JOIV : Int. J. Inform. Visualization, 8(2) - May 2024 897-905

INTERNATIONAL JOURNAL ON INFORMATICS VISUALIZATION

journal homepage : www.joiv.org/index.php/joiv

Leveraging Various Feature Selection Methods for Churn Prediction Using Various Machine Learning Algorithms

Kusnawi Kusnawi [a,*], Joang Ipmawati [b], Bima Pramudya Asadulloh [a], Afrig Aminuddin [a], Ferian Fauzi Abdulloh [a], Majid Rahardi [a]

[a] *Faculty of Computer Science, Universitas Amikom Yogyakarta, Depok, Sleman, Indonesia*
[b] *Faculty of Information Technology, Universitas Nahdlatul Ulama Yogyakarta, Gamping, Sleman, Indonesia*
Corresponding author: [*]khusnawi@amikom.ac.id

***Abstract*— This study aims to examine the effect of customer experience on customer retention at DQLab Telco, using machine learning techniques to predict customer churn. The study uses a dataset of 6590 customers of DQLab Telco, which contains various features related to their service usage and satisfaction. The data includes various features such as gender, tenure, phone service, internet service, monthly charges, and total charges. These features represent the demographic and service usage information of the customers. The study applies several feature selection methods, such as ANOVA, Recursive Feature Elimination, Feature Importance, and Pearson Correlation, to select the most relevant features for churn prediction. The study also compares three machine learning algorithms, namely Logistic Regression, Random Forest, and Gradient Boosting, to build and evaluate the prediction models. This study finds that Logistic Regression without feature selection achieves the highest accuracy of 79.47%, while Random Forest with Feature Importance and Gradient Boosting with Recursive Feature Elimination achieve accuracy of 77.60% and 79.86%, respectively. The study also identifies the features influencing customer churn most, such as monthly charges, tenure, partner, senior citizen, internet service, paperless billing, and TV streaming. The study provides valuable insights for DQLab Telco in developing customer churn reduction strategies based on predictive models and influential features. The study also suggests that feature selection and machine learning algorithms play a vital role in improving the accuracy of churn prediction and should be customized according to the data context.**

***Keywords*— Machine learning; feature selection; customer experience.**

Manuscript received 30 Dec. 2023; revised 31 Jan. 2024; accepted 23 Feb. 2024. Date of publication 31 May. 2024.

I. INTRODUCTION

Customer experience is crucial for service businesses such as communications, entertainment, etc. Customer retention with a particular service provider is vital for running the business. A good customer experience creates a good impression for the customer, increasing the chances of a customer staying with a service provider. Conversely, a lousy customer experience makes a wrong impression on the customer, increasing the likelihood of the customer switching to a competitor. Switching customers to competitors is not what a company wants. Therefore, companies must study the factors or conditions supporting switching customers to competitors [1].

DQLab Telco is a company that has a business in telecommunication services. In its development, DQLab Telco has several branches spread out in various places. Established in 2019, the company consistently focuses on customer experience [2]. DQLab Telco's consistency towards customer experience aims to prevent DQLab Telco from being abandoned by customers. Since a little more than one year of operation, DQLab Telco has been abandoned by many of its customers, and many customers have switched subscriptions to competing companies. DQLab Telco management plans to reduce the number of customers who switch from DQLab Telco to competing companies.

DQLab Telco has a valuable data set, which includes the data of customers who use DQLab Telco services. This valuable data set has a solid potential to support the company's various needs. Proper data processing and analysis can assist the company in drawing the correct conclusions [3]. The data processing and analysis findings can help the company take a stance or decision regarding its direction and policy. Reducing the number of customers who switch to

making the right decisions based on data, will positively impact the company [4].

Applying technology to data presents machine learning as a suitable technique for analyzing data. The ability of machine learning to perform modeling can be applied to study DQLab Telco customer data. Through this research, we propose churn prediction in the case of DQLab Telco using machine learning techniques. Machine learning techniques are applied to obtain predictions on whether users will switch to competitors or not based on data trained on machine learning models. By studying the data and making predictions, the company can determine its concrete steps in dealing with the possibility of churn [5].

Diverse data conditions require researchers to choose to manage data and select features carefully to obtain an optimal prediction model. Good feature selection will result in excellent modeling, which affects prediction accuracy [6]. There are various feature selection methods in machine learning modeling. Various feature selection methods can be compared to obtain optimal results. In this study, researchers used several feature selection methods such as ANOVA, recursive feature elimination (RFE), feature importance, Pearson correlation, and models without feature selection [7].

Chang et al. [8] proposed experimental results and a modeling procedure to analyze the factors affecting the defects of centrifugal pumps. Researchers collected various signals such as pressure, flow, motor current, and vibration and extracted features in the time and frequency domains. Researchers applied random forest modeling to perform multi-fault classification and suggested variables that significantly affected the anomaly detection of the pump. This paper also presented a correlation matrix of cosine similarity to analyze the experimental results of feature importance. This paper was supported by the National Research Foundation of Korea and the Korea Institute of Energy Technology Evaluation and Planning. In this research, Chang applied one machine learning algorithm, Random Forest. The best classification results can be obtained by comparing the results of several algorithms, so this research can still be improved by adding several algorithms as a comparison. This research also applies a feature selection method, Feature Importance. Model performance can be enhanced by selecting features and leaving the most informative and valuable features for modeling. Researchers can develop by applying several feature selection methods to obtain the best model performance based on several feature selection methods.

Pang et al. [9] reported a study investigating the independent risk factors affecting the prognosis of patients with bladder pain syndrome/interstitial cystitis (BPS/IC) after hydrodistension surgery. Researchers analyzed the clinical data of 1006 BPS/IC patients and identified age and the expression of CD117, P2X3R, NGF, and TrkA as independent prognostic factors. Researchers then developed a column chart and a random forest model to predict the clinical outcomes based on these factors. This study evaluated and validated the performance of the models using various statistical methods and found that they had good predictive accuracy and clinical benefits. They concluded that the models could help clinicians assess the risk and guide the treatment of BPS/IC patients. This research applies logistic regression and random forest algorithms to modeling. As a development, researchers can add additional algorithms to obtain the best accuracy from several algorithms. The researchers applied the Feature Importance method to perform feature selection. The application of several feature selection methods can be carried out as development so that research obtains the most optimal model performance.

Chen et al. [10] proposed a fault diagnosis method for rotating machinery based on improved multiscale attention entropy and random forests. The method used a nonlinear dynamics technique called multiscale attention entropy to measure the signal complexity at multiple time scales and a composite multiscale attention entropy to overcome the problem of insufficient coarse-graining. The method also used t-distributed stochastic neighbor embedding to reduce the dimensionality of the extracted features and random forests to classify the fault patterns. The method was tested on two fault datasets and an actual hydropower unit, achieving better diagnostic performance and adaptability than conventional methods. In this study, Chen used one machine learning algorithm to perform classification. This research can be developed by adding several algorithms for comparison. The addition of several algorithms as a comparison can show the best results that can be obtained from several algorithms used. In addition, this research can also add feature selection methods to obtain informative features to obtain the most optimal model performance.

Al-Haddad et al. [11] proposed an innovative approach to fault diagnosis in permanent magnet synchronous motors (PMSMs) by fusing vibration and current data. Researchers simulated stator faults as inter-turn short circuits and collected vibration and current signals from a PMSM test rig. Researchers extracted statistical features from the signals and used information gained for feature selection. This study employed a gradient-boosting-based machine learning model to classify different fault states using the selected features. This study achieved an impressive diagnostic accuracy of 90.7% and an area under the curve of 95.1%, demonstrating the efficacy of data fusion and gradient boosting for fault diagnosis. In this study, the researchers used Gradient Boosting, one of the machine learning algorithms for classification. Applying multiple algorithms in classification can increase the chances of obtaining classification results with the best results. In addition, the classification in this study can also be improved by adding feature selection methods so that the model's performance improves. We evaluated the performance of the model using accuracy metrics and ROC curves. Performance evaluation can be improved by adding additional metrics such as precision, recall, and f-1 score to obtain more diverse evaluation conclusions.

Nhat-Duc and Van-Duc [12] proposed and verified a computer vision-based method for automatically classifying raveled areas and their severity on asphalt pavement surfaces. The technique used gradient-boosting machines integrated with lightweight feature extractors, such as local binary patterns and their variants, to categorize visual data into three classes: non-raveling, minor raveling, and severe raveling. The method was tested on a dataset of 6600 image samples collected from field trips in Da Nang, Vietnam. The experimental results showed that the technique achieved high classification accuracy, low computational cost, and minimal

requirements on computer hardware. The method outperformed the state-of-the-art GoogleNet deep transfer learning model and demonstrated its potential as a tool to assist road maintenance agencies in performing raveling condition surveys. This research can be developed by adding several other classification algorithms to obtain the highest classification results among several algorithms. Development can also be done by adding feature selection methods to obtain informative features that optimize model performance. The evaluation in this research can be improved by adding several other evaluation metrics.

Palmese et al. [13] developed and internally validated a Logistic Regression model to predict the probability of 1-year readmission to the emergency department (ED) for acute alcohol intoxication (AAI). Palmese used a retrospective cohort of 3304 patients with AAI who were admitted to the ED of a hospital in Italy from 2005 to 2017. This study found that sex, age, homelessness, previous admission for trauma, and mental or behavioral disease were independent predictors of 1-year readmission for AAI. They suggested that their model be externally validated and tested in a randomized controlled trial to evaluate its clinical utility. However, in this study, researchers used one machine learning algorithm, Logistic Regression, which can be improved by adding several additional algorithms. This research also needs to perform optimization, such as adding feature selection methods to obtain informative features and optimize the model's performance in performing classification.

Kharsa and Al Aghbari [14] presented a method that used association rules to select features from medical data, which reduced the dimensionality of the input feature space. The selected features were then fed to a deep neural network, specifically ResNet, which performed classification tasks with high accuracy. The paper compared the proposed method with other traditional machine learning models on various medical datasets and showed that it outperformed them regarding classification accuracy. The paper also discussed the benefits of using association rules for feature selection and deep learning for classification. This research applies one of the feature selection methods, namely association rules. Several feature selection methods can be added to find the one that best suits the data and model. The appropriate feature selection method produces the best features that can be used in model training so that the model learns the most informative features and performs well.

Mahto et al. [15] proposed a hybrid method for cancer classification using gene expression data based on a novel combination of Cuckoo Search and Spider Monkey Optimization algorithms for feature selection and deep learning for classification. The paper evaluated the performance of the proposed method on eight benchmark microarray datasets and compared it with other existing methods. The paper reported that the proposed method achieved higher or competitive accuracy, reduced overfitting, and enhanced model performance. The paper also discussed the proposed method's advantages, limitations, and future directions. This research applies the Spider Monkey Optimization algorithm as a feature selection method. Feature selection methods can be added to obtain the best feature selection method for better model performance.

Machine learning has various algorithms that can be used to make predictions. Applying the correct algorithm affects the model's performance in making predictions. Good data requires an excellent algorithm to get good performance. Researchers applied several algorithms in this study to compare and get the best performance prediction. Some of the algorithms that researchers use in this study are Logistic Regression, Random Forest, and Gradient Boosting [16].

This paper is organized as follows: Section I introduces the problem and proposes solutions in the application of machine learning. Section II presents works related to the application of machine learning. Section III describes the proposed Exploratory Data Analysis (EDA) Data Pre-processing, Classification Modeling, and Performance Evaluation. Section IV presents a visual analysis of EDA and a performance analysis of logistic Regression, random forest, and gradient boosting algorithms. Finally, this research study is concluded in Section V.

II. Materials and Method

The method proposed in this research consists of four stages: Exploratory Data Analysis (EDA), Data Pre-processing, Classification Modeling, and Performance Evaluation. The dataset used in this research is DQLab Telco's customer data. This dataset has 11 columns and 6590 rows. The research flow diagram is presented in Fig. 1, and the sample dataset is presented in Table I

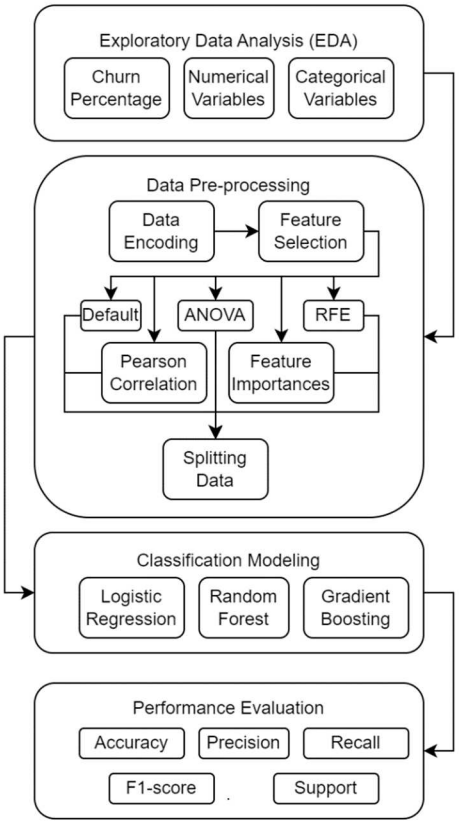

Fig. 1 Research flow diagram

The sample dataset shows some customer data that can potentially affect customer churn. Some data shows several things about the customer, such as their gender, whether they are a senior citizen, whether they have a spouse, how long

their tenure is, whether they use phone service, TV streaming services, and others. This data would be used to predict customer churn so that it can predict whether the customer could switch subscriptions or not. The top 5 data samples from the dataset are shown in Table I, which are taken in order starting from the topmost data in the dataset.

TABLE I
SAMPLE OF DATASET

Column	Data *n*th				
	D1	**D2**	**D3**	**D4**	**D5**
Gender	Female	Male	Male	Female	Female
SeniorCitizen	No	No	No	No	No
Partner	Yes	Yes	No	Yes	Yes
Tenure	1	60	5	72	56
PhoneService	No	Yes	Yes	Yes	Yes
StreamingTV	No	No	Yes	Yes	Yes
InternetService	Yes	No	Yes	Yes	Yes
PaperlessBilling	Yes	Yes	No	Yes	No
MunthlyCharges	29.85	20.50	104.10	115.50	81.25
TotalCharges	29.85	1198.80	541.90	8312.75	4620.40
Churn	No	No	Yes	No	No

The research begins with visual analysis at the Exploratory Data Analysis (EDA) stage, and then the data is further processed at the data pre-processing stage [17]. The processed data is then trained using several machine-learning algorithms. Researchers used accuracy, precision, recall, F1-score, and support metrics to evaluate the model's performance.

A. Exploratory Data Analysis (EDA)

Researchers explored the data at this stage to obtain an initial dataset analysis. The analysis is done by visualizing the data as needed. Researchers visualize several parts of the data: the percentage of customer churn, numeric variables, and categorical variables. Researchers visualize using the pyplot module in the Matplotlib library, which is commonly used to visualize data. In addition, researchers also use the NumPy module to manipulate numeric data, such as operations or transformations performed on data [18].

Researchers explored the data by conducting univariate and bivariate data visualization. Univariate visualization aims to determine the percentage of customer churn in the DQLab Telco company. Bivariate visualization is conducted between numeric and categorical variables with churn variables. Bivariate visualization of churn variables aims to determine how much the variables in the dataset affect customer churn. Variables that affect customer churn will be a focus for companies in improving customer experience. By understanding the influence of supporting variables on customer churn, companies can make the right decisions according to the conclusions obtained [19].

B. Data Pre-processing

At this stage, researchers apply several methods to the data so that the data is ready to be used in modeling. The methods include removing unnecessary columns, encoding data, selecting features, and splitting data. These four methods were applied sequentially to obtain good data. In the first stage, researchers removed unnecessary columns. The columns in the dataset contained data that could be a factor in customer churn. However, some columns only served as data identity and did not affect customer churn. Data that had no potential influence on customer churn was removed and not used in modeling [20].

The next step performed on the data is data encoding. Data that can be processed in modeling is numeric because calculations are carried out, which requires data in numeric format. Then, the data with a type other than numeric must be converted to numeric. Researchers use the LabelEncoder module in the sklearn library to convert data of types other than strings into string types [21].

Not all features are used in modeling. Some selected features are used in modeling, while the rest are ignored. Researchers applied several feature selection methods at the data pre-processing stage, such as ANOVA, Recursive Feature Elimination (RFE), Feature Importance using the Random Forest model, Pearson Correlation, and data without feature selection. Feature selection is used to select a subset of the features available in the dataset for use in modeling. Feature selection only alleviates model performance by focusing on informative features [22]. In this research, applying different feature selection methods to the dataset results in different dataset feature diversity. Various datasets with different feature diversity will produce classification models with different performance.

The Analysis of Variance (ANOVA) technique is used to compare between two or more data groups so that researchers can find out if there is a significant difference between at least one pair of group means. Recursive Feature Elimination (RFE) eliminates less important features iteratively until only the best features remain. Considering the contribution of each feature individually results in a model that is more efficient and effective in predicting churn behavior. Feature Importance using the Random Forest model evaluates the importance of each feature in churn prediction. When the model is trained, each feature is assigned a critical value based on how often or how much it affects the quality of the prediction. Pearson Correlation measures the strength and direction of a linear relationship between two variables. It provides information about the direction and strength of the relationship between two numerical variables.

Data pre-processing ends at data splitting. Data splitting divides data into training and testing data with a certain proportion. The training data contains 70% of the dataset, while the testing data comprises 30%. The division is done randomly using the train_test_split module from the sklearn library. Generally, the proportion of training data is more significant than the testing data. The more considerable amount of training data compared to testing data is related to the model's recognition of the data. The more data trained on the model, the better the model recognizes the data. Conversely, the less data trained on the model, the worse the model recognizes the data. The better the model acknowledges the data, the better the model predicts the testing data [23].

C. Classification Modelling

In this stage, modeling is done by applying machine learning algorithms to the data and storing them in variables. Training the model using data that has been labeled gives the model the ability to recognize data patterns. Diverse data increases the ability of the model to make predictions on a variety of data in the testing data. The condition of the data

also affects the results of model training. Excellent data produces models with good performance in making predictions. Meanwhile, insufficient data produces models with poor prediction performance [24].

Balanced data is required to obtain optimal predictions for each class. Unbalanced data will recognize the dominant class more than other classes. As a result of data imbalance, the model will predict more data as the dominant class. In addition to the amount of data, the balance of the numerical scale of the data is also required. When the numerical scale of the data is unbalanced, the model will give more importance to features with a larger numerical scale than data with a small numerical scale. Due to the unbalanced numerical scale in the data, the model will make predictions based on the model's assessment of features with a larger scale [25].

The correct algorithm also affects the performance of the model. Algorithms suitable for specific data types will produce models with better performance than other algorithms. Combining a prime dataset and a suitable algorithm will ensure optimal model performance. In addition to data conditions and algorithm suitability, feature selection also affects model performance. Features that are informative to the model result in better classification [26].

Logistic Regression is a statistical model used to predict the probability of two or more classes. In the context of churn prediction, Logistic Regression generates the probability that a customer will churn or not churn. This algorithm has several important parameters such as regularization (model complexity), penalty, and C parameter (regularization strength).

Random Forest consists of many decision trees that work independently and generate predictions. In the case of churn prediction, Random Forest combines predictions from many decision trees to improve accuracy and reduce overfitting. The algorithm has some important parameters such as the number of decision trees (n_estimators), maximum depth (max_depth) and features considered (max_features).

Gradient Boosting builds a predictive model through a series of alternating decision trees. Each decision tree compensates for the prediction error of the previous tree. In the case of churn prediction, Gradient Boosting iteratively corrects previous prediction errors to improve the model's accuracy. The algorithm has several essential parameters, such as the number of decision trees (n_estimators), maximum depth (max_depth), and learning rate.

D. Performance Evaluation

Various evaluation metrics can assess model performance. The confusion matrix displays the result of the model's prediction against the test data. The confusion matrix compares the original label's value and the predicted label's value. Thus, through the confusion matrix, researchers can observe the amount of data correctly and incorrectly classified in a particular class. Some of the matrices used as performance evaluation in this research are accuracy, precision, recall, F1-score, and support.

The precision, recall, F1-score, and accuracy formulas are presented in the following formula [27]–[30]:

$$precision = \frac{TP}{TP + FP} \tag{1}$$

$$recall = \frac{TP}{TP + FN} \tag{2}$$

$$F1 = \frac{2 \times precision \times recall}{precision + recall} \tag{3}$$

$$accuracy = \frac{TP + TN}{TP + FN + TN + FP} \tag{4}$$

III. Results and Discussion

This section presents the results and analysis of the application of the proposed method. This section presents the researcher's analysis of data processing both visually and through machine learning algorithm modeling. Fundamental data analysis is required to understand the dataset used in modeling—visual observation of the data results in assessing the correlation level of a feature to the predictor feature. The correlation level of a feature can be a basic description of various features that can be used in modeling.

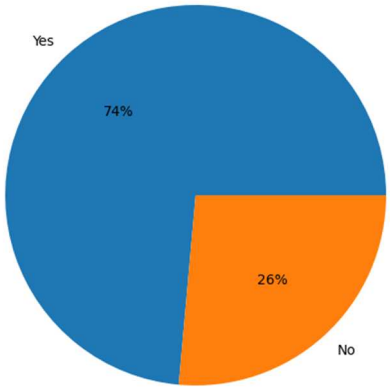

Fig. 2 Churn percentage

After obtaining a basic overview of the correlation of features to predictor features, researchers conducted further feature selection using several methods. Several feature selection methods were compared and applied to various modeling algorithms to obtain the best feature selection method and algorithm. Various feature selections produce different amounts of data in training data and training data. The different features used also affect the performance of the model in classification. The number of features used in the default feature selection method (without feature selection), ANOVA, Recursive Feature Elimination (RFE), Feature Importance (FI), and Pearson Correlation are 10, 7, 5, 9, and 7, respectively. In modeling, the algorithms used are Logistic Regression, Random Forest, and Gradient Boosting. In all algorithms, we compared several performance evaluation metrics such as accuracy, precision in class 0, precision in class 1, recall in class 0, recall in class 1, F1-score in class 0, F1-score in class 1, support in class 0, and support in class 1.

A. Correlation Analysis

In the first analysis, researchers visualized the percentage of customer churn to determine how significant the customer churn is at DQLab Telco. Based on the distribution of customer churn data in Fig. 2, overall, customers do not churn, with a percentage of no churn more significant than churn.

This shows that the tendency of customers to churn is smaller than no churn. The slight tendency of churn is undoubtedly not something that can be ignored. This remains a focus for companies to reduce the percentage of customer churn. So, further analysis is needed to understand the patterns between customer churn and its supporting factors.

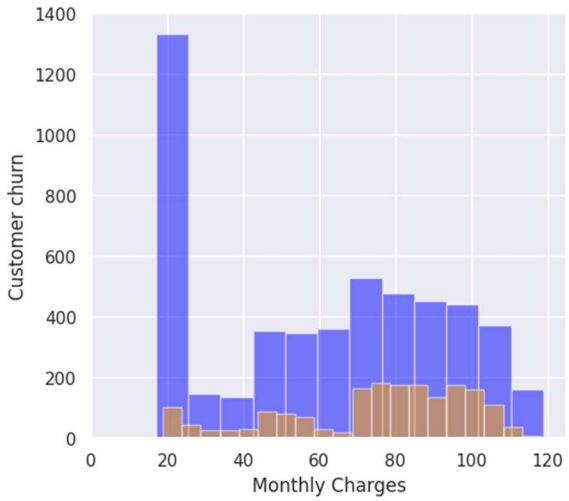

Fig. 3 Correlation between monthly charges and customer churn

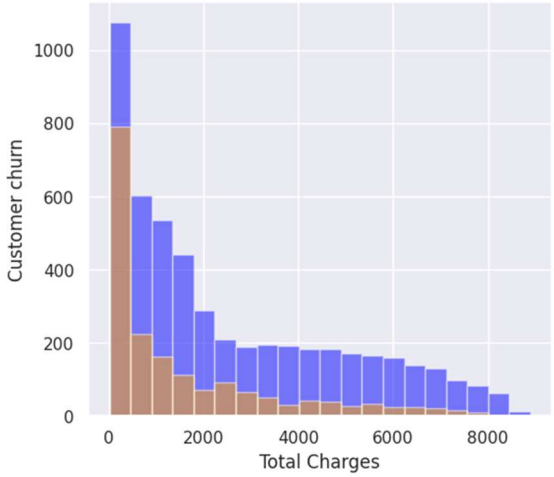

Fig. 4 Correlation between total charges and customer churn

Fig. 3 shows the relationship between monthly charges and customer churn. Based on the visualization, there is a tendency that the smaller the value of monthly charges charged, the smaller the propensity to churn. This shows that monthly charges have a direct effect on customer churn. Thus, company policymaking related to customer churn can use monthly charges as a reference. Fig. 4 shows the relationship between total charges and customer churn. Based on the visualization, there is no significant trend between the value of monthly charges charged and the tendency of customers to churn. This shows that monthly charges do not affect customer churn. Thus, total charges are not a good reference for companies in making company policies related to customer churn.

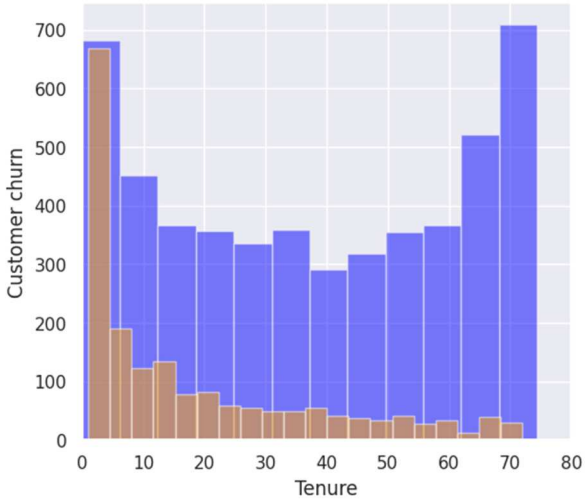

Fig. 5 Correlation between tenure and customer churn

Fig. 5 presents a visualization of the relationship between tenure and customer churn. Based on the visualization, there is a tendency that the longer the customer subscribes, the smaller the tendency to churn. This shows that tenure has a direct effect on customer churn. Thus, company policymaking related to customer churn can use tenure as a reference.

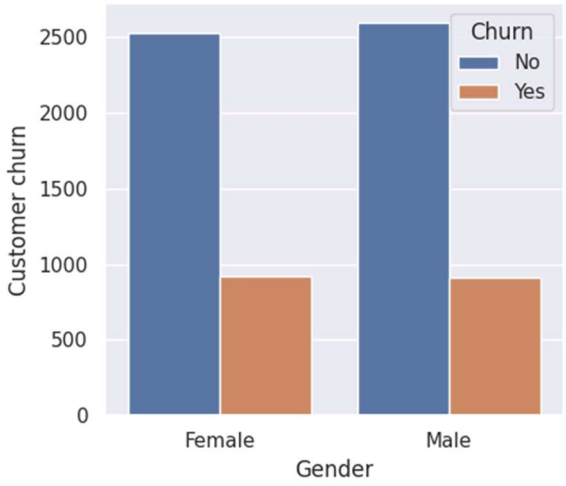

Fig. 6 Correlation between gender and customer churn

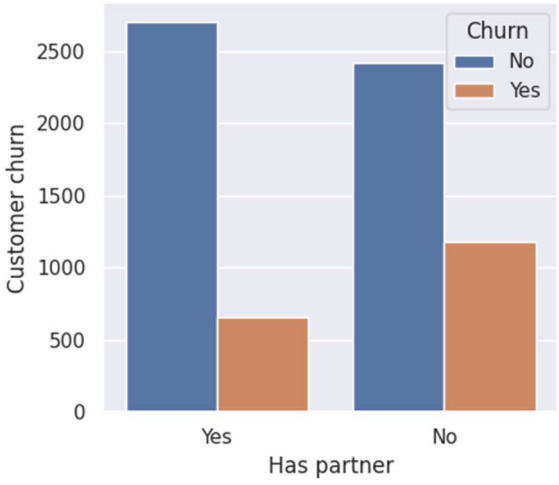

Fig. 7 Correlation between partner and customer churn

Fig. 6 presents a visualization of the relationship between gender and customer churn. Based on the visualization, there is no significant trend between gender and the tendency of customers to churn. This shows that gender does not affect customer churn. Thus, gender is not a good reference for companies in terms of customer churn policies. Fig. 7 presents a visualization of the relationship between partners and customer churn. Based on this visualization, customers who do not have a partner tend to churn. This shows that partners have a direct effect on customer churn. Thus, company policymaking related to customer churn can use partners as a reference.

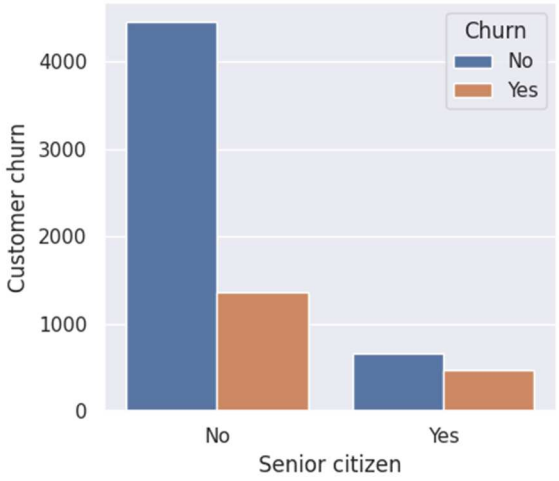

Fig. 8 Correlation between senior citizen and customer churn

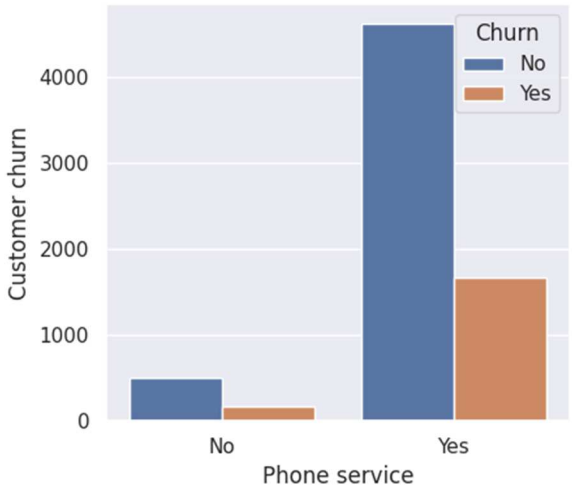

Fig. 9 Correlation between phone service and customer churn

Fig. 8 shows the relationship between senior citizens and customer churn. Based on this visualization, customers who are senior citizens tend to churn. This indicates that senior citizens have a direct effect on customer churn. Thus, company policymaking related to customer churn can use senior citizens as a reference. Fig. 9 shows the relationship between phone service and customer churn. Based on the visualization, there is no significant trend between phone service and the tendency of customers to churn. This shows that phone service does not affect customer churn. Thus, phone service is not a good reference for companies when making company policies related to customer churn.

Fig. 10 shows the relationship between TV streaming and customer churn. Based on this visualization, customers with TV services tend to churn. This shows that TV streaming has a direct effect on customer churn. Thus, company policymaking related to customer churn can use TV streaming as a reference.

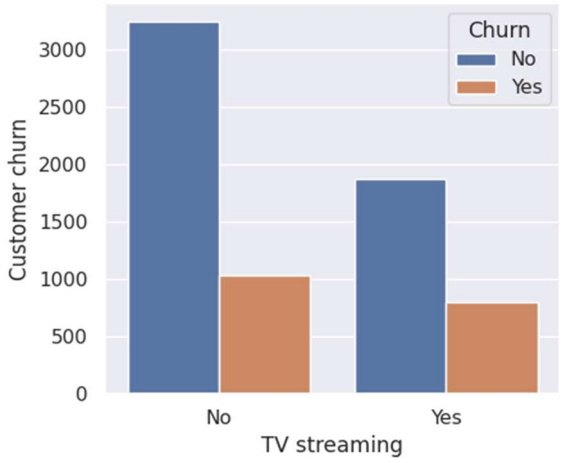

Fig. 10 Correlation between TV streaming and customer churn

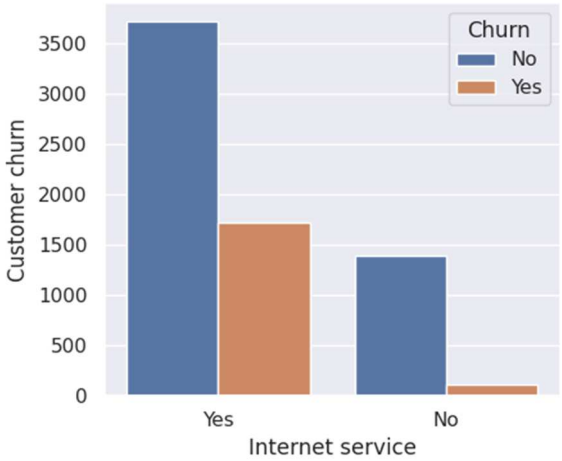

Fig. 11 Correlation between Internet service and customer churn

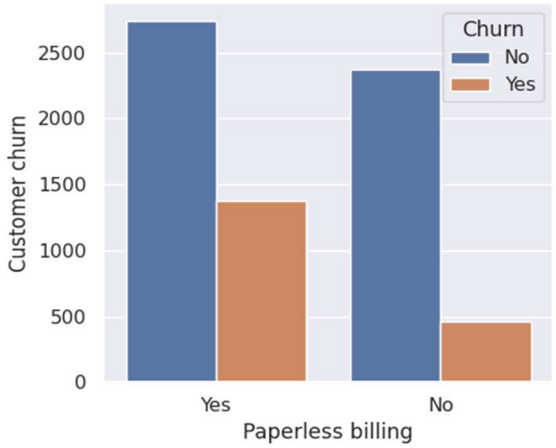

Fig. 12 Correlation between paperless billing and customer churn

Fig. 11 shows the relationship between internet service and customer churn. Based on this visualization, customers with internet service tend to churn. This shows that internet service

has a direct effect on customer churn. Thus, company policymaking related to customer churn can use Internet service as a reference. Fig. 12 shows the relationship between paperless billing and customer churn. Based on this visualization, customers with paperless bills tend to churn. This shows that paperless billing has a direct effect on customer churn. Thus, company policymaking related to customer churn can use paperless billing as a reference.

B. Logistic Regression

Logistic Regression is a practical, simple, and easy-to-interpret classification algorithm. Logistic Regression makes it easy to understand the contribution of each variable to the outcome, as it produces probabilities as output and coefficients that are easy to interpret. In addition, Logistic Regression is efficient for datasets with a small number of variables. Table II presents a comparison of Logistic Regression performance evaluation on several feature selections, namely Default, ANOVA, Recursive Feature Elimination (RFE), Feature Importance (FI), and Pearson Correlation (PC).

TABLE II
LOGISTIC REGRESSION PERFORMANCE EVALUATION

Metric	Default	ANOVA	RFE	FI	PC
Acc.	**79.47**	78.90	78.32	78.90	78.90
Prec. 0	**83.28**	82.84	82.60	82.88	82.84
Rec. 0	**90.32**	90.06	89.47	89.99	90.06
F1-0	**86.66**	86.30	85.90	86.29	86.30
Supp. 0	1539	1539	1539	1539	1539
Prec. 1	**64.18**	62.86	61.24	62.80	62.86
Rec. 1	**48.90**	47.43	46.89	47.61	47.43
F1-1	**55.51**	54.07	53.11	64.17	54.07
Supp. 1	546	546	546	546	546

Based on Table II, the overall performance evaluation obtained the highest value without using feature selection. This explains that the feature selection method does not significantly affect the Logistic Regression model; instead, the data without feature selection obtained the best evaluation results. Different types and characteristics of data allow different data pre-processing needs. The diversity and characteristics of various data make it necessary for research to apply several other methods for comparison, such as feature selection.

C. Random Forest

Random Forest excels with its ability to produce accurate predictions. An exciting part of this algorithm is using multiple decision trees working in tandem. Using multiple decision trees helps Random Forest overcome the problem of overfitting, one of the challenges in complex modeling. In addition, the algorithm can manage each feature's importance and provide additional insight into how each variable contributes to the decision. The following table compares Random Forest performance evaluation on several feature selection features: Default, ANOVA, Recursive Feature Elimination (RFE), Feature Importance (FI), and Pearson Correlation. Table III shows the performance evaluation of Random Forest.

Based on Table III, the overall performance evaluation obtained the highest value in the Feature Importance (FI) method. The feature importance selection method using Random Forest obtained the best results on classification using the Random Forest algorithm. This shows that the classification algorithm and related feature selection methods produce good performance in classification.

TABLE III
RANDOM FOREST PERFORMANCE EVALUATION

Metric	Default	ANOVA	RFE	FI	PC
Acc.	77.36	76.74	76.88	**77.60**	76.59
Prec. 0	82.55	82.06	82.28	**82.80**	82.14
Rec. 0	**87.91**	87.65	87.52	**87.91**	87.26
F1-0	85.15	84.76	84.82	**85.28**	84.62
Supp. 0	1539	1539	1539	1539	1539
Prec. 1	58.30	56.92	57.14	**58.76**	56.44
Rec. 1	47.62	45.97	46.89	**48.53**	46.52
F1-1	52.42	50.86	51.51	**53.16**	51.00
Supp. 1	546	546	546	546	546

D. Gradient Boosting

Gradient Boosting takes a different approach by correcting errors in the previous model. The advantages of Gradient Boosting are its high accuracy and ability to handle complex datasets. It applies a gradient-based technique, optimizing models faster than other approaches. Gradient Boosting is quite resilient to outliers as it can give smaller weights to misclassified data, making it a good choice for datasets with anomalies. The following table compares Gradient Boosting's performance evaluation on several feature selections: Default, ANOVA, Recursive Feature Elimination (RFE), Feature Importance (FI), and Pearson Correlation. Table IV shows the performance evaluation of Gradient Boosting.

TABLE IV
GRADIENT BOOSTING PERFORMANCE EVALUATION

Metric	Default	ANOVA	RFE	FI	PC
Acc.	79.38	78.90	**79.86**	79.28	78.90
Prec. 0	83.03	82.65	**83.13**	82.85	82.65
Rec. 0	90.58	90.38	**91.23**	90.71	90.38
F1-0	86.64	86.34	**86.99**	86.60	86.34
Supp. 0	1539	1539	1539	1539	1539
Prec. 1	64.29	63.18	**65.91**	64.25	63.18
Rec. 1	**47.80**	46.52	**47.80**	47.07	46.52
F1-1	54.83	53.59	**55.41**	54.33	53.59
Supp. 1	546	546	546	546	546

Based on Table IV, the overall performance evaluation obtained the highest value in the Recursive Feature Elimination (RFE) method. This shows how feature selection affects the model's performance in training data. The selection of more informative data makes the model better trained, thus obtaining more optimal results.

IV. CONCLUSION

This study found that proper feature selection and machine learning algorithms play a vital role in improving the accuracy of customer churn prediction at DQLab Telco. Logistic Regression without feature selection produced the highest accuracy of 79.47%. In comparison, Random Forest with Feature Importance achieved an accuracy of 77.60%, and Gradient Boosting with Recursive Feature Elimination (RFE) attained an accuracy of 79.86%. These results show that not all features contribute equally to churn prediction, and each algorithm has strengths in specific data contexts. These findings provide important insights for DQLab Telco in

customer churn reduction strategies by leveraging customized machine learning techniques to predict and reduce the chances of customers churning to competitors.

REFERENCES

[1] K. Prasad, A. S. Tomar, T. De, and H. Soni, "A Conceptual Model for Building the Relationship Between Augmented Reality, Experiential Marketing & Brand Equity," International Journal of Professional Business Review, vol. 7, no. 6, p. e01030, Dec. 2022, doi:10.26668/businessreview/2022.v7i6.1030.

[2] R. F. A. Aziza, A. Aminuddin, A. N. Widianingsih, and D. I. S. Saputra, "User Experience Analysis of Student Assistant Application Using The Five Planes Method," *2023 7th Int. Conf. New Media Stud.*, 2023.

[3] Z. Mustaffa, M. H. Sulaiman, D. Rohidin, F. Ernawan, and S. Kasim, "Time Series Predictive Analysis based on Hybridization of Meta-heuristic Algorithms," International Journal on Advanced Science, Engineering and Information Technology, vol. 8, no. 5, pp. 1919–1925, Oct. 2018, doi: 10.18517/ijaseit.8.5.4968.

[4] E. Y. Sari, A. D. Wierfi, and A. Setyanto, "Sentiment Analysis of Customer Satisfaction on Transportation Network Company Using Naive Bayes Classifier," 2019 International Conference on Computer Engineering, Network, and Intelligent Multimedia (CENIM), Nov. 2019, doi: 10.1109/cenim48368.2019.8973262.

[5] B. P. Asadduloh, A. Aminuddin, M. Rahardi, F. F. Abdulloh, A. Yaqin, and M. I. Hasani, "Machine Learning Techniques to Predict Rain Tomorrow for Automated Plant Watering System," *2023 1st Int. Conf. Adv. Eng. Technol.*, 2023.

[6] A. Yaqin, M. Rahardi, and F. F. Abdulloh, "Accuracy Enhancement of Prediction Method using SMOTE for Early Prediction Student's Graduation in XYZ University," International Journal of Advanced Computer Science and Applications, vol. 13, no. 6, 2022, doi:10.14569/ijacsa.2022.0130652.

[7] N. Sholihah, F. F. Abdulloh, M. Rahardi, A. Aminuddin, B. P. Asadduloh, and A. Y. A. Nugraha, "Feature Selection Optimization for Sentiment Analysis of Tax Policy Using SMOTE and PSO," *2023 3rd Int. Conf. Smart Cities, Autom. Intell. Comput. Syst.*, 2023.

[8] K. Chang and S. H. Park, "Random forest-based multi-faults classification modeling and analysis for intelligent centrifugal pump system," Journal of Mechanical Science and Technology, vol. 38, no. 1, pp. 11–20, Dec. 2023, doi: 10.1007/s12206-023-1202-2.

[9] L. Pang, Z. Ding, H. Chai, and W. Shuang, "Construction and evaluation of a column chart model and a random forest model for predicting the prognosis of hydrodistention surgery in BPS/IC patients based on preoperative CD117, P2X3R, NGF, and TrkA levels," BMC Medical Informatics and Decision Making, vol. 23, no. 1, Dec. 2023, doi: 10.1186/s12911-023-02396-w.

[10] F. Chen et al., "A fault diagnosis method of rotating machinery based on improved multiscale attention entropy and random forests," Nonlinear Dynamics, vol. 112, no. 2, pp. 1191–1220, Dec. 2023, doi:10.1007/s11071-023-09126-x.

[11] L. A. Al-Haddad, A. A. Jaber, M. N. Hamzah, and M. A. Fayad, "Vibration-current data fusion and gradient boosting classifier for enhanced stator fault diagnosis in three-phase permanent magnet synchronous motors," Electrical Engineering, Dec. 2023, doi:10.1007/s00202-023-02148-z.

[12] H. Nhat-Duc and T. Van-Duc, "Computer Vision-Based Severity Classification of Asphalt Pavement Raveling Using Advanced Gradient Boosting Machines and Lightweight Texture Descriptors," Iranian Journal of Science and Technology, Transactions of Civil Engineering, vol. 47, no. 6, pp. 4059–4073, Jun. 2023, doi:10.1007/s40996-023-01138-2

[13] F. Palmese et al., "Development and internal validation of a multivariable model for the prediction of the probability of 1-year readmission to the emergency department for acute alcohol intoxication," Internal and Emergency Medicine, vol. 19, no. 3, pp. 823–829, Dec. 2023, doi: 10.1007/s11739-023-03490-7.

[14] R. Kharsa and Z. Al Aghbari, "Leveraging Association Rules in Feature Selection for Deep Learning Classification," SN Computer Science, vol. 5, no. 1, Dec. 2023, doi: 10.1007/s42979-023-02397-6.

[15] R. Mahto et al., "A novel and innovative cancer classification framework through a consecutive utilization of hybrid feature selection," BMC Bioinformatics, vol. 24, no. 1, Dec. 2023, doi:10.1186/s12859-023-05605-5.

[16] M. I. Akbar, A. Aminuddin, F. F. Abdulloh, M. Rahardi, S. N. Wahyuni, and B. P. Asadduloh, "Comparison of Machine Learning Techniques for Heart Disease Diagnosis and Prediction," *2023 Int. Conf. Adv. Mechatronics, Intell. Manuf. Ind. Autom.*, 2023.

[17] M. M. Hamed, M. G. Khalafallah, and E. A. Hassanien, "Prediction of wastewater treatment plant performance using artificial neural networks," Environmental Modelling & Software, vol. 19, no. 10, pp. 919–928, Oct. 2004, doi: 10.1016/j.envsoft.2003.10.005

[18] F. Wunderlich and D. Memmert, "A big data analysis of Twitter data during premier league matches: do tweets contain information valuable for in-play forecasting of goals in football?," Social Network Analysis and Mining, vol. 12, no. 1, Dec. 2021, doi: 10.1007/s13278-021-00842-z.

[19] M. W. Ahmad, M. Mourshed, and Y. Rezgui, "Tree-based ensemble methods for predicting PV power generation and their comparison with support vector regression," Energy, vol. 164, pp. 465–474, Dec. 2018, doi: 10.1016/j.energy.2018.08.207.

[20] K. Maswadi, N. A. Ghani, S. Hamid, and M. B. Rasheed, "Human activity classification using Decision Tree and Naïve Bayes classifiers," *Multimed. Tools Appl.*, vol. 80, no. 14, pp. 21709–21726, Jun. 2021, doi: 10.1007/S11042-020-10447-X/tables/3.

[21] H. Zakiyyah and S. Suyanto, "Prediction of Covid-19 Infection in Indonesia Using Machine Learning Methods," Journal of Physics: Conference Series, vol. 1844, no. 1, p. 012002, Mar. 2021, doi:10.1088/1742-6596/1844/1/012002.

[22] R. Nair and A. Bhagat, "Feature Selection Method To Improve The Accuracy of Classification Algorithm," *Int. J. Innov. Technol. Explor. Eng.*, vol. 8, no. 6, pp. 124–127, 2019.

[23] A. M. Rahat, A. Kahir, and A. K. M. Masum, "Comparison of Naive Bayes and SVM Algorithm based on Sentiment Analysis Using Review Dataset," 2019 8th International Conference System Modeling and Advancement in Research Trends (SMART), Nov. 2019, doi:10.1109/smart46866.2019.9117512.

[24] K. B. Newhart, R. W. Holloway, A. S. Hering, and T. Y. Cath, "Data-driven performance analyses of wastewater treatment plants: A review," Water Research, vol. 157, pp. 498–513, Jun. 2019, doi:10.1016/j.watres.2019.03.030.

[25] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," Journal of Artificial Intelligence Research, vol. 16, pp. 321–357, Jun. 2002, doi:10.1613/jair.953.

[26] N. M. Nawi, W. H. Atomi, and M. Z. Rehman, "The Effect of Data Pre-processing on Optimized Training of Artificial Neural Networks," Procedia Technology, vol. 11, pp. 32–39, 2013, doi:10.1016/j.protcy.2013.12.159.

[27] J. Singh and P. Tripathi, "Sentiment analysis of Twitter data by making use of SVM, Random Forest and Decision Tree algorithm," 2021 10th IEEE International Conference on Communication Systems and Network Technologies (CSNT), Jun. 2021, doi:10.1109/csnt51715.2021.9509679.

[28] D. A. Anggoro and D. Permatasari, "Performance Comparison of the Kernels of Support Vector Machine Algorithm for Diabetes Mellitus Classification," International Journal of Advanced Computer Science and Applications, vol. 14, no. 1, 2023, doi:10.14569/ijacsa.2023.0140163.

[29] B. B, A. Pal, and P. Muruganandam, "https://www.ijrte.org/portfolio-item/C5403098319/," International Journal of Recent Technology and Engineering (IJRTE), vol. 8, no. 3, pp. 3236–3242, Sep. 2019, doi:10.35940/ijrte.c5404.098319.

[30] Y. B. P. Pamukti and M. Rahardi, "Sentiment Analysis of Bandung Tourist Destination Using Support Vector Machine and Naïve Bayes Algorithm," 2022 6th International Conference on Information Technology, Information Systems and Electrical Engineering (ICITISEE), Dec. 2022, doi: 10.1109/icitisee57756.2022.10057802.