Predictive analytics of Churn Customers Calling Details Records using Classification by Clustering (CBC) dealing with Supervised Machine Learning Algorithms

Faroug A. Abdalla, Saife E Osman, Jamal S. Mohamed

Faculty of Computer Science and Information Technology, Al Neelain University, Khartoum, Sudan
farougali@hotmail.com

Received: 10/11/2022
Accepted: 06/12/2022

ABSTRACT- Telecom companies generate enormous amounts of data regularly. The telecom Decision makers that obtaining new customers is more challenging than sustaining existing ones. Furthermore, data from existing churn customers may be utilized to detect churn clients and their patterns of behavior. This research develops a model of churn prediction for the telecommunication business, which uses NB, SVM, DT, and RDF to detect churn clients. The proposed model churns customers' data using classification techniques, with the Random Forest (RDF) method performing well (95.94 % correctly categorized instances), the Decision Tree (DTs) providing classification accuracy (95.40 %), the Naïve Bayes (NB) provided classification accuracy (89.58 %), and the Support Vector Machine (SVMs) provided classification accuracy (71.08 %). The four different classification algorithms' predictions and observations are compared, with a percentage of 71 percent equality and 29 percent variation.

Keywords: *Machine Learning (ML), Naïve Bayes (NB), Support Vector Machine (SVM), Decision Tree (DT), Random Forest (RDF), Calling Details Records (CDR), and Classification by clustering (CBC).*

المستخلص - تنتج شركات الاتصالات كميات هائلة من البيانات بشكل منتظم. أكد صانعو القرار في مجال الاتصالات أن الحصول على عملاء جدد يمثل تحديًا أكبر من استمرار العملاء الحاليين. علاوة على ذلك، يمكن استخدام البيانات الواردة من العملاء الحاليين للكشف عن العملاء الاساسين (churn clients) وتحديد أنماط سلوكهم. يطور هذا البحث نموذجًا للتنبؤ بالعملاء الاساسين في ريادة عالم الاتصالات، والذي يستخدم الخوارزميات [RDF، DT، SVM،NB] لاكتشاف العملاء الاساسين. يقوم النموذج المقترح بإخراج بيانات العملاء باستخدام تقنيات التصنيف، مع أداء طريقة الغابة العشوائية Random Forest (RDF) بشكل جيد بنسبة (95.94٪ حالات مصنفة بشكل صحيح)، قدمت شجرة القرار(DTs) Decision Tree تصنيف بدقة (95.40٪)، وطريقة (NB) Naïve Bayes مقدمة تصنيف بدقة (89.58٪)، وآلة المتجهات الداعمة Support Vector Machine (SVMs)قدمت تصنيف بدقة (71.08٪). وتمت مقارنة التوقعات والملاحظات الخاصة بخوارزميات التصنيف الأربعة المختلفة، متساوية بنسبة (71%) واختلاف بنسبة (29%).

INTRODUCTION

In this study, we use an unsupervised learning technique called k-means clustering to label customers with their clusters and then use this labeled data as the training set for supervised models like Support Vector Machine (SVM), Nave Bayes (NB), Decision Tree (DT's), and Random Forest (RDF) to develop a prediction model to forecast a loyal customer using the best performance among the given approaches using calling Details records (CDR) datasets.

Such models assist telecommunications industries in analyzing customer patterns in terms of the extent to which they consume corporate services, as these industries must identify (expect) the customers who consume the most corporate services and focus on them to provide high-qualities products, motivate them, and maintain their loyalty, as customers. Loyalty is a corporate asset, and all telecommunications firms strive to maintain it in the face of fierce competition for consumers.

A developing technique called machine learning enables computers to learn autonomously from historical data. Numerous algorithms are used by machine learning to create mathematical models and make predictions based on previous knowledge or data. It is currently widely utilized in a variety of sectors.

The technological revolution has advanced to the level of machine learning, where technology and human knowledge and intelligence combine. Machine learning is at the heart of current transformations that which is going to be the Fourth Industrial Revolution. How the institutions and some industries recently use machine learning applications. Machine learning applications are going to make changes in many fields like the economy and financial system, data science, healthcare, and including how industries operate.

Machine learning techniques, which is, divided into Supervised, Unsupervised, and Reinforcement learning. In this paper, we use unsupervised learning techniques which are k-means clustering techniques which a k-means clustering to label the customer with its cluster and use this labeled data as a training set to train a supervised model such as (which are) Naïve Bayes (NB), Support Vector Machine (SVM), Decision Tree (DT's) and Random Forest (RDF) .to address the best performance among the above-mentioned techniques on calling Details records(CDR) datasets to build a prediction model to predict a loyal customer. Our goal in this study focused are to Compare the performance accuracy of four algorithms and choose the best one and evaluate the method of classification by clustering.

RELATED WORK

Customer loyalty in the telecom industry

The study aimed to investigate how customer satisfaction and service loyalty in the telecom industry were related to various service quality metrics; the current study identifies the areas on which the practitioner should concentrate while attempting to service customers [1]. The results of this study offer empirical proof of how customer experience affects customers' loyalties in telecom companies. The study identifies the 3rd elements of core service, pricing, and brands which go into creating a great customer experience and, consequently, customer loyalty [2]. The findings indicate that four factors: service quality, customer satisfaction, trust, and corporate image, have a major impact on consumer loyalty. Switching costs are a negligible factor in fostering customer loyalty, and studies suggest that service quality is the most important factor in Bangladesh's telecom sector [3]. In the cellular telephone industry in Nigeria, this study looks at the moderating impact of customer satisfaction on the correlations between service quality parameters and customer loyalty. [4]

The results of the analysis, which were based on 138 replies obtained from experienced mobile phone users in one of the major cities in southeastern Nigeria, show factor loadings ranging from 0.636 to 0.898. [5] to investigate how service fairness, customer trust, and customer engagement affect customer loyalty and to validate the aspects of customer engagement. The findings suggest that trust and customer involvement have a direct impact on customer loyalty, and the factor loading score ranges from 0.774 to 0.952 [6].

Comparative studies of machine learning techniques

A comparison study was made for the most common machine learning approaches when applied to the difficult task of predicting customer turnover in the telecoms sector, the study performs BPN, SVM, and DT classification methods for the churn prediction problem of telecom customers based on a publicly available dataset and they acquire. The best overall classifier was the SVM-POLY using AdaBoost with an accuracy of almost 97% and an F-measure of over 84% [7].

Comparative Analysis of Machine Learning Methods for Knee Abnormality Classification The accuracy for the extra tree classifier was 91.3% while it was equal to 70.1%, 70.0%, 79.3%, and 88.8% for the support vector machine, decision tree, k-nearest neighbor, and random forest classifiers, respectively. Five different machine learning classifiers, including the k-nearest neighbor, support vector machine, decision tree, random forest, and extra tree, were studied for the classification step [8-9].

Customer churn is a significant issue and one of the top issues for big businesses. Companies are attempting to create methods to predict possible customer churn because it has a direct impact on their revenues, particularly in the telecom field. The model's performance was improved from 84 to 93.3% in comparison to the AUC benchmark thanks to the application of SNA [10]. employing Support Vector Machines (SVMs) and Decision Trees (DTs) to classify records of customer call details, according to the accuracy computation results, the accuracy rates for DTs and SVMs were 0.9836 and 0.3304, respectively [11].

Customer churn prediction in telecommunications

One of the most difficult issues for large businesses to solve is customer churn. Organizations are striving to develop methods to estimate likely customer turnover because it has such a direct impact on a company's revenue, especially in the telecom industry. To take the appropriate action to reduce churn, it is crucial to identify the factors that influence customer churn. The main contribution of our work is the creation of a churn prediction model that aids telecom providers in identifying the most probable consumers to leave. To establish a novel method for feature engineering and selection, the model developed in this research uses machine learning techniques on a large data tool. used to evaluate the model's effectiveness. The Area of Under Curve (AUC) standard is calculated by assessing the model's performance, and the value of AUC achieved is 93.3 % [12].

Business is becoming extremely saturated in today's cutthroat global marketplace. Due to the abundance of active, competitive service providers, the telecommunications industry is particularly faced with difficult issues. As a result, keeping their current clientele has grown to be exceedingly challenging for them. The telecom sectors need to take the required actions to retain customers to stabilize their market value because the cost of recruiting new customers is substantially higher than the cost of maintaining existing customers [13].

Customer Churn Prediction (CCP) is a difficult task for decision-makers and the machine-learning community because churn and non-churn clients frequently have similar characteristics. It is clear from various studies on customer churn and related data that a classifier exhibits varying levels of accuracy for various zones of a dataset. The phenomenal expansion of digital data and related technologies has spawned a trend where sectors are fast going digital. Particularly in TCI, these technologies are offering excellent chances to pinpoint and address the pervasive problem of client attrition. We have determined the classifier decision's perceptive level of certainty based on the distance component using a novel CCP approach. We have also classified customers into several customer groups based on lower zone and distance [14].

RESEARCH METHODOLOGY

Data description

The open-source Kaggle website provided the data used in this study, which included mobile phone activity datasets made up of one week's worth of Call Details Records from Milan and the Province of Trentino (Italy) [15]. The CDRs data consists of 10,000 records about Call -in, Call-outs, SMS-in, SMS-out, and internet usage. The dataset contains five main features; including SMS_-n, SMS-out activity, Call-in, Call-out activity, and usages of internet activity, and the Describe of variables are shown in Table 1

TABLE 1: DATASET DESCRIPTION

Variable Names	Description
Call in	A customer receives a call
Call out	A customer makes a call
SMS in	A customer receives a message
SMS out	A customer sends a message
Internet usages activity	A customer starts to connect to the internet

DATA PREPROCESSING

The following procedures are used to preprocess data:

a) Zeros (0) are used to fill in missing data

b) To create new datasets, the original datasets are grouped by cell Id

c) The grouping was carried out by computing the total minute number of SMS-in, SMS-out, call-in, call-out, and internet for each cell Id

d) Since the cost of calls, SMSs and internet are not equal it was necessary to weigh them subject to their contribution to the revenue of the company, we added a new field of the contribution for each cell Id, such that the calls-out have the biggest weight followed by SMSs-out, internet activities, calls-in, and then SMS-in.

Algorithmic flowcharts

Figure 1, shows the many steps of our study. The dataset for our predictive model was first preprocessed, then separated into 2 parts: training and testing data. In this work, we used 4 classification techniques ((NB, SVM, DT, and RDF) as well as K-means clustering. As a result, the best predictive model has been accepted as a technique for making future predictions.

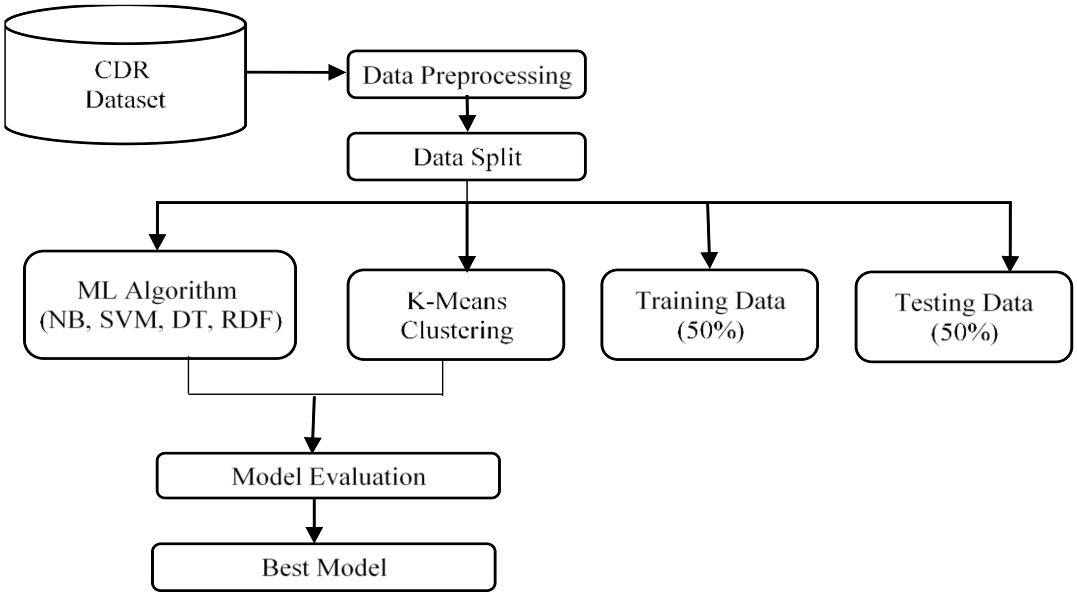


Figure 1. Flowchart for Proposed Model

ML Algorithms

With the use of data and algorithms, machine learning (ML), a subfield of artificial intelligence (AI) and computer science, strives to simulate how individuals learn. Numerous algorithms, such as supervised machine learning algorithms, unsupervised machine learning algorithms (such as support vector machines, neural networks and decision trees, and Random Forest), semi-supervised learning algorithms, clustering algorithms, regression algorithms, Bayesian algorithms, and many others are used in machine learning applications. A sophisticated data mining method called machine learning is used to extract characteristics from massive amounts of data [16].

K-MEANS is an unsupervised machine learning technique like clustering that is used to solve clustering problems in machine learning. It is a well-known aggregation (clustering) strategy that ends up at a solution no matter where it begins. Not all beginning points provide the same result from it. To identify the optimal response for the selected criterion in this situation, the calculations are frequently repeated.

To provide a starting point for the first iteration, the centers of the k classes and the k objects are correlated (either taken at random or not). The objects are then assigned to the centers that are closest to them based on the distance between the items and the k centers. The objects that have been assigned to the different classes are then used to redefine the centers. Then, based on their distances, the objects are transferred to new centers. The process continues until convergence is achieved [17-18]. Unsupervised learning algorithm K-Means Clustering divides the unlabeled dataset into various clusters. In this case, K specifies how many pre-defined clusters must be produced during the procedure; for example, if C=5, it will result in five clusters, and so on and we grouped the CDR data into five clusters with each cluster including the k-means. Clustering is Determinant (W) is a classification criterion for determining the best solution.

Algorithm 1: Mathematical background of K-means clusters:

```
Input:
D={t1, t2, …..Tn}    // Set of elements
K                    // The required number of clusters
Output:
K                    // Set of elements
Algorithm of K-Means:
determine the initial values of a1, a2,... ak
repeat
The clusters with the closest means should receive each item ti;
compute a fresh mean for every cluster;
until convergence standards are satisfied.
```

NAIVE BAYES [NB] A popular supervised machine learning method that relies on predictor independence is the Naive Bayes approach. A probabilistic machine-learning model called the Naive Bayes classifier is used for classification tasks. The classifier's core is the Bayes theorem [19-20].

$$\text{Bayes Theorem: } P(A\backslash B) = \frac{\boldsymbol{P(B\backslash A)P(A)}}{\boldsymbol{P(B)}} \qquad (1)$$

SUPPORT VECTOR MACHINE [SVM] - A supervised approach to machine learning that excels even in non-linear circumstances is the support vector machine uses this technique to do a regression, a multi-class classification, or a binary classification on a set of observations that are described by qualitative and quantitative variables (predictors). The primary goal of the SVM method is to establish the optimum binary classifier or line that can divide n-dimensional space into subclasses such that we can quickly classify new data points in the future. This best decision boundary is called a hyperplane.

The Support Vector Machine (SVM) is a supervised machine learning technique that was invented in the context of the It took until the mid-1990s for an algorithm implementation of the SVM to be given, due to the discovery of the kernel technique and the generalization to the non-separable situation. [21- 22]. Since then, the SVM has known numerous developments and gained popularity in various areas such as Machine Learning, optimization, neural networks, and functional analysis.

Categorization in binary classifications of SVM with the idea that the bigger the separation, the more accurate the classification, the SVM is made to identify a difference between two classes of things. In its simplest form, the linear and separable case, the approach will select a hyperplane that splits the observation set into two different classes in a method that optimizes the distance between the hyperplane and the training set's nearest observation.

Multiple-class categorization SVM can only address binary issues; other approaches for solving multi-class issues have been created. All of them employ the same technique to break down a multi-class problem into numerous binary problems, except One vs One, which generates one binary model per pair of classes. Regression in the SVM method was generalized such that it could be used to solve regression problems or predict time sequences. Let the training set be {x{p}, y{p}} for p = 1,... N, where x is the set of observation predictors and y{p} in R. [23-24].

DECISION TREE [DT] is a type of non-parametric SLM algorithm that is used for regression and classification tasks. DT aims to create a model that predicts how many different variables will affect the outcomes of the target instances. Here are the children's edges for each of the internal nodes, which each correspond to one of the input variables' values. Each branch executes a value of the target variable provided by the input possible values seen along the branch's path from the root [25-27].

CLASSIFICATION AND REGRESSION RANDOM FORESTS [RDF]: This sophisticated machine learning method enables you to create predictions based on a variety of decision trees. Random forests Predictive models for categorization and regression are provided. The approach employs binary decision trees, specifically. In classification (qualitative response variable): The model allows predicting depending on explanatory quantitative and/or qualitative characteristics, the classification of observations.

In regression (continuous response variable): The model allows building a predictive model for a quantitative response variable based on explanatory quantitative and/or qualitative variables [28-32].

RESEARCH FRAMEWORK

Figure 2, shows that the datasets were divided into five classes, each of which assigned a weight to each customer based on SMS-in, SMS-out, call-in, call-out, and internet usage. We used sampling without replacement to divide the datasets into two parts, each representing 50% of the datasets (5000 records).

Then, we use K-Means clustering to classify the consumers into five groups. The K-Mean clustering method used in this study was used to divide consumers into five clusters (C1, C2, C3, C4, and C5). At the end of the process, we obtain 5 groups, each one representing the type of customers. This sample of data will be used as a training set for the classification.

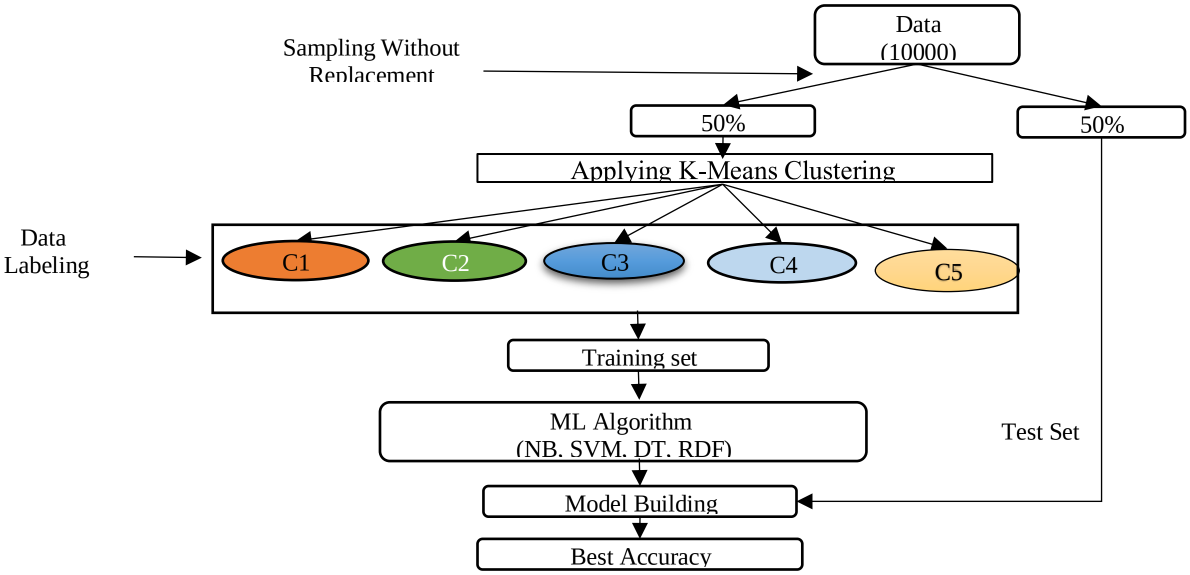


Figure 2. The research framework of Classification by Clustering (CBC)

Then, we implement the 4 classification algorithms (SVM, DT, NB, and RDF) to test the classification model. Those algorithms use the training data provided by the clustering model, whereas the test model's data is utilized to test the classification model. Then, among the four techniques used in this framework, the Classification by Clustering (CBC) model is used to obtain the best accuracy.

RESULTS AND DISCUSSIONS

The first step is involved preprocessing the dataset and utilizing the (random sampling with replacement) approach to split it into two parts, the main goal of the aforementioned procedure is to get two datasets, The K-means technique is used to illustrate clustering in the first dataset, this step's objective was to divide the instances into homogenous clusters, so each cluster now represented a certain class of customers, each customer will have their category selected after this procedure, which produced the five clusters (C1, C2, C3, C4, and C5). The second step was produced as a result of the first step where the dataset was suitable for classification, and each object in the dataset was assigned to a category, the datasets used in 4 models were trained to utilize algorithms (SVM, NB, DT, and RDF).

According to the accuracy computation findings, the accuracy rates for RDF, DT, NB, and SVM were 0.9594, 0.9540, 0.8958, and 0.7108, respectively. This shows that the RDF classifier provides more accurate predictions than the SVM classifier. The accuracy ratings for the proposed methods used to get this result to demonstrate that SVM has the lowest accuracy rate. Overall, the findings showed that the RDF approach of categorization performed better than the other methods in the current study. This conclusion can be supported by the fact that SVM requires more training time than RDF, DT, and NB do, making it less effective for large datasets. Additionally, the SVM approach is sensitive to the kind of kernel being utilized and performs badly with overlapping classes. However, when it comes to the implementation of the algorithms utilized in the study, the RDF and DT algorithms are faster than the NB and SVM algorithms.

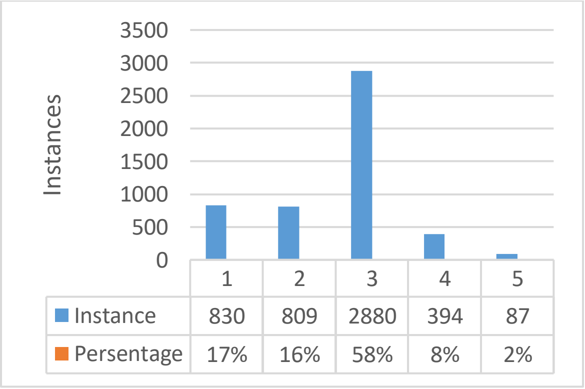
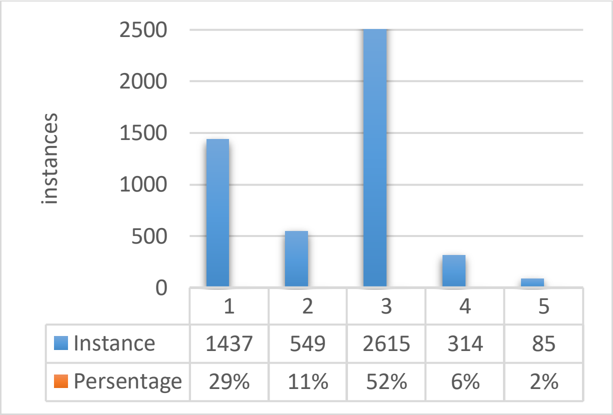
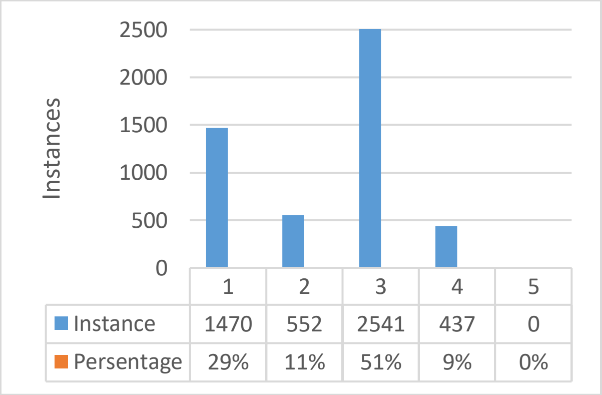


Figure 3. Distribution of the instances among 5 clusters in SVM.



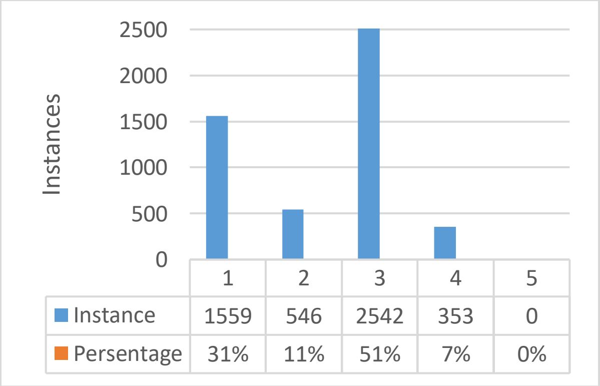
	1	2	3	4	5
🟦 Instance	1437	549	2615	314	85
🟧 Persentage	29%	11%	52%	6%	2%

Figure 4. Distribution of the instances among 5 clusters in NB.



	1	2	3	4	5
🟦 Instance	1470	552	2541	437	0
🟧 Persentage	29%	11%	51%	9%	0%

Figure 5. Distribution of the instances among 5 clusters in DT.



	1	2	3	4	5
🟦 Instance	1559	546	2542	353	0
🟧 Persentage	31%	11%	51%	7%	0%

Figure 6. Distribution of the instances among 5 clusters in RDF.

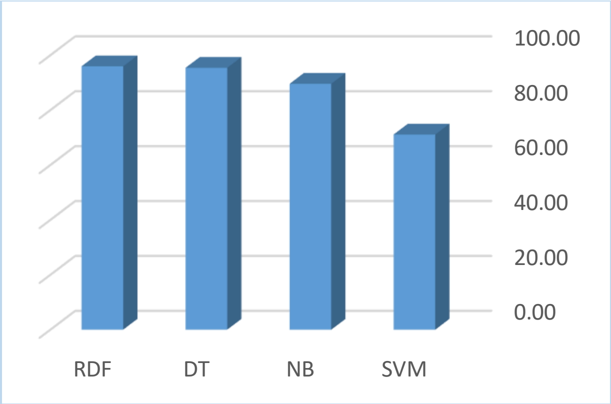


Figure 7. Distribution of the Classification Accuracy among 4 Classifications Algorithms.

Figure 3,4,5, and 6 show the number of instances belong to specific class for each classifier.

Figure 7, shows that Random Forest (RDF) provided the highest classification accuracy 95.940 % and the Support Vector Machine (SVM) provided the lowest classification accuracy 71.08 %**.** For practical experiments, XLSTAT tools were used to interact with the dataset and analyze the data [33].

The results obtained from the various experiments that were conducted are better than the results obtained in previous studies. The results showed that the SVM classifier gave less accuracy compared to the rest of the classifiers.

Table (II) shows these processes where the 1st prediction object (PredObs [1]) is represented in the cluster (3) for SVM, NB, DT, and RDF algorithms used in this model, and the 13th prediction object (PredObs [13]) represented in the cluster (2) for SVM, NB, and DT algorithms, as well as 13th prediction object, shows in the cluster (1) for the RDF algorithm. We found that the comparison processes of predictions observations of 4 classification algorithms have a percentage of 71 % equality and 29 % variations in Table II, below. The four algorithms are compared. As a sample, make a prediction based on the (5000 records) dataset.

TABLE II: A SAMPLING OF 4COMPARES ALGORITHMS PREDICTION OBSERVATION BASED ON THE DATASET OUTCOMES.

PredObs	SVM	NB	DT	RDF	equality	variance
1	3	3	3	3	√	
2	1	1	1	1	√	
3	3	3	3	3	√	
4	1	1	1	1	√	
5	2	2	2	2	√	
6	3	3	3	3	√	
7	4	4	4	4	√	
8	3	3	3	3	√	
9	1	1	1	1	√	
10	3	3	3	3	√	
11	1	1	1	1	√	
12	3	3	3	3	√	
13	2	2	2	1		×
14	2	2	2	2	√	
15	4	4	4	4	√	
16	3	3	3	3	√	
17	3	3	3	3	√	
18	3	3	3	3	√	
19	3	3	3	3	√	
20	1	1	1	2		×
---	**---**	**---**	**---**	**---**		
5000	3	3	3	3	√	

CONCLUSION

In this study, the CDRs Dataset has been used to implement a comparative analysis of the SVM, NB, DT, and RDF algorithms to analyze the consumption of customer services. The outcome shows that the RDF classifier performs better than other classifiers while taking less time. Future studies will focus on the choice of various datasets, various behavioral patterns, and reality mining. RDF and DT Classifiers offer good accuracy and perform faster prediction compared to NB and SVM algorithms. This finding can be explained by the fact that, because of its lengthy training process, SVM is not better for big datasets. In addition, SVM consumes more time in training compared to RDF and DTs. It is sensitive to the kind of kernel and performs terribly with overlapping classes. The four different classification algorithms' predictions and observations are compared, with a percentage of 71 percent equality and 29 percent variation.

REFERENCES

[1] Chen, W., Xie, X., Wang, J., Pradhan, B., Hong, H., Bui, D. T., . . . Ma, J. (2017). A comparative study of the logistic model tree, random forest, and classification and regression tree models for spatial prediction of landslide susceptibility. *Catena, 151*, 147-160.

[2] Kumar, A. (2017). Effect of service quality on customer loyalty and the mediating role of customer satisfaction: an empirical investigation for the telecom service industry. Journal of Management Research and Analysis, 4(4), 159-16

[3] Imbug, N., Ambad, S. N. A., & Bujang, I. (2018). The influence of customer experience on customer loyalty in the telecommunication industry. International Journal of Academic Research in Business and Social Sciences, 8(3), 103-116.

[4] Hafez, M., & Akther, N. (2017). Determinants of customer loyalty in the mobile telecommunication industry in Bangladesh. Global Journal of Management and Business Research.

[5] Palladan, A. A., & Ahmad, M. A. (2019). Leveraging customers loyalty in telecommunication industry: The role of service quality and customer satisfaction a PLS approach. International Journal of Marketing Research Innovation, 3(1), 1-10.

[6] Izogo, E.E. (2017), "Customer loyalty in the telecom service sector: the role of service quality and customer commitment", The TQM Journal, Vol. 29 No. 1, pp. 19-36.

[7] Hapsari, R., Hussein, A. S., & Handrito, R. P. (2020). Being Fair to Customers: A Strategy in Enhancing Customer Engagement and Loyalty in the Indonesia Mobile Telecommunication Industry. Services Marketing Quarterly, 41(1), 49-67. doi:10.1080/15332969.2019.1707375

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

[9] A. Vijayvargiya, R. Kumar, N. Dey, and J. M. R. S. Tavares, "Comparative Analysis of Machine Learning Techniques for the Classification of Knee Abnormality," 2020 IEEE 5th International Conference on Computing Communication and Automation (ICCCA), 2020, pp. 1-6, DOI: 10.1109/ICCCA49541.2020.9250799.

[10] Ahmad, A. K., Jafar, A., & Aljoumaa, K. (2019). Customer churn prediction in telecom using machine learning in the big data platform. Journal of Big Data, 6(1), 1-24.

[11] Abdalla, F. A., & Ali, S. E. O. (2021). Classification of customer call details records using Support Vector Machines (SVMs) and Decision Trees (DTs). Journal of Engineering and Computer Science (JECS), 22(3), 1-8.

[12] Ullah, I., Raza, B., Malik, A. K., Imran, M., Islam, S. U., & Kim, S. W. (2019). A churn prediction model using random forest: analysis of machine learning techniques for churn prediction and factor identification in the telecom sector. IEEE Access, 7, 60134-60149.

[13] Umayaparvathi, V., & Iyakutti, K. (2016). A survey on customer churn prediction in telecom industry: Datasets, methods, and metrics. International Research Journal of Engineering and Technology (IRJET), 3(04).

[14] Adnan Amin, Feras Al-Obeidat, Babar Shah, Awais Adnan, Jonathan Loo, Sajid Anwar, Customer churn prediction in telecommunication industry using data certainty, Journal of Business Research, Volume 94,2019, Pages 290-301, ISSN 0148-2963,

[15] Kaggle Datasets, Available at:https://www.kaggle.com/datasets,mobile-phone-activity

[16] Alican Dogan, Derya Birant, Machine learning and data mining in manufacturing, Expert Systems with Applications, Volume 166,2021,114060, ISSN 0957-4174.

[17] Sinaga, K. P., & Yang, M.-S. (2020). Unsupervised K-means clustering algorithm. IEEE Access, 8, 80716-80727.

[18] Havlíček, V., Córcoles, A. D., Temme, K., Harrow, A. W., Kandala, A., Chow, J. M., & Gambetta, J. M. (2019). Supervised learning with quantum-enhanced feature spaces. Nature, 567(7747), 209.

[19] Yang, F.-J. (2018). An implementation of naive Bayes classifier. Paper presented at the 2018 International conference on computational science and computational intelligence (CSCI).

[20] Jadhav, S. D., & Channe, H. (2016). Comparative study of K-NN, naive Bayes, and decision tree classification techniques. International Journal of Science and Research (IJSR), 5(1), 1842-1845.

[21] Fradkov, A. L. (2020). Early history of machine learning. *IFAC-PapersOnLine, 53*(2), 1385-1390.

[22] Cortes, C., Mohri, M., & Storcheus, D. (2019). Regularized gradient boosting. *Advances in Neural Information Processing Systems, 32*.

[23] Blanco, V., Japón, A., & Puerto, J. (2019). Optimal arrangements of hyperplanes for SVM-based multiclass classification.

[24] Advances in Data Analysis and Classification, 1-25.

[25] Pisner, D. A., & Schnyer, D. M. (2020). Support vector machine Machine learning (pp. 101-121): Elsevier.

[26] Phan, H. T., Tran, V. C., Nguyen, N. T., & Hwang, D. (2019). Decision-Making Support Method Based on Sentiment Analysis of Objects and Binary Decision Tree Mining. Paper presented at the International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems.

[27] Charbuty, B., & Abdulazeez, A. (2021). Classification based on decision tree algorithm for machine learning. Journal of Applied Science and Technology Trends, 2(01), 20-28.

[28] Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (2017). *Classification and regression trees*: Routledge.

[29] Patel, H. H., & Prajapati, P. (2018). Study and analysis of decision tree-based classification algorithms. International Journal of Computer Sciences and Engineering, 6(10), 74-78.

[30] Muppavarapu, V., Rajendran, A., & Vasudevan, S. K. (2018). Phishing detection using RDF and random forests. Int. Arab J. Inf. Technol., 15(5), 817-824.

[31] Soliman, H. (2020). Random forest-based searching approach for RDF. IEEE Access, 8, 50367-50376.

[32] Jaiswal, J. K., & Samikannu, R. (2017). Application of random forest algorithm on feature subset selection and classification and regression. Paper presented at the 2017 world congress on computing and communication technologies (WCCCT).

[33] Addinsoft (2020) XLSTAT Statistical and Data Analysis Solution. New York. Available at:https://www.xlstat.com