 *Business Understanding*

Tahap pemahaman terhadap tujuan penelitian yang selanjutnya diubah menjadi rencana *data mining* untuk mencapai tujuan tersebut [15]. Tujuan penelitian ini yaitu mengidentifikasi pengaruh dari jumlah *record dataset* terhadap beberapa algoritma klasifikasi yang digunakan. Hal ini dilakukan untuk menentukan algoritma yang sesuai dengan kondisi dari *dataset* berdasarkan jumlah *record* pada *dataset* tersebut.

3.2 *Data Understanding*

Data yang digunakan yaitu dua *dataset customer churn* yang berbeda, diambil dari situs *public dataset Kaggle*. *Dataset* yang pertama berjumlah 7043 *record* dan *dataset* kedua berjumlah 3333 *record*. *Dataset* yang digunakan berupa tabel dengan setiap kolomnya merepresentasikan atribut yang berisikan informasi mengenai pelanggan dan tiap barisnya berisi mengenai *record* atau observasi setiap pelanggan. Berikut keterangan pada *dataset*-1 yaitu:

a. *Churn*: pelanggan yang berhenti berlangganan selama satu bulan terakhir.
b. Layanan yang digunakan oleh pelanggan: *phone*, *multiple lines*, internet, *online security*, *online backup*, *device protection*, *tech support*, dan *streaming* TV/*movies*.
c. Informasi akun pelanggan: waktu menjadi pelanggan, kontrak, metode pembayaran, tagihan, dan total biaya.
d. Demografi pelanggan: jenis kelamin, rentang usia, serta memiliki pasangan dan tanggungan.

Adapun keterangan pada *dataset*-2 yaitu:

a. *Churn*: pelanggan yang berhenti berlangganan.
b. Demografi pelanggan: kode negara, dan kode area.
c. Informasi akun pelanggan: waktu aktif akun, dan jumlah biaya panggilan (siang, sore, malam hari, dan panggilan internasional).
d. Layanan yang digunakan setiap pelanggan: jumlah panggilan (siang, sore, malam hari, dan panggilan internasional), jumlah menit (siang, sore, malam hari, dan panggilan internasional), jumlah *voice mail* pelanggan, panggilan ke *customer service*.

3.3 *Data Preparation*

Tahap ini merupakan tahap membangun *dataset* akhir yang sesuai dengan proses *mining* yang akan dilakukan berdasarkan algoritma klasifikasi yang digunakan. Berikut proses yang dilakukan pada setiap dataset:

1. *Dataset*-1
   a. *Attribute reduction*
      Menghapus atribut yang tidak berpengaruh terhadap proses *mining*, seperti *customerID*.
   b. *Handling missing value*
      Mengatasi *missing value* pada TotalCharges sebagai atribut biaya pelanggan dengan cara menghapus *record* data *missing value*.
   c. *Replace name*
      Hal ini dilakukan dengan mengganti nama *record* yang sesuai dan diasumsikan sama berdasarkan *record* sebelum maupun sesudahnya. Pada kasus ini terdapat *record* '*Yes*', '*No*', '*No phone service*', dan '*No internet service*' yang berarti bahwa pelanggan tidak menggunakan fasilitas yang tersedia, sehingga diganti sebagai nilai '*No*'.
   d. *Data transformation*
      Mengubah atribut dengan tipe data nominal menjadi numerik.

DOI : https://doi.org/10.35316/jimi.v6i1.1223

P-ISSN: 2549-7480
Anita dkk/ JIMI 6 (1) pp. 1-10 E-ISSN: 2549-6301

Pengubahan tidak dilakukan pada atribut *Churn* karena merupakan label pada *dataset* dan digunakan untuk melakukan klasifikasi.

Setelah tahap *preparation* dilakukan, terdapat perubahan pada *dataset*-1 menjadi 7032 baris dengan 20 atribut.

Tabel 1. *Dataset*-1 Akhir

Multiple Lines	Online Security	Online Backup	Device Protection	…	Churn
0	0	0	0	…	0
0	1	1	1	…	0
0	1	0	0	…	1
0	1	1	1	…	0
0	0	1	0	…	1
…	…	…	…	…	…
1	1	1	1	…	0
1	0	0	1	…	0
0	1	1	0	…	0
1	0	1	0	…	1
0	1	1	1	…	0

2. *Dataset*-2
   a. *Attribute reduction*

      Menghapus atribut *phone* karena tidak diperlukan pada proses *mining* yang dilakukan.
   b. *Transformation*

      Mengubah tipe data nominal pada atribut *State* menjadi numerik.

Tidak ada perubahan pada *dataset*-2 setelah dilakukannya tahap *preparation*, yaitu tetap berjumlah 3333 baris dengan 20 atribut.

Tabel 2. *Dataset*-2 Akhir

Day Mins	Eve Mins	Night Mins	Intl Mins	…	Churn
265,1	197,4	244,7	10.0	…	0
161,6	195,5	254,4	13,7	…	0
243,4	121,2	162,6	12,2	…	0
299,4	61,9	196,9	6,6	…	0
166,7	148,3	186,9	10,1	…	0
…	…	…	…	…	…
156,2	215,5	279,1	9,9	…	0
231,1	153,4	191,3	9,6	…	0
180,8	288,8	191,9	14,1	…	0
213,8	159,6	139,2	5,0	…	0
234,4	265,9	241,4	13,7	…	0

3.4 *Modeling*

Tahap ini dilakukan dengan menguji jumlah *record* dua *dataset* dengan menerapkan serta membandingkan tiga algoritma klasifikasi yaitu *logistic regression*, *naive bayes*, dan *decision tree* c4.5 sehingga terdapat enam perbandingan pada hasil akhir pemodelan.

DOI : https://doi.org/10.35316/jimi.v6i1.idpublikasi

P-ISSN : 2549-7480
Anita dkk / JIMI 6 (1) pp. 1-10 E-ISSN : 2549-6301

Logistic Regression merupakan algoritma klasifikasi biner yaitu hanya terdapat dua nilai pada atribut targetnya dengan nilai interval 0 hingga 1. Jika label memiliki nilai lebih tinggi dari 0,5 maka diklasifikasikan sebagai kelas 1, jika kurang maka diklasifikasikan sebagai kelas 0 [16].

Naïve Bayes merupakan salah satu algoritma klasifikasi *data mining* berdasarkan pada teorema bayes yang digunakan dalam menghitung peluang dari setiap atribut dan menentukan kelas yang optimal [17]. Pada persoalan klasifikasi, algoritma ini melakukan pendekatan statistik terhadap inferensi induksinya [18].

Decision Tree merupakan algoritma yang terdiri dari *node* (*root*, *branch*, *leaf*) dan *edge*. *Tree* C4.5 merupakan bagian dari algoritma *decision tree*. Secara umum, algoritma ini memiliki dua proses yaitu melakukan penyusunan terhadap pohon keputusan (*decision tree*) dan membuat aturan, serta menghitung entropi untuk memilih atribut yang memiliki nilai *gain* tertinggi [19].

Pada prosesnya data akan dibagi menjadi *data training* sebanyak 70% dan *data testing* sebanyak 30%.

3.5 *Evaluation*

Pada tahap ini dilakukan evaluasi kualitas dari teknik pemodelan yang telah diterapkan. Pada penelitian ini evaluasi dilakukan dengan melihat nilai akurasi berdasarkan *confusion matrix*, nilai *area under curve* (AUC), dan *execution time*. *Confusion matrix* mengkategorikan hasil klasifikasi menjadi empat jenis yaitu *true positive* (TP), *true negative* (TN), *false positive* (FP), dan *false negative* (FN) [20].

Tabel 3. Confusion Matrix

	Actual True	***Actual False***
Prediction True	TP	FP
Prediction False	FN	TN

Dimana:

TP = Jumlah *record* positif, klasifikasi positif
FP = Jumlah *record* negatif, klasifikasi positif
FN = Jumlah *record* positif, klasifikasi negatif
TN = Jumlah *record* negatif, klasifikasi negatif

Nilai akurasi dapat diketahui berdasarkan *confusion matrix* dengan persamaan sebagai berikut.

$$Accuracy = \frac{TP+TN}{TP+FP+FN+TN} \quad (1)$$

Nilai AUC digunakan untuk mengetahui kriteria pemodelan berdasarkan teknik yang telah dilakukan dengan pernyataan keberhasilan sebagai berikut [21].

Tabel 4. Kriteria Nilai *Area Under Curve*

Nilai	**Kriteria**
0,90 – 1,00	Sangat baik (*excellent classification*)
0,80 – 0,90	Baik (*good classification*)
0,70 – 0,80	Wajar (*fair classification*)
0,60 – 0,70	Buruk (*poor classification*)
< 0,60	Gagal (*failure*)

DOI : https://doi.org/10.35316/jimi.v6i1.1223

P-ISSN: 2549-7480
E-ISSN: 2549-6301

Berikut adalah hasil evaluasi pemodelan yang telah dilakukan terhadap *dataset*-1 dan *dataset*-2.

1. *Dataset*-1

a. *Logistic regression*

Berikut merupakan *confusion matrix* dari pemodelan *logistic regression* pada *dataset*-1.

Tabel 5. *Confusion Matrix Logistic Regression Dataset*-1

	Actual Yes	***Actual No***
Prediction Yes	326	174
Prediction No	246	1364

$$Accuracy = \frac{326 + 1364}{326 + 174 + 246 + 1364}$$
$$= \frac{1690}{2110} = 0{,}8009 = 80{,}09\%$$

Nilai AUC yang diperoleh yaitu sebesar 0,728 dan *execution time* yang dibutuhkan yaitu 0,0678 detik.

b. *Naïve bayes*

Berikut merupakan *confusion matrix* dari pemodelan *naïve bayes* pada *dataset*-1.

Tabel 6. *Confusion Matrix Naive Bayes Dataset*-1

	Actual Yes	***Actual No***
Prediction Yes	387	366
Prediction No	164	1193

$$Accuracy = \frac{387 + 1193}{387 + 366 + 164 + 1193}$$
$$= \frac{1580}{2110} = 0{,}7488 = 74{,}88\%$$

Nilai AUC yang diperoleh yaitu sebesar 0,733 dan *execution time* yang dibutuhkan yaitu 0,00798 detik.

c. *Decision tree* c4.5

Berikut merupakan *confusion matrix* dari pemodelan *decision tree* c4.5 pada *dataset*-1.

Tabel 7. *Confusion Matrix Decision Tree c4.5 Dataset*-1

	Actual Yes	***Actual No***
Prediction Yes	282	290
Prediction No	296	1242

$$Accuracy = \frac{282 + 1242}{282 + 290 + 296 + 1242}$$
$$= \frac{1524}{2110} = 0{,}7222 = 72{,}22\%$$

Nilai AUC yang diperoleh yaitu sebesar 0,649 dan *execution time* yang dibutuhkan yaitu 0,0435 detik.

2. *Dataset*-2

a. *Logistic regression*

Berikut merupakan *confusion matrix* dari pemodelan *logistic regression* pada *dataset*-2.

DOI : https://doi.org/10.35316/jimi.v6i1.idpublikasi

P-ISSN : 2549-7480
Anita dkk / JIMI 6 (1) pp. 1-10 E-ISSN : 2549-6301

Tabel 8. *Confusion Matrix Logistic Regression Dataset*-2

	Actual Yes	***Actual No***
Prediction Yes	7	9
Prediction No	152	832

$$Accuracy = \frac{7 + 832}{7 + 9 + 152 + 832}$$
$$= \frac{839}{1000} = 0{,}839 = 83{,}9\%$$

Nilai AUC yang diperoleh yaitu sebesar 0,516 dan *execution time* yang dibutuhkan yaitu 0,0458 detik.

b. *Naïve bayes*

Berikut merupakan *confusion matrix* dari pemodelan *naïve bayes* pada *dataset*-2.

Tabel 9. *Confusion Matrix Naive Bayes Dataset*-2

	Actual Yes	***Actual No***
Prediction Yes	69	68
Prediction No	38	825

$$Accuracy = \frac{69 + 825}{69 + 68 + 38 + 825}$$
$$= \frac{894}{1000} = 0{,}894 = 89{,}4\%$$

Nilai AUC yang diperoleh yaitu sebesar 0,784 dan *execution time* yang dibutuhkan yaitu 0,00403 detik.

c. *Decision tree* c4.5

Berikut merupakan *confusion matrix* dari pemodelan *decision tree* c4.5 pada *dataset*-2.

Tabel 10. *Confusion Matrix Decision Tree c4.5 Dataset*-2

	Actual Yes	***Actual No***
Prediction Yes	112	42
Prediction No	39	807

$$Accuracy = \frac{112 + 807}{112 + 42 + 39 + 807}$$
$$= \frac{919}{1000} = 0{,}919 = 91{,}9\%$$

Nilai AUC yang diperoleh yaitu sebesar 0,846 dan *execution time* yang dibutuhkan yaitu 0,0299 detik.

Berdasarkan pengujian yang telah dilakukan terhadap *dataset*-1 dan *dataset*-2, dapat dilihat rekap nilai dari *accuracy*, AUC, dan *execution time* di bawah ini.

Tabel 11. Hasil Pengujian *Dataset*-1

	Accuracy	**AUC**	***Time***
LR	80,09%	0,728	0,0678
NB	74,88%	0,733	0,00798
DT	72,77%	0,649	0,0435

DOI : https://doi.org/10.35316/jimi.v6i1.1223

P-ISSN: 2549-7480
Anita dkk/ JIMI 6 (1) pp. 1-10 E-ISSN: 2549-6301

Tabel 12. Hasil Pengujian *Dataset*-2

	Accuracy	AUC	*Time*
LR	83,9%	0,516	0,0458
NB	89,4%	0,784	0,00403
DT	91,9%	0,846	0,0299

3.6 *Deployment*

Pada tahap ini dilakukan kegiatan pembuatan laporan dalam bentuk *paper* mengenai informasi dari hasil penelitian yang dilakukan yaitu pengaruh jumlah *record* pada dataset *customer churn* terhadap algoritma klasifikasi. Laporan ini juga dapat digunakan sebagai referensi untuk penelitian yang akan dilakukan selanjutnya.

4. KESIMPULAN

Berdasarkan penelitian yang telah dilakukan dapat disimpulkan bahwa semakin banyak jumlah *record* pada *dataset* maka semakin tinggi nilai *accuracy* yang didapatkan dan setiap *dataset* memiliki kesesuaian algoritma yang berbeda.

Pada *dataset*-1, *logistic regression* merupakan algoritma yang lebih baik berdasarkan nilai *accuracy* sebesar 80,09%, sedangkan *naïve bayes* lebih unggul berdasarkan nilai AUC sebesar 0,733 (*fair classification*) dan *execution time* yang dibutuhkan lebih sedikit yaitu 0,00798 detik.

Pada *dataset*-2 didapatkan bahwa *decision tree* merupakan algoritma yang lebih sesuai dibandingkan algoritma *logistic regression* dan *naïve bayes*, dengan *accuracy* sebesar 91,9% dan nilai AUC sebesar 0,846 yang termasuk pada kriteria *good classification*. Akan tetapi pada *execution time*, algoritma *naïve bayes* hanya membutuhkan waktu proses 0,00403 detik sehingga dapat dikatakan bahwa algoritma *naïve bayes* merupakan algoritma yang membutuhkan waktu proses yang lebih sedikit pada kedua *dataset customer churn*.

5. REFERENSI

[1] A. K. Ahmad, A. Jafar, and K. Aljoumaa, "Customer churn prediction in telecom using machine learning in big data platform," Journal of Big Data, vol. 6, no. 1, 2019.

[2] V. Kavitha, G. H. Kumar, S. V. M. Kumar, and M. Harish, "Churn Prediction of Customer in Telecom Industry using Machine Learning Algorithms," International Journal of Engineering Research and Technology, vol. 9, no. 5, pp. 181–184, 2020.

[3] J. Pamina et al., "An Effective Classifier for Predicting Churn in Telecommunication," Journal of Advanced Research in Dynamical and Control Systems, vol. 11, no. 1, pp. 221–229, 2019.

[4] Y. He, Y. Xiong, and Y. Tsai, "Machine Learning Based Approaches to Predict Customer Churn for an Insurance Company," Systems and Information Engineering Design Symposium, pp. 1–6, 2020.

[5] I. A. Nikmatun and I. Waspada, "Implementasi Data Mining untuk Klasifikasi Masa Studi Mahasiswa Menggunakan Algoritma K-Nearest Neighbor," Jurnal SIMETRIS, vol. 10, no. 2, pp. 421–432, 2019.

[6] M. A. Macmudi, "Uji Pengaruh Karakteristik Dataset," Journal of Computer, Information System, & Technology Management, vol. 1, no. 2, pp. 7–11, 2018.

[7] N. Sagala and H. Tampubolon, "Komparasi Kinerja Algoritma Data Mining pada Dataset Konsumsi

DOI : https://doi.org/10.35316/jimi.v6i1.idpublikasi

P-ISSN : 2549-7480
E-ISSN : 2549-6301

Alkohol Siswa," Khazanah Inform. Jurnal Ilmu Komputer dan Informatika, vol. 4, no. 2, pp. 98–103, 2018.

[8] N. W. Wardani and N. K. Ariasih, "Analisa Komparasi Algoritma Decision Tree C4 . 5 dan Naïve Bayes untuk Prediksi Churn Berdasarkan Kelas Pelanggan Retail," International Journal of Natural Sciences and Engineering, vol. 3, no. 3, pp. 103–112, 2019.

[9] N. Yahya and A. Jananto, "Komparasi Kinerja Algoritma C.45 Dan Naive Bayes Untuk Prediksi Kegiatan Penerimaan mahasiswa Baru (Studi Kasus : Universitas Stikubank Semarang)," Prosiding SENDI, pp. 221–228, 2019.

[10] A. R. Wibowo and A. Jananto, "Implementasi Data Mining Metode Asosiasi Algoritma FP-Growth Pada Perusahaan Ritel," Jurnal Teknologi Informasi dan Komunikasi, vol. 10, no. 2, pp. 200–212, 2020.

[11] C. Schröer, F. Kruse, and J. M. Gómez, "A Systematic Literature Review on Applying CRISP-DM Process Model," Procedia Computer Science, vol. 181, pp. 526–534, 2021.

[12] E. P. A. Akhmad, "Data Mining Menggunakan Regresi Linear untuk Prediksi Harga Saham Perusahaan Pelayaran," Jurnal Aplikasi Pelayaran dan Kepelabuhanan, vol. 10, no. 2, pp. 120–131, 2020.

[13] K. A. Pratama, G. A. Pradnyana, and I. K. R. Arthana, "Pengembangan Sistem Cerdas Untuk Prediksi Daftar Kembali Mahasiswa Baru Dengan Metode Naive Bayes (Studi Kasus: Universitas Pendidikan Ganesha)," SINTECH (Science and Information Technology) Journal, vol. 3, no. 1, pp. 22–34, 2020.

[14] I. Sutoyo, "Implementasi Algoritma Decision Tree Untuk Klasifikasi Data Peserta Didik," Jurnal Pilar Nusa Mandiri, vol. 14, no. 2, pp. 217–224, 2018.

[15] A. Purwanto, A. Primajaya, and A. Voutama, "Penerapan Algoritma C4.5 dalam Prediksi Potensi Tingkat Kasus Pneumonia di Kabupaten Karawang," Jurnal Sistem dan Teknologi Informasi, vol. 8, no. 4, pp. 390–396, 2020.

[16] A. Wicaksono, Anita, and T. N. Padilah, "Uji Performa Teknik Klasifikasi untuk Memprediksi Customer Churn," Bianglala Informatika, vol. 9, no. 1, pp. 37–45, 2021.

[17] I. Riadi, R. Umar, and F. D. Aini, "Analisis Perbandingan Detection Traffic Anomaly Dengan Metode Naive Bayes Dan Support Vector Machine (SVM)," ILKOM Jurnal Ilmiah, vol. 11, no. 1, pp. 17–24, 2019.

[18] H. Annur, "Klasifikasi Masyarakat Miskin Menggunakan Metode Naive Bayes," ILKOM Jurnal Ilmiah, vol. 10, no. 2, pp. 160–165, 2018.

[19] E. Budiman, Haviluddin, N. Dengan, A. H. Kridalaksana, M. Wati, and Purnawansyah, "Performance of Decision Tree C4.5 Algorithm in Student Academic Evaluation," International Conference on Computational Science and Technology, pp. 380–389, 2017.

[20] A. Nurmasani and Y. Pristyanto, "Algoritme Stacking Untuk Klasifikasi Penyakit Jantung Pada Dataset Imbalanced Class," Jurnal Pseudocode, vol. 8, no. 1, pp. 21–26, 2021.

[21] A. Bisri and R. Rachmatika, "Integrasi Gradient Boosted Trees dengan SMOTE dan Bagging untuk Deteksi Kelulusan Mahasiswa," Jurnal Nasional Teknik Elektro dan Teknologi Informasi (JNTETI), vol. 8, no. 4, pp. 309–314, 2019.

DOI : https://doi.org/10.35316/jimi.v6i1.1223