Vol. 8 No. 2, 2020 **Jurnal Komputasi**

Penerapan Algoritma C4.5 Untuk Prediksi *Churn Rate* Pengguna Jasa Telekomunikasi

[1]Yohana Tri Utami, [2]Dewi Asiah Shofiana & [3]Yunda Heningtyas

[1,2,3] Jurusan Ilmu Komputer FMIPA Universitas Lampung
Jl. Prof. Dr. Soemantri Brodjonegoro No. 1 Bandar Lampung 35145
e-mail : [1]yohana.utami@fmipa.unila.ac.id, [2]dewi.asiah@fmipa.unila.ac.id, [3]yunda.heningtyas@gmail.com

Abstract *— Telecommunication industries are experiencing substantial problems related to the migration of customers due to a large number of competing companies, dynamic circumstances, as well as the presence of many innovative and attractive offerings. The situation has resulted in a high level of customer migration, affecting a decrement toward the company revenue. Regarding that condition, the customer churn is one well-know approach that can help in increasing the company's revenue and reputation. As to predict the reason behind the migration of customer, this study proposed a data mining classification technique by applying the C4.5 algorithm. Patterns generated by the model were implemented using 10-fold cross-validation, resulting in a model with an accuracy rate of 87%, precision 87.5%, and a recall of 97%. Based on the good performance quality of the model, it can be stated that the C4.5 algorithm succeeded to discover several causes from the migration of telecommunication users, in which price holds the top place as the primary reason.*

Keywords**:** *C4.5 Algorithm; Classification; Customer Churn*; *Data Mining*

1. PENDAHULUAN

Pelanggan dianggap sebagai salah satu aset terpenting bagi bisnis di berbagai perusahaan yang dinamis dan kompetitif dalam *marketplace* [1]. Dalam menghadapi pasar yang kompetitif, pelanggan memiliki banyak pilihan penyedia layanan, mereka dapat dengan mudah beralih layanan. Pelanggan tersebut disebut sebagai pelanggan churned [2]. Penyebab *churn* pelanggan bisa disebabkan karena adanya ketidakpuasan, biaya lebih tinggi, kualitas rendah, kurangnya fitur, dan masalah privasi [3]

Pada penelitian [2] dilaporkan bahwa perusahaan industri telekomunikasi mengalami masalah substansial terkait dengan perpindahan pelanggan karena adanya persaingan yang ketat, kondisi dinamis dan adanya inovasi penawaran terbaru yang lebih menarik. Hasil pengamatan menyatakan bahwa memperoleh pelanggan baru dapat lebih mahal untuk perusahaan dibandingkan denganretensi pelanggan yang ada [4]. Para peneliti juga telah menyatakan bahwa pendekatan prediksi *customer churn* dapat meningkatkan pendapatan perusahaan dan reputasi yang baik di pasar [5]. Saat ini, perusahaan telekomunikasi memiliki banyak informasi tentang pelanggan mereka. Hal tersebut dapat menjadi peluang dalam memprediksi tehnik pemodelan untuk menangani perpindahan pelanggan di perusahaan telekomunikasi. Pendekatan ini dapat membantu dalam menentukan langkah-langkah dari data pengguna telekomunikasi untuk mencegah pelanggan mereka berpindah layanan dengan menawarkan promosi dan penawaran yang lebih baik [6].

Pada penelitian Amin, et al. [7] mengemukakan tentang sebuah pendekatan untuk memprediksi perpindahan pelanggan, yaitu pendekatan *Customer Churn Prediction (CCP)*. Pendekatan ini berfokus pada faktor jarak yang terbagi menjadi zona atas (nilai faktor jarak lebih besar) dan zona bawah (nilai faktor jarak terkecil) dengan memperkirakan klasifikasi tertentu. Selain itu, dapat menghitung tingkat prediksi kepastian *customer churn* dengan mengevaluasi hasil klasifikasi. Adapun hasil dari penelitian yang dilakukan menawarkan dua kontribusi, yaitu mengenalkan pendekatan baru untuk CCP berdasarkan faktor jarak dan mengungkapkan efek faktor jarak pada zona jarak yang berbeda. Selain itu, tingkat efektivitas ditunjukkan dalam hal *precision, recall, accuracy, dan f-measure.*

Berdasarkan pemasalahan yang ada, diperlukan tehnik dan metode pemecahan masalah yang dapat digunakan untuk mengidentifikasi penyebab adanya perpindahan pelanggan. Untuk mengidentifikasi pola tersebut,

penelitian ini menjelaskan mengenai pemodelan *data mining* dengan menggunakan tehnik klasifikasi dan menerapkan algoritma C4.5 untuk metode pemecahan masalahnya. Weka juga akan digunakan sebagai alat *(tools)* untuk mengolah data. Hasil penelitian ini menegaskan bahwa algoritma tersebut dapat digunakan sebagai model alternatif. Hasil identifikasi pola digunakan untuk membantu mempresiksi penyebab terjadinya perpindahan pelanggan pengguna jasa telekomunikasi agar perusahaan dapat meningkatkan layanan ke pelanggan sehingga dapat meminimalisir jumlah perpindahan pelanggan.

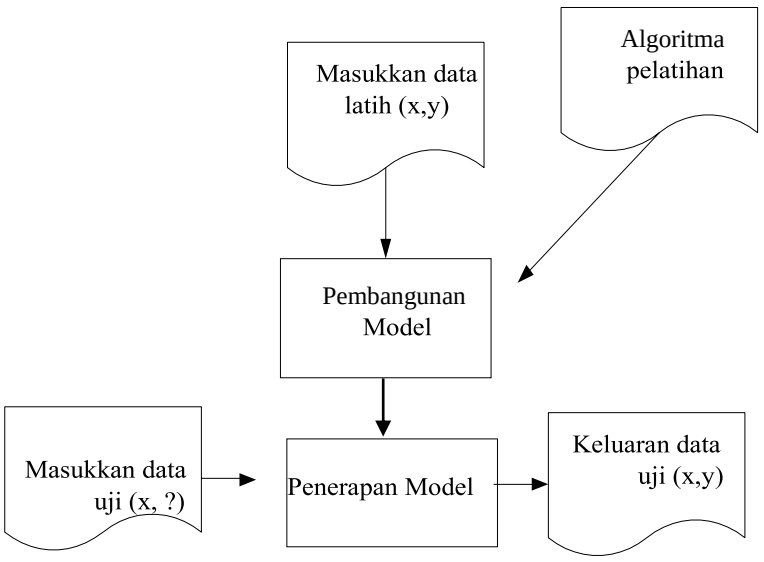
2. LANDASAN TEORI

2.1. *Data Mining*

Data Mining atau penggalian data adalah proses menemukan pengetahuan menarik dari sejumlah besar data yang tersimpan baik di *database*, gudang data, atau repositori informasi lainnya [8]. Penggalian data juga didefinisikan sebagai proses menemukan pola dalam data [9]. Penggalian data (langkah analisis penemuan pengetahuan dalam basis data) merupakan teknologi baru yang kuat ditingkatkan dan begitu cepat berkembang. Ini adalah teknologi yang dengan potensi besar untuk membantu bisnis dan perusahaan untuk berfokus pada informasi yang paling penting dari data yang mereka punya dan harus mengumpulkan untuk mengetahui perilaku pelanggan mereka [10].

2.2. Metode Klasifikasi

Klasifikasi merupakan proses menemukan model atau fungsi yang menggambarkan dan membedakan kelas atau konsep data. Model tersebut diperoleh dari analisis sekumpulan *data training* yang digunakan untuk memprediksi label kelas dari objek yang label kelasnya tidak diketahui [8]. Kerangka kerja klasifikasi ditunjukkan pada Gambar 1. Pada gambar tersebut, disediakan sejumlah data latih (x, y) untuk digunakan sebagai data membangun model, kemudian menggunakan model tersebut untuk memprediksi kelas dari data uji (x, ?) sehingga data uji (x, ?) diketahui kelas y yang seharusnya. Kerangka kerja yang ditunjukkan pada Gambar 2 meliputi dua langkah proses yaitu induksi dan deduksi. Induksi merupakan suatu langkah untuk membangn klasifikasi dari data latih yang diberikan atau disebut juga dengan proses pelatihan, sedangkan deduksi merupakan suatu langkah untuk menerapkan model tersebut pada data uji sehingga data uji dapat diketahui kelas yang sesungguhnya atau disebut juga dengan proses prediksi.



Gambar 1. Kerangka kerja klasifikasi [11]

2.3. Algoritma C4.5

Algoritma C4.5 merupakan algoritma yang dipergunakan dalam membentuk *decision tree* (pengambilan keputusan) [12]. Algoritma C4.5 adalah salah satu algoritma dalam induksi decision tree yaitu ID3 *(Iterative Dichotomiser* 3) yang dikembangan oleh J. Ross Quinlan. Dalam prosedur algoritma ID3, input berupa sampel

training, label *training* dan atribut. Algoritma C4.5 ini merupakan pengembangan dari ID3. Ide dasar dari algoritma ini adalah pembuatan pohon keputusan berdasarkan pemilihan atribut yang memiliki prioritas tertinggi atau dapat disebut memiliki nilai *gain* tertinggi berdasarkan nilai entropy atribut tersebut sebagai poros atribut klasifikasi. Kemudian secara rekursif cabang-cabang pohon diperluas sehingga seluruh pohon terbentuk. Terdapat empat langkah dalam proses pembuatan pohon keputusan pada algoritma C4.5, yaitu:

a. Memilih atribut sebagai akar
b. Membuat cabang untuk masing-masing nilai
c. Membagi setiap kasus dalam cabang
d. Mengulangi proses dalam setiap cabang sehingga semua kasus dalam cabang memilki kelas yang sama.

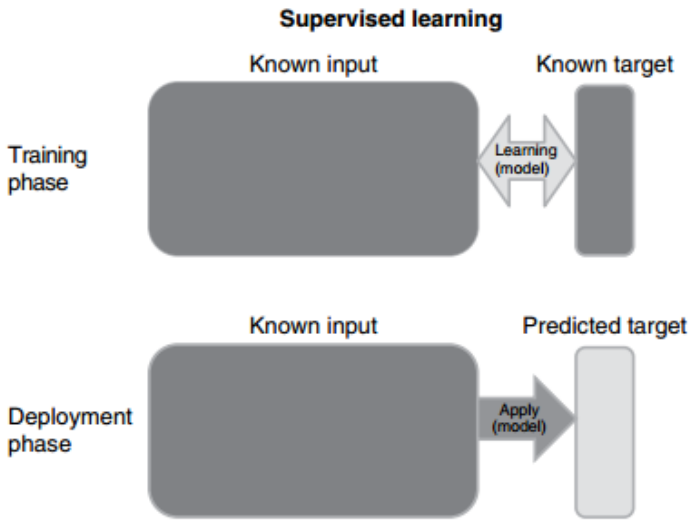
2.4. *Training, Validation and Testing Set*

Dalam pembangunan model *machine learning*, terdapat istilah *training* dan *testing*. *Training* merupakan proses konstruksi model sedangkan *testing* adalah sebuah proses untuk menguji kinerja model pemebelajaran dengan menggunakan *dataset* [13]. Secara umum, *dataset* dibagi menjadi tiga jenis, yaitu:

a. Training set merupakan kumpulan data yang digunakan untuk melatih atau membangun model.
b. Validation set merupakan kumpulan data yang dioptimisasi saat melatih model.
c. Testing set merupakan kumpulan data yang digunakan untuk menguji model setelah model selesai dilatih.

2.5. *Supervised Learning*

Dalam proses *data mining* menggunakan berbagai metode analisis data untuk menemukan yang tidak diketahui, tak terduga, menarik dan relevan pola dan hubungan di data yang dapat digunakan untuk membuat prediksi yang valid dan akurat [14].



Gambar 2. *Supervised learning* [14]

Supervised learning merupakan algoritma yang membutuhkan bantuan eksternal yang didasarkan dengan melatih sampel data dari sumber data dengan klasifikasi yang benar, untuk menghasilkan hipotesis umum atau model dari distribusi label kelas dalam hal prediksi atau klasifikasi [15]. Pada algoritma ini *dataset* masukan dibagi menjadi data latih (*data training)* dan data uji (*data testing)*. Data latih memiliki variabel keluaran yang

perlu diprediksi atau diklasifikasikan. Algoritma-algoritma yang sering digunakan dalam *supervised learning* ini misalnya *decision tree, naïve bayes, support vector machine*, dan *extreme gradient boosting* (XGBoost).

2.6. *Cross Validation*

Cross Validation merupakan salah satu metode statistik yang digunakan untuk menguji keefektifan model dalam pembelajaran mesin. Menurut Refaeilzadeh, Tang, & Liu [16] *cross validation* mengevaluasi dan membandingkan algoritma pembelajaran dengan membagi data menjadi dua segmen: satu digunakan untuk mempelajari atau melatih suatu model dan yang lainnya digunakan untuk memvalidasi model. *K-fold cross validation* merupakan bentuk dasar dari *cross validation*. Pada *k-fold cross validation*, pertama kali data dipartisi menjadi segmen atau lipatan yang sama (atau hampir sama). Selanjutnya k iterasi pelatihan dan validasi dilakukan sedemikian rupa sehingga dalam setiap iterasi lipatan data yang berbeda digunakan untuk validasi atau testing sedangkan k - 1 (k minus 1) lipatan lainnya digunakan untuk pembelajaran atau training.

2.7. *Confusion Matrix*

Confusion Matrix merupakan tentang informasi aktual dan prediksi klasifikasi yang dilakukan oleh sistem klasifikasi. Kinerja atau performa sistem klasifikasi tersebut biasanya dievaluasi menggunakan data dalam matriks. Tabel 1 merupakan tabel *confusion matrix* untuk du akelas *classifier*.

Tabel 1. *Confusion Matrix*

		Actual	
		Positive	*Negative*
Predicted	*Positive*	*True Positif (TP)*	*False Positive (FP)*
	Negative	*False Negative (FN)*	*True Negative (TN)*

Acuan dalam perhitungan matrix memiliki empat nilai, yaitu:
True Positive (TP) = kelas yang diprediksi positif, dan faktanya positif.
True Negative (TN) = kelas yang diprediksi negatif dan faktanya negatif
False Positive (FP) = kelas yang diprediksi positif dan faktanya negatif
False Negative (FN) = kelas yang diprediksi negatif dan faktanya positif

Berdasarkan nilai TP, TN, FP, dan FN akan diperoleh nilai akurasi dari penerapan algoritma. Nilai akurasi menggambarkan seberapa akurat sistem dapat mengklasifikasi data secara benar. Dari nilai akurasi, diperoleh persamaan sebagai berikut:

$$Accuracy = \frac{TP+TN}{(TP+TN+FP+FN)} x100\% \quad (1)$$

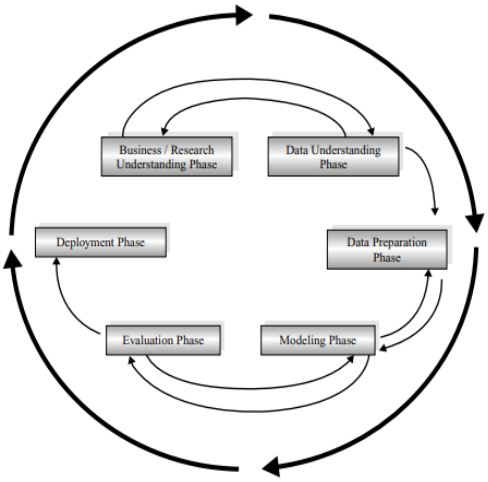
$$Precision = \frac{TP}{(TP+FP)} x100\% \quad (2)$$

$$Recall = \frac{TP}{(TP+FN)} x100\% \quad (3)$$

3. METODOLOGI PENELITIAN

3.1. CRISP-DM

Cross Industry Standard Process for Data Mining (CRISP-DM) merupakan salah satu kerangka kerja yang digunakan untuk menerjemahkan masalah bisnis menjadi tugas *data mining* dan melaksanakan proyek *data mining* yang independen dari area aplikasi dan teknologi yang digunakan **[17]**. Gambar 2 menunjukkan enam fase dan interaksinya dari model proses CRISP-DM.



Gambar 2. Proses CRISP-DM [17]

Enam fase dalam metodologi CRISP-DM, yaitu:

1. Fase Pemahaman Bisnis/Penelitian *(Business/Research Understanding Phase)*
   Proyek *data mining* dimulai dengan mendefinisikan pemahaman tujuan proyek, kemudian mengubah pengetahuan menjadi definisi masalah *data mining* dan rencana awal yang dirancang untuk mencapai tujuan [18].
   Adapun tujuan dalam penelitian ini adalah menerapkan tehnik klasifikasi untuk mengukur nilai akurasi dalam memprediksi penyebab perpindahan pengguna jasa telekomunikasi.
2. Fase Pemahaman Data *(Data Understanding Phase)*
   Tahap pemahaman data dimulai dengan mengumpulkan data awal, mengidentifikasi kualitas data, dan mendeteksi subset yang menarik untuk membentuk hipotesis [18]. Tahap Pemahaman Data dapat diidentifikasi sebagai kombinasi dari Seleksi dan *Preprocessing.*
   Data yang digunakan dalam penelitian ini menggunakan penyebaran kuesioner terhadap pengguna jasa telekomunikasi. Berdasarkan data tersebut, dilakukan proses *data cleaning* dengan menghapus serta mengoreksi data yang salah dan tidak lengkap agar data tersebut dapat di proses pada tahapan selanjutnya. Variabel yang tidak terpakai yaitu: nama, usia, alamat, status perkawinan, pekerjaan, dan pendidikan terakhir.
3. Fase Persiapan Data *(Data Preparation Phase)*
   Tahap persiapan data diidentifikasi dengan transformasi data. Transformasi data merupakan proses mengubah data dari satu format ke format lainnya [19].
   Data responden yang terkumpul dari hasil kuesioner akan diubah formatnya menjadi data yang berofrmat *comma separated values (*csv) agar dapat diproses pada tahap pemodelan.
4. Fase Pemodelan *(Modelling Phase)*
   Fase pemodelan dilakukan dengan memilih dan menerapkan tehnik pemodelan. Algoritma klasifikasi C4.5 digunakan pada pemodelan penelitian ini untuk menguji nilai akurasi, *presicion* dan *recall* dalam memprediksi penyebab perpindahan pengguna jasa telekomunikasi. Pemodelan dilakukan dengan membandingkan model menggunakan *k-fold cross validation* dan menggunakan model *training set.* Selanjutnya hasil pemodelan ini disajikan dalam tabel *confusion matrix* untuk dilakukan perhitungan nilai akurasi, *presicion* dan *recall*
5. Fase Evaluasi *(Evaluation Phase)*
   Pada tahap ini model yang dihasilkan akan dievaluasi secara menyeluruh dan melakukan peninjauan atas langkah-langkah yang telah dilakukan untuk membangun model tersebut agar mencapai tujuan bisnis dengan benar.
   Berdasarkan perbandingan pemodelan yang telah dilakukan dan penerapan algoritma klasifikasi yang telah dilakukan menghasilkan nilai akurasi sebesar 87%, *precision* sebesar 87,5% dan 96% untuk nilai *recall.*
6. Fase Penyebaran *(Deployment Phase)*

Berdasarkan hasil evaluasi, tahap ini akan menentukan strategi penerapan termasuk langkah yang diperlukan dan cara melakukannya. Hasil akurasi penerapan algoritma klasifikasi C4.5 masih menunjukkan presentase sebesar 87%. Hal tersebut menunjukkan bahwa terjadi banyak perpindahan pengguna jasa telekomunikasi yang diidentifikasi dengan atribut harga yang dianggap sebagai atribut yang paling menarik. Pola yang dihasilkan dapat membantu perusahaan dalam mengetahui penyebab perpindahan pelanggan, sehingga perusahaan dapat memperbaiki layanan kepada pelanggan untuk meminimalisir tingkat perpindahan pelanggan.

4. HASIL DAN PEMBAHASAN

Pada bagian ini menjelasakan tentang pengujian variabel evaluasi prediksi perpindahan pelanggan dengan menggunakan model *tree* C4.5. Penerapan algoritma C4.5 menggunakan *k-fold cross validation* dan *training set.* Pengujian menggunakan *k-fold* menghasilkan akurasi sebesar 83%, sedangkan menggunakan *training set* menghasilkan akurasi sebesar 87%. Berdasarkan hasil pengujian yang dilakukan, maka model klasifikasi yang dihasilkan dengan menggunakan *training set* dianggap lebih baik dengan perbandingan akurasi sebesar 4%. Tabel *confusion matrix* yang dihasilkan dari model ini disajikan dalam Tabel 2.

Tabel 2. Hasil *Confusion Matrix*

		Actual	
		Positive	*Negative*
Predicted	*Positive*	524	22
	Negative	75	127

Berdasarkan penerapan klasifikasi model algoritma C4.5 dihasilkan table *confusion matrix* dengan perhitungan sebagai berikut:

$$Accuracy = \frac{524+127}{(524+127+75+22)} \mathbf{x100\% = 87{,}03\%} \quad (1)$$

$$Precision = \frac{524}{(524+75)} \mathbf{x100\% = 87{,}5\%} \quad (2)$$

$$Recall = \frac{524}{(524+22)} \mathbf{x100\% = 96\%} \quad (3)$$

Proses evaluasi yang dilakukan dalam penelitian ini adalah mengukur tingkat akurasi dari algoritma C4.5 dalam memodelkan data perpindahan pengguna jasa telekomunikasi. Penggunaan *tools* weka membantu proses perhitungan akurasi algoritma dan diketahui bahwa akurasi algoritma yang dihasilkan sebesar 87%, *precision* sebesar 87,5% dan *recall* sebesar 96%. Penelitian Saito [20] merekomendasikan *presicion/recall* sebagai alat analisis visual yang paling informatif.

5. KESIMPULAN

Berdasarkan penelitian yang telah dilakukan, diperoleh beberapa kesimpulan bahwa algoritma *decision tree* C4.5 untuk melakukan klasifikasi pada set data pengguna jasa telekomunikasi menghasilkan nilai akurasi sebesar 87%, *precision* sebesar 87,5% dan *recall* sebesar 96% yang dianggap memperoleh hasil yang cukup baik. Atribut yang dianggap paling menarik dari hasil pengujian ini adalah atribut harga yang selanjutnya dapat diidentifikasi sebagai prediksi pola perpindahan pelanggan *(customer churn)*.

UCAPAN TERIMA KASIH

Terima kasih kami sampaikan kepada Lembaga Penelitian dan Pengabdian kepada Masyarakat (LPPM) Universitas Lampung yang telah mendanai penelitian ini sehingga penelitian dapat terlaksana dengan baik.

DAFTAR PUSTAKA

[1] K. Coussement, S. Lessmann and G. Verstraeten, "A comparative analysis of data preparation algorithms for customer churn prediction: A case study in the telecommunication industry," *Decision Support System*, pp. 27-36, 2017.

[2] M. Óskarsdóttir, C. Bravo, W. Verbeke, C. Sarraute, B. Baesens and J. Vanthienen, "Social Network Analytics for Churn Prediction in Telco: Model Building, Evaluation and Network Architecture," *Expert Systems with Applications*, pp. 204-220, 2017.

[3] R. R. Sharma and S. Rajan, "Evaluating prediction of customer churn behavior based on artificial bee colony algorithm.," *International Journal Of Engineering And Computer Science*, p. 20017–20021, 2017.

[4] A. Athanassopoulos, "Customer satisfaction cues to support market segmentation and explain switching behavior," *Journal of Business Research*, p. 191–207, 2010.

[5] O. Maria, V. W. , B. B. and J. V. , "A comparative study of social network classifiers for predicting churn in the telecommunication industry," *IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, p. 1151–1158, 2016.

[6] S. Jamil and K. A., "Churn comprehension analysis for telecommunication industry using ALBA," *International Conference on Emerging Technologies*, pp. 1-5, 20116.

[7] Amin, Adnan, F. A.-O. . S. Babar, A. A. J. L. and S. A. , "Customer churn prediction in telecommunication industry using data," *Journal of Business Research*, 2018.

[8] J. Han, M. Kamber and J. Pei, Data Mining Concepts and Techniques, USA: Morgan Kaufmann, 2012.

[9] I. Witten, E. Frank and M. Hall, Data Mining Practical Machine Learning Tools and Techniques, USA: Morgan Kaufmann, 2011.

[10] D. Al-Nabi, L. D. and S. S. Ahmed, "Survey on Classification Algorithms for Data Mining: (Comparison and Evaluation)," *Computer Engineering and Intelligent Systems*, pp. 18-25, 2013.

[11] E. Prasetyo, ata Mining Mengolah Data Menjadi Informasi Menggunakan Matlab, Yogyakarta: Andi Offset, 2014.

[12] Purushottam, K. Saxena and R. Sharma, "Efficient Heart Disease Prediction System using Decision Tree," in *International Conference on Computing, Communication and Automation (ICCCA)*, Noida, India, 2016.

[13] J. W. Gotama Putra, "Pengenalan Konsep Pembelajaran Mesin dan Deep Learning," Tokyo, Jepang, 2020.

[14] A. Ahlemeyer-Stubbe and S. Coleman, A Practical Guide to Data Mining for Business and Industry, USA: Wiley, 2014.

[15] M. Iqbal and Y. Zhu, "SUPERVISED MACHINE LEARNING APPROACHES: A SURVEY," *ICTACT JOURNAL ON SOFT COMPUTING*, vol. 05, no. 03, 2015.

[16] R. P. &. T. L. &. L. and H. , "Cross-Validation," *Encyclopedia of Database Systems*, pp. 532–538. 532-538. 10.1007/978-0-387-39940-9_565, 2009.

[17] R. Wirth and J. Hipp, "CRISP-DM: Towards a standard process model for data," *Proceedings of the 4th international conference on the practical*, pp. 29-39, 2000.

[18] A. Azevedo and M. F. Santos, "KDD, SEMMA and CRISP-DM: a parallel overview," in *Proceedings of the IADIS European Conference on Data Mining 2008*, pp 182–185, 2008.

[19] "https://www.cio.com/article/2378615/data-management/agile-comes-to-data-integration.html," Agile Comes to Data Integration, 2020. [Online]. [Accessed 2020].

[20] T. Saito and M. Rehmsmeier, "The Precision-Recall Plot Is More Informative than the," *PLoS ONE*, no. DOI:10.1371/journal.pone.0118432, 2015.