Institut Riset dan Publikasi Indonesia (IRPI)
IJATIS: Indonesian Journal of Applied Technology and Innovation Science
Journal Homepage: https://journal.irpi.or.id/index.php/ijatis
Vol. 2 Iss. 1 February 2025, pp: 1-6
ISSN(P): 3032-7466 | ISSN(E): 3032-7474

Comparison of Support Vector Machine, Random Forest, and C4.5 Algorithms for Customer Loss Prediction

Bima Maulana[1*], Dany Febrian[2], Irgie Rachmat Fachrezi[3], Muhammad Ferdi Zeen[4]

[1,2,3]Program Studi Sistem Informasi, Fakultas Sains dan Teknologi
Universitas Islam Negeri Sultan Syarif Kasim Riau, Indonesia
[4]International University of Africa Khortum, Sudan

E-Mail:[1]12050313080@students.uin-suska.ac.id, [2]12050313116@students.uin-suska.ac.id,
[3]12050310381@students.uin-suska.ac.id, [4]zeenferdy@gmail.com

Received Dec 23th 2023; Revised Nov 16th 2024; Accepted Nov 28th 2024; Available Online Feb 28th 2025, Published Feb 28th 2025
Corresponding Author: Bima Maulana

Abstract

Loss of customers has been discussed and many studies have been conducted, starting from using the Bayesian network algorithm, Decision tree, random vorest, Support vector machine, and neyral network Algorithms Support Vector Machine (SVM), Random Forest, and Decision Tree or C4.5 are algorithms used for prediction and have several advantages Random forest has the advantage of being able to combine many predictions from decision trees that have a tendency to reduce overfitting. This research uses the C4.5 algorithm, SVM and random forest. Research shows that the Random Forest method has the highest accuracy of 87.02% compared to the Support Vector Machine and Decision Tree methods. In contrast, Decision Tree gets low accuracy results with a value of 78.52%. Experimental results show that the Random forest method for customer loss prediction achieves an average classification accuracy of 4% - 9% higher than the Support Vector Machine and Decision Tree methods.

Keyword: Customer Loss Prediction, C4.5, Random Forest, Support Vector Machine

1. INTRODUCTION

Very significant technological developments affect people's lifestyles and habits, starting from shopping, learning and surfing the web to get more information so that it indirectly changes people's habits [1]. The many benefits offered by this technological development include companies in the telecommunications sector that utilize technological sophistication for marketing. With the existence of marketing via digital has also been tried by many other companies so as to create high competitiveness and require each telecommunications company to create the latest innovations and new strategies. The number of competitors who use and utilize technology for marketing their products as well [2].

The high competitiveness also provides a high possibility of a high level of possibility that customers will be lost or what is commonly called not subscribing or moving to another company that is more attractive to customers [3]. Strategy is needed to increase competitiveness. The high competitiveness of many competitors creates new similar competitors. The possibility of bankruptcy of a company that is unable to compete due to customers who move subscriptions is carried out research to avoid customers who are likely to be lost as well as to obtain more data related to the running of the company [4].

Customer loss has also been widely discussed and many studies have been conducted, starting from using the Bayesian network algorithm, decision tree, random vorest, Support vector machine, and neyral network [4], [5] The Support Vector Machine (SVM), Random Forest, and C4.5 algorithms are algorithms used for prediction and have several advantages [6]. SVM is very effective for dealing with data that has many features and high dimensions and has good outlier handling capabilities because this algorithm tries to find the largest margin between different classes. Random forest has the advantage of being able to combine many predictions from decision trees which has a tendency to reduce overfitting [7]. C4.5 is an algorithm that is easy to understand and interpret. The three algorithms used for predicting customer loss are very likely to be used to increase competitiveness and increase input for future steps for the company [7], [8].

Previous research has been conducted by Stevan Desena Damanik and Muhammad Ihsan Jambak in 2023 on the classification of costumer churn in the telecommunications industry for customer retention using the C4.5 algorithm, this research is based on Cross-Industry Standard Process for Data Mining (CRISP-DM). Using Rapid Miner tools and the C4.5 algorithm with classification techniques as a solution to analyze the characteristics of customer churn. From the results of the study it is known that there are 5 attributes that have a considerable influence on customer churn, namely contract, InternetService, TotalChares, tenure, PaperlessBilling, MultipleLines, StreamingMovies.And from the results of this study has an accuracy rate of 79.53% [9].

The previous research that has been done, it is the basis of this research. Namely by comparing the accuracy of the support random machine, random forest and C4.5 algorithms. The novelty of this research is the comparison of the three algorithms using several experimental techniques, including the comparison of data sharing and several other parameters in the three algorithms.

2. MATERIAL AND METHOD

2.1. Data Collection

Data collection is done using the kaggle API with google colab. This research uses a dataset containing customer data in a telecommunications company. The dataset used is 955 KB in size. This dataset has 2000 customer data and 12 columns. The 12 columns are variables that will be used as churn predictions. The variables contained in the churn dataset are as follows.

1. SeniorCitizen: The customer is a senior citizen. This field has 2 values of 0 and 1.
2. Depedents: The customer has dependents. This column has 2 values, viz: Yes and No.
3. Tenure: The number of months the customer has used the company's services.
4. PhoneService: The customer has a phone service. This column has 2 values, Yes and No.
5. MultipleLines: The customer has a multi-line service. This column has 2 values, Yes and No.
6. InternetService: Customer's internet service provider. This column has 3 values which are DSL, Fiber Optic and No.
7. Contract: Customer's contract terms. This column has 3 values viz: Month-to-month, One year, Two year.
8. PaperlessBilling: Customer has paperless billing. This column has 2 values viz: Yes and No.
9. PaymentMethod: Customer's payment method. This field has 4 values, namely: Electronic check, Mailed check, Bank transfer (automatic) and Credit card (automatic).
10. MonthlyCharges: This is the amount charged to the customer every month.
11. TotalCharges: The total amount of services charged to the customer.
12. Churn: Customer category churn or not. This column has 2 values, namely Yes and No.

2.2. Preprocessing

After the data is collected, the next process is preprocessing. Data preprocessing is a process that aims to transform data into a format that is easier and more effective for users. One of the most important issues in data preprocessing is how we find out what valuable information is in the raw data so that we can ensure the information is retained [10]. This may depend on our definition of data preprocessing. Some of the data preprocessing methods we use are:

1. Data cleaning, is the process of cleaning data that has missing values.
2. Data adjustment, is the process of adjusting the amount of data for each target.
3. Data separation, is the process of separating data into two groups, namely train and test.

2.3. Transformation

Transformation is the process of changing the selected data, so that the data is suitable for the data mining process. This process is a creative process and depends heavily on the type or pattern of information to be searched in the database. In this process, the data will be grouped using the One Hot Encoding method [11].

2.4. Data Mining

Data mining is the process of extracting added value from a data set in the form of knowledge that has not been known manually. This is done through a series of stages, including data collection, data extraction, data analysis, and data statistics [12]. There are many variations of techniques, methods, or algorithms in data mining. This research uses the C4.5 algorithm, Support Vector Machine (SVM) and random forest.

2.5. C4.5 Algorithm

The C4.5 algorithm is an algorithm used in forming decision trees. The C4.5 algorithm is one of the algorithms in decision tree induction, Iterative Dichotomizer 3 (ID3) developed by J. Ross Quinlan. In the ID3 algorithm procedure, inputs are training samples, training labels and attributes. The C4.5 algorithm is a development of ID3 [13]. The basic idea of this algorithm is the creation of a decision tree based on the selection

of attributes that have the highest priority or can be called having the highest gain value based on the entropy value of these attributes as the axis of classification attributes [14]. Then recursively the branches of the tree are expanded so that the entire tree is formed. There are four steps in the process of making a decision tree in the C4.5 algorithm, namely:

1. Selecting the attribute as the root
2. Creating branches for each value
3. Dividing each case in the branch
4. Repeating the process in each branch until all cases in the branch have the same class.

2.6. Support Vector Machine (SVM)

An SVM is an algorithm that works using nonlinear mapping to transform the original training data to a higher dimension. In the new dimension, it will then find the optimal linear separating hyperplane (i.e., the "decision boundary" that separates tuples from one class to another) [15]. This method uses nonlinear mapping to transform the original training data to a higher dimension. In this new dimension, a linear optimization that separates the two target classes with a hyperplane is sought. A hyperplane is a decision boundary that separates types from one class from another. SVM finds the hyperplane using support vectors and margins [16].

2.7. Random Forest

Random Forest is a learning method for classification and regression. It creates a series of decision trees at the same time as training. Classify new cases by assigning new cases to each tree. Each tree performs classification and produces a class [17]. The output class is selected based on the most votes, that is, the maximum number of similar classes generated by various trees is considered the output of Random Forest [18].

3. RESULTS AND DISCUSSION

3.1. Results and Analysis

A dataset containing customer data from a telecommunications company, such as those available on Kaggle, is a valuable resource for analyzing customer loss prediction. This dataset typically includes various customer attributes that can be examined to understand patterns, behaviors, and factors influencing customer loyalty toward the services offered by the company. This dataset is highly relevant for customer loss prediction research as it provides variables reflecting customer behavior. Using these attributes, predictive models can be developed to identify customers at high risk of leaving the service.

Using this dataset in customer loss prediction research can provide significant contributions, including: Enhancing Retention Strategies: Prediction results can be used to identify specific customer groups needing targeted attention. Resource Efficiency: By focusing retention efforts on high-risk customers, companies can optimize resource allocation. Developing High-Accuracy Models: The dataset supports testing various predictive algorithms, such as Random Forest, SVM, and Decision Tree, to identify the best-performing method with the highest accuracy.

Confusion Matrix Testing Results from 3 modeling, Description:

1. True Positive (TP): The result predicts a positive customer turnover in the telecommunication used and it is true that the customer turnover is positive in the telecommunication used.
2. True Negative (TN): The result of predicting negative customer turnover in the telecommunication used and it is true that the negative customer turnover in the telecommunication used.
3. False Positive (FP): The result of predicting a positive customer turnover in the telecommunication used and the prediction is wrong, it turns out to be a negative customer turnover in the telecommunication used.
4. False Negative (FN): The result of predicting negative customer turnover in the telecommunication used and the prediction is wrong, it turns out that the customer turnover is positive in the telecommunication used. As explained above, FN is a type 2 error where this error is quite detrimental to the company. A customer is predicted to be negative using telecommunications when in fact the customer is positive using telecommunications then the customer that the cost of acquiring new customers is far greater than the cost of maintaining existing customers [19], [20].

The accuracy of the modelling algorithm in classifying how accurate the process is. The results of the modelling obtained confusion matrix of SVM accuracy is 83,18%, while for Random Forest has an accuracy of 87,02% and Decision Tree with 78,52% accuracy. The confusion matrix of the three algorithms can be shown in Figures 1, Figure 2 and Figure 3.

Based on the evaluation results of the modeling algorithms, it can be concluded that the accuracy achieved for each algorithm shows significant variation. The results obtained from the confusion matrix are as follows:

1. SVM achieved an accuracy of 83.18%, indicating a fairly good performance in classifying the data correctly, although there are still some classification errors that need further improvement.

2. Random Forest showed better results with an accuracy of 87.02%, making it the algorithm with the highest accuracy among the three tested algorithms. This suggests that Random Forest is more effective in handling data variability and reducing classification errors.
3. Decision Tree had a lower accuracy compared to SVM and Random Forest, with an accuracy of 78.52%. Although it still shows reasonable accuracy, this algorithm is less effective in achieving higher classification accuracy.

From this analysis, Random Forest proves to be the most accurate algorithm in classification compared to SVM and Decision Tree. However, the differences in accuracy highlight the strengths and weaknesses of each algorithm, which should be considered when choosing the appropriate algorithm based on the needs and complexity of the data to be processed. For the visualization of the accuracy comparison results of each model, refer to Figure 4, which illustrates the comparison between the three algorithms based on the accuracy values obtained.

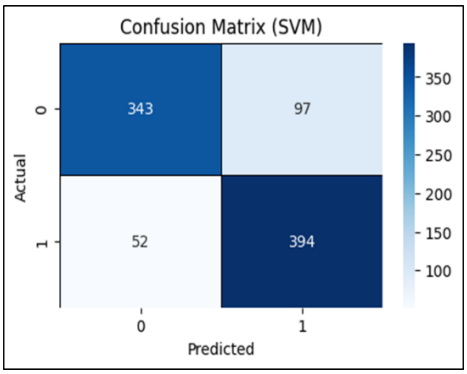

Figure 1. Visualization of confusion matrix of SVM algorithm

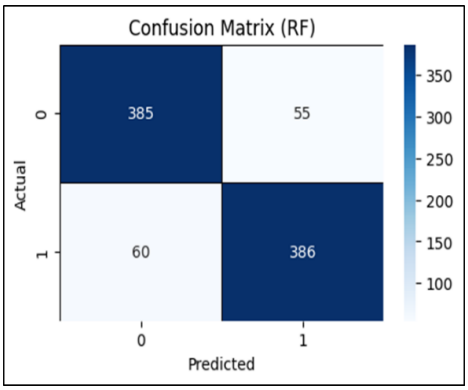

Figure 2. Random-Forest algorithm confusion matrix visualization

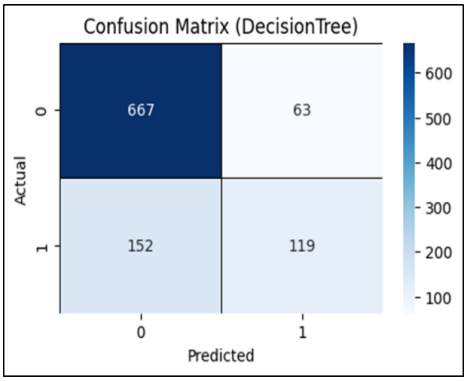

Figure 3. Visualization of confusion matrix of Decision Tree algorithm

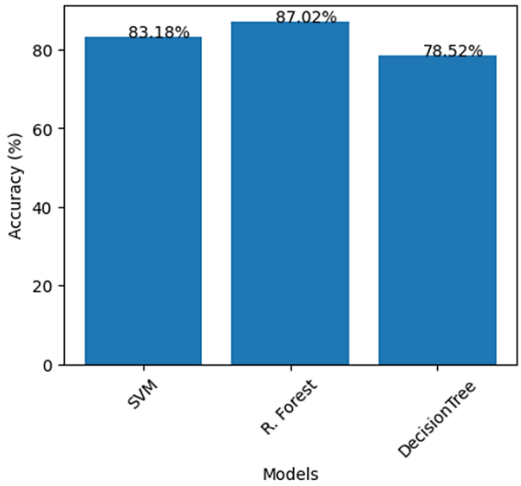

Figure 4. Visualization of Results from Accuracy Comparison of SVM, Random Forest, and Decision Tree Algorithms

So in the application of kaggle data at PT.Telco by implementing 3 algorithms and with the best parameter model of the Decision Tree model with an accuracy of 78.52%, SVM model with an accuracy of 83.18%, and Random-Forest with an accuracy of 87.02%. So from PT.Telco data with these 3 models with the best accuracy is Random-Forest with an accuracy of 87.02%.

Potential Research Developments: (1) Incorporating External Data: Combining customer data with external sources, such as market trends or social media reviews, to improve model accuracy. (2) Visualization of Insights: Providing visualizations such as heatmaps or feature importance charts to help interpret customer behavior patterns. (3) Practical Implementation: Integrating predictive models with CRM systems to provide automated recommendations for customer retention actions.

3.2. Discusion

The research results indicate that the Random Forest method outperforms the SVM and Decision Tree (C4.5) methods in predicting customer loss. Random Forest achieved the highest accuracy of 87.02%, significantly better than Decision Tree, which only reached 78.52%. This difference highlights the significant advantage of Random Forest in handling datasets related to customer loss prediction. The strength of Random Forest lies in its ability to combine predictions from multiple decision trees. This ensemble approach helps reduce the risk of overfitting, which often occurs in individual algorithms like Decision Tree. Furthermore, leveraging multiple decision trees enables the model to capture complex patterns in the data, resulting in more accurate and reliable predictions.

On the other hand, the Support Vector Machine method performed lower than Random Forest, although it still outperformed the Decision Tree. This might be due to SVM's sensitivity to kernel parameters and data distribution, which can influence the model's effectiveness. While SVM is suitable for high-dimensional data, it is not as efficient as Random Forest in managing data variations in this scenario. This study also utilized a Confusion Matrix to evaluate the models. The testing results based on True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) provide deeper insights into each method's ability to classify the data. The high TP and TN values in Random Forest confirm that this method effectively predicts both customers who will stay and those likely to leave the telecommunications service.

Overall, the experimental results indicate that Random Forest provides an average 4% - 9% higher classification accuracy than SVM and Decision Tree. This establishes Random Forest as a superior choice for customer loss prediction applications, particularly for data with complex characteristics. However, the low accuracy of Decision Tree highlights the need for alternative approaches, such as ensemble methods, to improve its performance. Additionally, this research can be further developed by evaluating other methods or combining algorithms to enhance prediction accuracy and reduce the risk of model bias toward specific datasets. Moreover, a deeper analysis of classification errors (FP and FN) can help identify the factors contributing to the model's misclassification.

4. CONCLUSION

A comparison of Support Vector Machine, Decision Tree, and Random Forest methods for customer loss prediction is presented in this paper. The research shows that the Random Forest method has the highest accuracy of 87.02% compared to the Support Vector Machine and Decision Tree methods. In contrast, Decision Tree has the lowest accuracy result with 78.52%. The experimental results show that the Random forest method

for customer loss prediction achieves an average classification accuracy of 4% - 9% higher than the Support Vector Machine and Decision Tree methods. However, the difference is not statistically significant.

The authors expect the results of this study to improve predictive modeling more significantly than changing to a better classifier, which has important implications for telecom operators facing competition. Telecom operators handling larger data assets, more so than predictive techniques, could potentially gain a huge competitive advantage over telecom operators without access to so much data.

REFERENCES

[1] S. Mitrović, B. Baesens, W. Lemahieu, and J. De Weerdt, "On the operational efficiency of different feature types for telco Churn prediction," Eur J Oper Res, vol. 267, no. 3, pp. 1141–1155, 2018.

[2] D. H. Tisantri, R. C. Wihandika, and S. Adinugroho, "Prediksi Keputusan Pelanggan Menggunakan Extreme Learning Machine Pada Data Telco Customer Churn," Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer, vol. 3, no. 11, pp. 10516–10523, 2019.

[3] K. Kelvin, C. Cindy, C. Charles, D. P. Leonardo, and Y. Yennimar, "Customer Churn's Analysis In Telecomunications Company Using Fp-Growth Algorithm: Customer Churn's Analysis In Telecomunications Company Using Fp-Growth Algorithm," Jurnal Mantik, vol. 4, no. 2, pp. 1285–1290, 2020.

[4] R. Yu, X. An, B. Jin, J. Shi, O. A. Move, and Y. Liu, "Particle classification optimization-based BP network for telecommunication customer churn prediction," Neural Comput Appl, vol. 29, pp. 707–720, 2018.

[5] B. Huang, M. T. Kechadi, and B. Buckley, "Customer churn prediction in telecommunications," Expert Syst Appl, vol. 39, no. 1, pp. 1414–1425, 2012.

[6] S. Huang, N. Cai, P. P. Pacheco, S. Narrandes, Y. Wang, and W. Xu, "Applications of support vector machine (SVM) learning in cancer genomics," Cancer Genomics Proteomics, vol. 15, no. 1, pp. 41–51, 2018.

[7] H. Nalatissifa and H. F. Pardede, "Customer Decision Prediction Using Deep Neural Network on Telco Customer Churn Data," Jurnal Elektronika dan Telekomunikasi, vol. 21, no. 2, pp. 122–127, 2021.

[8] N. Hashmi, N. A. Butt, and M. Iqbal, "Customer churn prediction in telecommunication a decade review and classification," International Journal of Computer Science Issues (IJCSI), vol. 10, no. 5, p. 271, 2013.

[9] S. D. Damanik and M. I. Jambak, "Klasifikasi Customer Churn pada Telekomunikasi Industri Untuk Retensi Pelanggan Menggunakan Algoritma C4. 5," KLIK: Kajian Ilmiah Informatika dan Komputer, vol. 3, no. 6, pp. 1303–1309, 2023.

[10] A. Famili, W.-M. Shen, R. Weber, and E. Simoudis, "Data preprocessing and intelligent data analysis," Intelligent data analysis, vol. 1, no. 1, pp. 3–23, 1997.

[11] S. Manikandan, "Data transformation," J Pharmacol Pharmacother, vol. 1, no. 2, p. 126, 2010.

[12] A. Febriani, T. T. Rahmawati, E. Sabna, P. Studi, T. Informatika, and H. T. Pekanbaru, "Implementation of Data Mining to Predict the Feasibility of Blood Donors Using C4.5 Algorithm 1," Indonesian Journal of Artificial Intelligence and Data Mining (IJAIDM), vol. 1, no. 1, pp. 41–46, 2018.

[13] W. Katrina, H. J. Damanik, F. Parhusip, D. Hartama, A. P. Windarto, and A. Wanto, "C. 45 classification rules model for determining students level of understanding of the subject," in Journal of Physics: Conference Series, 2019, p. 12005.

[14] H. Hasanah, "Perbandingan Tingkat Akurasi Algoritma Support Vector Machines (SVM) dan C4.5 dalam Prediksi Penyakit Jantung," 2023.

[15] S. Huang, N. Cai, P. P. Pacheco, S. Narrandes, Y. Wang, and W. Xu, "Applications of support vector machine (SVM) learning in cancer genomics," Cancer Genomics Proteomics, vol. 15, no. 1, pp. 41–51, 2018.

[16] K. Prima Wijaya and A. Muslim, Peningkatan Akurasi pada Algoritma Support Vector Machine dengan Penerapan Information Gain untuk Mendiagnosa Chronic Kidney Disease. 2016.

[17] A. Paul, D. P. Mukherjee, P. Das, A. Gangopadhyay, A. R. Chintha, and S. Kundu, "Improved random forest for classification," IEEE Transactions on Image Processing, vol. 27, no. 8, pp. 4012–4024, 2018.

[18] A. Primajaya and B. N. Sari, "Random forest algorithm for prediction of precipitation," Indonesian Journal of Artificial Intelligence and Data Mining, vol. 1, no. 1, pp. 27–31, 2018.

[19] I. Tachtsidis and F. Scholkmann, "False positives and false negatives in functional near-infrared spectroscopy: issues, challenges, and the way forward," Neurophotonics, vol. 3, no. 3, p. 31405, 2016.

[20] S. T. Brookes et al., "Subgroup analysis in randomised controlled trials: quantifying the risks of false-positives and false-negatives," 2001.