International Research Journal of Advanced Engineering and Science
ISSN (Online): 2455-9024

Customer Churn Classification Using Support Vector Machine (SVM) Algorithm

Masrina Manalu[1], Lily Wulandari[2]
Magister Manajemen Sistem Informasi, Sistem Informasi Bisnis, Universitas Gunadarma
Kampus III/5, Jl. Kenari I, RT.4/RW.5, Kenari, Senen, Jakarta Pusat, 10430
Email address: [1]masrinamanalu95@gmail.com, [2]lily@staff.gunadarma.ac.id

***Abstract*—** *Customers are an asset in running a business, this is a driving force for how important it is to have a strategy in retaining customers. Companies must be able to retain customers to maintain their business, with customer data owned by the company, it is expected to be able to identify the characteristics of all its customers and be able to retain existing customers so as not to stop buying or not switch to competing companies (churn). The classification of customer loyalty aims to identify customers who tend to switch to competing companies which are often called customer churn. The Support Vector Machine (SVM) algorithm is a classification algorithm in data mining that can function to predict customer loyalty. This study aims to analyze the prediction of customer churn using machine learning algorithms. The algorithm can construct predictive models, predict customer churn and display prediction results. The identification of churn prediction variables is based on a prediction model with selected variables, namely age, total purchase, account manager, years, num sites, and location. The total data used is 300 data with the distribution of 80% training data and 20% testing data. SVM produces a visual model that represents churn and non-churn customer behavior patterns. Tests conducted using customer data with linear kernel modeling resulted in a prediction model accuracy rate of 80.42%.*

***Keywords*—** *Classification, customer churn prediction, machine learning, support vector machine (SVM).*

I. INTRODUCTION

An increasingly competitive business environment and a lot of competition in the same field make companies have to maintain customer loyalty. Each company has its own way of offering quality and best services so as not to lose customers, this is of course troubling to the company, because if it is not prevented and handled it will result in a decrease in company revenue. Customers are the most important asset of any type of business. Business prospects are only possible with the presence of satisfied customers who are always loyal and build their relationship with the company. Customer churn is the percentage of customers who stop using a business's products and services during a certain period of time (Zeniarja et al., 2015). When a company loses its customers (customers stop using the products & services provided) the higher the churn rate percentage, so the growth rate of a business will be lower. Companies are competing to innovate in offering attractive products to customers. The customer discontinues service with the company, then the customer switches to the services of another company or competitor. This customer behavior is known as Churn. Customer churn is a condition that tends to discontinue subscriptions from certain products or services within a company. Customer churn occurs when a customer or subscriber stops doing business with a company or service. There are three types of Churn (Nurzahputra et al., 2016) :

- Active: Customers switch to other products or services due to dissatisfaction with service quality.
- Rotation / Incidental: The customer stops using the service but does not change or switch services that require the service.
- Discontinues: Passive or non-voluntary towards the contract itself. One way to deal with this problem is to predict which customers will Churn.

For many companies, finding reasons for losing customers, measuring customer loyalty and getting customers back are very important concepts to prevent customer churn. Churn is a process that can reduce company profits. Therefore, churn management becomes a crucial weapon in competition, and a foundation on a customer-oriented marketing effort. Agency companies that provide services in the field of predicting customer potential churn are very much needed. One way to predict Churn is to create a Churn prediction model based on the dataset. This is done by classifying the total customers into customers who are indicated to Churn and those who are not Churn. This method can also identify variables that influence customers to Churn. Thus, the company can adopt a retention policy that fits the needs of Churn's potential customers and is not late in making an offer. Churn prediction is an important business strategy for companies, by predicting customer churn, the company can immediately take action to retain customers, where the cost of getting new customers is much higher than retaining repeat customers. Therefore, the ability to predict customer churn is a must. Marketing agencies faced an increase in customer churn over the past period due to increased competition in the market. Marketing agencies generally have large customer data sets, but do not use them effectively, while marketing needs to identify outgoing customers in order to provide target-oriented advertising tools to retain those customers. In addition, agencies want to understand the factors that influence churn so that they can be more proactive in dealing with the problem rather than just reacting after the incident has occurred. The real problem and need is to reduce customer churn, stabilize business and increase profits, in retaining customers, telecommunications companies need a way to predict to find out the risk when customers will become churn (IMADUDDIN, 2014) Forecasting customer churn can be done with Data Mining techniques.

Masrina Manalu and Lily Wulandari, "Customer Churn Classification Using Support Vector Machine (SVM) Algorithm," *International Research Journal of Advanced Engineering and Science*, Volume 9, Issue 3, pp. 74-79, 2024.

International Research Journal of Advanced Engineering and Science
ISSN (Online): 2455-9024

Data mining is a process used to find hidden information through patterns or trends that exist in the database so as to find new information that is very useful. Data mining is an important step in doing Knowledge Discovery from Data. There are several stages of the process that must be passed according to (Suntoro, 2019):

- Data Cleansing: Stages of removing irrelevant data
- Data Integration: Combining a lot of data from different sources
- Data Selection: Relevant data analyzed
- Data Transformation : Where data is transformed and consolidated into a form suitable for mining by performing summary or aggregation operations
- Data Mining: The stage in which the process to extract certain patterns in the data
- Pattern Evaluation: Identify interesting patterns that can represent knowledge
- Knowledge Presentation: The last stage where the results of knowledge will be visualized and present knowledge to present the results that have been achieved.

Machine learning is a series of techniques that can assist in handling and predicting very large data by presenting the data with learning algorithms (Permana & Sahara, 2019). Machine learning algorithm is one of the analytical tools for customer churn where machine learning algorithms can find the characteristics that cause customer churn. This pattern is revealed automatically by the algorithm. Machine Learning algorithms can support in predicting customer data, when they will churn, and the level of the prediction. Companies can create promotions that aim to increase customer loyalty and improve strategies to get new customers and perform customer retention by predicting the number of customers who will replace services continuously.

Research predicting Churn customers has been done before. Research conducted by (Risky Novendri, 2021) Testing Data using RFM (R(recencies), F(frequency) and M(monetary value) Analysis. The Naïve Bayes algorithm produces the highest accuracy of 83.02%. The results of this accuracy predict customers The correct non-churn customer prediction results are 38551. The incorrect non-churn customer prediction results are 9449. From the non-churn customer prediction results, the results obtained are 84.90% precision and 80.31% recall resulting in an F1-measure of 82.54% Then for customer churn predictions, the model predicts the correct customer churn predictions are 41147 predictions and the wrong prediction results have a total of 6853 predictions. Then from the results of the customer churn predictions, the precision is 81.43% and the recall is 85.56 so as to produce F1 -measure of 83.44%.Other research is research conducted by the Support Vector Machine Method as a Determinant of Student Graduation, the highest test results using will support vector with an accuracy value of 85.02%, and an AUC value of 0.610. Compared to using only the Support Vector Machine (SVM) this result is higher with a difference in accuracy of 1.77% and an AUC difference of 0.004. variables that are significant to the model. (Rizqi Agung Permana, 2019).

II. LITERATUR REVIEW

The following is a picture of 1 step or stage in research which can be described as follows:

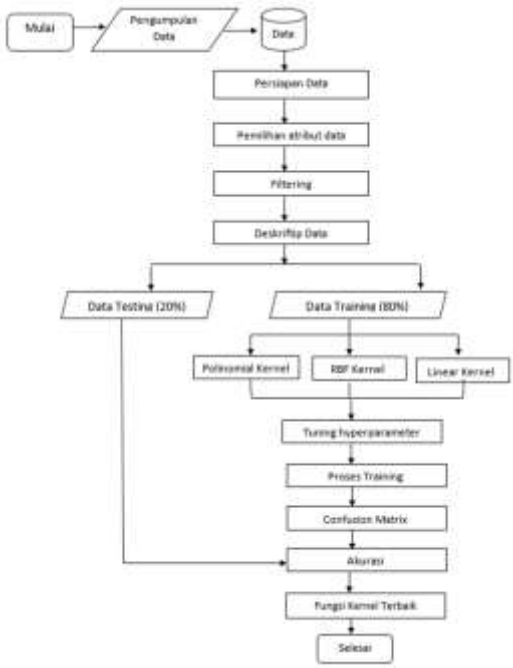

Figure 1. Research Stages

1. Data Collect

Data collect in this study was obtained by downloading the customer churn dataset in CSV format from the Kaggle page by hassanamin. The dataset obtained is then placed in a working program folder. The goal is to simplify the data input process. Figure 2 is a display of the data path settings.

Figure 2. Data Path Settings

2. Data

The data used is customer churn of a marketing agent, consisting of 10 variables, "Churn" as a response variable, which can be seen in table 1.

3. Data Preparation

This study uses a public dataset which is the data of a marketing agency. The goal is to find patterns of customer behavior based on demographic, billing and behavioral data so that customers can be distinguished into loyal and unfaithful customers. In the final result, there are 10 attributes (Table 2). Of the 10 attributes, an attribute selection is made to separate the label attribute (the attribute that will be used as a

Masrina Manalu and Lily Wulandari, "Customer Churn Classification Using Support Vector Machine (SVM) Algorithm," *International Research Journal of Advanced Engineering and Science*, Volume 9, Issue 3, pp. 74-79, 2024.

prediction key) with the predictive variable attribute in predicting a data.

TABLE 1. Data *Customer*

No	Variable	Description
1	*Name*	Nama terakhir yang menghubungi perusahaan
2	*Age*	Umur *customer*
3	Total *Purchase*	Total iklan yang dibeli
4	*Account_Manager*	*Binary* 0=*No manager*, 1= *Account manager assigned*
5	*Years*	*Totaly Years as a customer*
6	*Num sites*	Jumlah situs *web* yang menggunakan layanan
7	*Onboard_date*	*Date that the name of the latest contact was onboarded*
8	*Location*	Alamat *client*
9	*Company*	Nama perusahan *client*
10	*Churn*	0=*No*, 1=*Yes*

In this dataset, the attributes with values of no and yes are selected as labels and the others are used as variables. The data is selected and has been reduced to input variables and then the data is integrated into a single table that combines the input variables and their output variables. Once the integrated dataset has been created. The next step is to equalize the entry format for the existing input variables. The following is a table that describes the characteristics of each input variable. Caterogical data types are data types in the form of text while continuous data types are numeric data types.

TABLE 2. Input Variable Format

No	Variable Input	Type	Unit
1	*Name*	*Caterogical*	
2	*Age*	*Continous*	Tahun
3	Total_*Purchase*	*Continous*	Tahun
4	*Account_Manager*	*Continous*	
5	*Years*	*Continous*	
6	*Num sites*	*Continous*	
7	*Onboard date*	*Continous*	
8	*Location*	*Continous*	
9	*Company*	*Continous*	
10	*Churn*	*Continous*	

4. Data Attribute Selection

Churn indicator which contains information on whether or not a customer is active or not taken from the customer history database is used as an output variable. This variable will be predicted by the model based on the existing inputs.

The sales data collection contains information about all sales made by the customer. This dataset lists the purchase value of each customer based on the services used. However, in making this churn prediction model the input variable taken from this database is Account manager. The input variable is used as an average per customer because it makes it easier to input data because this model can be carried out on an ongoing basis. Customers can continue to be evaluated because there is a total_purchase per month. Each customer has complete data from these two input variables. Unlike Num_site where not all customers use this facility. Therefore, the variable is not included as an input variable for the model. Another step is to select a variable that determines customer turnover.

TABLE 3. Table of selected variables

No	Variable	Description
1	*Age*	Umur *customer*
2	Total *Purchase*	Total iklan yang dibeli
3	*Account_Manager*	*Binary* 0 = *No Manager,* 1 = *Account Manager Assignment*
4	*Years*	*Totally as a customer*
5	*Num sites*	Jumlah situs *web* yang menggunakan layanan
6	*Churn*	0=No, 1=Yes

5. Filtering

After the data has been selected and has been reduced to input variables, then the data is integrated into a single table that combines the input variables. The selected variables are listed in table 3. The script to perform filtering is shown in Figure 3. The results of the filtering process are shown in Figure 4.

Figure 3. Filtering

	Age	Total_Purchase	Account_Manager	Years	Num_Sites	Churn
0	42.0	11066.80	0	7.22	8.0	1
1	41.0	11916.22	0	6.50	11.0	1
2	38.0	12884.75	0	6.67	12.0	1
3	42.0	8010.76	0	6.71	10.0	1
4	37.0	9191.58	0	5.56	9.0	1

Figure 4. Filtering data results

III. RESEARCH METHOD

Descriptive analysis was conducted to describe the percentage of active and churn customers.

Figure 5. Descriptive Analysis

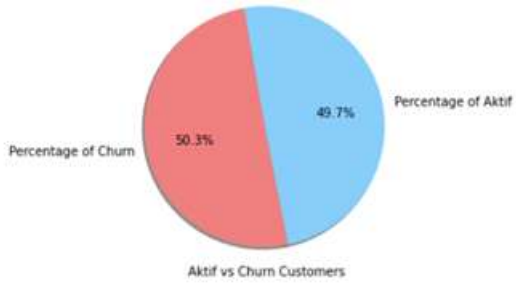

Figure 6. Distribution of Customer Status

Based on Figure 6, it is known that the active customer status data is 50.3% of the total customers, while the churn customers are 49.7% of the total customers. This data will then be the main focus in this study, namely measuring the accuracy of predicting customer status based on other

Masrina Manalu and Lily Wulandari, "Customer Churn Classification Using Support Vector Machine (SVM) Algorithm," *International Research Journal of Advanced Engineering and Science*, Volume 9, Issue 3, pp. 74-79, 2024.

International Research Journal of Advanced Engineering and Science
ISSN (Online): 2455-9024

variables, which in this data there are 6 variables that are thought to affect customer status.

IV. Results and Discussion

Data Classification

This study uses 300 dataset which are grouped into 20% testing data and 80% training data.

SVM Classification Modeling

The Support Vector Machine (SVM) method uses three SVM functions to find the best accuracy values, including linear SVM, polynomial kernel SVM and linear kernel RBF. The explanation of each kernel is:

1. Linear Kernel SVM

Linear kernel SVM is a good kernel algorithm to use when segmenting data. Analysis with linear kernel function, optimization of C or Cost is performed.

a. Linear Kernel Modeling SVM Data Training

SVM kernel linear modeling is carried out based on the results obtained from the best parameters. The best parameters are obtained from the previous hyperparameter tuning process. The parameters performed by tuning are linear kernel functions by choosing the best Cost value between 0.01, 0.1 and 1. Figure 7 shows the level of accuracy carried out on the training data.

```
In [9]: #pembentukan model Linear Kernel
        model_SVM = SVC(kernel='linear', C=1)
        model_SVM.fit(x_train, y_train)
        y_tr = model_SVM.predict(x_train)
        #akurasi data training
        acc = accuracy_score(y_train, y_tr)
        print("Accuracy: %.2f" % (acc*100), "%")

        Accuracy: 80.42 %
```

Figure 7. SVM Linear Kernel Parameters

The support vector machine (SVM) is trained using a training dataset, then its performance is evaluated into a testing dataset, when SVM analysis is performed using the Linear kernel function, the results show the model is able to classify customer churn with an accuracy rate of 80.42%.

b. SVM Data Testing Kernel Linear Modeling Validation

Model validation is done by using data testing. The goal is to see the performance of the model that has been formed using training data. Testing data is assumed to be new data to see class predictions using the previously obtained model. The following is the process carried out and the output obtained. Based on the output in figure 7, it is found that the model is able to predict with accuracy of 85%. The precisiom value class 0 (active customer) indicates that the model is able to predict the class of active customers with accuracy level of 96% and the recall value defines the proportion of actives customers who are correctly identified 74%.

While the precision value in class 1 (inactive customers) shows that the model predicts the class of inactive customers with accuracy in classification of 78% and the recall value explains the proportion of inactive customers who are correctly identified by 97%.

c. Confusion Matrix Support Vector Machine (SVM) Method

The confusion matrix explains in more detail the amount of data that is predicted to be correct and incorrect. Table 4 is the confusion matrix of the classification using the linear function of the SVM kernel.

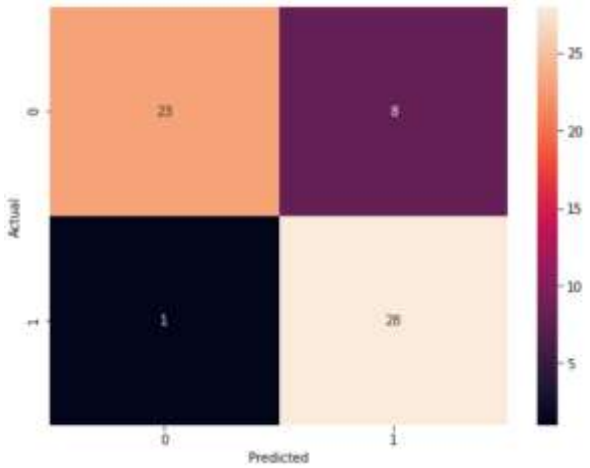

Figure 8. Confusion Matrix Linear Kernel

Figure 8 shows that from the total test data of 60 data entered in the SVM model using a linear kernel, there are 25 customers who are predicted to be active and 28 are predicted to be churn customers. The validation results show that there is 1 customer churn which is recognized as an active customer and 6 active customers are recognized as customer churn.

1. Polynomial Kernel SVM

A polynomial kernel is a non-linear kernel function that normalizes all training datasets.

a. Pemodelan Polynomial Kernel SVM Data Training.

Modeling using the polynomial kernel SVM function is carried out using training data with the best parameters, namely degree = 1 (d=1) and cost = 0.01 (c=0.01). The script for the Polynomial kernel SVM modeling process is shown in Figure 9.

Figure 9. SVM Ker Kernel Polynomial Parameters

The support vector machine (SVM) is trained using a training dataset, then its performance is evaluated in the testing dataset. When the SVM analysis is performed using the polynomial kernel function, the results show that the model is able to classify customer churn with an accuracy rate of 51.25%.

b. Validation of Kernel Polynomial Modeling SVM data testing.

Validation of the model with the kernel polynomial function is done using data tests. as with linear kernel functions, the goal is to see the performance of the model that has been formed using the training data. The script in Figure 10, to validate the Polynomial Kernel model

Masrina Manalu and Lily Wulandari, "Customer Churn Classification Using Support Vector Machine (SVM) Algorithm," *International Research Journal of Advanced Engineering and Science*, Volume 9, Issue 3, pp. 74-79, 2024.

International Research Journal of Advanced Engineering and Science
ISSN (Online): 2455-9024

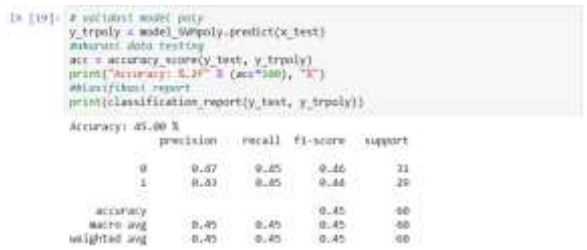

Figure 10. Validation of the SVM Linear Kernel Model

Based on these outputs, it was found that the model with the SVM kernel polynomial function was able to predict with an accuracy of 45%. The precision value in class 0 (active customers) explains that customers who are really active out of all customers who are predicted to be active are 47%, while the precision value in class 1 (customer churn) explains that customers who actually churn out of all customers who are predicted to churn by 43%. The recall value in class 0 (active customers) explains that customers who are predicted to be active compared to all actual active customers are 45% while the recall value in class 1 (customer churn) explains that customers who are predicted to churn compared to all students who actually churn are 45 %. The F1 score value is a comparison of precision and recall.

c. Confusion Matrix Polynomial Kernel SVM Method.

The confusion matrix explains in more detail the amount of data that is predicted to be correct and incorrect. Figure 11 is the confusion matrix of the classification using the polynomial function of the SVM kernel.

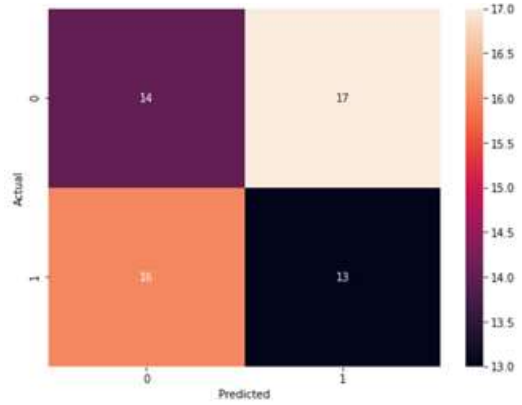

Figure 11. Confusion Matrix Polynomial Kernel

Figure 11 shows that from the total test data of 60 data entered in the SVM model using a polynomial kernel, there are 14 customers who are predicted to be active and 13 are predicted to be churn customers. The validation results show that there are 16 churn customers who are recognized as active customers and 17 active customers are recognized as churn customers.

3. Radial Basis Function (RBF) Kernel SVM

SVM kernel RBF which is a kernel function is used when the data is not linearly separated.

a) Modeling RBF kernel SVM Data Training

SVM kernel RBF modeling is done using training data. Figure 12 is the modeling process carried out.

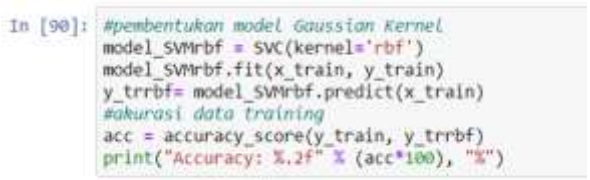

Figure 12. SVM Kernel RBF Modeling

SVM is trained using a training dataset, then its performance is evaluated into a testing dataset. When the SVM analysis is performed using the polynomial kernel function, the results show that the model is able to classify customer churn with an accuracy rate of 51.25%.

b. Validation of Polynomial Kernel SVM Data Testing Modeling.

Model validation is done using test data, the goal of which is to see how well the model will perform using the training data. The script and validation results with testing data are shown in Figure 13.

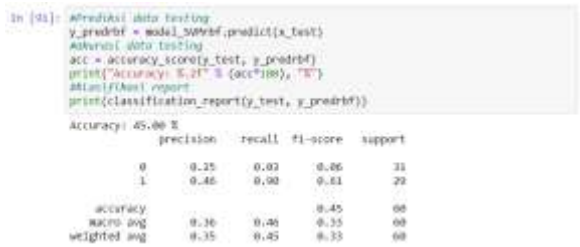

Figure 13. SVM Linear Kernel Model Validation

Model validation is done by predicting the testing data using the predict() command with model_SVMrbf which has been formed from the training data. Based on the output, it is found that the model is able to predict with an accuracy of 45%. The precision value in class 0 (active customers) explains that the true active customers of the total predicted active customers are 25%, while the precision value in class 1 (customer churn) explains that the customers who actually churn out of the total predicted churn customers are 46%. The recall value in class 0 (active customers) explains that customers who are predicted to be active compared to all actual active customers are 0.3% while the recall value in class 1 (customer churn) explains that customers who are predicted to churn compared to all students who actually churn are 90 %. The value of the f1 score is a comparison of precision and recall. If you look at the results of the accuracy obtained in the method with the RBF kernel function, the accuracy of the training data is 100%, but when you run on the test data, the accuracy is only 45%, which means that it is manufactured. the example above is correct.

c. Confusion Matrix RBF Kernel SVM

The confusion matrix explains in more detail the amount of data that is predicted to be correct and incorrect. Figure 14 is the confusion matrix of the classification using the SVM kernel RBF function.

Masrina Manalu and Lily Wulandari, "Customer Churn Classification Using Support Vector Machine (SVM) Algorithm," *International Research Journal of Advanced Engineering and Science*, Volume 9, Issue 3, pp. 74-79, 2024.

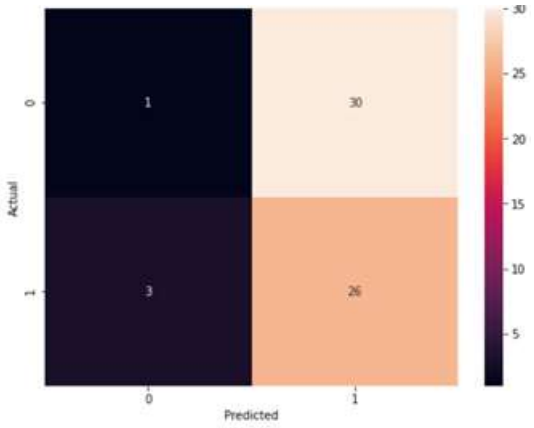

Figure 14 confusion matrix with Linear Kernel SVM

Figure 14 shows that from the total test data of 60 data entered in the SVM model using the RBF kernel, there is 1 customer that is predicted to be active and 26 are predicted to be customer churn. The validation results show that there are 3 classes of customer churn which are recognized as active customers and 30 active customers are recognized as churn customers.

Model Evaluation

Based on the method used, namely Support Vector Machines (SVM), which has been carried out, Table 4 is a comparison of the accuracy values of the classification model for each of these methods:

TABLE 4. Classification Performance Value

Method	Accuracy	Precision	Recall
Linear Kernel	80.42%	81.00%	80.00%
Polynomial Kernel	51.25%	51.00%	51.00%
RBF Kernel	51.25%	51.00%	55.00%

Table 4 shows that the best classification model for predicting customer status is the SVM (Kernel Linear) method with an accuracy of 80.42%, which means that the level of accuracy of the model for classifying testing data is 80.42% with the actual positive proportion that is correctly identified at 81%, meaning the level of accuracy in predicting customers The active and correct ones include active customers by 81% and the actual negative proportion that is correctly identified by 80% means that the level of accuracy in predicting customer churn and correct including customer churn is 80%. Therefore, it can be concluded that the SVM method with linear kernel functions has better performance than the SVM method with the other two kernel functions, namely polynomial and RBF.

V. CONCLUSION

The classification process to determine customer churn using the Support Vector Machine (SVM) method has been successfully carried out. The data used in this study amounted to 300 data consisting of 240 training data and 60 test data or testing data. The test was carried out by applying all the SVM kernel functions, namely linear SVM, polynomial kernel SVM and linear kernel RBF. Based on the accuracy value, it can be concluded that the Linear kernel function SVM method has the best performance, namely 80.42%.

REFERENCES

[1] Nurzahputra, A., Safitri, A. R., & Muslim, M. A. (2016). Klasifikasi Pelanggan pada Customer Churn Prediction Menggunakan Decision Tree. Prosiding Seminar Nasional Matematika X 2016, 717–722. https://journal.unnes.ac.id/sju/index.php/prisma/article/download/21528/10288/

[2] Permana, R. A., & Sahara, S. (2019). Metode Support Vector Machine Sebagai Penentu Kelulusan Mahasiswa pada Pembelajaran Elektronik. Jurnal Khatulistiwa Informatika, 7(1), 50–58. https://doi.org/10.31294/jki.v7i1.5743

[3] Suntoro, J. (2019). 22-DATA MINING Algoritma dan Implementasi Menggunakan Bahasa Pemrograman PHP. DATA MINING Algoritma Dan Implementasi Menggunakan Bahasa Pemrograman PHP, 9(9), 259–278.

[4] Zeniarja, J., Luthfiarta, A., Komputer, F. I., Dian, U., & Semarang, N. (2015). Prediksi Churn dan Segmentasi Pelanggan Menggunakan Backpropagation Neural Network. Techno.COM, 14(1), 49–54.

[5] IMADUDDIN, G. (2014). Evaluasi Dan Perbaikan Churn Model Dengan Mempertimbangkan Aspek Customer Value Dan Social Network: Studi Kasus Pt Xl Axiata Tbk.

[6] PHP. DATA MINING Algoritma Dan Implementasi Menggunakan Bahasa Pemrograman PHP, 9(9), 259–278.

[7] Zeniarja, J., Luthfiarta, A., Komputer, F. I., Dian, U., & Semarang, N. (2015). Prediksi Churn dan Segmentasi Pelanggan Menggunakan Backpropagation Neural Network. Techno.COM, 14(1), 49–54.