Addis Ababa University

Addis Ababa Institute of Technology

School of Electrical and Computer Engineering

Telecommunication Engineering Graduate Program

Classification of Top Call Reasons using Machine Learning in Call Center Service

A Thesis Submitted to the School of Electrical and Computer Engineering in Partial Fulfillment of the Requirements for the Degree of Master of Science in Telecommunication Engineering.

By: Atsede Abebe

Advisor: Dr. -Ing. Dereje Hailemariam

December 2021

Addis Ababa University
Addis Ababa Institute of Technology
School of Electrical and Computer Engineering

This is to certify that the thesis prepared by **Atsede Abebe**, entitled **Classification of Top Call Reasons using Machine Learning in Call Center Service** and submitted in partial fulfillment of the requirements for the degree of Master of Science in Telecommunication Engineering complies with the regulations of the university and meets the accepted standards with respect to originality and quality.

Signed by the Examining Committee:

Internal Examiner_____ Signature_____Date _____

External Examiner_____Signature_____Date _____

Advisor Dr.–Ing. Dereje Hailemariam_ Signature_____Date _____

_____
Dean, School of Electrical and Computer Engineering

Declaration

I, the undersigned, declare that the thesis consists of my own work in accordance with internationally accepted practices; I have fully recognized and referred all the materials used in this thesis.

Atsede Abebe
Name

Signature

Abstract

A call center (CC) connects customers to a service provider, such as telecom operators. The CC receives of customer requests, feedback, and complaints. These inputs from the customers provide an opportunity to comprehend the customer's needs, problems the customers face when using services, and the performance of the service provider. Meeting customer expectations by responding to complaints increases customer satisfaction, which translates to revenue maximization. Hence, the CC is critical to the success of the service provider.

Ethio-telecom, a telecom service provider in Ethiopia, operates a large CC that provides telecom-related services throughout the country. The CC accepts over two million calls per day via a service-free line. The center records and maintains a huge amount of customer-related data, which can further be analyzed using state-of-the-art machine learning algorithms for the purpose of proactively estimating call types and reasons.

This thesis proposes to map features from customer profile information into top call reasons so as to better understand customer call requests and map future calls to specific top call reasons. Data was extracted from Ethio-telecom's IP contact center, customer relational management, and customer billing system servers. To construct the classification models, J48, Random Forest (RF), and Naive Bayes (NB) algorithms are used. Accuracy, time to build a model, and model interpretation of each algorithm are used to compare their performance.

Results show that RF and J48 algorithms outperform NB, with scores of 97.46% and 97.4%, respectively. The NB model is the least accurate, with an accuracy of 83.6%.

However, the time spent building a model for NB is less compared to J48. During the model's interpretation, J48 algorithm is more interpretable than the NB and RF. J48 algorithm are best.

Keywords: Classification algorithm, Call Center, IP Contact Center, Customer Relational Management, Customer Billing System, J48, Naive Bayes, and Random Forest.

Acknowledgment

Foremost, I would like to thank God and St. Mary for the wisdom, strength and health in order to finalize my thesis work. Thank you, adviser Dr.-Ing. Dereje Hailemariam, for devoting significant time to guiding my work, supporting me, and responding quickly to any request. Another note of gratitude goes to my examiners, Dr. Ephrem and Dr. Surafel, for providing valuable feedback to make my work more valuable.

I would like to express my gratitude towards my family for the encouragement and being there for me on every journey of my life. I also thank all my friends for their support. Finally, I would like to thank Ethio-telecom, My Company, for giving me this glorious sponsorship in attending a master's program.

List of Figures

List of tables

List of Acronyms

AI	Artificial Intelligence
CBS	Customer Billing System
CC	Call Center
CDR	Call Detail Record
CRM	Customer Relation Management
CSD	Customer Service Division
DM	Data Mining
IG	Information Gain
IPCC	IP Contact Center
ISD	Information Service Division
IVR	Interactive Voice Response
KPI	Key Performance Identifier
MB	Mega Byte
NB	Naïve Bayes
RF	Random Forest
SIM	Subscriber Identity Module
SMS	Short Message Service
SVM	Support Vector Machine
VAS	Value Added Services
WEKA	Waikato Environment For Knowledge Analysis

Table of Contents

1. Introduction

A call center (CC) is a customer service in the form of an inbound call from the customer or an outbound call to the customer. Its aim is to help solve a customer's problem or to aid with the company's business process. A CC refers to help desks, information lines, and customer service centers as a whole. Customer assistance, operator services, inbound and outbound telemarketing, and web-based services are some of the services generally provided by these centers [1].

The CC can be broadly categorized into two types: Inbound and Outbound. Inbound CC is typically focused on providing support to clients that need to solve problems or follow instructions [2]. Instead of receiving calls from customers, agents in an outbound CC make calls to them [3]. The CC delivers service in two ways: direct voice calls and multi-channel service. A direct voice call is a service given in a telephone conversation. Multi-channel is a type of service that is delivered through Short Message Service (SMS), web chat, and various social media. So, CC service is critical in the telecom industry for responding to customer complaints, increasing customer satisfaction, and increasing revenue.

An IPCC (IP Contact Center) platform is used by the majority of telecommunications companies to accept incoming calls. The IPCC system includes an IVR (Interactive Voice Respond) platform and automatically routes calls to agents. When a customer calls a telecom CC, the system first connects the customer to IVR. The IVR acts as a go-between for the agent and the customer. The call handling process can be categorized into three steps. The first is the answering of the call by the IVR and the time spent in the queue if the line is busy. The second is the time that the agent spends handling the customer's request.

Finally, the closing time is anything the agent has to do to wrap up the call, such as registering a call reason and forwarding complaints to the respective offices. Call reasons are the reason that a customer reaches any service provider's CC. Call reasons are logged after each call by agents. Logging call reasons correctly could benefit service providers by unearthing user experience issues and taking reactive measures for incidents quickly. To achieve this, one would select top-call reasons. Top call reasons are the causes of customers' contacting a CC the most. Ethio-telecom generate a huge amount of call reason data that is collected each day.

The CC has been overloaded with customer complaints and requests regarding common occurrences. As a result, the department employs a large number of people. Despite the fact that CC employees are overburdened with customer requests, the service is difficult to access, resulting in a variety of call drops. Dropped calls have a significant impact on customer satisfaction levels. The CC has KPIs (Key Performance Indicators) such as answered calls, working time, average talk time, call drop, and so on. It is still struggling to meet the desired level of KPI. As a result, the customer is dissatisfied because they did not receive adequate service. Customers are repeatedly contacting the CC. However, Ethio-telecom CC has no knowledge of where the request originated. It only knows when they are confronted with it. But, there is a daily report of top calls from the past to address this issue, we examine the history of a customer's calls beginning with the request they make in the CC. Incoming call requests are classified based on a customer's historical top call reasons and other attributes.

The top call reason is determined based on previous call requests that can classify the re quest when the customer calls.

The Figure 1- 1 demonstrates the number of logged top call reasons on a daily basis. The chart shows the number of calls received for each top call reason. There are five top reasons displayed, ranging from the highest to the lowest.

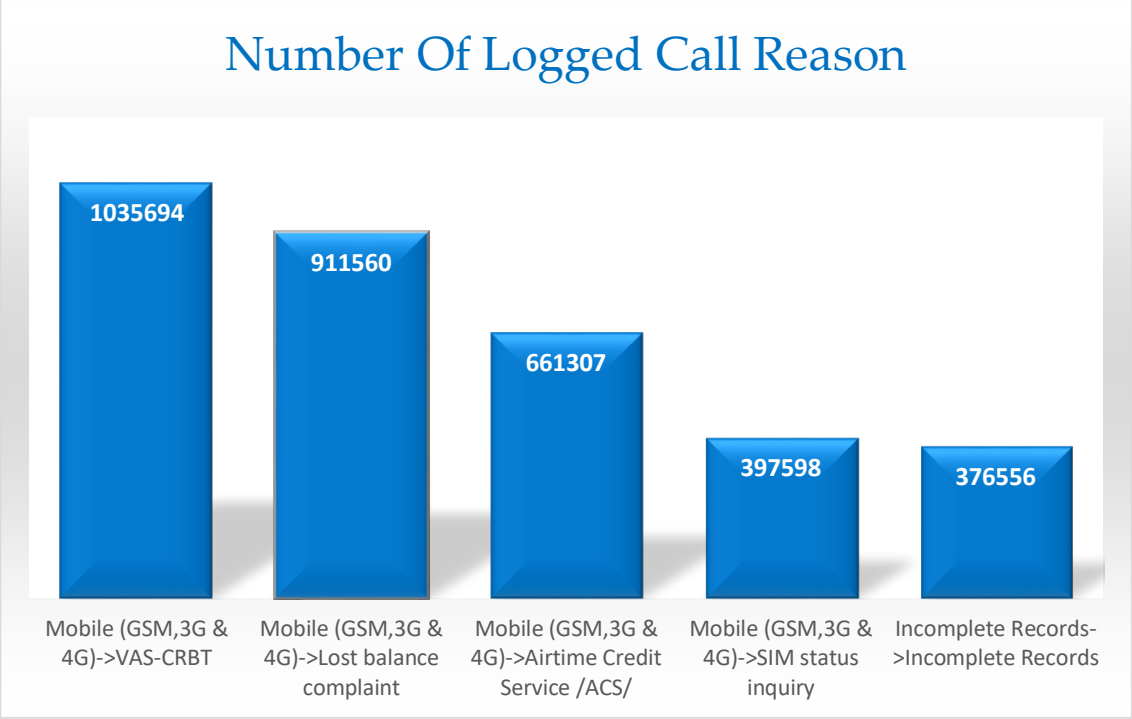

Figure 1- 1 Ethio-telecom a one day CC report

1.1 Statement of the Problem

Customers make a lot of requests in the CC service. The requests of these customers are recorded in the system, and a daily report is generated. On one day, more than one million call reasons were recorded, according to Ethio telecom CC report. It takes complaints, requests, and feedback from customers. Incoming calls to the CC are for wide range of reasons; from fault number registration to information request and many more. So, a CC serves as a conduit between a customer and a service provider. The agent who accepts the call and serves the customer according to the request then records the reason for the call. There is no special agent or remedy for customer requests that are already categorized based on the call reason. By its nature inbound calling is when the customer initiates the call. The agent accept the call, give feedback for request and register the call reason. After this process companies knows customer requests. This type of support tends to be reactive support. This results high amount of call drop. A call drop one of measure of performance of CC service.

The available data can be used to integrate features from customer profile information with history of top call reasons, so as to analyze and better understand the customers, call center and the company. State-of-the art machine learning supervised algorithms are used to analyze the huge data. In this thesis work classifying the incoming three selected top call reason using machine learning. Using a customer's previous call reason, usage history, and profile data as features. The goal of this study is to categorize the incoming top call cause. It facilitates the CC response to the request. The customer receives a quick response. The quality of CC service is improving. It also fixes the line's availability problem.

Research Questions

- Which feature or attribute are influences the incoming customer call request?
- Which customers' requests categorized on top call reason?
- Which classification algorithm is most suitable for forecast the incoming call?

1.2 Objective

1.2.1 General objective

The general objective of this thesis work is classification of top call reasons using machine learning in CC Service.

1.2.2 Specific Objectives

The specific objectives are:

- To associate top call requests is based on their characteristics.
- To determine the important features to classify top call causes.
- To set on the best methods for classifying incoming top call requests.
- To evaluate the algorithm's efficiency.
- To analyze incoming top calls in the CC.

1.3 Scope and Limitations

1.3.1 Scope

The scope of the thesis work is to provide for the Ethio-telecom CC based on prior call requests and some client characteristics; it categorized the incoming top call request. The information is gathered from IPCC's database of a customer's top call reasons in the past. Customer profile information and three consecutive months of consumption are collected from CRM and CBS servers, respectively. Three classification techniques are used to assess the data.

1.3.2 Limitations

It takes longer to retrieve data from a different department. Customer profiles, data from ISD (Information System Division), and data from the customer service division (CSD) top call cause. This thesis works only for classify call to the first three top call reason which is customer requests sim status requests, loose balance and internet package.

1.4 Contribution

Ethio-telecom has no way of knowing what requests are being made and can only respond to them after they have been made. The CC service efficiency suffers as a result of this. Forecasting forthcoming requests based on prior calls, usage, and profile data increases the chances of the customer getting a line and receiving the service they want quickly.

The company's getting average knowledge of a customer and implement based on result classify the incoming call. It also integrated respond for the customer request in IVR. This makes CC service efficient.

1.5 Literature Review

The study [12] investigates customer behavior by employing CRM and DM (Data Mining) methodologies to analyze user behavior. It creates client behavioral models by estimating attributes such as age, income, and lifestyle. The methodology used is rule-based DM, which extracts patterns from data and then uses those patterns for various purposes, such as prediction. The researchers who carry out the methodology's primary steps. The first step is to solicit client feedback. This step selects several inquiries raised by the customer. In the second phase, it groups clients depending on the features of their inquiries, history, and profile. Then it clusters based on similarity. In the third stage, rule induction with cluster data, it offered a model. The proposed model is utilized to increase sales, replay consumer inquiries, and develop marketing initiatives. The rule induction process also makes use of clustered data to generate new rules that may be used to better understand client needs and the organization's growth. The fourth step is customer knowledge, which entails applying a rule induction method to pattern data to better understand customer behavior, satisfaction, and loyalty. The article concludes that rule induction on customer clustered data is a critical factor enhancement for any organization's CRM improvement.

The primary goal of this paper [28] is to examine customer behavior in a telecommunications company's CC. It consists of four steps. It takes data from the corporation first, then the event logo. Case ids, which are defined in two columns, are the data attributes of the collected data.

It defines four activity categories in the activity attribute: type of action, request type of action, complaint type of action, and operational initiative type of action. In the paper, they evaluate customer behaviors using fuzzy mining and Disco and ProM5.2.

The paper [32] Predictive analytics is the use of data, statistical algorithms, and machine-learning (ML) techniques to identify likely future outcomes based on historical data. The effectiveness of predictive analytics is more about making better decisions. Previously, with a limited data volume, intuitive decision making was nevertheless successful. However, as the bulk of the data has grown to amazing dimensions, the human ability to make intuitive judgments has diminished. As a result, data-driven decision making is increasingly regarded as assuring a credible pathway to better decision making.

The goal of the study, as stated in [4], is to employ pattern analysis of customer actions to predict churn and for intelligent and targeted promotions. With the data acquired, the researcher begins the methodology. The quality of the data is the most difficult task in the data collection procedure. Customer type, recharging details, outgoing phone calls, incoming voice calls, and SMS sent are the data kinds used. Name, profession, gender, income demographic data, mobile number, voucher type, recharge kind, recharge amount, balance, accumulated call count, call duration, call type, amount, count all types, and rated amount are the main factors. The total quantity of recharges done each month has fallen, but the number of customers recharging has stayed constant, according to the publication.

In [5], the author analyzes customer behavior on a website based on click stream, which means how many times the customer clicks on the website. The goal is to comprehend the customer's requirements. The proposed solution in the paper is based on an artificial neural network. The model accepts input, output, and hidden layers.

The [6] study employs text categorization techniques to automatically analyze the use of internal service applications with the goal of reducing process delays. The system design has a six-step request. It collects service and then labels the document. WEKA's (Waikato Environment for Knowledge Analysis) experiment test employs Bayesian, SVM (Support Vector Machine**)**, and decision tree algorithms. The paper concludes that electronic systems will reduce delays in correctly assigning documents, resulting in significant enterprise-wide time savings.

In [7], the CC is one department in a telecom company that connects a customer to a service provider. To give effective service and to satisfy customers, the company must work on the CC's efficiency. To model, the call process uses different modeling techniques. By using those modeling techniques, it also knows about the economic impact and predicts service needs.

In [14], the study focuses on a model which uses ML techniques to find an intelligent route in CC. The author uses different solutions for routing inbound calls to improve customer satisfaction. One by using IVR, which is by interacting with the caller to get information and route the call to the appropriate queue. The second one is by using skills-based routing. In this case, an incoming call is assigned to the most appropriate agent. Another solution uses data analysis to optimize routing. The research question of this paper is "how can demographic and historical data be used in CC to find a better route to improve their customer satisfaction"? The result showed a predictive model for the outcome with reasonable performance.

1.6 Methodology

Literature Review

Relevant literature, such as books, journal articles, conference papers, and the internet, was reviewed.

Data collection

Ethio-telecom provided the data needed for the research. It is obtained from the ISD and CSD data centers. The top call reason from the CSD IPCC server whereas information about the service numbers and usage statistics are collected from CRM and CBS respectively.

Data preprocessing

Data collected from various departments can be aggregated. The missing values were removed, and the collected data was clean and ready for processing in the appropriate format. The average data usage for three consecutive months was then calculated.

Model building

WEKA open source tools were used in the experiment. For the training data, it employs three classification algorithms and cross-validation techniques.

Result analysis

After carrying out the experiment using different performance matrix, comparing the result of each algorithms.

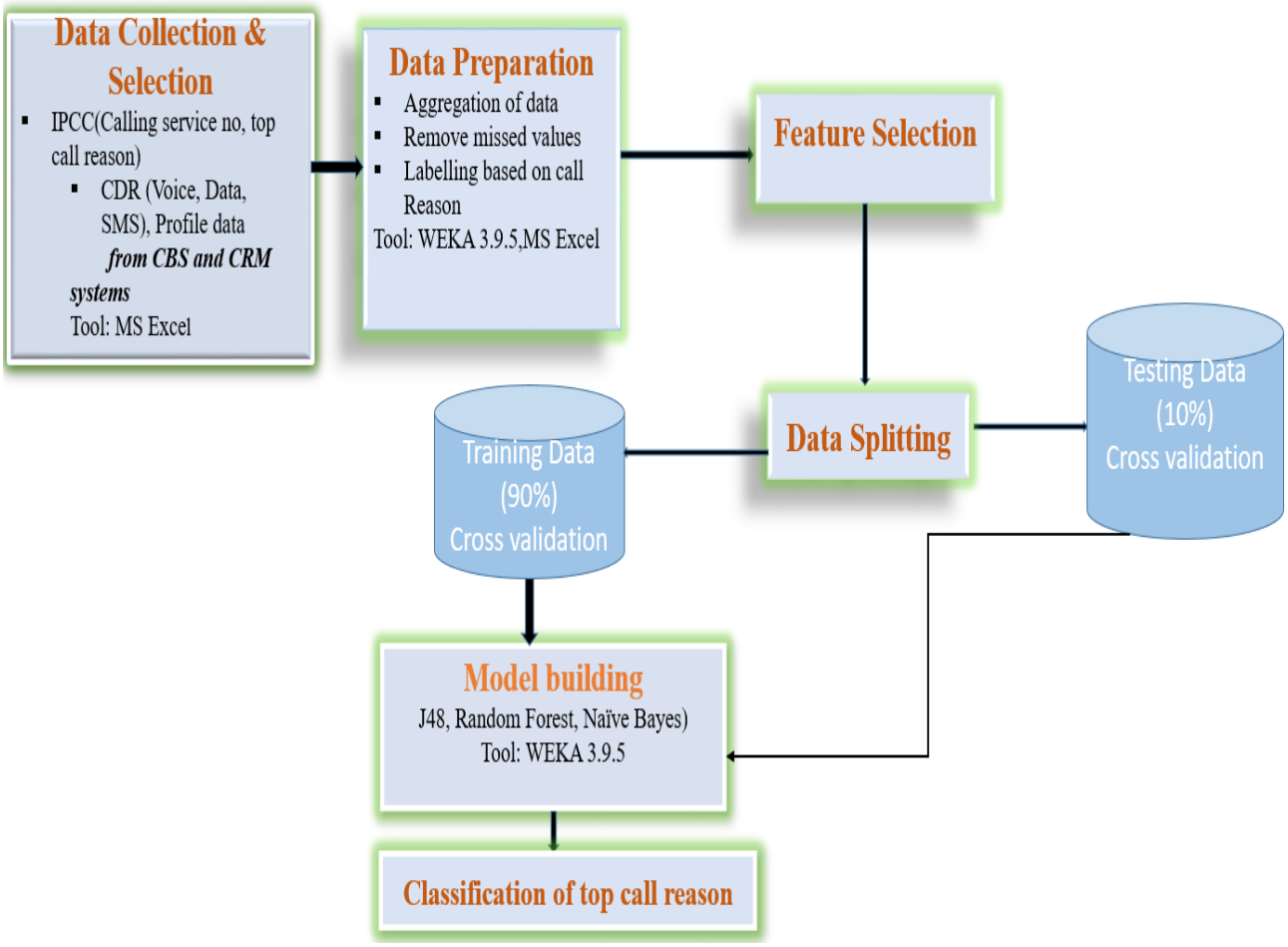

Figure 1- 2 Methodology process

The Figure 1- 2 show the methodology process.

1.7 Thesis Organization

The thesis is divided into six chapters: Chapter Two discusses CC services, Chapter Three discusses organized ML techniques, and Chapter Four discusses experimental analysis, which includes a detailed description of how data processing works in each step. Chapter 5: Discussion and Results based on the findings.

The findings demonstrate the research's findings. Chapter 6: Conclusion and Future work discusses the work's conclusion as well as some recommendations for future work.

2. Contact Center System

2.1 Overview

In [13] a CC also sometimes called a contact center or customer service center is a tunnel communication with customers (internal or external) through multiple channels email, phone, and live chat. The CC is the center for service provider to connect to customers. It is basic department in any service provider. It is a key to know about the customer need, the company performance and satisfaction of service. Service providers can use CC to get plenty of information such as service gaps, usage behavior of customers and overall performance out of a customer requests.

CC typically handle more than one type of call, with each distinct call type referred to as a queue [21].CC generally handles multiple types of calls, with each separate call type denoted by a queue. Inbound calls arrive at random inside each queue over time. Agents make outbound calls to customers in many centers, either proactively for telemarketing or collections activities, or as a follow-up to previous incoming calls.

Inbound calls can be directed to agents, groups, and locations using Automatic Call Distribution and Computer Telephony Interaction devices, with developments in routing technology allowing for more sophisticated logic to be supported over time. Individual agents can be skilled to handle one type of call, several types of calls, or all types of calls, with different priorities and preferences specified in the routing logic.

Each call can be sub-divided into three event groups. The IVR is the call distribution system that routes the call to an agent with the appropriate knowledgebase. The caller listens to pre-recorded messages instructing them to respond to key-presses corresponding to different departments of a CC.

Queue events are instances of the caller waiting in a queue to speak to an agent. Agent events are instances of the caller being put through to a live CC employee.
Thus, CC can be thought of as stochastic systems with multiple queues and multiple customer types.

2.2 Architecture CC service

In [18] IPCC manipulated the CC service. The Huawei IPCC solution offers different media access methods including phone, video, fax, email, online, and social media, and employs an intelligent routing platform to route all media in a unified manner and distribute calls to the most relevant agents IVR. As shown below on Figure 2- 1 IPCC solution, which is based on the browser/server architecture, includes complete management components such as a real-time monitoring system, an inspection system, and a report system. Management personnel can conveniently monitor the status of CC at any time.

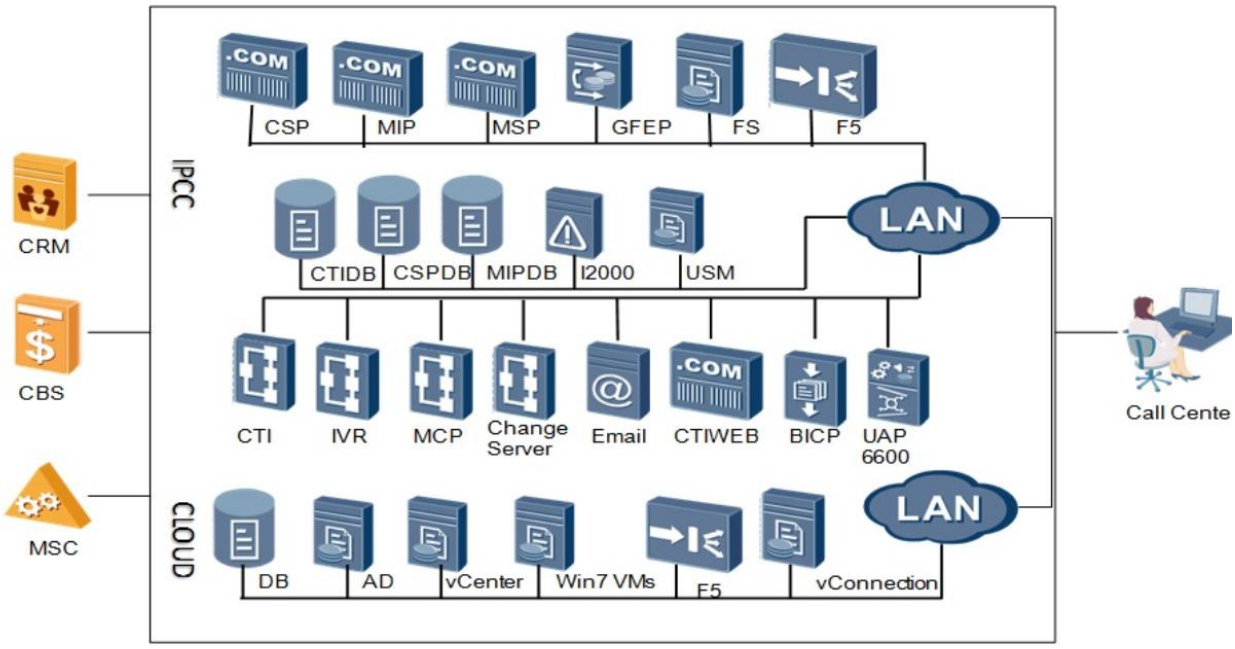

Figure 2- 1 Huawei IPCC Architecture [18]

2.3 CC service

In the CC service there is person who serve the customer is agent. The agents handle calls from customers with various complaints and inquiries about products and services, and escalate issues that cannot be resolved at the agent level to the appropriate bodies. The goal of the CC is to provide the best possible service to its customers at the lowest possible cost that is to minimize customer waiting time and maximize agent productivity [14].

3. Machine Learning

3.1 Overview

ML is a branch of AI (Artificial Intelligent) that deals with training a machine using data rather than developing a program. According to its definition, ML is a branch of computer science that arose from the study of pattern recognition and computational learning theory in AI [12]. ML is the scientific study of algorithms and statistical models that use patterns and inference to do a certain task without explicit direction [8].

In ML, a computer program is assigned to perform some tasks, and it is claimed that the machine has learned from its experience if its measured performance on these tasks increases as it obtains more and more experience in executing these jobs [16]. ML is used to train machines how to handle data more efficiently. Sometimes unable to analyze or extract information from data after viewing it. In that instance, it employ ML. With a plethora of datasets available, the demand for ML is increasing. Many industries use ML to extract important data. The goal of ML is to learn from data [13].

3.2 Type of ML

ML comes in four varieties. There are four types of ML:

- Supervised ML
- Unsupervised ML
- Reinforcement ML
- Semi-supervised ML

Learning type	Model building	Example
Supervised	Algorithm or models learn from labeled data (task – driven approach)	Classification, Regression
Unsupervised	Algorithm or models learn from unlabeled data (Data –driven approach)	Clustering, Association, Dimensionality reduction
Semi-supervised	Model are built using combined data(labeled unlabeled)	Classification ,Clustering
Reinforcement	Models are based on reward or penalty(environment – driven approach)	Classification, Control

Table 3- 1 Types of ML techniques with examples [11]

The Table 3- 1 shows type of ML with model building strategies and examples.

3.3 Supervised ML

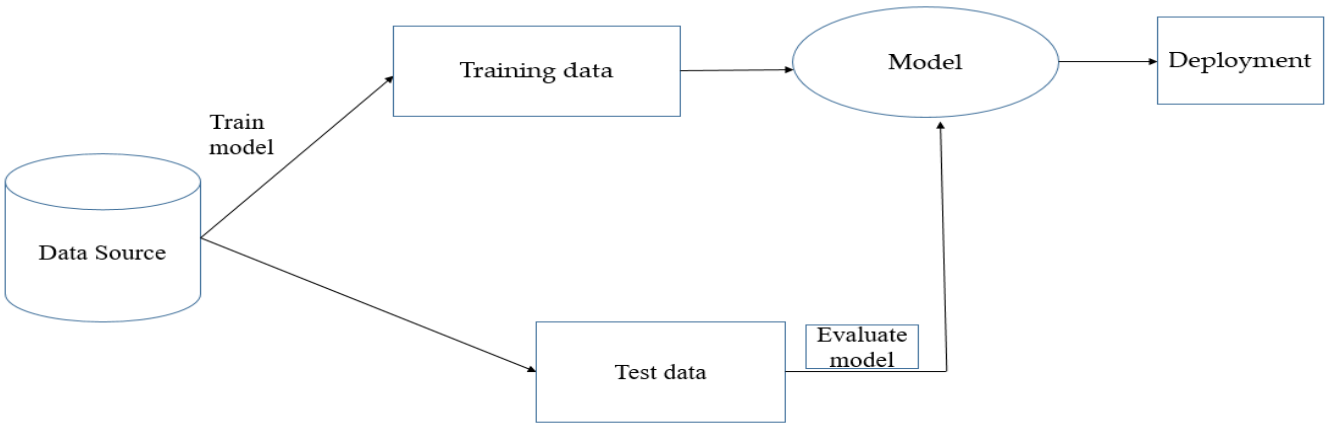

Figure 3- 1 supervised learning work flow [7]

When given new data, the computer is trained on a pre-defined set of training data, which improves its capacity to draw an accurate conclusion [9]. There are various supervised ML methods available. The classification algorithm is the most common. The Figure 3- 1 show the process of model building in supervised ML.

3.3.1 Naive Bayes

It is a classification technique that is based on the Bayes theorem and assumes predictor independence. In simpler terms, a NB classifier assumes that the presence of one feature in a class has no bearing on the presence of any other feature. The text classification industry is where NB shines. It is primarily used for clustering and classification, based on the conditional probability of occurrence [7].

3.3.2 Random Forest

A RF is widely used as a powerful new approach to data exploration, data analysis, and predictive modeling. In [33] RF is a collection of CART-like trees for growing, combining, testing, and post-processing. One is an ensemble of trees in which each tree grows while training on a sample obtained from the raining set via bagging without replacement. This is a known technique from ensemble learning methodology where generalization error is decreased due to combining decisions (or so-called votes) of multiple learners which are usually weak and unstable individually. The second approach is random split selection for a decision tree. This split is chosen randomly from a subset of best splits. Thus, these two ideas led finally to the basis for algorithm. It generally applies two mechanisms: building an ensemble of trees via bagging with replacement (bootstrap) and a random selection of features at each tree node. The first one means that any example selected from the training set can be selected again. Each tree is grown using the obtained bootstrap sample. The second mechanism performs random selecting a small fraction of features and further splitting using the best feature from this set.

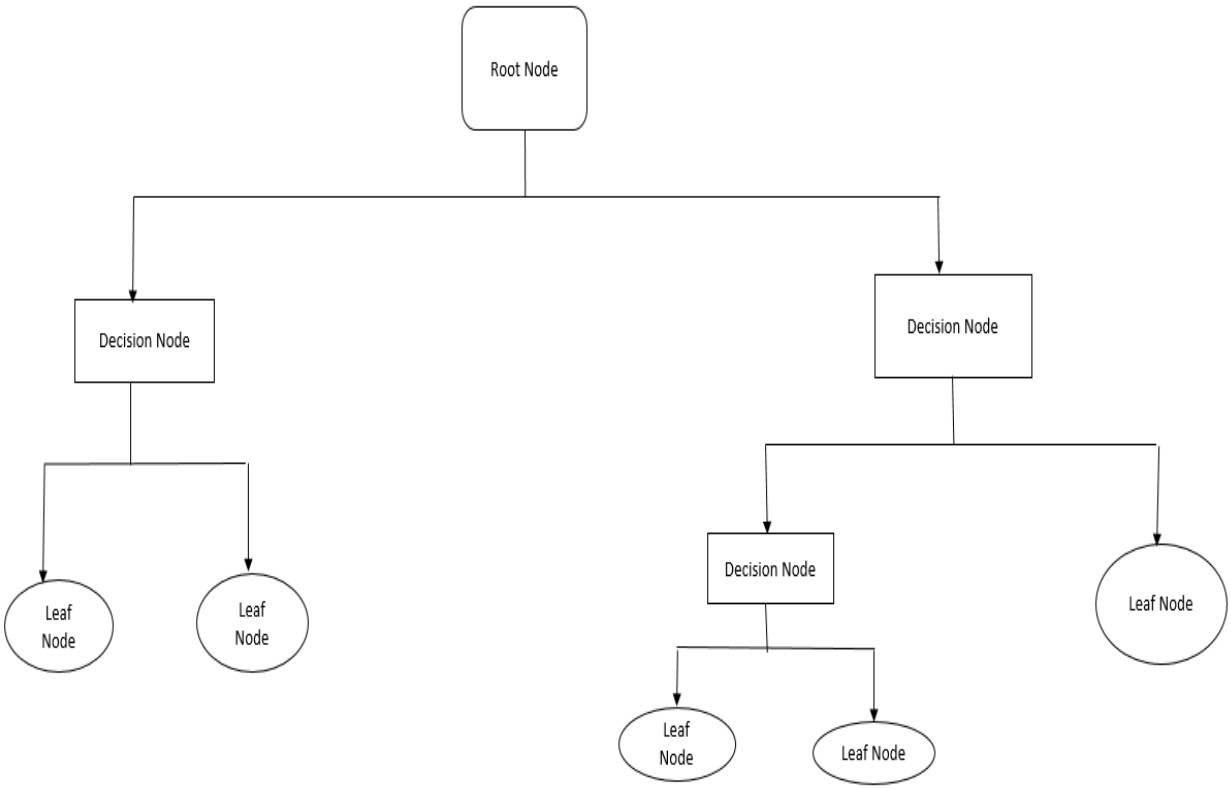

Figure 3- 2 Random forest [11]

The figure 3-2 show steps to of tree creation process of RF.

3.3.3 J48 Decision Tree

This algorithm is an extension of ID3 algorithm and possibly creates a small tree. It uses a divide and conquers approach to growing decision trees that was leaded by Hunt and his co-workers [21]. The J48 algorithm is popular ML based upon C4.5 algorithm [15]. J48 classifier is the classification algorithm used for detecting the novel and multi novel class.

For the problem to the classification the methodology of decision tree is used. For modelling the classification process tree is build.

While the tree is generated it is connected with each column of the database and results in classification for that column [11].

Instances are classified by checking the attribute defined by that node, starting at the root node of the tree, and then moving down the tree branch corresponding to the attribute value. For splitting, the most popular criteria are "Gini" for the Gini impurity and entropy [12].

J48 parameters	Descriptions
Binarysplits	Whether to use binary splits on nominal attributes when building the trees
ConfidenceFactor	Debug if set to true, classifier may output additional info to the console.
MinNumObj	The minimum number of instances per leaf.
NumFolds	Determines the amount of data used for reduced error pruning. One fold is used for pruning, the rest for growing the tree.
Seed	The seed used for randomizing the data when reduced error pruning is used.
SubtreeRaising	Whether to consider the subtree raising operation when pruning.
Unpruned	Whether pruning is performed.
UseLaplace	Whether counts at the leaves are smoothed based on Laplace.

Table 3- 2 Description of J48 algorithms to eliminate overfitting [20]

The table 3-2 show list parameters of J48 algorithm and its descriptions.

3.5 Unsupervised ML

Unsupervised ML is a type of ML not need of any supervision. This is widely used for extracting generative features, identifying meaningful trends and structures, groupings in results, and exploratory purposes.

4. Data Analysis

4.1 Data Collection

The data for this study came from Ethio-telecom, specifically CSD and ISD. First take service numbers which call to CC ask requests rank on top call reason. Profile information, including status and activation year, is obtained from the ISD of CRM servers based on the selected service numbers. Data from CBS servers' ISD is also collected, including voice usage, voice fee, data usage, data fee, and SMS fee. The top three call causes for this study were SIM (Subscriber Identity Module) status, balance loss, and data package. This category has 7,000 service numbers.

4.2 Data Preprocessing

4.2.1 Aggregation of data

Data preprocessing is the simple process of transforming raw data into an understandable format. Real world data is sometimes incomplete, inconsistent, redundant and noisy. Data preprocessing involves various steps that help to convert raw data into a processed and sensible format [20].

The steps involved in data preprocessing. The data used to extract the features is generated from different servers such as IPCC, CRM, and CBS. The extracted data is saved as a CSV file format for further analysis and aggregation.

Customer's service numbers are collected from the IPCC server for those service numbers, three consecutive months of usage statistics are mainly for voice, SMS, and data generated from CBS. The CRM server holds customer information such as activation year and status. The input for the experiments summarized the data for each service number. It sums up the status of the service number, the activation year of the service number, average voice usage for the service number, average voice fee for the service number, average SMS fee for the service number, average data usage in MB, and average data fee in birr for each service number.

4.2.2 Data Clearance

WEKA and Excel were used to collect the data, which consisted of unambiguous missing values. They began by gathering information from the customer calling number, which has the highest redundancy. Remove the unnecessary numbers with WEKA and Excel. Then take profile information on those cleansed numbers. For the past three months, some service numbers have had no activity. The service numbers have been removed.

4.2.3 Attribute selection

Attribute selection (Feature selection) is one of the basic tasks in ML. After aggregation and cleaning, select the best attribute. Attribute selection methods are used to reduce the dimensionality of the data through removing redundant and irrelevant attributes in a data set [23]. Feature selection is used to reduce the dimensionality, remove irrelevant and redundant data [19]. Attribute selection is a process in which a subset of M attributes out of N is chosen, complying with the constraint $M \leq N$, in such a way that characteristic space is reduced according to some criterion. Attribute selection guarantees that data getting to the mining phase is of good quality [22].

There are different attribute selection methods.

- **Entropy** is a measure of disorder of data. Entropy is measured in bits, nets or bans. This is also called measurement of uncertainty in any random variable [21].
- **IG**(Information Gain) is a feature selection measure introduced by J. R. Quinlan and used to select the test attribute at each node of the decision tree in the basic algorithm for building a decision tree (ID3) [23]. Let node N represent or hold the records of partition D from the dataset. Calculating IG helps in choosing the splitting attribute for node N, where the attribute of N having maximum IG is the one chosen for splitting. This attribute with higher IG minimizes the information needed to classify the objects in the resulting partitions and reflects the least randomness or "impurity" in these partitions. Such an approach optimizes the performance of the decision tree algorithm and guarantees that a simple tree is found [23].

The data was collected from different servers. The first customer dialed service number for the reason listed on the top call. After collecting and clearing the data, they calculate the usage.

Feature Name	Calculation	Data Type	Description
Activation Year	Year of activation Date	Numerical	Subscription year related to service request behavior.
Status	SIM Status	Nominal	Status of service number that the customer calls.
AVG _SMS_FEE	$= \frac{\text{Total SMS fee}}{\text{Number of month}}$	Numerical	To extract information about the usage range of the SMS service.
AVG_Data_ usage_MB	$= \frac{\text{Total data usage}}{Number\ of\ month}$	Numerical	To understand data usage behavior of customer.
AVG _Data _FEE ETB	$= \frac{\text{Total data fee}}{\text{Number of months}}$	Numerical	To understand Data fee behavior of customer.
AVG _voice _Usage	$= \frac{\text{Total voice usage}}{\text{Number of months}}$	Numerical	To extract information about the usage range of the voice service.
AVG_VOICE_FEE ETB	$= \frac{\text{Total voice fee}}{\text{Number of months}}$	Numerical	To understand voice fee behavior of customer.

Table 4- 1 Feature Extracted from CRM and CBS

4.2.3 Data Labeling

All data is a label based on the call request that asks and registers the call reason with the CC services. It also includes all features for each service number.

Labeling helps to understand the history of the customer calls in the CC. So ML learns based on a history of calls and features. It helps to know the pattern of data and to classify the coming calls based on history and selected features.

4.3 Experiment

For each top call request, a data collection of seven thousand data sets was picked for the experiment. It requires a dataset equal to the top three call reasons. There are three classes, which correspond to the three most common call causes, so it is multi-classification. Three classification algorithms were utilized in the experiment. These are J48, RF, and NB. There are various test options in each classification algorithm. Use the training set, given test set, cross-validation, and percentage. Use the cross-validation test option for the testing option.

There are different for testing option that described below [26]

- Use training set: It was trained on how well it predicts the class of the instance to evaluate the classifier.
- Supplied test set: Choose the file from the dialog box to test set, that the classifier is evaluated on how well it predicts the class of a set of instances loaded from a file.
- Cross-validation: The classifier is evaluated by cross-validation, using the number of folds that are entered in the Folds text field.

- Percentage split: Certain percentage of the data to classifier is evaluated on how well it predicts which is held out for testing. The amount of data thought out depends on the value entered in the percentage % textbox.

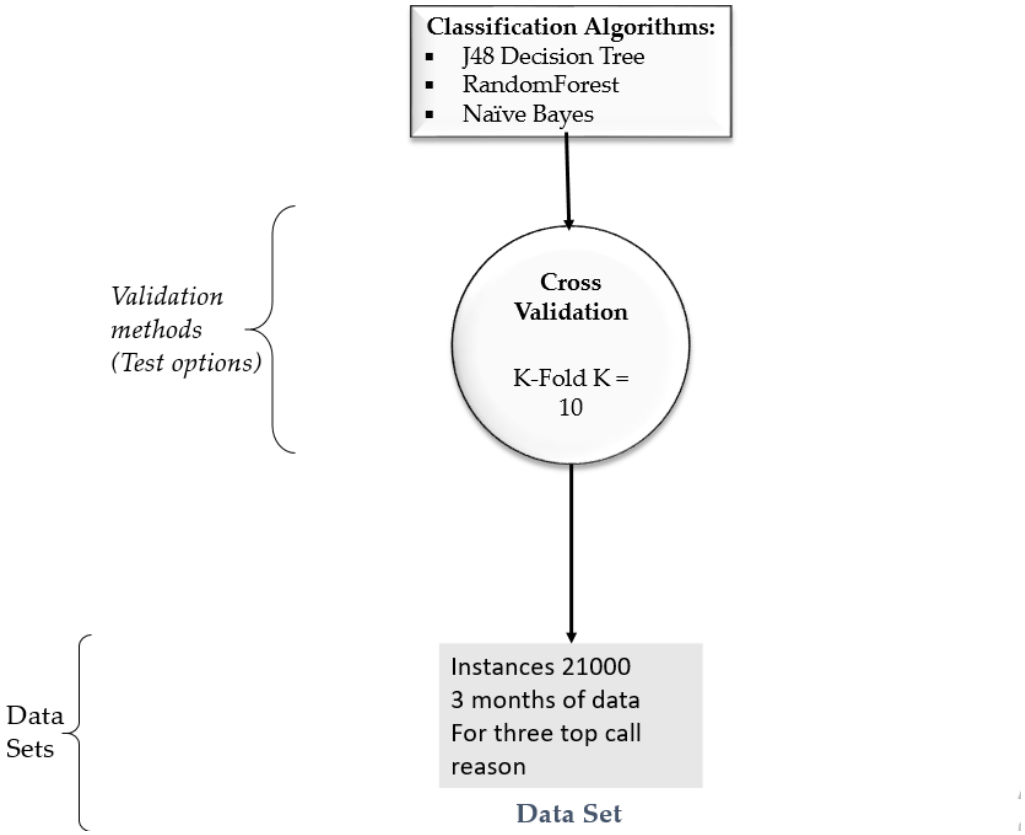

Figure 4- 1 Experiment process

4.4 Performance evaluation Matrices

There are various matrices for evaluating performance. To assess the accuracy of the classification algorithm.

4.4.1 Confusion Matrices

The n x n confusion matrix connected with a classifier displays the predicted and actual classification, where n is the number of possible classes [24]. The number of accurately classified classes can be used to assess the validity of a classification. True positives (TP) are those that are appropriately identified as belonging to the class.

True negatives (TN) are those that were either wrongly classified as false positives or were not identified as class examples of false negatives [25]. The matrix compares the actual target values to the ML model's predictions. The rows represent the target variable's expected values [33].

Accuracy

Accuracy refers to how close a measured value is to the real (true) value. Accuracy is the percentage of accurately classified instances [29].

$$\text{Accuracy}= \frac{TP+TN}{TP+FP+TN+FN} \quad (4.1)$$

Where

TP =True Positive

TN=True Negative

FP=False Positive

FN=False Negative

Precision

Precision is a value of the accuracy provided by a unique class that was predicted [29].

$$\text{Precision}=\frac{TP}{TP+FP} \quad (4.2)$$

Recall

Recall is a measure of the ability of a prediction model to select instances of a certain class from a data set. It is also called sensitivity, and points to the TP rate [29].

$$\text{Recall}=\frac{TP}{TP+FN} \quad (4.3)$$

F1-score

F-measure (or F-score or F1-score) has been introduced to balance between sensitivity and specificity. It is defined as the harmonic mean of the two scores, multiplied by 2 to obtain a score of 1 when both sensitivity and specificity equal 1 [30].

$$F = 2.\frac{1}{\frac{1}{Sensitivity}+\frac{1}{Specificity}} = 2.\frac{Specificity.Sensitivity}{Specificity+Sensitivity} \quad (4.4)$$

Kappa

The Kappa statistic (or value) is a metric that compares an observed accuracy with an expected accuracy (random chance) [33].

$$\frac{FP}{TP+FP+TN+FN}+FN\frac{FP+FN}{TP+FP+TN+FN} \quad (4.5)$$

4.4.2 Error Measurement

Metrics	Formula	Evaluation Focus
Error Rate	$\frac{\mathrm{FP}+\mathrm{FN}}{\mathrm{TP}+\mathrm{FP}+\mathrm{TN}+\mathrm{FN}}$	Misclassification error measure the ratio incorrectly predictions over the total number of instances evaluated.
Sensitivity	$\frac{TP}{TP+FN}$	This Metrix used to measure the fraction of positive pattern that are correctly classified.
Specificity	$\frac{TN}{TN+FP}$	This Metrix used to measure the fraction of negative pattern that are correctly classified.

Table 4- 2 Classification algorithms performance matrix [26]

The Table above shows three error measurement matrixes, its calculation and descriptions

5. Result and Discussion

The analysis section makes use of all of the data sets that were chosen based on the characteristics described. Use 7,000 service numbers for each top call request. It was a request based on the most important reason. The chosen data set employs three classification algorithms: J48, RF, and NB. There are various data set combinations available. There are three classes in the analysis work. It has a multi-classification. It employs an equal data set. Multi-classification makes the classification model more accurate based on the training data. For the experiment, choose the cross-validation test option. Then, using a different performance measurement matrix, to compare the results of each algorithm.

Accuracy

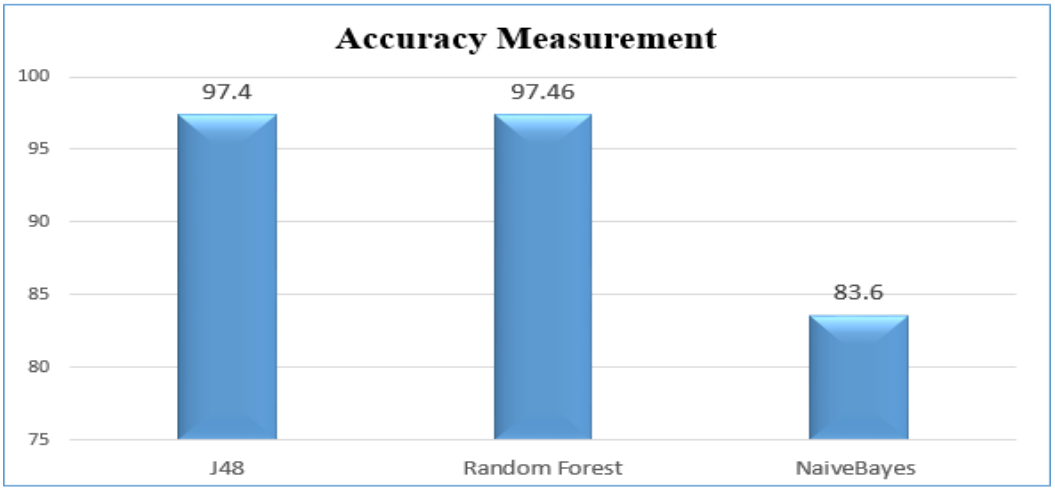

Figure 5- 1 Accuracy measurement comparisons

The accuracy comparison of each algorithm is one of the measures for the performance of the model. Therefore, for the experiment use, three supervised classification algorithms. Which are RF, J48 and NB employed and the result of each compare for accuracy. The Figure 5- 2 show that RF, J48 and NB produces accuracy score of 97.46%, 97.4% and 83.6% respectively.

The RF algorithm and J48 outperforms NB algorithms in terms of accuracy. The NB has the least accuracy of model building for the analysis.

Time taken to build model

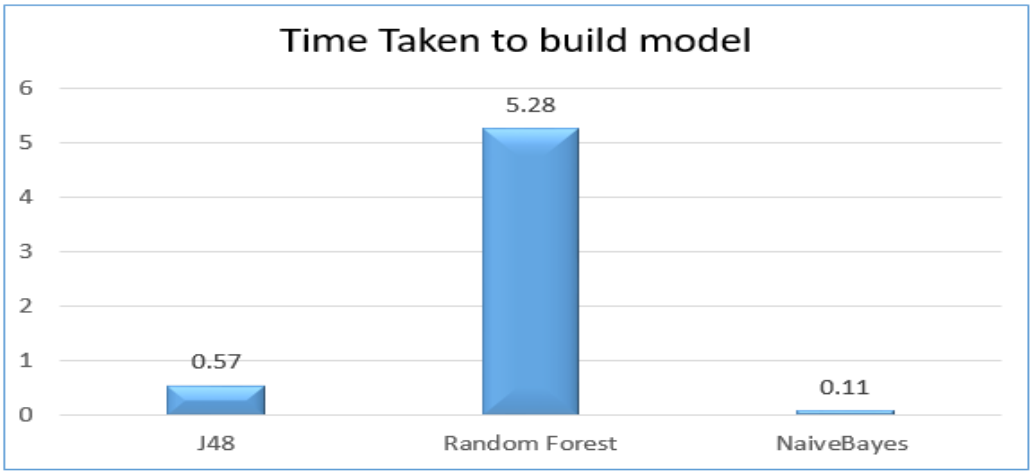

Figure 5- 3 Time taken to build model

One performance comparison matrix is the time it takes to build the model. It compares the time it takes to complete the model. Table 5-3 shows NB model took 0.11 seconds to develop, whereas the J48 and RF models took 0.57 and 5.25 seconds, respectively.
In this aspect, the NB method outperforms J48 and RF. In comparison to the J48 and NB algorithms, the RF approach takes more time to develop a model.

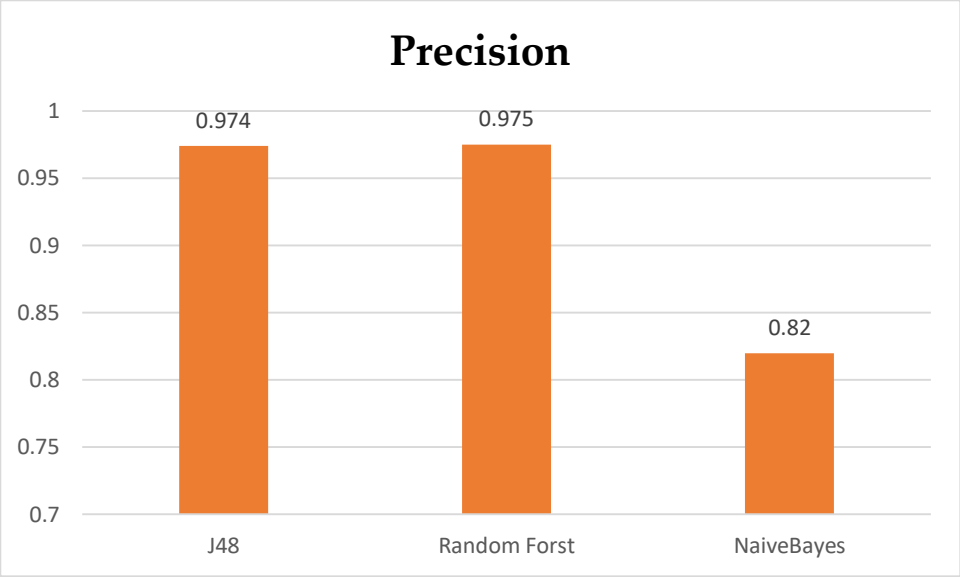

Figure 5- 4 Precision measurement

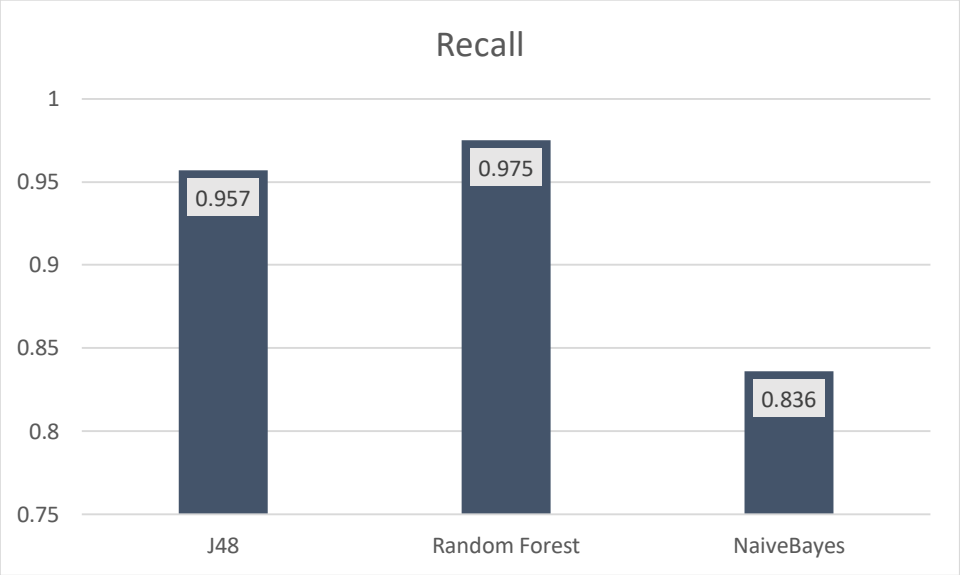

Figure 5- 5 Recall Measurement Comparison

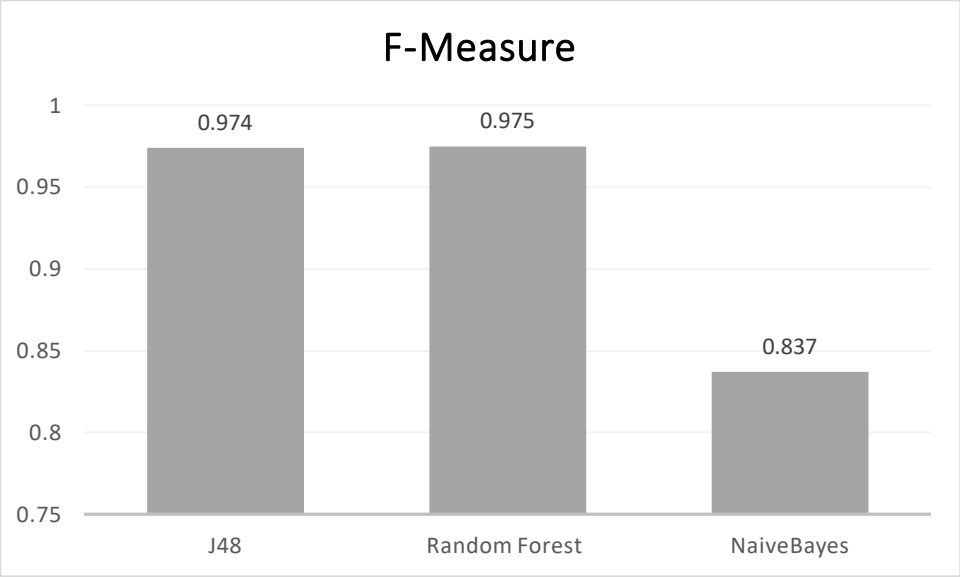

Figure 5- 6 F-measure Comparison

The figure 5.3, figure 5-4 and Figure 5.5 show the precision, recall and F measure result of each algorithm.

Errors

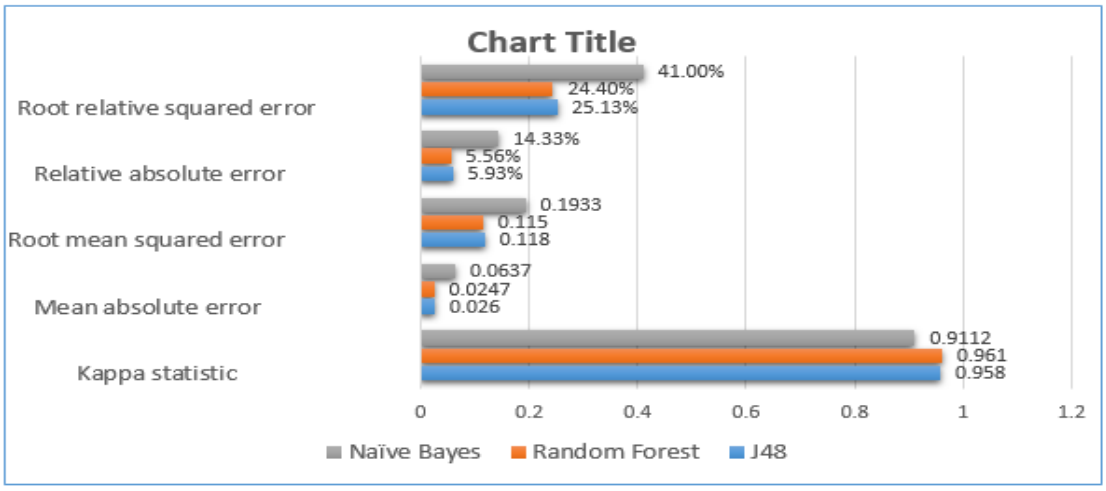

Figure 5- 7 Comparison Error Occurrence

As shown in the above J48 algorithm generates less errors than the NB algorithm. The RF algorithm has a low error rate. The NB model has a high rate of error. In all error metrics, both the J48 and RF methods exhibit fewer error occurrences. So, in order to address the third research question, which classification algorithm is most suitable for forecast the incoming call? J48 is the best algorithm since it has fewer error occurrences and easily interpretable and less computational time than other two algorithm. For the problem, the J48 algorithm is superior to RF and NB.

Model building and interpretation

J48 model building and interpretation

Model building is one parameter to compare the outcomes with is the model building and interpretation. The J48 method is easier to understand than both the RF and NB algorithms.

Attributes	IG	Ranked by IG
Service no Status	0.105	5
Activation Year	0.737	1
Average_SMS_Fee	0.266	4
Average_Data_Usage_MB	0.7198	2
Average_Data_Fee ETB	0.460	3
Average_Voice_Usage	0.023	7
Averge_Voice_Fee ETB	0.096	6

Table 5- 1 Attribute rank by IG

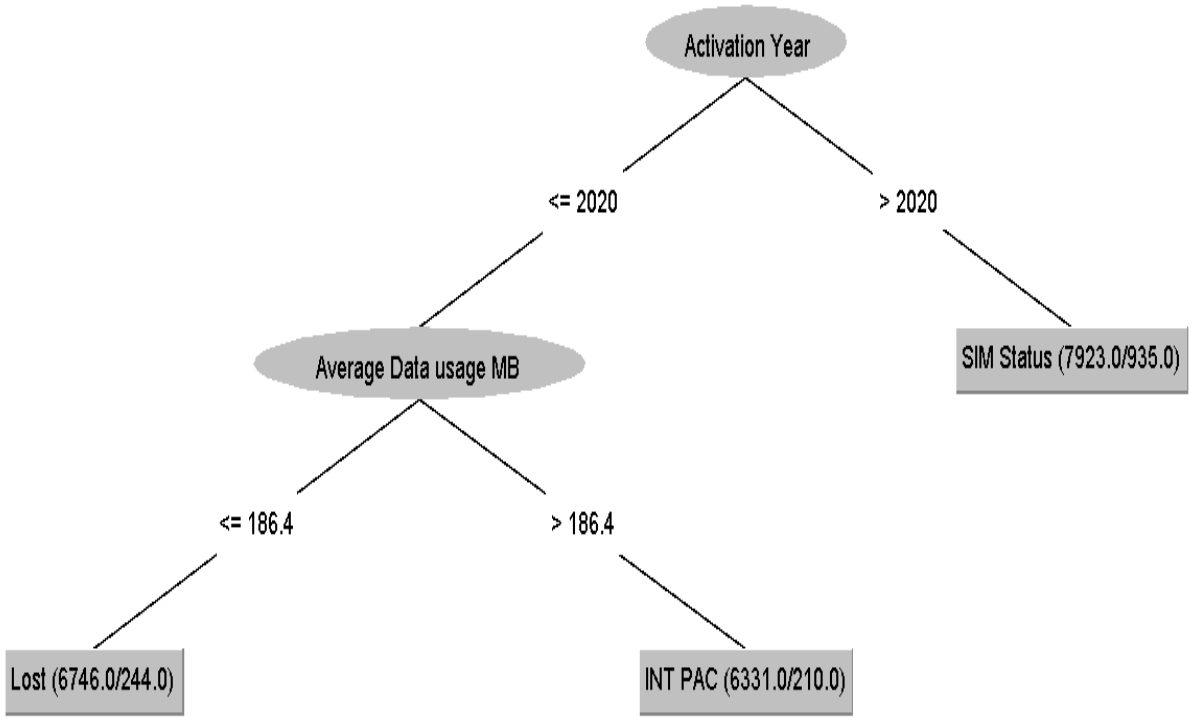

Figure 5- 8 J48 decision tree

As seen in the Table 5- 2, attribute ranking employs using IG. The activation year is rated first, followed by the average data usage in MB and the average data fee. In Figure 5.7 above show the decision tree outcome, three attributes are also displayed. The activation year is represented by the root node of the tree.

Employ pruning to gain a better understanding of the tree for interpretation. Pruning strategies, in general, seek to simplify decision trees that are over fit to the data [27].The tree had 42 leaves before it was pruned after it was pruned, it had 8 leaves.

Overfitting and interpretation are made more difficult as a result of this. The number of leaves on the tree decrees to three after pruning, and its size decrease to five. Despite its ease of interpretation, the model's accuracy has dropped following the trimming procedure. Before pruning, the model had 97.4% accuracy. Following pruning, the model's accuracy improved to 93.37%.

The first research question is which feature or attribute influences the incoming customer call request? There are seven features that are selected for the analysis part. Each feature has its own value for the target class. The activation year has a high influence on the incoming customer requests. The customer subscription year is highly related to customer requests. As shown in the Table 5.1, the three most IG features. Which is the activation year, average data usage MB, and average data fee. Also in the J48 decision tree show above on Figure 5.7, these three features decide on the incoming call request categorization. As seen on the tree activation year, the root node it is classified based on its activation. The customer's average usage of data is also the second feature to decide the customer's reason for calling the CC. The third feature is that the average data fee also affects the incoming call classification.

The second research question is: which customer’s requests are categorized as the top call reasons? As shown on J48 decision tree a customer calls to CC to request a balance loss if the calling number's activation year is before 2020, specifically 2020, and the customer's average data use is less than 186.4 MB. Whereas average data usage of more than 186.4MB needs the purchase of an internet plan, the activation year subscription after 2020 contacts the CC for a SIM status request. As a result, those clients' service numbers are not renewed on time, and they are no longer allowed to call.

RF model building and interpretation

Features	Importance of Features	Number of Node
Activation Year	39%	1192
Average Data usage MB	36%	8232
Average Data fee ETB	33%	9339
Average voice usage	31%	9216
Average SMS fee	26%	8208
Status	23%	8264
Average Voice fee	15%	732

Table 5- 2 The RF result before Pruning

The RF is not descriptive to interpret the result. As shown on Table 5.2 the three features high IG values also has an important feature in the RF algorithm result. RF also use pruning to manage the tree. It is adjusted on minim tree size from unlimited value to 4. As shown on the table number node is decreased after pruned. The accuracy of the model is also minimized from 97.4% to 95.6%.

Features	Importance of Features	Number of Node
Activation Year	57%	26
Average Data usage MB	29%	40
Average Data fee ETB	14%	36
Average voice usage	6%	4
Average SMS fee	6%	15
Status	6%	10
Average Voice fee	4%	11

Table 5- 3 The RF result after Pruning

NB model building and interpretation

Call reason	Sim status		Internet package		Lost	
Measurement	Mean	STD	Mean	STD	Mean	STD
Activation year	2020.9	0.793	2015.8	4.45	2015.6	3.98
Average data usage MB	246.6	1299.3	3064.1	8307.42	32.1	73.9
Average data fee	17.6	49	70.5	104.4	514.3	2346.0
Average Voice usage	265.3	265.6	325.4	286.9	361.4	325.2
Average SMS fee	2.33	5.82	4.57	6.0	2.97	4.57
Status	1755	118.6	56.0	193.0	118.0	21.0
Average voice fee	52.1	75.2	40.9	44.0	59.7	48.1

Table 5- 4 The NB result

The NB is not descriptive but in Table 5.4 based on mean NB algorithm outcome shows that a customer subscribed 2020 and above calls to the CC for a SIM status request. The customer subscribed less than 2020 mostly asks for lost balance and internet package. Average data usage 3064.1MB call to CC for an internet package. The customer Average data usage 32.1MB call to CC for complain lost balance.

6. Conclusion and Future work

6.1 Conclusion

The CC that connects a service provider with a customer. The information was obtained from IPCC, which is a database of service number calls to CC requests classified by top call reason. Then, from CRM, the status of the service number and the subscription year. Following that, there is data usage, voice usage, and SMS usage. Then, employ three popular classification algorithms: J48, RF, and NB. WEKA is being used as an experimental tool. The RF and J48 produced the most accurate results than NB. For the amount of time spent for building the model, NB is less computational time but also J48 is less time than RF algorithm. So, J48 algorithm is best algorithm for this work because score high accuracy, easily interpretable model and less computational time.

6.2 Future Work

Future works include:

- Working for other top call reason, this helps to give a solution for each customer's problem.
- Working on data more than three months.

Reference

[1] N. N. Annisa, D. I. Sensuse, and I. Wilarso, "Knowledge Base Model for Call Center Department: A Literature Review," *2018 Int. Conf. Inf. Technol. Syst. Innov. ICITSI 2018 - Proc.*, pp. 242–247, 2018, doi: 10.1109/ICITSI.2018.8696042.

[2] G. Mishne, D. Carmel, R. Hoory, A. Roytman, and A. Soffer, "Automatic Analysis of Call-center Conversations," pp. 453–459, 2005

[3] M. M. Dr. Bhargava N., Sharma G., Dr. Bhargava R., "International Journal of Advanced Research in Decision Tree Analysis on J48 Algorithm for Data Mining," *Int. J. Adv. Res. Comput. Sci. Softw. Eng.*, vol. 3, no. 6, pp. 1114–1119, 2013.

[4] R. Mahajan and S. Som, "Customer behavior patterns analysis in Indian mobile telecommunications industry," Proc. 10th INDIACom; 2016 3rd Int. Conf. Comput. Sustain. Glob. Dev. INDIACom 2016, no. April 2013, pp. 1165–1169, 2016

[5] G. Silahtaroglu and H. Donertasli, "Analysis and prediction of E-customers' behavior by mining clickstream data," *Proc. - 2015 IEEE Int. Conf. Big Data, IEEE Big Data 2015*, pp. 1466–1472, 2015, doi: 10.1109/BigData.2015.7363908.

[6] Y.-C. Huang, L.-C. Yu, and I.-C. Lin, "Text Categorization for Service Request Classification," *Int. J. Signal Process. Syst.*, vol. 1, no. 2, pp. 54–58, 2013, doi: 10.12720/ijsps.1.1.54-58.

[7] B. Rao, "Machine Learning Algorithms: A Review," *Int. J. Comput. Sci. Inf. Technol.*, vol. 7, no. 3, pp. 1174–1179, 2016, doi: 10.21275/ART20203995.

[8] D. S. Veena, T. Shankari, S. Sowmiya, and M. Varsha, "a Survey on Tools Used for Machine Learning," *Int. J. Eng. Appl. Sci. Technol.*, vol. 04, no. 09, pp. 116–119, 2020, doi: 10.33564/ijeast.2020.v04i09.012.

[9] A. Choon Tan and D. Gilbert, "Machine Learning and its Applications: An Overview," *Univ. Glas. Dep. Comput.*, no. January, 2003, [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.5.7839&rep=rep1&type=pdf.

[10] R. Patil and V. M. Barkade, "Class-Specific Features Using J48 Classifier for Text Classification," *Proc. - 2018 4th Int. Conf. Comput. Commun. Control Autom. ICCUBEA 2018*, pp. 1–5, 2018, doi: 10.1109/ICCUBEA.2018.8697473.

[11] I. H. Sarker, "Machine Learning: Algorithms, Real-World Applications and Research Directions," *SN Comput. Sci.*, vol. 2, no. 3, pp. 1–21, 2021, doi: 10.1007/s42979-021-00592-x.

[12] A. Al-Mudimigh, F. Saleem, and Z. Ullah, "Efficient implementation of data mining: Improve customer's behaviour," *2009 IEEE/ACS Int. Conf. Comput. Syst. Appl. AICCSA 2009*, pp. 7–10, 2009, doi: 10.1109/AICCSA.2009.5069289.

[13] R. C. Roberts, C. Tong, R. S. Laramee, G. A. Smith, P. Brookes, and T. D'Cruze, "Interactive analytical treemaps for visualisation of call centre data," *Ital. Chapter Conf. 2016 - Smart Tools Apps Comput. Graph. STAG 2016*, pp. 109–117, 2016, doi: 10.2312/stag.20161370.

[14] N. Mehrbod, A. Grilo, and A. Zutshi, "Caller-Agent Pairing in Call Centers Using Machine Learning Techniques with Imbalanced Data," *2018 IEEE Int. Conf. Eng. Technol. Innov. ICE/ITMC 2018 - Proc.*, 2018, doi: 10.1109/ICE.2018.8436314.

[15] S. Singaravelan, D. Murugan, and S. Mayakrishnan, "A Study of Data Classification Algorithms J48 and SMO on different Datasets," *Asian J. Res. Soc. Sci. Humanit.*, vol. 6, no. 6, p. 1276, 2016, doi: 10.5958/2249-7315.2016.00284.7.

[16] S. S. Raju and P. Dhandayudam, "Prediction of customer behaviour analysis using classification algorithms," *AIP Conf. Proc.*, vol. 1952, 2018, doi: 10.1063/1.5032060.

[17] C. Services, "Contact Center Work Instructions," pp. 1–89.

[18] "IPCC_service_topology." .

[19] Dinakaran S and Dr. P. Ranjit Jeba Thangaiah, "Role of Attribute Selection in Classification Algorithms," *Int. J. Sci. Eng. Res.*, vol. 4, no. 6, pp. 67–71, 2013, doi: June 2013.

[20] V. Agarwal, "Research on Data Preprocessing and Categorization Technique for Smartphone Review Analysis," *Int. J. Comput. Appl.*, vol. 131, no. 4, pp. 30–36, 2015, doi: 10.5120/ijca2015907309.

[21] H. Court, "Costs Service Quality Employee Satisfaction," *Time*, pp. 135–143, 2003.

[22] Dinakaran S and Dr. P. Ranjit Jeba Thangaiah, "Role of Attribute Selection in Classification Algorithms," *Int. J. Sci. Eng. Res.*, vol. 4, no. 6, pp. 67–71, 2013, doi: June 2013.

[23] K. B. Al Janabi and R. Kadhim, "Data Reduction Techniques: A Comparative Study for Attribute Selection Methods," Int. J. Adv. Comput. Sci. Technol., vol. 8, no. 1, pp. 1–13, 2018, [Online]. Available: http://www.ripublication.com.

[24] D. Visa Sofia, "Confusion Matrix-based Feature Selection Sofia Visa," ConfusionMatrix-based Featur. Sel. Sofia, vol. 710, no. January, p. 8, 2011.

[25] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," Inf. Process. Manag., vol. 45, no. 4, pp. 427–437, 2009, doi: 10.1016/j.ipm.2009.03.002.

[26] H. M and S. M.N, "A Review on Evaluation Metrics for Data Classification Evaluations," Int. J. Data Min. Knowl. Manag. Process, vol. 5, no. 2, pp. 01–11, 2015, doi: 10.5121/ijdkp.2015.5201.

[27] F. Esposito, D. Malerba, and G. Semeraro, "A comparative analysis of methods for pruning decision trees," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, no. 5, pp. 476–491, 1997, doi: 10.1109/34.589207.

[28] K. Rattanathavorn and W. Premchaiswadi, "Analysis of customer behavior in a call center using fuzzy miner," *Int. Conf. ICT Knowl. Eng.*, vol. 2015-December, pp. 137–141, 2015, doi: 10.1109/ICTKE.2015.7368485.

[29] S. Sumam, "Performance Evaluation By Artificial Neural Network Using WEKA," Int. Res. J. Eng. Technol., vol. 3, no. 03, pp. 1459–1464, 2016.

[30] P. Galdi and R. Tagliaferri, "Data mining: Accuracy and error measures for classification and prediction," *Encycl. Bioinforma. Comput. Biol. ABC Bioinforma.*, vol. 1–3, no. January, pp. 431–436, 2018, doi: 10.1016/B978-0-12-809633-8.20474-3.

[31] S. Echchakoui, "Addressing Differences Between Inbound and Outbound Agents for Effective Call Center Management," *Glob. Bus. Organ. Excell.*, vol. 36, no. 1, pp. 70–86, Nov. 2016, doi: 10.1002/JOE.21757.

[32] Asniar and K. Surendro, "Predictive analytics for predicting customer behavior," *Proceeding - 2019 Int. Conf. Artif. Intell. Inf. Technol. ICAIIT 2019*, pp. 230–233, 2019, doi: 10.1109/ICAIIT.2019.8834571.

[33] "What does the kappa statistic take into account? – Mvorganizing.org." https://www.mvorganizing.org/what-does-the-kappa-statistic-take-into-account/ (accessed Sep. 25, 2021).

Appendix

Classification of Top Call Reasons using Machine Learning in Call Center Service

Dereje Hailemariam (Phd)
derejehmr@gmail.com Atsede Abebe
Atse.mitu@gmail.com

School of Electrical and Computer Engineering,

Addis Ababa Institute of Technology, Addis Ababa University.

Abstract

A call center (CC) connects customers to a service provider, such as telecom operators. The CC receives of customer requests, feedback, and complaints. These inputs from the customers provide an opportunity to comprehend the customer's needs, problems the customers face when using services, and the performance of the service provider. Meeting customer expectations by responding to complaints increases customer satisfaction, which translates to revenue maximization. Hence, the CC is critical to the success of the service provider.

Ethio-telecom, a telecom service provider in Ethiopia, operates a large CC that provides telecom-related services throughout the country. The CC accepts over two million calls per day via a service-free line. The center records and maintains a huge amount of customer-related data, which can further be analyzed using state-of-the-art machine learning algorithms for the purpose of proactively estimating call types and reasons.

Proposes to map features from customer profile information into top call reasons so as to better understand customer call requests and map future calls to specific top call reasons. Data was extracted from Ethio-telecom's IP contact center, customer relational management, and customer billing system servers. To construct the classification models, J48, Random Forest (RF), and Naive Bayes (NB) algorithms are used. Accuracy, time to build a model, and model interpretation of each algorithm are used to compare their performance.

Results show that RF and J48 algorithms outperform NB, with scores of 97.46% and 97.4%, respectively. The NB model is the least accurate, with an accuracy of 83.6%. However, the time spent building a model for NB is less compared to J48. During the model's interpretation, J48 algorithm is more interpretable than the NB and RF. J48 algorithm are best.

Keywords—: Classification algorithm, Call Center, IP Contact Center, Customer Relational Management, Customer Billing System, J48, Naive Bayes, and Random Forest.

Introduction

A call center (CC) is a customer service in the form of an inbound call from the customer or an outbound call to the customer. Its aim is to help solve a customer's problem or to aid with the company's business process. A CC refers to help desks, information lines, and customer service centers as a whole. Customer assistance, operator services, inbound and outbound telemarketing, and web-based services are some of the services generally provided by these centers [1].

The CC can be broadly categorized into two types: Inbound and Outbound. Inbound CC is typically focused on providing support to clients that need to solve problems or follow instructions [2]. Instead of receiving calls from customers, agents in an outbound CC make calls to them [3]. The CC delivers service in two ways: direct voice calls and multi-channel service. A direct voice call is a service given in a telephone conversation. Multi-channel is a type of service that is delivered through Short Message Service (SMS), web chat, and various social media. So, CC service is critical in the telecom industry for responding to customer complaints, increasing customer satisfaction, and increasing revenue.

An IPCC (IP Contact Center) platform is used by the majority of telecommunications companies to accept incoming calls. The IPCC system includes an IVR (Interactive Voice Respond) platform and automatically routes calls to agents. When a customer calls a telecom CC, the system first connects the customer to IVR. The IVR acts as a go-between for the agent and the customer. The call handling process can be categorized into three steps. The first is the answering of the call by the IVR and the time spent in the queue if the line is busy. The second is the time that the agent spends handling the customer's request.

Finally, the closing time is anything the agent has to do to wrap up the call, such as registering a call reason and forwarding complaints to the respective offices. Call reasons are the reason that a customer reaches any service provider's CC.

Call reasons are logged after each call by agents. Logging call reasons correctly could benefit service providers by unearthing user experience issues and taking reactive measures for incidents quickly. To achieve this, one would select top-call reasons. Top call reasons are the causes of customers' contacting a CC the most. Ethio-telecom generate a huge amount of call reason data that is collected each day. The CC has been overloaded with customer complaints and requests regarding common occurrences. As a result, the department employs a large number of people. Despite the fact that CC employees are overburdened with customer requests, the service is difficult to access, resulting in a variety of call drops. Dropped calls have a significant impact on customer satisfaction levels. The CC has KPIs (Key Performance Indicators) such as answered calls, working time, average talk time, call drop, and so on. It is still struggling to meet the desired level of KPI. As a result, the customer is dissatisfied because they did not receive adequate service. Customers are repeatedly contacting the CC. However, Ethio-telecom CC has no knowledge of where the request originated. It only knows when they are confronted with it. But, there is a daily report of top calls from the past to address this issue, we examine the history of a customer's calls beginning with the request they make in the CC. Incoming call requests are classified based on a customer's historical top call reasons and other attributes.

The top call reason is determined based on pr evious call requests that can classify the requ est when the customer calls.

Literature Review

The study [3] investigates customer behavior by employing CRM and DM (Data Mining) methodologies to analyze user behavior. It creates client behavioral models by estimating attributes such as age, income, and lifestyle. The methodology used is rule-based DM, which extracts patterns from data and then uses those patterns for various purposes, such as prediction. The researchers who carry out the methodology's primary steps. The first step is to solicit client feedback. This step selects several inquiries raised by the customer. In the second phase, it groups clients depending on the features of their inquiries, history, and profile. Then it clusters based on similarity. In the third stage, rule induction with cluster data, it offered a model. The proposed model is utilized to increase sales, replay consumer inquiries, and develop marketing initiatives. The rule induction process also makes use of clustered data to generate new rules that may be used to better understand client needs and the organization's growth. The fourth step is customer knowledge, which entails applying a rule induction method to pattern data to better understand customer behavior, satisfaction, and loyalty. The article concludes that rule induction on customer clustered data is a critical factor enhancement for any organization's CRM improvement.

The primary goal of this paper [4] is to examine customer behavior in a telecommunications company's CC. It consists of four steps. It takes data from the corporation first, then the event logo. Case ids, which are defined in two columns, are the data attributes of the collected data. It defines four activity categories in the activity attribute: type of action, request type of action, complaint type of action, and operational initiative type of action. In the paper, they evaluate customer behaviors using fuzzy mining and Disco and ProM5.2.

The paper [5] Predictive analytics is the use of data, statistical algorithms, and machine-learning (ML) techniques to identify likely future outcomes based on historical data. The effectiveness of predictive analytics is more about making better decisions. Previously, with a limited data volume, intuitive decision making was nevertheless successful. However, as the bulk of the data has grown to

amazing dimensions, the human ability to make intuitive judgments has diminished. As a result, data-driven decision making is increasingly regarded as assuring a credible pathway to better decision making.

The goal of the study, as stated in [6], is to employ pattern analysis of customer actions to predict churn and for intelligent and targeted promotions. With the data acquired, the researcher begins the methodology. The quality of the data is the most difficult task in the data collection procedure. Customer type, recharging details, outgoing phone calls, incoming voice calls, and SMS sent are the data kinds used. Name, profession, gender, income demographic data, mobile number, voucher type, recharge kind, recharge amount, balance, accumulated call count, call duration, call type, amount, count all types, and rated amount are the main factors. The total quantity of recharges done each month has fallen, but the number of customers recharging has stayed constant, according to the publication.

In [7], the author analyzes customer behavior on a website based on click stream, which means how many times the customer clicks on the website. The goal is to comprehend the customer's requirements. The proposed solution in the paper is based on an artificial neural network. The model accepts input, output, and hidden layers.

The [8] study employs text categorization techniques to automatically analyze the use of internal service applications with the goal of reducing process delays. The system design has a six-step request. It collects service and then labels the document. WEKA's (Waikato Environment for Knowledge Analysis) experiment test employs Bayesian, SVM (Support Vector Machine), and decision tree algorithms. The paper concludes that electronic systems will reduce delays in correctly assigning documents, resulting in significant enterprise-wide time savings.

Description of the Call Center

Telecommunication network operators run call centers to accept customers' requests, comments and complaints. Call centers handle customer requests on a daily basis. The centers are organized such that incoming calls are assigned to agents (or adviser) responsible to handle customer's requests based on working procedures set aside by the company. From the call reason categories labeled on the system of

IPCC, the agent registers a specific call reason for the call made. Some requests are handled by the agent, while others are forwarded to concerned bodies.

IPCC Platform

Telecom call centers use state-of-the-art IP Contact Center (IPCC) platforms, that accepts customers calls and display profile information that includes:

- Customer profile information:- e.g., name, calling number, SIM status (like active, barred, suspended,), activation year, value added service usage (e.g., CRBT, voice mail), packages bought, balance, historical call reasons, callers geographic region.
- KPI parameters:- number of customers in queue, number of served calls.

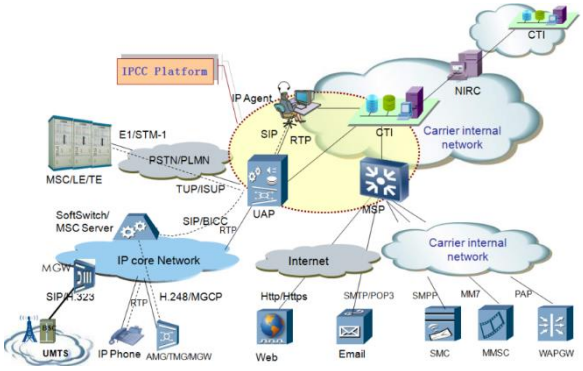

Figure 1: IPCC Structure

Customer Profile Information

Taking the mobile cellular service, common customer profile information include customer's name, calling number, Subscriber Identification Module (SIM) status (like active, barred, suspended), activation year, value added service usage.

Top call Reasons

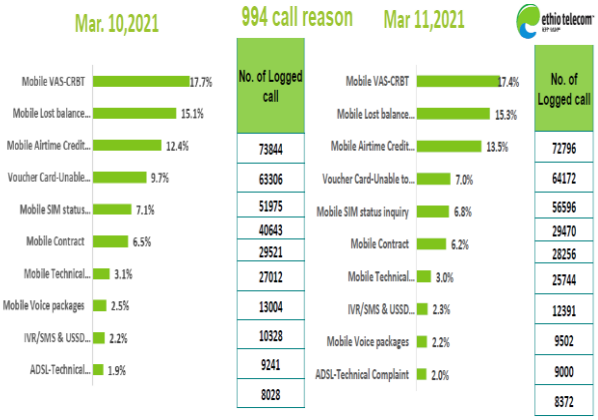

Figure 2: Top call reason report

Classification Algorithms

Naive Bayes

It calculate the possibility of whether a data point belongs within a certain category or does not. Naïve Bayes classifier assumes that the presence of one feature in a class has no bearing on the presence of any other feature.

J48

- The J48 algorithm is considered to be a fast classifier**.**
- Do not require any assumption of linearity for the variables in the data.
- It can also work on constant and categorical variables.
- It performs association**,** lacking far for calculation.
- It can generate understandable rules.
- Easy to use, build and interpret.

Random forest

- The process of averaging or combining the results of different decision trees helps to overcome the problem of overfitting.
- Random Forest also has less variance than a single decision tree. (It works correctly for a large range of data items than single decision trees.)
- Random Forest scores very high accuracy.
- It also maintains accuracy when outliers are occurred or even when a large proportion of the data are missing.

Features for classification

Feature selection

Feature Name	Calculation	Data Type	Description
Activation Year	Year of activation Date	Numerical	Subscription year related to service request behavior.
Status	SIM Status	Nominal	Status of service number that the customer calls.
AVG _SMS_FEE	$= \frac{\text{Total SMS fee}}{\text{Number of month}}$	Numerical	To extract information about the usage range of the SMS service.
AVG_Data_ usage_MB	$= \frac{\text{Total data usage}}{Number\ of\ month}$	Numerical	To understand data usage behavior of customer.
AVG _Data _FEE ETB	$= \frac{\text{Total data fee}}{\text{Number of months}}$	Numerical	To understand Data fee behavior of customer.
AVG _voice _Usage	$= \frac{\text{Total voice usage}}{\text{Number of months}}$	Numerical	To extract information about the usage range of the voice service.
AVG_VOICE_FEE ETB	$= \frac{\text{Total voice fee}}{\text{Number of months}}$	Numerical	To understand voice fee behavior of customer.

Table 1: Selected features and descriptions

Performance metrics

Correctly classified instances, incorrectly classified instances, Kappa statistics, mean absolute error, root mean squared error, relative absolute error, root relative squared error, total number of instances.

ML Algorithm Result		**Predicted**		
		D	**A**	**Total**
Actual	**D**	**a**	b	m_1
	A	c	**d**	m_0
	Total	n_1	n_0	**n**

Table 2: Confusion matrix

Outcome Accuracy

A true positive correctly predicts the positive value.

TP = a/m_1

A true negative correctly predicts the negative value.

TN = d/m_0

A false positive incorrectly predicts the positive value.

FP = c/m_0

A false negative incorrectly predicts the negative value.

FN = b/m_1

Accuracy = **(a + d)/n**

Precision = **a/n_1**

Recall = **a/m_1**

F-Measure: A combined measure for precision and recall.

F-measure = **2a / (2a + c + b)**

Data Collection

The data collecting from customer relational management, customer billing system and IP contact center. Customer profile information, Activation date IPCC Customer calling number to call center , Calling reason CBS Voice usage and fee, Data Usage and fee, short message usage and fee Check the missing and invalid value and remove it .

Aggregation data

Customer's service numbers are collected from the IPCC server for those service numbers, three consecutive months of usage statistics are mainly for voice, SMS, and data generated from CBS. The CRM server holds customer information such as activation year and status.

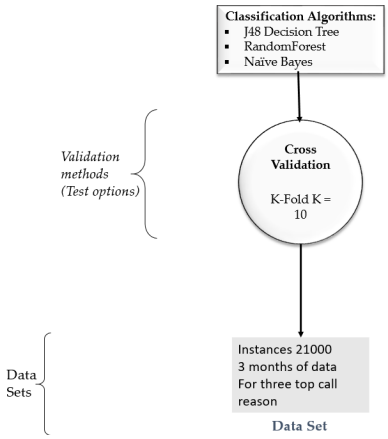

Figure 3: Experiment process

Result and Discussion

Accuracy

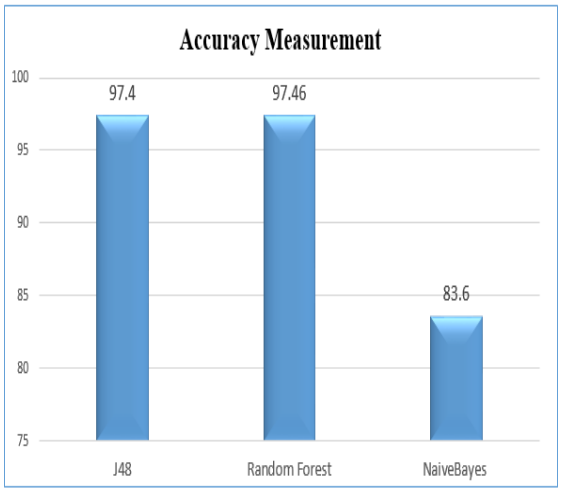

Figure 3: Accuracy measurement comparison

The accuracy comparison of each algorithm is one of the measures for the performance of the model. Therefore, for the experiment use, three supervised classification algorithms. Which are RF, J48 and NB employed and the result of each compare for accuracy. The figure show that RF, J48 and NB produces accuracy score of 97.46%, 97.4% and 83.6% respectively.

Attributes	IG	Ranked by IG
Service no Status	0.105	5
Activation Year	0.737	1
Average_SMS_Fee	0.266	4
Average_Data_Usage_MB	0.7198	2
Average_Data_Fee ETB	0.460	3
Average_Voice_Usage	0.023	7
Averge_Voice_Fee ETB	0.096	6

Table 3: Attribute rank by Information Gain

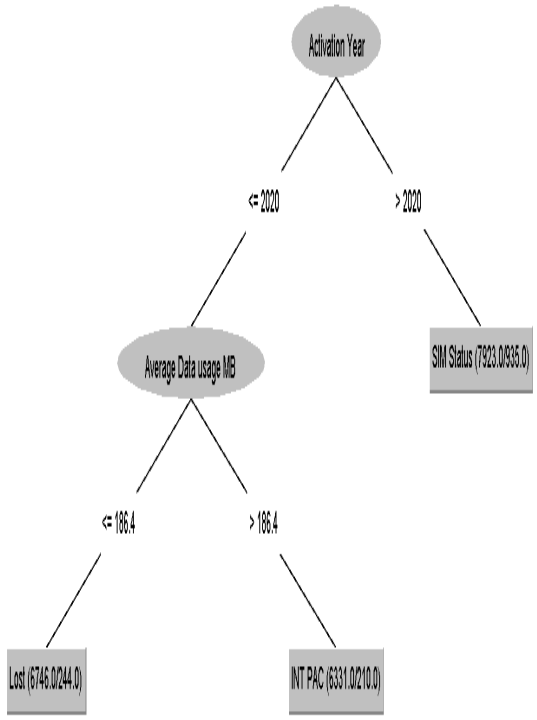

Figure 4: J48 decision tree result

As seen in the table attribute ranking employs using IG. The activation year is rated first, followed by the average data usage in MB and the average data fee. In Figure above show the decision tree outcome, three attributes are also displayed. The activation year is represented by the root node of the tree.

The first research question is which feature or attribute influences the incoming customer call request? There are seven features that are selected for the analysis part. Each feature has its own value for the target class. The activation year has a high influence on the incoming customer requests. The customer subscription year is highly related to customer requests. As shown in the Table, the three most Information Gain features. Which is the activation year, average data usage MB, and average data fee. Also in the J48 decision tree show above on Figure, these three features decide on the incoming call request categorization. As seen on the tree activation year, the root node it is classified based on its activation. The customer's average usage of data is also the second feature to decide the customer's reason for calling the CC. The third feature is that the average data fee also affects the incoming call classification.

The second research question is: which customer's requests are categorized as the top call reasons? As shown on J48 decision tree a customer calls to CC to request a balance loss if the calling number's activation year is before 2020, specifically 2020, and the customer's average data use is less than 186.4 MB. Whereas average data usage of more than 186.4MB needs the purchase of an internet plan, the activation year subscription after 2020 contacts the CC for a SIM status request.

As a result, those clients' service numbers are not renewed on time, and they are no longer allowed to call.

The third research question, which classification algorithm is most suitable for forecast the incoming call? J48 is the best algorithm since it has fewer error occurrences and easily interpretable and less computational time than other two algorithm. For the problem, the J48 algorithm is superior to RF and NB.

Conclusion

The CC that connects a service provider with a customer. The information was obtained from IPCC, which is a database of service number calls to CC requests classified by top call reason. Then, from CRM, the status of the service number and the subscription year. Following that, there is data usage, voice usage, and SMS usage. Then, employ three popular classification algorithms: J48, RF, and NB. WEKA is being used as an experimental tool. The RF and J48 produced the most accurate results than NB. For the amount of time spent for building the model, NB is less computational time but also J48 is less time than RF algorithm. So, J48 algorithm is best algorithm for this work because score high accuracy, easily interpretable model and less computational time

References

[1] N. N. Annisa, D. I. Sensuse, and I. Wilarso, "Knowledge Base Model for Call Center Department: A Literature Review," *2018 Int. Conf. Inf. Technol. Syst. Innov. ICITSI 2018 - Proc.*, pp. 242–247, 2018, doi: 10.1109/ICITSI.2018.8696042.

[2] G. Mishne, D. Carmel, R. Hoory, A. Roytman, and A. Soffer, "Automatic Analysis of Call-center Conversations," pp. 453–459, 2005.

[3] M. M. Dr. Bhargava N., Sharma G., Dr. Bhargava R., "International Journal of Advanced Research in Decision Tree Analysis on J48 Algorithm for Data Mining," *Int. J. Adv. Res. Comput. Sci. Softw. Eng.*, vol. 3, no. 6, pp. 1114–1119, 2013.

[4] K. Rattanathavorn and W. Premchaiswadi, "Analysis of customer behavior in a call center using fuzzy miner," *Int. Conf. ICT*

Knowl. Eng., vol. 2015-December, pp. 137–141, 2015, doi: 10.1109/ICTKE.2015.7368485.

[5] Asniar and K. Surendro, "Predictive analytics for predicting customer behavior," *Proceeding - 2019 Int. Conf. Artif. Intell. Inf. Technol. ICAIIT 2019*, pp. 230–233, 2019, doi: 10.1109/ICAIIT.2019.8834571.

[6] R. Mahajan and S. Som, "Customer behavior patterns analysis in Indian mobile telecommunications industry," Proc. 10th INDIACom; 2016 3rd Int. Conf. Comput. Sustain. Glob. Dev. INDIACom 2016, no. April 2013, pp. 1165–1169, 2016.

[7] G. Silahtaroglu and H. Donertasli, "Analysis and prediction of E-customers' behavior by mining clickstream data," *Proc. - 2015 IEEE Int. Conf. Big Data, IEEE Big Data 2015*, pp. 1466–1472, 2015, doi: 10.1109/BigData.2015.7363908.vol. 7, no. 3, pp. 1174–1179, 2016, doi: 10.21275/ART20203995.

[8] Y.-C. Huang, L.-C. Yu, and I.-C. Lin, "Text Categorization for Service Request Classification," *Int. J. Signal Process. Syst.*, vol. 1, no. 2, pp. 54–58, 2013, doi: 10.12720/ijsps.1.1.54-58.

[9] A. Choon Tan and D. Gilbert, "Machine Learning and its Applications: An Overview," *Univ. Glas. Dep. Comput.*, no. January, 2003, [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.5.7839&rep=rep1&type=pdf.

[10] R. Patil and V. M. Barkade, "Class-Specific Features Using J48 Classifier for Text Classification," *Proc. - 2018 4th Int. Conf. Comput. Commun. Control Autom. ICCUBEA 2018*, pp. 1–5, 2018, doi: 10.1109/ICCUBEA.2018.8697473.