ISTANBUL COMMERCE UNIVERSITY

GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES

CHURN PREDICTION IN TELECOMMUNICATION SECTOR

Faiza HASSAN MOHAMED

Supervisor
Assist. Prof. Dr. Mustafa Cem KASAPBAŞI

MASTER'THESIS
DEPARTMENT OF INDUSTRIAL ENGINEERING
ISTANBUL – 2021

ACEPTANCE AND APPROVAL PAGE

On 01/03/2021 **Faiza HASSAN MOHAMED** successfully defended the thesis, entitled "**Churn Prediction In Telecommunication Sector**" which she prepared after fulfilling the requirements specified in the associated legislation, before the jury members whose signatures are listed below. This thesis is accepted as a **Master's Thesis** by Istanbul Commercial University. Graduate School of Natural and Applied Sciences **Department of Industrial Engineering.**

Supervisor	**Assist. Prof. Dr. Mustafa Cem KASAPBAŞI** İstanbul Ticaret Üniversitesi
Jurry Member	**Assoc. Prof. Dr. Berk AYVAZ** İstanbul Ticaret Üniversitesi
Jurry Member	**Assist. Prof. Dr. Buket DOĞAN** Marmara Üniversitesi

Approved Date: 15.03.2021
Istanbul Commerce University, Graduate School of Natural and Applied Sciences, accordance with the 1st article of the Board of Directors Decision dated 15.03.2021 and numbered 2021/308, "Faiza HASSAN MOHAMED" (TC:99375225484P) who has determined to fulfill the course load and thesis obligation was unanimously decided to graduated.

Prof. Dr. Necip ŞİMŞEK
Head of Graduate School of Natural and Applied Science

DECLARATION OF ACADEMIC AND ETHIC INTEGRITY

I hereby declare that

- All the information in this thesis obtained and presented in accordance with academic and ethical integrity,
- I have fully cited and accurately and clearly referred to each source of inspiration and information, wherever and whenever applicable for all materials and results that are not original to this work,
- Neither whole nor any part of this document is not presented any other place previously

15/03/2021

Faiza HASSAN MOHAMED

TABLE OF CONTENTS

Page

ABSTRACT

M.Sc. Thesis

CHURN PREDICTION IN TELECOMMUNICATION SECTOR

Faiza HASSAN MOHAMED

Istanbul Commerce University
Graduate School of Natural and Applied Sciences
Department of Industrial Engineering

Supervisor: Assist. Prof. Dr. Mustafa Cem KASAPBAŞI
2021, 34 Pages

Churn Prediction has been implemented in the research works and more studies on it been published using different advanced mechanisms including Machine Learning, Data Mining, and Hybrid mechanism. These mechanisms found out to help enterprise companies and small businesses to classify and predict churning customers to be able to retain them to stay with their company using their services. Also, found out to help top managers and decision makers to be able take reliable decisions and Customer Relation Management CRM department as well. In this study, a telecom sector churn dataset named Orange which belongs to International Orange Telecom Company is used for customer churn prediction. Ensemble classifiers are used AdaBoostM1, PCA, Gain Ratio, Info Gain, Bagging in combination with J4.8, Naïve Bayes, Logistic Regression, Random Forest, KNN, LMT (Logistic model Tree). Highest accuracy of 94% is obtained by combination of Bagging and J48. The results are compared with other studies as well and this study performed as good as the surveyed literature and surpassed in same cases.

Keywords: Churn Prediction, KNN, LMT, CRM, PCA

ÖZET

Yüksek Lisans Tezi

Telekom Sektörleri için Topluluk Sınıflandırıcılarla Ayrılma Tahmini

Faiza HASSAN MOHAMED

İstanbul Ticaret Üniversitesi
Fen Bilimleri Enstitüsü
Endüstri Mühendisliği Anabilim Dalı

Danışman: Dr. Öğr. Üyesi Mustafa KASAPBAŞI
2021, 34 Sayfa

Literatürde, Makine Öğrenimi, Veri Madenciliği ve Hibrit teknikleri gibi farklı teknikler kullanılarak Ayrılma/Çalkalanma Tahmini gerçekleştirilmiştir. Bu teknikler, şirketleri ve işletmeleri, hizmetlerini kullanarak şirketlerinde kalabilmeleri için müşterileri belirleme ve tahmin etme ayrıca ayrılan müşteri konusunda destekler. Üst düzey yöneticilerin ve karar vericilerin güvenilir kararlar almasına ve Müşteri İlişkileri Yönetimi (CRM) departmanına da yardımcı olur. Bu çalışmada, müşteri kaybını tahmin etmek için Orange (Uluslararası Telekominikasyon firması) adlı bir telekom sektörü ayrılan müşteri veri seti kullanılmıştır. Topluluk sınıflandırıcıları AdaBoostM1, PCA, InfoGain, Gain Ratio, Bagging ile birlikte J4.8, Naive Bayes, Lojistik Regresyon, Rastgele Orman, KNN, LMT (Lojistik model Ağacı) sınıflandırıcıları kombinasyonları ile birlikte kullanılır. Torbalama ve J4.8 kombinasyonu ile en yüksek % 94 doğruluk elde edilir. Sonuçlar diğer çalışmalarla da karşılaştırılmış ve bu çalışma araştırılan literatür kadar iyi performans göstermiş ve bazı vakalarda daha başarılı olduğu görülmüştür.

Anahtar Kelimeler: Ayrılama Tahmini, kNN, LMT, CRM, PCA

ACKNOWLEDGEMENT

I would like to sincerely thank to my Supervisor Asst. Prof. Dr. Mustafa Cem Kasapbaşı, for his mentoring and support from the beginning till the end of the completion of my Thesis. Surely, it was a learning process.

My sincerely gratitude goes to Istanbul Commerce University from the top to the bottom staff, I am glad to have had the opportunity of being on of your student.
I dedicate this academic achievement to my sweetheart Parents, and also to my sisters and brothers for their endless support.

I would like to thank my colleagues and my friends for their encouragement and inspiration.

Faiza HASSAN MOHAMED

Istanbul, 2021

LIST OF FIGURES

LIST OF TABLES

ABBREVIATIONS

ANN	Artificial Neural Network
BP network	Back Propagation Network
CCP	Customer Churn Prediction
CRM	Customer Relationship Management
KNN	K Neural Network
LMT	Logistic Model Trees
PCA	Principal Component Analysis
WEKA	Waikato Environment Knowledge Analysis

1. INTRODUCTION

Recent years Data mining has been hot research area, which businesses in general using to solve so many problems, and also to use enhancing their businesses development, digital companies, specially service providers are the most businesses which has been used in Data mining technique to solve occurred problems and at the same time to use for the developments of their businesses functionality. Churn prediction has been one of problems which businesses suffering with, and needed to solve using so many different Data mining and Machine learning tools and techniques. In our study we would discuss Telecom churn prediction problem, using Orange telecom dataset while we testing the dataset an ensemble classifier with number of methodologies to find the best result method.

1.1 Background of the Study

Corporate businesses in the competitive market basically depend on the revenues which generated from the customers. According to that, CRM focuses on customers' data as it's the source which helps companies to take accurate decisions, because customers' data give decision makers hints as its summarized and organized, decision makers can easily get insights from the data which provide them all needed information of the customer with clear details. (Tsai and Lu. 2009).

Churn prediction is helpful tool for businesses to predict customers who churning or planning to unsubscribe businesses service. The aim of churn prediction is to provide companies the real situation of the customer by the collected and stored customer data, as that situation classifies with detail who are the customers who likely to churn and who are the customer who will existed and stay using the service. Therefore, businesses with accurate churn prediction results can easily organize their marketing campaigns to work on customer retention and creating best retention strategies which they will target to the churning customers and solving their shortcomings to satisfy and convince them to not churn by providing best services as they want. (Abbasimehr et al. 2011).

In the past years, mobile phone users increased over 60% in the world, as it made easy for world population to the communication.

The increase of mobile phone users caused the boom of the telecommunication service providers and also increased the competition in the market. Telecom Companies working on how to build marketing campaigns to attract customers in the hot competitive market. The rapid change of the market and the increasing competition caused to find developed tools and techniques to attract customers and to satisfy them, whether it takes effort to do researches to find out solutions and use of developed technologies to solve the issue, because the customer is primary value of the businesses as it affects the profit. (Verbeke et al. 2012).

Customer Churn Prediction is used to know, determine the customers who cease the subscription of the company services. Digital companies which provide services through online and offline are mostly facing the risk of losing a subscribed customer who subscribed their services, such as, Telecommunication sector, Banking sector, Technology and ICT (Information Communication Technology) sector, and Insurance companies sector. The problem of churn is increasingly boosting machine learning and data mining research areas, because they are the most applicable research fields and also provide most suitable solutions using developed tools and techniques to solve the churn problem. Data mining techniques allow companies to learn more about their customers, whether the companies can easily access the customer information which collected from customer transactions and stored in the company database. That collected data from the customer and stored in the company's database, allow companies to learn more about customers, their interests, and to know more about the customers' situation, whether they are satisfied and liked the services of the company they subscribed to use it or not, so, as the world is moved digitally and online today, churn prediction is booming and hot research topic, which most of the researchers are working on, and most of digital companies in the world are interesting to know more about the problem and how to solve it, as it is helping to get accurate and reliable customer data, which help them to take accurate and best decision through enhancing, developing, and continuing their businesses. Churn prediction which is knowing how many customers are going to leave or unsubscribe the service, and also prediction the customers who are likely to churn, is one of hot topics in Data mining and Machine learning areas of today's researches.

Customer satisfaction has been primary core value of the businesses today, and as rapidly changing technology and digital tools, businesses eagerly trying to find solutions to establish the best relations with their customers. That reason resulted to have created specially designated tools and systems for customer care and satisfaction.

Businesses not only created customer relationship office, but invested technology to build the best developed relation with the customer they could do.

Customer Relationship Management is a department which focuses on the customers' relationship with the company, today's businesses rapidly developing and customer is the one of business key elements, and companies working to have the best relationship with them. This department has created for customers, since they are primary value of the company, without customer companies will not generate any revenue expected, and as the customer is the core element in businesses existence, companies added this special section into the managerial levels, to build good relationship with the customers and also to learn more about the customers and to know their situations, as well as to know how satisfied their customers are. Satisfying the customer is one of the most activities in the business process, so to be sure for customer satisfaction, businesses must do whatever needed to successfully done that activity. CRM department sorted and put together the strategies needed to build a strong relationship between the customers and the business. Following the rapid change of digitalized world, businesses built a system which is called Customer Relationship Management System CRMS, which is designed for taking care of the customers' relations with the business.

The CRMS has functioning technologies which stores customer data, linking to the business activities and transactions done by them, and allow the business to easily use that data for enhancing business development strategies, and help the businesses to take reliable decisions.

1.2. Problem Statement

Telecommunication companies has been facing customer churn prediction, which is service providers, specially telecom service providers defined as serious problem as it causes losing the customers, which it causes losing profit.

Customers are part of the asset, as mentioned before, and their lose is revenue loss for the company, so that, companies are caring customers as they caring about their revenue and profit, because if one customer's subscription has lost, then the revenue of that subscription has lost with it too, also the company profit would face decrease of the asset and the profit, so businesses are caring customers as their lose and leaving from the company causes lose to the company and its profit too.

Churning the customer has been big problem to the company's assets, because of that, companies working on to know the customers who are likely to churn, before the churn activity has occurred, and that is the reason this topic has increased in the recent research areas in digital companies, machine learning, and data mining fields.

Our study structures, as we started the introduction part, after we discuss the literatures which has been published about our study, and after it, we outline the methodologies that we have used our study, collecting the results from the tested dataset with classifiers, and after, we will discuss the results comparing with other studies results, and at the end we conclude the study.

2. LITERATURE REVIEW

Telecom sector is a sector which customers play an important role whether they affect businesses work and revenue too, so the customer care and satisfaction must be the biggest goals and plans of the company to retain them and stay their customers, otherwise customers will churn and go to use other competitive service providers. Churn prediction is a problem most of telecom companies facing today, when the customer churn the company faces huge financial loss, as losing the customer causes the loss of the profit too, because when your customer leaves from you and go to another competitor, your company faces trouble for losing both revenue and customer too. For that reason, a lot of studies being published to solve this problem which more telecom companies nowadays facing, and researchers working on providing technologies and designing models to predict churners and then to prevent and retain churners for providing them needed services and solving their problems.

Andreea Dumitrache and Monic Mihaela Mear Matei (Dumitrache and Matei. 2019) have conducted a study on prediction customers who are going to defect in a Romanian mobile telecommunication companies. The churn analysis developed for post-paid customers. Study used in logistic regression to predict churn and a solution based on smooth bootstrap technique to correct for drawbacks of imbalanced classes. In the study, it analyzed churn behavior on a sample of 10701 subscribers randomly selected from a database of a large telecommunication companies operating on the Romanian market.

Businesses have been investing in marketing with high budget and planning time to time for marketing campaigns to attract customers. Thus, businesses are more eagerly trying to attract customers through those campaigns, which have invested in huge money, and businesses very curious to attract customers. Attracting a customer is something businesses could done by marketing campaigns and other strategies, but, what the businesses struggling with it attracting the existing customers, and preventing them to churn. Studies recent years found out that losing an existing customer is a huge lose to the company, because the existing customer is existing revenue, while the loss of that customer is loss of revenue. Thus, businesses are highly investing technology to be able to attract the existing customers and preventing them from churning.

Ullah et al (Ullah et al. 2019), published a study in churn prediction model which they used in the study they have done Random Forest, in their study, analyzed the machine learning techniques and factor classification in telecommunication sector. The study has used clustering technique which allowed researchers to predict easily the customers who are likely to churn, or churning customers, also, the caused reasoned for their churn, or to decide to unsubscribe the service of their subscribed telecom company. In this study, they have implemented feature selection by using with correlation attribute ranking filter and information gain. The implemented model classified first the customer data using classification algorithms, where random forest achieved better and resulted with 88.63% correctly classified instances. Assigning retention procedures is the most significant job of Customer Relationship Management CRM to stop churners and customers who are likely to churn their company subscription. This study has specified churn elements that are serious in defining the original caused sources and the reason of churn. Results found out after this model has implemented using random forest classifier produced better churn classification.

Data mining is a new technology procedure, which enables discovering unexpected data samples, to assist in the prediction of the upcoming directions. Recent times, this technology has been implemented in most of the world wide businesses, as businesses moved to digital, it is needed to apply the new technologies like data mining to solve business problems in general and develop business industries which cannot get away from using this technique, it is used for both, to solve businesses problems, and at the same time, to enhance and develop the services provided by the businesses to their end line customers to satisfy them. Decision Tree, is straightforward tool of data mining process, which used for predictions and to find out upcoming expectations. Decision Tree has different algorithms for generating the decision tree, these algorithms include C4.5, ID3, and other algorithms, which have been used to implement with different software tools.

Nijahwan et al (Nijahwan et al. 2019), have done a study, which they concentrated on implementing data mining in telecommunication sector, to enable predicting churn attitude of subscribed customers. They have used in their published work a data which collected from surveymonkey.com for mining purpose. After that, the collected data has been cleaned and processed, after that process, decision tree has generated to predict the customers who are likely to churn. Results found out after that, even showed some reasonable facts to point out, for example, if the customer who subscribed a specific telecom company and using the same number for two years and more duration time is less likely to churn and unsubscribe that telecom company to switch another company and subscribe new telecom company, that found out fact, strengthened the fact of retaining the existing customers and keeping them more difficult than attracting a new customer, so that found out data helped telecom companies to work on how they treated well and satisfy those existing customers and not allow them to churn for providing their needed and interested services..

Mihrimah Ozmen (Ozmen et al. 2020), in the study they have presented the importance of customer management and how telecommunication companies struggling the competition between them every company is trying to learn more about their customers and be able to manage them to keep them from churning and switching to another one.

V. Umayaparvathi and K. Iyakutti (Umayaparvathi and Iyakutti. 2012), have worked and published a study on data mining techniques in telecommunication company's churn prediction. Telecommunication companies facing many challenges in the competitive market, where every company changes rapidly as new technologies effecting to rapid changes, so that, customers are continuously looking for their need provider company to switch for a reason they finding whatever kind of service they needed, that switching stresses telecom companies to compete each other for the fear of losing their existing customer, each company trying to offer new and advanced services to their customers to prevent them from churning and switching to another company, constantly working on what satisfying most their service subscribed customers. So that, the work these researchers have published, searched the application of data mining procedures to predict customers who are likely going to churn and effect of attribute selection on classifying the churn.

Adnan Idris et al (Idris et al. 2013), have intended a study on an intelligent churn prediction system for telecommunication sector by utilizing functional feature extraction technique and ensemble method. İn their study, they have used ensemble classifications with minimum redundancy and maximum relevance mRMR, also, they have used, fisher's ratio and f-score methods to model the telecom churn prediction issue.

Utku Yabas et al (Yabas et al. 2012), have published a study on customer churn prediction for telecom services, they have tasked on data mining methods, for a purpose of to properly predict the customers who are likely to churn, while they have the tention to unsubscribe their current subscribed telecom service provider and to move and subscribe another similar one but different in terms of services and customer needs offered by it. They implemented their research task by using Orange Telecom dataset, which is one of the available telecom datasets in the internet, and a lot of researchers have used and studies have been done using it in churn prediction research issues.

Azeem and Usman (Azeem and Usman. 2018), have published a study churn prediction problem in telecommunication sector, which they not only focused on modeling the churn prediction and customer churners, but, they also, have implemented tools for customer retention. As the existing literatures have limitations, and the churn prediction issue getting serious problem which effecting telecom companies both of losing customer and revenue too, they have tried and implemented new narrative model, to be exact able to obtain the purpose of correct classification and to obtain the goal of retaining the churners. They worked on both narratively, to find the problem and also to draw the solution which is how company can retain customers to not churn and stay to use the service.

Alae Chouiekh and El Hassan Ibn El Haj (Chouiekh and El Haj. 2020), have published work on churn prediction problem. In their work, they have tried a new machine learning model, as they used a narrative method by implementing deep convolutional neural network, which they applied on the dataset they have implemented in their research as experiment to identify customer churn. They have found out that deep convolutional neural network achieved better results and performed better than other previously used machine learning algorithms.

Adnan Amin et al (Amin et al 2019), have presented a study on customer churn prediction, which they specially focused on finding out the real solution for churn prediction problem which most of telecom service providers suffering from, they found out that, there must be clear detailed and prove reasons behind the customer churn. Knowing that, finding the unseen factors must be the priority and it helps more to find the causes of the problem to be able to solve it later on. They have worked on determining relevance and dataset samples to know the unseen factor.

Yasser Khan et al (Khan et al. 2019), have conducted a customer churn prediction study. Churn prediction has been serious problem which most of the companies struggling with it recent years, specially service providing companies, for the reason of customers demand of the rapidly changing digital aspects and growing technology. Companies have been focusing to know better about their customers, using different tools and technologies which facilitate businesses to get the detailed informations and data of the customers. The competition is highly increasing, so, in their study, the authors have used Artificial Neural Network (ANN) approach to predict the customers who are likely to churn from the company to move to another one.

Clement Kirui et al (Kirui. et al 2013), have published a study on customer churn prediction in mobile telephony industry which they have used probabilistic classification in data mining. In their study, they aimed to enhance the ability of telecom companies to know customers who are likely to churn, in their research customer transactions recorded to including detailed customer information which enables them to know better and identify best results whether there is a churn or not. They have examined the new set of features of the customer data by using Naive Bayes and Bayesian Network probabilistic data mining algorithms, and then they have compared the found out results to the results that gained from C4.5 and Decision Tree classification algorithms.

J. Vijaya and E. Sivasankar (Vijaya and Sivansankar. 2018), have published research article on computing efficient features using rough set theory combined with ensemble classification techniques to improve the customer churn prediction in telecom sector. They have intended a methodology using rough set theory to classify functional characteristics for telecommunication churn prediction.

Zhong and Li (Zhong and Li. 2019), have published research paper on churn prediction, by using authentic customers' call data, they have modified the convolutional neural network predictive model in their research to classify telecom churn prediction problem.

Ruiyun Yu et al (Yu et al. 2018), have published their research study, in which they intended a particle classification optimization based back propagation BP network for telecommunication customer churn prediction algorithm.

Awodele Oludele et al (Oludele. 2020), have published study in enhanced churn prediction in telecommunication industry. In their research, they utilized Markov Chain Model to sample the customer churn prediction. The Markov Chain Model gives more adaptability than most other possible models, and can easily incorporate variables which most of other models cannot easily do.

Hossam Faris (Faris. 2018), have published research study in churn prediction. As the telecommunication companies facing churn prediction problem, the study analyzed and found out that the most effective churn prediction problem solution is knowing the customers who likely to churn, before they took they did churn and switch to another company. Thus, that solution and strategy need a powerful prediction model, so in their study, researchers have intended an intelligent hybrid model which it based on particle swarm optimization and feed forward neural network for churn prediction.

Recent years, more research studies have been done for customer churn prediction problem using ensemble classifiers, also, different models have proposed, to predict customer churn prediction problem, but, the issue still has been developing and researchers working on it, more studies been innovating new techniques and models for solving the problem, and some grate insights found out, genetic programming approach is one of proposed models to solve churn prediction problem.

3. MATERIAL AND METHODOLOGIES

Every research must have methodology to follow, and to answer how the research data collected and analyzed, we will discuss under this topic the data that we have collected to use on our research, and also the methodology that we have proceeded to get accurate results and solve the telecom churn prediction which was our research question.

3.1 Data Description

As we have been working on telecom churn prediction problem, we did not find any data provided by any company, then we have tried to find the available data sets on research platforms and on the internet. Fortunately, we have found Orange telecom data set which have been used in some of the previous researches on the field.

In our research, we have used the Orange Telecom Dataset which we have found from Kaggle platform (https://www.kaggle.com/mnassrib/telecom-churn-datasets), and we tested and trained the data by using Weka software which is a program for analyzing and interoperating the results of research data. We have trained and tested our dataset number of methodologies that we have applied and got different results. We will discuss the methodologies that we have applied on our research, one by one with deep details and explanations with the visual results that we have got during our research.

3.1.1 Brief history of Orange Telecom Company

Orange S.A is a France telecom company, which operates in several countries in the world wide. Orange company has 266 million customers over the worldwide, it also has around 145000 employees working on the company in French and other branches in the world. Orange is the tenth largest telecom company in the world, and the fourth largest in Europe. Orange Telecom company headquarter office is located in Paris. Orange company has been providing services like mobile, landline, internet, and Internet Protocol Television IPT services.

The company founded 1st January 1988 as France Telecom. The Orange name which is the official known name of the company was rebranded in 1st July, 2013. As the company is multinational company, and operates several countries over the worldwide, it has number of subsidiaries such as Orange UK, Telekom Kenya, Telekom Romania, Orange Marine, and more other branches in the worldwide.

3.1.2 Orange Telecom Dataset

Orange Dataset which is the data we used this study, is available on different machine learning resources sites and Data mining platforms. The one we used in this study specially, available in Kaggle.com, got and downloaded from there. Kaggle is website which provides different resources of machine learning and data sciences, datasets and other study resources.

Orange Dataset contains a cleaned customer activity data. The dataset file which used this study is telecom churn dataset. Each row represents a customer each; each column contains customer's attributes. The dataset has attributes, as list and described below.

Table 3.1. Dataset Attributes and its descriptions

Attribute Name	Data Type	Attribute Description
State	string	Customer's primary residence or location.
International plan	string	International plan of the customer, whether it is weekly, monthly, or yearly.
Total day minutes	double	The total call minutes made by the customer in the day.
Total eve minutes	double	The total call minutes made by the customer in the evening.
Total night minutes	double	The total call minutes made by the customer in the night.
Total intl minutes	double	Total international call minutes made by the customer.
Customer service calls	integer	Number of calls customer did to connect to the customer service center.

Account length	integer	How many accounts the customer has.
Voice mail plan	string	Customers' voice mail plan.
Total day calls	integer	The total calls made by the customer in the day.
Total eve calls	integer	The total calls made by the customer in the evening.
Total night calls	integer	The total calls made by the customer in the night.
Total intl calls	integer	The total international calls made by the customer.
Churn	string	Describes the customers status after the evaluated data whether its churn (to leave the company) or not.
Area code	integer	Geographic area code of the customer.
Number vmail messages	integer	The number of voice mail messages made by the customers.
Total day charge	double	The total day charges made by the customer.
Total eve charge	double	The total evening charges made by the customer.
Total night charge	double	The total night charges made by the customer.
Total intl charge	double	The total of the international charges made by the customer.

3.2 Methodologies

3.2.1 Naïve Bayes

Naïve Bayes is a classifier based on Bayes Theorem. This method uses independent assumption, which it is getting the results weather you have some missing values or lost some data, it is not making values and some parts of that depend on others, which will make the progress complicated and not proceed, instead it would process the available assumptions independently to produce accurate results. If some features missed or unknown, so other features could progress the process they not be depend on others to get or to know, thee existing could contribute enough the probability. Naïve Bayes method is easy to use also, and very useful for large data sets, large sized data sets and complex prediction problems use this method for its ability of easy to use. Bayes theorem provides a kind of calculating posterior probability, which is well known and useful way to calculate and to do probabilities and predictions. Naïve Bayes performs well in multi class prediction, as we mentioned above because it is useful for processing large sets of data, it is because of the ability of multi class prediction it has. It is better than other models and methods in terms of performance as its independence assumption. There are four applications of Naïve algorithms which are, first one is real time prediction, the second one is multi class prediction, the third one is text classification or spam filtering also known as sentiment analysis, and the fourth and the last one is recommendation system. The detailed information about this method can be found two text books which are mentioned the references list, one is "Introduction to Data Mining" written by Pang-Ning Tan, Michael Steinbach, and Vipin Kumar., and the other one is "Data Mining: Practical Machine Learning Tools and Techniques Morgan Kaufmann" written by H. Witte, E. Frank, M. Hill, and C. Pal.

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)} \tag{1}$$

Formulation of Naive Bayes

$$P(X|Y = y) = \prod_{i=1}^{d} P(Xi|Y = y), \tag{2}$$

3.2.2 J48

J48 classification algorithm is one of machine learning algorithm it is Java implementation of C4.5 methods in Weka. This classifier is used for predictions and to solve classification related problems, it generates a decision tree to list down in the nodes of the tree the assumptions and draw out the possibilities. J48 classifier results well accurately comparing with other classification algorithms.

Hunt's Algorithm

In Hunt's algorithm, a decision tree is grown in a recursive fashion by partitioning the training records into successively purer subsets. Let D_t be the set of training records that are associated with node t and $y = \{y_1, y_2, \ldots\ldots y_c\}$ be the class labels. The following is a recursive definition of Hunt's algorithm.

Step1: If all the records in D_t belong to the same class $y_{t,}$ then t is a leaf node labeled as $y_{t.}$

Step2: If D_t contains records that belong to more than one class, an attribute test condition is selected to partition the records into smaller subsets. A child node is created for each outcome of the test condition and the records in D_t are distributed to the children based on the outcomes. The algorithm is then recursively applied to each child node.

3.2.3 Random forests

Random Forest is a machine learning algorithm, which used for classification. Forest means compressed, so, Random Forests works as to create more decision trees, compressed to find out the best possible solution. This classification algorithm is flexible which can do both of regression and classification.

Random forests algorithm has different applications include: recommendation engine, image classification and feature selection. It is highly accurate and robust method. It uses mean decrease impurity (MDI) for calculation of the important of each feature. It contains set of multiple trees. The detailed information about this method can be found two text books which are mentioned the references list, one is "Introduction to Data Mining" written by Pang-Ning Tan, Michael Steinbach, and Vipin Kumar., and the other one is " Data Mining: Practical Machine Learning Tools and Techniques Morgan Kaufmann" written by H. Witte, E. Frank, M. Hill, and C. Pal.

3.2.4 K Nearest Neighbor (kNN)

K Nearest Neighbor is a machine-learning algorithm, which can use for both classification and regression. KNN is an algorithm, which classifies data point according to the similar relation of it. So many applications tried with this method because of its effectiveness, non-parametric and easy to implementation properties. It's a classification, which classifies instances based on their similarities. It is called in Weka IBK. The detailed information about this method can be found two text books which are mentioned the references list, one is " Introduction to Data Mining" written by Pang-Ning Tan, Michael Steinbach, and Vipin Kumar., and the other one is " Data Mining: Practical Machine Learning Tools and Techniques Morgan Kaufmann" written by H. Witte, E. Frank, M. Hill, and C. Pal.

3.2.5 Logistics regression

Logistic Regression is a machine-learning algorithm, which is used to allocate examinations to a separate group of classes. İt works as binary classification model, which uses mostly to result two possible results, for example yes or no. This classification algorithm used by most of online transaction dealing companies, because of it is simplicity and easiness.

The detailed information about this method can be found two text books which are mentioned the references list, one is " Introduction to Data Mining" written by Pang-Ning Tan, Michael Steinbach, and Vipin Kumar., and the other one is " Data Mining: Practical Machine Learning Tools and Techniques Morgan Kaufmann" written by H. Witte, E. Frank, M. Hill, and C. Pal.

3.2.6 Decision stump

Decision stump is a machine-learning algorithm, which structurally contains one main node of decision tree. That main single node connects directly to the leaves that are the rest of the tree structure in the decision stump.

3.2.7 Logistic model tree LMT

Logistic Model Tree LMT is a machine-learning algorithm, which is used for classification and to solve churn prediction problems. İt works together combined with logistic regression and decision tree.

3.2.8 AdaboostM1

AdaboostM1 classifier is the first successful boosting sophisticated for binary classification. Adaboost used to improve and support the execution of any machine learning algorithm.

3.2.9 Principal component analysis (PCA)

Principal Component Analysis PCA classifier is a popular technique that used for today's research areas of pattern recognition and visual classifications. This statistic method allows for minimizing the distance of large sets of data and summarizes in to small visual patter, which can be easily shown visually. PCA is strong data analysis, as it has functional computational technique. Formulation

3.2.10 Gain ratio

Gain Ratio is amendment of information gain that minimizes its bias. Gain Ratio improves the information gain by ensuring how much information needed to express which branch a sample belongs to.

3.2.11 Information gain

Information Gain determines the amount of information about the class that an attribute can provide, so, the attribute that has highest information gain will split first. Information Gain is crucial solution that decision tree algorithm uses.

3.2.12 Bagging

Bagging is a machine-learning algorithm that created to develop the settlement and precision of machine learning algorithms used in regression and classification.

3.2.13 WEKA

Weka is workbench is a set of algorithms and data mining appliance, which used to process large datasets to solve data mining problems. This data-mining tool is used for different fields and activities.

Weka stands for Waikato Environment for Knowledge Analysis. It was developed by Waikato University in New Zeland, in that time designed for New Zeland's agricultural sector; but currently it has been using in so many different places and areas to solve data mining problems.

Weka works on nearly any platform, and has been tested under Linux, Windows, and Macintosh operating systems. Weka has methods for all data mining problems to solve, these methods include regression, classification, clustering, association rule mining, and attribute selection.

Weka tool has well designed interface, which allows users to use very easy, as it is designed visualized graphical form, and the user just clicks the buttons which has clear human understandable written names. Users can easily access, whether they can easily upload the dataset file whether the file is arff format or excel format, then can use the buttons to apply the activities they want, and to try and test the dataset with the different methods, to get accurate results.

The below figure is GUI Graphic User Interface chooser:

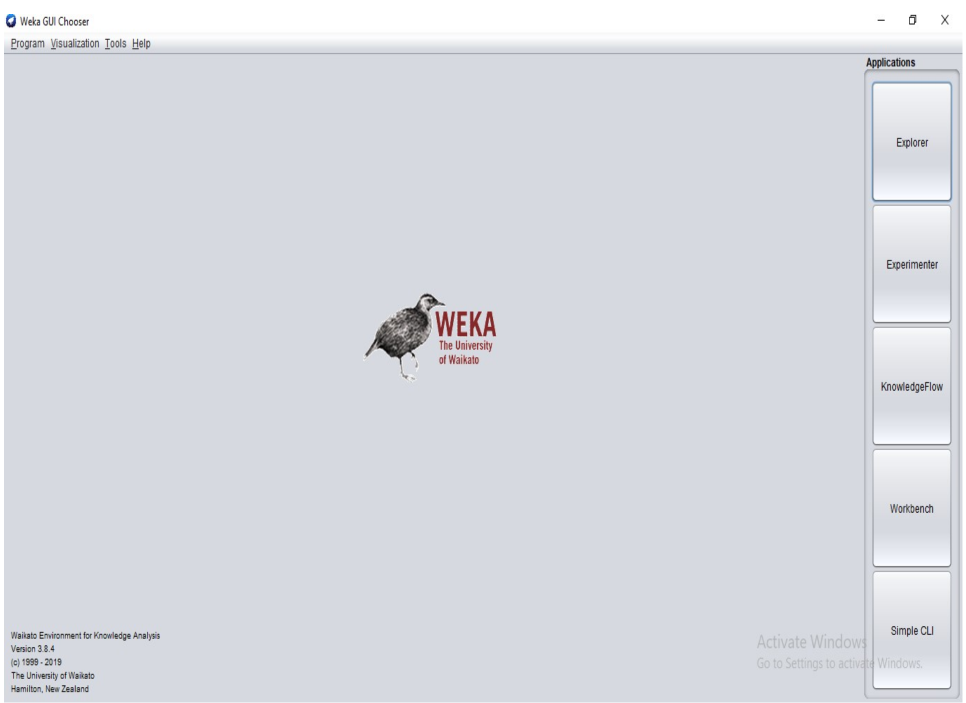

Figure 3.1. WEKA GUI

In our research we have used the Explorer Application, and most of the studies initially used in that, as it contains most of the data mining algorithms and enables the user to do data mining activities.

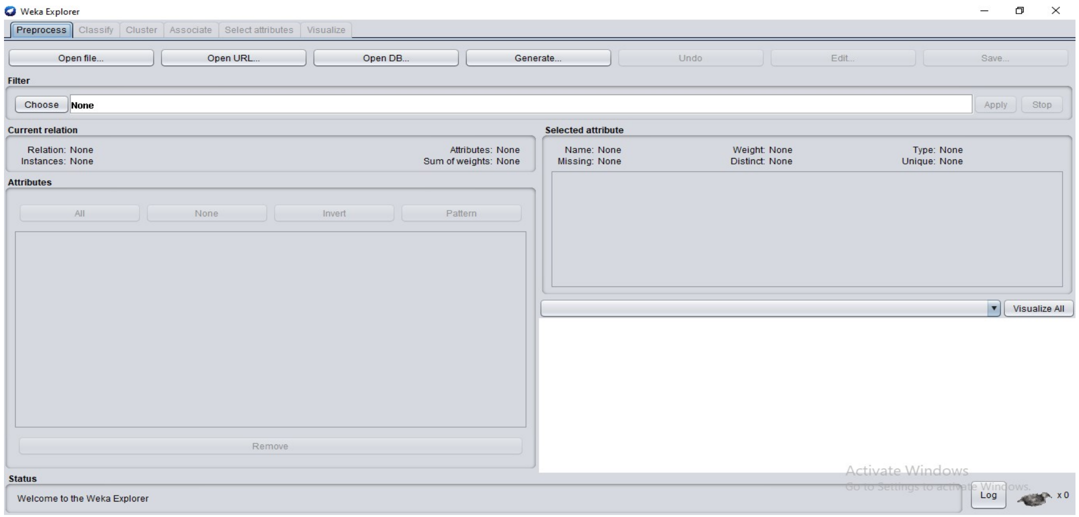

Figure 3.2. WEKA Explorer interface

4. EVALUATIONS AND DISCUSSION

In all evaluation process WEKA Software is utilized for implementation. As validation, K-fold cross validation is used to obtain confusion matrix. With Confusion matrix many evaluation metric is calculated namely: True Positive Rate (TP Rate/Recall), False Positive Rate (FP Rate), Precision, F-1 Score (F Measure), MCC (Matthews correlation coefficient), ROC (Receiver Operator Characteristic) Area and PRC (Precision Recall Curve) Area. The formulations of TP Rate, FP Rate, Precision, F1 Score and MMC metrics are given from Eq. 1-5 respectively. In these Equations TP means True Positive, TN means True Negative, FP means False Negative and FN means False negative. For further description one may consult (Wikipedia, 2021)

$$TP\ Rate = \frac{TP}{TP+FN} \tag{1}$$

$$FP\ Rate = \frac{FP}{FP+TN} \tag{2}$$

$$Precision = \frac{TP}{TP+FP} \tag{3}$$

$$F1 = \frac{2TP}{2TP+FP+FN} \tag{4}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \tag{5}$$

4.1 Implementation Results Evaluation

Table 4.1. indicates the results of Naïve Bayes classifier in combination with AdaBoost M1, PCA, Gain Ratio, Bagging and Info Gain is used to classify the churn data set. As it could be seen from the Table 4.1. For Naïve Bayes best result is obtained when Gain Ratio is used in conjunction with as indicated in the table.

Table 4.1. Naïve Bayes Classifier with ensemble methods

Attribute Selector	TP Rate	FP Rate	Precision	F-Measure	MCC	ROC Area	PRC Area
Ada boostM1	0.85	0.534	0.84	0.844	0.342	0.75	0.851
PCA	0.859	0.849	0.879	0.795	0.095	0.651	0.818
GainRAiot	0.882	0.459	0.872	0.875	0.471	0.793	0.879
InfoGain	0.882	0.459	0.872	0.875	0.471	0.793	0.879
Bagging	0.879	0.485	0.867	0.871	0.448	0.796	0.881

Table 4.2 indicates the results of J48 classifier in combination with AdaBoost M1, PCA, Gain Ratio, Bagging and Info Gain is used to classify the churn data set. As it could be seen from the Table 4.2. for J48 best result is obtained when Bagging is used in conjunction with as indicated in the table.

Table 4.2. J48 Classifier with ensemble methods

Ensemble Method/ Attribute Selector	TP Rate	FP Rate	Precision	F-Measure	MCC	ROC Area	PRC Area
Ada boostM1	0.894	0.5	0.882	0.882	0.495	0.867	0.915
PCA	0.841	0.579	0.827	0.833	0.289	0.652	0.807
GainRAiot	0.916	0.33	0.911	0.913	0.634	0.827	0.898
InfoGain	0.916	0.33	0.911	0.913	0.634	0.827	0.898
Bagging	0.934	0.309	0.931	0.93	0.708	0.9	0.943

Table 4.3. indicates the results of Random Forest classifier in combination with AdaBoost M1, PCA, Gain Ratio, Bagging and Info Gain is used to classify the churn data set. As it could be seen from the Table 4.3. for Random Forest best result is obtained when Gain Ratio is used in conjunction with as indicated in the table.

Table 4.3. Random Forest Classifier with ensemble methods

Ensemble Method/ Attribute Selector	TP Rate	FP Rate	Precision	F-Measure	MCC	ROC Area	PRC Area
Ada boostM1	0.88	0.696	0.877	0.845	0.369	0.777	0.871
PCA	0.861	0.831	0.845	0.802	0.135	0.764	0.861
GainRAiot	0.871	0.724	0.853	0.834	0.295	0.82	0.895
InfoGain	0.871	0.724	0.853	0.834	0.295	0.82	0.895
Bagging	0.87	0.759	0.859	0.825	0.27	0.816	0.894

Table 4.4. indicates the results of kNN classifier in combination with AdaBoost M1, PCA, Gain Ratio, Bagging and Info Gain is used to classify the churn data set. As it could be seen from the Table 4.4. For kNN best result is obtained when Bagging is used in conjunction with as indicated in the table.

Table 4.4. KNN k=1 Classifier with ensemble methods

Ensemble Method/ Attribute Selector	TP Rate	FP Rate	Precision	F-Measure	MCC	ROC Area	PRC Area
Ada boostM1	0.82	0.697	0.793	0.804	0.147	0.554	0.772

PCA	0.811	0.787	0.76	0.782	0.033	0.512	0.755
GainRAiot	0.815	0.736	0.778	0.793	0.101	0.54	0.763
InfoGain	0.815	0.736	0.778	0.793	0.101	0.54	0.763
Bagging	0.832	0.721	0.794	0.809	0.148	0.638	0.809

Table 4.5. indicates the results of Logistic Regression classifier in combination with AdaBoost M1, PCA, Gain Ratio, Bagging and Info Gain is used to classify the churn data set. As it could be seen from the Table 4.5. For Logistic Regression best result is obtained when Gain Ratio is used in conjunction with as indicated in the table.

Table 4.5. Logistic Regression Classifier with ensemble methods

Ensemble Method/ Attribute Selector	TP Rate	FP Rate	Precision	F-Measure	MCC	ROC Area	PRC Area
Ada boostM1	0.847	0.666	0.817	0.827	0.238	0.62	0.801
PCA	0.847	0.675	0.816	0.826	0.231	0.701	0.831
GainRAiot	0.847	0.666	0.817	0.827	0.238	0.688	0.821
InfoGain	0.847	0.666	0.817	0.827	0.238	0.688	0.821
Bagging	0.844	0.684	0.812	0.822	0.214	0.681	0.814

Table 4.6. indicates the results of Decision Stump classifier in combination with AdaBoostM1, PCA, Gain Ratio, Bagging and Info Gain is used to classify the churn data set. As it could be seen from the Table 4.6. For Decision Stump best result is obtained when Bagging is used in conjunction with as indicated in the table.

Table 4.6. Decision Stump Classifier with ensemble methods

Ensemble Method/ Attribute Selector	TP Rate	FP Rate	Precision	F-Measure	MCC	ROC Area	PRC Area
Ada boostM1	0.858	0.638	0.831	0.838	0.291	0.799	0.874
PCA	0.858	0.849	0.808	0.795	0.056	0.614	0.795
GainRAiot	0.853	0.613	0.831	0.838	0.296	0.631	0.803
InfoGain	0.853	0.613	0.831	0.838	0.296	0.631	0.803
Bagging	0.865	0.672	0.838	0.839	0.295	0.737	0.845

Table 4.7. indicates the results of LMT classifier in combination with AdaBoostM1, PCA, Gain Ratio, Bagging and Info Gain is used to classify the churn data set. As it could be seen from the Table 4.7. For LMT best result is obtained when Bagging is used in conjunction with as indicated in the table.

Table 4.7. LMT Classifier with ensemble methods

Ensemble Method/ Attribute Selector	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area
Ada boostM1	0.906	0.393	0.898	0.906	0.9	0.577	0.82	0.903
PCA	0.862	0.672	0.833	0.862	0.836	0.282	0.76	0.867
GainRAiot	0.906	0.376	0.899	0.906	0.901	0.583	0.859	0.918
InfoGain	0.906	0.376	0.899	0.906	0.901	0.583	0.859	0.918
Bagging	0.919	0.373	0.914	0.919	0.913	0.634	0.886	0.93

4.2 Discussion

In order to compare our study with others, which had used the same dataset, Table 4.8. is compiled. As it could be seen from the tablet that proposed methods are as good as the other methods and in some aspect it is better. On the other hand since in this study many comparison metrics are used only a few of them were common with others.

Table 4.8. Comparison with other studies

	TP Rate	F1 Geometric average Sensitity and Precision	Method Orange Data set
(Jain et al. 2020)	85.2385 %	%98	Logistic regression and Logistic Boost
(Azeem et al. 2018)	98%*	-	Fuzzy based Classifiers
	95%	85%	Random forest,
This Study	93%	90%	J48,

5. CONCLUSION

As we reached the conclusion chapter of this study, which was discussed churn prediction in telecommunication sector as research question. During this study, in the first pages of the introduction chapter discussed the problem of churn prediction which most of service providing companies facing, especially telecommunication companies which was the case study of this research. This study collected several studies which conducted before and discussed the churn prediction problem, where mentioned different tools and technologies which previous studies declared and found out as a solution to the churn prediction problem. Literature studies tested out different algorithms, and machine learning techniques to take part of the solution of telecom churn prediction problem. In the methodology chapter of this study we used Orange Telecom dataset as case study dataset, which we implemented in Weka software with number of ensemble classification methods, such as decision tree, naïve Bayes, logistic regression, j48, decision stump, random forest, and logistic model trees. We found out different results, and we compared with previous studies.

In this study, data mining classification is performed for Churn analysis in Telecom sector. In order to better understand and clarify the effects of the different methods namely J48, Naïve Bayes, Logistic Regression, Random Forest and Decision Dump they are applied on the data set. Moreover, ensemble methods such as AdaBoostM1, Bagging, PCA, Gain Ratio, and Info Gain are used in conjunction with aforementioned methods. It has been shown that better results are obtained in classification of the churn data set when applied with the ensemble methods. Bagging together with J48 has better result than the compared ones. Success of the methods can be attributed to the implementation and algorithm of the methods. While J48 is entropy based decision tree, Bagging is simply uses bootstrap aggregating. These results also compared with some of the studies using the same data set as well. The contribution of the study is introducing new aspect of analysis and better understanding to the Churn data set. To the best of our knowledge there is no other study that utilizes comparative results of aforementioned methods.

Finally, this study as it is not the first study conducted with the telecom churn prediction, as it will not be the last one. The study may have some changes, enhancements, and development in the future as the problem is still not solved completely. Here, we finally came to conclude this study, as we wishing to take part of solving telecom churn prediction problem.

REFERENCES

Abbasimehr, H., Setak, M., & Tarokh, M. J. (2011). A neuro-fuzzy classifier for customer churn prediction. International Journal of Computer Applications, 19(8), 35-41.

Amin, A., Al-Obeidat, F., Shah, B., Adnan, A., Loo, J., & Anwar, S. (2019). Customer churn prediction in telecommunication industry using data certainty. Journal of Business Research, 94, 290-301.

Azeem, M., & Usman, M. (2018). A fuzzy based churn prediction and retention model for prepaid customers in telecom industry. International Journal of Computational Intelligence Systems, 11(1), 66-78.

Awodele, Oludele., Adeniyi, Ben., Ogbonna, A.C., Kuyoro, S.O., Ebiesuwa, Seun. (2020). Enhanced churn prediction in the Telecommunication Industry, International Journal of Innovative Research in Computer Science & Technology (IJIRCST), 8, 6-15.

Chouiekh, A. (2020). Deep Convolutional Neural Networks for Customer Churn Prediction Analysis. International Journal of Cognitive Informatics and Natural Intelligence (IJCINI), 14(1), 1-16.

Dumitrache, A., & MAER MATEI, M. M. (2019). Churn Analysis in a Romanian Telecommunications Company. Postmodern Openings/Deschideri Postmoderne, 10(4), 44-53.

Faris, H. (2018). A hybrid swarm intelligent neural network model for customer churn prediction and identifying the influencing factors. Information, 9(11), 288.

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

Idris, A., Khan, A., & Lee, Y. S. (2013). Intelligent churn prediction in telecom: employing mRMR feature selection and RotBoost based ensemble classification. Applied intelligence, 39(3), 659-672.

Khan, Y., Shafiq, S., Naeem, A., Hussain, S., Ahmed, S., & Safwan, N. (2019). Customers churn prediction using artificial neural networks (ANN) in telecom industry. (IJACSA) International Journal of Advanced Computer Science and Applications, 10(9), 132-142.

Kirui, C., Hong, L., Cheruiyot, W., & Kirui, H. (2013). Predicting customer churn in mobile telephony industry using probabilistic classifiers in data mining. International Journal of Computer Science Issues (IJCSI), 10, 1694-0784.

Landwher and Frank., (2004). Logistic Model Trees LMT, 54, Germany.

Nijhawan, V. K., Madan, M., & Dave, M. (2019). An Analytical Implementation of CART Using RStudio for Churn Prediction. Information and communication technology for competitive strategies, 40, 109-120.

Özmen, M., Aydoğan, E. K., Delice, Y., & Toksarı, M. D. (2020). Churn prediction in Turkey's telecommunications sector: A proposed multiobjective–cost-sensitive ant colony optimization. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 10(1), e1338.

Pang-Ning Tan, Michael Steinbach, Vipin Kumar., 2014. Introduction to Data Mining, 719, USA.

Tsai, C. F., & Lu, Y. H. (2009). Customer churn prediction by hybrid neural networks. Expert Systems with Applications, 36(10), 12547-12553.

Umayaparvathi, V., & Iyakutti, K. (2012). Applications of data mining techniques in telecom churn prediction. International Journal of Computer Applications, 42(20), 5-9.

Ullah, I., Raza, B., Malik, A. K., Imran, M., Islam, S. U., & Kim, S. W. (2019). A churn prediction model using random forest: analysis of machine learning techniques for churn prediction and factor identification in telecom sector. IEEE Access, 7, 60134-60149.

Verbeke, W., Dejaeger, K., Martens, D., Hur, J., & Baesens, B. (2012). New insights into churn prediction in the telecommunication sector: A profit driven data mining approach. European journal of operational research, 218(1), 211-229.

Vijaya, J., & Sivasankar, E. (2018). Computing efficient features using rough set theory combined with ensemble classification techniques to improve the customer churn prediction in telecommunication sector. Computing, 100(8), 839-860.

Wikipedia Confusion Matrix https://en.wikipedia.org/wiki/Confusion_matrix: last access 15.01.2021.

Witten, I. H, E. Frank, M. Hill, C. Pal, Data Mining: Practical Machine Learning Tools and Techniques Morgan Kaufmann; 4. Press (17 November 2016).

Yabas, U., Cankaya, H. C., & Ince, T. (2012, July). Customer churn prediction for telecom services. In 2012 IEEE 36th Annual Computer Software and Applications Conference, 358-359.

Yu, R., An, X., Jin, B., Shi, J., Move, O. A., & Liu, Y. (2018). Particle classification optimization-based BP network for telecommunication customer churn prediction. Neural Computing and Applications, 29(3), 707-720.

Zhong, J., & Li, W. (2019). Predicting customer churn in the telecommunication industry by analyzing phone call transcripts with convolutional neural networks. In Proceedings of the 2019 3rd International Conference on Innovation in Artificial Intelligence, 1, 93-103.

BIOGRAPHY

Name and Surname : Faiza HASSAN MOHAMED

Place and Date of Birth : Mogadisho, Somalia, 25/7/1993

Foreign Languages : English, Arabic, Turkish

Email : Faaizadiini93@gmail.com

Education Status SIMAD University, Faculty of Computing
Department of Information Technology, 2017

Work Experience

School Teacher	2015
Term Video editor and Multimedia officer at JEEGAAN Company	2017
Web Administrator at IFTIN Foundation	2017

Publications

Mohamed, F.H, Kasapbaşı, M., 2021. Churn Prediction with Ensemble Classifiers for Telecom Sector, Technologies and Applied Science Journals, In Press.