DOI: 10.53704/fujnas.v13i2.469

A publication of College of Natural and Applied Sciences, Fountain University, Osogbo, Nigeria.
Journal homepage: www.fountainjournals.com
ISSN: 2354-337X(Online),2350-1863(Print)

Identifying Customer Churn in Telecom Sector: A Machine Learning Approach

***Hambali A. M., Emmanuel L., Olasupo Y. A., Ishaku A**

Federal University Wukari, PMB 1020, Katsina-Ala Road, Wukari, Nigeria

Abstract

Nowadays, there is no shortage of options for customers when choosing where to put their money. As a result, customer churn and engagement have become one of the top issues. With the increase in the number of service providers for the same targeted population, there is a need for service providers to try to find the changing customer behaviour and their rising expectations to retain them. Various studies have proposed customer churn. Data mining was routinely used to predict telecom customer attrition. Most researchers have compared and proposed different approaches for the prediction of customer churn, though, some of the Machine learning (ML) algorithms used were unable to provide the performance needed to identify customer churn. Therefore this paper presents a comparative analysis of Support Vector Machine (SVM), Decision Tree (DT) and Random Forest (RF) in Telecommunications Dataset. To prepare the dataset for machine learning algorithms, chi-square was used for feature selection to select the most informative features from the original dataset. We validate our model using a ten-fold cross-validation approach to test the performance of our models. RF model performed better than other models in terms of accuracy (94%), precision (94%) and F-measure (94%) respectively. Additionally, we compared our results with existing models that used the same dataset, the proposed strategy outperformed them.

Keywords: Customer Churn, Telecommunication, Machine Learning, Support Vector Machine (SVM), Decision Tree (DT), Random Forest (RF)

Introduction

The telecommunication market is very dynamic and highly competitive nowadays. The number of service providers has increased very rapidly in every business (Rahman & Kumar, 2020). The challenges of service providers are finding the changing customer behaviour and their rising expectations. These days, there is no shortage of options for customers when choosing where to put their money or chosen service provider. As a result, customer churn and engagement have become one of the top issues. Therefore, this is a big challenge for the new generation telecommunication service providers to think innovatively to fulfil and add value to the customers.

Several scientific journals and corporate publications report exciting developments in technology and competition in the mobile telecoms sector. Fujo et al. (2022) reported that Telecommunications has long been part of our

* Corresponding Author:https://orcid.org/0000-0002-2194-9376
Email address: hambali@fuwukari.edu.ng

culture but is attracting more attention now than ever before. A compelling case is the General Packet Radio Service (GPRS) - based medium broadband networks' rapid growth, which now hit a market penetration level of 70–80 per cent in many European Union (EU) countries.

An important discipline within database marketing is customer retention management (CRM), or the prevention of customer churn, defined as the propensity of customers to end the relationship with the company, and to switch to the competition. Churn prediction is the process of using transaction data to identify customers who are likely to cease their relationship with a company (Leung & Chung, 2020).

Nowadays, CRM systems are progressively utilizing the potential of machine learning (ML) to forecast customer churn. By scrutinizing historical customer data and employing predictive models, companies can pinpoint customers at risk and take proactive steps to keep them, ultimately elevating customer retention rates and bolstering long-term profitability. ML is a subset of artificial intelligence (AI) dedicated to creating algorithms and statistical models. These models empower computer systems to enhance their performance in particular tasks by gaining knowledge and experience from data, all without the need for explicit programming. In the realm of ML, systems are trained using data to discern patterns, provide predictions, or enhance processes. These systems gain knowledge from data, adjust to new information, and have the capacity to make decisions or predictions based on the insights they have acquired. ML finds applications across diverse domains, such as image and speech recognition, natural language processing, recommendation systems, and predictive analytics, among others.

Predicting customer turnover is a key component of database marketing, intending to identify consumers most likely to depart the firm (Tékouabou et al., 2022). Prediction models disclose the number of customers who will quit the influence of categorizing customers, and the program's profitability. Various studies have proposed customer churn prediction and detection using ML to routinely predict customer attrition in telecoms. Most researchers have compared and proposed different approaches for the prediction of customer churn (Jain, Khunteta, Srivastava, et al., 2020), though, these approaches were unable to provide the required accuracy needed to identify customer churn (Saheed & Hambali, 2021). Therefore, this paper presents a comparative analysis of Support Vector Machine (SVM), Decision Tree (DT) and Random Forest (RF) on Telecommunication Dataset. This paper is organised as follows. The related work is presented in section 2. Section 3 explains the methodology used in this work, while the results and discussion are presented in section 4, and section 5 concludes the work.

Review of Related Works

Some machine learning techniques have been used to overcome customer churn challenges in the banking and Telecommunication industries, a few among them being as follows. Ahmad et al., (2019) developed a churn prediction model designed to aid telecom operators in identifying customers most likely to churn. The proposed model leverages machine learning techniques in a big data platform and introduces novel methods for feature engineering and selection. To assess its performance, the authors employed the Area Under the Curve (AUC) as the standard metric, achieving an impressive AUC value of 93.3%. Another significant contribution of the approach is the incorporation of customer social networks into the prediction model, achieved by extracting features using Social Network Analysis (SNA). This addition substantially improved model performance, elevating the AUC from 84% to 93.3%. They implemented and tested the model within a Spark environment, using a large dataset derived from extensive raw data provided by SyriaTel, encompassing customer information spanning nine months. This dataset served as the basis for training, testing, and evaluating the system at SyriaTel. Their study explored four different algorithms, including Decision Tree, Random Forest, Gradient Boosted Machine Tree (GBM), and Extreme Gradient Boosting (XGBOOST). However, the most promising results were obtained with the application of the XGBOOST algorithm, which was used for the churn prediction model. The limitation of this work is that there is a decrease in results when new data (unseen data) was supplied to the model, this might be attributed to the non-stationary nature of the data model, which necessitates periodic retraining of the model.

Jain, Khunteta, & Srivastava, (2020) proposed a model for predicting customer churn in the telecommunications industry using logistic regression and logit boost algorithms. The methodology involves data filtering and cleaning, followed by applying the algorithms to updated data. The results are evaluated using different measurement criteria. The results showed that logistic regression outperformed logit boost with 0.1176 kappa, mean absolute error of 0.2237 and accuracy of 85.24%. One of the limitations is that the model is based on a specific dataset, and the results may vary for different datasets. Another limitation is that the model does not take into account external factors that may affect customer churn, such as changes in the market or competition. Finally, the authors acknowledge that the proposed model may not be suitable for all types of telecommunication companies, and further research is needed to generalize the findings.

Leung & Chung, (2020) introduced an innovative method for enhancing model specifications through the utilization of time-series predictors, data spanning various periods, and the identification of rare events, all aimed at improving the accuracy of churn prediction. The research employed a distinctive dataset encompassing three years' worth of records, comprising 32,000 transactions from a retail bank located in Florida, USA. The approach involved trend modelling to capture the evolution of customer behaviour over time. The findings indicate that incorporating data from multiple periods led to enhancements in both model precision and recall. The results showed that for 6 months, accuracy was 94.78%, precision 24.8% and recall was 31.14%. The result of the 4 months showed that accuracy was 90.81%, precision 13.89% and recall 30.32%. Furthermore, this dynamic approach to predicting churn can be extended to other domains that require the analysis of long-term customer data. One drawback of this approach is the potential issue of non-independence arising from having multiple observations of the same individuals. Because the training data is collected from various periods, data from the same customers is repeatedly incorporated. Even though customer behaviour might vary across different periods, the static predictors remain consistent, which could pose a challenge to the assumption of independence.

Alboukaey et al., (2020) developed an approach of predicting churn daily instead of the traditional monthly approach, leveraging the dynamic daily customer behaviour rather than their monthly patterns. The dataset was collected from the MTN operator in the country and it spans for150 days. To achieve this, the authors represent a customer's daily activity as a multivariate time series and put forth four models designed for daily churn prediction based on this representation. Two of these models rely on features derived from the multivariate time series: one is Recency, Frequency, and Monetary (RFM)-based, and the other is statistics-based. The remaining two models harness deep learning techniques for automated feature extraction, specifically Long Short Term Memory (LSTM)-based model and a Convolution Neural Network (CNN)-based model. The results highlight that the daily models surpass the monthly models significantly in terms of predicting churn earlier and with higher accuracy. Additionally, the (LSTM)-based model (Daily prediction: AUC 0.918 and F1-score of 0.571; Monthly prediction: 0.844 AUC and 0.463 F1-score) outperforms the Convolution Neural Network (CNN)-based model, although both match the predictive performance of the (RFM)-based model. Furthermore, all three of these models exhibit substantial improvements in predictive performance when compared to the statistics-based model. Two primary concerns could be explored in future research: interpretability and the effectiveness of customer retention. In terms of interpretability, the current models do not establish causality, which is essential for precise targeting. Therefore, future work can focus on creating a daily churn prediction model that incorporates causality, providing results that are more easily understood. This would empower companies to address customer churn based on the underlying reasons for their churn.

Rahman & Kumar, (2020) employed Machine Learning (ML) to estimate bank client turnover. Analyzing client behaviour helps researchers predict churn. Classifiers applied include K-nearest neighbour (KNN), Support Vector Machine, Decision Tree, and Random Forest. Some feature selection strategies were used to locate important characteristics and test system performance. Kaggle's churn modelling dataset was used. Comparing the results helps find a precise and predictable model. Random Forest is more accurate than other models after oversampling. The results showed that a combination of the Random Forest classifier with oversampling leads to a superior outcome, yielding an accuracy of 95.74%. It's important to note that feature selection techniques do not impact tree classifiers like Decision Tree and Random Fore. As the results show, reducing features (feature selection) tends to lower the predictive performance of tree classifiers.

Domingos et al., (2021) investigated the influence of various hyperparameters in the context of utilizing Deep Neural Networks (DNNs) for predicting churn within the banking sector. The findings from three experiments indicated that the Deep Neural Network (DNN) model outperformed the Multilayer Perceptron (MLP) when the activation function in the hidden layers was set to rectifier, and the output layer utilized a sigmoid function. The DNN exhibited better performance with a smaller batch size than the size of the test dataset, while the root mean square propagation (RMSprop) training algorithm displayed higher accuracy in comparison to other algorithms such as stochastic gradient descent (SGD), adaptive gradient algorithm (AdaGrad), Adaptive Movement Estimation (Adam), Adadelta, and AdaMax. The results showed that MLP achieved its highest performance accuracy (84.5%) when the training algorithm RMSProp was chosen, but its lowest performance (79.65%) was observed when AdaGrad was used as the training algorithm. Similarly, for the DNN, the highest performance (86.45%) was achieved when RMSProp was selected as the training algorithm, while the lowest performance (83.1%) was recorded when SGD was chosen as the training algorithm. The limitation of the research was that it exclusively used a fabricated dataset obtained from a public data repository, potentially originating from a single bank over a limited period. Consequently, the applicability of this dataset to other banks is questionable, and caution should be exercised when extrapolating the findings to other banking institutions. Linked to the initial constraint of the study was the dataset's imbalance in distribution, with 2000 churners and 8000 non-churners. Even though the stratified cross-validation technique was applied to maintain a proportional representation of each category, this could have had an impact on the accuracy of machine learning classifiers in predicting outcomes.

Saheed & Hambali, (2021) developed various machine learning (ML) techniques, including Support Vector Machine (SVM), Multi-Layer Perceptron (MLP), Random Forest (RF), and Naive Bayes (NB). This research introduces an innovative approach to feature selection, combining Information Gain (IG) and Ranker (R) methods. To assess the model's effectiveness, standard metrics such as accuracy, precision, and F-measure are utilized, alongside 10-fold cross-validation. The outcomes demonstrate an accuracy of 95.02% with feature selection and 92.92% without it. In comparison to existing methods, they concluded that their models exhibit competitive performance, particularly in terms of precision (94.9%) and F-measure (94.7%).

Sudharsan & Ganesh, (2022) introduced an innovative approach for predicting customer churn, employing a deep learning model referred to as the Swish Recurrent Neural Network (S-RNN). The S-RNN is subsequently tailored for the classification of Churn Customers (CC) and regular customers. In cases where a churn customer is identified, the network utilization history is examined to facilitate retention efforts. The results showed that S-RNN achieved classification accuracy of 95.99%, Specificity of 92.31% and Sensitivity of 98.27%. However, it's important to note that this framework does not account for the identification of churn customers based on network usage in specific geographic areas. The telecommunications industry has encountered a significant challenge of customer churn, attributed to factors such as market saturation, intense competition, evolving criteria, and the introduction of enticing new offers.

Sana et al., (2022) explored various machine learning models and data transformation techniques. To enhance the prediction models, the authors utilized a univariate feature selection method and identified the best hyperparameters through grid search. Subsequently, experiments were conducted on multiple publicly available TCI datasets to evaluate the proposed models using widely accepted metrics like AUC, precision, recall, and F-measure. The rigorous experimental analysis highlights the advantages of incorporating data transformation techniques and feature selection during the training of an optimised customer churn prediction (CCP) model. The best model was the Random Forest (RF) model with a logarithmic (LOG) transformation, achieving impressive scores of 0.858 for AUC and 0.82 for F-Measure. Their proposed approach led to significant improvements, with AUC and F-measure increasing by up to 26.2% and 17%, respectively.

Tékouabou et al., (2022) provided a complete Machine Learning (ML) model construction procedure incorporating cross-validation of synthetic minority oversampling technique (SMOTE) to balance data and ensemble modelling. Using balanced data, the random forest (RF) model has 0.86 accuracy and 0.86 f1-score. "Age" was the most important attribute in the created and optimized models, while "HasCrCard" was the least. The developed model was recommended to support bank customer loyalty decisions. The authors recognized that classification algorithms encountered difficulties stemming from factors such as data size constraints, data heterogeneity, non-numeric data, and class imbalance.

Mirabdolbaghi & Amiri, (2022) employed Principal Component Analysis (PCA), Autoencoders, Latent Dirichlet Allocation (LDA), T-Stochastic Neighbor Embedding (T-SNE), and Xgboost for Bank churn. They presented a model aimed at predicting churn in Gradient Boosting Machine (GBM) applications. The model consists of a structured five-step process. The first phase involved data preprocessing which included handling the missing and corrupted values. This step ensures that the input data is clean and appropriately scaled for modelling. In the second phase, the model employs feature selection techniques based on popular algorithms to reduce the dimensionality of the data and retain the most relevant features. Once the feature set is refined, the third step focuses on fine-tuning the hyperparameters of the Light GBM model using Bayesian and genetic optimization approaches, which helps optimize the model's performance. After the model is trained, an interpretability phase follows, where the behaviour of the model and its predictions are analyzed. This phase leverages SHapley Additive exPlanation (SHAP) to understand the influence of each feature on the model's output, providing insights into the model's decision-making process. Finally, the model is applied to rank potential churners based on their estimated customer lifetime value, assisting businesses in prioritizing their efforts. This comprehensive approach aims to enhance the effectiveness of GBM churn prediction and provide actionable insights for better decision-making in customer retention strategies. The proposed technique was also evaluated using four renowned datasets. It outperforms Adaptive Boosting (AdaBoost), SVM, and decision tree on seven evaluation metrics: accuracy, Area under the ROC Curve (AUC), Kappa, Matthews Correlation Coefficient (MCC), Brier score, F1 score, and Economic Model Predictive Control (EMPC). In churn prediction, their algorithm handles imbalanced datasets better based on assessment measures.

Fujo et al., (2022b) created a Deep-Back Propagation Artificial Neural Network (Deep-BP-ANN) using Variance Thresholding and Lasso Regression. Early halting improves the model to prevent overfitting. To prevent overfitting, dropout and activity regularization were investigated for IBM Telco and Cell2cell. The efficiency of the model was determined using holdout and 10-fold cross-validation. Random Oversampling was implemented to balance both datasets. The results demonstrated that the model constructed with lasso regression for feature selection, early stopping to select epochs, 250 neurons for the input and hidden layers, and activity regularization to prevent overfitting for both datasets performs well. Predicting customer turnover, the Deep-BP-ANN model outperforms XGBoost, Logistic Regression, Naive Bayes, and KNN. The Deep-BP-ANN model outperformed all others in terms of accuracy during a hold-out evaluation, achieving a score of 79.38% for Cell2cell. It also demonstrated outstanding precision at 74.5%, recall at 89.32%, F1-Score at 81.24%, and an AUC of 79.38%. Notably, Deep-BP-ANN's results were superior to those of XG Boost, which followed closely. When they considered a 10-fold cross-validation evaluation, the Deep-BP-ANN model once again excelled by achieving an accuracy of 86.57% for IBM Telco. Additionally, it displayed impressive precision at 81.59%, recall at 94.45%, F1-Score at 87.55%, and an AUC of 86.57%.

Zhang et al., (2022) proposed a model for forecasting customer churn in the telecommunications industry by utilizing customer segmentation. Data from three prominent Chinese telecom firms were gathered, and a churn prediction model was crafted using Fisher discriminant equations and logistic regression analysis. The findings indicate that the churn prediction model established through regression analysis displayed superior predictive accuracy at 93.94% and delivered better overall results whereas Fisher's discriminant equations yielded an accuracy of 75%. This research offers telecom companies a valuable tool for effectively anticipating the likelihood of customer churn and implementing specific strategies to prevent it, ultimately leading to increased profitability. The limitation of this work is that the dataset contains data on 4,126 clients spanning from 2007 to 2018. However, it's worth noting that almost four years have passed since that time. Due to the impact of the COVID-19 pandemic, the telecom market and customer behaviour may have undergone significant changes compared to the earlier period. Therefore, we plan to collect more recent data to enhance the model's accuracy and align it with the current market conditions.

Methodology

The methodology adopted by this study takes into consideration three phases for the development of the proposed customer churn prediction model. The first step, read the customer churn dataset that was sourced from the Kaggle machine learning repository. The second phase encompasses the data preprocessing technique that was targeted at reducing complexity while enhancing the model accuracy via the reduced computational overhead cost. The data-preprocessing steps include the check for missing data, label encoding and feature scaling, and more importantly the conductance of feature selection using the chi-square method. The last stage involves feeding the cleansed and selected features to the models, which are the random forest (RF), Decision Tree (DT) and Support Vector Machine (SVM), and also the conductance of performance evaluation analysis from the result produced by the model. The stepwise approach for the implementation of the proposed model is shown in Figure 1.

Dataset Description

The dataset used in this research work is a Telco dataset obtained from Kaggle repository at https://www.kaggle.com/datasets/becksddf/churn-in-telecoms-dataset. Each row represents a customer; each column contains the customer’s attributes. The dataset contained 3333 instances with 21 attributes. Table 1 shows features in the dataset.

Data Preprocessing

To enhance the performance of the machine learning model while reducing the computational cost and complexity of the model, it is essential to conduct data preprocessing. The data preprocessing involves the treatment of empty fields, the encoding of categorical features, scaling of the dataset feature, and more importantly the selection of relevant features based on the chi-square algorithm. The data preprocessing steps are shown in Figure 2

Chi-Square Feature Selection

Feature selection is an important problem in machine learning as it is useful in enhancing model performance with reduced computational cost due to the selection of the most important features in a dataset before feeding the model with the selected dataset attributes. The chi-square

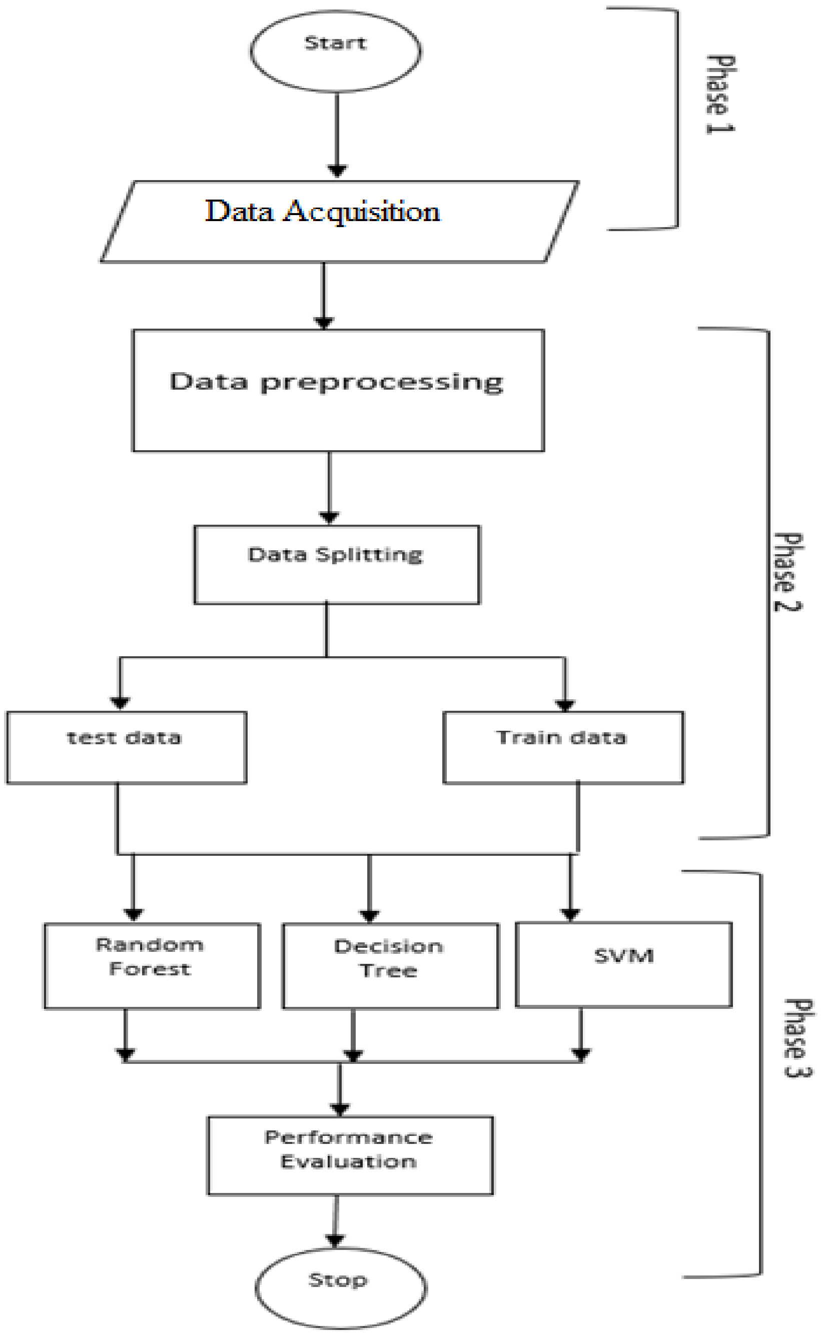

Figure 1: Framework of the proposed Model

Table 1: Dataset Characteristics

S/N	Attributes/features	Data type
1.	State	String
2.	Account length	Integer
3.	Area code	Integer
4.	Phone number	Integer
5.	International plan	string
6.	Voice mail plan	String
7.	Number Vmail messages	Integer
8.	Total day minutes	Double
9.	Total day calls	Integer
10.	Total day charge	Double
11.	Total eve minutes	Double
12.	Total eve calls	Integer
13.	Total eve charge	Double
14.	Total night minutes	Double
15.	Total night calls	Integer
16.	Total night charge	Double
17.	Total Int’l minutes	Double
18.	Total Int’l calls	Integer
19.	Total Int’l charge	Double
20.	Customers service calls	Integer
21.	Churn	String

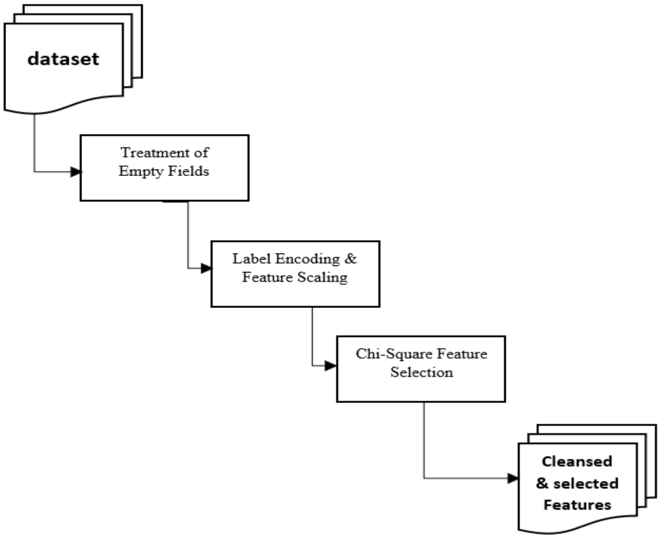
dataset
Treatment of Empty Fields
Label Encoding & Feature Scaling
Chi-Square Feature Selection
Cleansed & selected Features

Figure 2: Data Preprocessing

algorithm is very useful in feature selection as it tests the relationship between the features to deduce the importance and the correlation of features within a dataset (Hambali et al., 2022). Therefore, the chi-square algorithm tests the independence of features, by observing and measuring how the expected and observed features deviate or correlate from each other. The formula for chi-square is depicts in equation 1:

$$x_c^2 = \sum \frac{(o_i - E_i)^2}{E_i} \dots\dots\dots\dots\dots\dots 1$$

Where c is the degree of freedom, o observed values, E expected value.

```
Algorithm 1: Chi-Square Algorithm
Step 1: Define Hypothesis.
Step 2: Build a Contingency table.
Step 3: Find the expected values.
Step 4: Calculate the Chi-Square statistic.
Step 5: Accept or Reject the Null Hypothesis.
```

Algorithm 1: Pseudocode for Chi-Square

Classification Algorithm

This section presents brief explanation about the model used in this study.

Support Vector Machine

The support vector machine is a supervised machine learning algorithm applicable to both classification and regression problems. The goal of the proposed support vector machine algorithm is to separate class labels of customer churn while finding a hyperplane in the data space that produces the largest minimum distance (called margin) between the objects (samples) that belong to different classes of customer churn using a hyperplane. To separate the churn classes rather than using the differences in class means, the proposed SVM uses an object label on the edges of the margin (referred to as support vectors) since the separating hyperplane is supported (defined) by the vectors (data points) nearest the margin. Moreover, to find a hyperplane that separates the inputs into separate groups, the points closest to the hyperplane also known as support vectors are utilized by the support vector machine as there can be many that successfully divide the input vectors. Therefore, in the proposed support vector machine approach, the hyperplane having maximal distance from support vectors is chosen as the output hyperplane.

```
Algorithm 2: Support Vector Machine
SVM $closest = \{closest\ pair\ from\ opposite\ plane\}$
initialize $i, v = violator,\ vc = violator\ checker$
while violating points exist do
    Find $v$
    $closest = closest * v$
    if any cv$<$ 0 then
        $closest = closest/i$
        repeat till points are pruned
    end if
    increment $i$
end while
```

Algorithm 2: Pseudocode for SVM

Random Forest

Random forest (RF) is a supervised machine learning algorithm that was constructed from the decision tree algorithms. Just like the SVM, RF can be used to solve both regression and classification problems. The RF algorithm utilizes the concept of ensemble learning techniques which combines many classifiers to provide solutions to complex problems. This implies that the RF algorithm produces a result based on the predictions of the decision trees by taking the average mean of the output of the various tree. Mathematically, assuming a collection of randomized trees as M. For the $j-th$ tree in the family, the predicted value at the query point x is denoted by $M_n(x;\ \theta_i,\ D_n)$, where $\theta_i \ldots \theta_m$ are independent random variables, distributed the same as a generic random variable θ and independent of D_n. In practice, the variable θ is used to resample the training set before the growth of individual trees and to select the successive directions for splitting. In mathematical terms, the $j-th$ tree estimate takes the form:

$$M_n(x;\ \theta_i, D_n)$$

$$= \sum_{i \in D_n^*(\theta_j)} \frac{x_i \in A_n(x;\ \theta_i,\ D_n)Y_i}{N_n(x;\ \theta_j,\ D_n)} \ldots \ldots \ldots \ldots 2$$

Where $D_n^*(\theta_j)$ is the set of data points selected before the tree construction $A_n(x;\ \theta_i,\ D_n)$ is the cell containing x, and $N_n(x;\ \theta_j,\ D_n)$ is the number of preselected points that fall into $A_n(x;\ \theta_i,\ D_n)$. Having noted that the output of the random forest is from the decision of the majority tree. Hence, the trees are combined to form a finite forest as mathematically represented:

$$m_{M,n}(x;\ \theta_i \ldots \theta_m, D_n) = \frac{1}{M}\sum_{j=1}^{M} m_n(x;\ \theta_j, D_n) \ldots \ldots 3$$

Where M is the number of trees in the forest.

```
Algorithm 3: Random Forest
Inputs: X – input data t – number of trees S – subsampling size
Output: a set of t, $itree$
Initialize $Forest$
set the height limit l = ceiling ($log^2 s$)
for i = 1 to t do
    X = sample (X, S)
    $Forest = Forest \cup itree\ (X, o,\ l)$
end for
return $Forest$
```

Algorithm 3: Pseudocode for Random Forest

Decision Tree

According to Yang, (2019) a Decision Tree (DT) is a tree-based technique in which any path beginning from the root is described by a data separating sequence until a Boolean outcome at the leaf node is achieved. It is the hierarchical exemplification of knowledge relationships that contain nodes and connections. When relations are used to classify, nodes represent purposes (Priyanka & Kumar, 2020; Suresh et al., 2020).

The DT algorithm is part of the supervised learning algorithm family, and its main objective is to construct a training model that can be used to

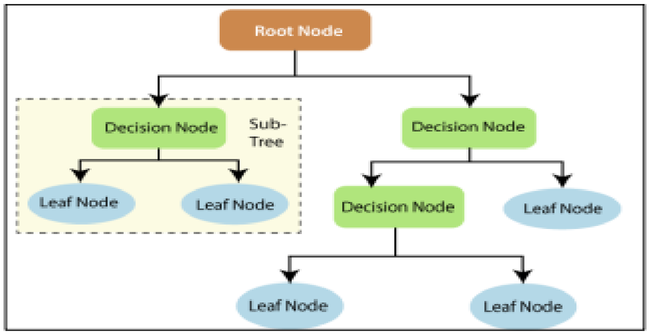

Figure 3: Decision Tree

Algorithm 4: Decision Tree

```
Input: GenDecTree(Sample S, Features F);
Output: root;
Steps:
    1. If stopping_condition(S, F) = true then
        a. Leaf = createNode()
        b. leafLabel = classify(s)
        c. return leaf
    2. root = createNode()
    3. root.test_condition =findBestSpilt(S,F)
    4. V = {v | v a possible outcomecf root.test_condition}
    5. For each value v Є V:
        a. $S_v$ = {s | root.test_condition(s) = v and s Є S };
        b. Child = TreeGrowth ($S_v, F$);
        c. Add child as descent of root and label the edge {root → child} as v
6. return root
```

Algorithm 4: Pseudocode for Decision Tree (Hambali et al., 2022)

predict the class or value of target variables through learning decision rules inferred from the regression and classification problems (Mittal et al., 2017; Priyanka & Kumar, 2020)(Mittal, Khanduja and Tewari, 2017; Priyanka and Kumar, 2020). In this study, DT algorithm is used to solve a classification problem. Figure 3 depicts image of DT while Algorithm 4 presents the pseudocode of it.

Performance Evaluation Metrics

The metrics employed to assess the performance of the adopted Random Forest and Support Vector Machine algorithm entail

Accuracy: is one of the most important metrics for the performance evaluation of machine classification models. It determines the fraction of prediction the utilized model got rightly. In more detail, it is measured as a percentage of the number of correctly predicted instances to the total number of instances present in the dataset. Thus, for binary classification of customer churns, the accuracy can be calculated in terms of positives and negatives as mathematically represented below:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \dots\dots\dots\dots 4$$

training data. The DT algorithm can be used to solv

Precision: measures the classifier's accuracy. It is the percentage of the number of correctly predicted positive instances divided by the total number of predicted positive instances:

$$\boldsymbol{precision} = \frac{TP}{TP + FP} \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots 5$$

Sensitivity or Recall: define the number of instances from the positive class that was correct in their predictions. It is a measure of the proportion of initial customer churn labels and was predicted to have the same customer churn label by the model.

$$\boldsymbol{Recall} = \frac{TP}{TP + FN} \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots 6$$

F-measure (or F-score): defines the harmonic mean of precision and recall. It combines recall and precision metrics to obtain a score.

$$\boldsymbol{F - Measure} = 2 \times \frac{Precision\ \times Recall}{Precision + Recall} \dots\dots\dots\dots\dots 7$$

Area under curve (AUC): AUC is a degree of how fit a parameter can distinguish between the binary classes classification (False/True). AUC varies between 0 and 1. The nearer the AUC value to 1 the better classification result.

Receiver Operating Curve (ROC): is a curve used generally to evaluate the performance of classification algorithm for two or more classification tasks. It is also applied to determine the best cut-off value to distinguish between customer churns whether False or Negative.

From the defined performance metrics, the description of the TP, TN, FP, and FN is as follows:

i. **True Positives (TP):** defines an instance where the actual class from a customer churn record is true and the model predicted it.

ii. **True Negatives (TN):** a scenario where a data record was false and hence predicted false by the model.

iii. **False Positives (FP):** an instance where the actual data record class is false but the model predicted true.

iv. **False Negatives (FN):** an instance where an actual data record point is true but the model predicted false.

4. Results and Discussion

This section presents the results of the experiments carried out. During implementation, each model was trained using dataset from the Telecom Churn that has 3333 instances and 21 features. Tenfold cross-validation mode was used to test and evaluate the algorithms. We carried out the experimental analysis with features selection using Chi-square to select relevant and most informative features based on correlation matrix.

In Figure 4 shows the distribution of dataset based on customer churn, 1 indicates the customers that are likely to leave the service provider that is, churned (14% of total instances which is about 483 instances), while 0 indicates the costumers that will stay (about 2850 instances). Figure 5 shows a correlation matrix of all the attributes in the dataset. The correlation matrix reveals the relationships between different attributes within the dataset. A correlation matrix is valuable in assessing how attributes are interrelated, which can help identify potential multicollinearity issues or indicate which attributes might have a strong influence on the target variable. In this context, it allows us to understand how important each attribute is in relation to the target class, which customer is likely to churn.

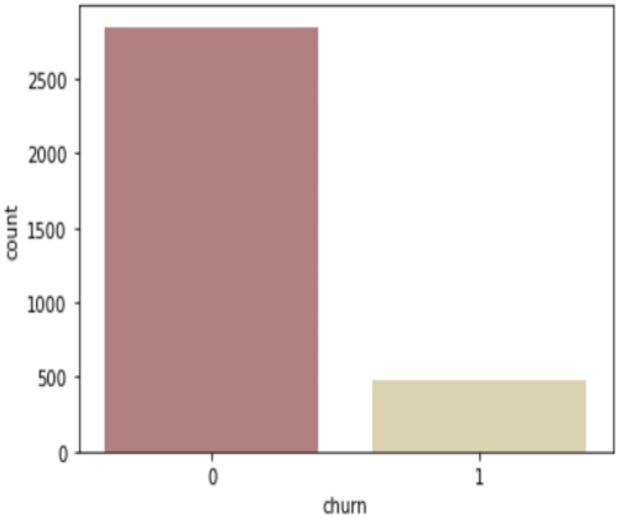

Figure 4: Data Distribution

This Chi-squared statistical test is commonly used for feature selection, especially with categorical data. It helps identify the attributes that have the strongest relationship with the target variable, in this case, churn. The next step in the process involved passing these selected features to classifier algorithms. This indicates that machine learning models were used to evaluate how well these 12 chosen features can predict which customers will churn. The outcome of this step is a set of predictive models that can assist in identifying potential churners.

To ensure the models' reliability, 70% of the dataset was allocated for training the models. This part of the dataset was used to teach the models how to predict churn based on the selected featuress. The remaining 30% of the dataset, known as the test dataset, was used to evaluate the models' performance. This division helps determine how well the models generalize to new, unseen data. In essence, it checks how accurately the developed models can predict churn in a real-world scenario.

4.1 Result Analysis

The result in terms of performance accuracy of the customers churn prediction using the three different algorithms such as Support Vector Machine, Decision Tree and Random Forest were presented in Table 2 - 3 and Figure 6 – 8.

Table 2: Training Result in Terms of Performance Accuracy for SVM, DT and RF

Algorithms	Training scores
Support Vector Machine (SVM)	0.8551
Decision Tree (DT)	0.9366
Random Forest (RF)	0.9768

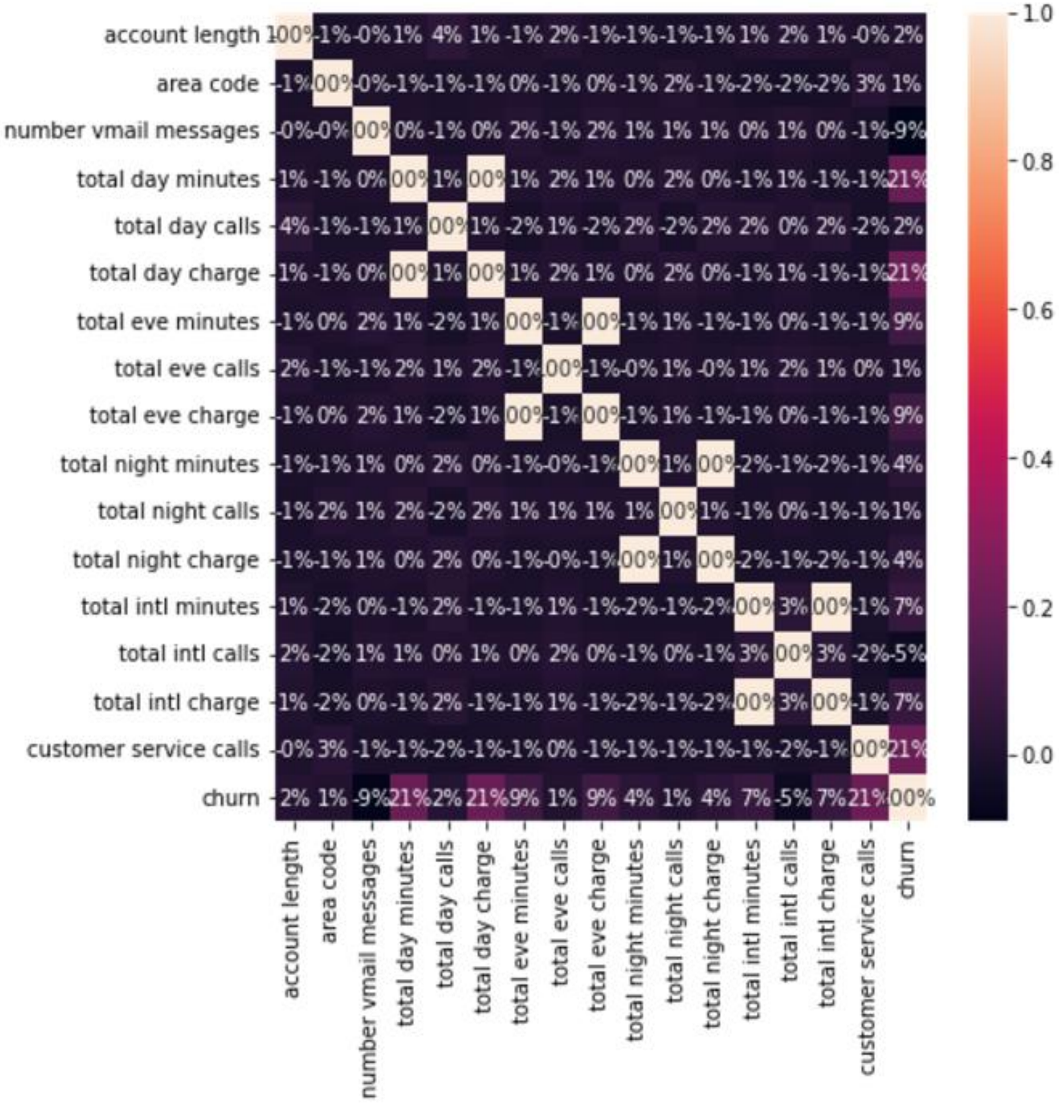

Figure 5: Correlation Matrix of Dataset Attributes

Table 2 shows training results of SVM with 85.51%, DT with 93.66% while RF has 97.68% accuracy. This indicates that RF has the best training accuracy follows by DT while SVM has the lowest performance. But all of them performed far above the average of 50% classification accuracy which is average benchmark for ML algorithm.

Table 3: Performance Test scores for SVM, DT and RF

Algorithm	Accuracy	Precision	Recall	F1-score	AUC
RF	0.94	0.94	0.94	0.94	0.83
DT	0.93	0.93	0.93	0.92	0.80
SVM	0.85	0.73	0.85	0.79	0.50

Table 3 shows that RF has highest accuracy of 94% with Recall, Precision and F1-score of 94% respectively and AUC of 83%. Follow by DT with 93% classification accuracy, Recall and Precision of 93% respectively, F1-Score of 92% and AUC of 80%. While SVM attained 85% classification accuracy, Precision of 73%, Recall of 85%, F1-score of 79% and AUC of 50%. Therefore, RF outperformed all other models in terms of classification accuracy, Precision, Recall, F1-score and AUC.

Figure 6 shows the ROC Area curve for RF classifiers with true positive rate closer to 1(that is above 0.8 and increase steadly). This shows that RF attained the best cut-off value to distinguish between customer churns whether False or Negative.

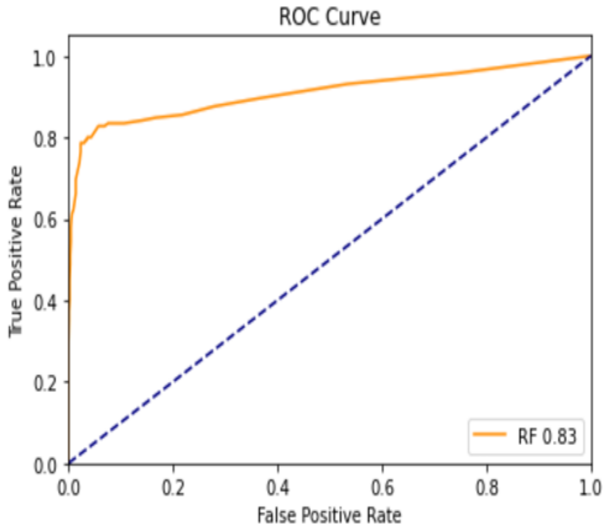

Figure 6: ROC for Random Forest Classifier

Figure 7 shows the ROC Area curve for DT classifiers with true positive rate far above the average (from above 0.6 and increase steadily till it reached 0.8). This indicates that DT showed the best cut-off value to distinguish between customer churns whether False or Negative.

Figure 8 shows the ROC Area curve for SVM classifiers with true positive rate increase from 0, with erratic curve. This indicates that SVM performed poorly to show the best cut-off value to distinguish between customer churns whether False or Negative.

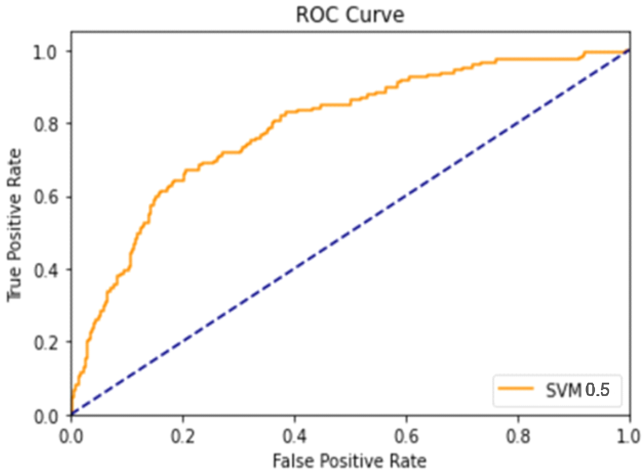

Figure 7: ROC for Decision Tree Classifier

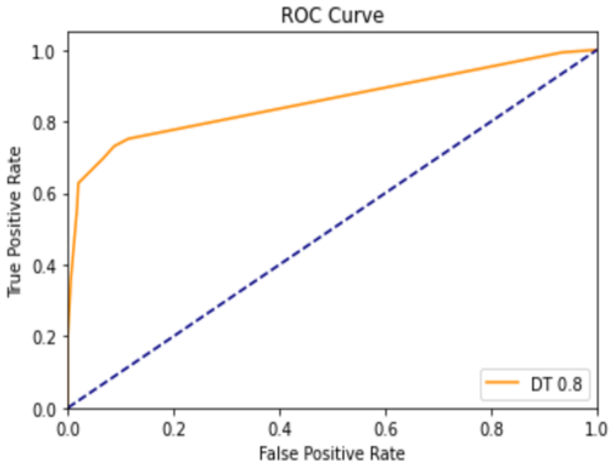

Figure 8: ROC for SVM Classifier

With all the performance metrics used for this experiment, it showed that RF outperformed other models employed for this study. Therefore, RF is performed better than both DT and SVM algorithms.

Comparison with Existing Studies

Comparison analysis was carried out with the existing models in the domain. The work of Saheed & Hambali, (2021) gave the best accuracy of 95.02%, the precision of 94.90% and F-score of 94.70% which is a little bit higher than the proposed model for this work. The proposed model attained an accuracy of 94% with the same value of precision and F-measure as shown in Table 5. Our comparison is based on those authors using the same dataset as ours, so as to maintain a level plain ground.

Table 5: Comparison with Previous Studies

Authors	Approach	Accuracy	Precision	F-measure
Sharma & Panigrah, (2011)	Neural Network (NN)	93.20	66.27	73.20
Shaaban et al., (2012)	SVM and NN	83.70	-	-
Jain, Khunteta, & Srivastava, (2020)	LogitBoost	85.24	-	-
Saheed & Hambali, (2021)	IG-R-RF	95.02	94.90	94.70
Fujo et al., (2022b)	Deep-BP-ANN	86.57	81.59	87.55
Proposed model	**Chi-RF**	**94.00**	**94.00**	**94.00**

Conclusion and Future Work

The aim of this study is to assist telecom industry develop a model that help them increase their profits. It is obvious that churn forecasting is an important task to improve the revenue for telecom firms. Therefore, this study tries to develop a forecasting model for telecommunications sector clients' attrition because there has been a growing emphasis on the development of a precise and efficient customer churn prediction model, which is seen as a significant research challenge for scholars and industry professionals alike. This study proposes that leveraging ML methods holds great potential for addressing the issue of customer churn management, allowing the creation of an early-warning system for this dynamic customer environment. The concluding summary of the model presented in this study indicates an impressive overall accuracy of 94% in predicting customer churn. Lastly, the study did not evaluate the characteristics of projected customer churns, although they may be significantly useful for organizations when making resolutions whether to let go of individual customers or to retain them. Thus, future research will focus on churn client characteristics because they may possess higher lifetime value to the organization.

References

Ahmad, A. K., Jafar, A., & Aljoumaa, K. (2019). Customer churn prediction in telecom using machine learning in big data platform. *Journal of Big Data*. https://doi.org/10.1186/s40537-019-0191-6

Alboukaey, N., Joukhadar, A., & Ghneim, N. (2020). Dynamic Behavior based Churn Prediction in Mobile Telecom. *Expert Systems With Applications*, 113779. https://doi.org/10.1016/j.eswa.2020.113779

Domingos, E., Ojeme, B., & Daramola, O. (2021). Experimental analysis of hyperparameters for deep learning-based churn prediction in the banking sector. *Computation*, *9*(3), 34.

Fujo, S. W., Subramanian, S., & Khder, M. A. (2022a). Customer churn prediction in telecommunication industry using deep learning. *Information Sciences Letters*, *11*(1), 24.

Fujo, S. W., Subramanian, S., & Khder, M. A.

(2022b). Deep Learning Customer Churn Prediction in Telecommunication Industry Using Deep Learning. *Information Sciences Letters Volume*, *11*(1), 185–198.

Hambali, M. A., Oladele, T. O., Adewole, K. S., Sangaiah, A. K., & Gao, W. (2022). Feature selection and computational optimization in high-dimensional microarray cancer datasets via InfoGain-modified bat algorithm. *Multimedia Tools and Applications*, *1213*, 1–45. https://doi.org/10.1007/s11042-022-13532-5

Jain, H., Khunteta, A., & Srivastava, S. (2020). Churn prediction in telecommunication using logistic regression and logit boost. *Procedia Computer Science*, *167*, 101–112.

Jain, H., Khunteta, A., Srivastava, S., Jain, H., Khunteta, A., & Srivastava, S. (2020). ScienceDirect ScienceDirect Churn Prediction in Telecommunication using Logistic Regression Churn Prediction in Telecommunication using Logistic Regression and Logit Boost and Logit Boost. *Procedia Computer Science*, *167*(2019), 101–112. https://doi.org/10.1016/j.procs.2020.03.187

Leung, H. C., & Chung, W. (2020). A Dynamic Classification Approach to Churn Prediction in Banking Industry. *Americas Conference on Information Systems*, 1–5.

Mirabdolbaghi, S. M. S., & Amiri, B. (2022). Model Optimization Analysis of Customer Churn Prediction Using Machine Learning Algorithms with Focus on Feature Reductions. *Discrete Dynamics in Nature and Society*, *2022*.

Mittal, K., Khanduja, D., & Tewari, P. C. (2017). An insight into 'decision tree analysis.' *World Wide Journal of Multidisciplinary Research and Development*, *3*(12), 111–115.

Priyanka, & Kumar, D. (2020). Decision tree classifier: a detailed survey. *International Journal of Information and Decision Sciences*, *12*(3), 246–269.

Rahman, M., & Kumar, V. (2020). Machine learning based customer churn prediction in banking. *2020 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA)*, 1196–1201.

Saheed, Y. K., & Hambali, M. A. (2021). Customer churn prediction in telecom sector with machine learning and information gain filter feature selection algorithms. *2021 International Conference on Data Analytics for Business and Industry (ICDABI)*, 208–213.

Sana, J. K., Abedin, M. Z., Rahman, M. S., & Id, M. S. R. (2022). A novel customer churn prediction model for the telecommunication industry using data transformation methods and feature selection. *PLoS ONE*, *17*(12), e0278095 1-21. https://doi.org/10.1371/journal.pone.0278095

Shaaban, E., Helmy, Y., Khedr, A., & Nasr, M. (2012). A proposed churn prediction model. *International Journal of Engineering Research and Applications*, *2*(4), 693–697.

Sharma, A., & Panigrah, P. K. (2011). A Neural Network based Approach for Predicting Customer Churn in Cellular Network Services. *International Journal of Computer Applications*, *27*(11), 26–31.

Sudharsan, R., & Ganesh, E. N. (2022). A Swish RNN based customer churn prediction for the telecom industry with a novel feature selection strategy. *Connection Science*, *34*(1), 1855–1876. https://doi.org/10.1080/09540091.2022.2083584

Suresh, A., Udendhran, R., & Balamurgan, M. (2020). Hybridized neural network and decision tree based classifier for prognostic decision making in breast cancers. *Soft Computing*, *24*(11), 7947–7953.

Tékouabou, S. C. K., Gherghina, Ștefan C., Toulni, H., Mata, P. N., & Martins, J. M. (2022). Towards explainable machine learning for bank churn prediction using data balancing and ensemble-based methods. *Mathematics*, *10*(14), 2379.

Yang, F.-J. (2019). An extended idea about decision trees. *2019 International Conference on Computational Science and Computational Intelligence (CSCI)*, 349–354.

Zhang, T., Moro, S., & Ramos, R. F. (2022). A Data-Driven Approach to Improve Customer Churn Prediction Based on Telecom Customer Segmentation. *Future Internet*, *14*(94), 1–19.