Classification of customer call details records using Support Vector Machine (SVMs) and Decision Tree (DTs)

Faroug A. Abdalla, Saife E Osman Ali

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

Received:09/09/2021
Accepted:25/11/2021

ABSTRACT- On a daily basis, telecom businesses create a massive amount of data. Decision-makers underlined that acquiring new customers is more difficult than maintaining current ones. Further, existing churn customers' data may be used to identify churn consumers as well as their behavior patterns. This study provides a churn prediction model for the telecom industry that employs SVMs and DTs to detect churn customers. The suggested model uses classification techniques to churn customers' data, with the Support Vector Machine (SVMs) method performing well 98.36 % properly categorized instances) and the Decision Tree (DTs) approach performing poorly 33.04 % and the decision tree algorithm deliver outstanding results.

Keywords*: Support vector machines (SVMs), Decision trees (DTs), Data mining; Call detail records (CDRs), Supervised Machine Learning (SLM), Total Contribution (T.C).*

المستخلص – في الحياة اليومية ، تنشئ شركات الاتصالات كمية هائلة من البيانات. شدد صانعو القرار على أن الحصول على عملاء جدد هو أكثر تكلفة من الحفاظ على العملاء الحاليين وأن بيانات العملاء الحاليين يمكن استخدامها للتعرف على العملاء الاساسين (churn customers) وكذلك أنماط سلوكهم. توفر هذه الدراسة نموذجًا للتنبؤ بالعملاء الاساسين في مجال الاتصالات الذي يوظف طريقة دعم آلة المتجهات (SVMs) ونهج نموذج شجرة القرار (DTs) للكشف عن العملاء الاساسين .يستخدم نموذج تقنيات التصنيف المقترح لمعرفة بيانات العملاء الاساسين(churn customers)، مع أداء طريقة دعم آلة المتجهات(SVMs) بشكل جيد بنسبة 98.36٪ من الحالات المصنفة بشكل صحيح ونهج شجرة القرار (DTs) الذي يحقق أداءً ضعيفًا بنسبة 33.04٪ وتقدم خوارزمية شجرة القرار نتائج رائعة.

INTRODUCTION

Data mining (DM) is the process of discovering knowledge in a database. Data mining is discovering meaningful, innovative, persistent, useful, and ultimately understood patterns in data is a difficult task. Data mining is one of the most enticing fields of study that is becoming more and more widespread in the telecom industry [1].

Artificial intelligence (AI) is a technique that lets computers and robots to replicate the human mind's ability to solve problems and make decisions for the telecom industry in the loyalty of the customer calling detail records [2-3].

Machine learning (ML) is a branch of artificial intelligence (AI) and computer science that focuses on simulating the way people learn and steadily improving its accuracy using data and algorithms. Many algorithms are used in Machine Learning applications, including supervised Machine Learning algorithms, unsupervised Machine Learning algorithms (support vector machine, neural network and decision tree, Random Forest), semi-supervised learning algorithms, clustering algorithms, regression algorithms, Bayesian algorithms, and many others. Machine Learning is an advanced data mining technique used to extract features from massive amounts of data [4].

Call Detail Record (CDRs) indicates a record that consists of detailed information about a telecom transaction, such as origin of the call, destination of call, duration of call, start time and end time. Furthermore, CDRs can provide each event details that can happen in the network. This studied take attention to find customer behavior from CDRs. Information based upon CDRs are captured by the telecommunication industries during Call in, Call out, SMS in , SMS out and internet activities of a client. This kind of information shows bigger ideas

about the customer's wants. Most of the Telecommunication's institutions uses CDRs info for fraud detection by clustering and/or classifying the user records [5-11].

Data mining is basic research that identifies the interesting relationships between groups of elements in datasets and predicts the interrelated behaviors of the new data. Data mining techniques are often able to classify a dataset based on CDRs for most accurately than traditional statistical works. This study discusses some data mining techniques about classifying customer action activities of user's profile in telecom for some day by using SVMs and DTs algorithms. These techniques were used to find part of customers with regard to their usage by few hours [12-20].

This manuscript is organized as follows: Section 2 presents algorithms methods to be used in this study, where in the article the two data mining techniques will be outlined are support vector machines and decision trees. In Section 3, we give research methodology, including a brief description of the dataset that used in the analysis as long as the application schemes. Section 4 display the results of the analysis. Finally, in section 5 the conclusion and future works are presents.

Research objectives

Our goals of this studies:

- In a huge data environment in telecom, customers' value was assessed by segmenting them using DTs and SVMs, and then determining the amount of loyalty for each segment.
- The telecom data was used to create a set of features.
- The best attributes for consumers with demographic data where the following classification algorithms were used and the classification models were developed based on these attributes and the amount of loyalty for each segment: Support vector classifiers and decision tree classifiers are two types of classifiers.
- Our models were assessed using a set of criteria to determine which model was the most accurate.
- The loyalty rules were generated from this model; these rules revealed the features of each degree of loyalty, allowing the reasons for loyalty to be identified in each segment and targeted in a representative manner. The ability to develop an approach to predict new customers by loyalty was another advantage of using classification algorithms.

RELATED WORK

Various approaches, such as machine learning and data mining, have been used in the literature to forecast churn and the most widely acknowledged tools for predicting difficulties connected with client turnover are decision trees

the supervised model uses support vector machine (SVM) classification stages to separate churn and non-churn customers into two groups [21].

The study compares SVM against BPANN, decision tree C4.5, logistic regression, and naïve Bayesian classifiers in order to predict customer attrition in the telecommunications industry. From the approach, it concludes that SVM's characteristics include a simple classification plane, strong generating ability, and excellent fitting precision, among others. When there are a lot of samples (abundant support vectors), a lot of attributes, a lot of churns, a lot of missing records, and nonlinearity data, SVM has an excellent prediction precision [22].

They used a case study to demonstrate the best use of data science in churn prediction and management. More precisely, the article discusses how to use data mining techniques instead of traditional churn management strategies to minimize churn, boost profitability, and, as a result, increase customer happiness. Although we have presented a method for optimizing churn prediction, with a focus on comparing classification methods on the collected dataset and influencing factors, additional mining results could be obtained by performing a cause-effect analysis of identifying the most influencing attributes from the customer details elaborated, and finally clustering and pattern finding of the customer churn logged behavior.

In the not-too-distant future, technology will unleash a revolution in management sciences based on practical algorithms refined by artificial intelligence, data mining, and machine learning approaches. The classification algorithms employed were decision trees, SVN, and NN [23].

DTs have been effectively utilized to forecast customer turnover in the telecommunications industry in a number of studies. obtained a high level of accuracy on their model for forecasting customer churn in the telecommunication services industry.) discovered that DTs are appropriate for

investigating customer churn in the telecoms industry. The use of DTs to examine customer turnover in the insurance industry has been investigated, though the model did not produce the most accurate results in her study [24].

For enhancing the accuracy of customer churn prediction, comparison research was conducted using three famous classifiers: K-NN, Random Forest, and XG boost. On the publicly accessible telecom dataset, the XG boost classifier scores well out of three. Using the XG boost classifier, the effort was then focused on identifying the characteristic with the highest cognition for churn.

The results of the trial reveal that Fiber Optic customers with higher monthly prices have a stronger impact on attrition. Expected directions can be predicted using a mix of classifiers that provides high accuracy and desired outcomes [25].

Supervised Machine learning(SML)

SML is to make the computer to do something; supervised learning is givens the right answers to the datasets.

Support Vector Machine(SVMs)

Support Vector Machine is very technique of all Machine Learning area and it is an example of "Kernel based technique" used for classifying was introduced by Boser, Guyon, Vapnik in 1992, it implements classification by making a Multidimensional hyperplane that as the most used divides the data into two divisions.

Support Vector Machines (SVMs) is a technique that depend on statistical learning methods and it is used to reduce the structural risk of machine learning [26-28]. The major function of SVMs is to find the highly edge hyperplane and a set of linearly separable data, classify data correctly, in order to increase the minimum distance between data and the hyperplane. Many studies of SVMs seek to propose simple and efficient methods to find the problem of maximal margin hyperplane [29-30].

The support vector machine classification approach was created by Vladimir Vapnik and Alexey criteria for classification and regression trees are different. Chervonenkis, and it is an effort to classify a dataset into two groups using a linearly separable hyperplane. Finally, the model will very certainly be able to predict the target groups (labels) for fresh examples. Assume we can properly segregate the data. Then we'll be able to improve the following: Minimize $\|W\|^2$, subject to

$$w.x_i + b \geq 1\,,\ \text{if}\ y_i = 1 \quad (1)$$

$$w.x_i + b \geq -1\,,\ \text{if}\ y_i = -1 \quad (2)$$

The last two constraints can be commutated to:

$$y_i(w.x_i + b) \geq 1 \quad (3)$$

Decision Trees Algorithm (DTs)

Decision tree are a non-parametric SLM Algorithms, which is used for classification and regression purposed. DTs aims to build a model that expects the outcome of the target instances with respect to many variables' elements. Each internal node corresponds to one of the input variables; here are children's edges for each of the respect to values of the variable input. Each branch performs a value of target variable offered by the input variables values appeared by the route from the root to the branch.

DTs are trees that classify cases instances by categorizing them based on parameter values. A decision tree is a simple structure where non-terminal nodes represent tests on one or more variables and the last nodes consider the decision results [31-33].

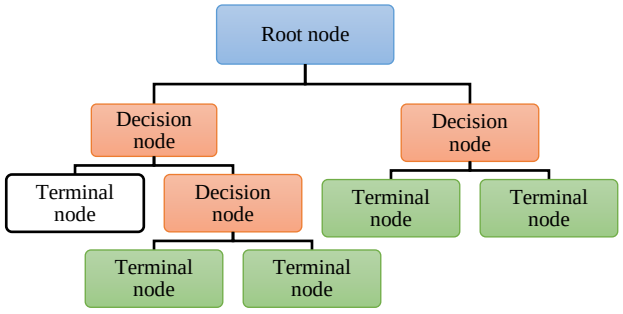

Figure 1. Working of Decision tree

The examples are classified using decision trees by sorting them along the tree from the root to a leaf/terminal node, with the leaf/terminal node providing the classification.

Each node in the tree represents a test case for some property, with each edge descending from the node corresponding to the test case's potential solutions. This is a cyclical procedure that occurs for each subtree rooted at the new node.

The Decision Trees and their ways of work to make strategic splits has a significant impact on a tree's accuracy.

RESEARCH METHODOLOGY

Data description

The datasets used in this study acquired from Kaggle Website, which is an open source, where the mobile phone activity datasets composed by one week of Call Details Records from the Milan city and the Province of Trentino (Italy) [34]. The CDRs file consists data for 10, 000 records about Call in, Call out, SMS in, SMS out, and internet usage. The dataset contains five main features; including about SMS in, SMS out activity, call in, call out activity and internet usages 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 internet

CDRs come in a variety of shapes and sizes, and Telecom Italia has documented the following activities: SMS was received. When a user gets an SMS, a CDR is produced. SMS has been sent. Each time a user sends an SMS, a CDR is created. *incoming phone call Each time a user receives a* call, a CDR is produced. Leaving a Call Every time, a user makes an Internet call, a CDR is created. Each time a user connects to the Internet or disconnects from it, a CDR is created.

Data preprocessing

a) In order to preprocess data, the following steps are used:
b) Missing data are filled-in by Zeros(0)
c) For construction of new datasets, the original datasets are grouped with a cell Id
d) The grouping was carried out by computing the total minutes number of sms-in, sms-out , call-in, call-out and internet for each cell Id
e) Since the cost of calls, smss and internets are not equal it was necessary to weight them subject to their contribution of the revenue of 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 finally smss-in.
f) The eq. 4 is used to find the customers contributions for the telecom companies:

The total of data has converted into nominal data type (five class labels) according to the following syntax in Table II:

TABLE II: CLASSES CLASSIFICATION

Class	Class range	Classifications /Categories
C1	0 to 2000	Very low
C2	2000 to 4000	Low
C3	4000 to 6000	Average
C4	6000 to 8000	High
C5	8000 to 10000	Very High

To interact with the dataset and analyze the data, we use a free Python environment (version 3.6.5) [35] with the Pandas and scikit-learn tools [36].

Mathematical Model

$$\text{Total Contribution} = \big((1 * \text{smsin}) + (2 * \text{callin}) + (3 * \text{internet}) + (4 * \text{smsout}) + (5 * \text{callout})\big)/15.0 \quad (4)$$

Mathematical modeling is defined as the process of applying mathematics to a real-world situation in order to get a better understanding of the problem. As such, mathematical modelling is obviously related to problem solving.

Algorithmic flowcharts

Figure 3 describes the different steps involved in our study. In the first step, we preprocessed the dataset for our predictive model, then divided it into two sections, training and testing datasets. In our study, the two classification techniques and a Total Contribution (T.C) equation have been applied. As a result, the best predictive model is approved as a method for future prediction.

Research Framework

In research framework the datasets were split into two sections: training sets and test sets. The training sets contain 75% of the datasets (7500 records) while the test sets contain the remaining data. In the next used a T.C equation to distribute the consumers into five classes (C1, C2, C3, C4, and C5). At the end of this procedure, we had five groups, each representing a different sort of churner Customer. This portion of the data is referred to as the training set, and we use two classification techniques (SVMs and DTs) to train the classification model. The above technique makes use of the training datasets generated T.C equation, as well as Part 2 of the datasets used to test the classification model.

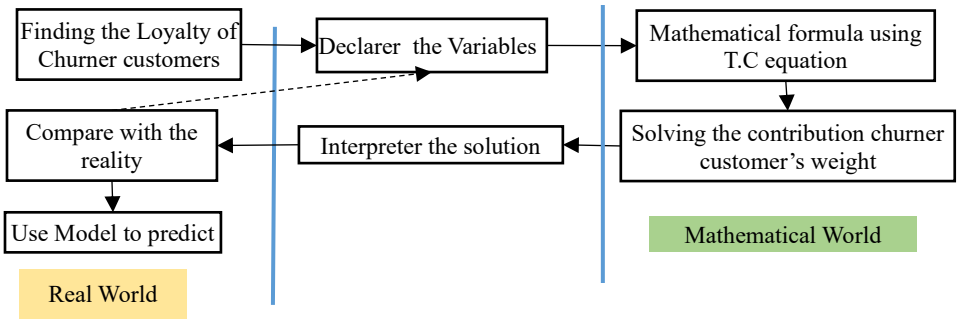

Figure 2. Mathematical Model

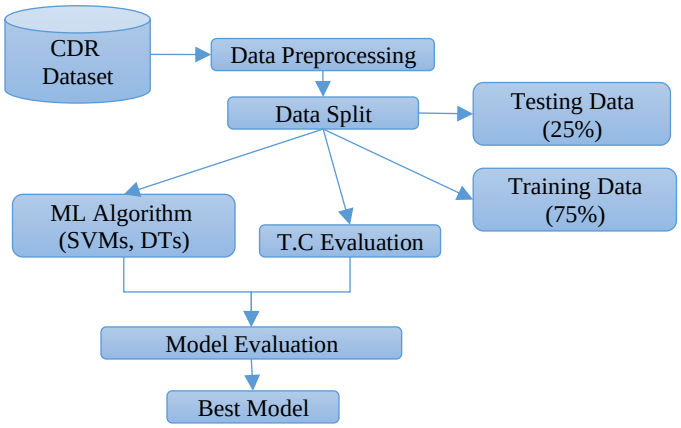

Figure 3. Flowchart for Proposed Model

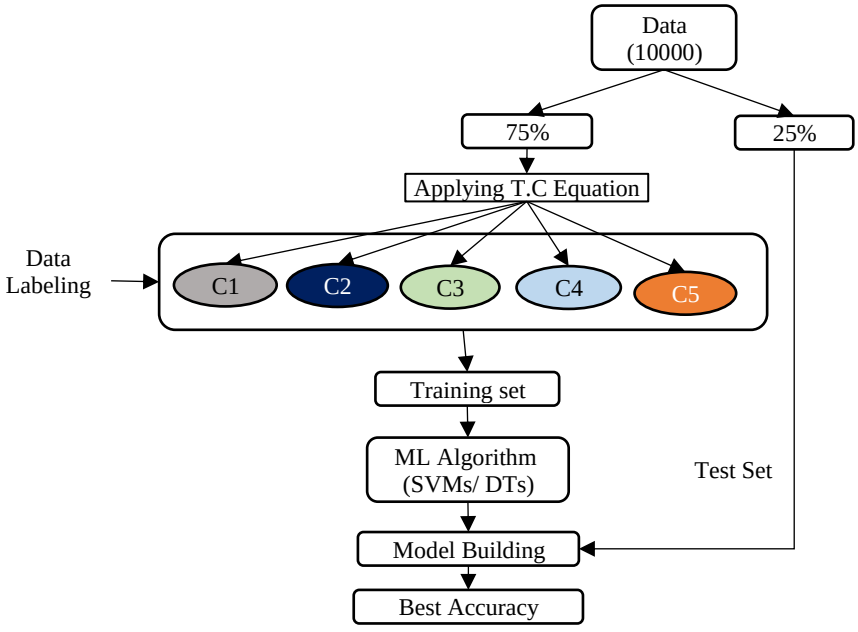

Figure 4. Research Framework

Evaluation Metrics

In this study, we focus on the so-called the confusion matrix. This matrix can be used to make a decision that is constructed by classifier. It contains four categories, including These are: (1) True positives (TP): these are instances correctly labeled as positives; (2) false positives (FP) correspond to negative instances incorrectly labeled as positive; (3) true negatives (TN) refer to negative correctly labeled as negative; and (4) false negatives (FN) correspond to positive instances incorrectly labeled as negative.

In addition, this matrix can be used to build three basic measures; these are recall, precision and F-score. Such measures are readily usable for the evaluation of any multi classifier. Therefore, to evaluate the performance of the call detail records (CDR), We use the above-mentioned criterions as well as the following accuracy measure, which will be calculate as the number of correct predictions accuracy divided by the total number of predictions

Results and Discussions

The results of computing the accuracy show that the accuracy rates were 0.9836 and 0.3304 for DTs and SVMs, respectively. This implies that the DT's classifier predicts better than the SVM's classifier. This result comes from the accuracy rates calculated from both algorithms, which show that SVMs present lowest accuracy rate. In general, the results yielded that classification using DTs method is better than the SVMs method at the present study. This finding can be justified by the fact that SVM is not better for a big dataset because of its much training time and it consumptions more time in training compared to DTs. Furthermore, SVMs method works poorly with overlapping classes and is sensitive to the type of kernel used. On the other hand, the DTs algorithm is faster, regarding the implementation of the algorithms used in the analysis, compared to SVMs algorithm.

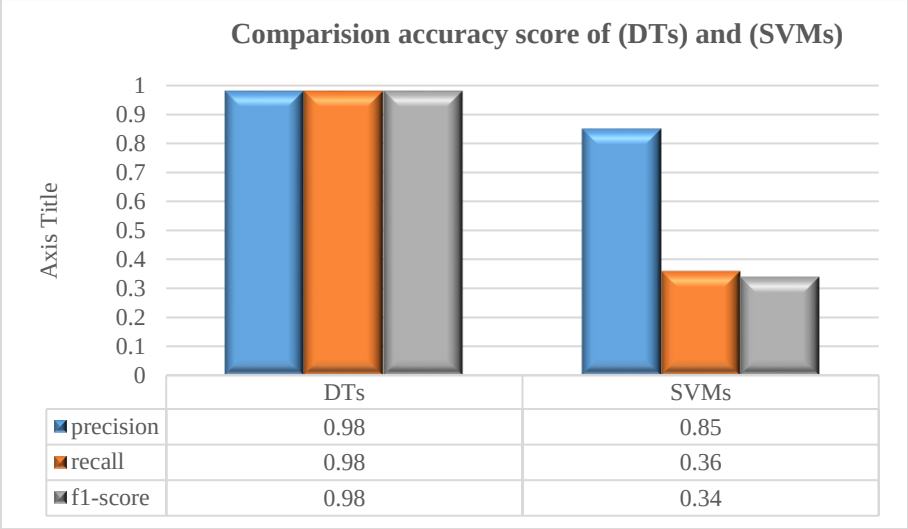

Figure 5. Comparison accuracy score of (DTs) and (SVMs)

TABLE III: TESTING OPTIONS (SVMS)

Training option	Correct classify instance %	Incorrect classify instance %
Training set	33.88%	66.12%
Cross validation folds=10	35.64%	64.36%
Percentage split 65%	34.76%	65.24%
Kapa = 0.0041		

TABLE IV: TESTING OPTIONS (DTS)

Training option	Correct classify instance %	Incorrect classify instance %
Training set	98.16%	01.84%
Cross validation folds=10	96.00%	04.00%
Percentage split 80%	19.44%	80.56%
Kapa = 0.0072		

Furthermore, Tables, III and IV present the training set that was used to build a predictive model. This was done to fit both algorithms; namely SVMs and DTs. The training set contains the predictor attributes and the class label attributes. Here, we have to make it clear that we first use the training set in the preprocess panel, and then we have selected the algorithms to be used. This is followed by selecting the so-called the choice of 10-fold cross validation. In the second stage, we run both algorithms to indicate the differences in accuracy. Note that when the instances are used as test data, the correctly/incorrectly classified instances can specify the case.

Tables III and IV yielded that 33.88% and 96.16% have been given for SVMs and DTs, respectively, which is to say that the DTs algorithm has a good percentage to achieve the main goal of this study

CONCLUSION

In this study, the comparative study of DTs and SVMs algorithms were implemented to the CDRs Dataset for analyzing the usage of customer services. The result indicates that DTs classifier better than other classifiers, Whereas DTs classifier consumes less time. In future, research will be directed towards selection of different datasets, different behavioral patterns and reality mining.

DTs Classifiers offer good accuracy and perform faster prediction compared to SVMs algorithm. This fact can be justified by the fact that SVMs is not better for a big dataset because of its much training time. Furthermore, SVMs consumptions more time in training compared to DTs. It works poorly with overlapping classes and is sensitive to the type of kernel used.

REFERENCES

[1] Baalaji, K., & Khanaa, V. (2020). A Review on Process of Data Mining Approaches in Healthcare Sectors. Indian Journal of Public Health Research & Development, 11(1), 80-84.

[2] Saura, J. R., Ribeiro-Soriano, D., & Palacios-Marqués, D. (2021). Setting B2B digital marketing in artificial intelligence-based CRMs: A review and directions for future research. Industrial Marketing Management, 98, 161-178.

[3] Hong Chen , Ling Li & Yong Chen (2021) Explore success factors that impactartificial intelligence adoption on telecom industry in China, Journal of Management Analytics, 8:1,36-68.

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

[5] Manero, K. M., Rimiru, R., & Otieno, C. (2018). Customer Behaviour Segmentation among Mobile Service Providers in Kenya using K-Means Algorithm. International Journal of Computer Science Issues (IJCSI), 15(5), 67-76.

[6] Garg, D., Kappla, S., & To, P. (2017). Methods and systems for call detail record billing systems: Google Patents.

[7] Gunavathi, C., Priya, R. S., & Aarthy, S. (2019). Big Data Analysis for Anomaly Detection in Telecommunication Using Clustering Techniques

[8] Information Systems Design and Intelligent Applications (pp. 111-121): Springer.

[9] Ruan, N., Wei, Z., & Liu, J. (2019). Cooperative Fraud Detection Model with Privacy-Preserving in Real CDR Datasets. IEEE Access, 7, 115261-115272.

[10] Jiang, D., Wang, Y., Lv, Z., Qi, S., & Singh, S. (2019). Big Data Analysis-based Network Behavior Insight of Cellular Networks for Industry 4.0 Applications. IEEE Transactions on Industrial Informatics.

[11] Rizwan, A., Nadas, J., Imran, M., & Jaber, M. (2019). Performance Based Cells Classification in Cellular Network using CDR Data. Paper presented at the ICC 2019-2019 IEEE International Conference on Communications (ICC).

[12] Doyle, C., Herga, Z., Dipple, S., Szymanski, B. K., Korniss, G., & Mladenić, D. (2019). Predicting complex user behavior from CDR based social networks. Information Sciences.

[13] Al-Zuabi, I. M., Jafar, A., & Aljoumaa, K. (2019). Predicting customer's gender and age depending on mobile phone data. Journal of Big Data, 6(1), 18.

[14] Rutkowski, L., Jaworski, M., & Duda, P. (2020). Basic Concepts of Data Stream Mining Stream Data Mining: Algorithms and Their Probabilistic Properties (pp. 13-33): Springer.

[15] Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J . .(2016)Data Mining: Practical machine learning tools and techniques :Morgan Kaufmann.

[16] Petrova, E., Pauwels, P., Svidt, K., & Jensen, R. L. (2019). In search of sustainable design patterns: Combining data mining and semantic data modelling on disparate building data Advances in Informatics and Computing in Civil and Construction Engineering (pp. 19-26): Springer.

[17] Leung, C. K.-S. (2019). Big data analysis and mining Advanced Methodologies and Technologies in Network Architecture, Mobile Computing, and Data Analytics (pp. 15-27): IGI Global.

[18] Rani, H., & Gupta, G. (2019). Prediction Analysis Techniques of Data Mining: A Review.
[19] Kadhim, A. I. (2019). Survey on supervised machine learning techniques for automatic text classification. Artificial Intelligence Review, 52(1), 273-292.
[20] Kadhim, A. I. (2019). Survey on supervised machine learning techniques for automatic text classification. Artificial Intelligence Review, 52(1), 273-292.
[21] Adebayo, A. O., & Chaubey, M. S. (2019). DATA MINING CLASSIFICATION TECHNIQUES ON THE ANALYSIS OF STUDENT'S PERFORMANCE. GSJ, 7(4).
[22] Sudharsan, R. (2020). SVM Based Churn Analysis for Telecommunication. International Journal of Advanced Research in Engineering and Technology, 11(6).
[23] Xia, G.-e., & Jin, W.-d. (2008). Model of customer churn prediction on support vector machine. Systems Engineering-Theory & Practice, 28(1), 71-77.
[24] Albadawi, S., Fraz, M. M., & Kharbat, F. (2017). Telecom churn prediction model using data mining techniques. Bahria University Journal of Information & Communication Technologies (BUJICT), 10(Special Is).
[25] Huigevoort, C., & Dijkman, R. (2015). Customer churn prediction for an insurance company. Master Thesis (Eindhoven University of Technology).
[26] Pamina, J. and Raja, Beschi and SathyaBama, S. and S, Soundarya and Sruthi, M. S. and S, Kiruthika and V J, Aiswaryadevi and G, Priyanka, An Effective Classifier for Predicting Churn in Telecommunication (June 6, 2019). Jour of Adv Research in Dynamical & Control Systems, Vol. 11, 01-Special Issue, 2019 , Available at SSRN: https://ssrn.com/abstract=3399937
[27] Mitrović, S., Baesens, B., Lemahieu, W., & De Weerdt, J. (2019). tcc2vec: RFM-informed representation learning on call graphs for churn prediction. Information Sciences.
[28] Arnold, T., Kane, M., & Lewis, B. W. (2019). A computational approach to statistical learning: Chapman and Hall/CRC.
[29] 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.
[30] Guo, H., & Wang, W. (2019). Granular support vector machine: a review. Artificial Intelligence Review, 51(1), 19-32.
[31] Blanco, V., Japón, A., & Puerto, J. (2019). Optimal arrangements of hyperplanes for SVM-based multiclass classification. Advances in Data Analysis and Classification, 1-25.
[32] 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.
[33] Pham, B. T., & Prakash, I. (2019). Evaluation and comparison of LogitBoost Ensemble, Fisher's Linear Discriminant Analysis, logistic regression and support vector machines methods for landslide susceptibility mapping. Geocarto International, 34(3), 316-333.
[34] Sun, X., Su, S., Huang, Z., Zuo, Z., Guo, X., & Wei, J. (2019). Blind modulation format identification using decision tree twin support vector machine in optical communication system. Optics Communications, 438, 67-77.
[35] Kaggle Datasets, Available at:https://www.kaggle.com/datasets,mobile-phone-activity
[36] VanRossum, G., & Drake, F. L. (2010). The python language reference: Python Software Foundation Amsterdam, Netherlands.
[37] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., . . . Dubourg, V. (2011). Scikit-learn: Machine learning in Python. the Journal of machine Learning research, 12, 2825-2830.