Addis Ababa University

Addis Ababa Institute of Technology

School of Electrical and Computer Engineering

Graduate Program in Telecommunication Engineering

CDR Based Recommender System for Mobile Package Service Users

Thesis

Submitted in partial fulfillment of the requirements for

The degree of Master of Science in Telecommunication Engineering

By: Saba Mulugeta

Advisor: Sosina Mengistu (PHD)

Addis Ababa, Ethiopia

September, 2023

DECLARATION

I declare that the thesis titled ***" CDR Based Recommender System for Mobile Package Service Users"*** submitted by me for the degree of Master of Science in Telecommunication Engineering at Addis Ababa University/Addis Ababa Institute of Technology my original work and has not been submitted previously for any degree or examination at any other university or institution. All sources of information and material I have used in the thesis, including any previously published material, have been duly acknowledged.

Saba Mulugeta Walle _____

Name Signature

September 2023

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

This is to certify that the thesis prepared by Saba Mulugeta, entitled ***CDR based recommender system for mobile package users*** and submitted in partial fulfillment of the requirements for the degree of Master of Science 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 _____

Internal Examiner _____ Signature _____ Date _____

Advisor Sosina Mengistu (PHD) Signature _____ Date _____

Dean, School of Electrical and Computer Engineering

DEDICATION

This paper is dedicated to my beloved father.

ABSTRACT

Due to increased competition, telecom operators are continually introducing new products and services. To make their services easy to use, to meet customer requirements, and to satisfy their customers' needs in terms of payment, operators have launched various telecommunication packages. Telecom operators have so many packages that customers are unaware of; some packages may go unnoticed even if they are useful. To overcome such problems, we need a recommender system that directly notifies customers based on their interests. Most research has been conducted on recommendation systems for web service users based on user ratings. In this paper, we propose a mobile package service recommendation system for customers.

The proposed recommendation system has two phases. The first is creating a relationship between customer usage and mobile packages by grouping customers based on their usage. To create a relationship between customers' usage and mobile packages, we have used the k-means clustering algorithm. The elbow method is used to determine the number of clusters for each service. The second phase is building a classification model that will recommend mobile packages for users. Two-month CDR data was used to build a classification model by using random forest (RF) and K-nearest neighbor (KNN) classifier algorithms.

The evaluation result shows KNN outperformed RF for weekly and monthly data usage plans with F1 scores of 90.4% and 96%, respectively, whereas RF outperformed for daily plans with an F1 score of 86.9%. On the other hand, RF outperformed KNN with F1 scores of 95.10% and 99.60% for daily and monthly voice usage plans, respectively. Similarly, KNN showcased better performance than RF on the weekly voice usage plan with an F1 score of 94.30%.

Generally, the strengths of each algorithm differ for different usage scenarios within the voice and data service domains.

Keywords: Recommender system, Collaborative filtering, Content-based recommender systems, Mobile recommender systems, Clustering and Classification.

ACKNOWLEDGMENT

Writing and finishing this thesis was a long journey and an excellent learning process with wonderful experiences. It's a pleasure to thank all the people involved in the journey who made the thesis possible.

First and foremost, I would like to say thank you to the almighty God, who gives me the strength and patience to finish this study. Secondly, I would like to say thank you to Ethiotelecom, which believes in me and gives me the chance to study telecom engineering.

Then I would like to express my great respect and thank my advisor, Dr. Sosina Mengistu, for the insightful discussion and guidance throughout this thesis.

Finally, I would like to express my deepest gratitude to my beloved husband and amazing son for their understanding, patience, and support in completing the study.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ACRONYMS

App	Application
AVG	Average
CDR	Call detail record
D	Day
EMS	Expectation-maximization algorithms
e.g.	Example
G	Group
GB	Giga bit
ID	Identification
IKNN	Improved Nearest neighbor
KNN	K-nearest neighbor
M	Month
min	Minute
MB	Megabit
P	Package label
RF	Random forest
RSs	Recommendation system
SMS	Short Message Service
USSD	Unstructured Supplementary Service Data
VAS	Value added services
W	Week

1 INTRODUCTION

1.1 BACKGROUND

Telecommunication services and products have been growing. Due to this reason, telecom operators use different methods to make their services and products available to their customers. Mobile package service can be mentioned as an example. Mobile package service is a service that enables customers to access all services (voice, data, voice plus data, and SMS) provided by a telecom company at a lower tariff than the regular one.

Ethiotelecom is one of the telecom operators in Ethiopia. Ethiotelecom, like other operators, offers various mobile packages and is accessible through Ethiogebeta (dialing *999#), the Ethiotele mobile app, and Tele Birr (app or USSD using 127). There are more than eleven package offers on Ethiotelecom that are very useful for customers in terms of charging. Using a package deal will result in a lower price than the normal rate. However, due to the information overload of the package, not all Ethiotelecom customers visit all package offers. Also, all packages are displayed to every customer whenever they try to access Ethiogebeta. For example, a customer whose phone does not support the internet and who does not use the internet will be uninterested in visiting a data package. To overcome such problems, a recommendation system will be important.

A recommendation system is the output of a process of analysis on a dataset of users' preferences, whose goal is to extract the most related or interesting items for a target user [1]. Recommender systems automate the process of recommending products, services, or information items to consumers (subscribers) based on several types of data concerning users, items, and previous interactions between users and items [2].The basic idea of recommender systems is to utilize various sources of data to infer customer interests [3].

The goal of a recommender system is to generate meaningful recommendations for a collection of users for items or products that might interest them. Suggestions for books on Amazon or movies on Netflix are real-world examples of the operation of industry-strength recommender systems. The design of such recommendation engines depends on the domain and the particular characteristics of the available data [4].

Recommendations typically speed up searches, make it easier for users to access content they're interested in, and surprise them with offers they would have never searched for [5].

Based on the types of input data, recommendation models can be categorized into collaborative filtering, content-based recommender systems, and hybrid recommender systems [6].

Collaborative filtering makes recommendations by learning from user or item historical interactions, either through explicit (e.g., user's previous ratings) or implicit feedback (e.g., browsing history). Content-based recommendations are based on comparisons across items' and users' information. Hybrid models are recommender systems that integrate two or more types of recommendation strategies [7, 8] .

Most studies have been done on various recommendation systems for different purposes, such as the telecom industry, e-commerce, education, tourism, and so on. However, the studies used data based on ratings, which is not applicable to mobile package service users. Therefore, the main objective of this study is to develop a mobile package service recommendation system to make the service easier for customers.

1.2 STATEMENT OF THE PROBLEM

Telecom operators launch various mobile-based products and services to satisfy customers, increase revenue, and compete more effectively with other operators. For customers to use services with a lower tariff than the regular tariff, operators introduce different mobile packages. There are so many mobile packages offered by operators that they go unnoticed by all customers, even if they are important.

Different recommender systems have been developed by different scholars. The basic models for recommender systems work with two kinds of data. The first one is user-item interactions, such as ratings or buying behavior, and the second is attribute information about the users and items, such as textual profiles or relevant keywords [1]. Most studies focus on web service users so that they can use rating data to capture the user's preferences, which is not applicable to mobile package users. Most companies that use web-based systems to sell their products have cookies that store the user's preferences whenever users visit that specific website, which helps the company easily identify the user's preferences. Similarly, a lot of products built on web-based systems have rating options to see their customer's feedback, which helps the company easily identify the user's interests. On the other hand, mobile package services are services that are offered to customers offline. On mobile package services, there is no rating data, which makes it difficult to understand and capture the user's preferences.

There is an ever-increasing complexity in understanding the behavioral patterns of subscribers' preferences, which makes the existing recommendation approaches inefficient in meeting

customer demands [2]. As per my knowledge, there is no research conducted on a specific service called the mobile package service recommendation system using CDR data. Taking the CDR data stored by telecom operators as an opportunity, we can build a recommendation system for mobile package service users.

Ethiotelecom is one of the telecom operators found in Ethiopia. Like other operators, Ethiotelecom offers mobile packages for its customers. These mobile packages are accessed by the USSD code *999#, namely Ethiogebeta, Ethiotele mobile app, and Tele Birr (app or USSD using 127). In Ethiotelecom, more than eleven packages are available, each with voice, SMS, and data services. There are so many clicks to see all the packages, which makes it difficult for customers to see all of the available packages, even if they are very useful. To overcome this problem, we need a recommendation system based on the customer's interests and preferences.

1.3 OBJECTIVE

1.3.1 General Objective

The main objective of the research is to build a mobile package service recommender model for customers by analyzing their CDR data using clustering and classification techniques.

1.3.2 Specific Objectives

1. To investigate mobile packages offered by Ethiotelecom.
2. To investigate appropriate algorithms and models applicable to the recommender system.
3. To group customers based on usage patterns.
4. To select features from CDR data.
5. To develop a model that recommends mobile package services for users.
6. To evaluate the effectiveness of the recommender model.

1.4 SCOPE

Collaborative filtering, content-based filtering, clustering, and classification are the different methods for developing recommender systems. The focus of our study is to develop a mobile package recommendation system for customers using clustering and classification techniques based on CDR data. On Ethiotelecom, there are different mobile package offers from those for

which our study focused only on daily, weekly, and monthly packages for both voice and data services.

1.5 CONTRIBUTION OF THE STUDY

The main contribution of this study is the development of a model for recommending mobile package services to customers based on their usage patterns. On the operator side, specifically within Ethiotelecom, this study can assist in revising the current mobile package services to better align with customer usage patterns. Additionally, Ethiotelecom can use this study to introduce new mobile package services that are personalized to customer needs. Ultimately, this research can serve as a valuable reference for future studies in the field.

1.6 METHODOLOGY

To address the general and specific objectives of the study, we have used quantitative experimental research methods. The flow diagram in Figure 1.1 illustrates the approach that was taken to complete the thesis. In the literature review, several works that are connected to our subject were reviewed. Problem formulation was done based on findings from the literature review in relation to Ethiotelecom. CDR data was collected from Ethiotelecom and then preprocessed accordingly using both Microsoft excel and Python programming tools. After preprocessing, features were used to group customers based on their usage patterns using the K-means algorithm. Cluster analysis was done using Microsoft Excel to label the grouped data. Using the output of the clustering algorithm and labels as features, a classification model using RF and KNN was built for recommending mobile package services. Finally, the mobile package service recommender system was evaluated based on precision, recall, and F1-score.

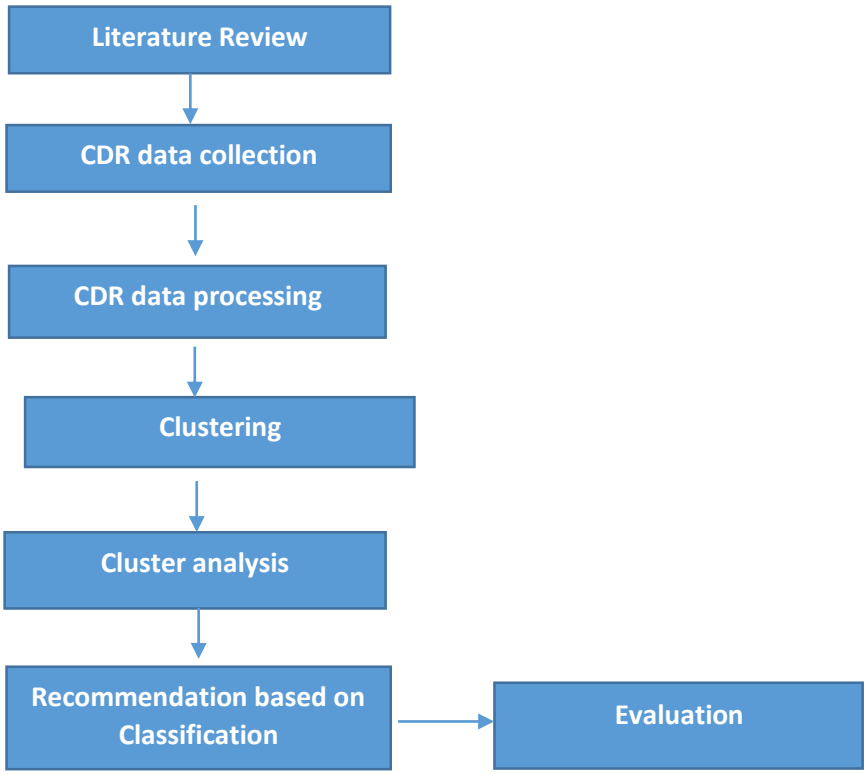

Figure 1-1:Methodology

1.7 THESIS ORGANIZATION

This research paper contains five chapters. The objectives, research methodology, and problem formulation are discussed in **Chapter 1**. In **Chapter 2**, the recommender system and its applicable areas are discussed. In this chapter, related works about the different methods applicable to the recommender system are briefly explained.

Chapter 3 describes the proposed system for the mobile package recommender system. In this chapter, the data set used, the preprocessing steps followed, the methods that were used for understanding customer usage behavior, and how mobile package assignment was done are briefly explained. Building classification-based recommender systems using RF and KNN, the selected features that are applicable for model building, and the evaluation metric selected for testing the effectiveness of the model are also described.

The model performance evaluation is discussed in **Chapter 4**. The researcher's conclusion and further research regarding the mobile package recommender system are pointed out in **Chapter 5.**

2 LITERATURE REVIEW

2.1 RECOMMENDER SYSTEM

The main driving force behind the development of a recommender system is the increase in electronic and business transactions to be held on a web platform. Whenever business transactions are held on the web, companies will get their customers interest and feedback very easily [3]. A recommendation system is mainly about understanding the data and your customers [9]. Recommender systems have different importance. The operational benefit of a recommendation system is relevance. The system recommends products that are relevant to the customers, and this is considered the main goal of the recommender system. The recommendation system is also helpful in recommending products that customers have not seen in the past. For example, in our case, most customers did not know that Flexi unit packages have daily, weekly, and monthly plans, so the system will recommend such packages based on their usage. Increasing recommender diversity is another important aspect of the recommender system. Customers will not be bored by getting similar products since it recommends top-quality products [3].

Recommendation systems are information-filtering and decision-support tools that provide product and service recommendations tailored to the user's specific needs and preferences [9].

The recommendation system is applicable in different areas, like tourism [10] , e-commerce [11], telecom [12], indoor shopping [13], and insurance [14].

2.2 METHODS FOR RECOMMENDER SYSTEM

Mobile information recommendation is becoming very prevalent due to the increasing range, convenience, and use of mobile information services [11]. The basic models for recommender systems work with two kinds of data. The first one is user-item interactions, such as ratings or buying behavior, and the second is attribute information about the users and items, such as textual profiles or relevant keywords [1].

Instead of using historical ratings or buying data, external knowledge bases and constraints are also used to create the recommendation, which is referred to as a knowledge-based recommendation system. Methods that use user-item interactions are referred to as

collaborative filtering methods, whereas methods that use attribute information about the users and items are referred to as content-based recommender methods. Some recommendation systems combine these different aspects to create hybrid systems. Hybrid systems can combine the strengths of various types of recommender systems to create techniques that can perform more robustly in a wide variety of settings [1].

2.2.1 Collaborative filtering

Collaborative filtering is the most common and widely used recommendation system [11]. Collaborative filtering is a technique that can filter out items that a user might like based on the reactions of other similar users. It works by searching a large group of people and finding a smaller set of users with tastes similar to a particular user's. It looks at the items they like and combines them to create a ranked list of suggestions [15]. Generally, it is based on user ratings and community ratings [3].

Miyahara et al. [16] used collaborative filtering based on a simple Bayesian classifier. They used two types of collaborative filtering: user-based filtering and item-based filtering. User-based collaborative filtering makes predictions based on user similarity, and item-based collaborative filtering makes predictions based on item similarity. The main limitation of the study was that it used a simple Bayes classifier with binary class data (likes and dislikes), and the data they used was synthetic. To overcome these limitations, Xiaoyuan SU and Taghi M. Khoshgoftaar [17] proposed collaborative filtering based on an advanced Bayes classifier that works in real-world multiclass data. Using an advanced Bayes classifier makes the prediction robust, and it is much simpler to deal with incomplete and sparse data. In general, it outperforms both the simple Pearson algorithm and collaborative filtering with a simple Bayes classifier.

A collaborative filtering and recommendation system can be used for a variety of purposes, including marketing, telecom, and e-commerce. Kridel et al. [18] created a mobile-based recommender system for selling services to customers. The recommendation system is based on customer purchase history, customer browsing history, and user segments, respectively. Similarly, the authors propose a recommender system for users of telecom services based on different collaborative filtering algorithms, namely GenericItemBased, GenericUserBased, ItemAverage, ItemUserAverage, TreeClustering, and SlopeOne recommendation algorithms provided by Taste libraries. The algorithms were then applied to a complex data set of telecom users. The authors then evaluate different correlation algorithms on a typical telecom data set.

The evaluation was based on time of computation, variation in sample size, neighborhood computation, scalability, and precision.

Linden, G., Smith, B., and York, J. [19] introduce Amazon's recommendation system, which is an item-to-item collaborative recommendation system for the huge Amazon.com website. The recommendation system focuses on finding similar items, not similar customers. For each of the user's purchased and rated items, the algorithm attempts to find similar items. It then aggregates similar items and recommends them.

Generally, collaborative filtering is the most widely used recommendation system that can be applied in different areas like e-commerce, telecom companies, and marketing services.

2.2.2 Content-based recommendation system

A content-based technique uses the distinct characteristics of the product or item that the customer had earlier bought or had shown a liking for so as to give recommendations to the customer of products with identical characteristics [20]. Content-based recommender systems recommend items based on the content information or description of the items [21].

Content-based filtering recommends items based on the textual information of an item under the assumption that users will like similar items to the ones they liked before. The textual description of items is used to build item profiles. User profiles can be constructed by building a model of the user's preferences using the descriptions and types of the items that a user is interested in, or a history of the user's interactions with the system can be stored (e.g., user purchase history, types of items he purchases together, his ratings, etc.) [21].

Michael J. Pazzani and Daniel [22] discussed how content-based recommendation systems in general work. Since content-based recommendation systems require describing the items that may be recommended, a means for creating a profile of the user that describes the types of items the user likes, and a means of comparing items to the user profile to determine what to recommend, The authors prepared a database table that contains item descriptions, and to prepare a user profile, they used the user's preferences and user history that may be recommended, a means for creating a profile of the user that describes the types of items the user likes, and a means of comparing items to the user profile to determine what to recommend. The authors prepared a database table that contains item descriptions. They prepared a user profile using the user's preferences and history. Instead of using rule-based recommendations, the author suggests different machine learning algorithms like decision trees, nearest neighbor

relevance feedback, and probabilistic methods to make comparisons between items and user profiles to determine what to recommend.

2.2.3 Hybrid recommendation system

A hybrid recommendation system is a combination of collaborative filtering and content-based recommendation systems. For recommendation to be done via collaborative filtering, there should be user feedback in the form of ratings, followed by similarities and differences among the profiles of several users in determining how to recommend an item. On the other hand, content-based methods provide recommendations by comparing representations of content contained in an item to representations of content that interest the user [23].

The main limitation of collaborative filtering is that since most users do not rate most items, the user-item rating matrix is typically very sparse, and an item cannot be recommended unless a previous user has rated it, making it difficult for a new user to be recommended. Similarly, content-based recommendations recommendation requires domain knowledge for the recommended item. So combining the very advantages of content-based and collaborative filtering methods gives better accuracy in recommendations [23, 24].

Ghazanfar et al. [21] proposed a unique switching hybrid recommendation approach by combining a Naive Bayes classification approach with collaborative filtering for movie recommendation.

2.2.4 Clustering and classification algorithms for recommendation system

Researchers use clustering and classification algorithms to make recommendation systems more accurate. Beregovskaya, Irina, and Koroteeev, Mikhail [25] examined and offered a novel method for increasing the accuracy of recommendation systems through the use of clustering algorithms.

Using the clustering technique as a first step, recommendation drawbacks like diversity, lack of consistency, and reliability issues can be solved [26-28] Tian et al. [26] developed a customized recommendation system by applying k-means clustering before evaluating similarity.The author employed a hybrid recommendation algorithm with information from a university library. To overcome the sparsity problem, the authors apply clustering of users to a user-category matrix (k = 15). The sparsity of the matrix was 99.99%, but after applying k-means clustering, it was reduced to 76.42%.

With a minor loss in accuracy, Shi et al. [28] introduced a novel technique called ClusDiv that can be used to broaden the diversity of suggestion lists. The goal was to group items and provide a list of suggestions by choosing items from several groups, maximizing diversity without significantly lowering accuracy. ClusDiv can be used in real-world recommender systems without requiring changes to current prediction algorithms.

Miranda et al. [29] did a systematic literature review on the use of clustering algorithms in recommender systems. 51 studies using partitional, hierarchical, fuzzy clustering, co-clustering, and adapted algorithms were reviewed. 21 of the 22 articles on the partition algorithm were based on the k-means method, which is used in recommendation systems because of its efficiency and simplicity. The author concluded that collaborative clustering algorithms based on rating features use mainly three types of clustering (partitional, hierarchical, and fuzzy). Co-clustering algorithms are usually used for relationships between users and items. Content-based RSs use the output of the clustering algorithm process. Thus, a content-based algorithm generates recommendations based on a specific cluster that is similar to the user who needs a recommendation. Similarly, for a hybrid recommendation system, the first step is to cluster users using different algorithms, then the system makes a content-based prediction on items that have not been rated and constructs a new rating matrix. The final rating is a combination of two sets of ratings.

K-means clustering is also widely applied in customer segmentation [30] and pattern identification [31] for the purpose of market analysis. One application of K-means is segmentation, which groups data points into distinct, non-overlapping subgroups to get a better understanding of the data [32].

Utilizing K-means and EMS clustering, the author [30] used usage-based segmentation for developing products. Because the Ethiotelecom segmentation model had holes in the dataset and lacked sufficient features to accurately reflect the customer's usage patterns, the author created an additional customer segment model for mobile service packaging using voice, SMS, and internet services.

Amri et al. [31] identified the electricity usage pattern of customers at a specified time using k-means clustering for the purpose of adjusting the amount of electricity production. In the study, electric consumption of 370 customers was used. K means clustering, which groups users that have similar usage as one group. The authors created four different scenarios regarding the data set and the number of clusters. The data sets without outliner detection and

the data sets with outliner detection have cluster groups of 4 and 5 interchangeably. They have used SSE and the number of iterations as selection criteria for the scenario they have created. Finally, they have selected the data set that didn't remove the outliner, and the number of clusters is 5. Based on their findings, authors established four different patterns of electric consumption and came to the conclusion that summer is when the most electricity is consumed. Winter is when certain clients use the least amount of electricity. However, the spring is when most customers use the least amount of electricity.

For pattern identification and classification, many researchers have tried to employ the K-Nearest Neighbor classifier, which compares a given test tuple with a group of training tuples that are similar to it [33]. A KNN-based recommendation system was also applied in social recommendation systems [34] , movie recommendation [35] and personalized music recommendation systems [36] .

According to [37], in order to combine the mining of web server logs and web contents to identify user navigation patterns and anticipate future request patterns, the K-Nearest Neighbor algorithm was employed in conjunction with five other classification algorithms. The final result demonstrates that the KNN performed better than three other algorithms. Furthermore, of the six algorithms, KNN achieves the greatest F-Score on the training set.

KNN-based recommendation systems had the drawbacks of diversity and a global effect factor whenever the size of the dataset increased. To overcome such an issue, the authors of [35, 36] used improved KNN algorithms for movie and personalized music recommendation systems. They have used a movie dataset of three different sizes: 100k, 1M, and 100M for the movie recommender system and 100k and 1M for the music recommender system. Li, Gang, and Zhang, Jingjing [36] created a new global effect factor formula to balance parameters like the average of all items, the time interval of scoring, and the scoring time of items. Their method gives a better RMS and precision score for IKNN than KNN whenever the data size increases from 100 thousand to 1 million to 100 million.

On the recommendation system, there is a major problem called the cold start problem that happens when new users or new items enter the existing recommender system. When new users enter the recommendation system, there is no data that will explain the interest of the user, which makes it difficult to make a recommendation. So to overcome such issues, the authors [33, 38] proposed a recommendation system using information from social networks and sentiment analysis using a random forest classification algorithm.

Herce-Zelaya et al. [33] extracted implicit information from user's social networks to understand their preferences, tests, and characters. This information is then used to generate a new user profile, which will be used for predicting ratings. After a user profile is formed, a random forest will classify users. Sentiment analysis is also another option to analyze and find the emotions and feeling that were expressed by human beings. Random forest algorithm will predict the polarity of sentiment that was given by users. then by considering sentiment analysis and user profile information product recommendation will be done.

Ajesh et al. [39] proposed recommender system that use clustering and random forest classification as a multilevel strategy to predict recommendation. Based on user ratings K-means and k-means ++ clustering algorithms were used so as it will be easy to generate labels. The labels for the users was predicted using random forest algorithm. After a user profile is formed, a random forest will classify users based on whether one movie is recommended or not. Similarly, random forest algorithm is used to predict insurance products and make recommendation based on the predicted result [14].

2.3 RELATED WORK

The recommendation systems in the telecommunications sector have different purposes. It can be used to recommend value-added services (VAS), telecom products and services, and so on.

Due to the increased number of services and offers in the telecom sector, most of the services or offers may go unnoticed by the customers. To overcome this issue, the author [20] proposed a big data analytics-based recommender system for value-added services like ringtones and games, using customer segmentation and meta-data details. Kridel et al. [18] created a mobile-based recommender system for selling telecom services to customers using a collaborative filtering approach. Sibanda, Elias M., and Zuva, Tranos R. [2] developed a prototype based on a query recommendation system using machine learning algorithms for classifying subscribers per usage stream and clustering all the subscribers with similar usage patterns. Tian et al. [26] developed a customized recommendation system by applying k-means clustering before evaluating similarity. The authors in [36] used a KNN-based recommender system for recommending personalized music and movies.

All related papers presented have their own role in developing recommendation systems. We can see and understand that there are different methods for developing recommendation systems that can be applied in different areas.

In general, there has been a lot of research done on recommendation systems; we can see from the literature that more research has been done on recommendation systems for e-commerce users or web service users based on rating and user history. Mobile recommender systems are an active research area. Not much research has been done on specific telecom services like mobile package services. Mobile package service is one broad service offered by telecom operators. It attracts customers since the tariff is lower than the normal rating. So recommending mobile package service for customers based on their usage will be very important. Recommending mobile package service will satisfy customers, attract new customers, improve the customer experience, and increase the operator's revenue. This work will develop a recommendation system for mobile package service users based on their usage history using the CDR data of the customers.

3 EXPERIMENTAL ANALYSIS

3.1 PROPOSED METHODOLOGY

The proposed methodology recommends mobile package bundles for users. The methodology we propose involves suggesting specific mobile package bundles to users based on their usage patterns. This approach is carried out in two distinct phases:

Phase 1: Customer-Service Relationship Building and Usage Behavior Analysis

In the initial phase of our study, we establish usage relationships among customers of Ethiotelecom. This involves building connections between users and the various services offered by Ethiotelecom, including voice and data services. These connections provide us with valuable insights into how different customers utilize these mobile services. To achieve this, we employ a clustering algorithm that groups users exhibiting similar usage behaviors and interests. Once users are organized into clusters based on their usage patterns, we conduct a comprehensive usage behavior analysis for each service. This analysis is carried out by calculating the average values within each user group, using Microsoft Excel as a tool. With these average values in hand, we formulate rules that align the existing Ethiotelecom service packages with the usage patterns of the grouped customers. This alignment ensures that the packages provide to customers' needs in a cost-effective manner, ultimately benefiting them.

Phase 2: Package Recommendations

The second phase of our study is to build classification models to recommend mobile package services to customers. The model uses both the usage patterns of customers and the attached package label to make mobile package recommendations. The model essentially acts as a decision-making tool that takes into account the usage behaviors and preferences of customers. It then recommends suitable mobile package bundles based on these insights. This way, we ensure that users receive personalized recommendations that match their unique usage preferences.

3.2 DATA COLLECTION

Call detail record data for 7899 active customers was collected randomly from the Ethiotelecom Information System Department. The data was two-month data, from January 2023 to February 2023. The requested CDR data contained voice in seconds, data in kilobytes,

and SMS in numbers. There are also other features stored on CDR data, like revenue generated on each service, off-peak voice, peak voice, and so on.

3.3 DATA PREPROCESSING

Microsoft excel and Python programming tools are used to preprocess our CDR data. Since the raw data was unorganized and has so many features, it needed to be preprocessed. Here are the main preprocessing steps:

1. ***Removing unnecessary features***: From CDR data, features that explain usage patterns are voice, data, and SMS usage from day one up to day 30, from week 1 up to week 9, and from month one up to month two for daily, weekly and monthly base respectively.

   Raw data features were SERV_NO, DATA_DAY, STATUS_NAME, OFF_PEAK_USG_SECONDS, PEAK_USG_SECONDS, DATA_USAGE_KB and SMS_LOCAL_USAGE.

   Among the available features, we specifically selected DATA_DAY, OFF_PEAK_USG_SECONDS, PEAK_USG_SECONDS and DATA_USAGE_KB. SERV_NO primarily serves the purpose of identifying customers within the same group and is not utilized as a feature for grouping or constructing classification models.

   One new feature was created by adding off peak voice usage and peak voice usage namely ***TOTAL_VOICE_USG_*** **SECONDS.**

2. ***Data conversion:*** Raw data was stored in seconds and kilobytes for voice and data services, respectively. Since the available mobile package services are in minutes and megabits, data conversion was done from seconds to minutes and from kilobytes to megabits.

1minute = 60 seconds, and 1MB = 1024 KB

3. ***Data organization:*** Raw data is the day-to-day usage of voice and data services. And it is from day one up to day sixty. Since we are recommending packages for the current Ethiotelecom mobile package, i.e., on a daily, weekly, and monthly basis. We need to organize our data on a daily, weekly, and monthly basis. Daily base means raw data by itself; weekly base is the sum of 7 days; and monthly base means 30 days of aggregated usage for voice, data, and voice plus services.

4. ***Data Aggregation:*** From the raw data, off-peak voice usage and peak voice usage were

used, so to get the total voice usage of customers, we aggregated both off-peak and peak usage.

$$TOTAL_{VOICE_{USG}\ \text{SECONDS}} = \text{OFF}_{\text{PEAK}_{\text{USG}_{\text{SECONDS}}}} + \text{PEAK\_USG}_{\text{SECONDS}} \quad \textbf{(3.1)}$$

5. ***Outliner detection and removal:*** Outliners are values that are very low or very high compared to the usual customer usage. Since we are using the K-means clustering algorithm, which will be discussed in brief in the coming sections, it is very sensitive to outliner values. We have set minimum and maximum threshold values for each column based on quantile values of 0.05 and 0.95, respectively. Values greater than the 95% quantile and less than 5% are treated as outliners and thus removed.
6. ***Normalization*** is the process of making data sets in an equal range. We used the Min-Max scalar method to make data points on the same scale. The Min-Max scalar method makes all our data range between 0 and 1. Since we are using K-means clustering, which is distance-based clustering, the data points need to be on the same scale. So it becomes an essential step before clustering because Euclidean distance is highly sensitive to changes in the differences [40].

After preprocessing our data set, phase one of our work will start, which is customer-service relationship building and usage behavior analysis. To create and understand customer usage patterns, we have grouped customers based on their usage behavior using a clustering algorithm.

3.4 CLUSTERING

Clustering is an unsupervised machine learning technique that is capable of grouping data that does not contain labels or previous knowledge. Clustering is an important tool of machine learning and pattern recognition. Data points in the same cluster have similar properties, while data points in different clusters will have different properties [29]. Clustering can be used in various areas like segmentation [30], recommendation systems [29], market analysis [18], and so on. There are different types of clustering algorithms, on this work to group customers based on their usage and understand their usage patterns, the K-means clustering algorithm was used.

3.4.1 K-Means algorithm

The K-means clustering algorithm is a centroid-based algorithm (which organizes data points into a non-hierarchal cluster) in which one data point fails only in one group. This algorithm uses Euclidian distance to divide the datasets into k clusters, where k is the number of groups created. It assigns the data points to the cluster closest to the centroid point in Euclidean space.

The K-means algorithm is very sensitive to noise data, outlier values, and missing values. To overcome such an issue, proper data preprocessing has to be done [2]. As explained in the preprocess section, voice and data sets have been preprocessed.

Generally, we selected the K-Means clustering algorithm because it has the advantages of less time complexity and high computing efficiency, is efficient in large datasets, and is very simple to study and run [41].

Our method focused on the application of the K-means clustering algorithm to the preprocessed data for grouping and understanding usage patterns. Figure 3-1 shows the flow chart for the K-means algorithm.

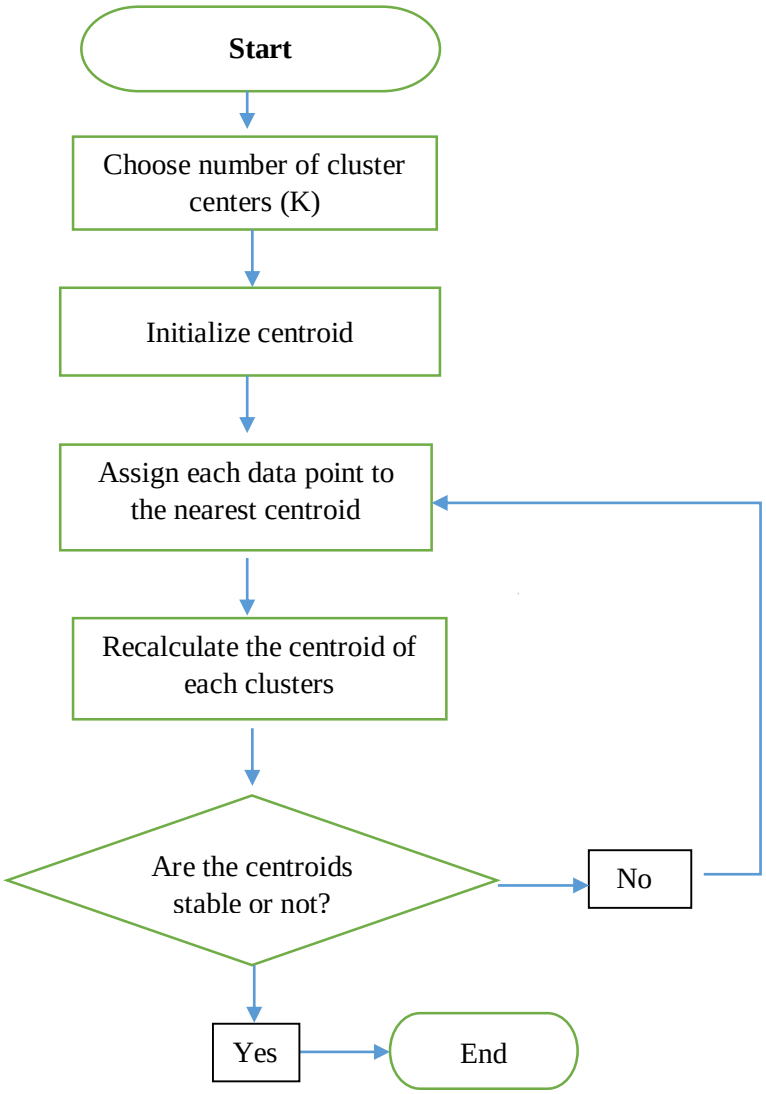

Figure 3-1: Flow chart of K-means algorithm

Voice, data, and voice plus data usage are used individually in a daily, weekly, and monthly way to group customers. To minimize the number of features on a daily plan, we used one month of data, and two months of data are used for the weekly and monthly plans.

The features for daily voice and data usage are from day 1 up to day 30 in minutes and megabits, respectively; similarly, the features for weekly voice and data usage are from week 1 up to week 9 in minutes and megabits, respectively. We had 9 weeks because we used a 31-day dataset. For monthly voice and data usage, month 1 and month 2 usage in minutes and megabits were used. Table 3.1 shows the data features given to the K-means algorithm.

Table 3-1:Data features given for K-means algorithm

Usage plan	Features	
Daily	Service number	Day1 usage - Day31 usage
Weekly	Service number	Week1 usage –Week9 usage
Monthly	Service number	Month1 usage - Month2 usage

If one has sufficient knowledge of the domain area, it will be easier to determine the optimal number of clusters. The number of clusters should not be too small or too large; if it's too large, it makes it difficult to interpret and understand the pattern in the data, and similarly, it will be computationally expensive since the algorithm requires more computational resources to maintain and update the centroid for each cluster. The number of clusters should also not be too small because the algorithm may group data points that are actually distinct and belong to an underlying subgroup, and data points within each cluster may be spread out widely.

Different methods, like elbow and silhouette scores, are used to determine the optimal number of clusters, i.e., the K value. For our study, we used the elbow method to determine the number of clusters for voice, data, and voice plus data services.

The elbow method determines the optimal number of clusters as a function of the distortion or cost of the k-means algorithm as a function of the number of clusters.

First, we apply the K-means algorithm to our dataset (voice, data, and voice plus data) for different values of K, then calculate distortion, or within-cluster sum squares, which is the sum square distance between each data point and its assigned centroid within a cluster, and lastly, plot the distortion. The plot will contain the location of the elbow point with a dashed line, as shown in Figure 3-2, which helps to decide the number of clusters very easily.

Figure 3-2 is the elbow plot or graph for daily data usage. Based on the graph, the optimal number of clusters for daily data usage is 6.

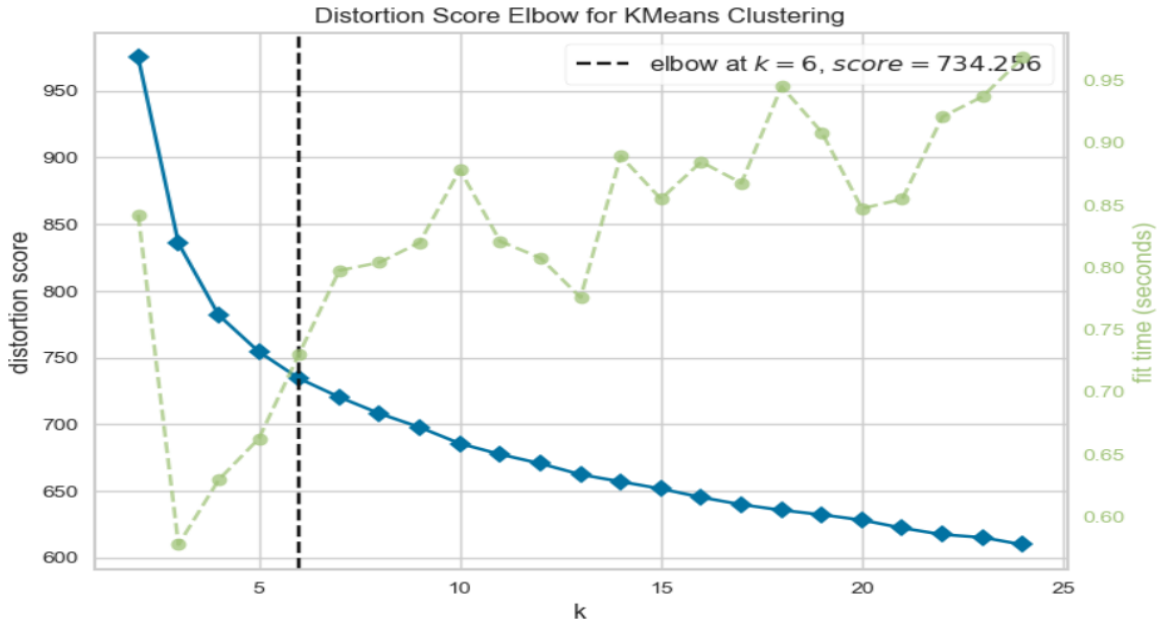

Figure 3-2: Elbow Method for daily data usage

Table 3-2 summarizes the number of clusters in voice, data and voice plus data services.

Table 3-2:Number of clusters based on elbow method

Services		K value
Total Voice Usage	Daily	3
	Weekly	5
	Monthly	5
Data Usage	Daily	6
	Weekly	5
	Monthly	6
Voice Plus Data usage	Daily	6
	Weekly	7
	Monthly	5

3.4.2 Clustering analysis

After we grouped customers based on their usage, we calculated the average value for each group and plotted it to see the usage distribution with reference to the average values. We used Microsoft Excel to do both the plotting and calculation of average values.

3.4.2.1 Voice usage clusters

On voice usage, we had daily voice usage cluster, weekly voice usage clusters and monthly voice usage cluster.

1. **Daily voice clusters:** The daily voice usage clusters had a total of 4768 customers. Based on the elbow method, we have three groups. Table 3-3 shows the number of customers in each group and the average usage per minute for each group.

Table 3-3:Number of customers and average minute usage for daily voice cluster

Groups	Number of customers in a group	Average Voice usage in minute
G1	2312	0.8 min
G2	1696	2.39 min
G3	759	15.64 min

2. **Weekly Voice clusters**: The weekly voice usage clusters had a total of 6076 customers. Based on the elbow method, we have five groups. Table 3-4 shows the number of customers in each group and the average usage per minute for each group.

Table 3-4: Number of customers and average minute usage for weekly voice cluster

Groups	Number of customers in a group	Average Voice usage in minute
G1	3515	3.8 min
G2	330	205 min
G3	1056	43 min
G4	542	143 min
G5	633	90 min

3. **Monthly Voice cluster:** The monthly voice usage clusters had a total of 6882 customers. Based on the elbow method, we have five groups. Table 3-5 shows the number of customers in each group and the average usage per minute for each group.

Table 3-5: Number of customers and average minute usage for monthly voice cluster

Groups	Number of customers in a group	Average Voice usage in minute
G1	3943	18 min
G2	434	1068 min
G3	736	462 min
G4	611	735 min
G5	1158	215 min

3.4.2.2 Data usage clusters

On data usage, we had daily data usage clusters, weekly data usage clusters, and monthly data usage clusters.

a. **Daily Data clusters:** The daily data usage clusters had a total of 6253 customers. Based on the elbow method, we have six groups. Table 3-6 shows the number of customers in each group and the average usage in megabits for each group.

Table 3-6: Number of customers and average megabit usage for daily data cluster

Groups	Number of customers in a group	Average data usage in megabit
G1	5077	4.3MB
G2	70	750MB
G3	246	275MB
G4	123	509MB
G5	134	363MB
G6	603	118MB

b. **Weekly Data clusters:** The daily data usage clusters had a total of 6076 customers. Based on the elbow method, we have six groups. Table 3-7 shows the number of customers in each group and the average usage in megabits for each group.

Table 3-7: Number of customers and average megabit usage for weekly data cluster

Groups	Number of customers in a group	Average data usage in megabit
G1	708	1464MB
G2	162	7007MB
G3	325	3234MB
G4	5473	65MB
G5	183	4624MB

C. Monthly data cluster: The monthly data usage clusters had a total of 7116 customers. Based on the elbow method, we have six groups. Table 3-8 shows the number of customers in each group and the average usage in megabit for each group.

Table 3-8: Number of customers and average megabit usage for monthly data cluster

Groups	Number of customers in a group	Average data usage in megabit	Standard Deviation of the group
G1	5510	316MB	832.5 MB
G2	170	28620MB	7120 MB
G3	173	42669MB	6812 MB
G4	723	6973MB	3681.5 MB
G5	371	363MB	4749 MB
G6	170	26520MB	6906.5 MB

3.4.2.3 Voice plus Data usage clusters

On voice plus data usage, we had daily usage clusters, weekly usage clusters, and monthly usage clusters.

a. **Daily Voice plus Data clusters:** The daily voice plus data usage cluster had totally 4011 customers. Based on the elbow method we had 6 groups. Table 3-9 shows number of customers in each group and average usage in minute and megabit for each group.

Table 3-9:Number of customers and average min/MB usage for daily voice plus data cluster

Groups	Number of customers in a group	Average voice usage of the group (Voice)	Average usage of the group (Data)
G1	1707	0.82 min	9.08 MB
G2	1118	1.22 min	10.56 MB
G3	313	18.8 min	36.8 MB
G4	107	1.45 min	448 .4 MB
G5	695	8.63 min	21.25 MB
G6	71	17.56 min	434.49 MB

b. **Weekly voice plus data clusters:** The weekly voice plus data usage cluster had totally 5819 customers. Based on the elbow method we had 7 groups. Table 3-10 shows number of customers in each group and average usage in minute and megabit for each group.

Table 3-10: Number of customers and average min/MB usage for weekly voice plus data cluster

Groups	Number of customers in a group	Average voice usage of the group (Voice)	Average usage of the group (Data)
G1	687	115 min	406 MB
G2	3203	5 min	65 MB
G3	160	6 min	6527 MB
G4	164	174 min	4799 MB
G5	278	10 min	2934 MB
G6	996	54 min	152 MB
G7	432	190 min	631 MB

c. **Monthly Voice plus data cluster:** The monthly voice plus data usage cluster had totally 6714 customers. Based on elbow method we had 5 groups. Table 3-11 shows number of customers in each group and average usage in minute and megabit for each group.

Table 3-11: Number of customers and average min/MB usage for monthly voice plus data cluster

Groups	Number of customers in a group	Average voice usage of the group (Voice)	Average usage of the group (Data)
G1	4173	49 min	887 MB
G2	1264	490 min	1878 MB
G3	425	46 min	34474 MB
G4	660	1078 min	4137 MB
G5	192	962 min	29156 MB

3.5 Package Assignment(Labeling)

We used Microsoft Excel to create plots that helped us decide how to assign packages to each group. We wanted to give multiple packages to each cluster, considering the mobile package choices from Ethiotelecom and our grouped data. By looking at how users' usage was distributed below and above the average usage for each group, we create a rule.

When the average usage values of each cluster fall below or exceed the available package range, we decide to assign two packages. If the average usage values lie between the minimum and maximum available package values, we assign three packages to each group. The rule first identifies the minimum and maximum values among the available packages provided by Ethiotelecom for each service. The minimum value is the smallest value from the available packages, and the maximum value is the highest value from the available packages. We didn't consider unlimited packages as the maximum value because not all mobile packages offer unlimited packages. The first package is selected close to the average usage value to benefit customers in terms of cost. The second and third packages are determined by choosing the first numbers from the available package options that are lower and higher than the value of the first assigned package, respectively. The rule is stated as follows:

If Average usage value less than minimum available package

P1: The first nearest package to the average usage value

P2: The package is one step greater than **P1.**

Else IF

If average usage value in between minimum available package and maximum available package

P1: The first nearest package to the average usage value

P2: The package is one step less than **P1**

P3: The package is one step greater than **P1**

Else

Average usage value is in between the maximum available package and Unlimited package

P1: The first nearest package to the average usage value

P2: The package is one step less than **P1**

End

Based on our rule, let's discuss how the package assignment works with examples. Due to the plenty of voice and data service options available in each plan (daily, weekly, and monthly), a random group was selected to showcase the labeling that has been implemented. We selected group 2 from daily voice usage, group 1 from weekly voice usage, and group 3 from monthly voice usage.

From the daily voice service, we chose group 2. The group's average voice usage was 2.39 minutes. This implies that the majority of the customers' usage distribution is centered around 2.39 minutes.

On daily voice service, Ethiotelecom offers a daily voice package of 15 minutes, 25 minutes, 50 minutes, and daily unlimited. So based on our rule, the minimum and maximum values are 15 minutes and 50 minutes, respectively. According to the rule, the first top-recommended package is the first available package that is near to the average usage value of the customers, which is 15 minutes; for those customers who have used greater than 2.39 minutes, the recommended package is 25 minutes. After assigning those packages, we gave the package label name P1.

Similarly, from weekly voice usage, we have selected group 2. The analysis shows that the majority of customers in this group used voice services for a total of 205 minutes. The available weekly voice packages are 75 minutes, 110 minutes, 140 minutes, and 1320 flexi units (220 minutes). The minimum value from the available package is 75 minutes, and the maximum value is 1320 flexi units (220 minutes). The first available package near 205 minutes is 220 minutes, which is the maximum value from the available packages, and the first nearest lower

value to 205 minutes is 140 minutes. Therefore, the recommended packages for these groups are weekly 1320 flexi units (220 minutes) and 140 minutes with the package label as P4.

Monthly voice package subscribers in Group 3 have an average voice usage of 462 minutes. Among the available monthly voice packages: 125 minutes, 220 minutes, 380 minutes (2280 Monthly Flexi Unite), 400 minutes, 500 minutes, 1400 minutes, 2000 minutes, and Monthly Unlimited, the closest option to the average usage of 462 minutes is 500 minutes. The first top-recommended package is 500 minutes. The second recommended package is one step lower than 500 minutes, which is 400 minutes, and the third recommended package is one step bigger than 500 minutes, which is 1400 minutes. The package label is P9.

In the case of data usage on a daily basis, we have selected Group 1. The average data usage in megabits was 4.3 MB. By referring to the available daily data packages, the first available package near 4.3 megabits is 50 MB, which is the first recommended package for this group. The value greater than the first recommended package, i.e., 50 MB, is 100 MB. Therefore, the recommended packages for these groups are 50 MB and 100 MB, with the package label P11.

On monthly data usage, for group three customers, the average usage value is 42669 MB. The available package value is close to the average usage value of 50 GB, and it is assigned as the first recommended package. The second and third assigned packages will have package values that are one step lower and one step higher than 50 GB, which are 20 GB and 100 GB, respectively. The package label for these groups is P19.

The same concepts apply to the rest of the groups. Tables 3-12, 3-13, and 3-14 summarize the assigned packages and label names for each group of voice and data services.

Table 3-12: Daily and weekly voice labeled packages

Daily Voice Recommended packages	
Package Label	**Package Description**
P1 **(AVG-0.8 min)**	Top Daily Voice Packages 1.15 minutes 2.25 minutes
P1 **(AVG-2.39 min)**	Top Daily Voice Packages 1.15 minutes 2.25 minutes
P2 **(AVG-15.59 min)**	Top Daily Voice Packages 1.25 minutes 2.15 minutes 3.50 minutes

Weekly Voice Recommended packages	
Package Label	**Package Description**
P3 **(AVG-3.8 min)**	Top Weekly Voice Packages 1.75 minutes 2.110 minutes
P4 **(AVG-205 min)**	Top Weekly Voice Packages 1.1320 weekly flexi unite 2.140 minutes
P3 **(AVG-43 min)**	Top Weekly Voice Packages 1.75 minutes 2.110 minutes
P5 **(AVG-143 min)**	Top Weekly Voice Packages 1.140 minutes 2.110 minutes 3.1320 weekly flexi unite
P6 **(AVG-90 min)**	Top Weekly Voice Packages 1.110 minutes 2.75 minutes 3.140 minutes

Table 3-13: Monthly voice and daily data labeled packages

Monthly Voice Recommended packages	
Package Label	**Package Description**
P7 **(AVG-18 min)**	Top Monthly Voice Packages 1.125 minutes
P8 **(AVG-1068 min)**	Top Monthly Voice Packages 1.1400 minutes 2.500 minutes 3.2000 minutes
P9 **(AVG-462 min)**	Top Monthly Voice Packages 1.500 minutes 2. 400 minutes 3. 1400 minutes
P9 **(AVG-735 min)**	Top Monthly Voice Packages 1.500 minutes 2.400 minutes 3.1400 minutes
P10 **(AVG-215 min)**	Top Monthly Voice Packages 1.220 minutes 2. 125 minutes 3. 2280 Monthly flexi unite

Daily data Recommended packages	
Package Label	**Package Description**
P11 **(AVG-4.3 MB)**	Top daily data Packages 1.50 MB 2.100 MB
P12 **(AVG-759 MB)**	Top daily data Packages 1.600 MB 2.250 MB 2.Daily unlimited
P13 **(AVG-275 MB)**	Top daily data Packages 1.250 MB 2.390 flexi unit 3.600 MB
P12 **(AVG-509 MB)**	Top daily data Packages 1.600 MB 2.250 MB 2.Daily unlimited
P13 **(AVG-363 MB)**	Top daily data Packages 1.250 MB 2.390 flexi unit 3.600 MB
P14 **(AVG-118 MB)**	Top daily data Packages 1.100 MB 2. 50 MB 3.390 flexi unit

Table 3-14: Weekly and Monthly data labeled package

Weekly data Recommended packages	
Package Label	**Package Description**
P15 **(AVG-1464 MB)**	Top Weekly Data Packages 1.1.5 GB 2.1 GB 3. Weekly unlimited data
P16 **(AVG-7007 MB)**	Top Weekly Data Packages 1. Weekly unlimited data 2. 1.5 GB
P17 **(AVG-3234 MB)**	Top Weekly Data Packages 1. 1.5 GB no expire 2. Weekly unlimited data
P18 **(AVG-65 MB)**	Top Weekly Data Packages 1.375 MB 2.390 flexi unit
P16 **(AVG-4624 MB)**	Top Weekly Data Packages 1. Weekly unlimited data 2. 1.5 GB

Monthly data Recommended packages	
Package Label	**Package Description**
P19 **(AVG-26520 MB)**	Top Monthly Data Packages 1.20 GB 2.10 GB 2.50 GB
P19 **(AVG-15957 MB)**	Top Monthly Data Packages 1.20 GB 2.10 GB 2.50 GB
P20 **(AVG-6073 MB)**	Top Monthly Data Packages 1.4 GB 2. 2 GB 3.10 GB
P21 **(AVG-42669 MB)**	Top Monthly Data Packages 1.50 GB 2.20 GB 3.100 GB
P19 **(AVG-28620 MB)**	Top Monthly Data Packages 1.20 GB 2.10 GB 2.50 GB
P22 **(AVG-316 MB)**	Top Monthly Data Packages 1.500 MB 2.1GB

When we see the usage patterns of voice plus data customers, most of the groups usage patterns don't match the available voice plus data packages.

For example, let's see the daily voice and data usage patterns of Group 4 customers. Customers in Group 4 had average voice usage of 1.45 minutes and 435 MB of data. The available voice plus data packages are 15 minutes + 160 MB and 50 minutes + 600 MB. Based on our rule, the nearest available package to the average voice usage for these groups is 15 minutes, but the average data usage for these groups is 435 MB, which is far from the available data package of 160 MB. So the available package is far from meeting customer demand.

Customers in Group 6 used an average of 434.9 MB of data and 17.56 minutes of voice time. The nearest available package to the average voice usage for these groups is 15 minutes, but the average data usage for these groups is 435 MB, which is far from the available data package of 160 MB. So the available package is far from meeting customer demand.

The same analysis went for weekly and monthly voice and data usage. From weekly voice plus data usage groups 2, 3, 5, and 6, they had average voice usage of 5, 6, 10, and 54 minutes, and the average data usage was 65, 6527, 2934, and 152 MB. The available weekly voice plus data

packages are 150 minutes plus 600 MB and 250 minutes plus 11 GB. The usage pattern of the customers and the available packages are still far apart.

Generally, we can simply observe by seeing Table 3-15, which shows the average usage and available packages of voice and data service for each group and plan. As we can see from Tables 3–15, most of the groups usage patterns are different from the available voice and data packages, so we didn't make a recommendation.

Table 3-15: Summarized average voice and data usage

Used plan	Groups	Average used voice plus data	Available voice plus data package
Daily	1	0.82 min + 9.08 MB	15 min + 160 MB 50 min + 600 MB
	2	1.22 min + 10.56 MB	
	3	18.8 min + 36.8 MB	
	4	1.45 min + 448 .4 MB	
	5	8.63 min + 21.25 MB	
	6	17.56 min + 434.49 MB	
Weekly	1	115 min + 406 MB	150 min + 600 MB 250 min + 1 GB
	2	5 min + 65 MB	
	3	6 min +6527 MB	
	4	174 min + 4799 MB	
	5	10 min + 2934 MB	
	6	54 min + 152 MB	
	7	190 min + 631 MB	
Monthly	1	49 min + 887 MB	200 min + 1 GB 400 min + 1 GB 500 min + 1 GB 350 min + 2 GB 500 min + 5 GB 1300 min + Unlimited Data
	2	490 min + 1878 MB	
	3	46 min + 34474 MB	
	4	1078 min + 4137 MB	
	5	962 min + 29156 MB	

3.6 CLASSIFICATION MODEL BUILDING

Apart from recommending packages, the classification model also serves as a tool for evaluating the initial phase of our project. This model enables the company to quickly categorize its customers by determining if their raw usage data matches the usage patterns identified through our clustering algorithm.

To construct the classification model, we employed two machine learning algorithms: random forest and k-nearest neighbor (KNN) classifiers. Both of these algorithms are well suited for multi-label classification tasks.

3.6.1 Random Forest Classification algorithm

The random forest algorithm is a supervised machine learning technique that is mostly applicable to regression and classification problems. It has the ability to handle overfitting and complex data sets with continuous variables.

The main parameter in the random forest algorithm is n_estimators," which refers to the number of decision trees to be used in the ensemble. Determining the optimal value for this parameter depends on a specific dataset and the trade-off between model complexity and performance. Increasing the number of estimator values will increase the complexity and computational time of the model.

To build a model using the random forest algorithm, subsets of data points and sub features are selected for the construction of the decision tree. That means n random records and M features are taken from a dataset that has K records [42].

3.6.2 K nearest neighbor classification algorithm

The K-Nearest Neighbor Classification Algorithm is a supervised machine learning algorithm that is simple and easy to implement and can be used to solve both classification and regression problems.

The k-nearest neighbors (KNN) classification algorithm has several key parameters, including the number of neighbors (k) and the distance metric. The choice of k can significantly impact the algorithm's performance, with small values making the model sensitive to noise and large values leading to over smoothing and misclassification. The choice of distance metric depends on the nature of the data, with Euclidean distance being suitable for continuous numerical features and Manhattan distance for categorical or binary features. The choice of distance

metric depends on the nature of the data, and feature scaling should be considered to ensure fair comparisons between features.

We have used the sklearn library, the most powerful library in Python to build a classification model. We have 31 features for daily base usage, 10 features for weekly base usage, and 3 features for monthly base usage for both data and voice services to build a classification model.

3.7 EVALUATION METRICS

As mentioned in Chapter 3, our work consisted of two distinct phases. The initial phase entailed the use of a clustering algorithm to effectively group customers according to their usage preferences. The subsequent phase involved the recommendation and evaluation of associated packages through the implementation of classification algorithms. An appropriate evaluation metric is required to accurately assess the efficiency of the grouping mechanism as well as the performance of the system in generating recommendations that are relevant to user needs. Below are the evaluation metrics that are used:

3.7.1 Silhouette score

The silhouette score is an evaluation method used to measure how well data points are grouped. It is used to measure and validate clustering. The values of the silhouette score range from -1 to 1. When the values approach one, this indicates that the data points within a given cluster exhibit remarkable similarity among themselves and are quite distinct from other clusters.

3.7.2 Precession

Precision in recommendation systems is the measurement of important packages among the recommended packages. In other words, it calculates how many of the recommended packages are actually useful or relevant to the user. Precision is often used to assess the accuracy of the system.

$$\textbf{Precision} = \frac{(\textbf{Number of relevant packages recommended})}{(\textbf{Total number of recommended packages})} \quad (3.2)$$

3.7.3 Recall

In recommendation systems, recall is the measurement of the number of important packages that were successfully recommended among all the important packages. It helps determine how well the recommendation system can retrieve all the important packages for a user.

$$\textbf{Recall} = \frac{(\textbf{Number of relevant package recommended})}{(\textbf{Total number of relevant package})} \quad (3.3)$$

3.7.4 F1-score

The F1 score evaluates how effectively the system identifies relevant items for a user. It combines precession and recall. The F1 score is the harmonic mean of precision and recall, providing a single value that balances the trade-off between precision and recall [43].

$$\textbf{F1 Score} = \frac{\mathbf{2} * (\textbf{Precision} * \textbf{Recall})}{(\textbf{Precision} + \textbf{Recall})} \quad (3.4)$$

A higher F1 score indicates a more balanced performance in suggesting relevant items while minimizing both false positives and false negatives.

For our study, we evaluated our classification models based on precession, recall, and F1 Score, but to compare them, we used F1 Score since it tells the accuracy of the recommendations made by the system and how well it balances precision and recall.

We prepared a relationship between the usage pattern of the customers and the available package offer by Ethiotelecom, and we also assigned the packages in the first phase of our experiment. The second phase of our experiment is building a classification model based on the assigned labels and usage patterns of the groups.

Since we have different groups in voice and data services, we have built separate models for each service and group using a usage pattern for each group and an attached label. To train the model, we used a separate test data model training technique. In these training techniques, there is no standard dataset ratio for testing and training, but the divided dataset should be enough for both testing and training. Therefore, we split the dataset into 80% of the data set for training purposes and 20% of the data set for testing purposes.

3.8 DATA SERVICE CLASSIFICATION

The number of customers in each group for the daily, weekly, and monthly plans is different for data services too. Three models for daily data usage, weekly data usage, and monthly data usage were built.

For daily data usage, we had 6 groups with the attached packages, so to build a daily data classification model, we used labels as one feature and data usage from day 1 up to day 31 as another feature.

Group 1 of daily data usage had 2312 customers, and the package label assigned for it was P11; group 2 of data usage had 1697 customers, and the assigned package was P12; group 3 of data usage had 759 customers, and the assigned package was P13; group 4 of data usage had 759

customers, and the assigned package was P12; group 5 of data usage had 759 customers, and the assigned package was P13; and similarly, group 6 of data usage had 759 customers, and the assigned package was P14. Table 3-16 summarizes the data that was given to the RF and KNN algorithms and builds a classification model for daily data usage.

Table 3-16: Data used to build daily data classification model using RF and KNN

Service	Labels (Package name)	Group	Number of customer on a group	Features
Daily Data usage	P11	G1	5077	P11,D1-D31 usage
	P12	G2	70	P12, D1-D31 usage
	P13	G3	246	P13, D1-D31 usage
	P12	G4	123	P12, D1-D31 usage
	P13	G5	134	P13, D1-D31 usage
	P14	G6	603	P14, D1-D31 usage

For weekly data usage, we had 5 groups with the attached packages, so to build a weekly data classification model, we used labels as one feature and data usage from week 1 up to week 9 as another feature.

Group 1 of weekly data usage had 708 customers, and the package label assigned for it was P15; group 2 of data usage had 162 customers, and the assigned package was P16; group 3 of data usage had 325 customers, and the assigned package was P17; group 4 of data usage had 5473 customers, and the assigned package was P18; and similarly, group 5 of data usage had 183 customers, and the assigned package was P16. Table 3-17 summarizes the data that was given to the RF and KNN algorithms and builds a classification model for weekly data usage.

Table 3-17: Data used to build weekly data classification model using RF and KNN

Service	Labels (Package name)	Group	Number of customer on a group	Features
Weekly Data usage	P15	G1	708	P15,W1-W9 usage
	P16	G2	162	P16, W1-W9 usage
	P17	G3	325	P17, W1-W9 usage
	P18	G4	5473	P18, W1-W9 usage
	P16	G5	183	P16, W1-W9 usage

For monthly data usage, we had 6 groups with the attached packages, so to build a monthly data classification model, we used labels as one feature and month 1 and month 2 data usage as another feature.

Group 1 of monthly data usage had 5510 customers, and the package label assigned for it is P19; group 2 of data usage had 170 customers, and the assigned package was P19; group 3 of data usage had 173 customers, and the assigned package was P20; group 4 of data usage had 723 customers, and the assigned package was P21; group 5 of data usage had 371 customers, and the assigned package was P19; and similarly, group 6 of data usage had 170 customers, and the assigned package was P22. Table 3-18 Summaries the data that was given to the RF and KNN algorithms and builds a classification model for monthly data usage.

Table 3-18: Data used to build monthly data classification model using RF and KNN

Service	Labels (Package name)	Group	Number of customer on a group	Features
Monthly Data usage	P19	G1	5510	P19, M1-M2 usage
	P19	G2	170	P19, M1-M2 usage
	P20	G3	173	P20, M1-M2 usage
	P21	G4	723	P21, M1-M2 usage
	P19	G5	371	P19, M1-M2 usage
	P22	G6	170	P22 , M1-M2 usage

3.9 VOICE SERVICE CLASSIFICATION

On voice service, we had different numbers of records or customers in each group based on the daily, weekly, and monthly plans. Therefore, we built three models for daily voice usage, weekly voice usage, and monthly voice usage.

For daily voice usage, we have segmented our customers into three distinct groups, which have been derived using the K-Means clustering technique. For each group, we assigned the appropriate package, as discussed in the package labeling section. So to build a daily voice classification model, we have used labels as one feature and voice usage from day 1 up to day 31 as another feature.

Group 1 of daily voice usage had 2312 customers, and the package labeled for it was P1. Group 2 of voice usage had 1697 customers, and the assigned package was P1, and similarly, Group 3 of voice usage had 759 customers, and the assigned package was P2. Table 3-19 summarizes the data that was given to the RF and KNN algorithms and builds a classification model for daily voice usage.

Table 3-19: Data used to build daily voice classification model using RF and KNN

Service	Labels (Package name)	Group	Number of customer on a group	Features
Daily Voice usage	P1	G1	2312	P1, D1-D31 usage
	P1	G2	1697	P1, D1-D31 usage
	P2	G3	759	P2, D1-D31 usage

For weekly voice usage, we have five groups: Group 1 of weekly voice usage had 3515 customers, and the package label assigned for it was P3. Group 2 of weekly voice usage had 330 customers, and the assigned package was P4. Group 3 of weekly voice usage had 1056 customers, and the assigned package was P3. Group 4 of weekly voice usage had 542 customers, and the assigned package was P5. Similarly, Group 5 of weekly voice usage had 633 customers, and the assigned package was P6. Table 3-20 shows the data that was given to the RF and KNN algorithms and builds a classification mode for weekly voice usage.

Table 3-20: Data used to build weekly voice classification model using RF and KNN

Service	Labels (Package name)	Group	Number of customer on a group	Features
Weekly Voice usage	P3	G1	3515	P3, W1-W9 usage
	P4	G2	330	P4, W1-W9 usage
	P3	G3	1056	P3, W1-W9 usage
	P5	G4	542	P5, W1-W9 usage
	P6	G5	633	P6, W1-W9 usage

For monthly voice usage, we have 5 groups: group 1 of monthly voice usage had 3943 customers, and the package label assigned for it was P7; group 2 of monthly voice usage had 434 customers, and the assigned package was P8; group 3 of monthly voice usage had 736 customers, and the assigned package was P9; group 4 of monthly voice usage had 611 customers, and the assigned package was P9; and similarly, group 5 of monthly voice usage had 1158 customers, and the assigned package was P10. Tables 3-21 summarize the data that was given to the RF and KNN algorithms and build a classification mode for monthly voice usage.

Table 3-21: Data used to build monthly voice classification model using RF and KNN

Service	Labels (Package name)	Group	Number of customer on a group	Features
Monthly Voice usage	P7	G1	3943	P7, M1-M2 usage
	P8	G2	434	P8, M1-M2 usage
	P9	G3	736	P9, M1-M2 usage
	P9	G4	611	P9, M1-M2 usage
	P10	G5	1158	P10, M1-M2 usage

4 RESULT AND DISCUSSION

The evaluation matrix used for the K-means algorithm is the silhouette score, which measures how well the clustering is done. Similarly, for evaluating classification models, precision, recall, and F1 scores are used.

Precision measures the relevance of the recommended packages. It indicates, on average, how much of the recommended packages are relevant to the users. On the other hand, recall on the recommender system measures the proportion of relevant packages that are successfully recommended. It determines how well the recommender system retrieves all relevant packages for the customers.

F1 Score is the harmonic mean of precision and recall; it balances the trade-off between precession and recall. To compare our models, we used F1 Score.

4.1 K MEANS CLUSTERING RESULT

The k-means algorithm was used to group customers based on their usage patterns. Grouping was done for voice, data, and voice plus data services on a daily, weekly, and monthly basis. We had 3 groups for daily voice usage customers and 5 groups for both weekly and monthly voice usage customers. On the weekly and monthly plans for data services, we had 5 groups and 6 groups, respectively. Similarly, for voice plus data, we had 6 groups on the daily plan, 7 groups on the weekly plan, and 5 groups on the monthly plan. So we will have individual silhouette scores to measure the grouping.

By averaging the scores from each group in each plan, we were able to better understand and evaluate k-means clustering. The results of the k-mean algorithm are shown in Table 4-1 for each service and plan.

Table 4-1:K means Result

Services		Average Silhouette Score
Total Voice Usage	Daily	0.41
	Weekly	0.55
	Monthly	0.64
Data Usage	Daily	0.74
	Weekly	0.75
	Monthly	0.80
Voice plus data	Daily	0.31
	Weekly	0.51
	Monthly	0.64

The silhouette score explains how customers in one group are similar within the group and how these customers are different from other groups in their usage patterns. Generally, values greater than 0.5 are considered good clustering. This means users in that group are similar in their voice, data, and voice plus data usage to those in other groups.

The grouping method employed for both total voice usage and voice plus data usage has demonstrated efficiency across both weekly and monthly plans. Specifically, it achieved an average silhouette score of 0.55 for total voice usage in weekly plans and 0.64 for total voice usage in monthly plans. Additionally, for voice plus data usage, the method yielded silhouette scores of 0.51 for weekly plans and 0.64 for monthly plans. These scores affirm the effective grouping of customers with similar patterns in their voice service usage and voice plus data service usage.

Customers that have the same data usage pattern are grouped efficiently, with an average silhouette score of 0.74, 0.75, and 0.80 on the daily, weekly, and monthly plans, respectively.

When we see clustering of the daily total usage and voice plus data usage, it is 0.41 and 0.31, which is low compared to weekly and monthly voice usage clustering and voice plus data usage clustering, respectively. This is due to the properties of the data set from the source. Most of the customers didn't use voice service day-to-day, and the usage property of the customers on a day-to-day plan is almost zero. This will affect clustering voice and voice plus data service usage.

4.2 CLASSIFICATION RESULT

After grouping customers, cluster analysis and package labeling were done. RF and KNN classification algorithms were used to make package recommendation systems. As mentioned in Section 1, since the available voice plus data package and the usage pattern of customers didn't match, the mobile package recommendation was done for voice and data services only.

Tables 4-2 and 4-3 show the RF and KNN classification model results for both voice and data services.

Table 4-2:RF and KNN model results for voice service

Algorithm	Evaluation Metrix	Daily Voice	Weekly Voice	Monthly Voice
RF	Precession	0.978	0.934	0.994
	Recall	0.929	0.926	0.997
	F1-Score	0.951	0.929	0.996
KNN	Precession	0.973	0.947	0.994
	Recall	0.904	0.939	0.993
	F1-Score	0.934	0.943	0.993

Table 4-3:RF and KNN model results for data service

Algorithm	Evaluation Metrix	Daily Data	Weekly Data	Monthly Data
RF	Precession	0.896	0.906	0.975
	Recall	0.845	0.892	0.926
	F1-Score	0.869	0.898	0.949
KNN	Precession	0.914	0.911	0.972
	Recall	0.793	0.897	0.950
	F1-Score	0.843	0.904	0.960

4.3 DISCUSSION

4.3.1 Data service classification

On the data service, we had groups of daily data usage, weekly data usage, and monthly data usage. On each plan, we assigned a package label for each individual group, as shown in Tables 3-12, 3-13, and 3-14, so that our model would learn from it and make recommendations based on it. Figure 4-1 shows the data usage classification results of our models.

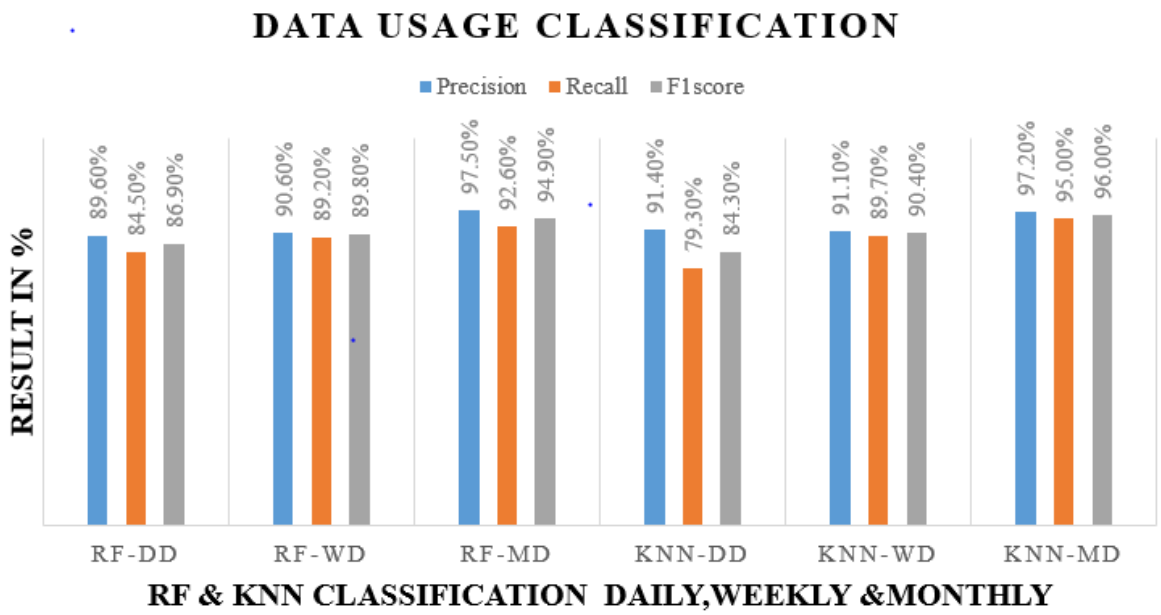

Figure 4-1:Data usage classification result

In the analysis of daily data usage, we categorized customers into six groups, each associated with specific package options (P11–P14). These packages were considered relevant for the customers of all six groups. Our model, created using RF and KNN algorithms, learned the usage patterns of these groups. The RF model demonstrated a recall of 84.5%, indicating that it accurately recommended relevant packages for 84.5% of the grouped users in the daily data usage group. It also achieved a precision of 89.6% and an F1 score of 86.9%. In comparison, the KNN model achieved a recall of 79.3%, a precision of 91.4%, and an F1 score of 84.3% for the same groups.

For weekly data usage, including 5 groups with package labels (P15–P18), the RF algorithm achieved a precision of 90.6% and a recall of 89.2%, while the KNN model had a precision of 91.1% and a recall of 89.7%. 89.2% and 89.7% of relevant packages are successfully recommended by RF and KNN models, respectively, for customers in weekly data usage groups. On average, 90.6% and 91.1% of the recommended packages are important for customers in these groups.

Turning to monthly data usage with six groups labeled as (P19–P22), the RF algorithm achieved a recall of 92.6% and a precision of 97.5%, while the KNN model had a recall of 95.0% and a precision of 97.2%.

In summary, KNN outperformed RF for weekly and monthly plans in terms of F1 scores of 90.4% and 96%, respectively, whereas RF outperformed for daily plans with an F1 score of

86.9%. This shows the strengths of each algorithm in different usage scenarios within the data service domain.

4.3.2 Voice service classification

Daily, weekly and monthly voice usage was grouped and for each individual group appropriate package was attached.

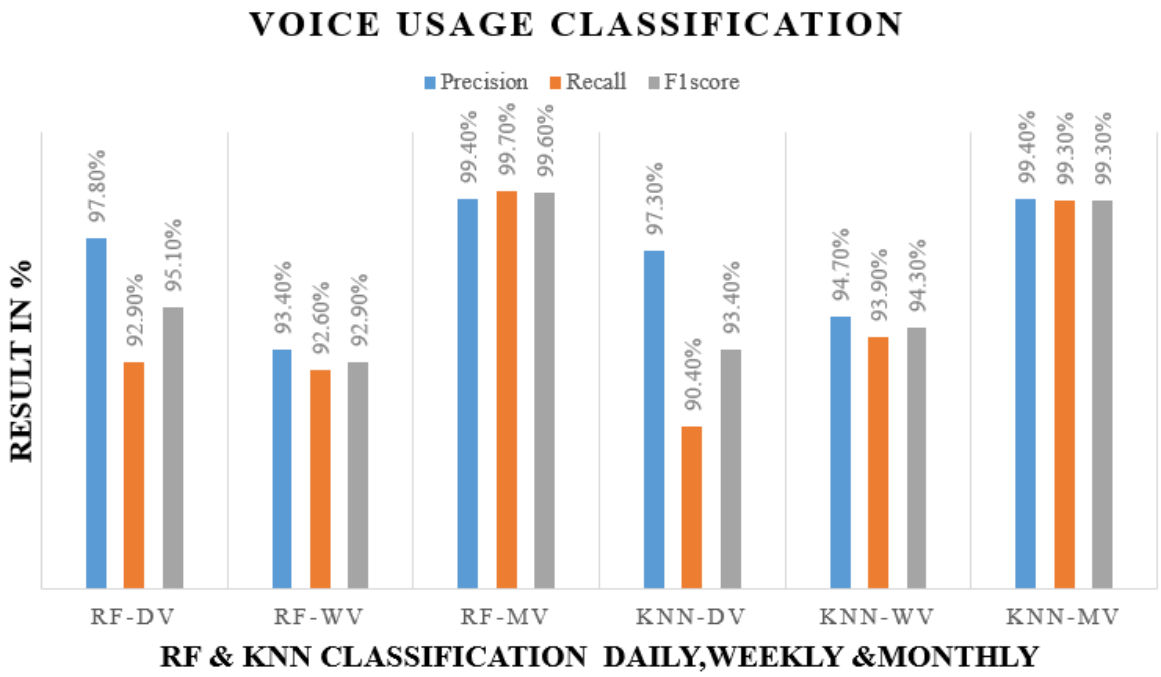

Figure 4-2: Voice usage classification result

In the domain of analyzing voice usage patterns, we categorized customers into distinct groups based on their usage habits, assigning specific package labels to each group. Our goal was to create classification models using both the RF and KNN algorithms to determine the most suitable package choices for customers in each group.

For daily voice usage, we divided customers into three groups labeled P1 and P2. The RF model achieved a recall of 92.90%, indicating that it accurately recommended relevant packages (P1–P2) for 92.90% of the grouped users in the daily voice usage group. Similarly, its precision was 97.80%, implying that 97.80% of the packages it suggested were appropriate for these customers. On the other hand, the KNN model achieved a slightly lower recall of 90.04% and a precision of 97.30%.

Moving to weekly voice usage with five categories labeled as (P3–P6), both RF and KNN models excelled. The RF model achieved a recall of 92.60%, indicating that it accurately recommended relevant packages (P3–P6) for 92.60% of the grouped users in the weekly voice

usage group. Similarly, its precision was 93.40%, implying that 93.40% of the packages it suggested were appropriate for these customers. On the other hand, 93.90% of the grouped users under weekly voice usage group accurately recommended relevant packages by KNN model. Similarly, its precision was 94.70%, implying that 94.70% of the packages it suggested were appropriate for these customers.

For monthly voice usage, including five groups labeled P7–P10, both RF and KNN models performed exceptionally well, with a precession of 99.40%. The recall values were also remarkable, with RF at 99.7% and KNN at 99.30%. This indicates their accuracy in recommending 99.7% and 99.30% of the relevant packages (P7–P10) for these groups.

In summary, RF outperformed KNN with F1 scores of 95.10% and 99.60% for daily and monthly voice usage plans, respectively. Similarly, KNN showcased better performance than RF on the weekly usage plan with an F1 score of 94.30%. This indicates the strengths of each algorithm for different usage scenarios within the voice service domain.

5 CONCLUSION AND FUTURE WORK

5.1 CONCLUSION

Telecommunications operators frequently introduce new products and services to satisfy consumer demands, maintain relevance in an ever-changing market, and capitalize on technological advancements. These services satisfy competitive pressures, customer expectations, and revenue generation. By promoting a culture of innovation and aligning their services with global practices, operators can ensure they remain competitive and meet the evolving needs of their customers. Mobile package services can be mentioned as a new service introduced by operators.

Mobile packages offer various voice, data, and SMS plans to help mobile customers make a voice call and use data or SMS services relative to cheap voice tariffs compared with PayGo usage tariffs.

A mobile package is one of the mobile services offered by Ethiotelecom. Customers can choose from a variety of mobile package options based on their interests, communication needs, and budget constraints. Customers can buy any of the mobile package offers via the ethiotele mobile app, telebirr (app or USSD using 127), or by dialing *999#. The mobile package offer has so many options. However, accessing the mobile package service requires an overflow of information and many clicks. Even if all mobile package offers are very important, they may go unnoticed by all customers. To overcome these issues, we proposed a mobile package service recommendation system.

In order to make a mobile package recommender system, clustering and classification techniques were used by understanding and referencing literature in the area of recommendation systems in two phases. 7899 CDR data points were collected from Ethiotelecom, and then it was preprocessed to make it suitable for the algorithms using Microsoft Excel and Python programming tools.

In the first phase, the preprocessed data was grouped based on a daily, weekly, and monthly basis for voice and data services to understand the usage patterns of the customers using the K-means clustering algorithm, and package labeling was done by aligning the customer's usage patterns with the available mobile package services on Ethiotelecom.

The second phase was building a model using RF and KNN classification algorithms. 80% of the dataset was used for training purposes, and 20% of the dataset was used for testing purposes. The usage pattern of the group data on the first phase with their attached package label was given to the model so that it would learn and recommend accordingly.

The evaluation result shows RF outperformed KNN for daily and monthly voice usage plans in terms of F1 score, whereas KNN excelled for weekly voice usage plans. On the other hand, KNN demonstrated superior performance to RF in terms of F1 score for weekly and monthly data usage plans. Conversely, RF showcased better performance than KNN on F1score for daily usage plans. This indicates the strengths of each algorithm for different usage scenarios within the voice and data service domains.

Finally, the deployment of a mobile package recommendation model is recommended for Ethiotelecom to satisfy and sustain its customers, attract new customers, and increase its revenue.

5.2 FUTURE WORK

In our study, active prepaid customers were randomly selected from Ethiotelecom to recommend mobile package services. Future researchers can enhance the recommendation system by utilizing different data collection mechanisms. Moreover, our study exclusively focuses on one type of service, namely mobile package services offered by telecom operators. However, the same approach can be applied to recommend other types of services, such as value-added services.

6 APPENDIX

CDR Based Recommender System for Mobile Package Service Users

Saba Mulugeta
Addis Abeba University
Addise Abeba Institute of Technology
School of Electrical and Computer Engenering
Addise Abeba, Ethiopia
sabamulugetaa@gmail.com

Sosina Mengestu
Addis Abeba University
Addise Abeba Institute of Technology
School of Electrical and Computer Engenering
Addise Abeba, Ethiopia
sosina.mengestu@aait.edu.et

***Abstract*— Telecom operators continually expand their product offerings in response to growing competition, resulting in a mass of packages. However, customers often remain unaware of these packages, leading to missed opportunities. Capturing mobile package services users interest and preference is difficult since they are offered offline. To address this issue, we propose a recommender system personalized to mobile packages. The system comprises two phases: The first phase is customer usage grouping. We employ the k-means clustering algorithm to group customers based on their usage patterns, determined using the elbow method to establish cluster counts for each service. The second phase was the classification model. We build a classification model using random forest (RF) and k-nearest neighbor (KNN) algorithms with two-month Call Detail Record (CDR) data. Evaluation results show that KNN outperforms RF for weekly and monthly data usage plans, with F1 scores of 90.4% and 96%, respectively. Conversely, RF performs better for daily data plans (F1 score: 86.9%). For voice usage plans, RF outperforms KNN in the daily (F1 score: 95.10%) and monthly (F1 score: 99.60%) plans, while KNN excels in the weekly voice usage plan (F1 score: 94.30%). The algorithm's strengths vary across different usage scenarios within the voice and data services domains.**

Keywords— Recommender system, collaborative filtering, content-based recommender systems, mobile recommender systems, clustering and classification.

I. INTRODUCTION

Telecom operators like Ethiotelecom offer mobile package services, which provide customers with voice, data, and SMS at lower rates. However, customers often face information overload with numerous package options. To address this, a recommendation system is proposed.

A recommendation system analyzes user preferences and interactions to suggest relevant products or services. It aims to simplify choices and enhance the user experience. The basic models for recommender systems work with two kinds of data. The first one is user-item interactions, such as ratings or buying behavior, and the second is attribute information about the users and items, such as textual profiles or relevant keywords [1]. Instead of using historical ratings or buying data, external knowledge bases and constraints are also used to create the recommendation, which is referred to as a knowledge-based recommendation system. Methods that use user-item interactions are referred to as collaborative filtering methods, whereas methods that use attribute information about the users and items are referred to as content-based recommender methods. Some recommendation systems combine these different aspects to create hybrid systems. Hybrid systems can combine the strengths of various types of recommender 7hsystems to create techniques that can perform more robustly in a wide variety of settings [1].

Collaborative filtering makes recommendations by learning from user or item historical interactions, either through explicit (e.g., user's previous ratings) or implicit feedback (e.g., browsing history).Content-based recommendations are based on comparisons across items' and users' information. Hybrid models are recommender systems that integrate two or more types of recommendation strategies [7, 8] .

Most studies have been done on various recommendation systems for different purposes, such as the telecom industry, e-commerce, education, tourism, and so on. However, the studies used data based on ratings, which is not applicable to mobile package service users. Therefore, the main objective of this study is to develop a mobile package service recommendation system to make the service easier for customers .

II. STATEMENT OF THE PROBLEM

Telecom operators, like Ethiotelecom in Ethiopia, offer a variety of mobile packages to satisfy to customer needs and boost revenue. However, the sheer number of packages often results in customers overlooking potentially beneficial options. Traditional recommender systems rely on user-item interactions and attribute information, primarily in web-based settings, making them less suitable for mobile package users who lack rating data and online interactions.

To address this gap, there's a need to develop a recommendation system tailored specifically for mobile package services using Call Detail Record (CDR) data, which telecom operators regularly collect. CDR data can serve as a

valuable resource for understanding subscriber preferences and behavior.

Ethiotelecom, like other operators, offers multiple mobile packages accessible via USSD codes like *999#, Ethiogebeta, Ethiotele mobile app, and Tele Birr. These packages encompass voice, SMS, and data services, but the overwhelming amount of information often leads to customers missing out on relevant packages.

To overcome this issue and enhance customer satisfaction, a recommendation system based on individual customer interests and preferences should be implemented. Such a system could utilize CDR data to provide personalized recommendations, ensuring that customers discover and benefit from the most suitable mobile packages.

III. RELATED WORK

Recommendation systems are information-filtering and decision-support tools that provide product and service recommendations tailored to the user's specific needs and preferences [9]. The recommendation system is applicable in different areas, like tourism [10], e-commerce [11], telecom [12], indoor shopping [13], and insurance [14].

Miyahara et al. [16] used collaborative filtering with two approaches: user-based and item-based. User-based considers user similarity, while item-based focuses on item similarity. However, they used a basic Bayesian classifier with binary data (likes/dislikes) and synthetic data, limiting their results.

To address these limitations, Xiaoyuan Su and Taghi M. Khoshgoftaar [17] proposed a more advanced Bayes classifier for real-world multiclass data. This advanced approach enhances predictions, particularly with incomplete or sparse data, outperforming simpler methods like the Pearson algorithm and basic Bayesian collaborative filtering.

Kridel et al. [18] developed a recommender system for selling services to customers. It uses customer purchase history, browsing history, and user segments to make recommendations. The authors also proposed a recommender system for telecom service users. They applied various collaborative filtering algorithms, including GenericItemBased, GenericUserBased, Item Average, ItemUserAverage, TreeClustering, and SlopeOne from the Taste libraries, to a complex telecom user dataset. They evaluated different correlation algorithms on a typical telecom dataset based on computation time, variation in sample size, neighborhood computation, scalability, and precision.

Ajesh et al. [39] proposed recommender system that use clustering and random forest classification as a multilevel strategy to predict recommendation. Based on user ratings Kmeans and k-means ++ clustering algorithms were used so as it will be easy to generate labels. The labels for the users was predicted using random forest algorithm. After a user profile is formed, a random forest will classify users based on whether one movie is recom1mended or not. Similarly, Tian et al. [26] developed a customized recommendation system by applying k-means clustering before evaluating similarity.

All related papers presented have their own role in developing recommendation systems. We can see and understand that there are different methods for developing recommendation systems that can be applied in different areas.

IV. PROPOSED METHODOLOGY

The proposed methodology suggests mobile package bundles for users in two phases:

Phase 1: Understanding Customer Behavior

In the first phase, we build relationships between Ethiotelecom customers and their usage of services like voice, data, and SMS. We group customers with similar usage patterns using a clustering algorithm. Then, we analyze how these groups use each service by creating usage patterns in Microsoft Excel. This analysis helps us identify which packages are best suited to customers' preferences.

Phase 2: Recommending Packages

Once we understand customer usage patterns and preferences, we create a classification model. This model uses both usage patterns and package information to make recommendations. It considers how customers behave and what they prefer to recommend the most suitable mobile package bundles. This ensures that customers get personalized recommendations tailored to their unique preferences.

1) DATA COLLECTION

Call detail record data for 7899 active customers was collected from the Ethiotelecom, Information System Department. The data was two-month data, from January 2023 to February 2023. The requested CDR data contained a day's summery of voice in seconds, data in kilobytes, and SMS in numbers. There are also other features stored on CDR data, like revenue generated on each service, off-peak voice, peak voice, and so on.

2) DATA PREPROCESSING

We used Microsoft Excel and Python programming tools to preprocess our CDR data. Here are the key preprocessing steps we followed:

A. Feature Selection

We started by selecting the most relevant features from the raw CDR data. We focused on voice, data, and SMS usage within specific time frames (daily, weekly, and monthly). We retained features such as SERV_NO, DATA_DAY, SMS_LOCAL_USAGE, DATA_USAGE_KB, and created a new feature called TOTAL_VOICE_USAGE_SECONDS by combining off-peak and peak voice usage.

B. Handling Missing Values

We carefully checked for any missing or null values in the raw data using both Microsoft Excel and Python. Fortunately, there were no missing values in our source data.

C. Data Conversion

To align the data with the available mobile package services, we converted the units of measurement. Voice usage, originally in seconds, was converted to minutes, and data usage, initially in kilobytes, was converted to megabits

D. Data Organization

Our raw data represented day-to-day usage of voice and data services spanning up to sixty days. To recommend packages for current Ethiotelecom mobile plans (daily, weekly, and monthly), we organized the data accordingly:

- Daily base: No changes were made, as this aligned with the raw data.
- Weekly base: We aggregated the data for 7 consecutive days.
- Monthly base: We aggregated the data for 30 consecutive days, considering usage patterns for voice, data, and voice plus services.

E. Data Aggregation

To calculate the total voice usage for each customer, we combined both off-peak and peak voice usage. We used the formula: TOTAL_VOICE_USAGE_SECONDS= OFFPEAKUSGSECONDS + PEAK_USGSECONDS.

E. Outlier Detection and Removal

Outliers are values that significantly deviate from typical customer usage. To prepare the data for the K-means clustering algorithm, which is sensitive to outliers, we set minimum and maximum threshold values based on the 5th and 95th quantiles, respectively. Values below the 5th quantile and above the 95th quantile were considered outliers and were removed.

F. Normalization

Normalization is the process of scaling data to the same range. We employed the Min-Max scaling method to ensure that data points were on the same scale. This step is crucial for K-means clustering, which relies on distance calculations. Keeping data on the same scale helps avoid the sensitivity of the Euclidean distance metric to variations in data differences.

These steps were taken to prepare the data for further analysis, particularly for K-means clustering in the Mobile Package Service Recommender system.

V.CLUSTERING AND CLUSTER ANALYSIS

A. K-means clustering

The K-means clustering algorithm is a way to group data points into clusters based on their similarity. It does this by finding centroids (central points) for each cluster and assigning data points to the cluster with the closest centroid. This helps us understand patterns in data.

K-means is simple, efficient, works well with large datasets, and is straightforward to use. This method helps us group customers based on their usage patterns, which can be valuable for various applications.

We applied K-means to customer usage data, which includes voice, data, and voice-plus-data usage. We looked at daily, weekly, and monthly usage patterns. For daily usage, we used data from one month to simplify things. For weekly and monthly patterns, we used data from two months.

The features we considered were minutes for voice and megabits for data. For daily usage, we looked at days 1 to 30. For weekly usage, it was weeks 1 to 9. And for monthly usage, we considered usage in months 1 and 2.

To find the right number of clusters for our data, we want to avoid having too few or too many clusters. If we have too many, it becomes hard to make sense of the data, and it's computationally demanding. But if we have too few, we might group together data that should be separate.

To figure out the best number of clusters, we used a method called the "elbow method." This method looks at how the K-means algorithm's cost changes as we vary the number of clusters. Here's how it works:

1. We run the K-means algorithm on our data with different values of K (the number of clusters).

2. For each K, we calculate something called "distortion" or "within-cluster sum squares." It measures how far each data point is from its cluster's center.

3. We then plot these distortion values for different K values.

4. When we look at the plot, we try to find the point where the distortion starts to level off, forming an "elbow" shape. This point is our best estimate for the optimal number of clusters.

By using the elbow method, we can make a more informed choice about how many clusters to use for our voice, data, and voice-plus-data services data.

TABLE I NUMBER OF CLUSTER BASED ON ELBOW METHOD

Services		K value
Total voice usage	Daily	3
	Weekly	5
	Monthly	5
Data usage	Daily	6
	Weekly	5
	Monthly	6
Voice plus data usage	Daily	6
	Weekly	7
	Monthly	5

B. Cluster analysis

Cluster analysis helps us understand how our data is grouped or clustered. To do this, we calculate the average values for each group to see how usage is distributed. Based on these averages, we

create rules that help us label packages for each group. Tables II and III show the attached package labels for voice data services in daily, weekly, and monthly plans.

TABLE II ATTCHED PACKAGE FOR VOICE SERVICE USERES

Voice service plan	Package label name
Daily	P1-P2
Weekly	P3-P6
Monthly	P7-P10

TABLE III ATTCHED PACKAGE FOR DATA SERVICE USERES

Data service plan	Package label name
Daily	P11-P14
Weekly	P15-P18
Monthly	P19-P22

VI. CLASSIFICATION MODEL BUILDING

We built a classification model using two machine learning algorithms: random forest and k-nearest neighbor (KNN).
Random Forest is great for dealing with complex data and avoiding overfitting. The critical parameter here is "n_estimators," which tells us how many decision trees are used in the model. Choosing the right number depends on your data and the balance between model complexity and performance. More trees mean a more complex model and longer computation.
K-Nearest Neighbor (KNN) is a simple and versatile algorithm for classification and regression. It relies on two key parameters: the number of neighbors (k) and the distance metric. Picking the right 'k' is crucial, as small values can make the model sensitive to noise, while large values can lead to smoothing and misclassification. The choice of distance metric depends on the data type, with Euclidean distance suitable for numbers and Manhattan distance for categorical data.
We used the powerful Scikit-Learn library in Python for this. We had 31 features for daily usage, 10 for weekly, and 3 for monthly, both for data and voice services, to create our classification model.

TABLE IV FEATURES GIVEN FOR RF & KNN MODELS

Usage plan for both voice &data services	Features given for RF &KNN
Daily	Package label, D1-D31 usage pattern
Weekly	Package label, W1-W7 usage pattern
Monthly	Package label, M1-M2 pattern

A. Evaluation metrics

An appropriate evaluation metric is required to accurately assess the efficiency of the grouping mechanism as well as the performance of the system in generating recommendations that are relevant to user needs. Below are the evaluation metrics that are used:

1. Silhouette Score: It measures how well data points are grouped in clustering. Scores range from -1 to 1, with 1 indicating strong similarity within clusters.

2. Precision: In recommendation systems, it assesses how many of the recommended items are actually relevant to the user, indicating system accuracy.

3. Recall: Measures how many relevant items were successfully recommended among all relevant items, indicating retrieval capability.

4. F1 Score: Combines precision and recall into a single value, balancing the trade-off between precision and recall in identifying relevant items for users**.**

VII. RESULT AND DISCUSSION

A. K means clustering result & discussion

We used the K-means algorithm to group customers based on their voice, data, and voice-plus-data usage patterns. Here's how we grouped them:

- For daily voice usage, we had 3 groups.
- For weekly voice usage, we had 5 groups.
- For monthly voice usage, we also had 5 groups.
- For weekly data usage, there were 5 groups.
- For monthly data usage, there were 6 groups.
- For voice-plus-data usage, we had 6 groups for daily, 7 for weekly, and 5 for monthly.

We used silhouette scores to evaluate these groupings. Silhouette scores above 0.5 are considered good, indicating that users in one group are similar in their usage compared to other groups. For voice and voice-plus-data, weekly and monthly plans had good groupings (scores around 0.55 to 0.64). For data usage, all plans (daily, weekly, and monthly) showed efficient groupings with high scores (around 0.74 to 0.80).
In summary, our clustering method effectively grouped customers with similar usage patterns, especially for data usage across different plans.

B. Data service classification

In our analysis, we grouped customers based on daily, weekly, and monthly data usage patterns, each with different package options. We built models using Random Forest (RF) and K-nearest neighbor (KNN) algorithms to recommend relevant packages to these groups. For daily data usage, RF achieved an 86.9% F1 score, while KNN scored 84.3%. In weekly data

usage, both RF and KNN performed well, with average precision above 90% and recall around 89%. For monthly data usage, RF achieved a recall of 92.6% and precision of 97.5%, while KNN had a recall of 95.0% and precision of 97.2%. In summary, KNN outperformed RF for weekly and monthly plans with high F1 scores, while RF excelled for daily plans, showcasing the strengths of each algorithm in various data service scenarios. Fig 1 shows data usage classification model result.

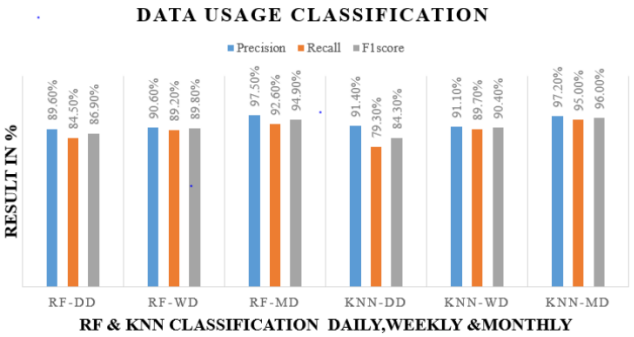

FIGURE I DATA USAGE CLASSIFICATION RESULT

C. Voice service classification

In the case of voice usage pattern analysis, we grouped customers into distinct categories based on their habits and assigned specific package labels. We aimed to create classification models using Random Forest (RF) and K-nearest neighbor (KNN) algorithms to recommend suitable packages to each group.

For daily voice usage, RF achieved a high recall of 92.90% and precision of 97.80%, indicating its accuracy in recommending relevant packages (P1–P2). KNN, although slightly lower, still performed well with a recall of 90.04% and precision of 97.30%.In weekly voice usage, both RF and KNN excelled with RF achieving a recall of 92.60% and precision of 93.40%, and KNN reaching a recall of 93.90% and precision of 94.70% for packages (P3–P6).For monthly voice usage, RF and KNN both performed exceptionally well with a precision of 99.40%, and their recall values were remarkable, with RF at 99.7% and KNN at 99.30% for packages (P7–P10).

In summary, RF outperformed KNN for daily and monthly voice usage plans with F1 scores of 95.10% and 99.60%, respectively. Conversely, KNN performed better than RF for the weekly usage plan with an F1 score of 94.30%. This highlights each algorithm's strength in different voice service scenarios. Fig II shows voice usage classification model result.

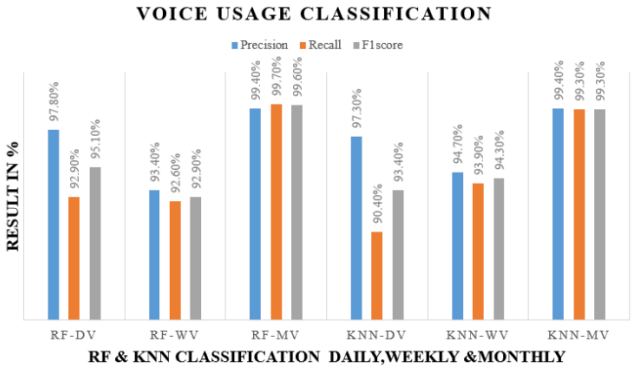

FIGURE II DATA USAGE CLASSIFICATION RESULT

VIII. CONCLUSION AND FUTURE WORK

A. CONCLUSION

Telecom operators has introduced mobile packages to cater to diverse customer needs, offering voice, data, and SMS plans at competitive rates. To enhance user experience and address the challenge of information overload, a mobile package recommendation system was developed. This system employed clustering and classification techniques on 7899 customer data records, grouping them by daily, weekly, and monthly usage patterns using K-means clustering. Subsequently, a model was built using RF and KNN classification algorithms, with RF performing better for daily and monthly voice plans and KNN excelling in weekly voice and data plans. This recommendation system, once deployed, will help telecom operators to attract and retain customers while boosting revenue by offering personalized mobile package suggestions.

B. FUTURE WORK

The researcher's proposed features for this study include the utilization of active prepaid customers to recommend daily, weekly, and monthly mobile packages, with the potential for applying similar methodologies to other package types and postpaid customers in the future. Furthermore, future work could involve integrating pricing information with customer usage patterns to offer more tailored recommendations. While the current study concentrates solely on mobile package services, it highlights the adaptability of these methodologies for providing recommendations in various service domains, allowing for a broader range of personalized suggestions to meet diverse customer needs.

ACKNOWLDGEMNT

First and foremost, I would like to say thank you to the almighty God, who gives me the strength and patience to finish this study. Secondly, I would like to say thank you to Ethiotelecom, which believes in me and gives me the chance to study telecom engineering.

Then I would like to express my great respect and thank my advisor, Dr. Sosina Mengistu, for the insightful discussion and guidance throughout this thesis.

Finally, I would like to express my deepest gratitude to my beloved husband and amazing son for their understanding, patience, and support in completing the study.

References

[1] F. Ricci, L. Rokach, and B. Shapira, "Introduction to recommender systems handbook," in Recommender systems handbook: Springer, 2 010, pp. 1-35Li, Xiupeng. “Differences between Proxy, VPN, and SSH.”(blog)http://blog.csdn.net/map_lixiupeng/article/details/4169504 5, 2014

[2] R. Burke, "Hybrid recommender systems: Survey and experiments," User modeling and user-adapted interaction, vol. 12, pp. 331-370, 2002.

[3] D. Jannach, M. Zanker, A. Felfernig, and G. Friedrich, Recommender systems: an introduction. Cambridge University Press, 2010.

[4] Min, Tong, Qingrong Li and Youqun, Mo. “Security Study of VPN”Computer Era, vol. 12, pp.1-3, 2002.

[5] K. Falk, Practical recommender systems. Simon and Schuster, 2019.

[6] D. Gavalas, C. Konstantopoulos, K. Mastakas, and G. Pantziou, "Mobile recommender systems in tourism," Journal of network and computer applications, vol. 39, pp. 319- 333, 2014.

[7] Z. Huang, D. Zeng, and H. Chen, "A comparison of collaborative-filtering recommendation algorithms for e-commerce," IEEE Intelligent Systems, vol. 22, no. 5, pp. 68-78, 2007.

[8] P. a. V. o. Falcarin, Antonio and Yu, Jian, "A Recommender System for Telecom Users," falcarin2020recommender, 2020

[9] Y. Guo, Y. Zhou, X. Hu, and W. Cheng, "Research on recommendation of insurance products based on random forest," in 2019 international conference on machine learning, big data and business intelligence (MLBDBI), 2019: IEEE, pp. 308-311.

[10] K. Miyahara and M. J. Pazzani, "Improvement of collaborative filtering with the simple Bayesian classifier," Information Processing Society of Japan, vol. 43, no. 11, 2002.

[11] X. Su and T. M. Khoshgoftaar, "Collaborative filtering for multi-class data using belief nets algorithms," in 2006 18th IEEE international conference on Tools with Artificial Intelligence (ICTAI'06), 2006: IEEE, pp. 497-504.

[12] D. J. Kridel, D. R. Dolk, and D. Castillo, "Recommender systems as a mobile marketing service," Journal of Service Science and Management, vol. 2013, 2013.

[13] . Wang, Y. Zhang, and Q. Wu, "College library personalized recommendation system based on hybrid recommendation algorithm," Procedia CIRP, vol. 83, pp. 490-494, 2019.

[14] A. Ajesh, J. Nair, and P. Jijin, "A random forest approach for rating-based recommender system," in 2016 International conference on advances in computing, communications and informatics (ICACCI), 2016: IEEE, pp. 1293-1297.

7 REFERENCE

[1] F. Ricci, L. Rokach, and B. Shapira, "Introduction to recommender systems handbook," in *Recommender systems handbook*: Springer, 2010, pp. 1-35.

[2] E. M. Sibanda and T. Zuva, "Call Data Record Based Recommender Systems for Mobile Subscribers," in *2018 International Conference on Intelligent and Innovative Computing Applications (ICONIC)*, 2018: IEEE, pp. 1-7.

[3] C. C.Aggarwal, *Recommender Systems*. 2016.

[4] P. a. S. Melville, Vikas, *Recommender Systems*. Boston,MA: Springer US, 2010.

[5] G. Rodríguez, "tryolabs.com," ed, 2018.

[6] S. Zhang, L. Yao, A. Sun, and Y. Tay, "Deep learning based recommender system: A survey and new perspectives," *ACM computing surveys (CSUR)*, vol. 52, no. 1, pp. 1-38, 2019.

[7] R. Burke, "Hybrid recommender systems: Survey and experiments," *User modeling and user-adapted interaction*, vol. 12, pp. 331-370, 2002.

[8] D. Jannach, M. Zanker, A. Felfernig, and G. Friedrich, *Recommender systems: an introduction*. Cambridge University Press, 2010.

[9] K. Falk, *Practical recommender systems*. Simon and Schuster, 2019.

[10] D. Gavalas, C. Konstantopoulos, K. Mastakas, and G. Pantziou, "Mobile recommender systems in tourism," *Journal of network and computer applications*, vol. 39, pp. 319-333, 2014.

[11] Z. Huang, D. Zeng, and H. Chen, "A comparison of collaborative-filtering recommendation algorithms for e-commerce," *IEEE Intelligent Systems*, vol. 22, no. 5, pp. 68-78, 2007.

[12] Falcarin, paolo, Antonio Vetro and Jian Yu,, "A Recommender System for Telecom Users," *falcarin2020recommender*, 2020.

[13] B. Fang, S. Liao, K. Xu, H. Cheng, C. Zhu, and H. Chen, "A novel mobile recommender system for indoor shopping," *Expert Systems with Applications*, vol. 39, no. 15, pp. 11992-12000, 2012.

[14] Y. Guo, Y. Zhou, X. Hu, and W. Cheng, "Research on recommendation of insurance products based on random forest," in *2019 international conference on machine learning, big data and business intelligence (MLBDBI)*, 2019: IEEE, pp. 308-311.

[15] A. Ajitsaria. "https://realpython.com/build-recommendation-engine-collaborative-filtering/#what-is-collaborative-filtering].. ." realpython.com. https://realpython.com/build-recommendation-engine-collaborative-filtering/#what-is-collaborative-filtering]. (accessed 12, 2022).

[16] K. Miyahara and M. J. Pazzani, "Improvement of collaborative filtering with the simple Bayesian classifier," *Information Processing Society of Japan*, vol. 43, no. 11, 2002.

[17] X. Su and T. M. Khoshgoftaar, "Collaborative filtering for multi-class data using belief nets algorithms," in *2006 18th IEEE international conference on Tools with Artificial Intelligence (ICTAI'06)*, 2006: IEEE, pp. 497-504.

[18] D. J. Kridel, D. R. Dolk, and D. Castillo, "Recommender systems as a mobile marketing service," *Journal of Service Science and Management*, vol. 2013, 2013.

[19] G. Linden, B. Smith, and J. York, "Amazon. com recommendations: Item-to-item collaborative filtering," *IEEE Internet computing*, vol. 7, no. 1, pp. 76-80, 2003.

[20] I. Singh, K. V. Singh, and S. Singh, "Big data analytics based recommender system for value added services (VAS)," in *Proceedings of Sixth International Conference on Soft Computing for Problem Solving: SocProS 2016, Volume 2*, 2017: Springer, pp. 142-150.

[21] M. Ghazanfar and A. Prugel-Bennett, "An improved switching hybrid recommender system using naive bayes classifier and collaborative filtering," 2010.

[22] M. J. Pazzani and D. Billsus, "Content-based recommendation systems," in *The adaptive web: methods and strategies of web personalization*: Springer, 2007, pp. 325-341.

[23] P. Melville, R. J. Mooney, and R. Nagarajan, "Content-boosted collaborative filtering for improved recommendations," *Aaai/iaai*, vol. 23, pp. 187-192, 2002.

[24] L. R. Francesco Ricci, Bracha Shapira, Paul B. Kantor, *Recommender Systems Handbook*. New York, NY: Springer New York, NY, 2011.

[25] I. Beregovskaya and M. Koroteev, "Review of Clustering-Based Recommender Systems," *arXiv preprint arXiv:2109.12839*, 2021.

[26] Y. Tian, B. Zheng, Y. Wang, Y. Zhang, and Q. Wu, "College library personalized recommendation system based on hybrid recommendation algorithm," *Procedia CIRP*, vol. 83, pp. 490-494, 2019.

[27] J. Zhang, Y. Lin, M. Lin, and J. Liu, "An effective collaborative filtering algorithm based on user preference clustering," *Applied Intelligence*, vol. 45, pp. 230-240, 2016.

[28] Y. Shi, M. Larson, and A. Hanjalic, "Collaborative filtering beyond the user-item matrix: A survey of the state of the art and future challenges," *ACM Computing Surveys (CSUR)*, vol. 47, no. 1, pp. 1-45, 2014.

[29] L. Miranda, J. Viterbo, and F. Bernardini, "Towards the Use of Clustering Algorithms in Recommender Systems," in *AMCIS*, 2020.

[30] D. BIRU, "Usage Based Clustering of Customers," 2019.

[31] Y. Amri, A. L. Fadhilah, N. Setiani, and S. Rani, "Analysis clustering of electricity usage profile using k-means algorithm," in *IOP Conference Series: Materials Science and Engineering*, 2016, vol. 105, no. 1: IOP Publishing, p. 012020.

[32] A. Sagar. towardsdatascience.com,. https://towardsdatascience.com/customer-segmentation-using-k-means-clustering-d33964f238c3. (accessed 15, 2023).

[33] J. Herce-Zelaya, C. Porcel, J. Bernabé-Moreno, A. Tejeda-Lorente, and E. Herrera-Viedma, "New technique to alleviate the cold start problem in recommender systems using information from social media and random decision forests," *Information Sciences*, vol. 536, pp. 156-170, 2020.

[34] R. Pan, P. Dolog, and G. Xu, "KNN-based clustering for improving social recommender systems," in *Agents and Data Mining Interaction: 8th International Workshop, ADMI 2012, Valencia, Spain, June 4-5, 2012, Revised Selected Papers 8*, 2013: Springer, pp. 115-125.

[35] B. Li, S. Wan, H. Xia, and F. Qian, "The research for recommendation system based on improved KNN algorithm," in *2020 IEEE International Conference on Advances in Electrical Engineering and Computer Applications (AEECA)*, 2020: IEEE, pp. 796-798.

[36] G. Li and J. Zhang, "Music personalized recommendation system based on improved KNN algorithm," in *2018 IEEE 3rd Advanced Information Technology, Electronic and Automation Control Conference (IAEAC)*, 2018: IEEE, pp. 777-781.

[37] H. Liu and V. Kešelj, "Combined mining of Web server logs and web contents for classifying user navigation patterns and predicting users' future requests," *Data & Knowledge Engineering*, vol. 61, no. 2, pp. 304-330, 2007.

[38] G. Khanvilkar and D. Vora, "Sentiment analysis for product recommendation using random forest," *International Journal of Engineering & Technology*, vol. 7, no. 3, pp. 87-89, 2018.

[39] A. Ajesh, J. Nair, and P. Jijin, "A random forest approach for rating-based recommender system," in *2016 International conference on advances in computing, communications and informatics (ICACCI)*, 2016: IEEE, pp. 1293-1297.

[40] V. R. Patel and R. G. Mehta, "Performance analysis of MK-means clustering algorithm with normalization approach," in *2011 World Congress on Information and Communication Technologies*, 2011: IEEE, pp. 974-979.

[41] Xu, D & Tian (2015), "A comprehensive survey of clustering algorithms," *Annals of Data Science(2015)*: 165--193, 2015.

[42] Sruthi E. R. "Understand Random Forest Algorithms With Examples." www.analyticsvidhya.com. https://www.analyticsvidhya.com/blog/2021/06/understanding-random-forest/ (accessed 5, 2023).

[43] Precision, Recall and F1-Score using R. www.geeksforgeeks.org. https://www.geeksforgeeks.org/precision-recall-and-f1-score-using-r/ (accessed 17, 2023).