MARMARA UNIVERSITY

INSTITUTE FOR GRADUATE STUDIES
IN PURE AND APPLIED SCIENCES

A COMPARATIVE STUDY FOR CUSTOMER CHURN ANALYSIS VIA MACHINE LEARNING ALGORITHMS

ZEYNEP UYAR ERDEM

MASTER THESIS
Department of Industrial Engineering

Thesis Supervisor
Prof. Dr. Seniye Ümit OKTAY FIRAT

Thesis CO- Supervisor
Arş. Gör. Dr. Banu ÇALIŞ USLU

ISTANBUL, 2021

MARMARA UNIVERSITY

INSTITUTE FOR GRADUATE STUDIES
IN PURE AND APPLIED SCIENCES

A COMPARATIVE STUDY FOR CUSTOMER CHURN ANALYSIS VIA MACHINE LEARNING ALGORITHMS

ZEYNEP UYAR ERDEM
524416022

MASTER THESIS
Department of Industrial Engineering

Thesis Supervisor
Prof. Dr. Seniye Ümit OKTAY FIRAT

Thesis CO- Supervisor
Arş. Gör. Dr. Banu ÇALIŞ USLU

ISTANBUL, 2021

MARMARA UNIVERSITY

INSTITUTE FOR GRADUATE STUDIES IN PURE AND APPLIED SCIENCES

Zeynep UYAR ERDEM, a Master of Science student of Marmara University Institute for Graduate Studies in Pure and Applied Sciences, defended her thesis entitled "**Customer Churn Analysis via Optimization Techniques**" on 27.09.2021, and has been found to be satisfactory by the jury members.

Jury Members

………… (Advisor)
Marmara University ....(SIGN) ....

…………….. (Jury Member)
Marmara University ....(SIGN) ....

Assist.Prof. Adı SOYADI (Jury Member)
Marmara University ....(SIGN) ....

APPROVAL

Marmara University Institute for Graduate Studies in Pure and Applied Sciences Executive Committee approves that Zeynep UYAR ERDEM be granted the degree of Master of Science in the Department of Industrial Engineering on …., 2021. (Resolution no: 12345).

Director of the Institute
Prof. Dr. Bülent Ekici

TABLE OF CONTENTS

ÖZET

OPTİMİZASYON TEKNİKLERİNİ KULLANARAK MÜŞTERİ KAYIP ANALİZİ

Bu tezin amacı, etkili bir müşteri kayıp tahmini (CCP) metodolojisine göre makine öğrenme algoritmalarının değerlendirilmesinin açıklayıcı bir analizini sağlamaktır. Hızla gelişen Müşteri İlişkileri Yönetimi (CRM) alanında, müşteri kaybetme eğiliminde olan müşterileri elde tutmak için uygun bir CCP metodolojisi önermek için, müşterilerden gelen hacimli bir veri setinden müşteri kaybını tahmin etmek için bir dizi veri madenciliği analizi yapılmıştır. Bu analizlerde açık kaynaklı bir veri madenciliği yazılımı olan WEKA'da oluşturulan makine öğrenimi (ML) algoritmaları kullanılmıştır. Ayrıca, sınıflandırma yaparken algoritmanın takip ettiği karar yolunu göstermek için bir karar ağacı modelinin Python uygulaması yapılmıştır. Çalışma boyunca, Türkiye'deki özel bir telekomünikasyon şirketinden sırasıyla 195712, 32905 ve 228617 olay sayılarına sahip bireysel, kurumsal ve birleşik veri kümeleri kullanılarak müşteri kaybı tahminine ilişkin bir dizi deneysel analiz yapılmıştır. Müşteri kayıp durumunun tahmini için altı veri madenciliği algoritması değerlendirildi: Lojistik Regresyon, Naive Bayes, J48 ve RandomForest, Bagging ve Boosting gibi ELM algoritmaları. RandomForest, RandomTree'yi kullanırken, Bagging temel öğrenme algoritması olarak J48'i kullanıyor. Deneysel analizler, CCP için uygulanan bu tür veri madenciliği analizlerine dayalı olarak gelecekteki müşteri kayıplarının olasılığının belirlenmesi için bazı karar ağaçlarının ve topluluk ML sınıflandırıcılarının etkinliğini doğrulamak için şirketin tarihsel veritabanından elde edilen gerçek dünya veri kümeleri ile gerçekleştirilir. Sonuçlar, J48'in tüm veri kümelerine göre Naive Bayes'ten daha iyi performans gösterdiğini ve Lojistik Regresyon sınıflandırıcı algoritmasına çok benzer sonuçlar verdiğini göstermektedir. Ayrıca, Bagging büyük boyutlu veritabanını çözmediğinden ve J48, bireysel ve eksiksiz veri setlerinde benzer doğru sonuçlar verdiğinden, müşteri kaybı tahmini için Bagging'in yanı sıra J48 karar ağacı sınıflandırıcısı da seçilebilir.

ABSTRACT

CUSTOMER CHURN ANALYSIS VIA OPTIMIZATION TECHNIQUES

The purpose of this thesis is to provide a descriptive analysis of the assessment of machine learning algorithms to an effective customer churn prediction (CCP) methodology. In the rapidly developing field of Customer Relation Management (CRM), to propose a convenient CCP methodology in order for retaining the customers who tends to churn, a set of data-mining analyses has been conducted to predict customer churn from a bulky dataset from customers with certain attributes in a telecommunication company by using machine learning (ML) algorithms built in an open-source data mining software, WEKA. Additionally, a Python implementation of a decision tree model has been made in order to show the followed decision path of the algorithm while classifying. Throughout the study, a set of experimental analyses with regards to customer churn prediction are conducted by using residential, corporate and combined datasets with number of incidences of 195712, 32905, and 228617 respectively from a private telecommunication company in Turkey. Six data mining algorithms have been evaluated for prediction of the customer churn status: Logistic Regression, Naive Bayes, J48, and ELM schemes such as RandomForest, Bagging, and Boosting. RandomForest uses RandomTree whereas the Bagging uses J48 as a base learner. The experimental analyses are conducted with real-world datasets acquired from the historical database of the company to validate the effectiveness of some decision tree and ensemble ML classifiers for determination of likelihood of future churning customers based on such data mining analyses implemented for CCP. The results show that the J48 outperforms Naïve Bayes based on all datasets and it provides very similar results as the Logistic Regression classifier scheme. Besides, since Bagging has not solved the large-sized database and J48 has given similar accurate results in the residential and complete data sets, J48 decision tree classifier can be chosen as well as Bagging for customer churn prediction.

DECLARATION OF NOVELTY

SYMBOLS

P **:** Possible outcomes

a_i **:** Input variable

w_i **:** Weight of the input variable

s_i **:** test quality

ABBREVIATIONS

ANN	**:** Artificial Neural Network
CCP	**:** Customer Churn Prediction
ML	**:** Machine Learning
Pr	**:** Probability
Exp	**:** Exponential

LIST OF FIGURES

LIST OF TABLES

1 INTRODUCTION

The concept of customer loss analysis is essential, especially in the telecommunication and banking sectors, which have a membership-based income model. These kinds of businesses should increase their stability to survive. Thus, they should prevent any income loss. For this instance, we can determine income loss analysis as customer churn analysis. Retain the current customers have increased its importance as the customer experience evolved over the past 20 years. In the past few decades, technology is overgrowing and supplying the increasing needs of customers. Comprehend relevant results from obtained data is requisite to meet customers increasing demands. Therefore, the importance of mass data is growingly increasing.

Data mining is the operation of discovering critical confidential data at the big data sets. Data mining can both use classical statistical methods and machine learning practices. One of the most important disciplines that data mining practices are categorising the customers who are probably lost and predicting it before it happens. Now, it is possible to determine customer profiles correctly, analyze the dataset from customers, and eventually prevent a possible customer loss according to these results of the Customer churn prediction (CCP) analysis. Accordingly, with the coordinated operations of marketing practices, customer experience management, and action plans of information technologies units, both current customers can be retained, and loss of customers can be prevented.

Thus, with the increase of the fidelity of customers and competitive rates of the company, the stability of the company will increase. In these circumstances, the importance of keeping current customers in the company and defining the possible customers who have ideation to replace their companies has increased. Understanding the needs of the customers has been an indispensable step to endure in the business. Identifying the customers who are willing to leave the company costs much to the business. Thus, it can be assumed that customer loss analysis has become an essential analysis, especially for large companies. A growing body of studies was carried out on customer loss in the telecommunication market.

According to one study, customer loss in the Turkish Telecommunication market could be obviated by predictive analysis through the Bayes network. Another study also suggested that customer loss in the Turkey telecommunications market was resolved with CCP analysis with the Bayes network. In the cited article, a casual map has formed the basis of the Bayes network. According to the results of the Bayes network, mean talking rates, monthly bills, the frequency to call other operators and to change their current tariff is determined as essential factors of customer loss (Kisioglu & Topcu, 2011).

Another structural modelling study, conducted with the dynamics of customer loyalty, satisfaction, trust, and the expense of changing the company, revealed that not only customer trust is enough to provide loyalty in Turkish GSM operators. The expenditure to prevent customers from changing their company and their faith in the company along with satisfaction is necessary to ensure long-term customer and company relationships. The result of this analysis stated that the reliability and validity of the changing price modelling are said to exist (Karaçuha et al., 2004). Geppert, in his study, has also implied that customer fidelity, satisfaction, changing cost, and trust factors of the general model include significant statistical relations, and its definition strength is high (Geppert, 2003).

1.1 CUSTOMER CHURN:

The ability of a company to gain or at least maintain a sustainable competitive advantage in any sector where there is intense competition depends on retaining existing customers and creating customer loyalty. The theoretical and empirical studies carried out reveal that customer satisfaction is the main factor in ensuring customer loyalty. However, customer satisfaction alone cannot be found sufficient in ensuring customer loyalty. In order to maintain the customer-brand relationship in the long term, besides satisfaction, additional costs and brand trust that will prevent the customer from turning to competitors must also be established. In this context, the aim of the study is to analyze the relationships modelled among the variables of customer satisfaction, replacement cost and trust, which are thought to affect customer loyalty and interact with each other.

Increasing competition in many sectors and rapid technological developments cause companies that aimed to increase their market share by gaining new customers in the past to focus on preserving their current market share. The main condition for maintaining market share is to create customer loyalty. Companies with a loyal customer portfolio; will earn more income from their customers' repeat purchases, increased price tolerance (the maximum price the customer can give), recommend the brand to consumers around them, and have less tendency to change. In other words, a loyal customer portfolio is both a very important competitive advantage for companies and the basic building block in maintaining their existence (Karaçuha et al., 2004).

Customer churn, one of the popular CRM scales at both company and customer level, refers to the customer shifting his focus out of the company in period t, although the customer makes his last purchase in the t-1 period. As a matter of fact, companies will focus primarily on stable income from their loyal customers while making certain decisions in an increasingly competitive environment, which will make it even more necessary to retain existing customers. At this point, as is widely known in marketing, keeping a customer at hand is much more costly than making a brand new customer loyal. As a matter of fact, what is desired to be achieved in the process called customer loss management is to manage the ratio of customers who are likely to stop choosing the company. That is why the correct management of customer loss will only be possible if the customer who has the potential to give up the company is correctly identified. At this point, the main purpose is to classify customers as customers who will and will not leave (Kaynar et al., 2017).

The concept of customer loss has become more on the plan with the increase in service businesses such as banks and telecommunications companies, changes in their customers' loyalty, and as a result of these changes, the customers abandon their choice of the company. The high potential of companies to attract customers who experience dissatisfaction in the rival company by creating a wide variety of campaigns with very different characteristics is very effective in the high customer loss. Customers have tended to switch to other companies due to access to the latest technologies, customer-friendly staff, low transaction fees, cost of change, advertising impact, geographic proximity and various service offerings. To anticipate this trend will enable the company to identify the customers that are likely to miss out on the company early and organize campaigns on them, as well as provide preliminary information about which points to take into account when reinstating customers who have fled the company (Kaynar et al., 2017).

However, since dissatisfied customers do not show a positive attitude towards a particular brand, customer satisfaction is the main factor in ensuring customer loyalty. In contrast, loyal customers do not mean satisfied customers, the loyalty tendency of satisfied customers increases. However, although customer satisfaction or dissatisfaction may affect subsequent behaviour, it alone may not guarantee customer loyalty. Within this framework, what needs to be done is to create structural barriers that make it difficult for the customer to choose another brand (Karaçuha et al., 2004).

Losing the customer is an important problem for companies where the loss occurs very easily. For example, banks, insurance companies and telecommunications companies. For these sorts of companies, the cost of gaining new customers is increasing day after day, so that a new era has begun in the marketing industry. Accordingly, companies are in search of various programs in order not to lose the customers they have, instead of organizing campaigns to win new customers. The only way to do this is to start taking action before losing the customer. At this point, modelling customer churn has created an important competitive opportunity and a new field of work. With good modelling, the firm reveals which customer is close to being lost and which customer is loyal. Part of modelling consists of measuring and determining customer value (Tosun, 2006).

Customer loss analysis, revenue loss analysis, or, more broadly, the notion of loss analysis is critical in business calculations, particularly in membership-based revenue models like insurance, telecom, and banking. It can be said that it is an analysis method that should be constantly on the spotlight of a business in many different dimensions, up to the market value of a business in financial calculations, in addition to recognizing existing customer profiles and analyzing customer losses, estimating customer loss, and working in coordination with marketing and customer relationship management units accordingly (Seker, 2016).

If a customer ends the contract with a company and begins a new one with another company, it's a churned customer (Richeldi & Perrucci, 2002). Customer churn is closely related to the loyalty of the customer. In nowadays economy, the only way is not to reduce the prices to gain customer loyalty. Accordingly, increase loyalty through the addition of new assets to the products has been a norm in various industries (Rud, 2001).

The aim of the companies to keep the customers in the company is to determine customers with leaving thoughts and cost analysis to keep them in the company. The most important point when making cost analysis is to define the customers with leaving thoughts.

In some cases, it is hard to make a definition. For instance, a customer may start using a credit card of another bank with not closing the current card. To determine the loss of these kinds of customers is to analyze the reduction of their monthly/yearly expenditure (Rud, 2001). Customer churn especially is very problematic for the companies which lose customers quickly — for instance, banks, insurance, and telecommunication companies (Yan et al., 2001).

In today's business world, where competition is intense, many customers leave the current company from which they receive products and services and prefer rival companies that are more suitable for their expectations. The best-known rule in marketing techniques is "approximately five times the cost of acquiring a new customer, the cost of holding existing customers; The cost of gaining a dissatisfied customer is approximately ten times the cost of holding the current customer. Income from existing customers is approximately twice that of new customers. Customer churn management can be defined as a process that guesses potential subscribers who are considering leaving and looks for ways to remove them from cancellation1. Acquiring new customers is an expensive process in the telecommunications market, which is one of the sectors where customer loss is most intense and rapid. The history of the Turkish telecommunications industry dates back to 1840. The meeting with Turkey's GSM technology has been possible on 23 February 1994. Despite being a young sector historically, the Turkish mobile telecommunications market has turned into a giant structure in a very short time. Turkey moved to implement mobile number portability on 9 November 2008. The application attracted great attention from consumers, and the number of subscribers benefiting from mobile number portability increased rapidly in a short time. Thanks to number portability, it is possible to switch to a different operator without changing the GSM number used and without the cost of changing it, thus further accelerating customer losses. Because 'number portability' creates a result that increases competition in the relevant market with its user dependency and its effect of reducing switching costs. Proportionally, 18% of the subscribers have changed their operator without changing their number. After the initiation of number portability, it has become even more important for businesses to gain new subscribers and to keep the current subscriber base. Customer losses in the telecommunications market are evaluated with "customer loss estimation". These analyzes are aimed to identify customers at risk of losing more, investigate the reasons for cancellation and increase the firm loyalty of effective customers. (Gülpınar, 2013).

Until recent years, the approach of selecting "influential customers" based on invoice or revenue histories was frequently used. However, this approach takes into account only the characteristics of individuals. That is, it considers the customers who make the most money to the company as the most effective customers. However, in telecommunication networks, some customers may have more real contributions to the company due to their connections and relationships rather than their individual contributions. When communication communities are considered as Social Networks (SN) where each individual can establish weak or strong relationships with other individuals in the network, Social Network Analysis (SNA) is considered a powerful method to reveal the influence of customers on the network. The customer loss event has a knock-on effect. This chain process can be initiated by a strong node of the SNA or even by a customer who generates less revenue but has a strong influence on the customer loyalty process. For this reason, companies should be able to reveal customers with high impact. Defined as a strong and central node in social network analysis, these customers can be selected by considering how effective they can be in chained events such as "customer loss". Thus, the marketing method to be applied can be determined according to individuals with high effects. The SN centrality measures and the weights of the links between the customers can reveal the effect of the chain process on the network. Understanding customer connections and identifying the central cores of each chain in the SN structure is the best way to avoid mass customer flight and thus avoid lost revenue. SNA reveals possible correlations in customer loss events within the community, demonstrating that the event has a stronger impact when triggered by a core node of SN, whereas less impact when triggered by a peripheral node (Gülpınar, 2015).

1.1.1 Type of Customer Churn

Customer churn can be described in two main chapters. The first can be defined as a voluntary churn, and the second one is an involuntary churn.

Voluntary Churn

Voluntary churn is the case if a customer voluntarily leaves the current company and decides to use another company's services. For instance, if a customer moves his/her choice to another telecommunication company, this loss of the customer can be determined as a voluntary churn. These kinds of situations occasionally occur if the customer is having problems with the current company and the advantages of the new company are more favourable.

Involuntary Churn

This situation develops without the choices of customers and gets affected by territorial circumstances. It can arise in cases like leaving the customer to a foreign country, death, or health problems. The instruments, statistical models, or data mining studies developed to analyze customer loss ignores these types of involuntary losses. The reason for this ignorance is because these losses are generally insuppressible. Most involuntary customer loss analysis includes statistical predictions, yet these methods are tended to predict the loss ratios in a time frame.

1.1.2 Churn Analysis Types

Most assessment agencies calculate and report gross and net loss numbers for churn analysis. For a certain time period during which the study is undertaken, gross loss generally indicates the value acquired from all services and goods to which the lost client brings value. The expense of regaining a client with comparable attributes (same segment, habits, market, and geography) into the system is called net loss. When calculating recurrent monthly revenue (RMR or monthly recurring revenue, MRR), many business intelligence software programs require customer loss analysis. RMR numbers essentially represent the amount of money that a customer will receive each month in a secure manner, so it's critical to look for potentially churned customers when calculating this figure.

Most statistical approaches are established as a forecast model based on historical data, with real practice data gathered on the field included in the model. Customer reviews, service satisfaction surveys, and appeal and dispute statements, for example, all offer valuable data for prediction models. Modern data mining models are far more sophisticated thus making it possible to perform many complex analyzes more easily today. This analysis include customer profiling, customer losses based on quality and demographic differences and interpreting effects of weakest customer.

One of the important values for businesses is the accurate prediction of losses in advance. For this goal, most business intelligence systems construct data mining-based apps. Prediction algorithms researched under data mining, for example, produce excellent results for this problem type. A fall in the ratio of customers and a decrease in the returns for the business are both considered losses for a business. Although the rate varies by industry, it can be claimed that acquiring new consumers is generally more expensive than keeping existing customers in all industries. As a result, it is critical for organizations to be able to conduct timely analyses in order to battle effectively and efficiently in the competitive business environment. In general, these analyses show that outcomes coordinated with customer relationship management and marketing units make significant contributions to enterprises. As a result, the three main criteria listed below play a significant role in loss estimation:

1. Identifying consumers who are about to leave by asking "who" and "how they behave" questions.
2. The possible impact of these customers if they leave, the position to take to keep them, the most efficient use of resources, and the best reflection of value in the customer profile.
3. Making strategies to avoid the separation of consumers who are difficult to forecast and have the ability to depart, in addition to individual actions, and ensuring that the measures to be implemented for these two groups of customers are compatible (because the measure in one of them often negatively affects the other).

1.2 DATA MINING:

Data mining creates the need to integrate different disciplines such as databases, artificial intelligence and statistics. For this reason, the analysis process is usually carried out by high-performance computers and expert users. It is imperative to use the correct technique for each process. Otherwise, unexpected or erroneous results may occur if the data is not cleaned properly and not analyzed well. The fact that the data source is generally very large can cause various problems in terms of system, hardware and time. Since the size of the databases is increasing size, the systems must be designed to handle this growth (Tosun, 2006).

In particular, the size of the raw data stored in the databases of large companies is increasing rapidly. However, raw data alone does not contain very large information. In today's competitive business world, companies need to turn this data into specific insights and useful information to guide marketing, investment and management strategies in a very short time. Data mining is the process of uncovering and discovering exciting information such as relationships, patterns, changes, deviations and trends, specific structures found between data in such large databases or data warehouses. The methods and tools used in data mining help in answering strategic questions about the nature of the job in a very short time. They can transform the patterns and relationships hidden in the raw data into predictive information. Although data mining is still under development, retail, finance, healthcare, shipping, aviation companies and especially banks are using data mining tools to take advantage of historical knowledge. With help, such organizations uncover undetected relationships, trends, expectations, or anomalies. Data mining is frequently used in the business world for the following topics (Tosun, 2006):

1. Market share segmentation: It reveals the common characteristics of customers who buy the same product/services from a company.
2. Loss of Customers: Predicts which customers may leave the company and move on to another competitor.
3. Detecting Fraud: It reveals which customer actions may be involved in fraud.
4. Direct marketing: It reveals which types of customers may have more return in the campaign lists made to gain customers.
5. Interactive marketing: It reveals the areas of interest to the person browsing a web page.
6. Market basket analysis: It reveals which products are most likely to be purchased together.

7. Trend Analysis: Examines the typical customer's behaviour differences at certain times.

Data mining technology can reveal new lines of business using:

1. Predicting trends and behaviour: Data mining helps in revealing predictable information in a large database, for example, predicting a bankruptcy or predicting which segment in a population will show common patterns of behaviour towards certain events.

2. Discovering previously unknown patterns: Data mining tools reveal patterns hidden in large databases. For example, the discovery of the possibility of buying two products that appear unrelated to each other at the same time or the identification of fraud from transactions on credit cards.

The discovery and prediction technique of data mining is called modelling. Modelling is the application of these rules and results on the data in environments with unknown answers, by drawing rules and conclusions from situations where there are known answers (Tosun, 2006).

The raw data stockings in big companies database are rapidly increasing. But raw data only does not contain significant information. These data should be converted useful data to guide marketing, investment, and management strategies in today's competitive business world. Data mining is a revealing and discovery process of unveiling the relationships, patterns, alterations, deviations, and tendencies between big databases and data storage. Consequently, data mining can be described as revealing meaningful information hidden in big data sets. The methods and implements used in data mining help to answer the strategic questions depending on the attribute of the line of work in a quick manner. They can convert the hidden patterns and relationships in raw data to estimated data (Yan et al., 2001). Pattern describing technologies, mathematical and statistical techniques, and data mining help unobtrusive relationships, tendencies, and expectancies for the companies. Various data mining techniques can be used in a data mining operation. Each technique has advantages and disadvantages. Most commonly used methods can be described as decision trees, nerve webs, sheaving algorithms, k-nearest algorithms, genetics algorithms, blurred logic, and web analysis (Han et al., 2014).

In data mining, the following steps are being used for estimation and description:

Cleanse of the data

Conversion of the data

Reduction of the data

Modelling of the data

Data mining

Information presentation

1.2.1 Algorithms for Data Mining

Neural Networks

Neural Networks are used to reveal relationships and patterns in data. This data can be the results of a market survey or results from a different production process. What reason it is used for, Neural Networks does differ from traditional methods. The person who analyzes with traditional methods describes the situations and rules to the computer one by one. Neural Networks do not need this kind of coding process. With the Neural Networks method, a prediction program is designed with raw data after a good testing process (Han et al., 2014).

Neural Networks are intelligent modelling systems that use the results of learning manners analysis with imitating the brain working process. They learn several rules from input values, evaluate the patterns and use them in new data (Tosun, 2006).

ANN are physical cellular systems that receive, store and use experimental information. ANNs are computer systems that can learn events by using "machine learning" and using realized examples and can be successfully applied in subjects such as learning, association, classification, generalization, feature determination, optimization and prediction, similar to the functional properties of the human brain.

Craven & Shavlik (1997) demonstrated that ANN is similar to the human brain in two aspects: First, the information is acquired by the network from its environment during the learning process, and the second is the use of the connection force between neurons known as synaptic to hide the acquired information. ANN is a nonlinear, multidimensional, noisy, complex, inaccurate, incomplete data, and a method that is frequently used when there is no mathematical model or algorithm to solve the problem.

The coming together of nerve cells is not random. Generally, cells form a network by coming together in 3 layers and in parallel within each layer. These layers are; input layer, intermediate layers and output layer. Neurons in the input layer are responsible for taking information from the outside world and transferring it to the intermediate layers. In some networks, there is no information processing at the input layer. Neurons in the intermediate layers process information from the input layer and send it to the output layer. There can be more than one intermediate layer for a network. Neurons in the output layer process the information from the intermediate layer and produce the output that the network must produce for the input set presented from the input layer. The output produced is sent to the outside world. In addition to the advantages listed above, ANN also has some disadvantages. The most important of these are; There are no specific rules in determining the parameter values of the network (learning coefficient, number of neurons in each layer, number of layers, etc.), and there is no method developed to decide when the training of the network will be completed (Karakurt, Erdal, Namli, et al., 2013).

Decision Tree

The decision tree is a data mining method that is used for the classification of the data and makes predictions. The more attracted side of the decision trees from nerve webs is their ability to make new rules. These rules are defined as the users can easily understand, which makes the analysis easier. Decision trees are classifier structures like a tree, and every node in the tree represents a leaf or decision node (Han et al., 2014).

Decision trees are classifiers in the form of trees. Each node in this tree indicates a leaf or decision node. A leaf node is the value of the target attribute. The decision node is the test value to be applied to an attribute, followed by all possible attribute values belonging to that attribute; these values form the branches of the tree.

The decision tree classifies a sample by examining it from root to leaf. Learning algorithms of decision trees use a set of decision trees to present a hypothesis. In the learning set, the raw data is examined and classified in the best possible way. The algorithm repeats this process recursively, and the last decision tree it generated creates the final hypothesis. The ideal decision tree creates the same rules for data outside the learning set or reveals the same hypothesis results with little margin of error.

Below, a decision tree algorithm is shown:

Input Variables: Sample set, attribute-list

ID3 Algorithm (Tosun, 2006):

1. Create nodes
2. If the sample values are all in the same class (Class C)
    a. Return node N as a leaf node in tag C
3. If the attribute-list is empty
    a. Return node N as a leaf node in the most used class tag in the Instance set
4. Select the most informative test attribute from the attribute-list
5. Name the tag of node N the test-attribute
6. Return for each known value of the test-attribute
    a. Draw a branch from node n such that test-attribute = ai
7. Let samples with si = test-quality = ai in the sample set
8. if s_i is blank
    a. Draw a leaf in the most used class label in the sample set
9. If not, return from Decision-Tree-Create (si, attribute-list, test-attribute)
10. Add the node value to the tree.

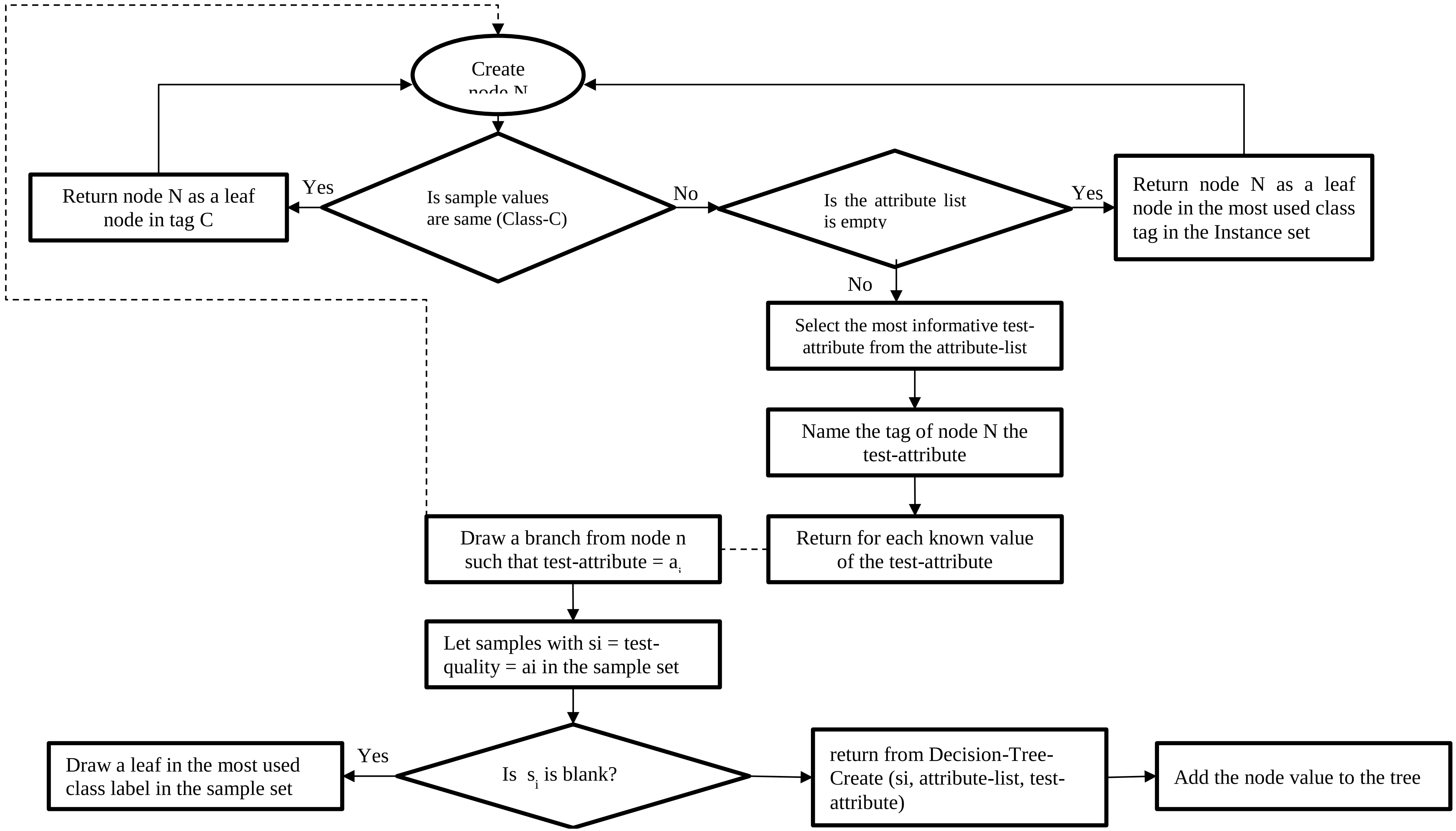

Figure 1.1 Schematic illustration of a version of the ID3 Decision tree algorithm

The algorithm mentioned above is a version of the ID3 algorithm (**Hata! Yer işareti başvurusu geçersiz.**), also known as decision tree induction. The basic strategy is as follows:

1. The tree begins with representing the sample set with a single node.
2. If all sample values in the sample set go to the same class (i.e. target value), this node becomes a leaf and gets the name of this target value. The tree does not grow further in that direction after the leaf value.
3. In the other case, the algorithm chooses the highest knowledge-gain attribute from the attribute list, and this attribute becomes the test attribute.
4. All possible values belonging to the test attribute and the values of the sample set that make up them are added to the tree in the form of a branch. The important point here is that these values are not in the form of continuous variables. In order for the dimensions of the tree to be controllable, continuous variables must be converted into categorical variables.
5. Algorithm In the next step, the program is returned with new sample sets and new attribute lists iteratively until it reaches the leaf node for each sample set value.
6. Recursive rotation stops only under the following conditions:
    a. If all sample set target values belong to the same class
    b. If there is no more list of qualifications for sampling to continue
    c. If there is no set of values left to sample.

Decision Tree Pruning

When a decision tree is created, many branches will have anomalies due to noise and losses in the learning data. The tree's pruning method can help eliminate this problem. This method typically consists of statistically calculating and removing the least reliable branch, resulting in a faster and more reliable classification. There are two pruning methods.

The first of these is the pre-pruning method. In this method, when classifying the learning data, it is decided in advance whether that branch of the tree will continue forward or not. If necessary, after the remaining division, the classification of the remaining data is stopped, and the value with the highest target value is made as a leaf. In this method, a threshold value is determined in advance. Attributes with knowledge gains that do not exceed this threshold value are grouped. While the program continues, at the point where this information gain falls, the other branch is passed without allowing the tree to grow. Either way, determining this threshold is the hardest part. Because if the threshold value is kept too high, the resulting tree becomes a tree consisting of many simple and general rules. If the threshold value is kept too low, the classification of the tree may become very specific, and correct results may not appear on the test data.

The second method is the post-pruning method. It is applied to a fully grown tree. The rules created by all branches are tested, and the branches that create the highest error rate are pruned. Thus, a simpler tree structure can be created. Alternatively, both pre- and post-pruning can be combined and used as a new method. The post-pruning method requires much more calculations than pruning beforehand but produces a more reliable tree (Grąbczewski, 2014).

K-Nearest neighbour algorithms

These algorithms are a kind of clustering. K value represents the number of neighbouring registries. The algorithm connects a number of prototype patterns to the closest neighbour group depending on the right classification. The accuracy of the classification increases with the rise of the K value. Furthermore, there is a need for retrospective data (Han et al., 2014)

Genetic Algorithms

Genetic algorithms work with imitating evolutionary progress. The optimum solution can be achieved by mutation and selection methods. Most favourable solutions are selected in this method, and by mutating them, they are being used to produce more profitable solutions. Generally, genetic algorithms are used in work tariffs or motor designs (Han et al., 2014)

Fuzzy Logic

Instead of certain variables, probable variables are being used in a fuzzy logic method. The value of the feature range from 0 to 1, and it defines the range of the feature to the accurate. This method usually is being used in control systems (Han et al., 2014).

Link Analysis

This method reveals the connections between data. Generally, it is used for market analysis, target-based marketing field or changing the stock prices, which are used to evaluate the relation of the products with the customers (Han et al., 2014)

OLAP

OLAP is used to present mass data and hierarchy. For instance, most variable sales of the products can be presented in a very high range of sales statistics. It can be used by the users to focus more on the more important and valuable. The most significant difference between OLAP and data mining is; OLAP is user-based, and the user makes a hypothesis and tests it with OLAP; on the contrary, in the data mining, the program makes the hypothesis itself. So, a data mining system can reveal a pattern that can be very hard to see without. OLAP gives the opportunity to allow examining the integration, calculations and result comparisons in a graphical format (Han et al., 2014).

In today's competitive business environment, considering that customers have many service provider options with many alternatives, it can be easily seen that they can change the direction of the service they received and their service provider quite easily (Amin et al., 2019). Considering that customers are the most important assets for businesses, the problem faced by companies is how to analyze loyal or lost customers (Kelvin et al., 2020). Specifically, for telecommunications companies, a lost customer can be defined as the person who terminates all business partnerships with the company by closing all of their accounts (Karvana et al., 2019). To retain the current customers – rather than gaining new ones – has increased its importance as the customer experience has evolved over the past 20 years. Therefore, the concept of customer loss analysis (Çiçek & Arslan, 2020) is essential, especially in the telecommunication and banking sectors (Fırat & Biçen, 2003), which have a membership-based income model. These business sectors are required to sustain their stability in order to survive by preventing any income loss (Biçen & Fırat, 2003). To achieve this, issues such as measuring the level of customer loyalty, determining the most important factors affecting customer churn, and predicting customer tendency should be identified through the customer churn analysis. In this context, many prediction models based on machine learning tools such as a decision tree algorithm have been developed in the literature to predict this possible future loss of income (Ahmad et-al., 2019; Al-Mashraie et-al., 2020; Machado et-al., 2019; Sjarif et al., 2019). These analyses are focused on issues such as keeping the customer loyal to the company or retaining the customers from the decision of churn.

When the sustainability of the company is taken into account, the purchasing processes and decision-making processes of the consumer are also required to be carefully examined (Çiçek & Arslan, 2020). Thus, with the increase of the fidelity of customers and competitive rates of the company, the stability of the company will increase. In these circumstances, the importance of keeping current customers in the company and defining the possible customers who have ideation to replace their companies has increased. Customers willing to leave the company have a high cost to the business. For this reason, understanding the needs of customers has been an indispensable step in minimizing the cost to arise from customer losses (Sahu et al., 2018). Customer churn analysis models, to be developed to prevent this loss, have become an indispensable analysis, especially for large companies.

Data mining became very significant in many fields such as supply chain management (Er Kara et-al., 2020), energy management (Akan et-al., 2015), logistic management (Es et-al., 2018), and customer relationship Management (CRM) in the recent years possessing a big portion of new Industry 4.0 applications in the new era leading as well as being led by advanced technology that enables attaining and storing Big Data from customers every single minute all day long. This interaction leads to keen opportunities for researchers to execute elaborate research regarding data mining in the field of CRM (Ćamilović, 2008) using Machine Learning (ML) algorithms (Çınar & Silahtaroğlu, 2013; Mitchell, 1997). Data mining (Er Kara et al., 2020; Özekes, 2003) is the operation of discovering critical confidential data at the big data sets (Gürsoy, 2010). It can both use classical statistical methods and machine learning practices. Various ML algorithms such as Multicriteria Decision Aiding (MCDA) (Es, 2018), clustering analysis (Anuşlu & Fırat, 2019; Es et-al., 2018) and grey cluster analyses (Karakoç et al., 2019) as an economic indicator that classifies different groups have been continuously developed to aid decision making in many fields in the age of Industry 4.0 (Anuşlu & Fırat, 2020).

One of the most important disciplines that data mining practices are to categorize the customers who are likely to be lost and predict it before it happens (Dahiya & Bhatia, 2015; Kayaalp, 2017). Now, thanks to the massive data collection and post-processing tools, it is possible to determine customer profiles correctly, analyze the customer churn, and according to these processed results, companies are enabled to prevent their customer loss before it may occur (Ahmad et al., 2019; Dahiya & Bhatia, 2015; Gülpınar, 2015). Accordingly, with the coordinated operations of marketing practices, customer experience management, and action plans of information technologies units, both current customers can be retained, and loss of customers can be prevented.

A growing body of studies has been recently carried out on customer loss in the telecommunication market (Li et al., 2016; Dalvi et al., 2016; Yeboah-Asiamah et al., 2018). It is significant to clarify the root cause of churned customers that may be due to one or more such reasons as dissatisfaction (Baier et-al., 2020), the higher cost compared to the competitor's (Yulianti & Saifudin, 2020), low quality (H. Celik & Güler, 2019), lack of desirable features (Deligiannis & Argyriou, 2020), privacy concerns (Bandara et-al, 2020), brand power (Al-Mashraie et-al., 2020) and so on. Thorough data analysis is required to develop an analytical method or to utilize one of the currently developed methods to maintain a long-lasting relationship with the existing customers. The more number of loyal customers a company maintains by not considering short term profits, the more advocates and the better reputation they have in the marketplace, which proves that Customer Churn Prediction (CCP) to be a substantial problem especially in the telecommunication industry (TCI) domain (Ganesh et-al., 2000). Although the retention of an existing customer is more efficient in comparison to acquiring a new one, it is also very challenging since TCI also suffers due to extremely intense contention, the concentrated marketplace, adaptive conditions, and initiating new appealing offers along with recent advanced development of technology (Amin et al., 2019).

Abundant data from TCI customers in terms of connection/disconnection, demographics, financial details, and other user attributes offer a great opportunity for the researchers in the fields of Machine Learning (ML) and Data Mining (DM) (Pelin Biçen, 2002) to foster CCP field with new predictive models in order to eventually provide aid tools to practitioners in TCI. Many ML approaches for data processing are available in the literature, such as Naïve Bayesian (Çiçek & Arslan, 2020), Artificial Neural Network

(ANN) (Dahiya & Bhatia, 2015; Gülpınar, 2015; Mitchell, 1997), Neuro-Fuzzy Classifier (Abbasimehr, 2011), Support Vector Machines (Oh & Sohn, 2009), Decision Tree (Ahmad et al., 2019; Dahiya & Bhatia, 2015; Hassouna et al., 2015), Genetic Algorithms (Goldberg & Holland, 1988), and Ensemble models (Coussement et-al., 2017; Wang et-al., 2019) to forecast whether a customer is a churner in anticipated future based on his/her activities. Bayesian networks are currently one of the primary tools in Customer Analytics (Sauro, 2015) for representing probabilistic knowledge from measuring customer attribution. General probability distributions could be concisely represented by Bayes' algorithm over a set of propositional stochastic quantities (Russell, 2013). Besides, the probabilities of hypotheses could be calculated based on Bayes' theorem. The Naive Bayes classifier algorithm that calculates explicit probabilities for hypotheses for estimating nominal values of unobserved variables is among the most practical approaches to the ML problems, such as CCP problems (Çiçek & Arslan, 2020). Decision trees, on the other hand, might be thought of as an unstable learning scheme because even minor differences in the training data can lead the model to deviate significantly. In other words, in a decision tree, slightly changing the training data helps obtain an entirely various decision tree structure. However, with Naïve Bayes, small changes in the training set do not deviate from the result of Naïve Bayes or instance-based learning schemes. Therefore, Naïve Bayes can be considered as a "stable" ML method. It can produce probabilities, and it can also function with respect to probabilities.

The sum of squared errors can be minimized by means of choosing Bayesian analysis that can be used to justify a key design choice in ANN learning algorithms (Es, 2013) as scanning the space of possible NN (Mitchell, 1997). As learning target functions by which probabilities can be predicted, the cross-entropy that is an alternative error function can be derived, which offers more proper results as compared to the sum of squared errors as the learning algorithm runs. The inductive bias of decision tree learning algorithms can be analyzed by the Bayesian perspective. It also favours short decision trees and examines the closely related Minimum Description Length rule (Aoga et al., 2018). A basic familiarity with Bayesian methods is of paramount importance for comprehending the way how many learning algorithms operate. Effective algorithms can be facilitated for training data in learning Bayesian belief networks (Yu et al., 2018). Many current researchers consider several different settings for this learning problem. First, the network structure might be inferred from the training data; second, in each training example, all the network variables might be directly observable. In the case where the variables of the network structure are fully observable in the training examples, it is somewhat straightforward to learn the conditional probability tables. The conditional probability table entries could be simply estimated just as a Naive Bayes classifier can do. In the case of some variable values of the network, the structure is observable in the training data, and the learning problem becomes more challenging. Learning the weights takes place in this case for the hidden unit, which resembles an ANN approach, where the input and output node values are given but the training examples leave the hidden unit values unspecified (Mitchell, 1997). Since the back-propagation algorithm has accomplished many practical solutions, it is among the most commonly used ANN learning technique (Es, 2013) and proper to resolve general problems such as those with;

- the instances represented by many pairs of attribute-values,
- the aim function output valued discretely, the vector of real-valued or several real or discrete-valued features that are robust to noisy data,
- the training examples possibly containing errors,
- acceptable long training times,
- fast evaluation of target function being learned,
- the training data possibly containing some missing attribute values

Due to its easy way of implementation and comprehension compared to other classification algorithms that exist in the literature, decision trees are the most commonly used algorithm (Mittal et-al., 2019) which can be implemented in a serial or parallel fashion based on the volume of data where memory space is available on the computer resource and scalability of the algorithm (Anyanwu & Shiva, 2011). Comparable accuracy can be often produced by ANN and decision tree learning, as Craven & Shavlik (1997) discussed in their experimental study; however, a practical method for concept learning and for learning other discrete-valued functions could be provided by decision tree learning. The decision tree algorithm is one of the most significant approaches that have been used as a classifier and forecasting tool in data mining studies. Decision tree methods existing in the literature are mainly ID3, C4.5, ASSISTANT, PUBLIC, CART, CN2, SLIQ, SPRINT and so on (Anyanwu & Shiva, 2011; Mitchell, 1997; Salzberg, 1994). ID3 is the most typical decision tree learning scheme which was originated in the concept learning system (CLS). It has addressed many accuracy issues improving its efficiency by avoiding issues such as over-fitting, mistraining due to missing values and so on; hence it was finally evoked as C4.5 (C5.0) so that it could deal with continuous attributes (Dai et-al., 2016). When the tree does not classify correctly all the given training parts of the dataset, it selects another training dataset adding to the original portion, and then it repeats this procedure until the decision set correctly avails. ID3 algorithm refines the search further to the tree below a leave node as long as the new erroneous example varies in an uncertain way from the other examples associated with the node, ID3 attempts until it eventually succeeds in finding a new decision attribute. Another approach in decision tree algorithms is the reduced error pruning for the purpose of avoiding over-fitting issue (Pham et al., 2019). In this approach, three nodes are iteratively pruned. The pruning approach always selects to remove the node to most increase the accuracy over the validation set and it continues to prune such nodes until their removal results in the inaccuracy of the tree over the validation set. The over-fitting issue can be also prevented by another successful method called Rule-post pruning that could find high-accuracy hypotheses. The approach first infers the decision tree from the training set then it grows the tree until the training subset agrees with minimal or no permutation allowing the over-fitting. Then the learned tree is adopted into an equivalent set of rules, from the root to a leaf node, it generates one rule for each path. To eventually improve the accuracy, it follows next by pruning (generalizing) each rule by simply removing any preconditions. Finally, it classifies the pruned rules by their accuracy and considers them in the same order when sorting later examples (Mitchell, 1997).

The number of open-source software has increased gradually with the development of the independent software movement since the software that has emerged could produce extremely safe results. Some of the open-source software is called SPSS Clementine (Altay, 2005), KNIME, Tanagra, RapidMiner and WEKA, and they can contribute to the data mining solution of companies of all sizes operating in different sectors as well as in academic fields (Çınar & Silahtaroğlu, 2013). In order to evaluate, Dahiya & Bhatia (2015) compared two data-mining algorithms –for churn prediction analysis – such as Decision Tree (J48 or, i.e. C4.8) and Logistic Regression which is built-in WEKA machine learning software. They found out that the J48 Decision Tree classifier algorithm has given slightly better-précised prediction in comparison to the Logistic classifier. J48 pruned decision tree algorithm generally offers fairly good accuracy compared to its counterpart algorithms. For some datasets, J48 may outperform even Ensemble (committee) techniques. WEKA (Waikato Environment for Knowledge Analysis) data mining software is rich in terms of the decision tree, Bayes, and ensemble learning algorithms nevertheless;, it does not involve any ANN algorithm whereas RapidMiner (Celik & Basarir, 2017; Geetha & Nasira, 2014) does. Both WEKA and RapidMiner have support vector machines (SVM) and ensembles as classifiers for data-mining the big data. They also have a user-friendly graphical user interface (GUI) for eventually conducting graphical analysis in the post-processing stage.

Decision tree algorithms have top-down recursive induction by selecting attributes for the root node, and it generates branches for each possible attribute value (Grąbczewski, 2014). Then it breaks up instances of a dataset into subsets, one for each branch extending from the node. It recursively repeats this for each branch in the decision tree using only instances that reach the branch. Once all instances have the same class, the algorithm stops. The approach to how the algorithm decides which one of the attributes end up with the purest nodes on pursuing the purity for eventually having the smallest tree is by utilizing the information theory (Shannon, 1948) based on entropy which is to be explained in more detail in the following section.

Decision tree algorithms such as J48 and Decision Stump use the measure of entropy for classification (Ullah et al., 2019). A boosting ensemble algorithm in WEKA called AdaBoostM1 (adaptive boosting), by default, uses Decision Stump as its base learner. With boosting, the accuracy generally improves as the number of iterations increases and then flattens out. The idea with bagging, several different decision structures are desired to be produced, such as using J48 to produce decision trees, and then with several different training sets of the same size, sampling the original training set can generate slightly different decision trees. The set could be sampled through replacement, which means that sometimes two of the same [instances] might be sampled in bagging. Several different training sets could be produced, and then a model could be built for each one – such as a decision tree – employing the same or some other ML schemes. Then voting combines the predictions made by different models, or in regression situations, the numeric result could be averaged rather than voting on it.

A combination of such algorithms as ANN, DT, Bayes, etc., has also been developed called ensemble models that combine multiple algorithms including decision tree regression and classification methods as well as any other models mentioned previously to gain better predictive performance. Ensemble learning methods (ELM) have received remarkable attention (Ma et al., 2015), and the performance of the regression and classification problems have had considerable advancement by using ELM techniques in recent years. Techniques based on regression are mostly associated with good results in CCP. The methods in ensemble learning can be classified into four groups: bagging, randomization, boosting, and stacking; each works with different approaches. Running different accompanying ML algorithms, allowing them to vote on classifiers for the same problem so that an unknown test instance is classified and can often improve the predictive performance. In general, ELM forecasting models such as bagged (bootstrap aggregated) regression trees (BRT) and gradient boosted regression/decision trees (GBDT) as well as bagged artificial neural network (BANN) and gradient boosted artificial neural network (GBANN)) are some of the most widely used ELM strategies for reducing ML algorithm prediction error. (Karakurt et-al., 2013; Zhang et-al., 2008).

In this thesis, customer churn analysis has been carried out in the telecommunication sector, and machine learning methods have been discussed to detect possible customers who might be in the leaving process since keeping the current customers takes less time and expenditure according to obtain new customers. Thus, the company can avoid more significant problems in regard to the company's stability. With the data mining methods, appropriate analysis of the current data has been performed, and the customers were decomposed and specified. Consequently, the customers were classified correctly, and appropriate diverse plans could be handled for each customer type according to the situation of the customer behaviour. Furthermore, this thesis will offer assistance, restraints, and implementations and potential further research fields to support in different sectors. For the purpose of evaluating the value of data mining to prevent customer churn and investigate potential customers who are likely to leave the company, CCP analysis has been carried out by using authentic, but anonymous Big Data collected from customers of a private company in the telecommunication sector, and the methodology of data mining is discussed to detect possible customers who are in the leaving process and to prevent customer churn. Keeping the current customers not only takes less time and expenditure in comparison to gaining new customers, but this also leads the companies to avoid more significant problems with regards to their stability. Through the data mining methods, an appropriate analysis of the current data can be performed, and the customers can be properly decomposed and specified. Consequently, the results from the customer churn analysis through data mining algorithms such as decision tree can indicate when the customers are classified correctly, and appropriate diverse plans are handled for each customer type according to the customer behaviour against the situation that they face with, the attributes of the incidences can be predicted and changed to good for the company. In the field of Customer Relation Management (CRM), to propose a convenient CCP methodology in order for retaining the customers who tend to churn, a set of data-mining analyses has been conducted to predict customer churn from big data in terms of services offered to the customers of a telecommunication company to find out the most suitable ML scheme built in an open-source data mining software, WEKA. For this purpose, six data mining algorithms have been evaluated for CCP analysis: Logistic Regression, Naive Bayes, J48, and ELM schemes such as RandomForest, Bagging and Boosting. RandomForest uses RandomTree, whereas Bagging uses J48 as a base learner. The experimental datasets acquired from the historical database of the TCI Company to validate the effectiveness of some ML classifiers such as decision tree and ensemble algorithms for determination of likelihood of customers who tends to leave based on such data mining analyses. The ML schemes have been run first on testing the dataset with two options to assess their stability. Then, ZeroR scheme is used as a baseline to benchmark the accuracy performance of the other schemes. Thus first, ZeroR baseline data-mining scheme is run on each dataset to have the least classification accuracy that the other data mining schemes are required to provide. Then Logistic Regression, Naive Bayes, and J48 decision tree schemes are run and followed by the ELM schemes such as RandomForest, Bagging, and AdaBoostM1.

Although this thesis aims to conduct a set of analyses for the telecommunication sector, the prediction methods mentioned in this study, can be used in any sector for preventing customer churn. The practice will be executed in the biggest and leading telecommunication company, Turkcell. A new method was developed to assess customer loss management. For the benefits from this study, Turkcell holds all the right to use the outcomes of our research to refine their methods. Furthermore, this study offers assistance for possible future implementations and potential further research applications to support the researchers in different fields and practitioners in different sectors.

2 STATE OF THE ART

In this section, the classifier algorithms that have been used herein in the prediction of customer churn are discussed in more detail, including their mathematical bases, their origins and their features in the built-in schemes of WEKA, the data mining open-source software package (Figure).

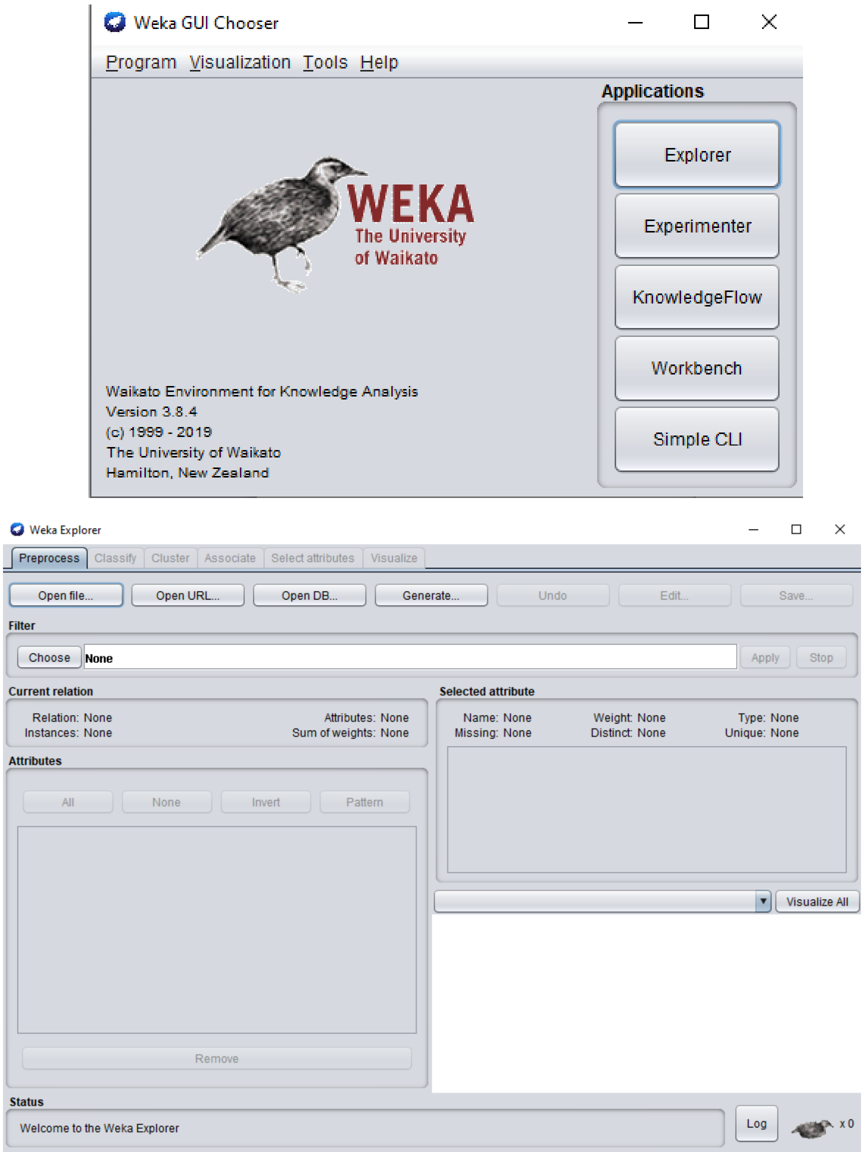

Figure 2.1 WEKA, the data mining open-source software package

2.1 Decision Tree

Claude Shannon (1916-2001), an American mathematician and scientist, came up with "Information theory", which measures information in bits by quantifying the entropy

$$\left(p_1, p_2, ..., p_n\right) = -p_1 \log p_1 - p_2 \log p_2 ... - p_n \log p_n \tag{0.1}$$

Eq. (0.1) is the entropy formula for each of the possible outcomes summing $p \log p$. All of those minus signs are used due to the negative logarithms since the numbers are less than 1, which results in the probabilities being less than 1. This eventually results in the entropy being a positive number.

Information gain:

- Knowing the value of the attribute gives the amount of information gained
- (Distribution entropy before the split) – (distribution entropy after the split)

Decision tree algorithms use this theory to choose the best attribute by checking the information gain: knowing the value of an attribute, the information amount in bits is taken into account by subtracting the distribution entropy after the split from that before the split (Witten, 2020).

J48 is a decision tree algorithm that carries out top-down induction based on information theory creating a strategy of recursive divide-and-conquer by selecting and splitting an attribute at its root node and the best attribute for this is chosen by information gain, and then this leads to the generation of a branch corresponding to each possible attribute value. This procedure eventually divides the instances into subsets, one corresponding to each branch extending from the root node. Recursively this procedure repeats with respect to each branch, making the selection of an attribute at each node, which uses only instances reaching the branch. This recursive event carries on until all instances have the same class, then it stops.

Once J48 runs, it also provides probabilities in the result panel on WEKA – the negative and positive probabilities, respectively. Where the errors can also be seen. These are all different probabilities that are internally used by J48 in order to perform the pruning operations, which was the same approach as discussed in ID3's pruning,

2.2 Logistic Regression

Respectively, a linear function and a threshold are calculated by the Linear regression method, which has a linear sum. Nevertheless, the Logistic regression performs direct estimation of class probabilities by carries out the same linear sum as shown in Eq. (0.2) – the type of linear sum same as that in the linear regression. This formula is called “logit transform”, which is multi-dimensional with the arbitrary number of different a’s. If only one dimension, one variable a_1, then if this is the input to the logit transform, the output as seen in Figure 2. appears to be an S-shaped curve between 0 and 1 generates by a softer function including an exponential term that leads to never drop below 0, never exceed 1 and leads a smooth transition in between (Witten, 2020).

$$\Pr[1|a_1, a_2, ..., a_k] = 1/\left(1+\exp\left(-w_0 - w_1 a_1 - ... - w_k a_k\right)\right) \qquad (0.2)$$

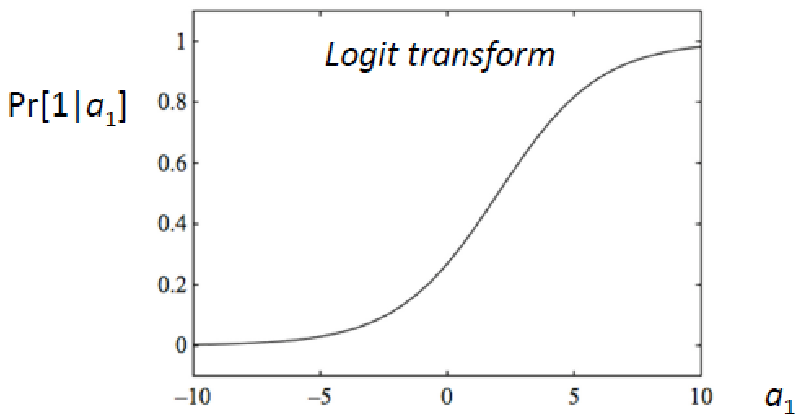

Figure 2.2 Logistic regression graphical demonstration (Witten, 2020)

The squared error is not minimized as done in linear regression. Instead, the log-likelihood (not minimize the squared error) is maximized with a logit transform by choosing weights. Choosing weights for maximizing a probabilistic function called the “log-likelihood function” (Eq. (0.3)) is better off:

$$\sum_{i=1}^{n}\left(1-x^{(i)}\right)\log\left(1-\Pr[1|a_1^{(1)}, a_2^{(2)}, ..., a_k^{(k)}]\right) + x^{(i)}\log\left(1-\Pr[1|a_1^{(1)}, a_2^{(2)}, ..., a_k^{(k)}]\right) \qquad (0.3)$$

For the prediction of probabilities directly, Logistic regression performing the logit transform is a popular and powerful ML method that works internally with probabilities as Naïve Bayes does.

2.3 Ensemble Learning

Over the last decade, ensemble machines have gained in popularity. Principal determinants are the first members of a community machine, and the basic learning algorithm commonly used in creating community machines are artificial neural networks (ANNs) and decision trees (DT). Bagging and boosting are two prominent community approaches that are based on the same philosophy and are intended to solve problems with poor predictors. Bagging (short for bootstrap aggregating) is one of the oldest methods presented by Breiman (1996) to improve learning machines' prediction accuracy (Karakurt, Erdal, Namlı, et al., 2013).

Mainly in the categories of Boosting (AdaBoost), Bagging, Randomizing (RandomForest, RandomTree), Stacking, many ELM algorithms exist in the open-source data-mining software.

In bagging: The training set is sampled to have another set at the same size by choosing the Bag size of 100%, but in each time sampled "with replacement", which means that different sets of the same size, but each set may include the duplicate of the original training [instances]. This bagging event happens to build a model concerning each one – using the same ML scheme, then predictions can be combined by voting on classifiers for which one to bag. In the bagging approach, the settings of the random-number seed and the number of bagging iterations can also be set.

In the randomizing approach: The training data is not randomized., in the "RandomForests" algorithm the algorithm is randomized instead [30]. Depending on what the algorithm is, the way how the algorithm is randomized. Random forests – using decision tree algorithms that select the greatest quality for dividing each time, for example, J48. This randomizing step selects a few of the best options rather than selecting the very best and then randomly picking among the classifier algorithms that give different trees every time. Generally, to gain better performance, the decision trees are bagged, and then randomized and eventually the result is bagged. In RandomForest, the trees' maximum depth created – by default, 0 is chosen which means unlimited depth. By choosing the number of features that could be 4; it could be chosen from the top four characteristics – each time a decision is made about which decision to put in the tree, the top four candidates are considered.

Boosting is iterative: The performance of previously established models influences new models by essentially establishing a model, and then taking into account the instances that are misclassified by that model. These are the difficult cases to categorize; more weight is given to those that go wrong in order to provide a training set for the next model in the iteration. The new model is urged to become an "expert" on occurrences that have been misclassified by all of the previous models. They are eventually combined by voting along with weighting models concerning their performance. WEKA has a great boosting scheme implementation named AdaBoostM1 that consistently provides decent results.

In the stacking ensemble learning method: Base learners are not combined by voting herein but through a meta-learner that is another learner scheme uniting the output of the base learners. The base learners are called level-0 models, whereas the meta-learner is a level-1 model. The classifying predictions performed by the base learners become the input to the meta-learner. ML schemes are typically employed as base learners to obtain diverse expertise in numerous fields. Special care is required to be paid in the way data is generated to train the level-1 model, which may involve somewhat cross-validation.

3 METHODOLOGY

In this research, customer churn analysis is carried out by using authentic, but anonymous Big Data collected from customers of a private company in the telecommunication sector, and the methodology of data mining is discussed to detect possible churning customers and to retain customers churn. The methodology that is used to carry out of research is divided into four main stages that are;

1. Literature survey in order to identify the concept of Churn Analysis

2. Definition of key characteristics of Customers and by using these characteristics building a decision tree that reveals the relationship between customer churn and cases.

3. Development of an appropriate optimization model using the findings.

4. Testing alternative scenarios in the developed model (via simulation or heuristic approaches)

5. Analysis of results

For this purpose, three sets of data are analyzed, dividing the main dataset into two: the residential members and corporate members. The third data set is the combination of both groups. In all datasets, we aim at predicting the same nominal classes, that is, churn status, whether active or passive, along with the other seven attributes the same in all. However, in the combined dataset, one more nominal attribute is added with respect to the customer type: residential and corporate, as shown in Table 3.1.

Table 3.1 Number of attributes and incidences of three datasets

	residential	*corporate*	*Combination*
Number of Attributes	7	7	8
Number of incidences	195712	32905	228617

3.1 Data mining software, WEKA

On the test options panel of WEKA, there are four test schemes such as 1. Use training set, 2. Supplied Test Set, 3. Cross-Validation, and 4. Percentage Split. The first test option uses the training dataset for testing it, which sometimes gives the best prediction precision percentage; however, it is not much recommended by the experts. For the second option, a dataset for testing needs to be uploaded into the software since we have not prepared or been supplied with a test set. In running most of the classifiers, we have not chosen this testing option. Instead, the other three options are generally chosen for all the datasets and the best results obtained are presented in the following section. In cross-validation, the default value is 10 folds which means the dataset is divided into 10 pieces, and 1 piece is used for testing, and the rest is used for training, and the model runs, and this repeats 10 times each time the test piece shifts to the next one. This causes us to take a longer time to obtain the classification results in comparison to others; however, this testing option is more robust and generally gives the best prediction percentage. In percentage split, the default value is 66%, which means two-third of the dataset is used for training, and one-third of it is for testing as the model runs only once. In this study, we have selected a 90% percentage split in the option for our data mining procedure. For the stability analysis of the schemes, we have not experienced any overfitting issue causing instability in any of the data mining schemes because we have not seen much deviation among the results that are obtained from different test options.

3.2 Python programming language and machine learning libraries

We have used Python programming language while implementing a decision tree classifer for presenting the decision path of a machine learning model. Python is a general purpose language which had been designed by a Dutch programmer Guido van Rossum in 1991 and has increased its popularity in 2010s. It has several features that a machine learning practitioner would want to use in the first place. First of all, it is an easy-to-learn and easy-to-write language with its clear and understandable syntax. Second, and most importantly, it has an incredible ecosystem for machine learning and scientific use cases.

One of the most prominent machine learning Python libraries is Scikit-learn which we have used in our project. Scikit-learn includes a broad collection of state-of- the-art machine learning algorithms for supervised and unsupervised problems (Pedrogose et.al., 2011). Another scientific library we have used is Pandas. Pandas, an Excel-like library for Python, is used for manipulating and analyzing the data (McKinney, 2011). We have used Pandas library while reading the data and presenting it to the model.

We have used DecisionTreeClassifier class in Scikit-learn. In a decision tree classifier, the model derives simple decision rules from data features while predicting the value of a target variable. Several parameters are utilized while creating a model instance; criterion, to choose the method to measure the split quality; max_depth, to specify the maximum depth of the tree; max_features, to determine the number of features to consider while looking for the best split. Our choices are Gini for criterion, 4 for max_depth, and 1 for max_features parameter. After the model instance is created, the fit method, which is a function of this class that the model learns from the dataset, is used. This method expects two arrays; X holding the training samples and Y holding the class labels. After training our tree model with the dataset by using the fit method, we have used the export_graphviz function to export our tree model in dot format and Pydotplus and Graphviz libraries in order to visualize it.

3.3 Datasets

Table 3.2 presents an example of the data anonymously recorded by the telecommunication company based on customer membership length, number of failures, and their response length to repair in their internet connection service. This dataset is required to be determined to comprehend the probability of customer churn based on failure in the service. Figure 3.1 Customer number distribution of the telecommunication company over years of the customer lifetime length illustrates the customer number distribution over the length of customer lifetime in terms of years. Figure 3.2 Churned customer number against the length of customers' lifetime and Figure 3.3 also respectively represent the graphical illustrations of the amount of the churned customers and percentages of the churned and non-churned customers against the lifetime length of the customers.

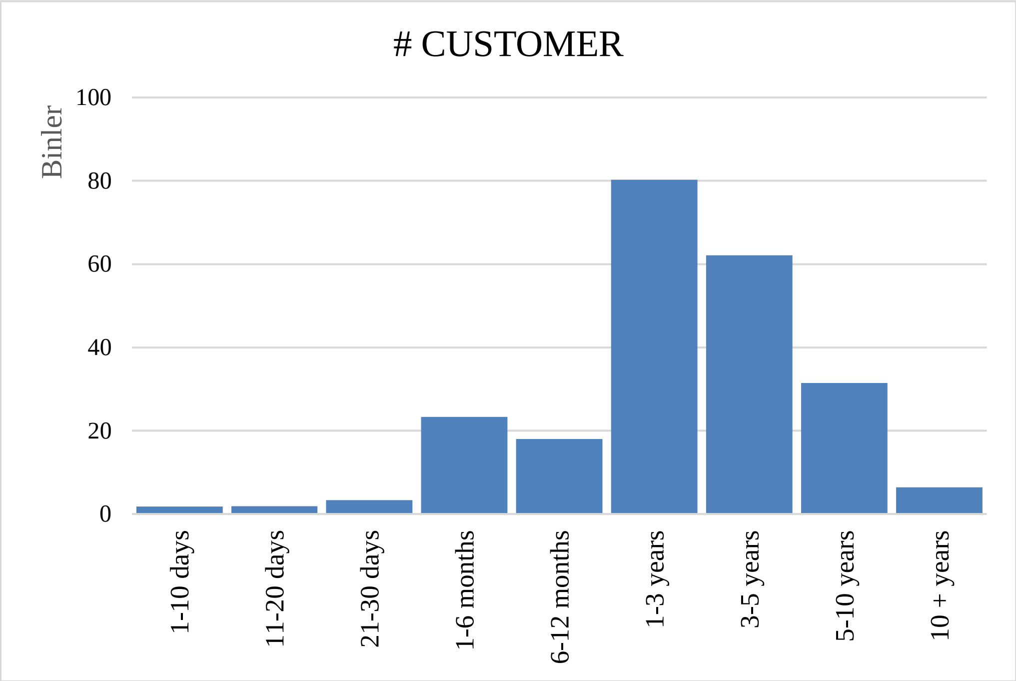

Figure 3.1 Customer number distribution of the telecommunication company over years of the customer lifetime length

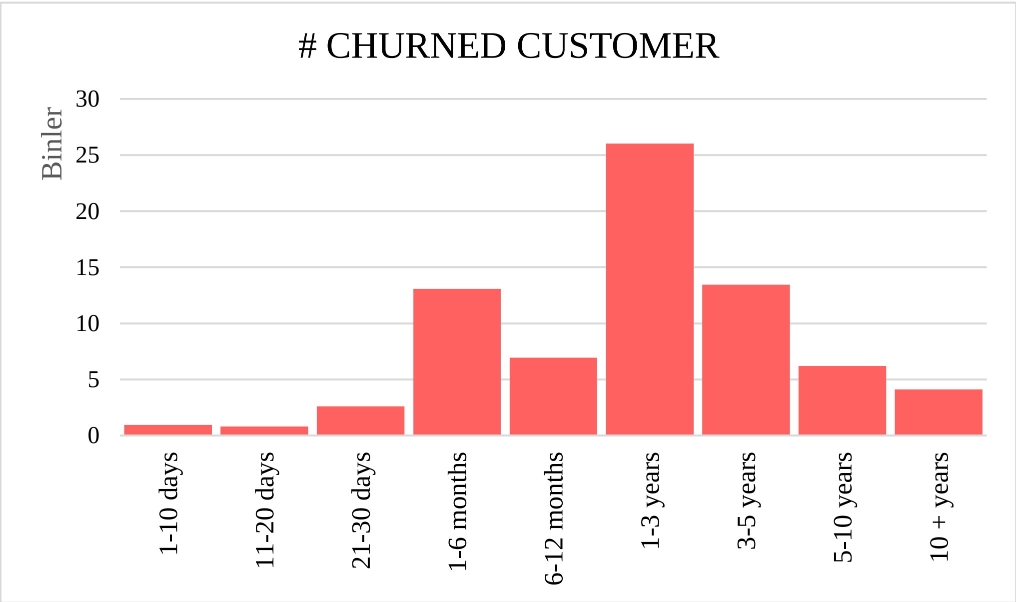

Figure 3.2 Churned customer number against the length of customers' lifetime

In Figure 3.2 Churned customer number against the length of customers' lifetime, 1-3 year length of the customer lifetime seems very critical in order to determine the retain or loss of the customers. After three years, the longer the customer is accomplished to remain in the company, the higher possibility of gaining his/her loyalty to the telecom company since fewer customers tend to churn as the time goes on over the years.

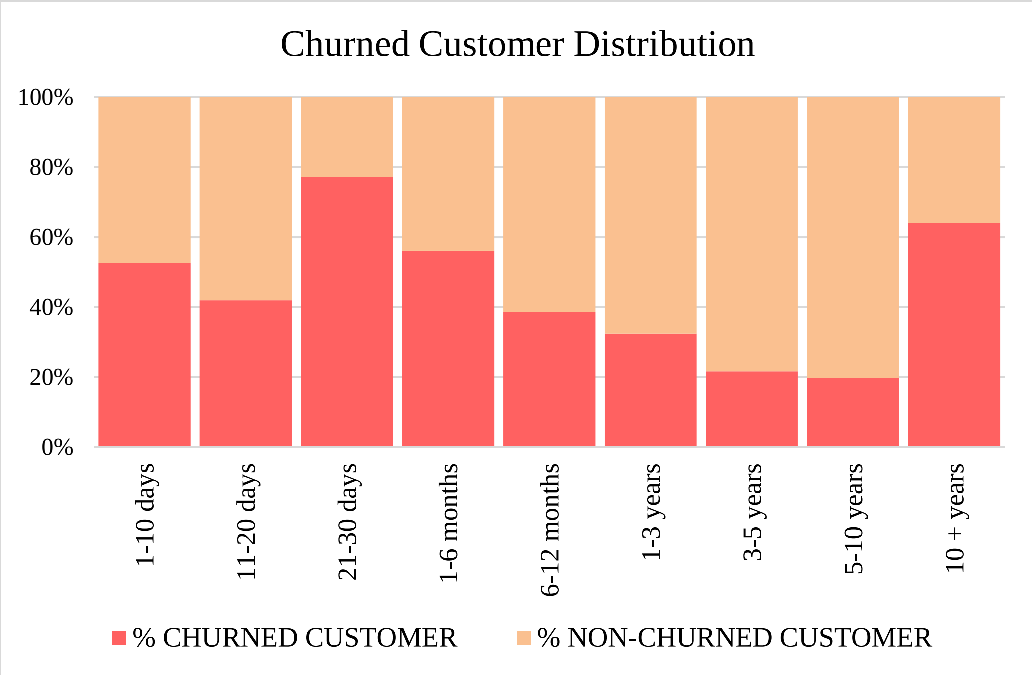

Figure 3.3 Percentage of the churned customers vs that of non-churned customers against the length of customers' lifetime

On the other side, as can be seen in Figure 3.3, churned customer distribution over the years has different tendencies versus non-churned customer distribution. Between the third week and the first month, the percentage of the churned customer is the highest, i.e. that of the non-churned customer is the lowest. However, interestingly the percentage of churned customers is the second-highest in the period of 10+ years, even though before then, the churn tendency becomes lower and lower as time passes.

Figure 3. and Figure 3.5 illustrate the churned customer distributions over different periods of customer lifetime for both corporate and residential customers. Both residential and corporate types of customers have the highest number of churn in the period of 1-3 year period, and the churned customer numbers decrease as the period

increases.

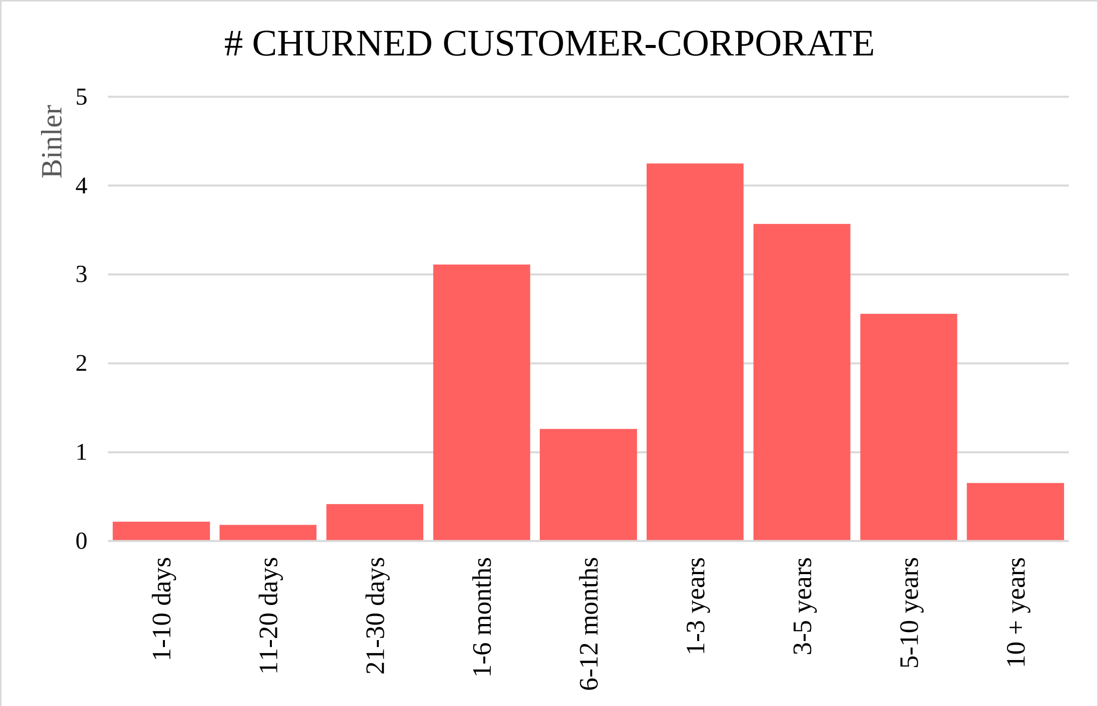

Figure 3.4 Churned customer distribution over different periods of customer lifetime for coroporate customers

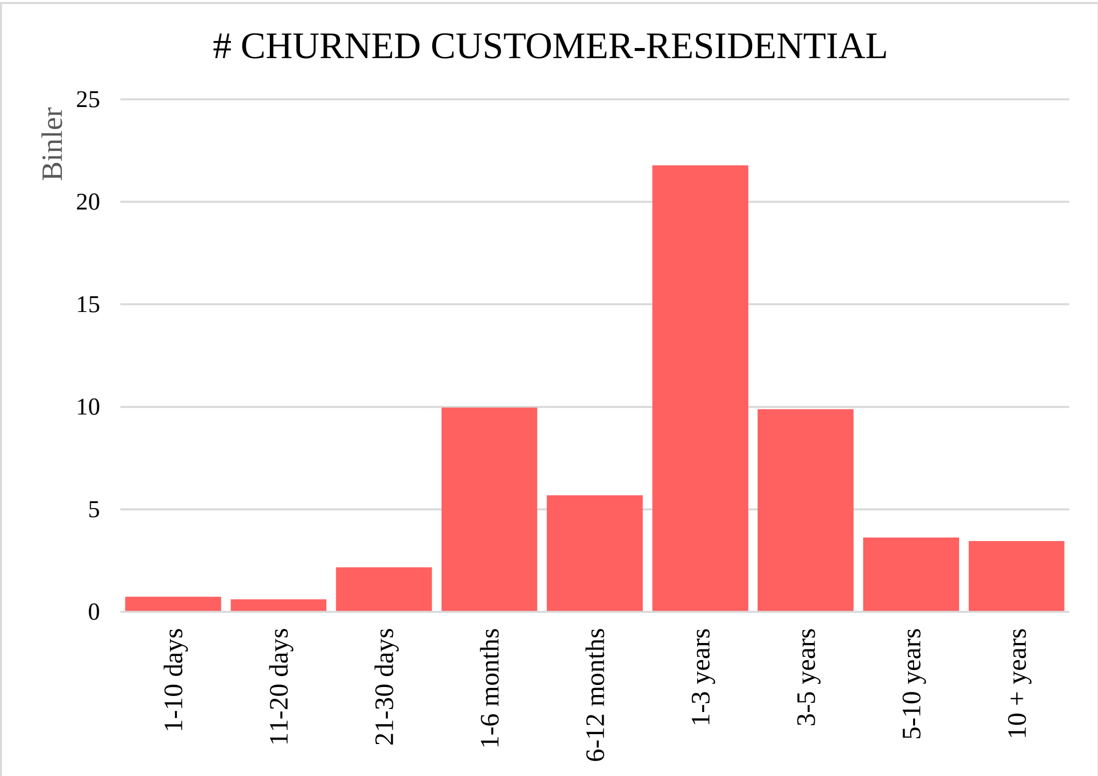

Figure 3.5 Churned customer distribution over different periods of customer lifetime for residential customers

Figure 3. shows the segmentation of customers in terms of total numbers. According to these total customers over the past ten years, the company has gained 87% residential customers vs 13% corporate companies. Figure 3. also shows the total percentage of churned and non-churned customers, and churned customer percentage seems relatively high, with 32% in total over the years.

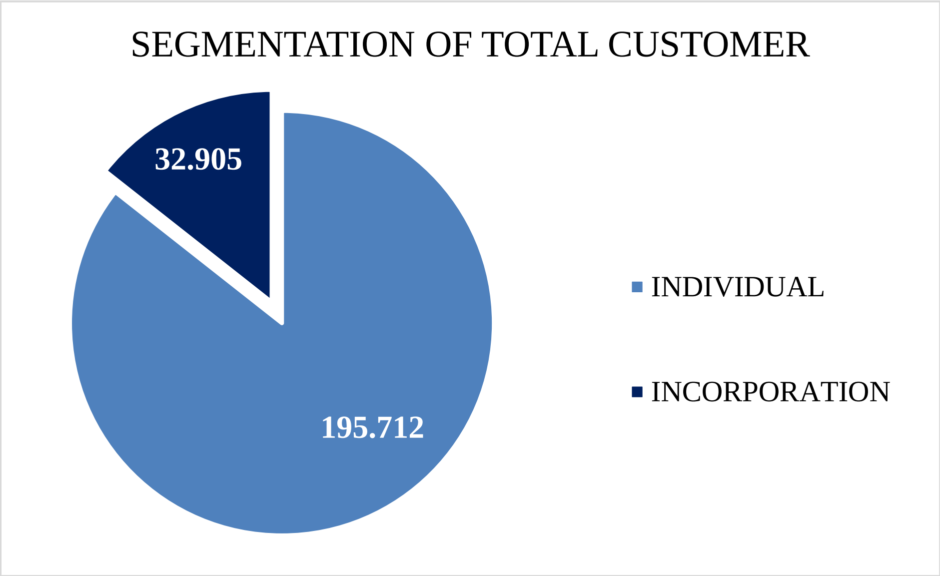

Figure 3.6 Total numbers of customers in segment of residential and corporate

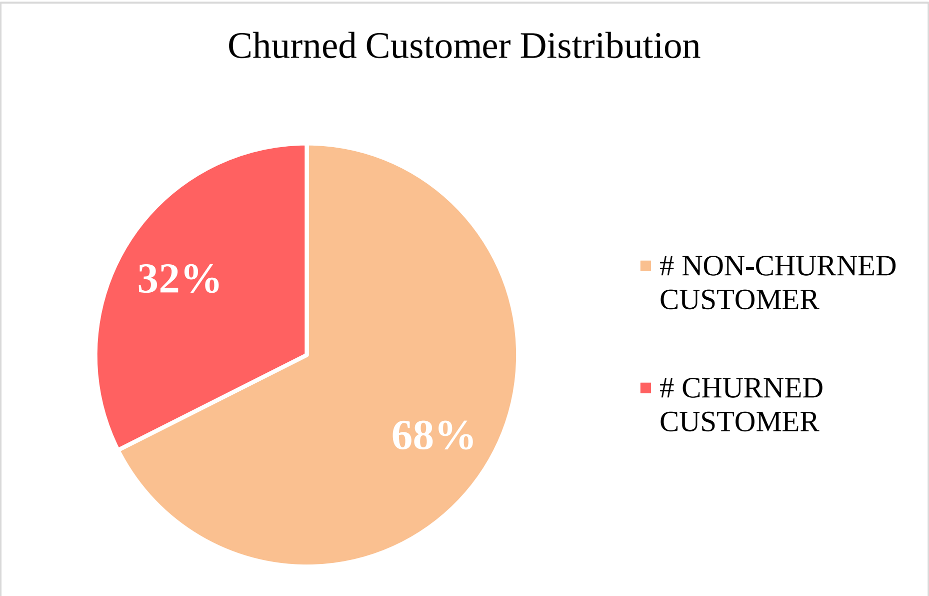

Figure 3.7 Percentage of total churned customer vs total non-churned customer for both residential and corporate segmentation

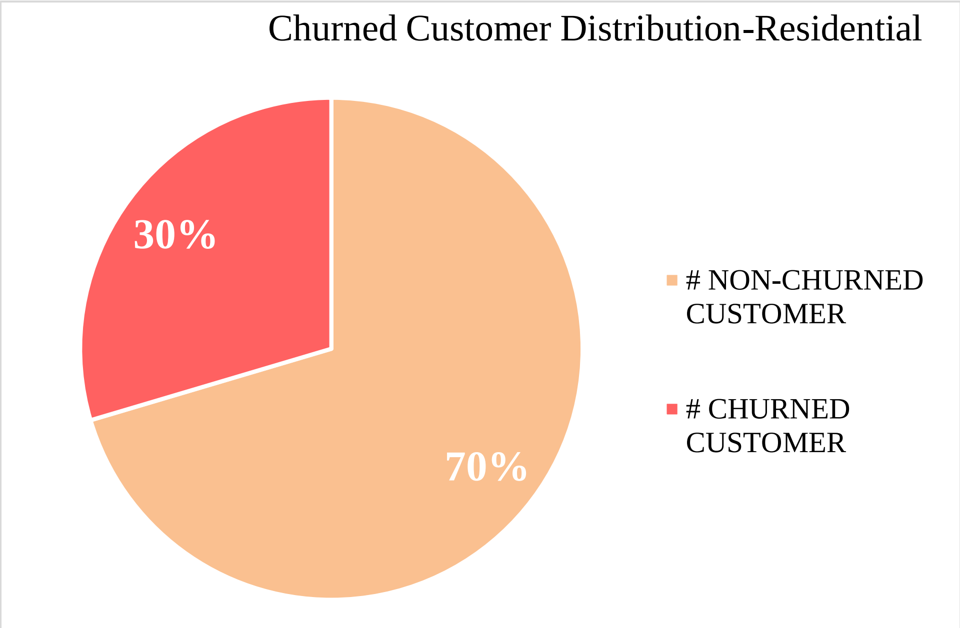

Figure 3.8 Percentage of churned residential customers vs that of non-churned residential customers

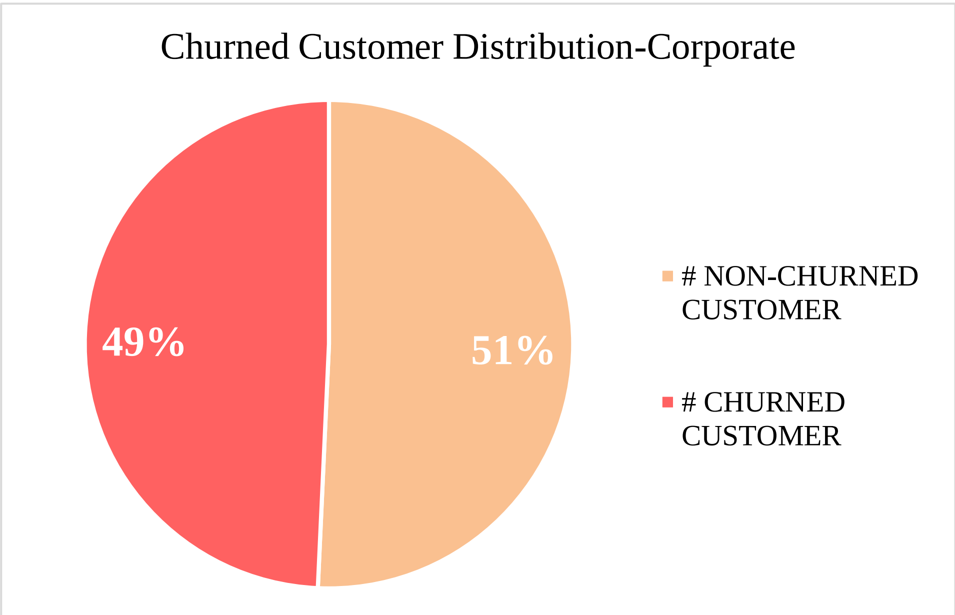

Figure 3.9 Percentage of churned corporate customers vs that of non-churned corporate customers

Figure 3., Figure 3. and Figure 3. show the percentage values of churned and non-churned customers for total, residential and corporate types of customer segmentation, respectively. This graphical analysis of the dataset, including all the lifetime periods, indicates that the corporate customers tend more to churn in comparison to the residential customers.

Table 3.2 Sample of a dataset used for customer churn prediction analysis

Total failure no	Unrepaired	Membership day	Unit time failure no	Ave repair hrs	Customer Life time	Customer type	Churn Status
0	0	1	0	(1,000)	1-10 days	residential	PASSIVE
0	0	1	0	(1,000)	1-10 days	residential	PASSIVE
0	0	1	0	(1,000)	1-10 days	corporate	PASSIVE
0	0	1	0	(1,000)	1-10 days	corporate	PASSIVE

Table 3.3. The amount and percentage distribution of all and churned customers over the periods of customer lifetime

CUSTOMER LIFETIME	NUMBER OF CHURNED CUSTOMER	NUMBER OF CUSTOMER	PERCENTAGE OF CHURNED CUSTOMER	PERCENTAGE OF NON-CHURNED CUSTOMER
1-10 days	952	1.808	**53%**	**47%**
11-20 days	790	1.884	**42%**	**58%**
21-30 days	2.592	3.359	**77%**	**23%**
1-6 months	13.081	23.305	**56%**	**44%**
6-12 months	6.943	18.005	**39%**	**61%**
1-3 years	26.023	80.247	**32%**	**68%**
3-5 years	13.453	62.124	**22%**	**78%**
5-10 years	6.187	31.459	**20%**	**80%**
10 + years	4.112	6.426	**64%**	**36%**
TOTAL	74.133	228.617		

The numbers of active and passive residential and corporate customers against the customer lifetime are presented in Table 3.4 and

Table 3.5, respectively. Even though customer lifetime of 3-5 years and 5-10 years have a good enough percentage of the grand totals, 10+ years-long customer lifetime is in the band of 2.5-3 % of grand totals of both residential and corporate customers. The loyalty of the customers needs to be increased by increasing their satisfaction with the service by means of avoiding or at least minimizing the matters of disturbance

Table 3.4 Numbers of an active and passive residential customer with respect to the membership time interval

Customer Life-Time	***Active***	***Passive***	***Grand Total***
10 + years	1,913	3,458	5,371
1-10 days	748	735	1,483
11-20 days	922	608	1,530
1-3 years	48,385	21,774	70,159
1-6 months	8,689	9,968	18,657
21-30 days	666	2,177	2,843
3-5 years	45,413	9,886	55,299
5-10 years	21,561	3,631	25,192
6-12 months	9,495	5,683	15,178
Grand Total	**137,792**	**57,920**	**195,712**

Table 3.5 Numbers of active and passive corporate customers with respect to the membership time interval

Customer Life-Time	***ACTIVE***	***PASSIVE***	***Grand Total***
10 + years	401	654	1,055
1-10 days	108	217	325
11-20 days	172	182	354
1-3 years	5,839	4,249	10,088
1-6 months	1,535	3,113	4,648
21-30 days	101	415	516
3-5 years	3,258	3,567	6,825
5-10 years	3,711	2,556	6,267
6-12 months	1,567	1,260	2,827
Grand Total	**16,692**	**16,213**	**32,905**

4 RESULTS AND DISCUSSION

In this chapter, we have discussed the data mining analysis of churn prediction and evaluated the decision tree classification model.

4.1 Data Mining Analysis

In this section, theoretical data mining analysis for churn prediction is conducted by using the experimentally obtained datasets from customers of a telecommunication company. ZeroR scheme is used as a baseline to benchmark the accuracy performance of the other schemes used in CCP since their accuracy must be higher as compared to the ZeroR. Thus first, ZeroR baseline datamining scheme is run on each dataset to have the least classification percentage to benchmark what accuracy the other data mining schemes are supposed to provide at least. Then Logistic Regression, Naive Bayes and J48 decision tree schemes are run and followed by the ELM schemes such as RandomForest, Bagging, and AdaBoostM1.

Since the Bagging scheme requires a large heap size in running Java in the background and when sometimes the memory size may not be enough, these schemes are not compatible to solve the models for large-sized datasets, e.g., in this particular case, the residential and combined dataset with around 200,000 instances. As a solution to this issue, the instances in the datasets of residentials and combined are shuffled by using the unsupervised incidence filter named 'Randomize' with a Random seed number of 42 (arbitrarily chosen by default) and then truncated from its 50% and 60% respectively by means of another unsupervised incidence filter called 'RemovePercentage' so that the datasets became prepared to be handled by the computationally expensive schemes such as Bagging.

As can be observed in the results presented in Table 4.1, J48 obviously outperforms Naïve Bayes based on all datasets, and it provides very similar results as the Logistic Regression classifier scheme in terms of both the accuracy and classifying the instances into the nominal attribute of Churn Status. The accuracy is determined by means of the confusion matrix. The classified instances indicate how many of the customers are classified to be in passive and active status. Indeed, Naïve Bayes seems to be worse than even ZeroR in terms of accuracy in two of the datasets.

One can conclude that J48 is a good choice for predicting the customer churn for these datasets; however, further investigation into its stability is required.

Table 4.1 Comparison of accuracy and number of classified instances from Naïve Bayes and J48 Decision Tree data mining algorithms based on three datasets

Datasets	*Result variable*	*ZeroR*		*Logistic Regression*		*Naïve-Bayes*		*J48*	
		Passive	Active	Passive	Active	Passive	Active	Passive	Active
Corp.	Accuracy %	51		61		61		67	
	Class. Instan.	0	16692	9045	11062	15933	3976	11466	10454
Res.	Accuracy %	70.4		73		65.4		75	
	Class. Instan.	0	68864	8002	63757	7156	14598	18228	128150
Combin.	Accuracy %	68		71		67		74	
	Class. Instan.	0	61866	10321	54840	18813	42422	29439	138935

Table 4.2 Comparison of accuracy and number of classified instances from ELM data mining algorithms RandomForest, Bagging, and AdaBoost based on three datasets

Datasets	*Result variable*	*RandomForest*		*Bagging*		*AdaBoost*	
		Passive	Active	Passive	Active	Passive	Active
corporate	Accuracy %	75		73		61	
	Classified Instances	12527	12178	12544	11375	15801	4289
residential	Accuracy %	77		75		73	
	Classified Instances	11808	63423	8553	64518	4993	66405
Combination	Accuracy %	77		73.5		71	
	Classified Instances	14778	55714	10840	56340	7166	57607

AdaBoost using DecisionStump decision tree as a classifier based on a weight threshold of 100 has provided the lowest accuracy for all the datasets. Therefore, other schemes should be further investigated for evaluating the customer churn choosing the best data mining scheme among the ELM classifiers that have been used herein in this study.

RandomForest datamining scheme classifies the dataset for constructing a forest of random trees bagging with 100 iterations using RandomTree as the base-learner scheme. Table 4.3 and Table 4.4 regarding the RandomForest classifier scheme are presented to show its stability. The result summaries and the confusion matrices are obtained from the ‘corporate dataset’ by running the Random Forest scheme based on testing the dataset choosing ‘10 fold cross-validation’ and ‘use training set’ options. Since the accuracy from the testing options has given large discrepancy from each other, the scheme may not be considered to be stable for this dataset even though evaluation with ‘on training set’ test mode has provided relatively somewhat high accuracy in comparison to the other ELM classifiers. By varying the test mode, RandomForest was attempted on the individual (relatively larger) data set. One of the results from the evaluation on training set has given the accuracy of 77% as presented in Table 4.2. Another result in terms of the accuracy of 74% was obtained from the evaluation on cross-validation, which is considerably different results. In conclusion, The RandomForest data mining scheme may not be the best option among other ELM algorithms for predicting customer churn since it seems unstable. However, in the same test, Bagging provides much better stabilization by giving almost the same results of accuracy as the test option is varied.

Table 4.3 Comparison of the result summaries from Random Forest scheme dataset based on selection for testing

	Random Forest			
	Use Training Set		10 Fold Cross-Validation	
Correctly Classified Instances	24705	75.08%	21860	66.43%
Incorrectly Classified Instances	8200	24.92%	11045	33.57%
Kappa statistic	0.5019		0.3289	
Mean absolute error	0.3238		0.3774	
Root mean squared error	0.3991		0.4643	
Relative absolute error	64.77%		75.49%	
Root relative squared error	79.84%		92.86%	

Table 4.4 presents the confusion matrices that show the numbers of correctly and incorrectly classified attributes from the RandomForest scheme testing based on both options. Even though passive and active customer numbers seem different due to different accuracy and misclassified customers, the ratio between the passive and active customer numbers are very similar: 10983/10877=1.01 and 12527/12178=1.03. Also, similarly, the ratio of the passive and active customer numbers with respect to the correctly classified instances (Table 4.3) are very similar: 10983/21860=0.502 and

12527/24705=0.507. The bagging scheme has been chosen J48 as a base learner. Bagging scheme with 73% accuracy has also classified the corporate customers into active and passive classes with similar manner at close accuracy to the RandomForest scheme with 75% : 12544 / 11375 = 1.102 and 12544 / (11375 + 12544) = 0.524. corporate customer churn status ratio is around the band of 50% in predictions. In residential customers, this ratio is much lower in results in terms of classified instances by the Bagging scheme, which means the number of incidences in the churn status of passive is much lower than those of active. The ratio given in Table 4.2 is simply 8553/64518=0.13.

Table 4.4 Confusion matrices for Random Forest classifier testing based on Cross-Validation and Use Training Set

Random Forest Confusion Matrix				
Cross-Validation		Use Traning Set		
a	b	a	b	classified as
10983	5230	12527	3686	a = PASSIVE
5815	10877	4514	12178	b = ACTIVE

4.2 Decision Tree Evaluation

We have analyzed the most promising model - decision tree - in order to explain the customer churn.

First of all, decision tree model has been applied to the two different customer segments – residential and corporate- separately.

In the decision tree algorithm, each node applies a question and data is split into sub-nodes except the bottom nodes. The top node is the root node and this question is applied to the entire data. The nodes at the bottom are the leaf nodes where the algorithm reaches the final result.

The most important criterion to be considered when splitting the nodes is how to come up with the most purity in the sub-node and how to ensure that the node is pure in terms of target variable. In other words, the algorithm aimed to split the data into homogeneous(pure) sub-nodes. Purity is calculated by gini index which measures the incorrect classification probability of a randomly chosen element. In each question(node) algorithm calculates the Gini index value of every possible boundary and choose the boundary that minimizes it. Gini index varies between 0 and 1 where 0 means all elements are in a certain class, 0.5 means elements are equally distributed over the classes and 1 means elements are randomly distributed. In the figures, if the algorithm predicts the samples as not churned nodes are colored orange nodes represents not churned decision and blue nodes represents churned decision of the algorithm. The hue of the color represents the confidence level of the decision, and a more reliable decision is made in dark colors.

4.2.1 Residential customer analysis

First, we examined the decision tree result of residential customers illustrated in Figure 4.1. To explain the tree easily, we applied our algorithm by setting the maximum depth to 4. It did not changed the performance of our model substantially.

In the root node, there are 156569 samples which is our train set size. 110114 data points represent the not churned customers and 464455 data points represent the churned customers. For the root node the Gini index of the samples is 0.417. Algorithm asks as a first question that if total failure is less or equal to 0.5 which means if it is equal 0. Left branch is followed for true and right branch is followed for false answers for each questions. Therefore we reached to the right node if the total failure of the data point is greater than 0.

At the arrived node (right first level node) the Gini index is 0.079 and sample size is 19271. The question here is that if the total failure is less than 1.5 which means if it is equals 1. If the total failure value of our data point is greater than 1 we reached to the right node again.

The Gini index of the reached node (the rightmost second level node) is 0.073, and there are 15607 samples. The question is that if the total failure is less or equal to 24.5. If the total failure value of selected data point is greater we move to the right node.

At the arrived node (the rightmost third level node) Gini index is 0.107 and sample size is 511. The last question is that if the subscription time is less or equal to 104 days. If the subscription time of selected data point is greater we move to the right node. As a result our data point is predicted as a not churned customer at the leaf node.

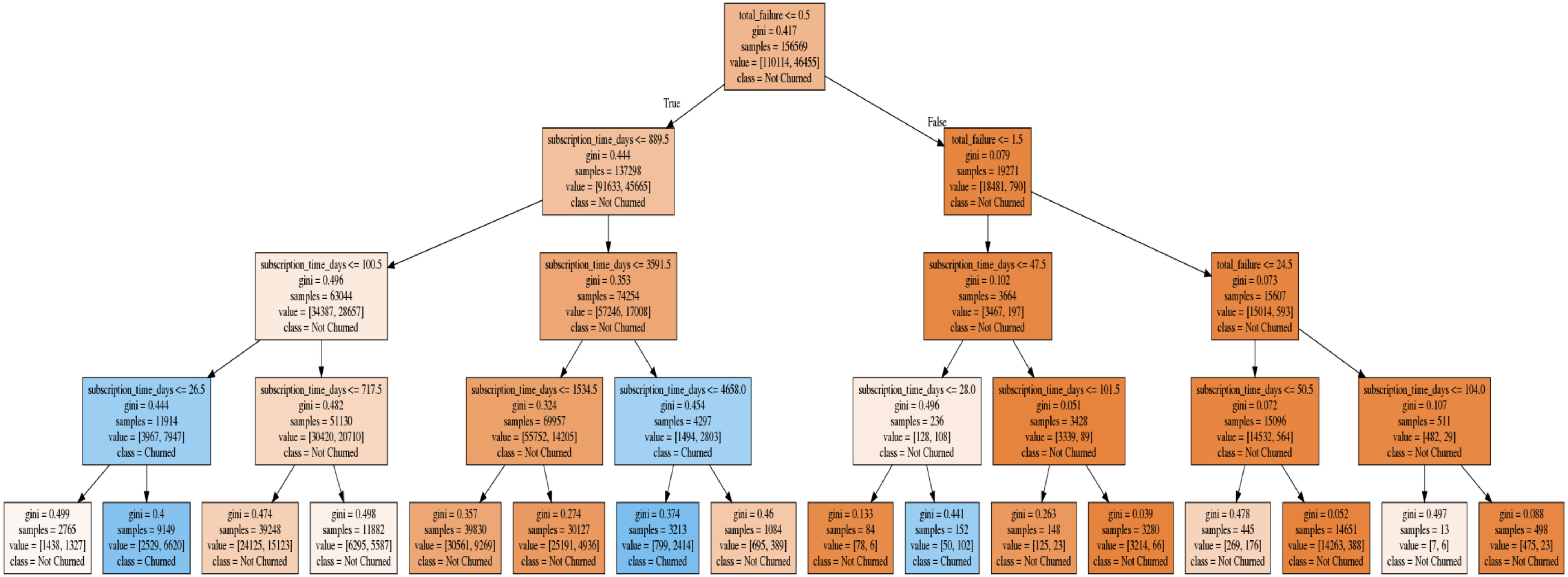

Figure 4.1 Schematic illustration of residential Customer Decision Tree Analysis

4.2.2 Corporate Customer Analysis

Secondly, we examined the decision tree result of corporate customers illustrated in Figure 4.2. We applied our algorithm by setting the maximum depth to 4 again for the sake of simplicity.

In the root node, there are 26324 samples which is our train set size. 13356 data points represent the not churned customers and 12968 data points represent the churned customers. For the root node the Gini index of the samples is 0.5 because of equal distribution. Algorithm asks as a first question that if total failure is less or equal to 0.5 which means if it is equal to 0. Therefore we reached to the right node if the total failure of the data point is greater than 0.

At the arrived node (right first level node) the Gini index is 0.126 and sample size is 3482. The question here is that if the subscription time is less than 70.5. If the subscription time value of our data point is greater than 70.5 we reached to the right node again.

The Gini index of the reached node (the rightmost second level node) is 0.093, and there are 3283 samples. The question is that if the subscription time is less or equal to 146.5 days. If the subscription time of selected data point is greater we move to the right node.

At the arrived node (the rightmost third level node) Gini index is 0.085 and sample size is 3123. The last question is that if the subscription time is less or equal to 1399.5 days. If the subscription time of selected data point is greater we move to the right node. As a result our data point is predicted as a not churned customer at the leaf no

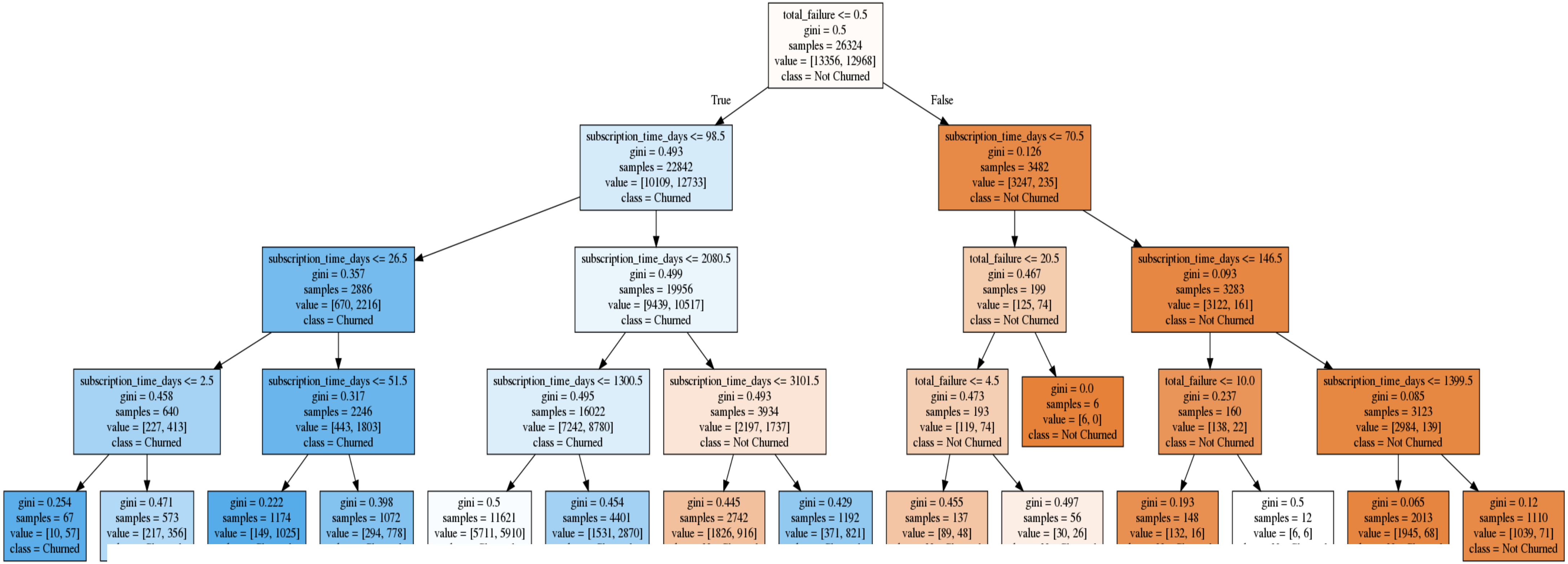

Figure 4.2 Schematic illustration of Corporate Customer Decision Tree Analysis

5 CONCLUSION

In this study, CCP analysis is carried out by using a dataset collected from customers of a private company in the telecommunication sector, and the data mining methodologies which can be utilized to detect and prevent possible customers who tend to churn are discussed. By using the experimental data from a private telecommunication company in Turkey, a set of data-mining analyses classifying a nominal attribute with regards to CCP has been conducted. Six data mining algorithms have been evaluated for the prediction of the customer churn status: Logistic Regression, Naive Bayes, J48, and ELM schemes such as RandomForest, Bagging and Boosting. RandomForest has used RandomTree, whereas the Bagging J48 is a base learner. All algorithms have been run on testing with two options to assess their stability. ZeroR scheme is used as a baseline to benchmark the accuracy performance of the other schemes. Thus first, ZeroR baseline datamining scheme is run on each dataset to have the least classification accuracy the other data mining schemes are supposed to provide at least. Then Logistic Regression, Naive Bayes and J48 decision tree schemes are run and followed by the ELM schemes such as RandomForest, Bagging, and AdaBoostM1.

RandomForest seemed to be unstable as providing considerably different results in terms of classes and prediction accuracy as the decision tree algorithms have the potential to suffer from instability with some datasets; therefore, they are required to be evaluated with this concern. The other ELM scheme named Bagging using J48 as a base learner has provided somewhat good accuracy with proper stability compared to others and J48 itself, especially in classifying the nominal attribute of churn status of the corporate dataset. However, the Bagging scheme is computationally much more expensive, requiring much higher performance and much longer time than J48. Since Bagging has not solved the large-sized database, they were edited by shuffling and truncating via relevant unsupervised incidence filters. Therefore since J48 has given similar accurate results in the residential and complete (combination of residential and corporate) data sets, depending on the capabilities in terms of computer performance and time, J48 decision tree classifier can be chosen as well as Bagging for customer churn prediction of such case studies.

Furthermore, to demonstrate the path of an algorithm while classifying the customers, a decision tree model has been implemented with the same dataset in Python. With this implementation, we have seen certain yes/no questions that the algorithm created while splitting the data and assigning a class for each instance.

As a future study, algorithms of ELM with regards to ANN such as bagged artificial neural network (BANN) and gradient boosted artificial neural network (GBANN) that are embedded in RapidMiner can be suggested to attempt as a classifier for CCP analyses in comparison to Decision Tree and other ELM forecasting algorithms such as bagged (bootstrap aggregated) regression trees (BRT), and gradient boosted regression/decision trees (GBDT).

REFERENCES

Abbasimehr, H. (2011). A Neuro-Fuzzy Classifier for Customer Churn Prediction. *International Journal of Computer Applications*, *19*(08), 35–41.

Ahmad, A. K., Jafar, A., & Aljoumaa, K. (2019). Customer churn prediction in telecom using machine learning in big data platform. *Journal of Big Data*, *6*(1). https://doi.org/10.1186/s40537-019-0191-6

Akan, M. Ö. A., Selam, A. A., Oktay Firat, S. Ü., Er Kara, M., & Özel, S. (2015). A Comparative analysis of renewable energy use and policies: Global and turkish perspectives. *Sustainability (Switzerland)*, *7*(12), 16379–16407. https://doi.org/10.3390/su71215820

Al-Mashraie, M., Chung, S. H., & Jeon, H. W. (2020). Customer switching behavior analysis in the telecommunication industry via push-pull-mooring framework: A machine learning approach. *Computers and Industrial Engineering*, *144*(October 2019), 106476. https://doi.org/10.1016/j.cie.2020.106476

Altay, T. (2005). *Knowledge Discovery in Databases and Data Mining Techniques: An Applied Study*. Marmara University.

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*(March), 290–301. https://doi.org/10.1016/j.jbusres.2018.03.003

Anuşlu, M. D., & Fırat, S. Ü. (2019). Clustering analysis application on Industry 4.0-driven global indexes. *Procedia Computer Science*, *158*(2019), 688–695.

Anuşlu, M. D., & Fırat, S. Ü. (2020). Ülkelerin Endüstri 4.0 Seviyesinin Sürdürülebilir Kalkınma Düzeylerine Etkisinin Analizi. *Endüstri Mühendisliği*, *1*(0), 44–58.

Anyanwu, M. N., & Shiva, S. G. (2011). Comparative Analysis of Serial Decision Tree Classification Algorithms. *Journal of Computer Science*, *3*(3), 230–240.

Aoga, J. O. R., Guns, T., Nijssen, S., & Schaus, P. (2018). Finding Probabilistic Rule Lists using the Minimum Description Length Principle. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *11198 LNAI*, 66–82. https://doi.org/10.1007/978-3-030-01771-2_5

Avni Es, H., Hamzacebi, C., & Oktay Firat, S. U. (2018). Assessing the logistics activities aspect of economic and social development. *International Journal of Logistics Systems and Management*, *29*(1), 1–16. https://doi.org/10.1504/IJLSM.2018.088577

Baier, L., Kühl, N., Schüritz, R., & Satzger, G. (2020). Will the customers be happy? Identifying unsatisfied customers from service encounter data. *Journal of Service Management*. https://doi.org/10.1108/JOSM-06-2019-0173

Bandara, R., Fernando, M., & Akter, S. (2020). Explicating the privacy paradox: A qualitative inquiry of online shopping consumers. *Journal of Retailing and Consumer*

Services, *52*(May), 101947. https://doi.org/10.1016/j.jretconser.2019.101947

Biçen, P., & Fırat, S. Ü. (2003). Knowledge Discovery in Databases KDD and Data Mining An Application of Customer Segmentation Analysis in Banking Sector. *International Statistical Institute 54 Th Session*.

Biçen, Pelin. (2002). *Veri madenciliği: Sınıflandırma ve tahmin yöntemlerini kullanarak bir uygulama / Data mining: Application by using predictive and classification modelling*. Yıldız Teknik Üniversitesi / Sosyal Bilimler Enstitüsü.

Ćamilović, D. (2008). Data Mining and Crm in Telecommunications. *Serbian Journal of Management*, *3*(1), 61–72.

Celik, H., & Güler, M. (2019). *The Importance of Customer Loyalty With Corporate Governance in the Telecommunication Sector*. 50–73. https://doi.org/10.4018/978-1-5225-9265-5.ch003

Celik, U., & Basarir, C. (2017). The Prediction of Precious Metal Prices via Artificial Neural Network by Using RapidMiner. *Alphanumeric Journal*, *5*(1), 45–45. https://doi.org/10.17093/alphanumeric.290381

Çiçek, A., & Arslan, Y. (2020). Müşteri Kayıp Analizi İçin Sınıflandırma Algoritmalarının Karşılaştırılması. *İleri Mühendislik Çalışmaları ve Teknolojileri Dergisi*, *1*(1), 13–19.

Çınar, A., & Silahtaroğlu, G. (2013). *Veri Madenciliği Teknikleri ile Müşteri Memnuniyetine Etki Eden Gizli Nedenlerin Keşfi*.

Coussement, K., Lessmann, S., & Verstraeten, G. (2017). A comparative analysis of data preparation algorithms for customer churn prediction: A case study in the telecommunication industry. In *Decision Support Systems* (Vol. 95). Elsevier B.V. https://doi.org/10.1016/j.dss.2016.11.007

Craven, M. W., & Shavlik, J. W. (1997). Using neural networks for data mining. *Future Generation Computer Systems*, *13*(2–3), 211–229. https://doi.org/10.1016/s0167-739x(97)00022-8

Dahiya, K., & Bhatia, S. (2015). Customer churn analysis in telecom industry. *2015 4th International Conference on Reliability, Infocom Technologies and Optimization: Trends and Future Directions, ICRITO 2015*, 1–6. https://doi.org/10.1109/ICRITO.2015.7359318

Dai, Q., Zhang, C., & Wu, H. (2016). Research of Decision Tree Classification Algorithm in Data Mining. *International Journal of Database Theory and Application*, *9*(5), 1–8. https://doi.org/10.14257/ijdta.2016.9.5.01

Deligiannis, A., & Argyriou, C. (2020). Designing a Real-Time Data-Driven Customer Churn Risk Indicator for Subscription Commerce. *International Journal of Information Engineering and Electronic Business*, *12*(4), 1–14. https://doi.org/10.5815/ijieeb.2020.04.01

Er Kara, M., Oktay Fırat, S. Ü., & Ghadge, A. (2020). A data mining-based framework for supply chain risk management. *Computers and Industrial Engineering*, *139*(December 2018). https://doi.org/10.1016/j.cie.2018.12.017

Es, H. A. (2013). *Yapay Sinir Aglari ile Turkiye Net Enerji Talep Tahmini*. Gazi University.

Es, H. A. (2018). *A novel classification approach based on multicriteria decision aiding / Çok kriterli karar destekli yeni bir sınıflandırma yaklaşımı*. Marmara University.

Fırat, S. Ü., & Biçen, P. (2003). Veri Madenciliği Tekniklerini Kullanarak Banka Müşterileri Bölümlendirmesi ve Kredi Skorlama Modeli. *Türkiye İstatistik Kurumu İstatistik Araştırma Dergisi*, *2*(2), 135–150.

Ganesh, J., Arnold, M. J., & Reynolds, K. E. (2000). Understanding the customer base of service providers: An examination of the differences between switchers and stayers. *Journal of Marketing*, *64*(3), 65–87. https://doi.org/10.1509/jmkg.64.3.65.18028

Geetha, A., & Nasira, G. (2014). Artificial neural networks' application in weather forecasting – Using RapidMiner. *International Journal of Computational Intelligence and Informatics*, *4*(3), 177–182. http://www.periyaruniversity.ac.in/ijcii/issue/Vol4No3December2014/IJCII 4-1-152.pdf

Geppert, C. (2003). *Customer Churn Management: Retaining High-Margin Customer with Customer Relationship Management Techniques*.

Goldberg, D. E., & Holland, J. H. (1988). Genetic Algorithms and Machine Learning. *Machine Learning*, *3*(4), 95–99. https://doi.org/10.1177/0954408914531118

Grąbczewski, K. (2014). *Meta-Learning in Decision Tree Induction* (J. Kacprzyk (ed.); Studies in). Springer International Publishing. https://doi.org/DOI 10.1007/978-3-319-00960-5

Gülpınar, V. (2013). Yapay Sinir Ağlari Ve Sosyal Ağ Analizi Yardimi ile Türk Telekomünikasyon Piyasasinda Müşteri Kaybı Analizi. *Marmara Üniversitesi İ.İ.B Dergisi*, *34*(1), 331–350. https://doi.org/10.14780/iibd.08553

Gülpınar, V. (2015). Yapay Sinir Ağlari Ve Sosyal Ağ Analizi Yardimi İle Türk Telekomünikasyon Piyasasinda Müşteri Kaybı Analizi. *M U Iktisadi ve Idari Bilimler Dergisi*, *34*(1), 331–350. https://doi.org/10.14780/iibd.08553

Gürsoy, U. T. Ş. (2010). Customer churn analysis in telecommunication sector. *İstanbul Üniversitesi İşletme Fakültesi Dergisi*, *39*(1), 35–49.

Han, J., Kamber, M., & Pei, J. (2014). Data mining Concepts and Techniques. In *Proceedings - 2013 International Conference on Machine Intelligence Research and Advancement, ICMIRA 2013*. https://doi.org/10.1109/ICMIRA.2013.45

Hassouna, M., Tarhini, A., Elyas, T., & Abou Trab, M. S. (2015). Customer Churn in Mobile Markets: A Comparison of Techniques. *International Business Research*, *8*(6), 224–237. https://doi.org/10.5539/ibr.v8n6p224

Karaçuha, E., Özer, G., Arasıl, Ö., & Aydın, S. (2004). Türk Gsm Sektöründe Müşteri Sadakati, Memnuniyeti, Güven Değiştirme Maliyeti Arasındaki Dinamik İlişkiler: Yapısal Denklem Modelleme Tekniği. *İktisat İşletme ve Finans*, *19*(219). https://doi.org/10.3848/iif.2004.219.7456

Karakoç, Ö., Avni Es, H., & Firat, S. Ü. (2019). Evaluation of the development level of provinces by grey cluster analysis. *Procedia Computer Science*, *158*, 135–144. https://doi.org/10.1016/j.procs.2019.09.036

Karakurt, O., Erdal, H. I., Namli, E., Yumurtaci Aydogmus, H., & Turkkan, Y. S. (2013). Comparing Ensembles Of Decision Trees And Neural Networks For One-day-ahead Stream Flow Predict. *Scientific Research Journal*, *1*(4), 1–12. https://doi.org/10.9780/23218045/1172013/41

Karakurt, O., Erdal, H. I., Namlı, E., Yumurtacı-Aydoğmuş, H., & Türkkan, Y. S. (2013). Comparing Ensembles Of Decision Trees And Neural Networks For One-day-ahead Stream Flow Predict. *Scientific Research Journal (SCIRJ), Volume I, Issue IV, November 2013*, *1*(4), 43–55. https://doi.org/10.9780/23218045/1172013/41

Karvana, K. G. M., Yazid, S., Syalim, A., & Mursanto, P. (2019). Customer Churn Analysis and Prediction Using Data Mining Models in Banking Industry. *2019 International Workshop on Big Data and Information Security, IWBIS 2019*, 33–38. https://doi.org/10.1109/IWBIS.2019.8935884

Kayaalp, F. (2017). Review of Customer Churn Analysis Studies in Telecommunications Industry. *Karaelmas Science and Engineering Journal*, *7*(2), 696–705.

Kaynar, O., Tuna, M. F., Görmez, Y., & Deveci, M. A. (2017). Makine Öğrenmesi Yöntemleriyle Müşteri Kaybı Analizi. *C.Ü. İktisadi ve İdari Bilimler Dergisi*, *18*(1), 1–14.

Kelvin, K., Cindy, C., Charles, C., Leonardo, D. P., & Yennimar, Y. (2020). Customer Churn's Analysis In Telecomunications Company Using Fp-Growth Algorithm: Customer Churn's Analysis In Telecomunications Company Using Fp-Growth Algorithm. *Jurnal Mantik*, *4*(2), 1285–1290.

Kisioglu, P., & Topcu, Y. I. (2011). Applying Bayesian Belief Network approach to customer churn analysis: A case study on the telecom industry of Turkey. *Expert Systems with Applications*, *38*(6), 7151–7157. https://doi.org/10.1016/j.eswa.2010.12.045

Li, H., Yang, D., Yang, L., Lu, Y., & Lin, X. (2016). Supervised massive data analysis for telecommunication customer churn prediction. *Proceedings - 2016 IEEE International Conferences on Big Data and Cloud Computing, BDCloud 2016, Social Computing and Networking, SocialCom 2016 and Sustainable Computing and Communications, SustainCom 2016*, 163–169. https://doi.org/10.1109/BDCloud-SocialCom-SustainCom.2016.35

Machado, M. R., Karray, S., & De Sousa, I. T. (2019). LightGBM: An effective decision tree gradient boosting method to predict customer loyalty in the finance industry. *14th International Conference on Computer Science and Education, ICCSE 2019*, *Iccse*, 1111–1116. https://doi.org/10.1109/ICCSE.2019.8845529

McKinney, W. (2011). Pandas: A foundational Python library for data analysis and statistics. Python for high performance and scientific computing, 14(9), 1-9.

Mitchell, T. M. (1997). *Machine Learning*. McGraw-Hill Science/Engineering/Math.

Mittal, M., Goyal, L. M., Sethi, J. K., & Hemanth, D. J. (2019). Monitoring the Impact of Economic Crisis on Crime in India Using Machine Learning. *Computational Economics*, *53*(4), 1467–1485. https://doi.org/10.1007/s10614-018-9821-x

Oh, C. K., & Sohn, H. (2009). Damage diagnosis under environmental and operational variations using unsupervised support vector machine. *Journal of Sound and Vibration*, *325*(1–2), 224–239. https://doi.org/10.1016/j.jsv.2009.03.014

Özekes, S. (2003). Veri Madenciliği Modelleri ve Uygulama Alanları. *İstanbul Ticaret Üniversitesi Fen Bilimleri Dergisi*, *2*(3), 65–82.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. the Journal of machine Learning research, 12, 2825-2830.

Pham, B. T., Prakash, I., Singh, S. K., Shirzadi, A., Shahabi, H., Tran, T. T. T., & Bui, D. T. (2019). Landslide susceptibility modeling using Reduced Error Pruning Trees and different ensemble techniques: Hybrid machine learning approaches. *Catena*, *175*(December 2018), 203–218. https://doi.org/10.1016/j.catena.2018.12.018

Preeti K. Dalvi, Siddhi K. Khandge, Ashish Deomore, Aditya Bankar, V. A. K. (2016). Analysis of Customer Churn Prediction in Telecom Industry using Decision Trees and Logistic Regression. *2016 Symposium on Colossal Data Analysis and Networking (CDAN)*, *88*(5), 613–618. https://doi.org/10.1177/019459988008800516

Richeldi, M., & Perrucci, A. (2002). Churn analysis case study. *Enabling End-User Datawarehouse Mining: Deliverable No. D17.2*. http://www-ai.cs.uni-dortmund.de/PublicPublicationFiles/richeldi_perrucci_2002b.pdf%0Ahttp://www-ai.cs.uni-dortmund.de:8080/PublicPublicationFiles/richeldi_perrucci_2002b.pdf

Rud, O. P. (2001). Data Mining, Cookbook: Modeling Data for Marketing, Risk and Customer Relationship Management. In *John Wiley & Sons, Inc. Canada* (Vol. 44). http://portal.acm.org/citation.cfm?id=515828

Russell, S. (2013). Chapter 4 : Machine Learning. In *Handbook ofPerception and Cognition* (Vol. 14, Issue 510).

Sahu, M. K., Pandey, R., & Silakari, S. (2018). Analysis of Customer Churn Prediction in Telecom Sector Using Logistic Regression and Decision Tree Keywords : *Journal of Applied Science and Computations*, *5*(6), 62–67. http://www.j-asc.com/gallery/11-june-756.pdf

Salzberg, S. L. (1994). Book Review : C4 . 5 : Programs for Machine Learning by J. Ross Quinlan. *Machine Learning*, *240*, 235–240.

Sauro, J. (2015). *Customer Analytics for Dummies* (2015th ed.). John Wiley & Sons, Inc.

Seker, S. E. (2016). Müşteri Kayıp Analizi (Customer Churn Analysis). *YBS Ansiklopedi*, 1–4.

Shannon, C. E. (1948). A Mathematical Theory of Communication. *Bell System Technical*

Journal, *27*(4), 623–656. https://doi.org/10.1002/j.1538-7305.1948.tb00917.x

Sjarif, N. N. A., Yusof, M. R. M., Wong, D. H. Ten, Ya'akob, S., Ibrahim, R., & Osman, M. Z. (2019). A customer Churn prediction using Pearson correlation function and K nearest neighbor algorithm for telecommunication industry. *International Journal of Advances in Soft Computing and Its Applications*, *11*(2), 46–59.

Tosun, T. (2006). *Veri Madenciliği Teknikleriyle Kredi Kartlarında Müşteri Kaybetme Analizi*. İstanbul Teknik Üniversitesi.

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. https://doi.org/10.1109/ACCESS.2019.2914999

Wang, Q. F., Xu, M., & Hussain, A. (2019). Large-scale Ensemble Model for Customer Churn Prediction in Search Ads. *Cognitive Computation*, *11*(2), 262–270. https://doi.org/10.1007/s12559-018-9608-3

Witten, I. H. (2020). *Data-mining-with-weka*. University of Waikato.

Yan, L., Miller, D. J., Mozer, M. C., & Wolniewicz, R. (2001). Improving prediction of customer behavior in nonstationary environments. *Proceedings of the International Joint Conference on Neural Networks*, *3*, 2258–2263. https://doi.org/10.1109/ijcnn.2001.938518

Yeboah-Asiamah, E., Narteh, B., & Mahmoud, M. A. (2018). Preventing Customer Churn in the Mobile Telecommunication Industry: Is Mobile Money Usage the Missing Link? *Journal of African Business*, *19*(2), 174–194. https://doi.org/10.1080/15228916.2018.1440462

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. https://doi.org/10.1007/s00521-016-2477-3

Yulianti, Y., & Saifudin, A. (2020). Sequential Feature Selection in Customer Churn Prediction Based on Naive Bayes. *IOP Conference Series: Materials Science and Engineering*, *879*(1). https://doi.org/10.1088/1757-899X/879/1/012090

Zhang, C. X., Zhang, J. S., & Wang, G. W. (2008). An empirical study of using Rotation Forest to improve regressors. *Applied Mathematics and Computation*, *195*(2), 618–629. https://doi.org/10.1016/j.amc.2007.05.010