Research Article / Araştırma Makalesi

https://academicopinion.org/

Determining Customer Risk Factors in an Insurance Company through Data Mining Analysis

Veri Madenciliği Analizi ile Bir Sigorta Şirketinde Müşteri Risk Faktörlerinin Belirlenmesi

Deniz MERDİN[1] Yasemin SAĞLAMCI[2]

[1] Tokat Gaziosmanpaşa University, Faculty of Engineering and Natural Sciences, Department of Industrial Engineering, Tokat/Türkiye, deniz.merdin@gop.edu.tr

[2] Tokat Gaziosmanpaşa University, Pazar Vocational School, Department of Social Services and Counseling, Tokat/Türkiye, yasemin.yildirim@gop.edu.tr

ABSTRACT

In recent years, data mining applications have been widely used in academic and sectoral studies as they provide speed and flexibility to sectors in areas such as decision support systems, market strategy, and financial forecasts. Data mining tools, which enable access to meaningful information in data stacks, contribute to increasing the sustainability level of companies. Data mining techniques are widely used in customer relationship management to improve flexibility, agility, and the ability to meet changing dynamics in customer demands. In this way, while enhancing the customer experience of companies, customer satisfaction and loyalty are also increased. This study aims to investigate the variables that affect the insurance premium value that customers will pay in an insurance company and the effect of the services provided by agencies on the insurance premium. Within the scope of the study, information on 150 customers of an insurance company operating in Ankara who have traffic insurance and the vehicle to be insured was collected, analyzed, and classified using data mining techniques. As a result of the study, the variable that affects the insurance premium the most is the traffic insurance level, i.e. the driver's no-claims status. Whether the services provided by the agency affect the determination of the insurance premium was investigated. As a result, no difference was found in the utilization of the services.

Keywords: *Data Mining, Insurance Risk Factors, Classification*

ARTICLE INFOS

Article History

Received: May 02, 2024

Accepted: June 08, 2024

Article Type

Research Article

Responsible Author

Deniz MERDİN

ÖZET

Son yıllarda karar destek sistemleri, pazar stratejisi, finansal tahminler gibi alanlarda sektörlere hız ve esneklik kazandırması nedeniyle, veri madenciliği uygulamaları akademik ve sektörel çalışmalarda yaygın olarak kullanılmaktadır. Veri yığınlarının içindeki anlamlı bilgiye ulaşılmasını sağlayan veri madenciliği araçları, şirketlerin sürdürülebilirlik düzeyinin artmasına katkı sağlanmaktadır. Giderek ön plana çıkan esneklik, çeviklik, müşteri taleplerindeki değişen dinamikleri karşılayabilme yeteneklerinin gelişmesi için müşteri ilişkileri yönetiminde yaygın olarak veri madenciliği teknikleri kullanılmaktadır. Bu sayede şirketlerin müşteri deneyimleri iyileştirilirken, müşteri memnuniyeti ve sadakati de arttırılmaktadır. Bu çalışmada, bir sigorta şirketinde müşterilerin ödeyeceği sigorta prim değerini etkileyen değişkenler ve acentelerin sağladığı hizmetlerin sigorta primine olan etkisini araştırmak amaçlanmıştır. Çalışma kapsamında Ankara'da faaliyet gösteren bir sigorta şirketinin trafik sigortası yaptıran 150 müşterisine ve sigorta yapılmak istenen aracına ilişkin bilgi toplanarak analiz edilmiş ve veri madenciliği teknikleriyle sınıflandırılmıştır. Çalışma sonucunda sigorta primini en çok etkileyen değişkenin trafik sigorta basamağı yani sürücünün hasarsızlık durumu olduğu görülmüştür. Sigorta priminin belirlenmesinde acente tarafından sağlanan hizmetlerin etkisinin olup olmadığı araştırılmış ve hizmetlerden faydalanma durumuna göre farklılık bulunmadığı tespit edilmiştir. Çalışmanın potansiyel müşterileri ve risklerini belirleme açısından sigorta sektöründe faaliyet gösteren şirketlere katkı sağlaması beklenmektedir.

Anahtar Kelimeler: *Veri Madenciliği, Sigorta Risk Faktörleri, Sınıflandırma*

MAKALE BİLGİSİ

Makale Geçmişi

Başvuru Tarihi: 02 Mayıs 2024

Kabul Tarihi: 08 Haziran 2024

Makale Türü

Araştırma Makalesi

Sorumlu Yazar

Deniz MERDİN

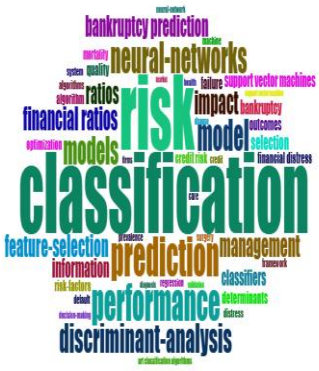
Atıf/Cite as: Merdin, D. & Sağlamcı, Y. (2024). Determining Customer Risk Factors in an Insurance Company through Data Mining Analysis. *Journal of Academic Opinion*, 4(1): 7-15.

1. Introduction

The insurance sector, which has an important place among the service sectors, is a global sector that has an important role in the economy due to its functions of sharing and reducing risk. The sector, which includes large funds around the world, operates widely in many areas in Turkiye, and some insurance such as earthquake insurance has become mandatory (Erol, 2013: 104). However, like many developing countries, insurance in Turkiye is not yet a manageable sector at the desired level. Insurance sector in Turkiye; The country faces some basic problems such as the fatalism of the people due to social and cultural characteristics, inadequacy of insurance education, gaps in legal regulations encouraging companies to unfair competition, low premium production level of the sector, inflation and imbalances in income distribution (Taşkın, 2004: 15; Ömürbek and Altın, 2008: 105-127; Güvel and Öndaş Güvel, 2004: 41).

Many academic studies have been carried out to solve these and similar problems in the sector. In the bibliometric study prepared on the keywords "financial risk analysis" and "classification", a total of 910 articles belonging to the years 2013-2023 were obtained from the ISI Web of Science database. RStudio Software was used to analyze the data. The words that occur with high frequency according to the keywords of the 910 articles examined in the study are given in Figure 1.

Figure 1
Word cloud



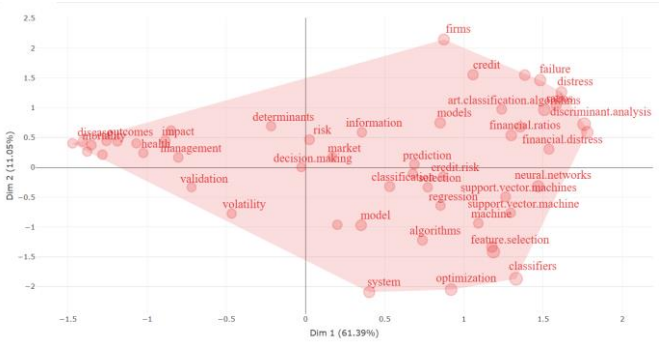
Accordingly, it is possible to say that “classification” is 10%, “risk” is 9%, “prediction” is 5%, “performance” is 4%, “model” is 4%, “neural networks” is 4%, “discrimination analysis” is 4%. Especially the keywords "risk" and "classification" are the words that have become the most popular in the last four years, as seen in Table 1.

Table 1
Usage frequency of words by years

Year	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022	2023
Classification	6	16	30	40	53	64	82	109	127	149	153
Risk	6	13	19	33	47	57	68	85	106	130	134
Prediction	4	13	18	23	30	33	43	50	54	73	75
Performance	4	8	9	12	14	24	29	41	51	67	68
Model	1	4	5	13	20	25	28	41	45	51	52
Neural Networks	3	7	10	17	22	27	34	46	52	58	58
Discriminatio n Analysis	1	8	11	18	22	28	34	41	49	57	57

When the factor analysis results of the keywords financial “risk analysis” and “classification” are examined, it is possible to say that these words are clustered in the same group.

Figure 2
Conceptual structure map



At the same time, in the studies conducted in this field, it is seen that the concepts of "credit risks", "classification" and "prevention" are very close to each other in the conceptual structure map given in Figure 2. Accordingly, it can be concluded that the keywords "credit risks", "classification" and "prevention" are used together in the studies.

When the studies in the literature are analyzed, it can be seen that Fernández-Fernández et al. (2015), Kang and Yao (2019) make classification using the clustering method. In their study, Fernández-Fernández et al. (2015) made a classification of Spanish financial institutions to contribute to improving banking supervision. They benefited from four parameters: solvency, liquidity, size, and profitability, and used the K-Means (Cluster) Analysis method (Fernández-Fernández et al., 2015: 1196-1210). Kang (2019) adopted an adaptive fuzzy weighted control method for information fusion of financial risk assessment data and big data classification (Kang, 2019: 1950021). In his study, Yao (2019) grouped the financial risk early warning model and industries into "profitable industries" and "active industries" through K-means cluster analysis in the classification (Yao, 2019: 5915-5923). Hamidi et al. (2016) and Zhao (2022) used the decision tree method. Hamidi et al. (2016) presented a new classification approach based on decision trees using parameters such as debt, equity, operating income and net income to evaluate credit risk in the Moroccan financial market. (Hamidi et al., 2016: 1-6). Zhao (2022), on the other hand, through the analysis of the financial risk source factors of the block chain digital currency, the evaluation index was created and then used the C&RT classification algorithm to analyze it and evaluate and test the model (Zhao, 2022: 1356480).

Zhang et al.(2018) and Gao et al.(2017) used Bayesian classification in their studies. Zhang et al. (2018) used Bayesian classification and Delphi method in their study to establish a financial risk intelligent early warning system for colleges and universities. (Zhang, 2018: 274-279). Gao et al., in their studies conducted in 2017 and 2018, suggest that data mining can be used for financial stress testing. They propose a new approach for stress testing financial portfolios based on Suppes-Bayes Causal Networks (SBCNs) and machine learning classification tools (Gao et al., 2017: 294-304). Huang and Wei (2021) and Hu et al. (2021) They made classifications based on deep learning. In their study, Huang and Wei (2021) proposed a new financial sensitivity classification method based on deep learning (Huang and Wei, 2021: 1-8). Hu et al. (2015) created a new risk rating method based on distance to default (DD) and order statistics (OS) to divide listed companies into three ratings according to their financial risks (Hu et al., 2015: 58-63).

Smiti et al.(2018), Çığşar and Ünal (2019) and Jabeur et al. tried to determine the most appropriate method by using different classification algorithms (Simiti et al., 2018, 658-667; Çığşar and Ünal, 2019: 8756505; Jabeur et al., 2020: 1161-1179).

It is seen that methods such as K-Means (Clustering) Analysis, Decision Tree, Bayes Classification, Regression, and Deep learning are mostly used in the studies.

Rapid changes in customer needs, trends and demands and the development of technology not only cause radical changes in all areas but also require changes in customer and service management in the insurance industry. By storing large and complex data stacks securely, meaningful data can be accessed. To predict the risks that may be encountered in the insurance sector, to increase the company's sales, and to adapt to change, different technologies must be used in the sector, especially in customer relations management. In this way, creating value for customers, ensuring customer loyalty, and identifying potential risks of customers is possible with technologies such as data mining and artificial intelligence used today (Ömürbek and Altın, 2008: 105-127, Erol, 2013: 104, Bollier and Firestone, 2010: 1-66).

Different data mining applications have been made in the insurance industry. Muslu (2009), Tosun (2006), Doğan et al (2018), Cömert, Kaymaz (2019) and Karataş (2009) used the decision tree method in their studies. In the study of Muslu, the rules for the negative outcome of damage notices, which is one of the important steps of the insurance industry, were determined and the outcome of new notices was tried to be predicted. Decision trees, one of the data mining methods, were used to determine the risks that would cause negative outcomes. For this purpose, Orange software, one of various data mining software, was used. The application was developed and the rules of the decision tree resulting from the application were evaluated. It will be determined from the resulting rules that it will be meaningful as a risk item. The results of this study will help the company predict whether new notifications will result in positive or negative results (Muslu, 2009: 93-94). In the study of Tosun (2006), it was aimed to reach results by using data mining methods to find the reasons for the churn of credit card customers. Thus, in addition to information about why customers were lost, an attempt was made to predict which types of customers were lost more frequently (Tosun, 2006: 36-37). In their study, Doğan et al., (2018) analyzed customer data of an insurance company operating in Turkey with the K-Means algorithm. With their analysis, they determined the characteristics of the company's similar customers and made suggestions to develop new marketing strategies accordingly (Doğan et al., 2018: 11-18).

In their study, Cömert and Kaymaz (2019) obtained data from an agency about its customers in its database in order to explain how to use data mining as an auxiliary tool in managing the risk of fraud in insurance companies, and those who were suspected of fraud and those who were not in the damage claims were evaluated using the J.48 algorithm of the decision tree model. It was tried to be estimated through (Cömert and Kaymaz, 2019: 364-390). In the study of Karataş (2021), a survey form was created to determine the approaches of insurance customers to the insurance concept and sector, their opinions on traffic and automobile insurance, and to determine the factors related to traffic accidents by performing risk analysis, and the data were obtained and analyzed with the SPSS package program. The data collected with the survey form was also analyzed with Decision Tree, one of the Data Mining models. In the analysis study where the Decision Tree model was used, the C&RT algorithm and CHAID algorithms were tested and it was seen that the C&RT algorithm had fewer errors than the CHAID algorithm and the C&RT algorithm was used (Karataş, 2021: 8).

Erol (2013), Kasap (2007), Hsieh (2004), Izadparast et al. (2022), Ata et al.(2008), Gep and Kumar(2012), Şahin (2018), Ata (2008), Seferzade and Dönmez (2020) used more than one data mining method in the insurance sector in their studies and explained which method was more successful in analyzing the available data. In his study, Erol (2013) reveals the stages of the Atadata and knowledge discovery process in the insurance sector for customer relationship management with example studies. Data were taken from the databases of a leading insurance company in its sector operating in Turkey, and Apriori, K-Means, and Kohonen Networks algorithms, which are among the main algorithms of VM, were applied to the data sets. Following the application, information regarding Customer Relationship Management was obtained (Erol, 2013: 104). In Kasap's (2007) study, the results of using data mining in the insurance sector were evaluated using customer data of an insurance company. When we look at the main headings of the analyses applied to the data set, these are association rules analysis, classification analysis, and clustering analyses. In the study, customer relationship management and data mining techniques, which are widely used especially in the banking sector, were tried to be applied in the insurance sector. By revealing the relationships between product-customer and company-customer, an effort was made to increase policy sales according to customers' preferences (Kasap, 2007:132-134). In his study, Şahin (2018) aimed to estimate the risk level of a new customer in the automobile insurance branch of the insurance industry by making a risk assessment in line with the information contained in the customers' automobile insurance policies. He preferred decision tree and artificial neural network methods for his research. When he compared the insurance risk prediction performances of the models he obtained with his analysis, he found that both were at an acceptable level and the prediction success of decision tree management was higher (Şahin, 2018: 55-56). In his study, Ata (2018) used Association Rules, one of the Clustering Analysis and Predictive Data Mining algorithms, to determine the customer profile of a company operating in the insurance brokerage sector and the suitable products for the company's customers. With this study, it was determined what kind of profile the best customer base draws and what kind of campaigns can be made for which products (Ata, 2018: 57-58). Hsieh, (2004) created a behavioral scoring model for a bank's credit card customers using neural networks and association rules, and aimed to increase customer loyalty by dividing customers into different groups according to their behaviors and characteristics and recommending management strategies appropriate to the characteristics of each group (Hsieh, 2004: 623-633). The aim of Izadparast et al.'s (2022) study is to classify customers with similar characteristics and estimate the approximate risk for each category according to these characteristics. For this, they used decision trees and clustering methods. According to the results obtained, the decision trees model gave better results

(Izadparast, 2012: 699-722). Ata et al. (2008) examined survival analysis methods within the framework of data mining and then examined survival probabilities, damage probabilities, and regression models for a data set of credit card holders. Risk factors that affect customers' decision to stop using credit cards were tried to be determined using regression models. It was concluded that the Weibull regression model was the most appropriate regression model for the data set. Accordingly, the study found that age, income, and marital status are important risk factors affecting customers' decision to stop using credit cards (Ata et al., 2008: 33-42). In their study, Gepp and Kumar (2012) compared digital data mining techniques and decision trees for fraud detection in automobile insurance (Gepp and Kumar, 2012: 537-561). In their study, Seferzade and Dönmez (2020) aimed to cluster insurance risk groups and insurance customers using machine learning and data mining methods. K-means clustering and hierarchical Agglomerative Clustering Algorithms were used as clustering methods (Seferzade and Dönmez, 2020: 1-7).

In this study, data regarding the customer, vehicle, and services provided, according to the vehicle's insurance premium, were classified and interpreted using decision trees, naive bayes, and random forest methods used in data mining. The study differs from other studies in the literature in that it investigates the impact of the services provided by agencies on insurance premiums and reveals the studies conducted in the field through bibliometric analysis.

1.1. Insurance Industry

The insurance sector is both a national and global sector that plays a key role in the financial system and real economy in terms of risk sharing and risk reduction functions. The insurance market, which is an integral part of the financial market as well as banking and capital markets, includes very large funds around the world. With this feature, the share of the insurance sector in the economy is increasing depending on regional and global developments (Umut, 2006; cited in Ömürbek and Altın, 2008: 108).

It is seen that Turkiye's insurance sector has changed in parallel with the developments in the economy in recent years. The insurance sector can't develop on its own without economic development and without making the economy competitive with foreigners.

Premium production in Turkiye is constantly increasing in real terms, but when compared to international data around the world, it remains well below that of developed European and OECD countries. Many reasons can be considered as the reason for this situation, such as socio-cultural factors, economic and human factors, inadequacies and marketing problems in insurance companies, intense competitive conditions in the sector reducing profitability, and the financial efficiency of insurance companies (Atilla and Gülay, 2022: 30-45).

It is possible to classify the main problems faced by the insurance sector in Turkiye as awareness, ethics, economics, human resources, legal and legal product and marketing, structural, competition, and trust problems. These problems are explained as follows (Karaman, 2018: 29-37):

Awareness: The fact that awareness about the insurance sector is not fully established in Turkiye, insurance is seen as a luxury, individuals have false and incomplete knowledge about the insurance sector, individuals approach price-oriented and do not question its contents, and the low level of education as a society is among the problems related to awareness in insurance.

Ethics: Reasons such as incomplete and insufficient information of customers, abuse of customers, seeing customers as economic objects, unethical competitor sales, agencies giving discounts to customers by cutting their commissions, and therefore acting against the rules of the competition are some of the ethical problems experienced in insurance. In addition, moral hazard is an ethical problem in which customers engage in risky behavior after taking out insurance, and as a result, insurance companies have to bear the accident costs. This situation is especially seen in company vehicles where collective agreements are made (Weisburd, 2015: 301–313). Insurance companies are exposed to high-risk customer portfolios due to customers providing incorrect information or information asymmetry. This situation, called adverse selection, results in high premium levels for customers and low profit margins for insurance companies. As a result, both situations are seen as unethical problems for the sustainability of the insurance industry (Einav and Finkelstein, 2011: 115-38).

Economic factors: Inflation in Turkiye's economy, imbalance in income distribution, and the constant increase in exchange rates negatively affect the insurance sector. This is why there are many problems. These can be listed as the agencies not being able to collect their collections on time, their commissions falling, not being able to pay a satisfactory wage to their employees, commission and income tax being different, price fluctuations, and the gap between the car insurance figures being very wide.

Human resources: The reluctance of qualified personnel to work in the insurance sector due to low wages, the reluctance of employees, and the inability to train well-equipped individuals in schools are considered as problems experienced in the insurance sector from the human resources perspective.

Legal factors: According to TOBB's report titled Insurance Agencies World Practices Research and Determination of 2023 Vision, the regulation requires agencies to operate within a more corporate structure, and increasing costs put pressure on agencies. For this reason, stronger corporate agencies can survive in the sector (TOBB Insurance Agencies Executive Committee: 57).

Product and marketing: The insurance industry has not yet fully transitioned to a customer-marketing-oriented structure. For this reason, customer needs and expectations cannot be fully met and an environment cannot be created for the spread of insurance awareness. Certain products are marketed, but there are deficiencies in developing products that will best meet the needs of the customer (Özüdoğru and Çetin, 2017: 61).

Structural problems: Lack of an independent organization for the insurance sector, insurance companies putting pressure on agencies, not making claim payments on time, and problems in acquiring dealerships are among the structural problems.

Competition problems: With the development of technology in recent years, the inability of agencies to keep up with digitalization, the spread of internet insurance, the prices given by rival companies being available on the internet platform, and insurance companies trying to attract customers by keeping their prices low to attract more customers create a fiercely competitive environment (Yayla, 2019, 11).

Trust issues: An environment of distrust is created in the insurance industry due to agencies delaying payments, not notifying owners of accident-ridden vehicles on time, and not fulfilling the conditions in the policies.

Premium Production Level of the Turkish Insurance Sector: When the premium production of the Turkish insurance sector is evaluated over the years, it is observed that it has increased both nominally and in real terms. However, it is still at very low levels (Güvel and Öndaş Güvel, 2004: 41).

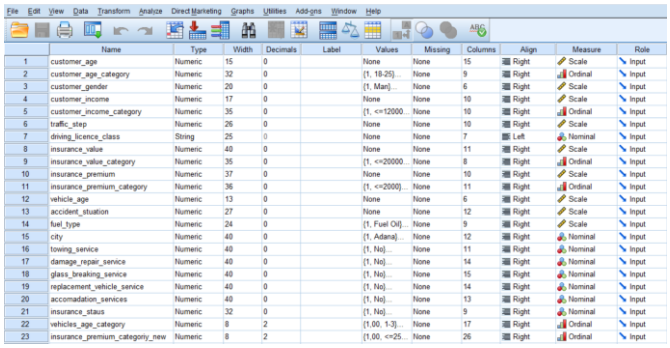
In the EU and developed countries, the insurance sector is among the indispensable sectors of the capital market. In these countries, one of the most important functions of insurance is to create the funds necessary for economic development. Life insurance companies create long-term funds needed by the economy, and non-life insurance companies create short and medium-term funds (Alkan, 2006; cited in Ömürbek and Altın, 2008: 109).

The Turkish insurance sector has a similar structure to the insurance sector in Poland, Hungary, and the Czech Republic when compared to EU countries in terms of criteria such as the number of companies and employment. However, Turkey is at a competitive disadvantage against EU countries due to reasons such as the insufficient number of insurance companies operating in the insurance sector, the low employment rate in the sector, and the low total direct premiums and insurance premiums per capita. When Turkey and EU countries are compared in terms of the asset size of insurance companies, it is seen that the total assets of all insurance companies in Turkey are much less than the total assets of a company in Germany, which is also an EU member (Ömürbek and Altın, 2008: 105-127).

2. Methodology

In the study, which aims to define and classify low and high-risk customer profiles for the insurance industry, data on 150 customers was taken from an insurance agency in Ankara. It was created using data on the services used by a total of 150 insured people, as shown in Figure 3.

Figure 3
Identification of Customer Data



	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
1	customer_age	Numeric	15	0		None	None	15	Right	Scale	Input
2	customer_age_category	Numeric	32	0		{1, 18-25}...	None	9	Right	Ordinal	Input
3	customer_gender	Numeric	20	0		{1, Man}...	None	6	Right	Scale	Input
4	customer_income	Numeric	17	0		None	None	10	Right	Scale	Input
5	customer_income_category	Numeric	35	0		{1, <=12000...	None	10	Right	Ordinal	Input
6	traffic_step	Numeric	26	0		None	None	10	Right	Scale	Input
7	driving_licence_class	String	25	0		None	None	7	Left	Nominal	Input
8	insurance_value	Numeric	40	0		None	None	11	Right	Scale	Input
9	insurance_value_category	Numeric	35	0		{1, <=20000...	None	8	Right	Ordinal	Input
10	insurance_premium	Numeric	37	0		None	None	10	Right	Scale	Input
11	insurance_premium_category	Numeric	36	0		{1, <=2000}...	None	11	Right	Ordinal	Input
12	vehicle_age	Numeric	13	0		None	None	6	Right	Scale	Input
13	accident_stuation	Numeric	27	0		None	None	12	Right	Scale	Input
14	fuel_type	Numeric	24	0		{1, Fuel Oil}...	None	9	Right	Scale	Input
15	city	Numeric	40	0		{1, Adana}...	None	12	Right	Nominal	Input
16	towing_service	Numeric	40	0		{1, No}...	None	11	Right	Nominal	Input
17	damage_repair_service	Numeric	40	0		{1, No}...	None	14	Right	Nominal	Input
18	glass_breaking_service	Numeric	40	0		{1, No}...	None	15	Right	Nominal	Input
19	replacement_vehicle_service	Numeric	40	0		{1, No}...	None	14	Right	Nominal	Input
20	accomadation_services	Numeric	40	0		{1, No}...	None	13	Right	Nominal	Input
21	insurance_staus	Numeric	32	0		{1, No}...	None	9	Right	Nominal	Input
22	vehicles_age_category	Numeric	8	2		{1.00, 1-3}...	None	17	Right	Ordinal	Input
23	insurance_premium_categoriy_new	Numeric	8	2		{1.00, <=25...	None	26	Right	Ordinal	Input

As seen in Figure 3, in the data obtained from the customers, insurance number, age, gender, income level, traffic level, driving license class, in the data obtained from the vehicle, the insurance value of the vehicle, insurance premium, age, accident status, and fuel type used, tow truck, damage repair. It includes information about glass breakage, replacement vehicles, availability of accommodation services, and availability of insurance.

An explanation of some concepts in the data obtained from the insurance agency is also included to better understand the purpose of the study. These concepts are the customer's traffic level, the vehicle's insurance value, and the insurance premium.

Customer's Traffic Level According to the amendment made on April 4, 2023, in the Regulation on Tariff Application Principles in Highways Motor Vehicles Compulsory Financial Liability Insurance, the difference between the traffic insurance premiums of people who are risk-free or with low damage frequency and the drivers of vehicles with high risk or high damage frequency has been increased. The indicator regarding the change is given in Table 2 (Resmi Gazete 4 April 2023).

According to Table 2, a 50% discount is applied to drivers who have not had an accident for five years and are included in the 8th Step. Drivers who are in Step 0 and cause a lot of damage are subject to a 20% increase.

The opposite of this situation will be valid for stage 0 drivers. Damaged drivers who are at level 1 will move to level 0 if they continue to have accidents and will have to pay a 200% increase in their traffic insurance.

Insurance Premium: The monetary value paid by the insured in return for the guarantees included in the insurance contract drawn up as a result of mutual agreements between the insurance company and the insured, is called an insurance premium. Even if all the conditions of the contract are complied with, if the premium fee is not paid, the insurance contract does not come into force. Increasing the probability of the risk occurring or increasing the insurance cost causes the premium price to increase (Çipil, 2013: 57; Cited by Doğru, 2019: 25).

Table 2
Traffic insurance steps

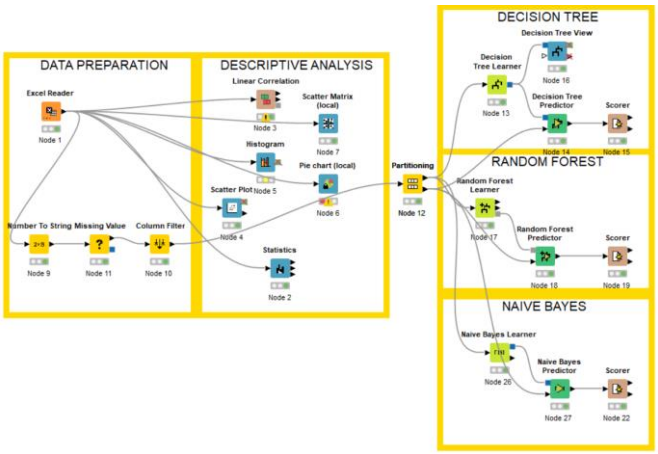
Step Number	Discount (%)	Increase (%)
8	50	-
7	40	-
6	20	-
5	5	-
4	-	10
3	-	45
2	-	90
1	-	135
0	-	200

Source: Resmî Gazete

Insurance Value of the Vehicle: It is the amount of money that the insurance company will pay in cases where the vehicle is completely damaged and is calculated according to the brand, model, year, mileage, condition, and loss of value. The insurance value list is a database of vehicle values used by insurance companies to determine the value of the vehicle. There are two basic types of car insurance value: actual cash value (ACV) and replacement cost. ACV is how much the vehicle is currently worth, while replacement cost is the cost of replacing the customer's vehicle with a similar vehicle (Generali, 2024).

For modeling, information about customers' demographic information which services they benefit from, and under what conditions was received from an insurance company. The model created with the data received from customers is given in Figure 2.

Figure 4
Modeling data



As can be seen from the model shown in Figure 4, for this study, the data was categorized according to a dependent variable determined by choosing the classification method from data mining techniques. In insurance, the premium value determined for customers is one of the most important criteria that shows the importance and reliability of customers. Therefore, when classifying the study, the "insurance premium" variable was taken as the dependent variable (output variable) and the data were classified

according to this variable. In addition, it was also investigated which variables depend on having insurance and benefiting from the services provided by the insurance company.

It was compiled by selecting from the customer data of the agency used in the research in 2023. Knime Analytics Platform was used to analyze the data. The data were first examined for missing data and extreme outliers, and some variables were filtered (Data preparation). Since there were no outliers or extreme values in the compiled data, the data did not need to be cleaned. In addition, this variable has been filtered since customers only have a Class B driving license. Then, a descriptive analysis of the remaining variables was performed. Scatter plot and box plot tests were carried out to examine the imbalances in the number of data in the class ranges of some variables and to detect data that disrupted the distribution, and as a result of the analysis, it was decided to continue with the existing data. And, the data were classified with data mining techniques using a decision tree, random forest, and naive bayes algorithms. In the algorithms, 75% of the data was used for learning and 25% for predictor. By comparing the obtained algorithm results, the best algorithm results were analyzed.

Whether the insurance premium level differs according to the services used was analyzed with the Mann-Whitney U test using the IBM SPSS Statistics 20 program, as the data was not normally distributed and the dependent variable consisted of two categories.

3. Analysis of Data

An analysis was made of the data collected from the insurance company to determine the customer profile. Table 3 contains information about the characteristics of people who have traffic insurance.

Table 3
Information regarding customer profile

	Variable	Group Sub-Variable	Number of People (N)	Percentage (%)
CUSTOMER INFORMATION	**Age**	18-25	1	0,7
		26-35	29	19,3
		36-45	41	27,3
		46-55	35	23,3
		over 55 years old	44	29,3
		Total	**150**	**100,0**
	Gender	Male	129	86,0
		Female	21	14,0
		Unspecified	0	0
		Total	**150**	**100,0**
	Income	12000 and below	14	28,0
		12001-15000	47	20,0
		15001-18000	42	11,3
		18001-21000	30	9,3
		above 21000	17	31,3
		Total	**150**	**100,0**
	Traffic Step	3	1	0,7
		4	8	5,3
		5	14	9,3
		6	22	14,7
		7	95	63,3
		8	10	6,7
		Total	**150**	**100,0**
	Driving License Class	B Class	150	100,0
VEHICLE	**Insurance Value**	200000 and below	32	21,3
		200001-400000	69	46,0
		400001-600000	29	19,3
		600001-800000	10	6,7
		above 800000	10	6,7
		Total	**150**	**100,0**
	Insurance Premium	2500 and below	5	3,3
		2501-3500	97	64,7
		3501-4500	22	14,7
		4500 and above	26	17,3
		Total	**150**	**100,0**
	Vehicle Age	1-3	17	11,3
		4-6	16	10,7
		7-11	42	28,0
		12-15	34	22,7
		16 and above	41	27,3
		Total	**150**	**100,0**
	Accident Situation	No	133	88,7
		Yes	17	17
		Total	**150**	**100,0**
	Insurance Ownership Status	No	58	38,7
		Yes	92	61,3
		Total	150	100,0
	Fuel Type Used	Gasoline	25	16,7
		Diesel	65	43,3
		LPG	60	40,0
		Hybrid	0	0,0
		Electric	0	0,0
		Total	**150**	**100,0**
SERVICES USED	**Towing Service**	No	139	92,7
		Yes	11	7,3
		Total	**150**	**100,0**
	Damage Repair Service	No	137	91,3
		Yes	13	8,7
		Total	**150**	**100,0**
	Glass Breaking Service	No	147	98,0
		Yes	3	2,0
		Total	**150**	**100,0**
	Replacement Vehicle Service	No	139	92,7
		Yes	11	7,3
		Total	**150**	**100,0**
	Accommodation Service	No	150	100,0
		Yes	0	0
		Total	**150**	**100,0**

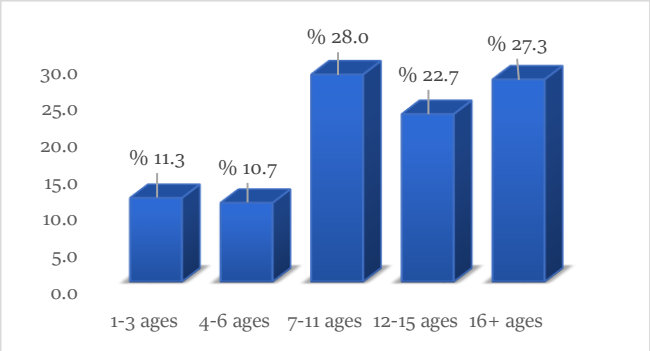
According to Table 3, the customer profiles are mostly those who are over 55 years old (29.3%), male (86.0%), have an income of 21000 and above (31.3%), and have a traffic level of 7 (63.3%). In addition, it was determined that the insured vehicles are mostly 7-11 years old (28.0%), mostly accident-free (88.7%), have insurance (61.3%), diesel (43.3%), and the insurance value is mostly 200001- 400000 TL (46.0%), and the insurance premium was in the range of 2501-3500 TL(64.7%).

Figure 5
Distribution of the provinces where the vehicle is found

According to Figure 5, it was seen that it was obtained from 150 customers residing in Turkey's Marmara, Aegean, Mediterranean, Central Anatolia, Eastern Anatolia, and Black Sea regions. According to the data collected, the city with the highest participation is Ankara.

The age distribution of customers' vehicles for which traffic insurance will be insured is given in Figure 6.

Figure 6
Age distribution of the vehicle to be insured



According to Figure 6, it is seen that the maximum age of the customer's vehicle to be insured is 22 years old and the minimum age is 1 year old, and the average age of the cars is 11.51 years old, approximately 12 years old.

Decision trees, random forests, and naive bayes algorithms were used when classifying the risk factors of people with traffic insurance. The error values for the results found in the comparison of these three methods are as in Table 4.

Table 4
Comparison of results

	Accuracy (%)	Cohen's kappa (%)
Decision Tree	91,892	0,837
Random Forest	81,579	0,633
Naive Bayes	84,211	0,692

According to Table 4, the premium value, which determines the relationship level between the dependent variable premium value and the independent variables, was predicted with 91.892% success in the decision tree algorithm, 81.579% in the random forest algorithm, and 84.211% in the naive bayes algorithm. According to these results, it is seen that the best model success is achieved with the decision tree algorithm.

The results of the analysis made with the Decision Tree algorithm are given in Figure 7.

Figure 7
Insurance premium value decision tree

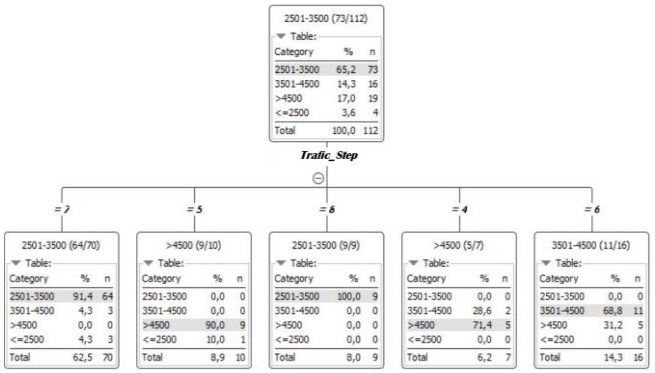
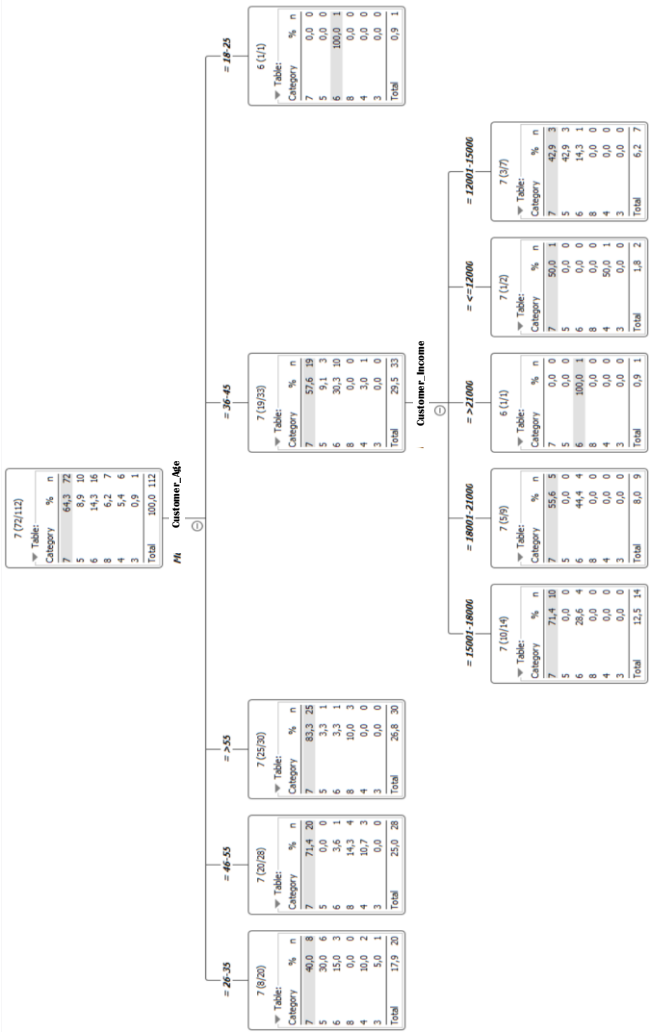


Figure 7 shows that the biggest factor affecting the determination of the insurance premium is the traffic level of the customer. It has been observed that as the traffic level increases, the insurance premium paid decreases. The results obtained support the results obtained with the regression model.

Based on these results, it was investigated which variables the traffic step depends on and a second decision tree was created.

The variables affecting the traffic level of the customer, which is the most important factor in the insurance premium value of the customers, are given in Figure 8.

Figure 8
Customer's traffic step decision tree



It was observed that the services provided by the researched insurance companies and the status of having insurance were not normally distributed (sig<0.05). For this reason, the comparison of the utilization of these services in terms of insurance premium value was made using the Mann-Whitney U test. The research results are given in Table 5.

According to Table 5, the insurance premium value does not differ depending on whether the service is provided by the agencies. It differs only according to the status of having insurance (p=0.001). While the average insurance premium of customers who do not have insurance is 3632.74 TL, the average insurance premium of customers who have insurance is 3270.02 TL. Accordingly, it has been observed that the insurance premiums of customers who do not have insurance are higher.

Table 5
Services used

Services used	Yes/ No	Average Insurance Premium (TL)	Std. Deviation	MWU	Sig (p)
Towing Service	No	3352,28	828,379	626,000	0,317
	Yes	4143,09	1798,969		
Damage Repair Service	No	3377,67	848,686	859,500	0,836
	Yes	3753,85	1678,250		
Glass Breaking Service	No	3375,84	848,603	140,500	0,282
	Yes	5097,33	3129,180		
Replacement Vehicle Service	No	3376,17	844,307	709,000	0,688
	Yes	3841,18	1812,646		
Insurance Ownership Status	No	3632,74	930,055	1836,500	**0,001***
	Yes	3270,02	934,382		

* p<0,005

4. Conclusion

Considering the insurance sector as a luxury by customers, and inadequate information about customers leads to ethical problems. There is a problem where agencies give special discounts to unethical customers by reducing their commissions. The existence of too many insurance companies, the fact that customers have the opportunity to access companies that provide services over the internet from the same portal, and adverse selection arising from insufficient information about the customer, pose a problem for agencies in terms of pricing. Therefore, accurate pricing is very important for customer satisfaction and agency continuity. Considering these problems, it is necessary to determine the factors affecting the insurance premium and reveal their impact. For these purposes, data from 150 customers of an insurance agency were collected and analyzed in the study.

As a result of the factor analysis conducted in the study, it was seen that most studies between 2013 and 2023 were conducted on classification and financial risk issues. In addition, in the studies conducted in these years, it was determined that methods such as decision trees, Bayesian classification, and regression were widely used in the classification of data. For this reason, three of the data mining techniques (decision trees, naive bayes and random forest) commonly used in the literature were used in the study to reveal financial risks and to classify and make the data meaningful. In this regard, it was aimed to determine the factors affecting the insurance premium variable by analyzing the data received from insurance agencies with these techniques. In addition, by comparing the model success of the methods used, the method that gave the best prediction result was determined. In the study, it was observed that the decision tree technique had higher model success compared to naive bayes and random forest techniques (91.89%). In addition, in the study, the factors affecting the customer's traffic level were determined and it was investigated whether the benefit of the services provided by the agency caused a difference in the insurance premium. This research contributed to the literature and according to the study, it has been observed that the most important variable affecting the insurance premium is the customer's traffic level, and as the traffic level increases, the insurance premium paid decreases. It was concluded that the traffic level varies depending on the age and income of the customer, and as the age of the customers' increases, the traffic level increases. In addition, it has been observed that the value of the insurance premium does not vary according to the status of benefiting from the services provided, but varies depending on the status of having insurance. It has been observed that insurance premiums are higher for customers who do not have insurance.

These results show that insurance companies focus on the right factors when determining insurance premiums. However, it also reveals the necessity for agencies to collect more detailed and in-depth data on customers. There are limitations in this research, such as the scarcity of data and the analysis being conducted with data obtained from a single agency. In order to increase the number of existing customers of the agencies, Doğan et al. (2018)'s conclusion that more customers can be attracted and sales can be achieved by organizing campaigns on the most preferred products for the customer is also consistent with this research (Doğan, Buldu, Demir and Erol Ceren 2018, 11-18). In addition, the research topic offers new areas to which new researchers can focus. It is especially important to learn the educational status of customers. If customers with low education levels do not question the contents, it increases the possibility of being defrauded and damages the trust in the insurance industry. In addition, it is expected that creating a more detailed customer profile will allow insurance companies to make customer-specific pricing and prevent agencies from giving unethical discounts to customers independent of insurance companies. In addition, it may be recommended to create integrated portals for agencies to follow technological developments. In this way, customer experiences can be improved by checking the accuracy of customer information. It will be possible to prevent unethical practices and to conduct risk analyzes more effectively and quickly according to the customer profile. In this way, customer expectations, customer-specific needs and pricing studies, and customer consumption habits can be regulated It is expected that the dissemination and integration of these applications will increase the strength and reliability of the insurance sector in Turkiye.

5. Acknowledgement

We would like to express our gratitude to the insurance agency that helped obtain the data.

References

Ata, N., Özkök, E. and Karabey, U. (2008). Survival Data Mining: An Application to Credit Card Holders. *Sigma Mühendislik ve Fen Bilimleri Dergisi*, *26*(1): 33-42.

Ata, F. (2018). *Understanding Customer Value Using Data Mining Applications: A Case Study Of An İnsurance Broker*. Yüksek Lisans Tezi. İstanbul: İstanbul Arel Üniversitesi Fen Bilimleri Enstitüsü.

Atilla, İ. and Gülay, A. (2022). Türkiye'de Sigorta Prim Üretimlerinin Dünya Sigortacılık Sektöründeki Yeri. *Uygulamalı Sosyal Bilimler ve Güzel Sanatlar Dergisi*, 4(8): 30-45.

Bollier, D. and Firestone, C. M. (2010). The Promise and Peril of Big Data. Washington, DC: Aspen Institute, Communications and Society Program, 1-66

Cömert, N., and Kaymaz, M. (2019). Araç Sigortası Hilelerinde Veri Madenciliğinin Kullanımı. *Marmara Üniversitesi İktisadi ve İdari Bilimler Dergisi*, 41(2): 364-390.

Çığşar, B. and Ünal, D. (2019). Comparison of Data Mining Classification Algorithms Determining the Default Risk. *Scientific Programming,* 8706505

Doğan, B., Buldu, A., Demir, Ö. and Erol Ceren. (2018). Sigortacılık Sektöründe Müşteri İlişki Yönetimi İçin Kümeleme Analizi. *Karaelmas Fen ve Mühendislik Dergisi*, 8(1): 11-18.

Doğru, Z. (2019). *Türk Sigortacılık Sektöründe Etkinlik Analizi ve Bir Uygulama*. Yüksek Lisans Tezi. Burdur: Burdur Mehmet Akif Ersoy Üniversitesi Sosyal Bilimler Enstitüsü

Einav L. and Finkelstein A. (2011). Selection In Insurance Markets: Theory And Empirics In Pictures. *J Econ Perspect*. Winter, 25(1): 115-38.

Erol, B. (2013) *Müşteri İlişkileri Yönetimi İçin Veri Madenciliği Kullanılması ve Sigortacılık Sektörü Üzerine Bir Uygulama*, Yüksek Lisans Tezi İstanbul: Marmara Üniversitesi Fen Bilimleri Enstitüsü

Fernández-Fernández, J. A., Berajano-Vázquez, V. and Vicente-Virseda, J. A. (2015). Classification of Spanish Credit Institutions for the Purposes of Financial Supervision, *ECORFAN Journal-Mexico,* 6(14): 1196

Gao, G., Mishra, B. and Ramazzotti, D. (2018). Causal Data Science for Financial Stress Testing. *Journal of Computational Science*, 26: 294-304.

https://www.generali.com.tr/hesaplayicilar/arac-kasko-degeri-hesaplama (Erişim Tarihi:16.01.2024)

Gepp, A and Kumar, K. (2012) A Comparative Analysis of Decision Trees Vis-_a-vis Other Computational Data Mining Techniques in Automotive Insurance Fraud Detection *Journal of Data Science*, 10: 537-561

Güvel, E. A. and Öndaş Güvel (2004). *Sigortacılık*. Ankara: Seçkin Yayınları, 2. Baskı

Hamidi, K. A., Berrado, A., Benabbou, L. and Tarmouti, A. (2016). A Classification Based Framework for Credit Risk Assessment in the Moroccan Financial Market. *In , October 2016 11th International Conference on Intelligent Systems: Theories and Applications* (SITA), IEEE, 1-6

Hsieh, N. (2004). An Integrated Data Mining and Behavioral Scoring Model For Analyzing Bank Customers, *Expert Systems with Applications*, 27(4): 623-633.

Hu, W. H., Gao, F. and Huang, C. (2015,). Financial Crisis Prediction Based on Distance to Default and Feature Weighted Support Vector Machine. *In August 2015 11th International Conference on Natural Computation (ICNC), IEEE*, 58-63,

Huang, B. and Wei, J. (2021). Research on Deep Learning-Based Financial Risk Prediction. *Scientific Programming*, 1-8.

Izadparast, S.M., Farahi, A., Nejad, F. F. and Teimourpour, B. (2022). Using Data Mining Techniques to Predict the Detriment Level of Car Insurance Customers. (English).*Journal of Information Processing & Management*, 27 (3).

Jabeur, S. B., Sadaaoui, A., Sghaier, A. and Aloui, R. (2020). Machine Learning Models and Cost-Sensitive Decision Trees for Bond Rating Prediction. *Journal of the Operational Research Society*, 71(8): 1161-1179.

Kang, Q. (2019). Financial Risk Assessment Model Based on Big Data. *International Journal of Modeling, Simulation, and Scientific Computing*, 10(04),

Karaman, D. (2018). Sigortacılık Sektörünün Güncel Sorunlarının Belirlenmesi: Alanya'da Bir Araştırma. *Uluslararası Yönetim ve Sosyal Araştırmalar Dergisi*, 5(10): 29-37.

Karataş, C. (2021). *Trafik ve Kasko Müşteri Eğilimleri ve Trafik Kazasını Etkileyen Faktörlerin Veri Madenciliği ile Risk Analizi* Yüksek Lisans Enstitüsü. Karabük: Karabük Üniversitesi Lisansüstü Eğitim Enstitüsü.

Kasap, E. (2007) *Sigortacılık Sektöründe Müşteri İlişkileri Yönetimi Yaklaşımıyla Veri Madenciliği Teknikleri ve Bir Uygulama* Yüksek Lisans Tezi. İstanbul: Marmara Üniversitesi Bankacılık Ve Sigortacılık Enstitüsü Sigortacılık Bölümü

Muslu, D. (2009) *Sigortacılık Sektöründe Risk Analizi: Veri Madenciliği Uygulaması*. İstanbul: İstanbul Teknik Üniversitesi Fen Bilimleri Enstitüsü, Yüksek Lisans Tezi

Ömürbek, N., and Altın, F. G. (2008). Sigortacılık Sektöründe Bilgi Teknolojilerinin Uygulanmasına İlişkin Bir Araştırma. *Suleyman Demirel University Journal of Faculty of Economics & Administrative Sciences*, 13(3): 105-127.

Özüdoğru, H. ve Çetin, Ç. (2017). Türkiye'de Sigortacılıkta Güncel Sorunlar. *Third Sector Social Economic Review*, 52(2): 57.

Seferzade, A. and Dönmez, İ. (2020). Sigorta Müşteri Risk Gruplarinin Kümeleme Yöntemi ile Analizi *International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA)*, *Ankara, Turkey,* 1-7

Resmi Gazete, (2023), https://www.resmigazete.gov.tr/eskiler/2023/04/20230404-2.htm

Smiti, S., Soui, M. and Gasmi, I. (2018). A Comparative Study of Rule Based Classification Algorithms for Credit Risk Assessment. *In 31st International Business Information Management Association Conference: Innovation Management and Education Excellence through Vision 2020*, IBIMA 2018. 658-667

Şahin, M. (2018). *Karar Ağaçları Ve Yapay Sinir Ağları Kullanarak Kasko Sigortalarında Risk Değerlendirme*. Yayımlanmamış Yüksek Lisans Tezi, İstanbul: Yıldız Teknik Üniversitesi,

Taşkın, E. Şener, H.Y. (2004). Küreselleşme Sürecinde Türk Sigorta Sektörünün Önüne Çıkabilecek Sorunlar, Bu Sorunları Asabilmek İçin Alınması Gereken Önlemler- "Global Normlu Sigorta", *Reasürör Dergisi, Milli Reasürans T.A.S. Yayını*, Sayı 51: 15.

TOBB Sigorta Acenteleri İcra Komitesi, Sigorta Acenteleri Dünya Uygulamaları Araştırma ve 2023 Vizyonunu Belirleme https://mobil.tobb.org.tr/DuyuruResimleri/2496-1.pdf (Erişim Tarihi : 10.02.2024)

Tosun, T. (2006) *Veri Madenciliği Teknikleriyle Kredi Kartlarında Müşteri Kaybetme Analizi*. Yüksek Lisans Tezi, İstanbul: İstanbul Teknik Üniversitesi Fen Bilimleri Enstitüsü

Weisburd, S. (2015). Identifying Moral Hazard in Car Insurance *Contracts. The Review of Economics and Statistics*, 97 (2): 301–313.

Yao, L. (2019). Financial Accounting Intelligence Management of Internet of Things Enterprises Based on Data Mining Algorithm. *Journal of Intelligent & Fuzzy Systems*, 37(5): 5915-5923.

Yayla, Ş. O. (2019). Sigortacılık ve Türkiye'de Sigorta Sektörünün Durumu. *Liberal Düşünce Dergisi*, 24(94): 107-125.

Zhang, Y. (2018). Food Safety Risk Intelligence Early Warning Based on Support Vector Machine. *Journal of Intelligent & Fuzzy Systems*, 38(6): 6957-6969.

Zhao, A. (2022). Financial Risk Evaluation of Digital Currency Based on CART Algorithm Blockchain. *Mobile Information Systems*, 1356480