International Journal of Information Technology and Computer Science Applications (IJITCSA)
p-ISSN: 2964-3139 e-ISSN: 2985-5330
Vol. 1, No. 2, page 96 – 102
Submitted 07/05/2023; Accepted 10/05/2023; Published 02/06/2023

Comparative Study of Classification Algorithms for Customer Decisions on Telecommunication Products Using Supervised Learning

[1]John Kristian Vieri, [2]Tb Ai Munandar, [3]Dwi Budi Srisulistiowati, [4]Dwipa Handayani and [5]Achmad No'eman and [6]Tyastuti Sri Lestari
[1,2,3,4,5,6] Informatics Department, Universitas Bhayangkara Jakarta Raya, INDONESIA
e-mail : [1]johnchristianvieri@gmail.com,
[2]tbaimunandar@gmail.com,
[3]dwi.srisulistiowati@gmail.com,
[4]dwipa.handayni@dsn.ubharajaya.ad.id,
[5]achmad.noeman@dsn.ubharajaya.ac.id,
[6]tyas@ubharajaya.ac.id

Corresponding Autor: John Kristian Vieri

Abstract
Customers are the main goal of all business fields, without customers the company will not be able to continue or compete in the business field it is in, even though the company has brilliant products, if it does not have an increase in the number of customers the business will not be able to develop or even go bankrupt. Therefore, it is necessary to make observations and make applications that are able to predict customers who will subscribe so that companies can predict customers who will subscribe correctly without having to wait for confirmation from customers whose possibilities are still unknown. This can be very useful for any company because companies no longer need to look for random customers where it only takes time to find customers. PT. Telekomunikasi Indonesia with its product (Indihome) which is struggling to compete in the business world in the telecommunications and internet sector. Therefore research and development of this application are carried out so that PT. Indonesian telecommunications can get its customers quickly without having to spend a lot of money and effort. Making this application uses a classification method from machine learning technology based on customer historical data. The classification method has many strong algorithms for predicting variables that have more than 1 label. Some of the algorithms used are Logistic Regression, Random Forest Classifier, Support Vector Machine and Decision Tree which are provided by modules in the python programming language, namely SkLearn. The four algorithms will be tested with data balanced using the Oversampling method from the Smote algorithm to get optimal results in automatically predicting customers.

Keywords— Classification, Logistic Regression, Random Forest, Support Vector Machine, Decision Tree Classifier, Smote, Sk-Learn..

1 Introduction

PT Telekomunikasi Indonesia Tbk (Telkom), hereinafter referred to as PT Telkom, is a large company engaged in the information and communication sector that provides complete telecommunication network services. Currently they are trying to increase the number of customers in the country, especially the Bekasi city area because many customers have stopped subscribing to one of their products, namely Indihome. The development of internet services in Indonesia is increasing rapidly, marked by the emergence of many companies that provide the same product and the same product quality. This has indirectly become a very important concern for PT. Telkom.

All providers compete with each other in getting consumers to subscribe to their products in terms of price, service, quality and other offers. Some of these providers include MNC Play, Biznet, and ICONNet. These three providers compete seriously with PT. Telkom to get consumer attention. In terms of price, quality and even service, PT. Telkom is able to compete, but it is quite difficult to maintain and get customers. This problem was found in reports from Indihome customer management in Bekasi City from 2016 to 2021 which explained that there were indications that many customers chose to stop subscribing rather than continue subscribing. The condition experienced by the company is based on the decrease in the number of Indihome customers if it continues will have a negative impact on the sustainability of existing business at PT. Telkom through its indihome product. Therefore, an approach is needed to be able to predict customers who will subscribe so that companies can get customers quickly and precisely.

In the development of modern technology, data is an important resource for every company, where data can be used for any needs, one of which is the use of data to assist business decisions. There are many techniques that can be used to make business decisions, especially data-driven based, one of which is the Supervised Learning technique, namely Classification. Supervised Learning is one of the techniques that machine learning has in reading patterns in data that already has a certain label. There are so many algorithms that this technique has because the human need to predict something is very high, with predictive abilities, humans are able to anticipate future problems and even humans can get something faster. In making predictions, of course humans must have information that has high quality and quantity, therefore Big Data is needed with a balance on the portion of the label so that in its application it can provide optimal results and does not have a detrimental impact on the company due to errors in predicting or only being able to predict partially. labels only. However, in this study the data owned had an unbalanced portion of the label, so a technique was needed to solve the problem. The research carried out is an attempt to develop applications to be able to predict customers based on PT. Telkom customer data with their products (Indihome) with a classification approach, the expected results can be used as an alternative for making business decisions, especially at PT. Telkom. Thus PT. Telkom can anticipate the possibility of inactive and inactive customers or get new customers.

2 Research methods

The research design was carried out with several important stages based on the stages described by previous researchers. The stages in this research design will be explained as follows;

1) Data Preparation. The first stage is to prepare the data by collecting data from companies to be uploaded to the tools.
2) Data Analysis and Visualization. Analysis of the data is an important stage, this stage is carried out to be able to understand and identify certain conditions that exist or are experienced in the data. Such as understanding and knowing the distribution of data, the correlation between variables or called descriptive analysis and the influence on the label or called analysis correlation.
3) Preprocessing. Preprocessing is the stage where the data will be converted into data that is ready to be applied to the model, at this stage there will be stages such as handling missing data, handling extreme data, handling an unbalanced number of labels, and taking sample data by separating the data will be trained and the data will be tested.
4) Normalization. Data Normalization is the process of converting data into numerical data that can be understood or processed by the model.
5) Modeling. Modeling is the stage where the model will be implemented using fully prepared data.
6) Evaluation. Evaluation is the stage where the model will be tested and validated for each value using the metrics used in the classification model.
7) Determination. Determination of the model is the stage where the model is determined based on valid evaluation results.

3 Results and Discussion

The prepared data is raw data that will be processed into data that will be ready to be trained on the model. Data must be carefully prepared and understood as a whole. This is done in order to understand how existing data can help solve existing problems.

1) Data Analysis.

Analysis of the data is an important stage, this stage is carried out to be able to understand and identify certain conditions that exist or are experienced in the data. Such as understanding and knowing the distribution of data, correlation between variables, influence on labels.

a. Numerical Variables

Several numerical variables namely, Tenor, Monthly Payment and Total Payment. The three variables have a normal and positive distribution. Because each variable has an increase. And also obtained if the variable is based on a blue label or customers who do not subscribe tend to have a wider distribution than those who subscribe. Correlation itself simply means the strength of the linear relationship between two variables. The correlation value shows how the pattern of movement of the two variables is. If one variable increases and the other variables also increase, then the correlation will show a positive value, whereas if the motion of the two variables is opposite then the correlation will produce a negative value. But if the pattern of movement of this data moves randomly or randomly. And data that has a relationship but is non-linear (quadratic), then the correlation will be close to 0.

b. Categorical Variables

The categorical variables will be analyzed for the results of the target, this aims to determine the distribution of each categorical variable. The results of the analysis of the target label variable are obtained if more customers have unsubscribed than those who have subscribed. Analysis of the technician's assistance variable shows results if more customers do not use this feature. Analysis of the backup variable results if more customers do not use the features provided as well as backup variables, many customers do not use this feature. The same thing also happened for internet service features, subscription packages, streaming movies, streaming TV and telephone services. In terms of payment, many customers use monthly contracts rather than annual or bi-annual contracts, with most payment methods using electronic payments.

2) Preprocessing.

At this stage, the process of data processing will be explained before transforming the data and applying algorithms to the data or modeling. The first stage in this process is data separation or splitting by separating two data, namely Train data or data to be trained as much as 80% and Test data or test data as much as 20%, then proceed with handling missing data, extreme data and the number of missing data labels. unbalanced, this is applied so that the accuracy results on both labels are optimal and there are no errors or errors when implementing or testing the model.

a. Split and Separation of data. Separation of 80% data on training data and 20% on test data will be carried out using the Train-Test Split method, with the following code;

```
from sklearn.model_selection import tr ain_test_split
X_train, X_test, y_train, y_test = tra in_test_split(X, y, test_size=0.2, random_state=0)
```

b. *Handling at extreme values (Handling Outliers).* This stage is carried out by identifying outliers to find out the number of extreme values in the numerical variables. Identification of extreme values is done by creating a function that can identify outlier data and the distribution of the data. The first thing in identifying outlier data is to find the 1st and 3rd quartiles of the data. The process uses quantile() from the Numpy module.

c. *Handling missing data (Handling Missing Value).* The next stage is identifying the missing values (missing values) in each data variable because missing values can give errors during the process. If there is a missing value, the value will be imputed according to the missing value handling technique, but at this stage no value is found.

d. *Unbalanced data handling (Handling Imbalance).* The final stage of Preprocessing is handling unbalanced target label data using the Over Sampling method. Oversampling is a method of adding the minority data as much as the majority data.

3) Normalization.

Normalization is an existing stage in data transformation. Data transformation is the stage where the data will be converted into a matrix form that can be processed by the classification algorithm. At this stage normalization is carried out for both categorical and numeric data types.

a. *Normalization of data on categorical data (Encoding).* In categorical data, the results obtained from the previous visualization are if the variable has ordinal properties or has a class in its value. Normalization uses the OrdinalEncoder method from the Sklearn module because this method is used for categorical data types that have classes.

b. *Normalization of data on numerical data (scaling).* On categorical data normalization is performed using the MinMaxScaler method from the Sklearn module. This is because this method is usually used for numerical data types that have a normal distribution of data. It should be noted that based on the results of data analysis, numeric type data has a normal distribution.

4) Modeling

After the preprocessing stage is complete, the next stage is to build a classification model for customer decisions based on ready-to-use data. There are 4 models at this stage, namely models from the Logistic Regression algorithm, Support Vector Machine, Decision Tree and Random Forest.

a. *Modeling with Logistic Regression*. The modeling results obtained an accuracy of 80% with a precision of 79% in predicting customers who will subscribe and 80% for customers who will not subscribe. The modeling results are shown in Figure 1.

```
Accuracy - Train Set   :  0.7950641180740382

Classification Report :
                    precision    recall  f1-score   support

      Berlangganan       0.79      0.81      0.80      4133
Tidak berlangganan       0.80      0.78      0.79      4133

          accuracy                           0.80      8266
         macro avg       0.80      0.80      0.80      8266
      weighted avg       0.80      0.80      0.80      8266
```

Figure 1 : Logistic Regression Model Accuracy

b. *Modeling with Support Vector Machine*. The results of this modeling obtained an accuracy of 82% with a precision of 82% in predicting customers who will subscribe and 81% for customers who will not subscribe. The modeling results are shown in Figure 2.

```
Accuracy - Train Set   :  0.8158722477619162

Classification Report :
                    precision    recall  f1-score   support

      Berlangganan       0.82      0.81      0.81      4133
Tidak berlangganan       0.81      0.82      0.82      4133

          accuracy                           0.82      8266
         macro avg       0.82      0.82      0.82      8266
      weighted avg       0.82      0.82      0.82      8266
```

Figure 2 : Support Vector Machine Model Accuracy

c. *Modeling with Decision Trees*. The results of this modeling obtained 99% accuracy with 99% precision in predicting customers who will subscribe and 100% for customers who will not subscribe. The modeling results are shown in Figure 3.

```
Accuracy - Train Set   :  0.993225260101621

Classification Report :
                    precision    recall  f1-score   support

      Berlangganan       0.99      1.00      0.99      4133
Tidak berlangganan       1.00      0.99      0.99      4133

          accuracy                           0.99      8266
         macro avg       0.99      0.99      0.99      8266
      weighted avg       0.99      0.99      0.99      8266
```

Figure 3 : Decision Tree Model Accuracy

d. *Modeling with Random Forests*. The results of this modeling obtained 99% accuracy with 100% precision in predicting customers who will subscribe and 99% for customers who will not subscribe. The modeling results are shown in Figure 4.

```
Classification Report :
                    precision    recall  f1-score   support

      Berlangganan       0.99      1.00      0.99      4133
Tidak berlangganan       1.00      0.99      0.99      4133

          accuracy                           0.99      8266
         macro avg       0.99      0.99      0.99      8266
      weighted avg       0.99      0.99      0.99      8266


Accuration Prediction :
0.993225260101621
```

Figure 4 : Random Forest Model Accuracy

5) Evaluation.

This section discusses model evaluation to determine the right method for the application. Evaluation is carried out using the Confusion Matrix metric to test model performance and Cross Validation to test the validity of model accuracy.

a. Evaluation based on the results of the Confusion Matrix

For the logistic regression model, the results of this modeling show an accuracy of 80% with a precision of 79% in predicting customers who will subscribe and 80% for customers who will not subscribe. This algorithm can predict as many as 3353 correct positives and 780 incorrect positives, 3219 correct negatives and 914 wrong negatives. From these results it can be said that the performance of this model is quite good because there are only a few errors in predicting negatives.

Meanwhile, the Support Vector Machine model yields 82% accuracy with 82% precision in predicting customers who will subscribe and 81% for customers who will not subscribe. This algorithm can predict as many as 3351 correct positives and 782 incorrect positives, 3393 correct negatives and 740 wrong negatives. From these results it can be said that the performance of this model is quite good because there are only a few errors in predicting negatives.

In the Decision Tree model, 99% accuracy is obtained with 99% precision in predicting customers who will subscribe and 100% for customers who will not subscribe. This algorithm can predict 4122 correct positives and 11 incorrect positives, 4088 correct negatives and 45 wrong negatives. From these results it can be said that the performance of this model is quite good because there are only a few errors in predicting negatives. However, this result is doubtful because results that are too perfect can indicate Overfitting or too much learning model which can lead to errors.

Meanwhile, the Random Forest model obtained an accuracy of 99% with 100% precision in predicting subscribers who will subscribe and 99% of customers who will not subscribe. This algorithm can predict 4117 correct positives and 16 incorrect positives, 4093 correct negatives and 40 wrong negatives. From these results it can be said that the performance of this model is quite good because there are only a few errors in predicting negatives. As with the decision tree model, the results from the random forest are a little doubtful because the results are close to perfect so that it can be indicated as Overfitting or the model learns too much which can result in errors.

b. Evaluation based on Cross Validation

Based on the results of validation using Cross Validation with 10 validation tests based on test data, it is found that the logistic regression model has a range from 77% to 81% with an average accuracy of 79.2% and a standard deviation of 0.02% (see Figure 5). Meanwhile for the support vector machine algorithm, with 10 validation tests based on test data, it is obtained if the model has a range from 75% to 83% with an average accuracy of 79.3% and a standard deviation of 0.03% (see Figure 6). Validation results using Cross Validation for the Decision Tree are obtained if the model has a range from 73% to 83% with an average accuracy of 78.4% and a standard deviation of 0.05% (see Figure 7), while the Random Forest has a range from 78% to 90% with an average accuracy of 84.3% and a standard deviation of 0.05% (see Figure 8).

```
Accuracy - All - Cross Validation  :  [0.77750907 0.75090689 0.76783555 0.7980653  0.7980653  0.7859734
 0.81590863 0.79782082 0.81961259 0.80992736]
Accuracy - Mean - Cross Validation :  0.7921696906172138
Accuracy - Std - Cross Validation  :  0.020700137307575993
Accuracy - Range of Test-Set       :  0.7714695533096377 - 0.8128698279247898
CPU times: user 1.61 s, sys: 1.2 s, total: 2.81 s
Wall time: 1.55 s
```

Figure 5 : Cross Validation Results for Logistic Regression

```
Accuracy - All - Cross Validation  :  [0.74002418 0.73035067 0.74727932 0.78718259 0.81620314 0.80773881
 0.82566586 0.82687651 0.8401937  0.81840194]
Accuracy - Mean - Cross Validation :  0.7939916732786612
Accuracy - Std - Cross Validation  :  0.038325662377512386
Accuracy - Range of Test-Set       :  0.7556660109011488 - 0.8323173356561735
CPU times: user 22.9 s, sys: 128 ms, total: 23 s
Wall time: 22.9 s
```

Figure 6 : Cross Validation Results for Support Vector Machine

```
Accuracy - All - Cross Validation  :  [0.70133011 0.7267231  0.69770254 0.76783555 0.83071342 0.80290206
 0.8220339  0.8401937  0.82929782 0.82687651]
Accuracy - Mean - Cross Validation :  0.7845608708509124
Accuracy - Std - Cross Validation  :  0.05371627259775443
Accuracy - Range of Test-Set       :  0.730844598253158 - 0.8382771434486669
CPU times: user 555 ms, sys: 2.74 ms, total: 558 ms
Wall time: 566 ms
```

Figure 7 : Cross Validation Results for Decision Tree

```
Accuracy - All - Cross Validation  :  [0.75574365 0.7617896  0.76541717 0.808948   0.88875453 0.87908102
 0.8874092  0.89830508 0.89709443 0.89225182]
Accuracy - Mean - Cross Validation :  0.8434794510922232
Accuracy - Std - Cross Validation  :  0.05929859170820139
Accuracy - Range of Test-Set       :  0.7841808593840218 - 0.9027780428004246
CPU times: user 9.12 s, sys: 49 ms, total: 9.17 s
Wall time: 9.19 s
```

Figure 8 : Cross Validation Results for Random Forest

6) Determination.

Although based on the test results, there is an assumption that the Random Forest is indicated to be over-fitting, but by looking at the level of accuracy of the algorithm, information is obtained that the Random Forest has the best accuracy among the other three algorithms. Thus, Random Forest is the most appropriate algorithm for classifying customer decisions.

4 Conclusion

Based on the results of the research conducted, the conclusions are as follows: 1) The model can be applied properly if the data handling process is right on the right data; 2) Random Forest is the most appropriate algorithm model for this data; 3) For unbalanced data, the Oversampling method can be applied; 4) Determining the right algorithm is by testing several algorithms by comparing the results of valid model performance using the Confussion Matrix and Cross Validation methods; and 5) From 7000 data with 20 variables, it turns out that the model can be trained properly.

There are several suggestions that can be given based on this research, including: 1) There are many variables with discrete data types that make it difficult to convert, so more handling of the data is needed; 2) In testing the model, proper parameter testing or Hyper Parameter Tuning was not carried out on the exclusive clustering

algorithm model, only using default parameters, so further testing is still needed by conducting parameter experiments; 3) In future studies, it is hoped that the data collected will be more numerous and more balanced so that the modeling results will provide perfect results.

BIBLIOGRAPHY

[1] Adrian, M. R. dkk. (2021) Jurnal Informatika Upgris, 7(1). doi: 10.26877/jiu.v7i1.7099 Perbandingan Metode Klasifikasi Random Forest dan SVM Pada Analisis Sentimen PSBB

[2] Ahn, J.H., S.P. Han, and Y.S. Lee. 2019. Policy. 30: 552-568.Customer churn analysis: Churn determinants and mediation effects of partial defection in the Korean mobile telecommunications service industry. Telecomunications

[3] Adiwijaya, Said Al-Faraby (2020) Jurnal Skripsi Universitas Telkom BandungAnalisis Churn Prediction pada Data Pelanggan PT. Telekomunikasi Menggunakan Underbagging dan Logistic Regression

[4] Nabi-Abdolyousefi, R. (2019). Conversion Rate Prediction in Search Engine Marketing. Grad. School of Natural and Applied Sci. - Thesis & Dissertations. Istanbul Conversion Rate Prediction in Search Engine Marketing. Grad. School of Natural and Applied Sci

[5] Aldi Nurzahputra, Afifah Ratna Safitri, Much Aziz Muslim FMIPA, Universitas Negeri Semarang (2019) Klasifikasi Pelanggan pada Customer Churn Prediction Menggunakan Decision Tree

[6] Meilina, Popy., 2020.. Jurnal Teknologi. 7(1), 12-30 Penerapan Data Mining dengan Metode Klasifikasi Menggunakan Decision Tree dan Logistic Regresion19

[7] Govindaraju, R., Simatapung, T., & Samadhi, A, T., 2008. Jurnal Informatika. 9(1), 33-42. Perancangan Sistem Prediksi Churn Pelanggan PT. Telekomunikasi Seluler Dengan Memanfaatkan Proses Data Mining.

[8] Jurnal Informatika Vincent Angkasa, Jefri Junifer Pangaribuan Volume 7 No.1 Tahun 2022 Komparasi Tingkat akurasi Random Forest Untuk Mendiagnosis Penyakit Kanker Payudara

[9] L. Indah Prahartiwi, W. Dari, and S. Nusa Mandiri, J. Tek. Komput. AMIK BSI, vol. 7, no. 1, 2021, doi: Komparasi Algoritma Naive Bayes, Decision Tree dan Support Vector Machine untuk Prediksi Pelanggan

[10] Samsudiney. Artikel Ilmiah: Penjelasan apa itu SVM ? https://medium.com/@samsudiney/penjela san-sederhana-tentang-apa-itu-svm- 149fec72bd02

[11] Anwar hidayat. Jurnal Statistika: Regresi Logistik Vol.3 No.1 2015 https://www.statistikian.com/2015/02/regresilogistik.html

[12] Rimbun Siringoringo Universitas Methodist Indonesia Klasifikasi Data Tidak Seimbang Menggunakan Algoritma SMOTE dan KNN Jurnal ISD Vol.3 No.1 Januari – Juni 2018 https://ejournalmedan.uph.edu/index.php/isd/article/viewFile/177/63

[13] Jianfeng Xu, Yuanjian Zhang, Duogian Miao.Three-way confusion matrix for classification: A measure driven view. Information Sciences Vol. 507, January 2020. Elsevier. https://doi.org/10.1016/j.ins.2019.06.064

[14] Daria, Devi. Analisis Kelompok Wilayah Rawan Penyakit Malaria di Provinsi Nusa Tenggara Timur Tahun 2014 Vol. 1 No. 1 2016.https://dspace.uii.ac.id/bitstream/handle/123456789/537/05.3%20bab%203.pdf?sequence=9&isAllowed =y

[15] Vercellis, Carlo. Business intelligence : Datamining and optimization for decision making. Chichester: John Wiley& Sons. 2009

[16] Fayyad. Knowledge Discovery and Data Mining: Towards a Unifying Framework. KDD-96 Proceedings. 1996 https://www.aaai.org/Papers/KDD/1996/KDD96-014.pd