Analysis And Classification of Customer Churn Using Machine Learning Models

Muhammad Maulana Sidiq[1], Dyah Anggraini[2]
[1,2]Master of Information Systems Management, Department of Information System Software, Gunadarma University, Jakarta, Indonesia
[1]maulanasidiqwork@gmail.com, [2]dyahangg@staff.gunadarma.ac.id,

Abstract

Analysis studies of customer loss (customer churn) have been used for years to increase profitability and build customer relationships with companies. Customer analysis using exploratory data analysis (EDA) for visualizing data and the use of machine learning for the classification of customer churn are often used by past analysts. This study uses several machine learning models that can be used for customer churn classification, namely Logistic Regression, Random Forest, Support Vector Machine (SVM), Gradient Boosting, AdaBoost, and Extreme Gradient Boosting (XGBoost). However, there is a class imbalance factor in the dataset, which is the biggest challenge that is usually faced by analysts to get good results in the classification of machine learning models. The Synthetic Minority Over-sampling Technique (SMOTE) method is a popular method applied to deal with class imbalances in datasets. The results of the analysis show that the classification of churn customers using the XGBoost algorithm has the best level of accuracy compared to other algorithms, with an accuracy value of 0.829424, and the oversampling method with SMOTE tends to reduce the accuracy value of each classification algorithm. The Permutation Feature Importance (PFI) technique from the XGBoost model gets the result that tenure, monthly contracts, and TV streaming are the features that affect customer churn the most.

Keywords: data mining; machine learning models; imbalance data; SMOTE; confusion matrix

1. Introduction

Economic growth and development began to grow and recover since the COVID-19 virus spread rate decreased in mid-2022. With this opportunity, many new companies have emerged with innovations to lure consumers to use their products or services. Of course, this makes the atmosphere of business competition so intense that it makes businesspeople have to use various strategies to survive. As a result of this massive business competition, business actors sometimes legalize various kinds of ways to gain profits.

To survive in this competitive market depends on several strategies. There are three common strategies used to increase the company's revenue, which are: (1) acquiring new customers, (2) upselling existing customers, and (3) increasing customer retention period. Comparing these strategies by considering the return on investment (RoI) value of each strategy shows that the third strategy is the better strategy to pursue. Retaining existing customers costs much less than acquiring new customers. in addition, it is also considered much easier than the upselling strategy.

To implement the third strategy, companies must reduce the potential for losing customers, known as "switching customers from one provider to another".

It is important that every business needs customers to remain viable and profitable [1]. Getting customers who are willing and able to buy the business's products is one of the main objectives of setting up a business. Every step taken in the management of the company is focused on customer needs. This includes the implementation of various marketing strategies.

Companies also cannot abandon customers who already know their business. How to retain old customers is just as important as getting new customers. If the company thinks that only loyal customers are prioritized, then the company is mistaken. Loyal customers also come from new customers who get the best quality service from the company.

The switching of customers from one service provider to another, which refers to the loss of customers within a company, which can result in a significant loss of revenue for the company is also commonly referred to as customer churn [2]. Predicting customer churn is one

way to identify churn before customers switch services, which can be done by analyzing data and finding useful patterns in customer behaviour [3]. Of course, this can be done with a data mining approach [4] and classification of whether the customer is churn or loyal.

Examples of research that has been done with churn analysis are, in retail [5]-[7], subscription television companies [1], [8] telecommunications companies [9], [10], etc. Several prediction methods have been used for customer analysis including Gradient Boosting, Extreme Gradient Boosting, and Random Forest [11], [12]. The dataset used for retail used a dataset of 74,088 customers for 2 years[5].

The study proposed by Alrence Santiago Halibas et al., [13] applied exploratory data analysis and feature engineering in a collection of public domain Telecommunication datasets and applied seven classification techniques namely, Naïve Bayes, Logistic Regression, Generalized Linear Model, Deep Learning, Random Forest, Decision Tree, and Gradient Boosted Trees. The proposed study revealed that all the classifiers have achieved more than 70% accuracy and also examined the use of oversampling on minority classes. The study recommends the GBT algorithm as the best option for churn prediction.

In a journal written by Iqbal Muhammad Latief [10], Using data mining techniques with the AdaBoost algorithm to predict customer churn. The research used a customer dataset in the telecommunications sector totalling 7403 customer data. The data is validated using a ratio of 80% for training and 20% for testing. The highest accuracy results obtained using the AdaBoost algorithm reached 80%.

In analyzing and predicting using machine learning models, the toughest challenge for analysts is usually faced with the imbalance of classes in the dataset [14]. This can allow wrong predictions to occur in the machine learning model that has been trained. Therefore, there is a need for techniques to handle this class imbalance problem. One of the techniques that can be used for class imbalance problems is the Synthetic Minority Over-sampling Technique (SMOTE) [15], [16].

Based on the previously described background, this study aims to determine whether machine learning can classify customer churn problems using the IBM Telco dataset, which contains data from 7043 customers in California. The study will also examine how the results of data analysis related to the classification results of the Machine Learning model and whether data imbalance handling techniques can improve accuracy. To achieve these goals, customer data will be analyzed using Exploratory Data Analysis (EDA) and customer churn will be classified using machine learning models. The Synthetic Minority Over-sampling Technique

(SMOTE) will be used to handle class imbalance in the dataset. Several machine learning models will be used, and each machine learning model will be compared.

2. Research Methods

The research method used in this study uses the Cross Industry Standard Process for Data Mining (CRISP-DM)[17]-[19] framework without deployment. CRISP-DM (Cross-Industry Standard Process for Data Mining) is an open standard process model that describes common approaches used by data mining experts [20]. The described phases of the CRISP-DM cycle are often too general [21]. CRISP-DM is deeply grounded in the practical, real-world experience of how people conduct data mining projects [22]. CRISP-DM method covers the steps for analyzing data. Starting from the stages of understanding the business, understanding the data, preparing data, and building models, to the evaluation stage. The CRISP-DM Method is presented in Figure 1.
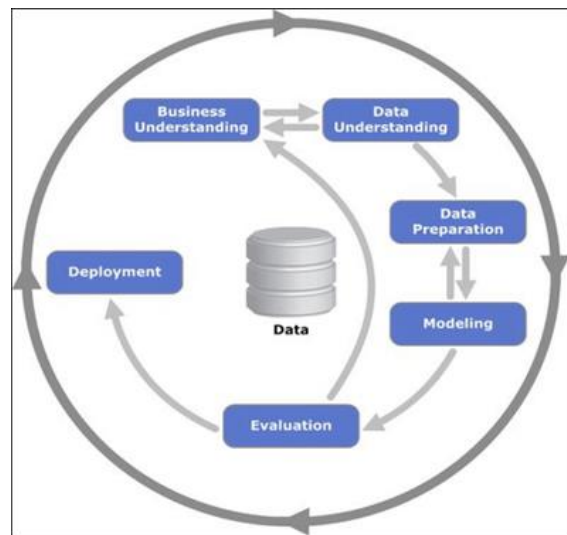

Figure 1. Phases of the CRISP-DM [23]

3.1 Business Understanding Phase

The sample dataset for customer churn in telecommunications, provided by IBM, is well-known and used as a resource for practising data analysis and training machine learning algorithms. The dataset provided contains information about a fictitious telecommunications company that provided landline and Internet services to 7043 customers in California in the 3rd quarter (Q3). The dataset shows which customers have left, stayed, or signed up for their services.

2.2 Data Understanding Phase

The dataset for customer churn in telecommunications provided by IBM has each column representing a customer. Each column contains attributes that describe the metadata in the column. The list of each column can

be seen in Figure 2. The information in the column in the dataset contains the following

Customers who left in the last month: this column is called Churn. Services that each customer has signed up for: such as phone columns, multiple lines, internet, online security, online backup, device protection, technical support, and TV and movie streaming. Customer account information contains how long they have been a customer, contracts, payment methods, paperless billing, monthly fees, and total fees. Demographic info about the customer contains gender, age range, and whether they have a spouse and dependents

```
telecom_cust.columns.values

array(['customerID', 'gender', 'SeniorCitizen', 'Partner', 'Dependents',
       'tenure', 'PhoneService', 'MultipleLines', 'InternetService',
       'OnlineSecurity', 'OnlineBackup', 'DeviceProtection',
       'TechSupport', 'StreamingTV', 'StreamingMovies', 'Contract',
       'PaperlessBilling', 'PaymentMethod', 'MonthlyCharges',
       'TotalCharges', 'Churn'], dtype=object)
```

Figure 2. List of columns/features in the dataset

2.2 Data Preparation Phase

In the data preparation stage consists of 2 stages including Feature engineering, and Exploratory Data Analysis (EDA) [13], [24], [25]. The Feature Engineering stage is divided into 3 stages, namely removing duplicate data, deleting empty-value data, and Encoding [26]. The EDA stage consists of 4 stages, namely demographic analysis, customer data analysis, customer distribution based on services, and customer cost distribution. The stages of the data preparation phase can be seen in Figure 3.
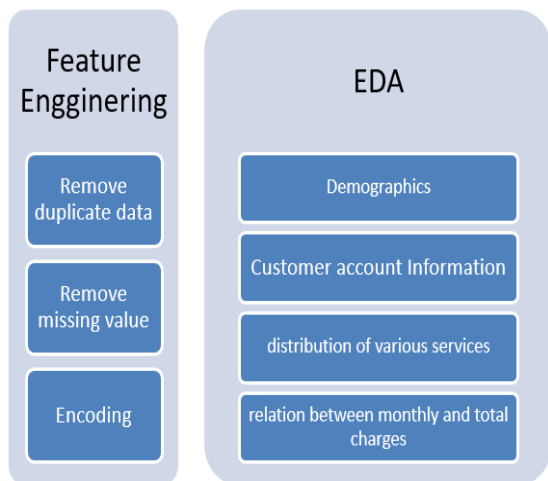

Figure 3. Data Preparation Flow

2.3. Modeling Phase

The modeling process is done after going through the data preparation stage where there is a data transformation phase and the data is considered ready. This modelling stage performs activities including determining training data and test data, determining machine learning algorithms, conducting test tests of several machine learning algorithms, and finding the best setting method for machine learning models such as resampling. For this research, the division of training data and test data is divided into 80% training data and 20% test data based on the total data in the dataset.

The total amount of training data that has been divided is 5625 data, while the total amount of test data is 1407 data. After performing the data division stage, the data is checked whether the classes in the data are balanced or not. In this study, it is known that the classes in the data are still unbalanced so it is necessary to do resampling to balance the classes in the data using the SMOTE method. After resampling, the results became balanced. Where it becomes 4121 majority classes and 4121 minority classes. Later, a comparison of classification results will be carried out on models whether the accuracy value is better than those that are not resampled with SMOTE.

2.4 Evaluation Phase

The result evaluation process uses a Confusion Matrix [27] which will get accuracy [28], precision, recall, and F1-score values from each machine learning algorithm model used. The confusion matrix is a 2x2 table that shows the number of correct and incorrect predictions from the model. The confusion matrix can be seen in Figure 4.

Figure 4. Confusion matrix

True Positive (TP) means the number of correct and positive predictions, False Positive (FP) means the number of false and positive predictions, True Negative (TN) means the number of correct and negative predictions, and True Negative (TN) means The number of correct and negative predictions.

Accuracy is the ratio of the number of correct predictions compared to the total data (the percentage of customers correctly predicted to churn and not churn out of the total customer data). The accuracy value is obtained based on Formula 1.

$$\frac{TP+TN}{TP+TN+FP+FN} \tag{1}$$

Precision is the ratio of positive true predictions compared to the overall positive predicted outcomes.

(the percentage of customers who churn out of all customers predicted to churn). The precision value is obtained based on Formula 2.

$$\frac{TP}{TP+FP} \qquad (2)$$

Recall is the ratio of true positive predictions compared to the overall true positive data. (percentage of customers predicted to churn compared to all customers who churned). The Recall value is obtained based on Formula 3.

$$\frac{TP}{TP+FN} \qquad (3)$$

The F1 score is a comparison of weighted average precision and recall. The F1-Score value is obtained based on Formula 4.

$$2 * \frac{Precission}{Precission+Recall} \qquad (4)$$

With permutation feature importance (PFI) can know a feature is considered 'important' if shuffling its values increases the model's error, indicating that the model relies on the feature for making predictions. Conversely, a feature is considered 'unimportant' if shuffling its values does not change the model's error, indicating that the model does not use the feature for making predictions[29].

The PFI has the following algorithm: Input = Trained model f, feature matrix X, target vector y, error measure L(y,f). The original model error estimating is shown in Formula 5.

$$e_{orig} = L(y, \hat{f}(X)) \qquad (5)$$

For each feature $j \in \{1, \dots, p\}$ do: Generate feature matrix Xperm by permuting feature j in the data X. Estimate error the predictions of the permuted data, showing in Formula 6.

$$e_{perm} = L(y, \hat{f}(X_{perm})) \qquad (6)$$

Calculate permutation feature importance shown in Formula 7. Finally, sort features by descending FI.

$$FI_j = e_{perm}/e_{orig} \qquad (7)$$

3. Results and Discussions

The evaluation results are divided into 2, which are the results of the analysis of customer data carried out at the EDA phase and the results of the analysis of the machine learning algorithm model classification.

3.1 Customer Data Analysis

Demographic Analysis (Gender distribution, seniority level, relationship status) can be seen in Figure 5. For churn distribution based on gender, women and men have the same percentage for churn at 27%.
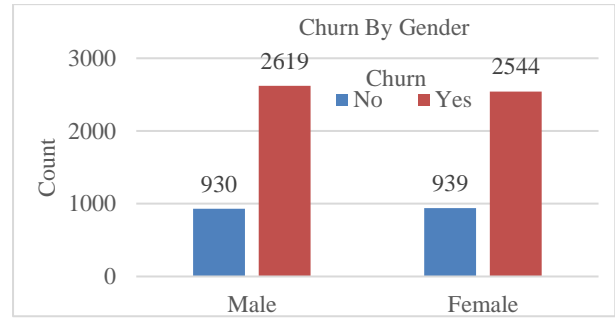
Figure 5. Churn distribution by gender.

The distribution of churn based on the seniority level in the community in Figure 6 has a percentage of 42% churn for the senior (old).
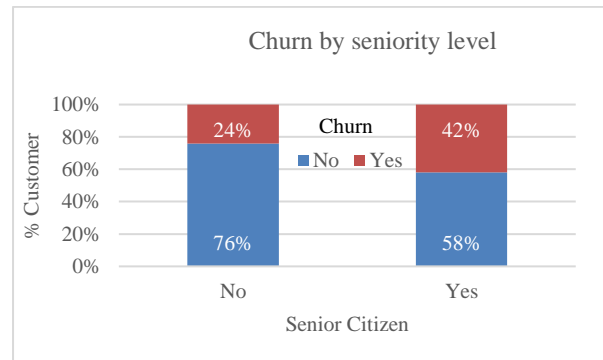
Figure 6. Churn Distribution by Seniority Level

The distribution for customers who already have a partner and have dependents in Figure 7 has a percentage of 14% churn.
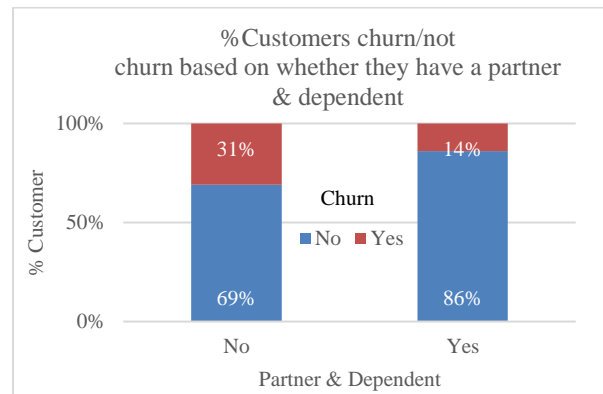
Figure 7. Distribution of customer churn against partner & dependent features

For the distribution of churn based on contract type in Figure 8, we found that customers who choose the monthly contract type, tend to churn more easily. The percentage of churn for customers who choose the monthly contract type reaches a total of 31% of all customers, compared to customers with the one-year contract type with only 14% of customers churn.

The tenure of customers using the service shows that customers who take longer contracts are more loyal to

the company and tend to stay for a longer period. Customers with monthly contract types last for 1-2 months, while customers with 2-year contract types tend to last for around 70 months. The result can be seen in Figure 9.
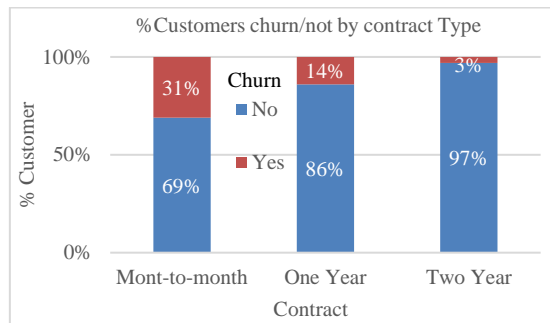

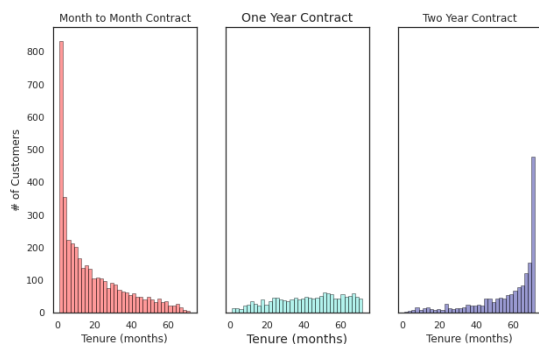Figure 8. Distribution of customer churn by contract type

Figure 9. Distribution of customer tenure by contract type

The distribution of customer churn based on the services used, the results shows that customers who use a small number of services tend to churn more easily. In Figure 10, the most influential services for customer churn are, customers who do not use online security services, customers who use internet services with Fiber optic cables, and customers who do not use tech support services.
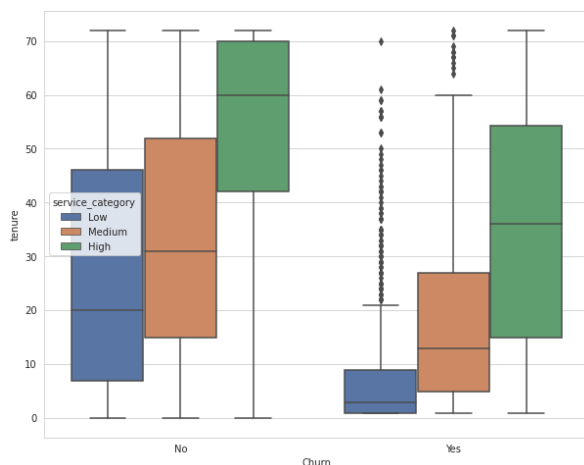

Figure 10. Customer churn based on the number of services used

The distribution of churn by cost shows in Figure 11, that customers with monthly subscription fees ranging from $70 - $100 have a significant increase in churn.
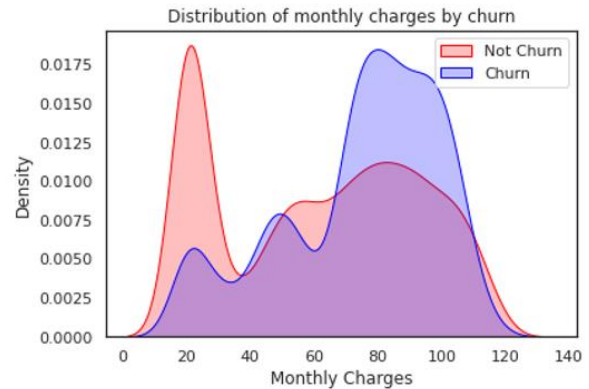

Figure 11. Churn distribution based on monthly fee

Customers with total subscription costs between $3786 and $8550 in Figure 12 have a longer tenure time compared to others. Then, customers with lower total subscription costs tend to have a shorter tenure time.
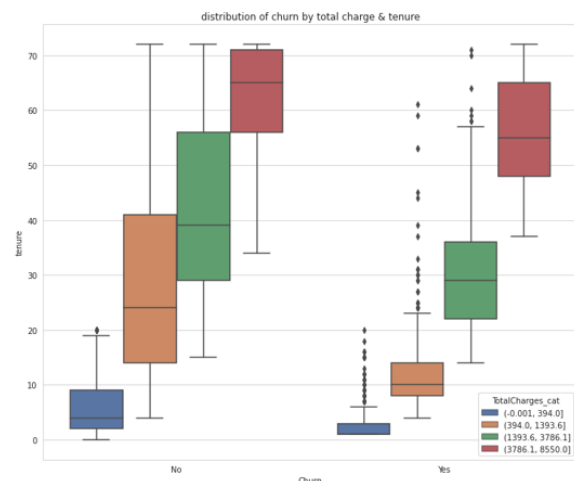

Figure 12. Distribution of churn based on total cost and tenure

3.2 Machine Learning Model Classification Result Analysis

The accuracy value of the XGBoost algorithm for customer churn classification has the highest result compared to other algorithms. XGBoost gets an accuracy value of 0.829424 or if the percentage and rounded up to 83% accuracy of success in customer churn classification against all data. For the precision value, the Logistic Regression algorithm with SMOTE resampling gets the highest precision value compared to other algorithms observed in this study. LR+SMOTE gets a precision value of 0.922078. The highest Recall value is in the Random Forest algorithm with a recall value of 0.845133. Then the highest F1-Score is obtained by the SVM algorithm with its F1-Score value of 0.840959.

The use of oversampling techniques with SMOTE does not increase the accuracy value of each classification algorithm model. The SMOTE method is more likely to reduce the accuracy value. However, the use of the SMOTE method increases the precision value of each algorithm model. Precision values are good to use to find consistent model results. The result is presented in Table 1.

Table 1. Evaluation of classification results of all models

Algoritma	Accuracy	Precision	Recall	F1-Score
Logistic Regression	0.820185	0.854447	0.824090	0.838994
LR + SMOTE	0.767591	0.922078	0.723148	0.810586
Random Forest	0.803127	0.834061	0.845133	0.839560
RF + SMOTE	0.795309	0.853659	0.813226	0.832952
SVM	0.819474	0.845009	0.836947	0.840959
SVM + SMOTE	0.784648	0.902944	0.750000	0.819396
AdaBoost	0.815920	0.845543	0.834495	0.839982
AdaBoost + SMOTE	0.785359	0.912027	0.741176	0.817773
XGBoost	0.829424	0.857398	0.824336	0.840542
XGBoost + SMOTE	0.797441	0.885831	0.774510	0.826438
Gradient Boosting	0.823028	0.853072	0.827288	0.839982
GBoost + SMOTE	0.798863	0.887640	0.773132	0.826438

By using Permutation Feature Importance (PFI), the XGBoost algorithm model, for the 3 features that most affect customer churn, namely: (1) Tenure, (2) Monthly contract, and (3) Streaming TV service. The PFI method normalizes the biased measure based on a permutation test and returns significant P-values for each feature [30]. The PFI XGB result is presented in Figure 13.
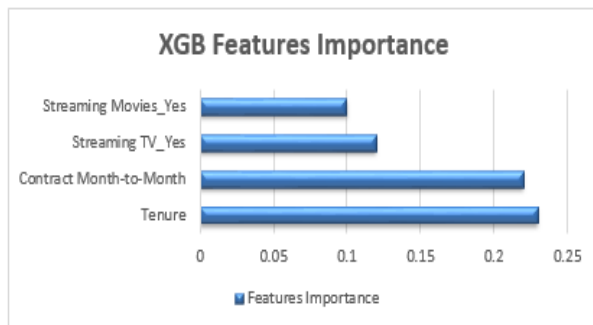

Figure 13. XGBoost features Importance

4. Conclusion

In this study, the use of machine learning models has proven invaluable for analysts and business professionals to effectively categorize customer churn. The analysis of the IBM Telco Customer churn dataset revealed crucial insights. Only 27% of customers were classified as churn customers based on gender, with senior community members being more susceptible to churning. Notably, having both a partner and dependents correlated with lower churn rates, while monthly contract types increased the likelihood of

churning. Additionally, using more services positively impacted customer retention, and larger customers with higher subscription costs tended to have longer tenure, especially with certain contract types. When employing the SMOTE resampling technique in machine learning models, accuracy may decrease, but precision improves, highlighting its value in addressing class imbalance. The XGBoost machine learning model achieved the highest accuracy score at 0.829424 in this study. Key factors influencing customer churn were Tenure, Contract month-to-month, and Streaming TV service, indicating that customers with shorter tenures and specific contract types were more likely to churn. As the research concludes, it encourages future studies on this topic, aiming for optimal results in classifying customer churn. For further research, the author suggests exploring additional techniques to handle class imbalance that can enhance accuracy.

References

[1] N. Suryana, " Prediksi Churn Dan Segmentasi Pelanggan Tv Berlangganan (Studi Kasus Transvision Jawa Barat)," *J. TEDC*, vol. 11, no. 2, pp. 185–191, 2019, [Online]. Available: https://ejournal.poltektedc.ac.id/index.php/tedc/article/view/77

[2] J. Pamina *et al.*, "An effective classifier for predicting churn in telecommunication," *J. Adv. Res. Dyn. Control Syst.*, vol. 11, no. 1 Special Issue, 2019.

[3] K. Matuszelański and K. Kopczewska, "Customer Churn in Retail E-Commerce Business: Spatial and Machine Learning Approach," *J. Theor. Appl. Electron. Commer. Res.*, vol. 17, no. 1, 2022, doi: 10.3390/jtaer17010009.

[4] B. Chen, Y. Liu, and J. Zheng, "Using Data Mining Approach for Student Satisfaction With Teaching Quality in High Vocation Education," *Front. Psychol.*, vol. 12, 2022, doi: 10.3389/fpsyg.2021.746558.

[5] Omer Faruk Seymen, Onur Doğan, Orhan Er, Kadir Hızıroğlu, and Emre Ölmez, "Customer Churn Prediction Using Ordinary Artificial Neural Network and Convolutional Neural Network Algorithms: A Comparative Performance Assessment," *J. Sci.*, vol. 36, no. 2, pp. 720–733, 2021.

[6] N. Bagul, P. Berad, P. Surana, and C. Khachane, "Retail Customer Churn Analysis using RFM Model and K-Means Clustering," *Int. J. Eng. Res. Technol.*, vol. 10, no. 03, 2021.

[7] N. W. Wardani, G. R. Dantes, and G. Indrawan, "Prediksi Customer Churn dengan Algoritma Decision Tree C4.5 Berdasarkan Segmentasi Pelanggan untuk Mempertahankan Pelanggan pada Perusahaan Retail," *J. Resist. (Rekayasa Sist. Komputer)*, vol. 1, no. 1, 2018, doi: 10.31598/jurnalresistor.v1i1.219.

[8] R. Venkatraman and R. Ragala, "A survey on churn analysis and prediction in video on demand," *Asian J. Pharm. Clin. Res.*, vol. 10, 2017, doi: 10.22159/ajpcr.2017.v10s1.19603.

[9] F. Kayaalp, "Review of Customer Churn Analysis Studies in Telecommunications Industry," *Karaelmas Sci. Eng. J.*, vol. 7, no. 2, 2017.

[10] I. M. Latief, A. Subekti, and W. Gata, "Prediksi Tingkat Pelanggan Churn Pada Perusahaan Telekomunikasi Dengan Algoritma Adaboost," *J. Inform.*, vol. 21, no. 1, 2021, doi: 10.30873/ji.v21i1.2867.

[11] I. Ullah, B. Raza, A. K. Malik, M. Imran, S. U. Islam, and S. W. Kim, "A Churn Prediction Model Using Random Forest: Analysis of Machine Learning Techniques for Churn Prediction and Factor Identification in Telecom Sector," *IEEE Access*, vol. 7, pp. 60134–60149, 2019, doi: 10.1109/ACCESS.2019.2914999.

[12] A. Mahalekshmi and G. H. Chellam, "Analysis of customer churn prediction using machine learning and deep learning algorithms," *Int. J. Health Sci. (Qassim).*, May 2022, doi: 10.53730/ijhs.v6ns1.7861.

[13] A. S. Halibas, A. Cherian Matthew, I. G. Pillai, J. Harold Reazol, E. G. Delvo, and L. Bonachita Reazol, "Determining the intervening effects of exploratory data analysis and feature engineering in telecoms customer churn modelling," in *2019 4th MEC International Conference on Big Data and Smart City, ICBDSC 2019*, 2019. doi: 10.1109/ICBDSC.2019.8645578.

[14] A. Amin *et al.*, "Comparing Oversampling Techniques to Handle the Class Imbalance Problem: A Customer Churn Prediction Case Study," *IEEE Access*, vol. 4, 2016, doi: 10.1109/ACCESS.2016.2619719.

[15] A. Fernández, S. García, F. Herrera, and N. V. Chawla, "SMOTE for Learning from Imbalanced Data: Progress and Challenges, Marking the 15-year Anniversary," *Journal of Artificial Intelligence Research*, vol. 61. 2018. doi: 10.1613/jair.1.11192.

[16] V. Flores, S. Heras, and V. Julian, "Comparison of Predictive Models with Balanced Classes Using the SMOTE Method for the Forecast of Student Dropout in Higher Education," *Electron.*, vol. 11, no. 3, 2022, doi: 10.3390/electronics11030457.

[17] W. Y. Ayele, "Adapting CRISP-DM for idea mining a data mining process for generating ideas using a textual dataset," *Int. J. Adv. Comput. Sci. Appl.*, vol. 11, no. 6, 2020, doi: 10.14569/IJACSA.2020.0110603.

[18] V. Plotnikova, M. Dumas, and F. P. Milani, "Applying the CRISP-DM data mining process in the financial services industry: Elicitation of adaptation requirements," *Data Knowl. Eng.*, vol. 139, 2022, doi: 10.1016/j.datak.2022.102013.

[19] C. Schröer, F. Kruse, and J. M. Gómez, "A systematic literature review on applying CRISP-DM process model," in *Procedia Computer Science*, 2021, vol. 181. doi: 10.1016/j.procs.2021.01.199.

[20] J. S. Saltz, "CRISP-DM for Data Science: Strengths, Weaknesses and Potential Next Steps," in *Proceedings - 2021 IEEE International Conference on Big Data, Big Data 2021*, 2021. doi: 10.1109/BigData52589.2021.9671634.

[21] F. Schafer, C. Zeiselmair, J. Becker, and H. Otten, "Synthesizing CRISP-DM and Quality Management: A Data Mining Approach for Production Processes," in *2018 IEEE International Conference on Technology Management, Operations and Decisions, ICTMOD 2018*, 2018. doi: 10.1109/ITMC.2018.8691266.

[22] C. P. Lopez, *Data Mining. The CRISP-DM Methodology. The CLEM Language and IBM SPSS Modeler*, vol. 7, no. 2. 2021.

[23] N. Mohd Nawi, M. Makhtar, M. Salikon, and Z. Afip, "A comparative analysis of classification techniques on predicting flood risk," *Indones. J. Electr. Eng. Comput. Sci.*, vol. 18, p. 1342, Jun. 2020, doi: 10.11591/ijeecs.v18.i3.pp1342-1350.

[24] K. Sahoo, A. K. Samal, J. Pramanik, and S. K. Pani, "Exploratory data analysis using python," *Int. J. Innov. Technol. Explor. Eng.*, vol. 8, no. 12, 2019, doi: 10.35940/ijitee.L3591.1081219.

[25] R. S. Oktavian and S. Budi, "Analisis Dataset Google Playstore Menggunakan Metode Exploratory Data Analysis," *J. Strateg. Maranatha*, vol. 2, no. 2, 2020.

[26] H. Wen *et al.*, "Multiparametric Quantitative US Examination of Liver Fibrosis: A Feature-Engineering and Machine-Learning Based Analysis," *IEEE J. Biomed. Heal. Informatics*, vol. 26, no. 2, 2022, doi: 10.1109/JBHI.2021.3100319.

[27] B. P. Pratiwi, A. S. Handayani, and S. Sarjana, "Pengukuran Kinerja Sistem Kualitas Udara Dengan Teknologi Wsn Menggunakan Confusion Matrix," *J. Inform. Upgris*, vol. 6, no. 2, 2021, doi: 10.26877/jiu.v6i2.6552.

[28] W. Rahayu and E. Wahyudi, "Classical Test Theory Of Innapropriate Index Score's Accuracy Comparison Using Confusion Matrix Accuracy Proportion In Educational Measurement," *IJER - Indones. J. Educ. Rev.*, vol. 4, no. 1, 2017, doi: 10.21009/ijer.04.01.08.

[29] A. Fisher, C. Rudin, and F. Dominici, "All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously," *J. Mach. Learn. Res.*, vol. 20, 2019.

[30] H. Kaneko, "Cross-validated permutation feature importance considering correlation between features," *Anal. Sci. Adv.*, vol. 3, no. 9–10, 2022, doi: 10.1002/ansa.202200018.