Journal of Emerging Information Systems and Business Intelligence
ISSN: 2774-3993
Journal homepage: https://ejournal.unesa.ac.id/index.php/JEISBI/

Prediction and Analysis Customer Churn at Telkomsel Using a Machine Learning Approach

Achmad Mauludi Asror[1], I Kadek Dwi Nuryana[2]

[1,2] *Universitas Negeri Surabaya, Surabaya, Indonesia*

achmad.19089@mhs.unesa.ac.id, mazz.asrorl@gmail.com, dwinuryana@unesa.ac.id

ABSTRACT

Customer churn is one of the main problems in the telecommunications industry, including Telkomsel, the largest cellular operator in Indonesia. This study aims to build a classification model to predict customer churn and analyze the factors influencing churn using the CRISP-DM approach. Data was obtained through an online questionnaire from 100 respondents who are active students of Universitas Negeri Surabaya. The research process includes stages of data preparation (normalization, encoding, and removal of irrelevant attributes) and the application of classification algorithms such as Logistic Regression, Decision Tree, Random Forest, K-Nearest Neighbors, Support Vector Machine, and Naïve Bayes. Evaluation was carried out using metrics such as accuracy, precision, recall, and F1-Score. The results show that Random Forest is the best-performing algorithm with an F1-Score of 87.50% using an 80:20 data split. Feature analysis indicates that the attribute of previous churn status has the greatest influence on churn prediction. This study is expected to help Telkomsel understand customer behavior patterns and develop more effective strategies to improve customer retention. The results of this model can serve as a basis for business decision-making, such as designing customer loyalty programs or adjusting services to better meet user needs.

Keyword: Customer Churn, Algorithm Comparison, Telkomsel, Classification, CRISP-DM.

Article Info:	**Corresponding Author**
Article history:	Achmad Mauludi Asror
Received January 23, 2025	Universitas Negeri Surabaya, Surabaya, Indonesia
Revised July 14, 2025	achmad.19089@mhs.unesa.ac.id
Accepted July 21, 2025	

1. INTRODUCTION

With the advancement of technology, the number of mobile phone users in Indonesia continues to grow. The total number of users has increased by 2.67%, rising from 210.03 million to 215.63 million individuals [1]. Telkomsel, as the leading mobile operator in Indonesia, serves 151.1 million customers as of 2023. Given its large market share, Telkomsel must ensure high-quality service to retain its users. A significant challenge in the telecommunications industry is customer churn, where customers stop using a company's services. A high churn rate can negatively impact revenue and increase marketing costs needed to attract new customers [2]. Therefore, leveraging technology-driven strategies is crucial to anticipate churn and improve customer retention. Customer churn not only provides significant protection for business benefits but also results in effective savings on employment costs. In this regard, customer churn depends on the extent to which customer needs and wants are met by the current operator [3].

To address this issue, this research applies machine learning techniques to predict customers likely to churn. The models implemented include six classification algorithms: Logistic Regression, Decision Tree, K-Nearest Neighbors, Naïve Bayes, Support Vector Machine (SVM), and Random Forest. These algorithms are selected for their ability to analyze customer historical data and identify patterns contributing to churn. Prior studies indicate that retaining existing customers is far more cost-effective than acquiring new ones, making accurate churn prediction a valuable asset for businesses [4]. This study focuses on two key research questions. First, how can the churn prediction process be implemented using the six classification models mentioned earlier? Second, how do these models compare in terms of performance in predicting churn rates? The evaluation process involves metrics such as accuracy, precision, recall, and F1-score to determine the most effective algorithm for handling customer churn [5].

The primary objective of this research is to develop a churn prediction system that is not only accurate but also capable of providing insights into the factors influencing customer turnover. Additionally, this study aims to benefit multiple stakeholders, including the researcher in gaining a deeper understanding of classification models, Telkomsel in improving customer retention strategies, and Universitas Negeri Surabaya (UNESA) as an academic reference for future studies in artificial intelligence and data analytics. The study has several limitations, such as using primary data collected through online questionnaires distributed to active UNESA students who have previously used Telkomsel services. The data is processed using Google Colab, with classification algorithms serving as the main method for churn analysis. These constraints help maintain the research focus and ensure that the findings remain relevant to the established scope.

2. METHODS

In this study, the CRISP-DM (Cross Industry Standard Process for Data Mining) methodology was used. The following are the steps of the CRISP-DM methodology, it explained in Figure 1.

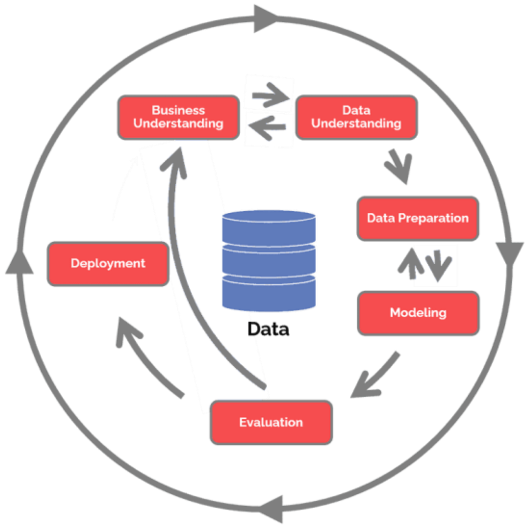


Figure 1. CRISP-DM Sequence of Steps

2.1. Business Understanding

Business understanding is the phase that requires an understanding from the business object perspective, how to build and acquire data, and align it with the modelling objectives to achieve business goals, ultimately obtaining the best model.

2.2. Data Understanding

Data understanding is the phase of understanding the condition of the dataset and identifying any potential issues within it. In this phase, researchers collect the necessary data to answer research questions or address the identified problems. This involves gathering, exploring, and describing the available data.

2.2.1. Data Collection

This phase is carried out to prepare several things needed before data collection. Data collection in research allows the researcher to be absent. One way to do this is through a questionnaire, where the researcher's questions and the respondents' answers can be presented in writing. This technique assigns respondents to read and answer the questions.

2.2.2. Attribute Selection

Attribute selection is performed to extract specific variables from the initial data collected. According to [6], the performance of predictive models can be influenced by feature extraction in terms of prediction accuracy. There is a set of new features for predicting customer churn in the telecommunications sector, which are explained as follows: Customer Name, Age, Gender, Location, Card Activation, Previous Churn, Number of Calls, Number of SMS, Data Package Usage, Expenditure Costs and Churn.

According to [6], the independent variables include Age, Gender, Card Activation, Previous Churn, Number of Calls, Number of SMS, Data Package Usage, and Expenditure Costs. The dependent variable used is the Customer Churn status with the output being either Churned (stopped) or Not Churned (continued).

2.3. Data Preparation

Data preparation is the process of preparing raw data so that it is suitable for further processing and analysis. This includes gathering, cleaning, and labelling the raw data into a suitable format. In this phase, researchers prepare the dataset that has been distributed and processed to be analyzed using existing algorithms. The process involves cleaning the data, merging data from various sources, filling missing values, and transforming the data into a format appropriate for the modelling being conducted.

2.4. Modelling

The modelling phase is the process of creating a model using machine learning algorithms that are appropriate for the characteristics of the dataset. In this phase, this research uses the Scikit-learn library to build and evaluate the generated model. This process involves several machines learning algorithms, including Logistic Regression, Random Forest, Decision Tree, K-Nearest Neighbours (KNN), Support Vector Machine (SVM), and Naïve Bayes.

In this research, each algorithm will be tested and compared for its performance. The purpose of using multiple algorithms is to determine which algorithm provides the best evaluation results in classifying the data. The modelling process is carried out through several stages as follows:

a. Algorithm selection: The algorithms used are Logistic Regression, Random Forest, Decision Tree, KNN, SVM, and Naïve Bayes.
b. Data splitting: The data is divided into three ratios, which are 90:10, 80:20, and 70:30 to observe the impact of the data ratio on the model's performance.

c. Model training and testing: Each algorithm is trained using the training data and then tested with the testing data.

2.5. Evaluation

The evaluation and validation phase are conducted after the model training and testing processes are complete. The data, which has been processed and split into three ratios (90:10, 80:20, 70:30), will be evaluated using specific performance metrics. The evaluation is performed on the prediction results of each algorithm, using a confusion matrix as a tool to calculate the model's performance metrics. The performance metrics used are Accuracy, Precision, Recall, and F1-Score.

2.6. Deployment

The deployment phase is the stage where the model is planned for use and integrated with operational system decisions. Although the model is deployed, it still needs to be monitored and replaced with a better model in the future. The final stage in adjusting the best model is finalization.

3. RESULTS AND DISCUSSION

This section explains the results of each stage outlined previously, from business understanding to deployment.

3.1 Business Understanding

This research aligns with the business challenges faced by telecommunications companies, particularly Telkomsel, which seeks to determine how many customers have churned. According [7], in an official statement from Telkom in May 2022, the number of Telkomsel customers at the end of March 2022 reached 175 million, with mobile data users accounting for 119.8 million, a growth of 4.3% compared to the previous year [7]. In a separate statement from Telkom in April 2023, Telkomsel reported serving 151.1 million customers. Therefore, with the implementation of a Data Mining algorithm model that can predict customer churn, Telkomsel will be able to identify how many of their customers have churned, specifically among students of Universitas Negeri Surabaya. If a predictive model with high evaluation metrics is achieved, it will assist company management in taking appropriate actions toward these customers.

3.2 Data Understanding

Data understanding is a crucial step in this research. The following presents the data analysis.

3.2.1. Gender

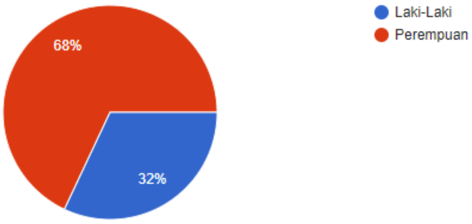


Figure 2. Respondent Based on Gender

In this study, out of 100 respondents based on Figure 2, female respondents dominated the questionnaire responses, accounting for 68%, while the remaining 32% were male respondents. Based on this data, it can be concluded that more females use Telkomsel services.

3.2.2. Telkomsel Service User Data

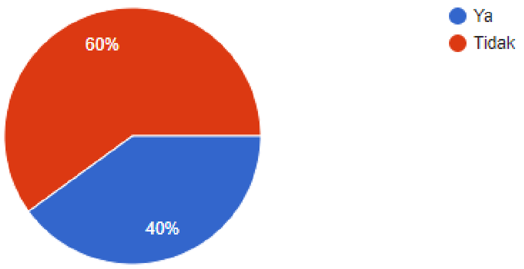


Figure 3. Telkomsel Service User Data

A total of 100 respondents were collected for this study, and the researcher categorized them into two groups to identify Telkomsel service users: respondents in the churn category and those in the non-churn category. As explained in Figure 3, respondents in the churn category accounted for 40%, while the non-churn category accounted for 60%. Based on this data, it can be concluded that while many respondents remain subscribed as Telkomsel users, a significant number have also stopped their subscriptions.

3.3 Data Preparation

3.3.1. Analysing Correlations Between Variables

In this stage, a correlation matrix is used to evaluate the relationships between variables in the dataset. This matrix provides an initial overview of the extent to which two variables have a linear relationship, measured using the correlation coefficient values. Correlation matrix shown in Figure 4.

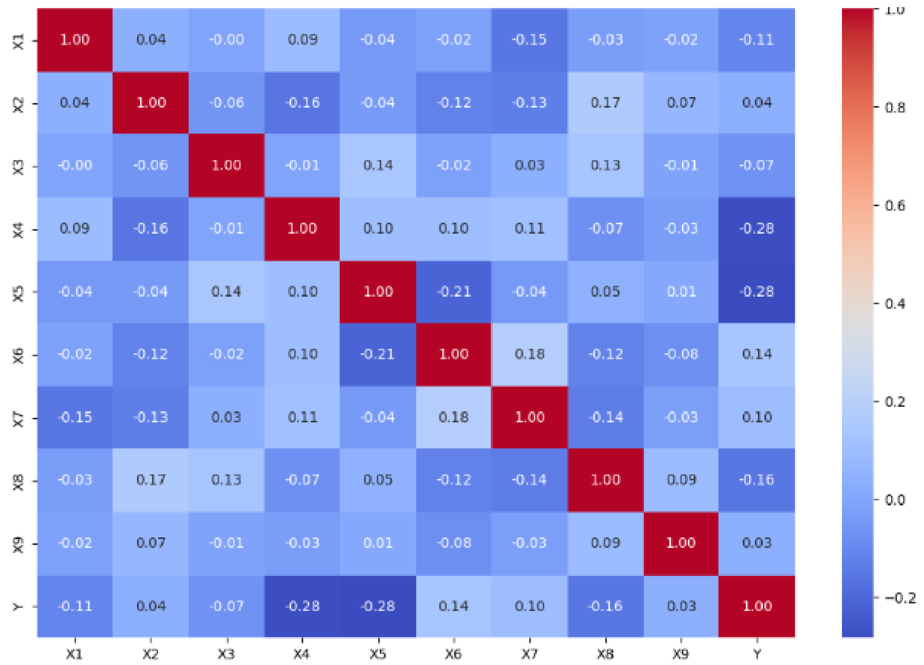
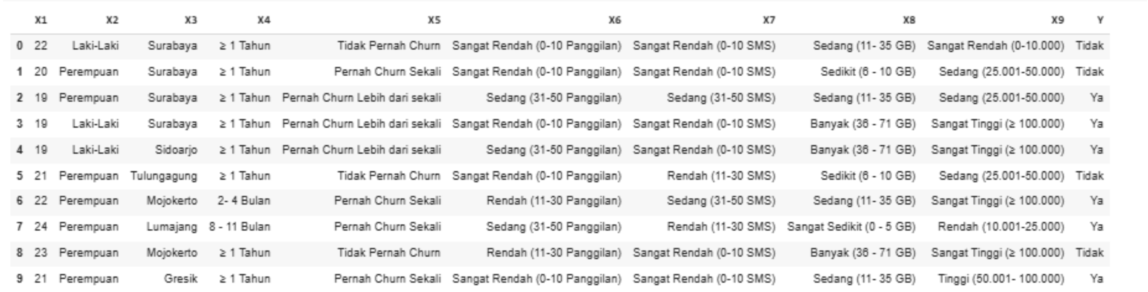


Figure 4. Correlation Matrix

3.3.2. Removing unnecessary columns

The removal of columns is an important step in data cleaning to eliminate irrelevant or redundant features. Some columns in the dataset, such as the Full Name column, are not relevant for analysis and modelling. Since this column does not significantly contribute to the churn prediction results, it is removed to make the dataset more efficient. The example of this process shown in Figure 5.

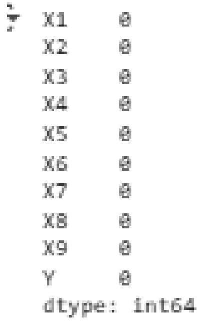


	X1	X2	X3	X4	X5	X6	X7	X8	X9	Y
0	22	Laki-Laki	Surabaya	≥ 1 Tahun	Tidak Pernah Churn	Sangat Rendah (0-10 Panggilan)	Sangat Rendah (0-10 SMS)	Sedang (11- 35 GB)	Sangat Rendah (0-10.000)	Tidak
1	20	Perempuan	Surabaya	≥ 1 Tahun	Pernah Churn Sekali	Sangat Rendah (0-10 Panggilan)	Sangat Rendah (0-10 SMS)	Sedikit (6 - 10 GB)	Sedang (25.001-50.000)	Tidak
2	19	Perempuan	Surabaya	≥ 1 Tahun	Pernah Churn Lebih dari sekali	Sedang (31-50 Panggilan)	Sedang (31-50 SMS)	Sedang (11- 35 GB)	Sedang (25.001-50.000)	Ya
3	19	Laki-Laki	Surabaya	≥ 1 Tahun	Pernah Churn Lebih dari sekali	Sangat Rendah (0-10 Panggilan)	Sangat Rendah (0-10 SMS)	Banyak (36 - 71 GB)	Sangat Tinggi (≥ 100.000)	Ya
4	19	Laki-Laki	Sidoarjo	≥ 1 Tahun	Pernah Churn Lebih dari sekali	Sedang (31-50 Panggilan)	Sangat Rendah (0-10 SMS)	Banyak (36 - 71 GB)	Sangat Tinggi (≥ 100.000)	Ya
5	21	Perempuan	Tulungagung	≥ 1 Tahun	Tidak Pernah Churn	Sangat Rendah (0-10 Panggilan)	Rendah (11-30 SMS)	Sedikit (6 - 10 GB)	Sedang (25.001-50.000)	Tidak
6	22	Perempuan	Mojokerto	2- 4 Bulan	Pernah Churn Sekali	Rendah (11-30 Panggilan)	Sedang (31-50 SMS)	Sedang (11- 35 GB)	Sangat Tinggi (≥ 100.000)	Ya
7	24	Perempuan	Lumajang	8 - 11 Bulan	Pernah Churn Sekali	Sedang (31-50 Panggilan)	Rendah (11-30 SMS)	Sangat Sedikit (0 - 5 GB)	Rendah (10.001-25.000)	Ya
8	23	Perempuan	Mojokerto	≥ 1 Tahun	Tidak Pernah Churn	Rendah (11-30 Panggilan)	Sangat Rendah (0-10 SMS)	Banyak (36 - 71 GB)	Sangat Tinggi (≥ 100.000)	Tidak
9	21	Perempuan	Gresik	≥ 1 Tahun	Pernah Churn Sekali	Sangat Rendah (0-10 Panggilan)	Sangat Rendah (0-10 SMS)	Sedang (11- 35 GB)	Tinggi (50.001- 100.000)	Ya

Figure 5. Removal of the Full Name Column

3.3.3. Check Missing Value

The next crucial step is to check for any missing values in the dataset. Missing values in a dataset can cause serious issues in analysis and modelling. In the dataset used, there were no missing values. Its shown in Figure 6.

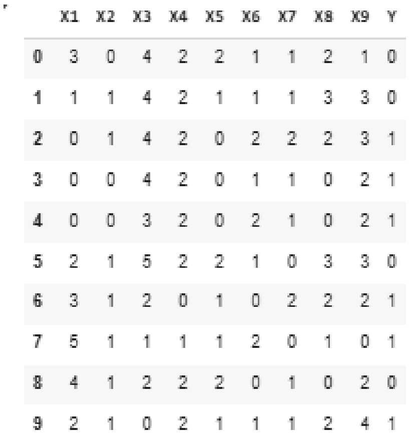


```
X1    0
X2    0
X3    0
X4    0
X5    0
X6    0
X7    0
X8    0
X9    0
Y     0
dtype: int64
```

Figure 6. Check Missing Value

3.3.4. Performing Encoding

The next step in data preprocessing is to convert categorical data into a numerical format. This is done to prepare the data for use with machine learning algorithms, which can only process numerical input. Each column, as shown in Figure 7, specified in categorical columns will be converted to a numerical format using the LabelEncoder method.

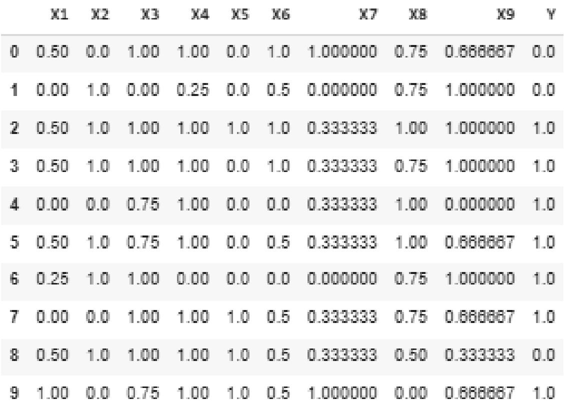


	X1	X2	X3	X4	X5	X6	X7	X8	X9	Y
0	3	0	4	2	2	1	1	2	1	0
1	1	1	4	2	1	1	1	3	3	0
2	0	1	4	2	0	2	2	2	3	1
3	0	0	4	2	0	1	1	0	2	1
4	0	0	3	2	0	2	1	0	2	1
5	2	1	5	2	2	1	0	3	3	0
6	3	1	2	0	1	0	2	2	2	1
7	5	1	1	1	1	2	0	1	0	1
8	4	1	2	2	2	0	1	0	2	0
9	2	1	0	2	1	1	1	2	4	1

Figure 7. Performing Encoding

3.3.5. Performing data normalization

Normalization is the process of transforming numerical features into a uniform scale. This is important to ensure that all features are within the same range, allowing the machine learning model to work with standardized data, which leads to more accurate predictions. The goal of this normalization process is to enhance the performance of the machine learning model, with the hope of reducing any bias that may arise from scale differences between variables.



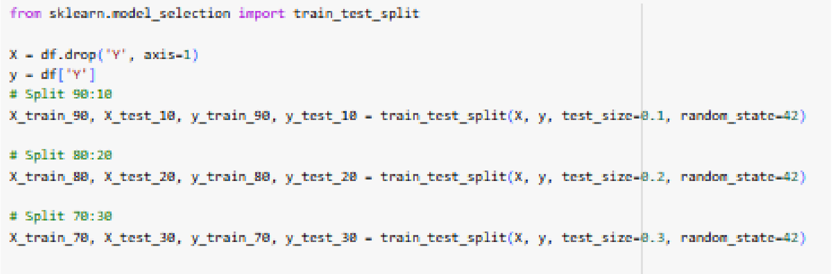
	X1	X2	X3	X4	X5	X6	X7	X8	X9	Y
0	0.50	0.0	1.00	1.00	0.0	1.0	1.000000	0.75	0.666667	0.0
1	0.00	1.0	0.00	0.25	0.0	0.5	0.000000	0.75	1.000000	0.0
2	0.50	1.0	1.00	1.00	1.0	1.0	0.333333	1.00	1.000000	1.0
3	0.50	1.0	1.00	1.00	0.0	1.0	0.333333	0.75	1.000000	1.0
4	0.00	0.0	0.75	1.00	0.0	0.0	0.333333	1.00	0.000000	1.0
5	0.50	1.0	0.75	1.00	0.0	0.5	0.333333	1.00	0.666667	1.0
6	0.25	1.0	1.00	0.00	0.0	0.0	0.000000	0.75	1.000000	1.0
7	0.00	0.0	1.00	1.00	1.0	0.5	0.333333	0.75	0.666667	1.0
8	0.50	1.0	1.00	1.00	1.0	0.5	0.333333	0.50	0.333333	0.0
9	1.00	0.0	0.75	1.00	1.0	0.5	1.000000	0.00	0.666667	1.0

Figure 8. Preforming Normalization Data

3.4 Modelling

3.4.1. Data splitting into test and training sets

In the modelling phase, the first step is to split the data into training and test sets. This split aims to evaluate the model's performance objectively by using data that was not utilized during the training process. In this study, the dataset contains a total of 100 observations. The researcher then prepares the feature and target data. In this dataset, (x) represents the feature or independent variables used to train the model, while (y) is the target or dependent variable that contains the values the model aims to predict. The data is split using three different ratios: 90:10, 80:20, and 70:30, to observe how varying data splits affect the model's performance. Figure 9 explains the example of data splitting.

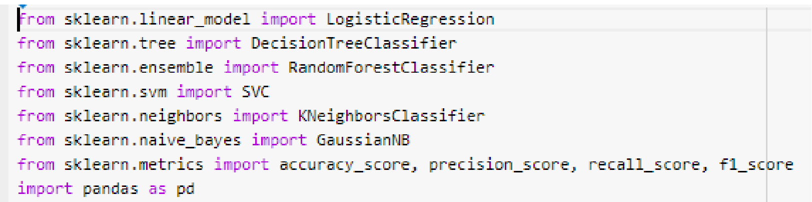


```
from sklearn.model_selection import train_test_split

X = df.drop('Y', axis=1)
y = df['Y']
# Split 90:10
X_train_90, X_test_10, y_train_90, y_test_10 = train_test_split(X, y, test_size=0.1, random_state=42)

# Split 80:20
X_train_80, X_test_20, y_train_80, y_test_20 = train_test_split(X, y, test_size=0.2, random_state=42)

# Split 70:30
X_train_70, X_test_30, y_train_70, y_test_30 = train_test_split(X, y, test_size=0.3, random_state=42)
```

Figure 9. Split Data

3.4.2. Comparing all algorithm models

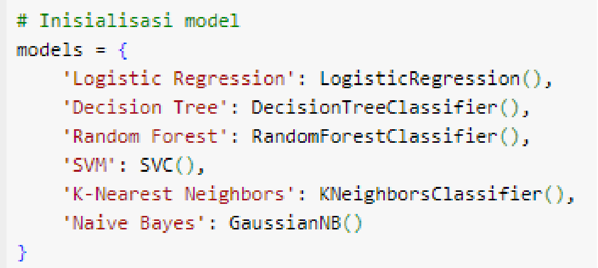
In this phase, a comparison is made among the six classification models selected for predicting Telkomsel customer churn. The models used are Logistic Regression, Decision Tree, Random Forest, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), and Naïve Bayes. The selection of these models is based on each algorithm's ability to handle binary classification problems, particularly in the context of customer churn. Figure 10 is shown the libraries used for modelling.



```
from sklearn.linear_model import LogisticRegression
from sklearn.tree import DecisionTreeClassifier
from sklearn.ensemble import RandomForestClassifier
from sklearn.svm import SVC
from sklearn.neighbors import KNeighborsClassifier
from sklearn.naive_bayes import GaussianNB
from sklearn.metrics import accuracy_score, precision_score, recall_score, f1_score
import pandas as pd
```

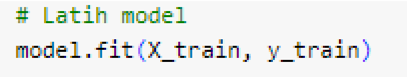
Figure 10. Import Algorithm Models

Next, as shown in Figure 11 and Figure 12, initialize several classification models in the form of a dictionary (models) :



```
# Inisialisasi model
models = {
    'Logistic Regression': LogisticRegression(),
    'Decision Tree': DecisionTreeClassifier(),
    'Random Forest': RandomForestClassifier(),
    'SVM': SVC(),
    'K-Nearest Neighbors': KNeighborsClassifier(),
    'Naive Bayes': GaussianNB()
}
```

Figure 11. Models Initialization used



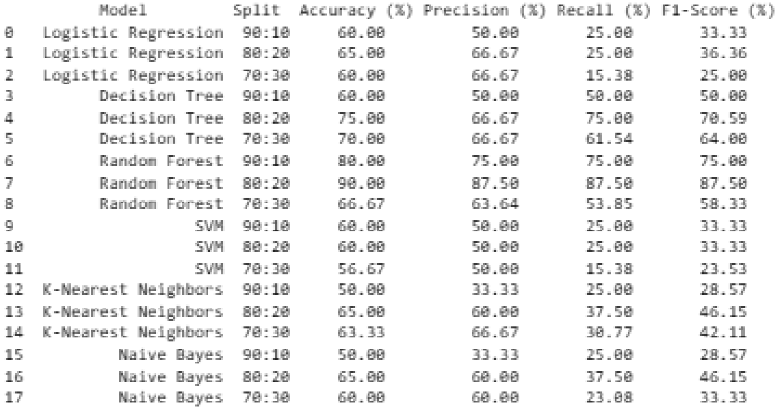
```
# Latih model
model.fit(X_train, y_train)
```

Figure 12. Training the model

3.5 Evaluation

3.5.1. Evaluation results of all algorithm models

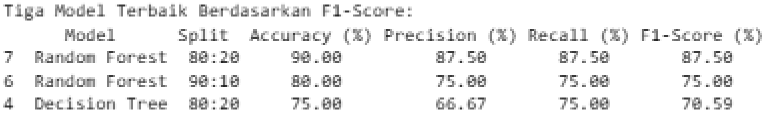
In this evaluation phase, I used several evaluation metrics to assess the performance of each model. The metrics used include Accuracy, Precision, Recall, and F1-Score. These metrics were chosen because they provide a comprehensive view of how well the model can predict the churn and non-churn classes in the dataset. Below are the evaluation results for each model.



```
                 Model  Split  Accuracy (%) Precision (%) Recall (%) F1-Score (%)
0   Logistic Regression  90:10      60.00          50.00      25.00        33.33
1   Logistic Regression  80:20      65.00          66.67      25.00        36.36
2   Logistic Regression  70:30      60.00          66.67      15.38        25.00
3         Decision Tree  90:10      60.00          50.00      50.00        50.00
4         Decision Tree  80:20      75.00          66.67      75.00        70.59
5         Decision Tree  70:30      70.00          66.67      61.54        64.00
6         Random Forest  90:10      80.00          75.00      75.00        75.00
7         Random Forest  80:20      90.00          87.50      87.50        87.50
8         Random Forest  70:30      66.67          63.64      53.85        58.33
9                   SVM  90:10      60.00          50.00      25.00        33.33
10                  SVM  80:20      60.00          50.00      25.00        33.33
11                  SVM  70:30      56.67          50.00      15.38        23.53
12  K-Nearest Neighbors  90:10      50.00          33.33      25.00        28.57
13  K-Nearest Neighbors  80:20      65.00          60.00      37.50        46.15
14  K-Nearest Neighbors  70:30      63.33          66.67      30.77        42.11
15          Naive Bayes  90:10      50.00          33.33      25.00        28.57
16          Naive Bayes  80:20      65.00          60.00      37.50        46.15
17          Naive Bayes  70:30      60.00          60.00      23.08        33.33
```

Figure 13. Comparison Evaluation Results

Next, identify the top 3 models based on their F1-scores, and the results are shown in the image below:



```
Tiga Model Terbaik Berdasarkan F1-Score:
          Model  Split  Accuracy (%) Precision (%) Recall (%) F1-Score (%)
7  Random Forest  80:20      90.00          87.50      87.50        87.50
6  Random Forest  90:10      80.00          75.00      75.00        75.00
4  Decision Tree  80:20      75.00          66.67      75.00        70.59
```

Figure 14. Top 3 Best Models Results

3.5.2. Confusion Matrix

The Confusion Matrix is used to measure the performance of each classification algorithm applied to the test data. It is a very useful tool for illustrating the model's performance by comparing the predicted results with the actual values. The results from the Confusion Matrix provide a more detailed picture of how well the model predicts each class (churn and non-churn). The Confusion Matrix shown will only be for the Random Forest algorithm because this model has the best performance compared to the others. This performance is measured using evaluation metrics such as Accuracy and F1-Score. Based on the testing results, Random Forest achieved the highest F1-Score of 87.50%, which indicates a balance between Precision and Recall in predicting customer churn. Below are the results of the Random Forest confusion matrix :

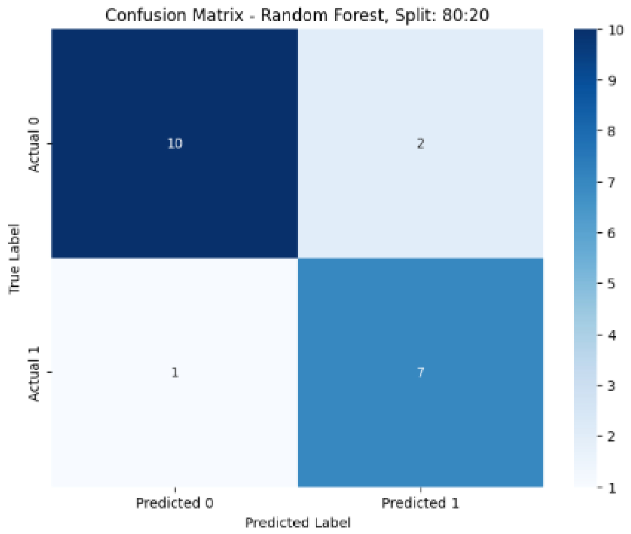


Figure 15. Confusion Matrix Results

The results from the confusion matrix that shown inf Figure 15, means:

a. **True Positive (TP)**: 7 → Positive data correctly predicted as positive.
b. **True Negative (TN)**: 10 → Negative data correctly predicted as negative.
c. **False Positive (FP)**: 2 → Negative data incorrectly predicted as positive.
d. **False Negative (FN)**: 1 → Positive data incorrectly predicted as negative.

3.5.3. Identifying the Most Significant Features

Analysis of the most significant features to evaluate the contribution of each feature in affecting the prediction results. Below are the results of the most significant features. The graph in Figure 16 shows that feature X5 has the greatest influence on the model, followed by feature X3, X1, and so on

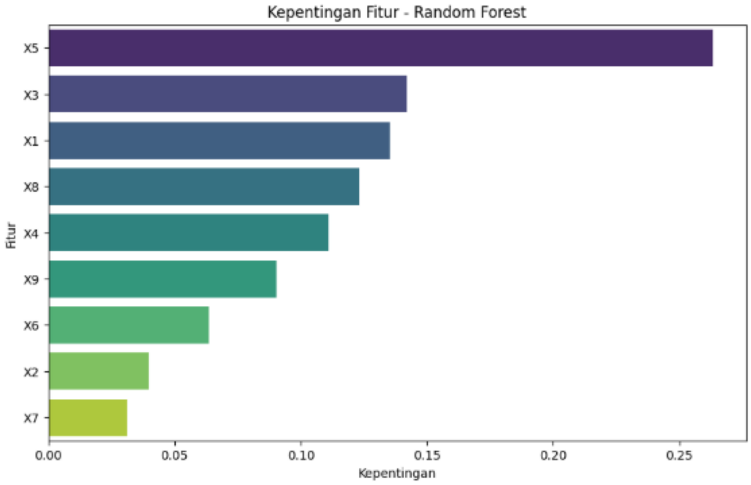
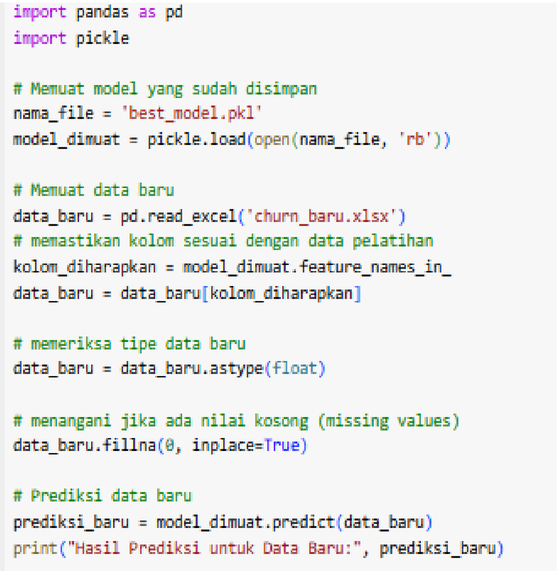


Figure 16. Graph of the Most Significant Features

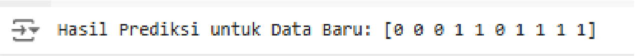
3.6 Deployment

Deployment is the process of applying a trained model into a production environment so that it can be used by users or systems to predict new data in real-time or in batches. This process ensures that the model is not just a prototype but can deliver real value in business decision-making. The machine learning model, after going through the training and evaluation processes, is saved in a specific file format, such as .pkl (pickle) or other formats like .joblib. Storing the model ensures that the structure and parameters remain intact, allowing it to be directly used without needing to retrain. In this research, the model is saved using the pickle file format.



```python
import pandas as pd
import pickle

# Memuat model yang sudah disimpan
nama_file = 'best_model.pkl'
model_dimuat = pickle.load(open(nama_file, 'rb'))

# Memuat data baru
data_baru = pd.read_excel('churn_baru.xlsx')
# memastikan kolom sesuai dengan data pelatihan
kolom_diharapkan = model_dimuat.feature_names_in_
data_baru = data_baru[kolom_diharapkan]

# memeriksa tipe data baru
data_baru = data_baru.astype(float)

# menangani jika ada nilai kosong (missing values)
data_baru.fillna(0, inplace=True)

# Prediksi data baru
prediksi_baru = model_dimuat.predict(data_baru)
print("Hasil Prediksi untuk Data Baru:", prediksi_baru)
```

Figure 17. Testing with New Data



```
Hasil Prediksi untuk Data Baru: [0 0 0 1 1 0 1 1 1 1]
```

Figure 18. Prediction Results for New Data

CONCLUSION

Based on the results and discussions, this research follows the Cross Industry Standard Process for Data Mining (CRISP-DM) methodology. In the data understanding phase, the data used was obtained through an online questionnaire, with a dataset of 100 respondents. The dataset includes several attributes, namely Age (x1), Gender (x2), Customer's Residence Area (x3), Subscription Duration (x4), Previous Churn Status (x5), Call Intensity (x6), SMS Intensity (x7), Internet Data Usage (x8), and Monthly Expenditure (x9). In the data preparation phase, several tasks were performed, including the removal of irrelevant columns, encoding categorical data, missing value inspection, and data normalization using the Min-Max Scaling method. In the modelling phase, the data was split into three ratios: 90:10, 80:20, and 70:30. Subsequently, six classification algorithms were applied to predict customer churn. Model evaluation was conducted using accuracy, precision, recall, and F1-Score metrics to assess the performance of each algorithm.

The comparison of algorithm models was evaluated using accuracy and F1-Score. A comparison of six algorithm models using data split ratios of 90:10, 80:20, and 70:30 on the customer churn dataset resulted in the best-performing model. The model with the highest evaluation score, an F1-Score of 87.50%, was the Random Forest model with a 80:20 data split ratio. Based on the attribute analysis, the feature that has the most significant influence on churn is the previous churn status

(X5). This indicates that customers who have previously churned are at a high risk of churning again

ACKNOWLEDGEMENTS

The author extends gratitude to Universitas Negeri Surabaya and all survey participants. Special thanks to academic mentors and colleagues for their valuable guidance and support

REFERENCES

[1] Ahdiat, (2023) Https://databoks.katadata.co.id/datapublish/2023/06/23/ini-operator-seluler-dengan-pengguna-terbanyak-di-indonesia-awal-2023.

[2] Husein, A. M., Harahap, M., & Fernandito, P. (2021). *Pendekatan Data Science untuk Menemukan Churn Pelanggan pada Sector Perbankan dengan Machine Learning. Data Sciences Indonesia* (DSI), 1(1), 8-13.

[3] Kun, L., Alli, H., & Abd Rahman, K. A. A. (2024). GA optimization-based BRB AI reasoning algorithm for determining the factors affecting customer churn for operators. Social Sciences & Humanities Open, 10, 100944.

[4] Utami, Y. T., Shofiana, D. A., & Heningtyas, Y. (2020). *Penerapan algoritma C4. 5 untuk prediksi churn rate pengguna jasa telekomunikasi. Jurnal Komputasi*, 8(2), 69-76.

[5] Shukla, S. N., & Marlin, B. M. (2019). *Interpolation-prediction networks for irregularly sampled time series*. arXiv preprint arXiv:1909.07782.

[6] Huang, B., Kechadi, M. T., & Buckley, B. (2012). Customer churn prediction in telecommunications. Expert Systems with Applications, 39(1), 1414-1425

[7] CNN Indonesia (2023) https://www.cnnindonesia.com/teknologi/20230727165359-213-978693/jumlah-pelanggan-telkomsel-anjlok-20-juta-dalam-setahun-cek-faktanya.

[8] Krishna, R., Jayanthi, D., Sam, D. S., Kavitha, K., Maurya, N. K., & Benil, T. (2024). Application of machine learning techniques for churn prediction in the telecom business. Results in Engineering, 24, 103165.

[9] Kirgiz, O. B., Kiygi-Calli, M., Cagliyor, S., & El Oraiby, M. (2024). Assessing the effectiveness of OTT services, branded apps, and gamified loyalty giveaways on mobile customer churn in the telecom industry: A machine-learning approach. Telecommunications Policy, 48(8), 102816.

[10] Maylani, I., Rochman, F., & Kurniasari, N. D. (2018). Seleksi Fitur pada Klasifikasi K-Nearest Neighbors untuk Data Churn for Bank Customers dengan Analisis Korelasi. Prosiding ISSN, 2775, 5126

[11] Wardani, N. W., Dantes, G. R., & Indrawan, G. (2018). Prediksi customer churn dengan algoritma decision tree C4. 5 berdasarkan segmentasi pelanggan untuk mempertahankan pelanggan pada perusahaan retail. Jurnal RESISTOR (Rekayasa Sistem Komputer), 1(1), 16-24.

[12] Nazar, Ridwan. "IMPLEMENTASI PEMROGRAMAN PYTHON MENGGUNAKAN GOOGLE COLAB." JIK: Jurnal Informatika dan Komputer 15, no. 1 (2024): 50-56

[13] Sihombing, J. (2021). Klasifikasi Data Antroprometri Individu Menggunakan Algoritma Naïve Bayes Classifier. BIOS: Jurnal Teknologi Informasi dan Rekayasa Komputer, 2(1), 1-10.

[14] Rohaeni, H., & Yuliyana, W. (2020). Pengaruh Harga dan Kualitas Produk Terhadap Loyalitas Pelanggan Telkomsel. Jurnal Sains Manajemen, 2(1), 37-44.

[15] Shukla, S. N., & Marlin, B. M. (2019). Interpolation-prediction networks for irregularly sampled time series. arXiv preprint arXiv:1909.07782