ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/357302836

# Classification methods comparison for customer churn prediction in the telecommunication industry

Article *in* International Journal of ADVANCED AND APPLIED SCIENCES · November 2021

DOI: 10.21833/ijaas.2021.12.001

CITATIONS
10

READS
837

**5 authors**, including:

Arif Bramantoro
Institut Teknologi Brunei
**56** PUBLICATIONS **305** CITATIONS

SEE PROFILE

Hasan J. Alyamani
King Abdulaziz University
**24** PUBLICATIONS **330** CITATIONS

SEE PROFILE

Ryan Alturki
Umm Al-Qura University
**64** PUBLICATIONS **822** CITATIONS

SEE PROFILE

All content following this page was uploaded by Ryan Alturki on 24 December 2021.

The user has requested enhancement of the downloaded file.

Contents lists available at Science-Gate

International Journal of Advanced and Applied Sciences

Journal homepage: http://www.science-gate.com/IJAAS.html

# Classification methods comparison for customer churn prediction in the telecommunication industry

Moh Makruf [1, *], Arif Bramantoro [2], Hasan J. Alyamani [3], Sami Alesawi [3], Ryan Alturki [4]

[1]*Faculty of Information Technology, Universitas Budi Luhur, Jakarta, Indonesia*
[2]*School of Computing and Informatics, Universiti Teknologi Brunei, Bandar Seri Begawan, Brunei*
[3]*Faculty of Computing and Information Technology in Rabigh, King Abdulaziz University, Jeddah, Saudi Arabia*
[4]*Department of Information Science, College of Computer and Information Systems, Umm Al-Qura University, Makkah, Saudi Arabia*

ARTICLE INFO

*Article history:*
Received 16 April 2021
Received in revised form
13 July 2021
Accepted 9 September 2021

*Keywords:*
Customer churn
Decision tree
Artificial neural network
Gaussian Naïve Bayes
Support vector machine
K-nearest neighbor

ABSTRACT

The need for telecommunication services has increased dramatically in schools, offices, entertainment, and other areas. On the other hand, the competition between telecommunication companies is getting tougher. Customer churn is one of the areas that each company gains more competitive advantage. This paper proposes a comparison of several classification methods to make a prediction whether the customers cancel the subscription to a telecommunication service by highlighting key factors of customer churn or not. The comparison is non-trivial due to the urgent requirements from the telecommunication industry to infer the most appropriate techniques in analyzing their customer churn. This comparison is often of huge commercial value. The result shows that Artificial Neural Network (ANN) can predict churn with an accuracy of 79%, Support Vector Machine (SVM) with 78% accuracy, Gaussian Naïve Bayes, and K-Nearest Neighbor (KNN) with 75% accuracy, while Decision Tree with 70% accuracy. Moreover, the technique with the highest F-Measure is Gaussian Naïve Bayes with 65% and the technique with the lowest one is Decision Tree with 49%. Hence, ANN and Gaussian Naïve Bayes are two methods with high recommendation to predict the customer churn in the telecommunication industry.

© 2021 The Authors. Published by IASE. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

## 1. Introduction

Over the past few decades, telecommunication service has gained their potential use dramatically in various areas, such as education, workplace, and entertainment. This leads to fierce competition amongst telecommunication companies. Detecting customers who are likely to cancel a subscription to the telecommunication service is important because it is more costlier to obtain a new customer than to retain an existing one and prevent it from leaving the service. It can also solve a business problem by identifying which customers are loyal and which ones are at risk of churning. The companies eventually prevent a churn with targeted actions, and, in the end, the number of churned customers can be reduced.

Several well-known methods used for prediction are Decision Tree, Support Vector Machine (SVM), Artificial Neural Network (ANN), Gaussian Naïve Bayes, and K-Nearest Neighbor (KNN). To the best of our knowledge, these five techniques are the most significant ones to analyze customer churn. Hence, we propose to evaluate these five techniques to find out the best one for customer churn in the telecommunication industry. The comparison is necessary because the telecommunication industry requires further knowledge on the most appropriate techniques in analyzing their customer churn. This comparison is also considered a key business value. Any errors in the analysis will ripple down throughout the business process of the telecommunication company.

To improve the flexibility and visualization of analysis, python is chosen as a modeling and implementation language due to the wide availability of its libraries. The accuracy, precision, and recall are calculated for the churn prediction model using a

* Corresponding Author.
Email Address: arif.bramantoro@utb.edu.bn (M. Makruf)

https://doi.org/10.21833/ijaas.2021.12.001
Corresponding author's ORCID profile:
https://orcid.org/0000-0003-2772-9427
2313-626X/© 2021 The Authors. Published by IASE.

This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/)

confusion matrix a comprehensive visualization to have a better understanding of the result as well as making the point as clearly and concisely as possible to many audiences.

## 2. Literature review

In previous research, customer churn in the telecommunication industry was analyzed with various techniques. Khan et al. (2019) presented customer churn prediction using Artificial Neural Network (ANN) in the telecommunication industry. It focuses on several churn factors and necessary steps to eliminate the reason that motivates customer churn in Pakistan where the market is claimed to be the most volatile in the world. The analysis conducted by using ANN provides an accuracy of 79% which is considered superb compared to other methods. However, there is no quantitative study provided to support this assumption. Moreover, it remains unclear which methods are considered inferior compared to ANN.

ApurvaSree et al. (2019) predicted customer churn possibility in the telecommunication industry by means of several classification methods, such as Random Forest, Logistics Regression, and SVM. It is interesting to note that the paper provided detailed pseudocodes for executing each method. Based on the thorough observation, the highest accuracy is achieved by SVM with 82%. It would be more interesting to see other evaluation metrics to compare the methods, such as recall and precision. Moreover, the visualization is only provided for the data understanding, while there is no visualization for the performance evaluation. There is no further explanation for the result of performance evaluation in this paper, hence, it is rather difficult to understand the discussion of the performance amongst these three techniques.

Dahiya and Bhatia (2015) presented customer churn analysis in the telecommunication industry by means of Decision Tree and Logistic Regression. It analyzed the selection of the most appropriate grouping of existing attributes as well as the setting of the proper threshold values that could give more accurate results. To the best of our knowledge, this work is the first one comparing different techniques of classification for the sake of customer churn prediction in telecommunication industry. However, it is important to investigate not only Decision Tree and Logistic Regression, but also other techniques that might inherit a better performance and a novel insight.

Jain et al. (2020) presented churn prediction for the telecommunication providers by combining Logistic Regression and Logit Boost provided by the Weka tool (Russell and Markov, 2017). The experimental results confidently show that both techniques have relatively equal performance. Logit Boost had an accuracy of 85.1785%, on the other hand, Logistic Regression had only 0.06%% more. It is interesting to note that Logit Boost is relatively new for predicting customer churn in the telecommunication industry, but the performance is quite promising. This paper also provided a comprehensive evaluation using ten performance metrics. Moreover, the accuracy result is detailed into eight further measures for two different classes and averages, although there is no explanation regarding all the abbreviations of the measures.

Pamina et al. (2019) presented an effective classification to predict the probability of existing customers leaving their providers within the telecommunication industry using KNN, Random Forest, and XGBoost. The dataset used in this paper is obtained from IBM Watson, which is published in 2015. It interestingly showed that the subscription of fiber optic service with a more monthly charge has a superior impact on the customer churn amongst telecommunication providers, although it remains unclear the impact of the purchase power parity of the country to the monthly charge. Overall, the XGBoost obtained the highest performance in terms of accuracy and F-score, followed by Random Forest and KNN.

Another comparison was proposed to present the finding of a better performance by the enhanced logistic regression technique (Coussement et al., 2017). Although the comparison was comprehensively conducted only on the data preprocessing step, this paper compared its modified method of the logistic regression model, built on the data set using an optimal data preparation treatment with almost all advanced classification methods, such as stochastic gradient boosting, Bagged Classification and Regression Trees, Bayesian network, Decision Tree with J4.8, multilayer perceptron neural network, Naïve Bayes, Random Forest, and radial basis kernel SVM. It is also interesting that this paper argued that the proposed data preparation technique has contributed 14.5% enhancement for the analysis performance.

The previous works give us the motivation to have a further comparison. Hence, this research compares five well-known classification methods to find the best accuracy, precision, and recall in analyzing the customer churn that is not yet presented in previous research. Moreover, we partly base the analysis work on our previous experience (Bramantoro et al., 2015; Murakami et al., 2012).

## 3. Theoretical comparison

Decision Tree is one of the graphical representations in modeling. The format is like a tree structure in the data structure. The node does not provide a specific value of variables, but it more likely characterizes an examination of a feature. The forking branch of the tree embodies an examination result. The nodes with no further forking are called the terminal nodes that name the predefined classes. Once the Decision Tree is formed, a new tuple is examined whether its features fit into the Decision Tree or not. In this research, the Decision Tree examines whether a tuple of the customer is considered a chunk or not.

In the Decision Tree, the link from the top of the tree to the terminal node represents the comprehensive analysis to fit a tuple into a class. Hence, the whole links of the Decision Tree are extracted as rules. The development of Decision Tree exploits it to match an observation on a customer to a conclusion of the objective value, i.e. customer chunk status.

The objective value that can handle a limited number of values is the main characteristic of classification. In the Decision Tree, the leave represents a class name, and the forking branch exemplifies a combination of the features that eventually form the class names. In terms of the development process, the Decision Tree is considered agile amongst classification methods. Moreover, when combined with the database system, the rules extracted from the Decision Tree are easily converted to SQL statements.

One of the Decision Tree methods is C4.5 (Quinlan, 2014), which is defined in Eq. 1 as follows:

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

where $S$ is the set of instances, $A$ is an attribute, $N$ is the number of partition attribute $A$, $|S_i|$ is the number of instances on $i$ partition, $|S|$ is the number of instances.

SVM is one of the learning techniques with supervision in analyzing data and pattern recognition. This technique is useful for classification and regression analysis. SVM reduces the underlying risk by discovering the finest hyperplane that divides two classes in the input area, as initially stated by Cortes and Vapnik (1995): "Support Vector Machine (SVM) originally separates the binary classes ($k$=2) with a maximized margin criterion."

The training data are represented by ($x_i$, $y_i$), where $i$ is the label or class with the values 1, 2, 3,.., $N$. $N$ is the number of data. The available data are denoted as $x_i \rightarrow\in E^d$, where $x_i$ = {$x_{i1}$, $x_{i2}$, $x_{i3}$,.., $x_{iq}$}. $T$ is the attribute (feature) for the $i^{th}$ training data. There should be an assurance that two classes -1 and +1 are entirely divided by a hyperplane with $d$ dimension, which is defined in Eq. 2 as follows:

$$w.x_i + b = 0 \qquad (2)$$

where $w$ and $b$ are model parameters for which we want to find the value, while $w.x_i$ is the inner product between $w$ and $x_i$. Data (occasionally referred to as patterns) $x_i$ are included in class -1 that can be articulated as a model that satisfies two constraints in Eq. 3 and Eq. 4 as follows:

$$w.x_i + b \leq -1 \qquad (3)$$
$$w.x_i + b \geq +1 \qquad (4)$$

The SVM classification in Eq. 3 and 4 can be combined with a notation $y_i(w.x_i + b) \geq 1, i = 1,2,..,N$. The largest surplus can be attained by exploiting the gap between the hyperplane and its nearest data.

ANN is one of the methods in artificial intelligence and is considered a popular breakthrough in a computational system. It is commonly utilized for assisting pattern recognition and data mining. The key concept of ANN is based on the science of biological neural, particularly the human brain that links a huge number of neurons (Abiodun et al., 2018).

Fig. 1 illustrates a generalized process of ANN from input to output through input, hidden, and output layers. The input and output layer has an easy pattern to grasp the process. The hidden layer is like a black box without knowing the pattern inside and, therefore, it requires several experimentations to obtain the best result, including in this research.

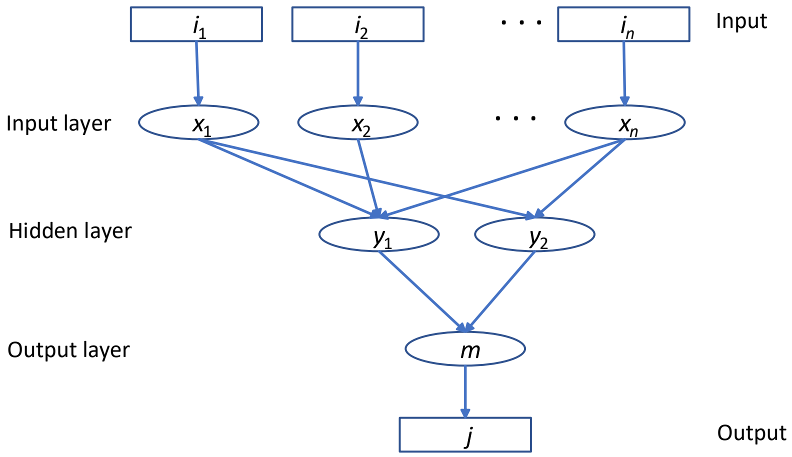


**Fig. 1:** ANN common process flow

The Gaussian Naïve Bayes is a classic classification technique by exploiting the statistical calculation proposed by the English statistician Thomas Bayes. This technique forecasts the possibility of upcoming events based on initial knowledge. The key feature of the Naïve Bayes classifier is the robust hypothesis that all parameters are independent of each other, like the Naïve world. Barus et al. (2020) argued that this hypothesis is suitable for parameters such as customer churn.

The Gaussian Naïve Bayes technique has the main benefit of its required number of training data to decide the projected parameter required during the classification. The required number can be less than other techniques due to its characteristic of parameter independence, hence, the variant of a parameter is only required to classify the customer churn. The Naïve Bayes formula is expressed in Eq. 5 as follows:

$$P(C|X) = \frac{P(X|C)P(C)}{P(X)} \tag{5}$$

where $P(C|X)$ is a posterior, $P(X|C)$ is a likelihood, $P(C)$ is a prior, and $P(X)$ is evidence.

KNN is a classifier that groups the data based on the proximity of the data to other data. One of the calculations for the distance between two objects is the Euclidean, although some defects have been lately found based on the singularity theory (Maxim et al., 2020). For object $A$ with observation value $x$=($x_1$, $x_2$,.., $x_p$) and object $B$ with observation value $y$=($y_1$, $y_2$,.., $y_p$), the Euclidean distance between objects $A$ and $B$ is shown in Eq. 6 as follows:

$$d_{AB} = d(x,y) = \sqrt{\sum_{i=1}^{p}(x_i - y_i)^2} \tag{6}$$

The performance of all these five models is evaluated by a confusion matrix that combines the accuracy, precision, recall, and F-Measure. The confusion matrix is sequentially formulated from Eqs. 7 to 10, where $TP$ refers to truly identified as positive result, $TN$ refers to truly identified as negative result, $FP$ refers to falsely identified as positive result, and $FN$ refers to falsely identified as negative result.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{7}$$

$$Precision = \frac{TP}{TP+FP} \tag{8}$$

$$Recall = \frac{TP}{TP+FN} \tag{9}$$

$$F-Measure = \frac{2\,x\,Precision\;x\;Recall}{Precision+Recall} \tag{10}$$

## 4. Implementation

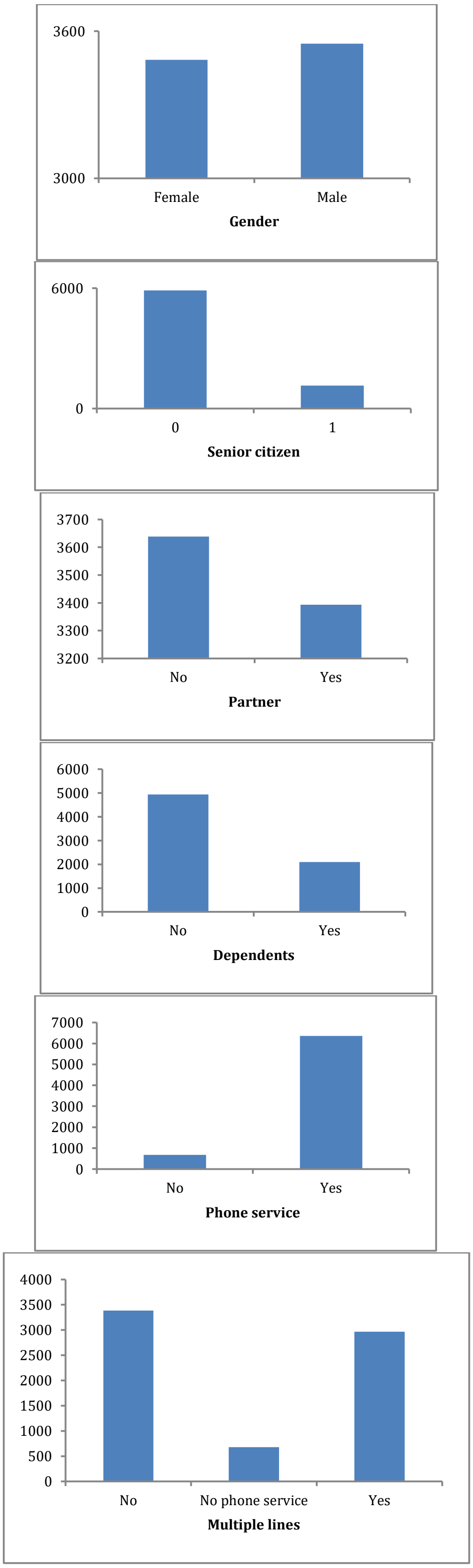
The proposed steps in this research are data preparation, exploratory data analysis, data cleaning, feature selection, and model building. Data preparation is non-trivial but time-consuming task. In this step, data are collected, combined, transformed, and cleaned. In detail, it includes checking for the completeness of the content, looking at the dimension, reviewing the structure of the input dataset, peeking into data, and checking for any missing data of particular customers. Python is chosen to be a modeling language because it provides robust libraries required in this research, such as Pandas, Matplotlib, and Seaborn.

The telecommunication customer churn dataset is provided by IBM Cloud Pak (Palmer, 2019). It consists of 20 predictor variables, 7,043 rows, and one response variable required as a label. The detail of the dataset is listed in Table 1 to give a more comprehensive understanding of the data. Out of 20 predictor variables, CustomerNo is the only one that is not utilized because it has a unique value. It is also important to check whether TotalCharges is an object instantiated from a class or not as well as MonthlyCharges is a floating-point or not. If yes, it is required to convert TotalCharges from ClassObject to Float64 during the data transformation.

**Table 1:** Dataset of telecommunication customer churn

| No | Columns | Type | Description |
|---|---|---|---|
| 1. | CustomerNo | ClassObject | The ID number of the Customer. |
| 2. | IsGender | ClassObject | Whether the customer is man or a woman. |
| 3. | IsSenior_Citizen | Int64 | Whether the customer has a status of a senior citizen or not (1 or 0). |
| 4. | IsPartner | ClassObject | Whether there is a spouse or not (Yes or No). |
| 5. | IsDependents | ClassObject | Whether there are children or not (Yes or No). |
| 6. | IsTenure | Int64 | The month period of the customer has been using the telecommunication provider. |
| 7. | IsPhone_Service | ClassObject | Whether there is phone service subscription or not (Yes or No). |
| 8. | IsMultiple_Lines | ClassObject | Whether there is a connection of multiple lines or not (Yes, No or No Telephone Service). |
| 9. | IsInternet_Service | ClassObject | The kind of internet service (DSL, Fiber optic or Not Applicable). |
| 10. | IsOnline_Security | ClassObject | Whether there is a subscription to online security or not (Yes or No). |
| 11. | IsOnline_Backup | ClassObject | Whether there is a subscription to online backup service or not (Yes or No). |
| 12. | IsDevice_Protection | ClassObject | Is there any subscription to a device protection service (Yes or No)? |
| 13. | IsTech_Support | ClassObject | Whether the customer deals with technical support or not (Yes or No). |
| 14. | IsStreaming_TV | ClassObject | Is there any subscription to a television streaming service or not (Yes or No)? |
| 15. | IsStreaming_Movies | ClassObject | Whether there is a subscription to streaming movies or not (Yes or No). |
| 16. | IsContract_Type | ClassObject | The type of contract (Monthly, Annually or Biannually). |
| 17. | IsPeperless_Billing | ClassObject | Is there any subscription to a paperless billing service (Yes or No)? |
| 18. | IsPayment_Method | ClassObject | The type of payment chosen by customer (Electronic check, Postal check, Bank transfer or Credit card). |
| 19. | Monthly_Charges | Float64 | The charges should be paid monthly. |
| 20. | Total_Charges | ClassObject | The total charges that should be paid. |
| 21. | IsChurn | ClassObject | Whether the customer churn or not (Yes or No). |

Fig. 2 shows the histogram to grasp the distribution of variables. It is interesting to note that only two variables, i.e., gender and partner, have almost equal distributions. The rest of the variables have partly unequal distributions for three feature values and total unequal distributions for two feature values. However, it remains unclear whether these unequal distributions affect the analysis or not. Moreover, there is no clear relationship between the number of features and the analysis result. This is due to the fact that the number of features in this research are only two and three.



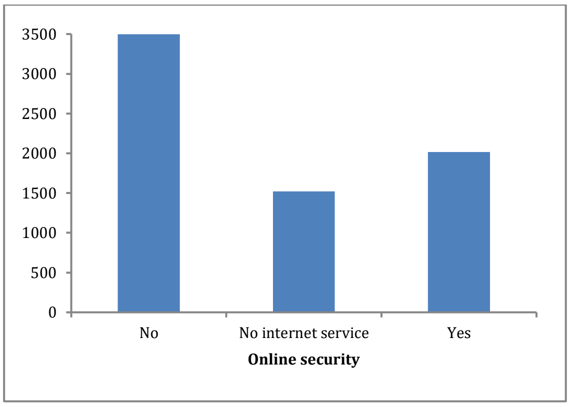


**Fig. 2:** Variables distribution

The next step is required to identify several interesting columns, such as tenure, and monthly charges. It is expected to have an interesting insight during this early step. As illustrated in Fig. 3, it is found that the number of customers who churn tends to reduce as the tenure period increases, unless for the longest period of tenure, i.e. 70 months and more. An extreme insight from the same visualization is that there are over a thousand customers with the longest tenure period, i.e. 70 months and more are not leaving their providers. On the other hand, almost the same number of customers who have the shortest period of tenure prefer not to leave. Among these two extreme numbers, the number of customers who are loyal to their providers fluctuates. This can be explained as the comfort zone existence amongst loyal customers and the rookie ones.

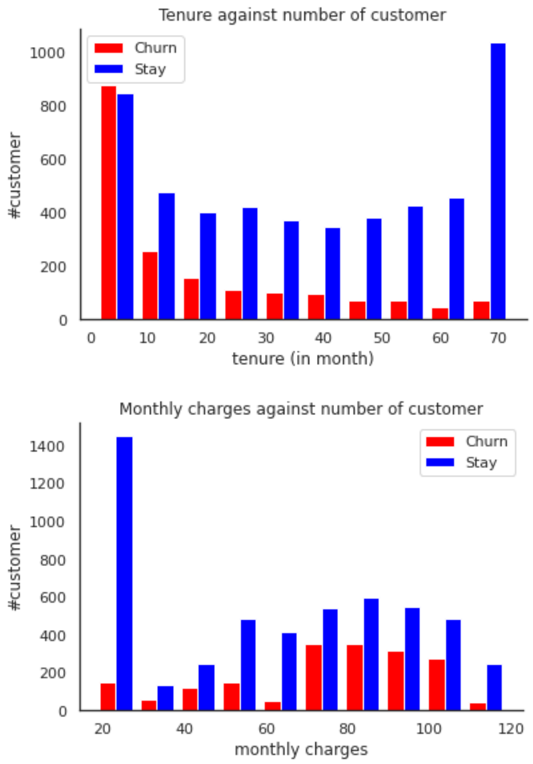
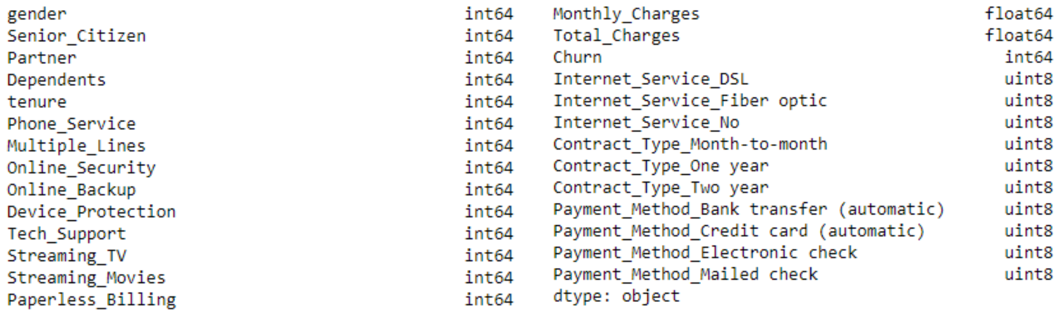


**Fig. 3:** Tenure and monthly charges against number of customers

Fig. 3 also shows that customers who have a relatively high charge, such as the ones between 70 and 110, have more possibility to churn. However, the highest charge does not reflect the highest possibility to churn amongst the customers. This is because there is a common sense of comfort zone amongst the rich people who tend to be this type of customer. It is also interesting to note that there is an extremely high value of the number of customers who chose to stay with their providers with the cheapest charges every month. This can be explained that the low-class customer who billed with low monthly charges have no further needs to extend the use of the telecommunication by leaving to another provider.

In this research, there is also a requirement to conduct data transformation and normalization. Several columns have *Yes* and *No* categorical values, that require to transform into 1 and 0. Several columns, such as Internet_Service, Contract_Type, and Payment_Method, have more than two categories that require a conversion from the categorical data to numerical data. There is also a requirement to scale on *Tenure*, Monthly_Charges, and Total_Charges columns.

Once the transformation of categorical data is completed, there are 27 columns obtained as illustrated in Fig. 4. There are three types of these 27 columns: int64, uint8, and float64. Amongst these three types, float64 is only applied to two columns: Monthly_Charges, and Total_Charges which require an advance analysis due to the complex nature of the type.



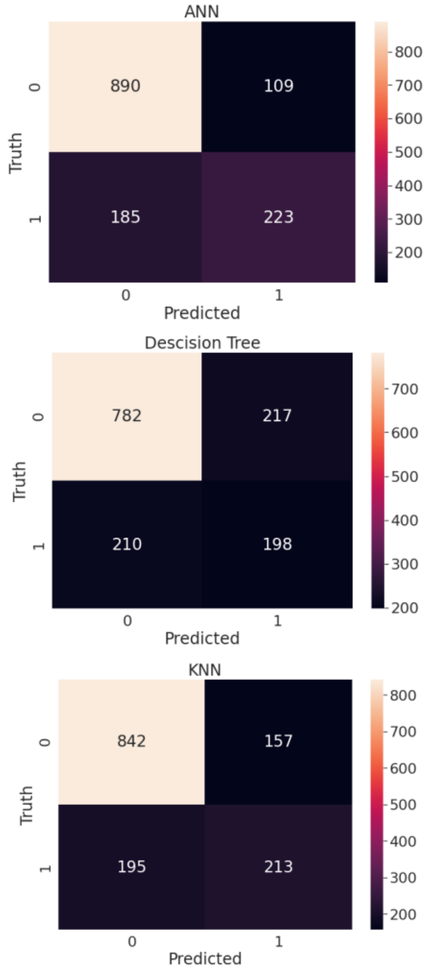
```
gender                int64    Monthly_Charges                          float64
Senior_Citizen        int64    Total_Charges                            float64
Partner               int64    Churn                                      int64
Dependents            int64    Internet_Service_DSL                       uint8
tenure                int64    Internet_Service_Fiber optic               uint8
Phone_Service         int64    Internet_Service_No                        uint8
Multiple_Lines        int64    Contract_Type_Month-to-month               uint8
Online_Security       int64    Contract_Type_One year                     uint8
Online_Backup         int64    Contract_Type_Two year                     uint8
Device_Protection     int64    Payment_Method_Bank transfer (automatic)   uint8
Tech_Support          int64    Payment_Method_Credit card (automatic)     uint8
Streaming_TV          int64    Payment_Method_Electronic check            uint8
Streaming_Movies      int64    Payment_Method_Mailed check                uint8
Paperless_Billing     int64    dtype: object
```

**Fig. 4:** Transformed dataset

## 5. Results and discussion

In this research, the data are allocated for training and testing set with the proportion of 80% and 20% correspondingly. This division is the best practice in most data mining experiments. There are five techniques used in this research: Artificial Neural Network (ANN), Decision Tree, K-Nearest Neighbor (KNN), Gaussian Naïve Bayes, and Support Vector Machine (SVM). The performance of all models is evaluated by utilizing the confusion matrix to provide a visual comparison as illustrated in Fig. 5. The ANN model accurately predicted that 223 customers would churn, and 890 customers would not churn. The decision Tree model correctly classified if 198 customers would churn in contrast to 782 customers who would not churn. KNN model accurately predicted that 213 customers would churn and 842 customers would not churn. Gaussian Naïve Bayes correctly classified 326 customers would churn and 728 customers would not churn. SVM model accurately predicted that 202 customers would churn and 894 would not churn.

In the evaluation, the ANN model has 79% accuracy, 67% precision, 55% recall, and 60% F-Measure. Decision Tree model has 70% accuracy, 49% precision, 49% recall, and 49% F-Measure. KNN model has 75% accuracy, 58% precision, 52% recall, and 55% F-Measure. Gaussian Naïve Bayes model has 75% accuracy, 55% precision, 80% recall, and 65% F-Measure. SVM model has 78% accuracy, 66% precision, 50% recall, and 57% F-Measure. The details of the measurement are shown in Table 2.



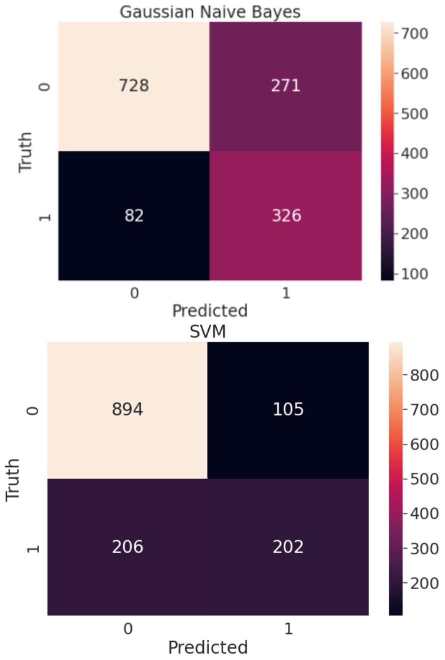


**Fig. 5:** Confusion matrixes

**Table 2:** The measurement result for all techniques

| Models | Accuracy | Precision | Recall | F-Measure |
|---|---|---|---|---|
| ANN | 79% | 67% | 55% | 60% |
| Decision Tree | 70% | 49% | 49% | 49% |
| KNN | 75% | 58% | 52% | 55% |
| G. Naïve Bayes | 75% | 55% | 80% | 65% |
| SVM | 78% | 66% | 50% | 57% |

The model with the highest accuracy is ANN with 79% and the model with the lowest one is Decision Tree with 70%. The model with the highest precision is ANN with 67% and the model with the lowest one is Decision Tree with 49%. The model with the highest recall is Gaussian Naïve Bayes with 80% and the model with the lowest one is Decision Tree with 49%. The model with the highest F-Measure is Gaussian Naïve Bayes with 65% and the model with the lowest one is Decision Tree with 49%.

Overall, ANN and Gaussian Naïve Bayes are two methods recommended to analyze customer churn in the telecommunication industry. This is because that these two methods outperform other methods in terms of four measurement values, i.e.: accuracy, precision, recall, and F-Measure. It is also interesting to note that the Decision Tree is the unsuggested method due to its lowest evaluation score in any measurements. However, it is commonly understood that the Decision Tree method has different algorithms on how to select the attributes in creating the tree and different mechanisms for pruning. Hence, there might be a better result obtained with different algorithms of the Decision Tree.

## 6. Conclusion

A robust method is needed by the telecommunication providers to analyze the customer churn to increase their competitive advantages. There are several methods in data analysis that have a broad range of prediction results. In this paper, we found that ANN is the method with the highest accuracy and precision, whereas Gaussian Naïve Bayes is the method with the highest recall and F-Measure. Hence, it is recommended to use these two methods for analyzing customer churn in the telecommunication industry. Moreover, it is not recommended to use Decision Tree for analyzing the customer churn due to its low results during the evaluation. However, it remains unclear whether this conclusion is due to the characteristic of the customer churn dataset or not. In the future, it is possible to enhance the model performance by combining these classification techniques as one workflow to complement each other and investigating different datasets of customer churn.

## Compliance with ethical standards

### Conflict of interest

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

## References

Abiodun OI, Jantan A, Omolara AE, Dada KV, Mohamed NA, and Arshad H (2018). State-of-the-art in artificial neural network applications: A survey. Heliyon, 4(11): e00938. https://doi.org/10.1016/j.heliyon.2018.e00938 **PMid:30519653 PMCid:PMC6260436**

ApurvaSree G, Ashika S, Karthi S, Sathesh V, Shankar M, and Pamina J (2019). Churn prediction in telecom using classification algorithms. International Journal of Scientific Research and Engineering Development, 5: 19-28.

Barus DT, Elfarizy R, Masri F, and Gunawan PH (2020). Parallel programming of churn prediction using Gaussian Naïve Bayes. In the 8th International Conference on Information and Communication Technology, IEEE, Yogyakarta, Indonesia: 1-4. https://doi.org/10.1109/ICoICT49345.2020.9166319 **PMid:32154151 PMCid:PMC7034176**

Bramantoro A, Hassine AB, Matsubara S, and Ishida T (2015). Multilevel analysis for agent-based service composition. Journal of Web Engineering, 14(1&2): 63-79.

Cortes C and Vapnik V (1995). Support-vector networks. Machine Learning, 20(3): 273-97. https://doi.org/10.1007/BF00994018

Coussement K, Lessmann S, and Verstraeten G (2017). A comparative analysis of data preparation algorithms for customer churn prediction: A case study in the telecommunication industry. Decision Support Systems, 95: 27-36. https://doi.org/10.1016/j.dss.2016.11.007

Dahiya K and Bhatia S (2015). Customer churn analysis in telecom industry. In the 4th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions), IEEE, Noida, India: 1-6. https://doi.org/10.1109/ICRITO.2015.7359318 **PMid:26827852**

Jain H, Khunteta A, and Srivastava S (2020). Churn prediction in telecommunication using logistic regression and logit boost. Procedia Computer Science, 167: 101-12. https://doi.org/10.1016/j.procs.2020.03.187

Khan Y, Shafiq S, Naeem A, Hussein S, Ahmed S, and Safwan N (2019). Customer churn prediction using artificial neural network (ANN) in telecom industry. International Journal of Advanced Computer Science and Applications, 10: 132-142. https://doi.org/10.14569/IJACSA.2019.0100918

Maxim LG, Rodriguez JI, and Wang B (2020). Defect of Euclidean distance degree. Advances in Applied Mathematics, 121: 102101. https://doi.org/10.1016/j.aam.2020.102101

Murakami Y, Tanaka M, Bramantoro A, and Zettsu K (2012). Data-centered service composition for information analysis. In the 9th International Conference on Services Computing, IEEE, Honolulu, USA: 602-608. https://doi.org/10.1109/SCC.2012.88

Palmer T (2019). Predict and optimize business outcomes with IBM decision optimization for Watson studio and IBM Cloud Pak for data. The Enterprise Strategy Group, Milford, USA.

Pamina J, Raja B, SathyaBama S, Sruthi MS, and Aiswaryadevi VJ (2019). An effective classifier for predicting churn in telecommunication. Journal of Advanced Research in Dynamical and Control Systems, 11(01-Special Issue): 221-229. https://doi.org/10.5373/JARDCS/V11SP11/20193050

Quinlan JR (2014). C4.5: Programs for machine learning. Elsevier, Amsterdam, Netherlands.

Russell I and Markov Z (2017). An introduction to the Weka data mining system. In the 2017 ACM SIGCSE Technical Symposium on Computer Science Education, Association for Computing Machinery, Seattle, USA: 742. https://doi.org/10.1145/3017680.3017821

View publication stats