# Application of Knowledge Discovery in Databases to Determine Car Sales Trends in Türkiye

**Orkun Eker[1] and Ensar Gul[2]**
[1]Ford Otosan, Istanbul, Türkiye
[2]Department of Software Engineering, Maltepe University, Istanbul, Türkiye

oeker1@ford.com.tr
ensargul@maltepe.edu.tr

**Abstract**: In today's world, organizations of all sizes, including companies, institutions, and governments, manage vast amounts of data. Organizations that effectively harness data and extract valuable insights can achieve their strategic objectives. This study exemplifies the application of data mining techniques by analyzing the trend of vehicle sales. It highlights the significance of data mining in the automotive sector and demonstrates how accurate and meaningful results can be produced.

**Keywords**: Data mining, Knowledge discovery, Vehicle sales trends

## 1. Introduction

The renowned Italian sculptor Michelangelo, when asked, "How do you create your sculptures?" responded, "I do not create sculptures; the sculpture is already within the stone, I merely remove the excess." The discovery of information from raw data in much the same way. The information uncovered, akin to Michelangelo's statue of David, brings new horizons, new solutions, and also new questions.

Data mining has become an essential tool for extracting meaningful information and valuable insights from large datasets in today's world. In data-intensive industries such as the automotive sector, it is possible to optimize business processes, gain a competitive advantage, and enhance customer experience by leveraging this abundance of data.

Since the early 2000s, with the advent DSL and 3G technologies, the widespread adoption of the internet has led to a rapid increase in the amount of digital data. Websites, social media platforms, e-commerce sites, and other online platforms have started generating large volumes of data. During this period, industries such as search engines and online advertising began using data mining and big data analytics techniques to analyze user behavior.

The aim of this study is to demonstrate how Knowledge Discovery in Database(KDD) process is being used in the automotive industry to determine car sales trends. We used a dataset gathered between 2020 and 2024 in Turkey and utilized the Rapidminer program (2024) and Python libraries to analyze the data and obtain meaningful results.

The following section presents the related works. Section 3 explains the methodology followed step by step. In Section 4, we use these methods to analyze car sales trends in Turkey between 2020-2024, and finally, Section 5 presents discussion and conclusions.

## 2. Related Works

One of the foundational works in the field of Knowledge Discovery in Databases (KDD) was presented by U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth (1996) . In this work, the authors provide a comprehensive overview of the KDD process and explain the relationship between data mining and KDD, highlighting how both are connected to related disciplines such as machine learning, statistics, and database systems.

The Gordon and Betty Moore Foundation's Data-Driven Discovery Initiative (DDD) launched an Investigator Competition (IC) to identify some of the leading innovators in data-driven discovery. As a result of this competition influential works in data driven discovery are listed and summarized by M. Stalzer and C. A. Mentzel (2016).

L. Cao (2017) provides a comprehensive survey and tutorial covering the core aspects of data science. This paper explores the evolution from traditional data analysis to modern data science, introduces key concepts, and presents an overarching view of the data science era. The paper also discusses major challenges and future directions in data innovation, the essence of data analytics, emerging industrial and service opportunities within

the data-driven economy, the professional competencies required for data science education, and insights into the future of the field.

Mikalef et al. (2018) highlight that to address the transition to a data-driven era and offer guidance to practitioners on implementing big data initiatives, researchers have introduced the concept of 'BVA (Business Value Analysis) capability.' This concept refers to a company's proficiency in obtaining strategic and operational insights from big data. Fundamentally, BVA capability aims to deeply explore the strategic potential of big data and maximize the benefits derived from this rich source of information. It encompasses not only technological tools and data analysis methodologies but also the organizational resources needed to support this process. In this way, it offers a more comprehensive and diverse perspective on big data.

C. Zehir and M. Bilgetürk (2022) state that Big data, with the capability to solve business operations in real-time, contributes significantly to decision-making processes by mining understandable information from scattered data.

Data analytics includes concrete resources like data and technology, human resources such as management and technical skills, and abstract resources like data-oriented culture and organizational learning (Gupta and George 2016).

Tekerek (2011) describes data mining as the process of obtaining information and analyzing data from large-scale datasets. This process involves discovering relationships, patterns, anomalies, and trends within data found in databases, data warehouses, or files. By uncovering and analyzing this information, data mining provides essential preliminary insights for decision support mechanisms in the business domain, utilizing Knowledge Discovery in Databases (KDD) applications.

The process of information discovery in the database is an interactive and iterative process consisting of several steps. Although the concepts of data mining and information discovery may appear similar, they have distinct differences. Information discovery encompasses the entire process followed to obtain meaningful information from data, whereas data mining is one of the crucial steps in this process ( Bölükbaş 2013).

There are various techniques and approaches for big data analytics. These include methods such as data mining, machine learning, artificial intelligence, deep learning, natural language processing, statistical analysis, and visualizationv(Hastie, Tibshirani and Friedman 2009). In this study, we focus on data mining in KDD.

## 3. Methodology

The KDD process is considered in a broader framework than data mining, encompassing stages such as data processing, analytics, and application of knowledge. This process enables organizations to derive meaningful and actionable insights from raw data, aiding their use in decision-making processes of significant strategic importance. Based on this point, we can outline the steps of the knowledge discovery process as follows: Data Selection, Data Cleaning and Pre-processing, Data Transformation, Data Mining, Evaluation and Interpretation, Information Presentation and Utilization.

### 3.1 Data Selection

At this stage, raw data obtained from various sources are examined, and the data to be analyzed is selected. This selection is made based on data quality, relevant features, and analysis objectives.

In the example we will consider, an automotive company wants to perform customer segmentation by analyzing customer behavior and developing personalized marketing strategies. For this purpose, the company's data analytics team needs to select appropriate data sets from the customer relationship management (CRM) database considering the following points: Goal Setting, Source Identification, Identifying Data Features, Data Selection and Data Integration.

Once the data selection stage is complete, the company will have a clean, consistent, and appropriate data set necessary for customer segmentation. This data set will be prepared for use in subsequent data mining and analysis stages and will form the basis for a better understanding of customer behavior.

### 3.2 Data Cleaning and Pre-Processing

Filling in missing data, removing noise, and correcting outliers are performed. This stage ensures that the data is consistent and analyzable.

The main purpose of data cleaning and pre-processing is to transform the data into high-quality data. The characteristics that data must have to be considered high quality are as follows:

- Data should convey certainty and should not contain vague or ambiguous information.
- Data should be complete, with no missing values.
- Data should be internally consistent and up to date.
- Data should be within credible ranges.
- Data should be easy to understand, interpret, and analyze.

Data cleaning and pre-processing are performed as follows.

Detection and Imputation of Missing Data: At this stage, these gaps can be filled with values that will not skew the average, or with data from other customers with similar characteristics, a process known as imputation.

y replacing the average of a variable to the missing data, the mean value of the variable remains unchanged. However, this method brings the missing data's value directly to the mean, thereby reducing its distance from the mean to zero and leading to an underestimation of the variance (Işıkoğlu and Atar2020).

Detection and Handling of Outliers: It is essential to check for incorrect or out-of-range values, which we refer to as 'noise.' Situations such as a customer being 12 years old, having a vehicle count of -2, or the vehicle price being extremely low or high should be identified and addressed.

Many statistical methods can be used to detect noise, but they all share the common feature of identifying data that falls outside the standard range. The most basic method among these is the Z-score method (Dayanıklı 2021).

### 3.3 Data Transformation

Data transformation is done to reduce the size of the dataset or to create the necessary features for analysis. This involves normalizing, aggregating, or encoding data to ensure it is in a suitable format for further processing and analysis.

To make the data suitable for the model to be used, a transformation process must be performed, as how the model and algorithm process the data is of great importance. When the means and variances of the variables differ significantly, variables with higher means and variances can overshadow others and reduce their impact. Therefore, it is necessary to apply normalization to the data (Tunç and Ülger 2016).

### 3.4 Data Mining

The mining process here refers to extracting meaningful relationships, patterns, and models from datasets using mining algorithms and techniques. To carry out these processes, there are specific methods for the mining operation. The previous steps involve preparing datasets that establish the necessary conditions for mining to be performed.

The mining process to be conducted should ensure that the data yields meaningful results. Therefore, the method used to process the data is very important. The mining method applied should be suitable for our objectives. The method to be used should be analyzed and determined at the beginning of the project to meet the needs. Additionally, the methods for preparing and transforming datasets will vary according to the model.

In this study only some of the data mining techniques are used. One of them is clustering. Clustering is the process of dividing data points into groups (clusters) based on their similarities. The data is grouped by considering the similarities between them. This similarity actually refers to the distance or proximity between the data points. The vehicles were grouped according to their types and places of manufacture using this method.

For predictive modeling, both **regression analysis** and **neural networks (NN)** were employed. Regression analysis enables the forecasting of future values by mathematically modeling the relationship between variables (Coşkun, 2016). In this study, it was used to explore the connection between historical vehicle sales (independent variable) and more recent sales figures (dependent variable), assessing whether this relationship could reliably support future predictions.

**Neural networks**, particularly those based on deep learning architectures such as **LSTM** (Maier 2023), were used to capture more complex, non-linear patterns in the data. Unlike traditional regression, NN models are capable of learning intricate temporal dependencies, making them particularly effective for time-series forecasting. The

accuracy and reliability of both methods were evaluated to determine their suitability for predicting future vehicle sales

### 3.5 Evaluation and Interpretation

The evaluation, which is the final stage of the knowledge discovery process, involves examining the model's performance and the accuracy of the findings. This stage is crucial for determining how well the model can solve business problems and making any necessary adjustments (Seker and Erdogan 2016).

There are specific techniques to determine model performance. Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE) are calculated to evaluate these models {Noyan 2020).

### 3.6 Information Presentation and Utilization

Even if the model's performance metrics are high, it is also important to evaluate whether the model meets business needs and objectives. Additionally, the results of the evaluation and findings from the knowledge discovery process should be shared with relevant stakeholders. Therefore, it is necessary for the obtained results to undergo a thorough reporting process, ensuring that the findings are presented in a clear and actionable manner. To achieve this, presentations should be organized using graphs, charts, and tables to convey complex data in a more understandable way.

## 4. Analyzing car Sale Trends

In this study we tried to answer the following questions:

Which brand stands out in vehicle sales in Turkey? How are the brands ranked in terms of passenger-commercial market share? What is the status of sales for domestically produced and imported vehicles?

Is it possible and meaningful to make predictions for 2024 based on the monthly sales figures from 2021 to 2023?

### 4.1 Data Selection of Vehicle Sales

We have a dataset containing the monthly and brand-specific list of vehicle sales in Turkey for the years 2020-2024. This dataset was created by combining the monthly data published on the Automotive Distributors and Mobility Association (ODMD) website. The monthly market sales figures for 2020 to 2024 were downloaded individually from the retail sales market section at https://www.odmd.org.tr/ and then merged into a single list, formatted to suit the results we aim to achieve. By mining this data warehouse, we will generate some meaningful results.

### 4.2 Data Cleaning and Pre-Processing of Vehicle Sales

Since we have set our goals, we need to define the characteristics of our data warehouse. We have a dataset consisting of five columns and approximately 10,000 rows. This table contains the sales figures for brands by month. To use the month value in comparisons, the 1st day of the month is entered as a date. The fields in the table are as follows: DATE, BRAND, MODEL, ORIGIN and QUANTITY.

After loading the data to Rapidminer we noticed that there are some missing values in Quantity fields. The reason for the empty sales quantity values is that the Automotive Distributors and Mobility Association does not enter data for vehicles that have not been sold. Here, what we need to do is enter zero '0' for the empty values. If this field were to be included in the average calculation, we would need to fill it with the average sales quantity value so that it would not affect the calculation.

Then we will find and clean the outliers. There are some values in the sales quantity that we have identified as outliers. We can perform an outlier analysis on this field and, if necessary, filter out the outliers. For this, we will use the Detect Outlier operator. We connect the Detect Outlier operator to follow the Replace Missing Values process. As parameters, we enter the neighborhood distance and the total number of outlier records we want to find. We want to identify 10 outliers in our dataset. To perform this detection, we use the Euclidean distance function.

The Detect Outlier operator has identified the 10 most extreme values for us, we used 2 extreme values as outlier values. A filter was used to remove the outliers we found from our list.

### 4.3 Data Transformation of Vehicle Sales

We want to see how data is arranged in the 0-1 range using Min-Max Normalization. To do this, we will use the Normalize operator in Rapid Miner. After grouping by the YEAR and BRAND fields, we use the Normalize operator to list the SUM(QUANTITY) field by normalizing the data within the range of 0.0 – 1.0 using the range transformation method. In the resulting list, we see many values that are 0 (zero). This result shows us that, when looking at the annual sales quantities of brands, brands with high sales volumes in certain years stand out from other brands

### 4.4 Data Mining for Vehicle Sales

To find out which brands are sold the most, filtering and grouping operations were performed sequentially on our dataset. The Generate Attributes operator was used to extract the month and year values from the DATE field. The sales quantities grouped by year were summed using the SUM function to determine the sales quantities for each brand. A grouping was performed with the Aggregate operator using the values BRAND, YEAR, and SUM(QUANTITY).

The annual sales quantities for each brand were drawn with bars. The bars were colored according to the years. However, since we saw that there were too many brands in the graph and it became very difficult to read, vehicles with annual sales quantities of less than 20,000 units were not included in the list as shown in Figure 1.

We can see that the FIAT brand has a significant advantage in the Turkish market. Especially for the year, 2023, with sales figures approaching 200,000, it appears to be far ahead of other brands. Additionally, it is possible to say that all brands increased their vehicle sales in 2024.

To find out how the brands are ranked in terms of passenger-commercial market share, first, a grouping was done based on vehicle model. It is known that some brands do not have any commercial models, and very few dealerships do not have passenger models. Considering overall sales, it is possible to say that passenger (automobile) sales are higher than light commercial vehicle sales.

As an initial listing, let's determine the sales quantities of Passenger Cars/Light Commercial Vehicles by year. Here, like we did with the brand-based list, we took the list and finally added the MODEL field to the Aggregate operator to enable the grouping of sales quantities into Passenger Cars/Light Commercial Vehicles.

Based on this, we can say that the sales of passenger cars/light commercial vehicles, which showed a notable difference, were significantly surpassed by passenger car sales in 2023 due to a surge in automobile sales. It seems reasonable to think that brands may have conducted campaigns to direct customers toward passenger vehicles or that the increase in the number of automobile models by brand has led to a rise in sales.

To address our main question, let's examine where the brands stand in this distinction. In this case, by looking at the classification of all vehicles sold between 2020 and 2024 as Passenger Cars/Light Commercial Vehicles, we can generate the pie chart as shown in Figure 2. This shows us that:

- FIAT emerges as a significant player in Turkey, dominating both the passenger car and light commercial vehicle segments.
- While FORD is the leader in light commercial vehicle sales in Turkey, it does not seem to have established any superiority in passenger car sales.
- On the other hand, RENAULT appears to have achieved success in passenger car sales but has not shown the same level of success in light commercial vehicle sales.

Not let's answer this question: What is our situation regarding the sales of domestically produced and imported vehicles. To answer this question, the dataset was grouped according to the origin of the vehicles. Among the brands sold between 2020 and 2024, only FIAT, FORD, HONDA, HYUNDAI, ISUZU, KARSAN, RENAULT, TOGG, and TOYOTA are produced locally in our country. Brands outside of these, in addition to imported vehicles, also have certain models that are marketed as imported products.

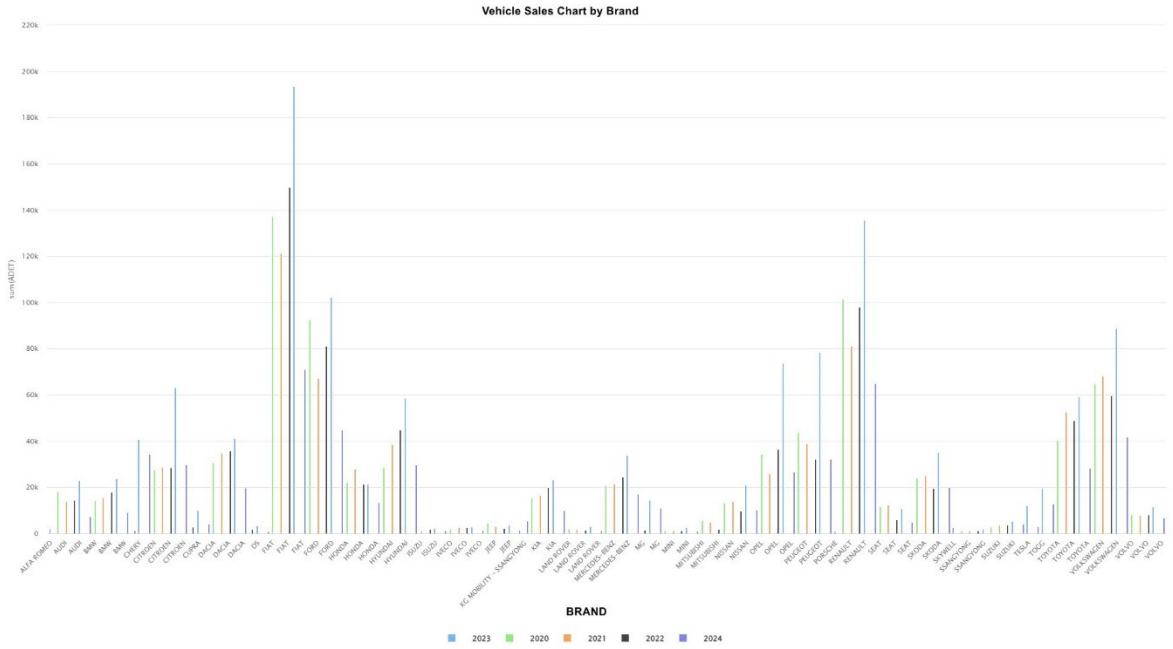


**Figure 1: Vehicle sales chart by brand**

It is observed that, except for Honda, all locally produced brands have significantly higher sales of domestically produced vehicles compared to imported vehicles. It is also possible to say that FIAT's imported models are not very popular in Türkiye.

Additionally, looking at the overall framework, we can say that nearly 40% of vehicle sales in Turkey are from domestically produced vehicles. FIAT holds the largest share in this segment. Following FIAT, RENAULT and FORD also have a significant proportion of domestic production. However, it is possible to say that the sales quantities of imported vehicles for these two brands are closer to average values compared to FIAT.

Now we will seek an answer to the following question: Is it possible and meaningful to make predictions for 2024 based on the monthly sales values from 2021 to 2023. If so, how accurate and meaningful is it? To make this analysis we do the following steps:

- First, we need a labeled and ordered dataset for the field to be predicted. Since predictions will be made based on monthly sales numbers, one group of the dataset will be used for testing, and another group will be used for analysis.
- The results obtained from the data group used for analysis will be tested on the data set aside for testing to determine the accuracy and reliability of the model

We prepared a regression model using Rapidminer and Scikit-learn (Scikit learn 2025) Python library with **FIAT** as an example brand. Among the two approaches tested, the **ARIMA model** produced more accurate results. A graphical comparison of the actual and predicted sales data for the last six months is presented in **Figure 3**.

As seen from Figure 3, although the predictions were close for the first three months, the predictions did not foresee the drop in Fiat's sales in the fourth month. The prediction performance was evaluated with an **RMSE of 6,555** and an **MAE of 6,136**..

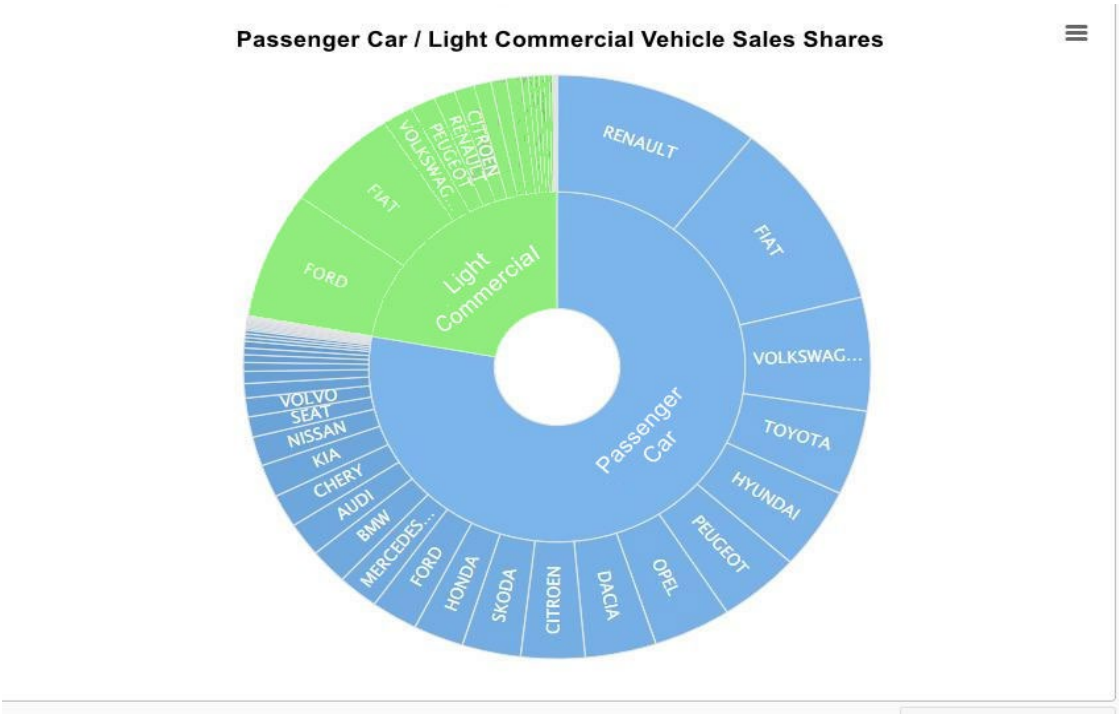


**Figure 2: Passenger car/light commercial vehicle sales shares**

In addition, we employed the **Keras (2025)** deep learning library to construct a **Neural Network LSTM model** for forecasting future sales. The results, presented in **Figure 4**, indicate that this model outperformed the regression-based approaches. The LSTM model achieved an **RMSE of 5,917** and an **MAE of 5,476**, reflecting improved accuracy than regression models. RMSE and MAE were calculated as 5917 and 5476 respectively.

However, it is not very feasible to make future sales predictions with the data we have on past monthly sales. There are many other factors affecting sales, and without data on these factors, our analysis will not produce accurate and reliable results. The results showed that past sales alone were insufficient for predictions. Incorporating diverse datasets, such as nationwide vehicle price changes or promotional campaign expirations, would improve accuracy.

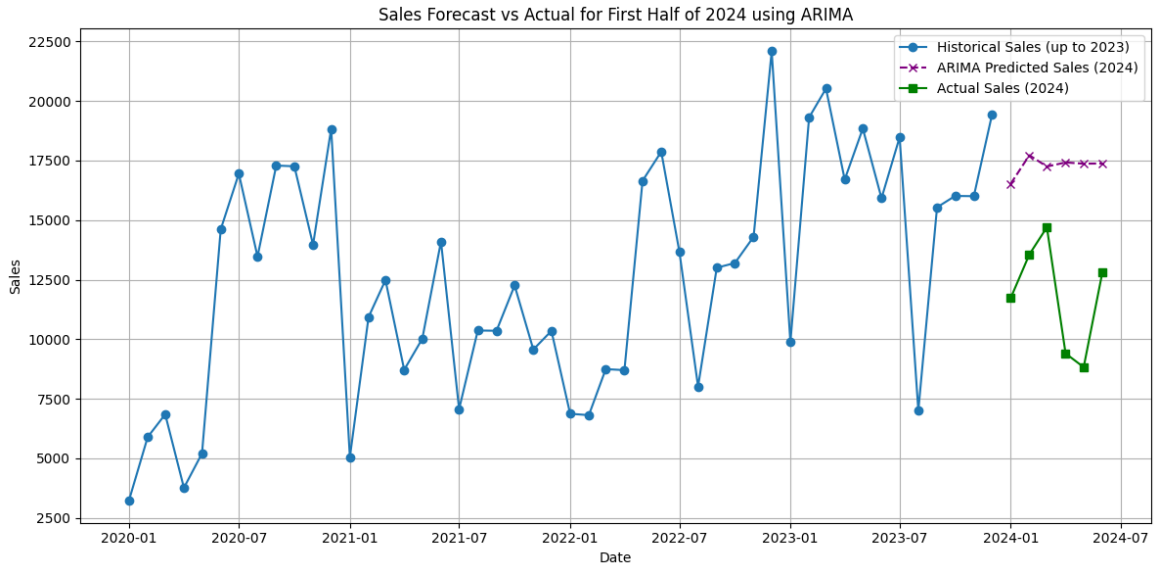


**Figure 3: Actual sales and predictions for the Fiat brand using ARIMA algorithm**

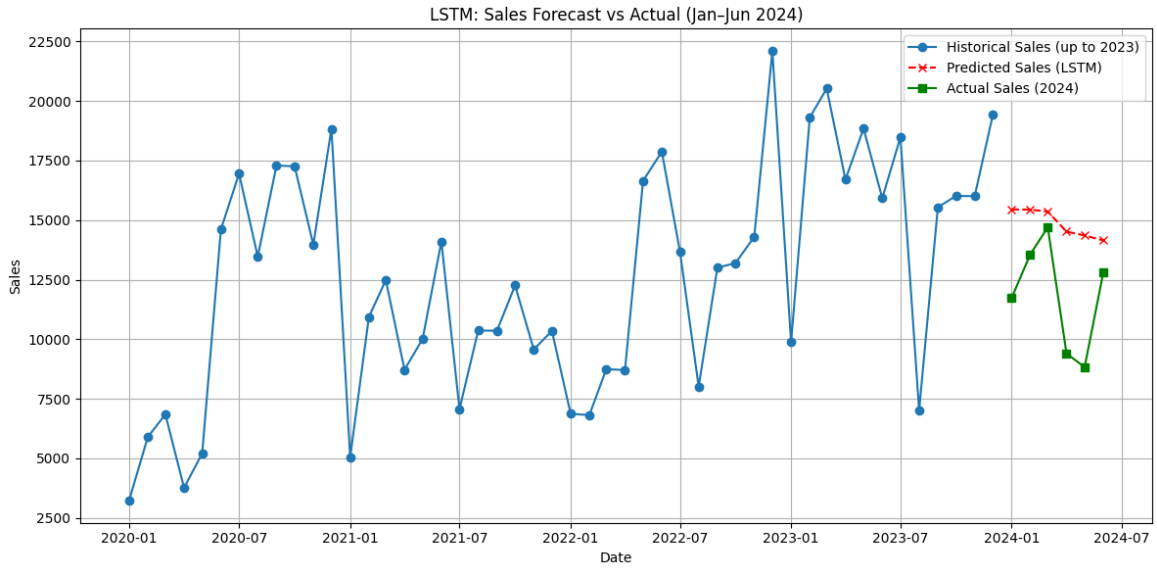


**Figure 4: Actual sales and predictions for the Fiat brand using NN LSTM model**

## 5. Discussion and Conclusions

Data mining is an indispensable method for producing meaningful results from large datasets, which can consist of thousands or even millions of rows. Data mining is used in all sectors, especially in the automotive marketing sector. The most important point here is optimizing large data sources in the best possible way. No matter how hard we try, we cannot obtain accurate results from a dataset that has not been well-optimized. Therefore, we can say that pre-processing is the most important step in mining.

There are many factors that affect car sales. Factors such as the country's tax rates, geographical conditions, and family and demographic structure greatly influence car sales. In this context, we can see that we need more data for critical issues such as sales forecasting. More complex models are necessary to make better predictions which considers not only the past sales but other economic factors and customer behavior. For this reason, automotive companies ask for a lot of information from their customers during sales. Their goal is to reach the right customer with the right options and achieve maximum sales. The most effective way to achieve this is through accurate data mining and marketing efforts.

While not everyone can create a work of art by removing the excess from a block of marble, anyone who wants to can extract meaningful results from large data piles using the right methods. The cost of learning data mining methods and applications is negligible compared to the cost of the product. The solutions produced by data mining go beyond marketing to make life easier and increase efficiency. However, alongside all its benefits, it also encompasses concerns such as privacy, security, and the misuse of data.

**Ethics declaration**: We declare that no ethical clearance was needed.

**AI declaration**: ChatGPT is used to enhance the English in this paper.

## References

Altair Home Page https://altair.com/altair-rapidminer [online] last accessed 2024/11/29

Bölükbaş, M.A., (2013), Çalışan Memnuniyetinin Veri Madenciliği ile Incelen-mesi(Examining Employee Satisfaction through Data Mining), Master Thesis, Mimar Sinan Güzel Sanatlar University Fen Bilimleri Enstitüsü, Istanbul, 2013.

Cao, Longbing (2017). Data Science: A Comprehensive Overview. ACM Computing Surveys (CSUR), Volume 50, Issue 3 ,DOI: https://doi.org/10.1145/307625

Coşkun, C. (2016) , Veri Madenciliği Algoritmaları Karşılaştırması(Comparison of Data Mining Techniques), Dicle Üniversitesi Fen Bilimleri Enstitüsü, Master Thesis, 2010.

Dayanıklı, A.S., (2021) "Aykırı değer (outlier) analizi nedir? Uç değer-ler nasıl tespit edilir?(What is outlier analysis? How to detect outliers?)", https://ravenfo.com/2021/02/11/aykiri-deger-analizi/ last accessed 2024/11/29

Fayyad, U., Piatetsky-Shapiro, G, and Smyth, P. (1996) “Knowledge Discovery in Databases”,AI Magazine Volume 17 Number 3 (1996).

Gupta, M. and George, J. F. (2016). "Toward the development of a big data analytics capability", Information and Management, 53(8), 1049–1064, 2016.

Hadoop Home Page https://hadoop.apache.org/ [online] last accessed 2024/11/29

Hastie T,, Tibshirani R., and Friedman J.(2009) , Elements of Statistical Learning, 2nd Ed., Springer, 2009.

Işıkoğlu, M.A., Atar, B. (2020) "Investigation of the effect of missing data handling Methods on Measurement Invariance of Multi-Dimensional Structures",Journal of Measure-ment and Evaluation in Education and Psychology, 11(3), pp.312-324,2020.

Keras Home Page https://keras.io/ last accessed 205/11/23

Holger R. Maier, Stefano Galelli, Saman Razavi, Andrea Castelletti, Andrea Rizzoli, Ioannis N. Athanasiadis, Miquel Sànchez-Marrè, Marco Acutis, Wenyan Wu, Greer B. Humphrey,(2023) Exploding the myths: An introduction to artificial neural networks for prediction and forecasting, Environmental Modelling & Software,Volume 167,2023, https://doi.org/10.1016/j.envsoft.2023.105776.

Mikalef, P., Pappas, I.O., Krogstie,J. and Giannakos,M. , (2018) "Big data analytics capa-bilities: a systematic literature review and research agenda" Information Systems and E-Business Management, 16(3), 547–578, 2018.

Noyan, M., (2024) , "Regresyon ve Sınıflandırmada Hata Metrikleri", [online] https://medium.com/deep-learning-turkiye/regresyon-ve-sınıflandırmada-hata-metrikleri-143a40c6b656 last accessed 2024/11/29

Scikit-learn Home Page, https://scikit-learn.org/stable/[online] last accessed 205/11/23

Stalzer M, Mentzel C. A (2016) “Preliminary review of influential works in data-driven discovery”. Springerplus. 2016 Aug 5;5(1):1266.

Şeker, S. E, and Erdoğan ,D (2016) , Rapid Miner ile Veri Madenciliği(Data Mining with Rapidminer), Bilgisayar Kavramları Yayınları, İstanbul, 2016.

Tekerek, A., (2011) "Veri madenciliği süreçleri ve açık kaynak kodlu veri madenciliği araçları(Data mining processes and open-source data mining tools), Akademik Bil-işim’11(Academic IT’11) - XIII. Akademik Bilişim Konferansı Bildirileri(Academic IT Conference Proceedings), İnönü University, 2 - 4 February 2011.

Tunç, A. and Ülger,I., (2016), "Veri madenciliği uygulamalarında özellik seçimi için fi-nansal değerlere binning ve five number summary metotları ile normalizasyon işlemi-nin uygulanması(Applying binning and five number summary methods with normal-ization for feature selection in data mining applications involving financial values)", VIII. Akademik Bilişim Konferansı(Academic IT Conference Proceedings), Adnan Menderes University, İzmir, 2016.

Zehir, C. and Bilgetürk, M., (2022) "Büyük veri analitiği yeteneği ve firma perfor-mansı ilişkisi: Firma büyüklüğünün düzenleyici rolü", Yıldız Sos Bil Ens. Der. Cilt. 6, Sayı. 2, 2022.