Telecom Churn Prediction Using Voting Classifier Ensemble Method and Supervised Machine Learning Techniques

O.Pandithurai[1*], *Sriman* B[1], *Hrudhai* Narayan S[1] and *Humaid* Ahmed H[1]

[1]Rajalakshmi Institute of Technology, Chennai, India

Abstract. In the current fast-paced world, there are a lot of changes and developments in the telecom sector, due to which the telecom companies find themselves in difficulties in retaining the customers who have availed of their services. In order to solve this problem, churn prediction system is needed to predict customer churn. So far, there are many supervised machine learning churn prediction models that compare various machine learning and deep learning models, select one model, and create a whole churn prediction model. The solution proposed has various supervised machine learning models like Support Vector Machine (SVM), Random Forest Classifier, Decision Tree Classifier, and Logistic Regression Classifier and combine all the models together using an ensemble method called Voting Classifier to produce a single model that considers all the mentioned algorithms and produces an optimum result. The above-mentioned model will be trained by the telecom dataset containing the records of 7043 customers, and the target field is classified into churned and stayed. The machine learning algorithm is evaluated using various performance metrics such as the F1 score, precision, confusion matrix, classification report, and accuracy.As the result the churn prediction model has shown 84% accuracy .

Keywords. Churn prediction, Dataset, Supervised Machine Learning Techniques (SMLT), Voting Classifier,Random Forest Classifier(RF).

1 Introduction

The churn prediction model is one of the most useful tool to retain their customers in the telecom industries [1]. There are many large scale telecom companies that are losing their customers because of extreme competitive market and poor services that are provided to the customers. hence , the customer leaves the services of that company and avail the services of another company that are providing better deals to the customers . To drastically reduce the customer churn telecom companies can use churn prediction model whose main work is to identify those customers who are most likely to leave the services [2],[3]. After identifying those customer who are about to leave the services , the telecom company can approach to these customers and persuade them to not to leave the services by rectifying the

* Corresponding author: pandics@ritchennai.edu.in

issues they are facing or provide better deals to them. If this churn prediction system is used by the telecom industries then they can increase their customer base an in the larger perspective they can generate lot of revenue. These churn prediction model works by training various machine learning models with the labelled datasets available in their company . These models will be ready to use for prediction. The datas that are present in the datasets are collect through various ways like using CRM or through getting feedbacks from the customers both online and offline.

2 Methodologies

The research for this paper is done by reading various research papers available on websites like Springer, IEEE, Researchgate and Science Direct. The research papers helped to understand various research objectives, working of various machine learning and deep-learning models and a criteria for inclusion and exclusion of various ideas in the survey **.**

2.1 Search Strategies

The literature reviews in this paper are done by reading various papers from well- reputed journal databases that are mentioned above. These reviews are mostly from the papers that are published in the conference, international journals and book chapters after 2017.

2.2 Requirements for Inclusion and Exclusion

Considering the following criteria, research and journal papers were selected.

- Techniques and quality of the churn prediction models in research papers.
- Papers which were published in the reputed journal mostly from the year 2017 to 2022.

To gather the large amount of information contained in the papers, the following exclusion criteria have been used.

- papers that were published in languages other than English.
- Those papers whose abstract is only available and to read whole papers request need to be send.
- Those papers whose end result is not according to the intended results and with no proof of their findings.

3 Related works

There are wide variety of papers that have examined various churn prediction models that have combined various machine learning and deep learning to somehow increase the efficing of the churn prediction models . In one paper authors have used various machine learning algorithm called Logistic regression(LR), Adaboost and Random forest classifier in the model and they have even used many deep learning models called CNN and ANN in their churn prediction model and the results that they have got is deep learning model had a better performance and accuracy compare to machine learning algorithm [1]. Another paper have worked on the models like Gradient boost and Random forest algorithm where they came to a conclusion that Gradient boost classifier have performed pretty well compare to other algorithm where it showed the accuracy of 91% [2]. In another paper the authors have

used hybrid models which combines both classification and clustering model using diiferent ensemble methods like stacking, bagging and voting where they got to know that gradient boost, clustering k-means ,decision tree classifier and deep learning models combined using ensemble methods have shown 94.30% in github datasets and 92.43 on bigml datasets [3].

In another paper the authors have used used various classification algorithms like random forest,Logistic regression(LR), SVM and some neural network models where they found that random forest has a better accuracy 91.73% and outperforms other algorithms in performance metrics [4]. In another paper authors have used a unique way to create a churn prediction models were they have used twitter hastags to judge the sentiments of the users and find out whether the user is about to churn or not. In this they used Naive bayesian classifier as there model to identity the potential churn customers [5]. In one paper the researcher have used datsets from a telecom industries in malaysia where she used various classification models like KNN, Logistic regression(LR) and Linear Discriminant Analysis(LDA), CART, SVM and Guassian Naive bayes where they got the result that CART has worked well in identifying the churn more accurately with an accuracy of 98% which is extremely high so far of the research that has been on this domain [6]. In another paper idea is pretty much the same where they have used telecom industry data in Nepal and used XGBoost algorithm to device their churn model where they got 86.30% as the accuracy [7]. In another paper the authors have clearly used two algorithms called Logistic regression(LR) and Logit Boost where they found that Logit boost had outperformed logistic regression with the accuracies of 79.40% and 75.1% [8].

4 Telecom churn prediction system

4.1 Dataset Information

The datasets that are being used in this designed model are from Kaggle and contain the data of 7043 customers. The target field in this dataset is the customer status, which is classified into two types: churned, stayed and joined.

4.2 Data Pre-processing Stage

In this process, the data contained in the dataset is analysed and checked to see if there are any missing or duplicate values that need to be removed so that the model can operate at its optimal level [4],[5]. For doing this data processing, a Python library called Pandas is used, which contains all the features needed to manipulate the data for the machine learning model. This process is necessary because machine learning does not work on those datasets that have a lot of irregularities [8]. It eases the work to visually analyse the large chunk of data using data visualisation. In this step, after dropping all those missing values, the shape of the data is checked to know the amount of data that has been deleted from the dataset.

4.3 Data Visualization Stage

This process is very crucial to understanding the large labelled dataset in a better way, which further helps to find the correlation between various fields. So that the feature can be used to train the enhanced machine learning model that has been created [11]. To do this data visualisation, various Python libraries like Matplotlib and Seaborn are used. These libraries are used to create various graphs like bar graphs and pie charts, as shown in Fig. 1. and Fig. 2., which are a better and easier way to extract common features that are closely

related to the target field, which is the customer status in this case [15]. Before creating any graph, the column name in the dataset needs to be renamed in such a way that there is no space between the words of the column name.

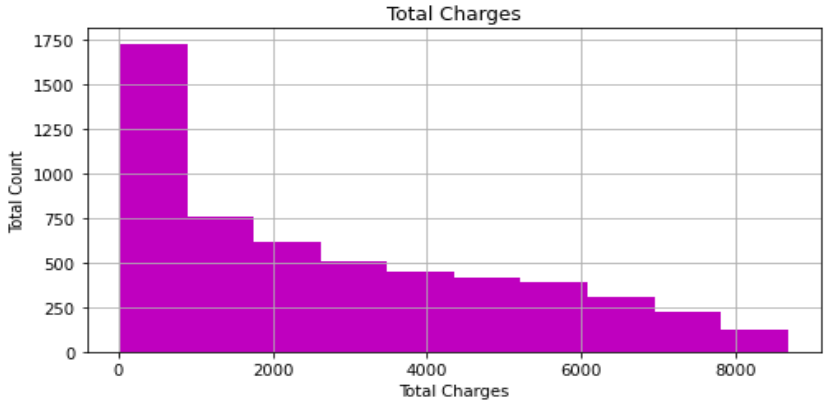

Fig. 1. Total number of customers having total charges

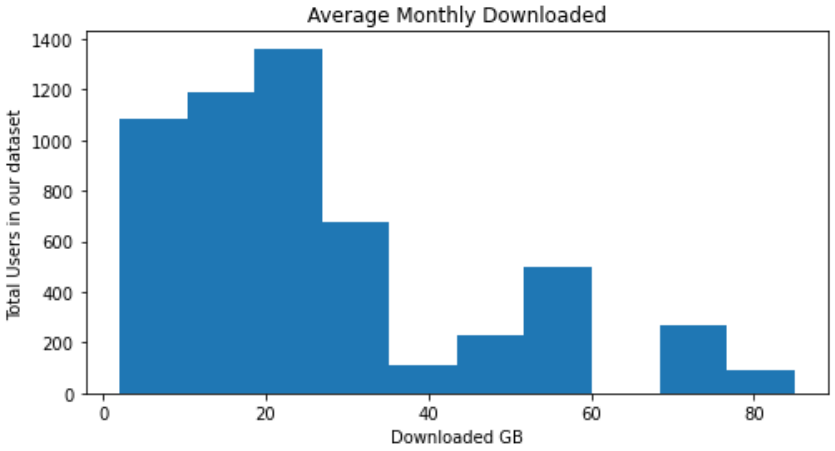

Fig. 2. Total number of users have average monthly downloads.

4.4 Algorithm Implementations

Before training the model, the compatibility of dataset is ensured. The first and foremost thing that needs to be done is to rename the column fields in such a way that if the column name has more than one word, replace the space with an underscore before converting the categorical data into an integer type because the machine learning models only understand numbers [14]. The conversion is done using the label encoder, which is in the sklearn library of Python,, convert categorical values into 0 or 1. Then the data is split so that 70% is used for training the model and the remaining 30% is used for testing purposes to check whether the algorithm is working perfectly or not.

After this, import various machine learning algorithms like support vector machines (SVM), random forests (RF), decision trees (DT), and logistic regression (LR) using the library called sklearn[17]. When the models are initialised, import the voting classifier

using Sklearn, where it will combine all four different algorithms together, which converts all the models into single-churn prediction models, as shown in Fig. 3. . The benefits of this churn model using a voting classifier are that it understands the shortcoming between various machine learning models. And produce optimum results.

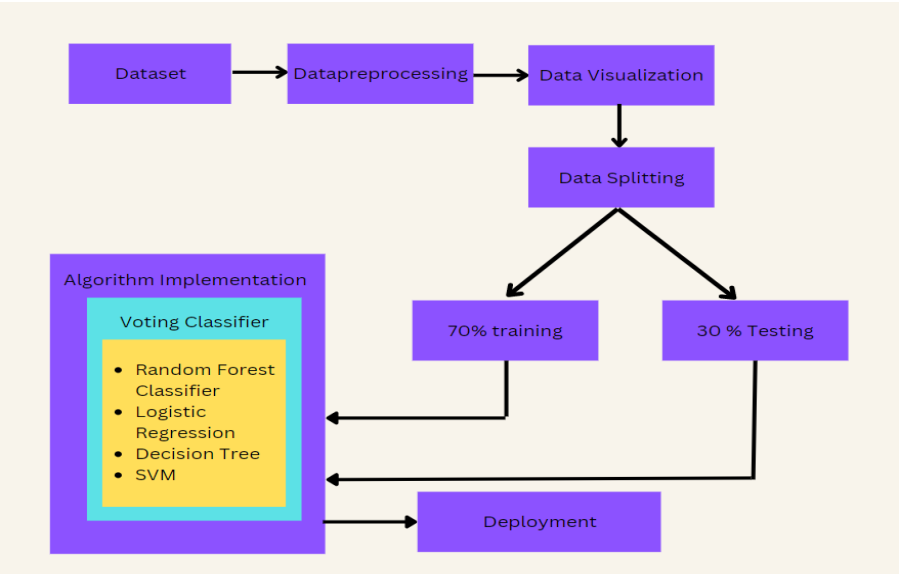

Fig. 3. Architechture of Telecom Churn Prediction Model

4.5 Performance Evaluation of the model

At this stage, the prediction models will be evaluated to see how well they performed after training the created model with the Training data[8],[9]. Various performance metrics are used to measure the working of these models, like the confusion matrix, classification report, and accuracy.

4.5.1 Confusion matrix

The confusion matrix is one of the ways to measure the performance of the machine learning model. Before any discussion about the confusion matrix, know about the terms like true positive, true negative, false positive, and false negative. True positive (TP) means that the value that the machine learning model has predicted is the same as the actual value[10],[11]. True negative (TN) means that the model has predicted a false value that matches the actual value. False positive (FP) means that the predicted value by the churn model is true but the actual value is found to be false. False negative (FN) means the model has predicted a false value that is not equal to the true actual value. The confusion matrix actually creates a 2D array of all these values, which helped to better understand the workings of the model , as shown in Fig. 4. .

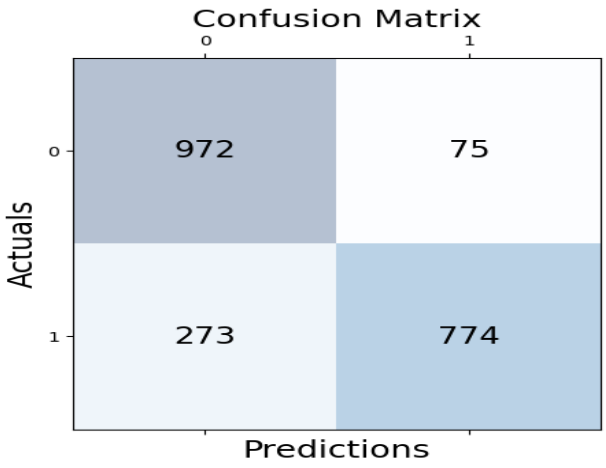

Fig. 4. Confusion Matrix of the Telecom Churn prediction model done

4.5.2 Classification Report

The classification report gives all the information about the machine learning classification model that has been created. It only contains three pieces of information: precision, F1 score, accuracy, and recall, as well as other data[10].

4.5.2.1 Precision

A statistical parameter called precision is used to assess the accuracy of a classification or predictive model [2]. Its percentage, out of all the cases the model correctly predicted as positive, is known as the true positive rate, as shown in Table 1. To put it another way, precision evaluates a model's ability to detect genuine positives while minimising false positives (incorrectly predicted positive instances) [1].

Table 1. Classification report of the Churn prediction model using voting classifier

	Precision	**Recall**	**F1-score**	**Support**
0	0.79	0.93	0.85	1047
1	0.92	0.75	0.82	1047
accuracy			0.84	2094
Macro avg	0.85	0.84	0.84	2094
Weighted avg	0.85	0.84	0.84	2094

4.5.2.2 F1 score

The F1-Score is a measurement that combines recall and precision. Generally speaking, it is referred to as the harmonic mean of the two [12]. Another method of determining a "average" of values is the harmonic mean, which is typically seen as better suited for ratios than the conventional arithmetic mean (such as recall and precision) [10].

4.5.2.3 Accuracy

A statistical parameter called accuracy is used to assess how well a classification algorithm or predictive model is doing [8]. It is described as the proportion of accurate forecasts to all of the model's predictions, as shown in Fig. 4. . In other words, accuracy assesses how accurately a model can place examples in the appropriate categories.

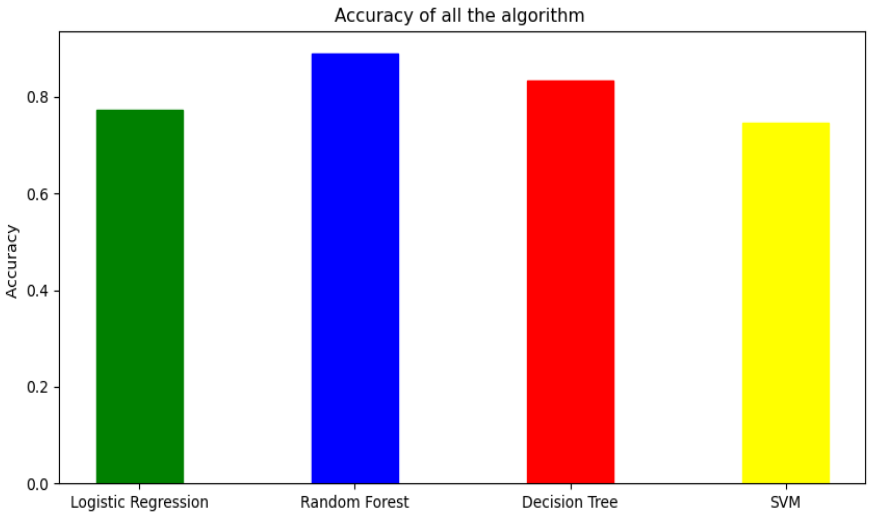

Fig. 5. Accuracy of various algorithms that has been used in the voting classifier.

4.5.2.4 Recall

It is a measure that is used to measure how many positive values the classifier was able to predict accurately by comparing the actual positive values in the dataset [10].

4.6 Deployment Stage

In the deployment part, the created model will be converted into .pkl format, which is used to create an interactive user interface to make the churn prediction work in real life. pkl is created using the python library called joblib. Using the Django framework, create churn prediction model and for further styling of the page CSS is used and for the structure of the website HTML is used. Through this Churn prediction model is ready to be used.

5 Results and Discussion

In the churn prediction model for the telecom churn, going through various phases like data preprocessing, data visualisation, and so on. where it was possible to build the model that can show the optimum solution for the required inputs that needed to be entered. According to the work done, it is found out that the voting classifier, which is an ensemble method that combines various predictions from the machine learning models, has an accuracy of 84%, as shown in Fig. 6.

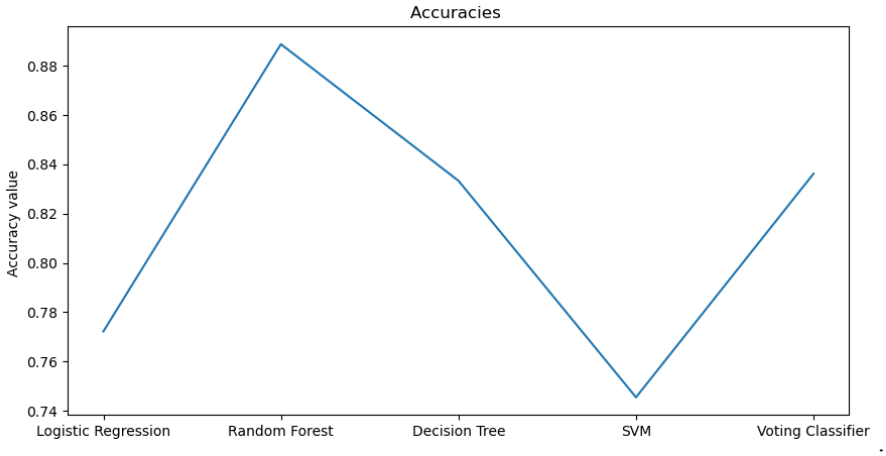

Fig. 6. Accuracy of various algorithms that has been used in the model.

There are others models that performed better; for instance, the random forest classifier had an accuracy of 88%. which, if compared, is better than the model created, but if there is any change in the datasets, then its accuracy might decrease, and the voting classifier has the upper hand in this area where it can combine the neglected features of all these algorithms, which were the reason for their accuracy being reduced, and has the capability to produce the optimum solution.

To further measure the proper working of the model, various evaluations using various well-known performance metrics that are used to measure the efficiency of the models, like recall, accuracy, precision, F1 score, and AUC-ROC values, as shown in table 2 and table 3.

Table 2. Performance metrics of various models in the Telecom churn prediction system

Models	**Precision**		**recall**		**F1 score**	
	Churned	**Stayed**	**Churned**	**Stayed**	**Churned**	**Stayed**
Voting Classifier	0.79	0.92	0.93	0.95	0.85	0.82
Logistic Regression	0.74	0.81	0.83	0.71	0.78	0.76
Random Forest	0.87	0.90	0.91	0.87	0.89	0.89
Decision Tree	0.80	0.88	0.89	0.78	0.84	0.82
Support vector machine	0.75	0.74	0.73	0.76	0.74	0.75

Table 3. ROC-AUC scores of all the models in the churn model

Models	ROC_ AUC score
Voting Classifier	0.83
Logistic Regression	0.77
Random Forest	0.88
Decision Tree	0.74
Support Vector Machine	0.83

A graph is aslo generated to show how the churn prediction models work. where marker x is the predicted value and marker 'o' is the actual test value, as shown in Fig.7. . If both values coincide, then the model has predicted the churn accurately, or else it has failed to identify the customer who is about to churn from the telecom services.

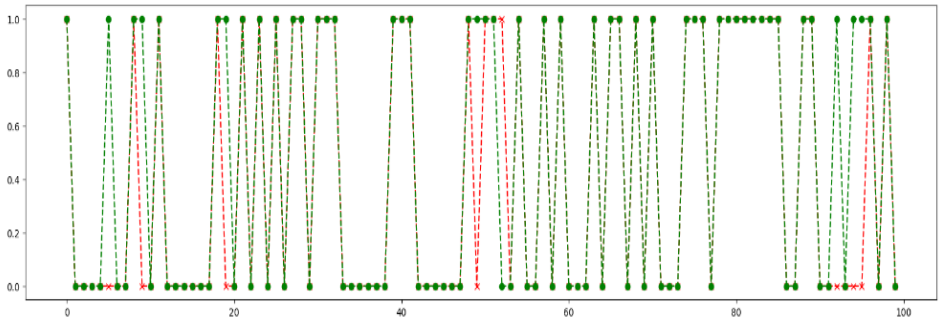

Fig. 7. Plottings of the prediction value and the actual values of the model.

6 Conclusion

The procedure for designing the machine learning models starts with data preprocessing, where all the missing values have been removed to make the data ready for the next step called data visualisation, various graphs are created to understand the dataset in a proper way and to extract certain features that will be used in training the churn prediction model. The features that used in my model are tenure in months, internet service, internet type, average monthly GB download by a user, unlimited data, payment method,monthly charge, total extra charges for the data, and the target field called customer status, which contains two values: churned and stayed.

After completing this process, the dataset is converted into an integer so that the data contained in the dataset could be used to train the machine learning prediction model. After this,the performance measure is done, where it is found out the model created has 84% accuracy. This model is useful for predicting those customers who are about to leave the service.

References

1. Chellam G.H, A., Mahalekshmi, "Analysis of customer churn prediction using machine learning and deep learning algorithms", Int. J. Health Sci., vol 11684-11693,2022, doi: https://doi.org/10.53730/ijhs.v6nS1.7861.
2. Cui &Ye, Xia, Duan & Bohan, Yunhuai, "Analysis and Prediction of Telecom Customer Churn based on Machine Learning ", Highlights in Science, Engineering and Technology, vol 16, pp 131-145, doi: 10.54097/hset.v16i.2495.(2022)
3. A.Almazroi, Syed Fakhar Bilal, Khan F.H, Bashir S, "An ensemble based approach using combination of clustering and classification algorithm to enhance customer churn prediction in telecom industry" , Peerj Computer Science, vol 8, doi: 10.7717/peerj-cs.854,(2022).
4. B.Raza, I.Ullah, A,K, Malik, S.U. Islam, M. Imran and S.W. Kim, "A Churn Prediction Model Using Random Forest Analysis of Machine Learning Techniques for Churn Prediction and Factor Identification in Telecom Sector", IEEE Access, vol 7,pp 60134-60149, doi: 10.1/ACCESS.2019.2194999.
5. Sumesh, Sandeep & Sood, Ranjan, "Sentimental Analysis Based Telecom Churn Prediction, Journal of Web Engineering and Technology", vol 7 Issue 1, pp 6-12, (2020).
6. Sook Ling,L., Mustafa, M.M.,& Abdul Razak, S.F., "Telecom churn prediction for telecommunication industry: A Malaysian Case Study", F1000 Research, vol 10, doi: 10.12688/f1000research.73597.1 ,(2021).
7. Aman Shakya,Sagar Maan Shrestha, "A Customer Churn Prediction Model Using XGBoost for the Telecom Industry in Nepal", Procedia Computer Science , vol 215,pp.652-661,doi: https://doi.org/10.1016/j.procs.2022.12.067,(2022)
8. Jain, Ajay & Shrivastava, Hemalatha & Khunteta, Sumit, "Churn Prediction in Telecommunication using Logistic Regression and Logit Boost", ICCDS (2019), vol 167 pp.101-112, (2020)
9. Dwivedi, A., "Telecom Industry: Customer Churn Prediction", (2019)
10. Jafar,A.,Ahmad, A.k & Aljoumaa,K., "Customer churn prediction in telecom using machine learning in big data platform", J big data vol 6 pp.28,doi: https://doi.org/10.1186/s40537-019-0191-6, (2019)
11. Muhammad& Jagurnauth, Jooflo, Rameshwar & Jooflo, Khalid, "Customer Churn Prediction in Telecom Using Machine Learning in Big Data Platform", Journal of Critical Reviews vol. 7 pp.1991, (2020).
12. D.Maheswari, Ammar A.Q., "Churn prediction on huge telecom data using hybrid firefly based classification", Egyptian Informatics Journal (EIJ) vol. 18 Issue 3 pp.215-220, (2017).
13. R.A. Jugurnauth, M.B.A. Joofloo, and K.M.B.A. Joofloo, "A Systematic Review of Algorithms applied for Telecom Churn Prediction" 3rd International Conference on Emerging Trends in Electrical, Electronic and Communications Engineering (ELECOM), Balaclava, Mauritius pp.136-140, doi: 10.1109/ELECOM.2020.9296999. (2020).
14. Mishra, M.K., Lalwani, P., chadha, J.S. et al., "Customer churn prediction system: a machine learning approach" Computing vol.104 pp.271-294, (2022).
15. Aditya& Bhoite, Kulkarni, Sachin, "Customer Churn Analysis and Prediction", International Journal of Computer Applications Technology and Research (IJCATR) vol 8, doi:10.7753/IJCATR0809.1005, (2019).

16. Babar Shah, Feras Al-Obeidat, Adnan Amin, Awais Adnan, Sajid Anwar, Jonathan Loo, "Customer churn prediction in telecommunication industry using data certainity", Journal of Business Research (JBR) vol. 94 pp.290-301, (2019).
17. B Sriman, P Baby Shamini, Annie Silviya SH, NV Keerthana, A Elangovan, "Deep Learning based Plant Leaf Disease Detection and Classification", ICIRCA pp.702-710, doi:0.1109/ICIRCA54612.2022.9985548, (2022)
18. J Pathmanaba, Annie Silviya, S.H., B Sriman, S Kingsley, "Prediction and Prevention Analysis Using Machine Learning Algorithms for detecting the Crime Data", ICCST pp.986-991, (2022)