SUSTAINABLE BUSINESS
INTERNATIONAL JOURNAL

Revista científica - ISSN 1807-5908
Avaliado pelo sistema Double Blind Review
Disponivel em https://periodicos.uff.br/sbijournal

PPGAd uff
Programa de Pós-Graduação em Administração - UFF

Data Science Process in Furniture Industry: A Case Study

Alex Fernandes da Veiga Machado
alex.machado@ifsudestemg.edu.br

Mario Antonio Ribeiro Dantas
mario.dantas@ice.ufjf.br

Victor Stroele de Andrade Menezes
victor.stroele@ice.ufjf.br

SUSTAINABLE BUSINESS
INTERNATIONAL JOURNAL

Revista científica - ISSN 1807-5908
Avaliado pelo sistema Double Blind Review
Disponivel em https://periodicos.uff.br/sbijournal

PPGAd uff
Programa de Pós-Graduação em Administração - UFF

Abstract

Data science has become increasingly relevant in the furniture industry to help companies overcome challenges such as the scarcity of raw materials and fierce competition in the globalized market. In addition, data analysis can also help companies better understand their customer profile and identify market trends. It can be applied to several management problems, such as sales analysis, customer behavior and targeted advertising. However, although Enterprise Resources Planning systems integrate knowledge and provide reporting tools for users to analyze data, supporting decision-making is not their primary purpose. This work presents a Data Science Trajectory (DST) model applied to commercial transactions in a case study of a real company in the furniture segment. It is intended to serve as a reference for data analysts who wish to incorporate this procedure into business routines based on commercial data. From the modeling of the DST implemented with machine learning techniques, we present differentiated results related to the discovery/understanding of the problems and proposition of interventions.

Keywords: data science trajectories model, clustering, time series prediction, association rule, furniture industry

SUSTAINABLE BUSINESS
INTERNATIONAL JOURNAL
Revista científica - ISSN 1807-5908
Avaliado pelo sistema Double Blind Review
Disponivel em https://periodicos.uff.br/sbijournal
PPGAd uff
Programa de Pós-Graduação em Administração - UFF

1. INTRODUCTION

The furniture industry has faced several challenges over the years, from the scarcity of raw materials to fierce competition in the globalized market. With the evolution of technology, data science has become increasingly relevant to help companies overcome these challenges. Data analysis is one of the main trends of Industry 4.0 [1] and adopting technologies related to data science can bring significant benefits to companies, such as optimizing production, reducing costs and increasing efficiency. In addition, data analysis can also help companies to understand their customers' profiles better and identify market trends [2]. As a result, using data and analytics can help companies improve the customer experience, increasing customer satisfaction and loyalty. Data science has become an essential tool for the furniture industry, allowing companies to make more strategic decisions and remain competitive in an increasingly demanding market [3][4]. It is even more complex when more data is captured in the industry edge environments, which requires more differentiated efforts, as shown in [5].

This work presents a Data Science Trajectory model for analyzing commercial transactions in the furniture industry, offering a reference for data analysts to incorporate into business routines. In this section the problem question, the relevance of the study and the organization of the article will be presented.

1.1 Problem description

Intelligent business management, supported by a consolidated, integrated business management system (Enterprise Resource Planning, ERP), must consider the large amount of data generated. Therefore, ERP systems are developed inspired by successful business practices, aiming to automate and integrate business processes and monitor and control the production process. However, although ERP systems integrate knowledge and provide reporting tools for users to analyze data, supporting decision-making is not their main objective [6].

To extract business insights from data, companies must consider two steps: data management and analysis. Management includes capturing, recording, extracting and cleaning data, as well as integration, aggregation and representation, which are inherent in ERP solutions. In addition, modeling, analysis and interpretation are part of the analysis. As a result, analysis is critical in developing effective strategies to improve managerial decision-making.

It is possible to rely on the techniques and methodologies present in data science to support

the analysis step. McAfee and Brynjolfsson [1] note that "the more data-driven companies are, the better they do on objective measures of financial and operational results. Researchers and managers must grasp the role and implications of data science in order to fully harness the potential of the big data revolution, as stated by [7]. On average, this particular company demonstrated a 5% higher level of productivity and a 6% greater profitability compared to its competitors.

Analytical procedures based on mathematical modeling, data mining, optimization and prediction can be applied to several management problems, such as calculation of order delivery time, inventory management, logistics optimization and customer sentiment analysis for engagement of brand [7].

A study of the data generated by the ERP of a company in the furniture sector can represent a competitive advantage for this industrial niche. Nevertheless, our research intends to be the pioneer related to the furniture pole in the interior of Minas Gerais, Brazil. However, it is worth noting that there are scientific works that address data science and ERP [1][7], as well as others that study data analysis in the furniture industry [3][4].

1.2 Research Contribution

The summary of our contribution is as follows:

- We present a systemic review of the data science process in the industry with an emphasis on the furniture sector.
- Based on ERP raw data from a company in this segment, we indicate features that should be considered to start the data science process.
- We propose in this research a Data Science Trajectories model to be applied. With business intelligence and data mining techniques, we present insights extracted to analyze sales, customer behavior and targeted advertising.
- More than presenting an example of the application of this process in data extracted from the ERP of a company in the furniture segment, this work seeks to justify the importance of these methods for discovering problems, understanding difficulties and proposing interventions.

1.3 Organization of research article

The rest of the paper is organized as follows. Section 2 presents studies on processes, methods and algorithms for extracting insights from business systems. Section 3 discusses the proposed process, data mining techniques and tools. A study of the business and raw data of the furniture segment is presente

d in Section 4. We present in Section 5 the application of the proposed process and its results. Finally, section 6 concludes the paper.

2.Research background

In this section, we present some of the main data science processes applied to sales and data exploration and mining methods that are covered in our case study.

2.1 Data Science Process

One of the main methodologies for carrying out data analysis processes is the CRISP-DM (CRoss-Industry Standard Process for Data Mining) [8]. This methodology helps ensure that data science projects follow a clear and well-defined structure, which can increase project efficiency and effectiveness. However, Mart´ınez-Plumed, F. et al. [8] explored the changes and challenges faced by it over the years, proposing the emergence of the Data Science Trajectories model (DST). The authors highlighted that this model is flexible and can be applied to different problems and contexts if properly adapted. Furthermore, DST becomes a more exemplary rather than prescriptive model by identifying more exploratory activities that are common in data science but not covered by CRISP-DM.

The model trajectories are also useful to illustrate the graphical notation that we use for the trajectory charts. Mart´ınez-Plumed, F. et al. [8] illustrates some trajectories on real cases of data science projects, using a precise notation on trajectory charts. An interesting model for sales OLAP (Online Analytical Processing) is presented in Figure 1.

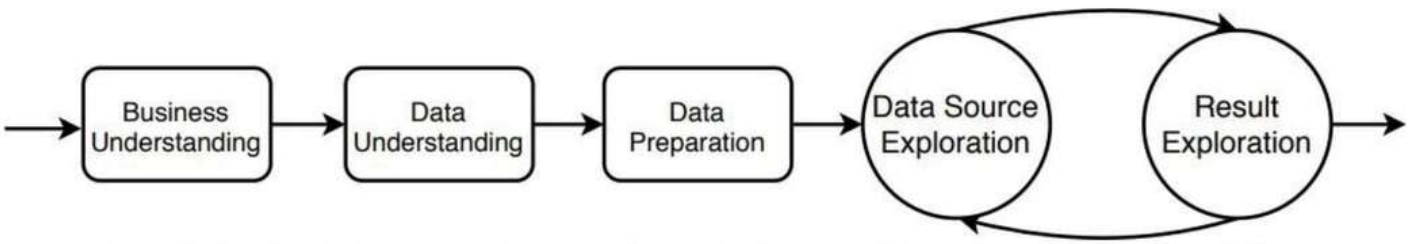


Fig. 1 Sales OLAP: Trajectory for the analysis of sales in retailing (adapted from source: [9]).

The trajectory consists of the analysis of sales in retailing. It involves the initial stages of a data mining project, spearheaded by a data scientist, which encompass business comprehension, data comprehension, and data preparation. To accomplish these tasks, ETL tools (Extract, Transform, and Load) can be employed in data warehousing, facilitating the seamless migration and integration of data from its original sources to the data warehouse. The subsequent phase of this process entails a team of analysts and managers who extract valuable insights from the data mart. By obtaining the appropriate data and thoroughly analyzing the outcomes, they engage in iterative cycles until they reach conclusive decisions [9].

1.Methods for data exploration and mining

DST is a methodology that can be applied to a wide range of analytical processes, including sales analysis, time series analysis, calculating customer retention rate, segmentation and sales recommendation. These analyses can be seen as a step within data mining, helping companies make more informed and strategic decisions.

Sales analytics involves collecting and preparing sales data to create predictive models that help better understand sales trends and patterns. For example, using time series analysis, you can identify seasonal patterns in sales data or predict future demand based on historical data. We can highlight for our bibliographic review the temporal analysis using charts with the help of aggregation data [10], for prediction the use of regression models [11], neural networks [12][13] and seasonal ARIMA model[13].

Customer retention, loyalty and churn are receiving attention in many industries, which is also an important subject in the customer lifetime value context. Customers have become more interested in the quality of service that organizations can provide them. Moreover, as the services offered by various vendors lack significant differentiation, organizations face intensified competition in their quest to uphold and enhance service quality. Customer retention can be calculated from the churn rate, considering the proportion of customers who stopped doing business given a period [14].

In the realm of customer relationship management, machine-learning models like Logistic Regression and Multi-layer perceptron can be employed to scrutinize customers' personal and behavioral data. This enables organizations to gain a competitive edge by enhancing customer retention rates [15][16]. By utilizing these models, it becomes feasible to forecast the likelihood of customers churning and identify the underlying factors contributing to churn.

Customer segmentation is a process of dividing customers into different groups based on similar characteristics, needs, or behaviors. It is a critical process for business decision-makers because it provides insights into customers' preferences, buying behaviors and how they interact with products or services. Among the methods investigated in this context, we can highlight the CLV value calculation according to the RFM parameter and also the use of RFM for segmentation with K-Means Algorithm [17] [18][19].

Developing a recommender system involves understanding the business problem, collecting and preparing data, selecting and building recommendation models, evaluating model performance and deploying the system into production. Most of the works investigated used collaborative filtering with the Apriori data mining algorithm to generate the association rules for collaborative recommendation [20][21][22].

SUSTAINABLE BUSINESS
INTERNATIONAL JOURNAL
Revista científica - ISSN 1807-5908
Avaliado pelo sistema Double Blind Review
Disponivel em https://periodicos.uff.br/sbijournal
PPGAd uff
Programa de Pós-Graduação em Administração - UFF

3. Methodology

Seeking to create a set of processes to be integrated into the ERP of companies in the furniture sector, in this section, we present the proposed Data Science Trajectories model and data mining methods, languages, frameworks and tools chosen for the case study.

3.1 Adopted data science steps

In general, the DST presented by MART´INEZ-PLUMED, Fernando et al. [9] for the problem of sales in retailing involves business understanding, data understanding, data preparation, data source exploration and result exploration, as can be seen in Figure 1. In search of a solution more adapted to the reality of the problem presented, we propose the following adjustments to this model:

- As the application will be in a specific industrial segment, the work of understanding the data source can be reduced (integrating data understanding with data preparation idea).
- We understand that it is necessary to highlight the new features construction by incorporating the Feature Engineering step. It allows the creation of new features from existing ones, often by combining or transforming them somehow. This process can involve domain knowledge, creativity and experimentation to identify new features that are more informative and relevant to the problem.
- Inserting a stage for model building can bring business advantages if incorporated into the process.
- Finally, the possibility of having fewer iterations in the process (such as the explicit cycle presented between the data source exploration and result exploration steps) can save time. We are not claiming that DST is a linearly rigid model, but interactions should not be mandatory.

Following these adjustments, we propose our case study's generic DST model shown in Figure 2.

Fig. 2 Proposed DST for commercial transactions.

Revista científica - ISSN 1807-5908
Avaliado pelo sistema Double Blind Review
Disponivel em https://periodicos.uff.br/sbijournal

The steps are defined as follows:

- Business Understanding - a general understanding of the business or a specific context. In our case, it could be related to sales, customer behavior and advertising.
- Data Acquisition - Basically, load raw data.
- Data Preprocessing - data process as select columns, grouping rows, pivot table creation, feature type transformation, reshaping table, moving average and seasonal adjustment over time series.
- Feature Engineering - as feature construction and analysis. Model Building - using data mining methods such as clustering, time series prediction and association rule.
- Visualization - creating dashboards using diagrams such as line charts, bar plots, histograms and scatter plots.

3.2 Data science and data mining methods

In our case study, according to the literature review carried out and demand from the data mining area, we listed specific methods for application in some steps as follows (Table 1).

Data Preprocessing

Method	Description
Creation of pivot table	To reshape the table based on column values to summarize the data from a more extensive table into a table of statistics. It allows us to identify patterns and trends in the data quickly. By aggregating and summarizing the data in different ways, analysts can gain insights into the relationships between variables and identify patterns that might not be apparent when looking at the data in its raw form.
One-hot encoding	Reshape table based on its column values in a transaction data process to prepare data for association rule learning. This algorithm works as follows: if we have a categorical variable with three categories, One-hot encoding will create three new columns. Each row in the dataset would have a 1 in the appropriate column and 0s in the other columns (if the respective category exists or not). This encoding preserves the information about the categories, while allowing the machine learning algorithm to use the data in its calculations. The importance of one-hot encoding lies in the fact that many data mining algorithms require numerical inputs to make predictions.
Moving average	To move the average over time series to compute a mean aggregation over a sliding window using, for example, months as aggregate time period of the time series. This prepares the data for a more comprehensive visualization.
Seasonal adjustment	It is a technique used to decompose a time series into its seasonal, trend, and residual components. By removing the seasonal differences, it provides a clearer perspective on nonseasonal trends and cyclical data that would otherwise be obscured. This adjustment enables a better understanding of the underlying base trends within the time series. The process involves applying an algorithm to eliminate noise from the data set, allowing significant patterns to become more prominent. It aims to disregard one-time outliers while considering the impact of seasonality.

Revista científica - ISSN 1807-5908
Avaliado pelo sistema Double Blind Review
Disponivel em https://periodicos.uff.br/sbijournal

Feature Engineering

Method	Description
Calculation of the Customer Binding Time (CBT)	It represents a simple and important key performance indicator (KPI) in the customer lifetime value context proposed in this study. It represents a quantifiable measure over time for the binding time of the customer since the first and the last time he/she buys a product.
Calculation of the retention rate	This function is designed to calculate the retention rate from the churn rate KPI for a business over a specified time span, using a sliding window approach. The churn rate is the percentage of customers who stop doing business with a company within a certain period of time. The relationship between retention and churn rate is [14]: RetentionRate = 1–ChurnRate
Multivariate analysis using Pearson algorithm	Computing all pairwise attribute correlations using the Pearson algorithm for multivariate analysis for ranking best features for the next step (Model Building). The algorithm works by calculating the covariance between the two variables, which is a measure of how much the variables change together. Then, the algorithm calculates the product of the standard deviations of the two variables. Finally, it divides the covariance by the product of the standard deviations to obtain the Pearson correlation coefficient. It ranges from -1 to +1, with -1 indicating a perfectly negative linear relationship, 0 indicating no linear relationship and +1 indicating a perfectly positive linear relationship.

Model Building

Method	Description
Time series prediction using vector autoregression model	Vector autoregression (VAR) models use linear regression techniques to model the relationships between multiple time series variables. It is used to forecast future values of a set of time series variables based on their past values. The VAR model extends the traditional autoregression (AR) model by allowing for multiple time series variables to be included in the model. In a VAR model, each variable is modeled as a linear function of its own past values and the past values of the other variables in the system. The model assumes that the variables are jointly stationary, meaning that their means, variances and covariances are constant over time. The VAR model is estimated using a set of historical time series data and the estimated model parameters are used to make predictions for future values of the variables.
Association rule learning with Apriori algorithm	It is the most used data mining technique to discover relationships between variables in a dataset [20][21][22]. Apriori algorithm works by iterating through the dataset to identify frequent itemsets, which are sets of items that occur together in transactions above a specified threshold and then generating association rules from those itemsets. Association rule learning with the Apriori algorithm has many applications in fields such as marketing, retail and e-commerce. It can be used to identify frequently purchased items together and offer bundle deals to customers or to understand the buying patterns of customers to improve product recommendations.
Clustering with K-Means	One of the most widely used clustering algorithms is the K-Means [17][19]. It is used to group together similar observations or data points into meaningful clusters. K-Means clustering is an iterative algorithm that partitions a set of data points into K clusters, where K is a user-defined number. The algorithm assigns each data point to the nearest cluster center and then recalculates the center of each cluster based on the mean of the data points assigned to it. The algorithm iterates until the assignment of data points to clusters no longer changes or reaches a maximum number of iterations.

Revista científica - ISSN 1807-5908
Avaliado pelo sistema Double Blind Review
Disponivel em https://periodicos.uff.br/sbijournal

3.3 Languages, frameworks and tools

To prototype the data modeling process flow, the Orange Data Mining tool [23] was used, allowing machine learning algorithms through Python scripts and visual programming. It uses widgets that are components in Orange Canvas, a visual programming environment of Orange. They represent some self contained functionalities and provide a graphical user interface (GUI). Widgets communicate with each other and pass objects through communication channels to interact with other widgets, allowing rapid prototyping of solutions.
The Python programming language was chosen due to its ecosystem of resources for data science [24]: NumPy for manipulation of homogeneous array-based data, Pandas for manipulation of heterogeneous and labeled data, SciPy for common scientific computing tasks, Matplotlib and Seaborn for visualizations and Scikit-Learn for machine learning, among others.
The development IDE is Google Collaboratory, a free notebook environment running in the cloud with a large community of developers worldwide [25].

4. Business and data understanding

Data science is a powerful tool that can help companies make better, more informed decisions. However, it is essential to understand the business and the problems the company faces. This understanding is necessary for any data science project, particularly in the furniture industry.

Companies in this segment face challenges, such as keeping up with design trends, managing a complex supply chain and improving operational efficiency, which can all be addressed using data science. It is important to delimit the database correctly, selecting the most relevant variables and ensuring that the data is clean and organized.

During the database delimitation stage, the most important tables for the problem in question and the characteristics of these data are determined to guarantee the quality and reliability of the information to be used in the analysis.

In this section, we present an overview of the case study's furniture segment, the database's delimitation and its main characteristics.

4.1 Framing the problem

This research intends to apply data science to the commercial transactions of a company, specifically in the furniture hub of Uba´, Minas Gerais, Brazil. This Local Productive Arrangement was selected due to the representation it has for the furniture industry in Minas Gerais, being considered the first furniture pole in the state and the second in the country [26], with a considerable capacity to generate jobs in the municipalities that comprise it.

However, it is worth mentioning that other scientific works study the furniture industry in its local context and the importance of commercial data analysis [3][4], although this one

4.2 Database delimitation

In order to protect the company and the researchers involved, both the data considered sensitive in the base (description of the products, name of the clients, among others) and the name of the company in the case study were omitted following the General Data Protection Law of Brazil (LGBD - Lei Geral de Proteção de Dados [27]). The summary of the data is presented in Table 2.

These attributes are the most relevant for the analysis of sales according to the company's managers. Despite representing a sampling of information from the ERP of a furniture company, the DST, the methods and the pattern of insights extracted here could be generalized to any retailer or wholesaler. However, the problem diagnosis, data exploration, analysis and intervention proposal will be exemplified on these real data, mainly to highlight the competitive differential potential that this segment can reach.

Table 2 Database delimitation.

Description	Registration of Sales and Technical Assistance Items
Extension	.csv
Origin	Company ERP
Collection interval	11/02/2016 to 11/02/2022
Records	389,985
Attributes	[numeric] INVOICE ISSUE DATE: the date when the sale was issued. [categorical] STATE: the federative unit where the sale was made. [categorical] CUSTOMER CODE: the code that identifies the customer who made the purchase. [categorical] ITEM CODE: the code that identifies the sold product. [numeric] QUANTITY: the amount of products sold in the transaction. [categorical] PAYMENT CONDITION TYPE: the type of payment condition agreed upon in the sale. [categorical] SALE ORIGIN: the channel or origin of the sale, such as physical store, online store, or commercial representative. [categorical] SITUATION TYPE: the payment method, such as check, cash or pix, for example. [categorical] COLOR CODE: the code that identifies the color of the sold product.

SUSTAINABLE BUSINESS INTERNATIONAL JOURNAL
Revista científica - ISSN 1807-5908
Avaliado pelo sistema Double Blind Review
Disponivel em https://periodicos.uff.br/sbijournal
PPGAd uff Programa de Pós-Graduação em Administração - UFF

5. Experimental Analysis

In this section, we apply the proposed Data Science Trajectories model in the database of commercial transactions delimited using the data mining methods mentioned above. To do so, we propose addressing questions such as sales analysis, customer behavior analysis and targeted advertising.
At the beginning of the description of each DST the steps will be the same: Business Understanding, associated with knowledge of the specific area (sales, customer and advertising) and Data Acquisition, representing the process of load data.

5.1 Sales analysis

We start by understanding the problem and investigating how product sales behave in the furniture sector.

5.2 Time Series Components

A time series is a sequence of observations of a variable over time. The main components are:

- Trend: general direction of the behavior of the time series over time. It can be increasing, decreasing, or constant.
- Seasonal variation: Regular, predictable fluctuations that occur over fixed periods, such as monthly, quarterly, or yearly seasonality.
- The DST of the extraction process for these components can be seen in Table 3 and the results in Figure 3.

Table 3 Description of the main steps related to the extraction of seasonal components.

Nº	Stage	Process	Description
1	Business Understanding	Sales	-
2	Data Acquisition	Load Data	-
3.1	Data Preprocessing	Select Columns	Manual selection of data attributes and composition of the data domain. The chosen columns were INVOICE ISSUE DATE and QUANTITY.
3.2	Data Preprocessing	Grouping	Groups data by selected variables and aggregates columns with selected aggregations. Grouped by INVOICE ISSUE DATE, using the sum aggregate function for the QUANTITY attribute.
3.3	Data Preprocessing	Moving average	Compute a mean aggregation over a sliding window.
3.4	Data Preprocessing	Seasonal Adjustment	Decomposition based on rates of change for decomposing the time series into seasonal and trend components. The chosen season period was 12 months. The time series decomposition model chosen was additive.
4	Visualization	Line Chart	Visualize the time series' sequence and progression.

Revista científica - ISSN 1807-5908
Avaliado pelo sistema Double Blind Review
Disponivel em https://periodicos.uff.br/sbijournal

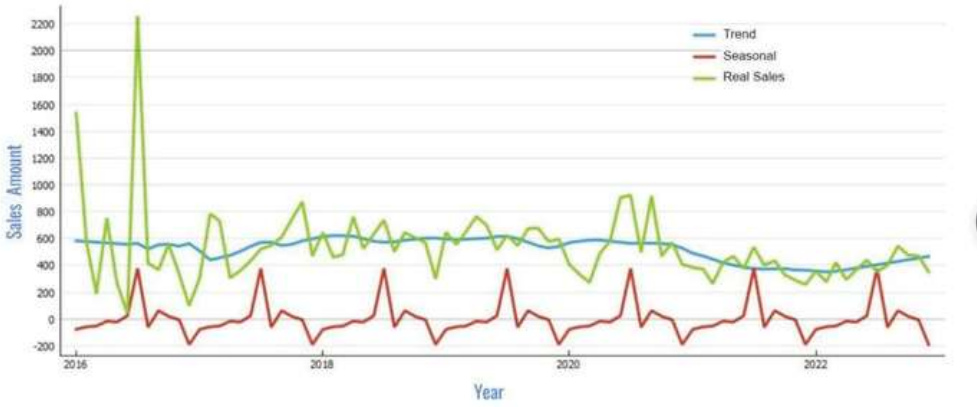


Fig. 3 Line chart showing seasonal components based on sales occurrences.

The seasonal component is clear, highlighting the sales peaks in September and July and the fall in December. We observe a slight general trend towards increased sales.

5.2.1 Sales Forecast

Sales forecast allows a company to identify product demand at different times of the year and adjust its production, inventory and marketing strategies to meet these demands. In addition, it can help the company make important decisions, such as pricing, new product selection and financial resource allocation. With an accurate sales forecast, the company can minimize costs and maximize profits, ensuring customer satisfaction and the business's survival in the highly competitive market of the furniture sector. For the general sales forecast based on the sales occurrence time series, we have the main processes described in Table 4 and results in Figure 4.

Revista científica - ISSN 1807-5908
Avaliado pelo sistema Double Blind Review
Disponivel em https://periodicos.uff.br/sbijournal

Table 4 Description of the main steps related to sales forecasting.

Nº	Stage	Process	Description
1	Business Understanding	Sales	-
2	Data Acquisition	Load Data	-
3.1	Data Preprocessing	Select Columns	Manual selection of data attributes and composition of the data domain. The chosen columns were INVOICE ISSUE DATE and QUANTITY.
3.2	Data Preprocessing	Grouping	Groups data by selected variables and aggregates columns with selected aggregations. Grouped by INVOICE ISSUE DATE, using the sum aggregate function for the QUANTITY attribute.
3.3	Data Preprocessing	Moving average	Compute a mean aggregation over a sliding window, using months as the aggregate time period of the time series.
4	Model Building	Time Series Forecasting	Time series prediction using Vector Autoregression (VAR) Model. Using 36 forecasting steps. We evaluate the model with Bayesian Information Criterion (BIC) as index.
5	Visualization	Line Chart	Visualize the time series' sequence and progression.

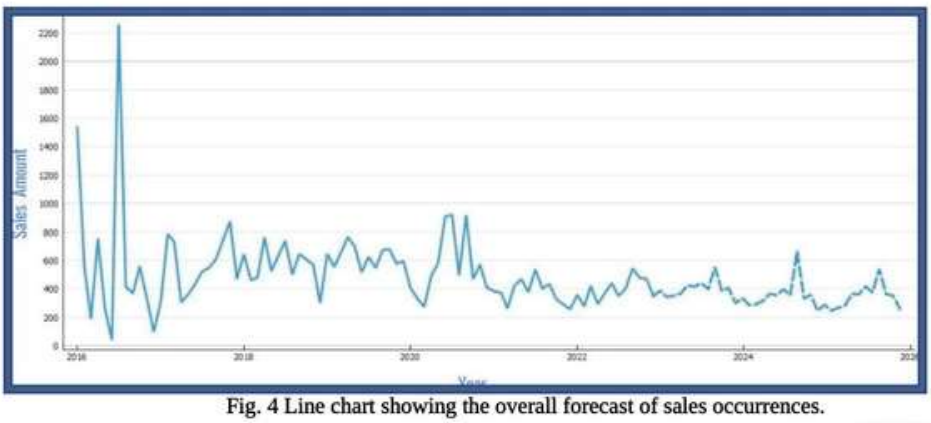


Fig. 4 Line chart showing the overall forecast of sales occurrences.

We noticed, at first, a large occurrence of sales at the beginning of the historical series, maintaining a characteristic pattern of sales from 2021 onwards.

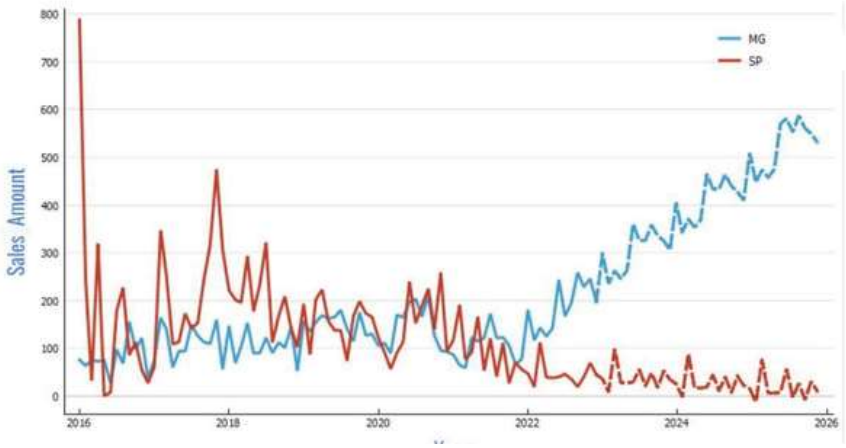
We isolated the states with the highest sales amount in a more detailed investigation. So, the states of MG and SP were separated to understand sales behavior using the historical series, as shown in Table 5.

Revista científica - ISSN 1807-5908
Avaliado pelo sistema Double Blind Review
Disponivel em https://periodicos.uff.br/sbijournal

Table 5 Description of the main steps related to sales forecasting.

Nº	Stage	Process	Description
1	Business Understanding	Sales	-
2	Data Acquisition	Load Data	-
3.1	Data Preprocessing	Select Columns	Manual selection of data attributes and composition of the data domain. The chosen columns were INVOICE ISSUE DATE, QUANTITY and STATE.
3.2	Data Preprocessing	Grouping	Groups data by selected variables and aggregates columns with selected aggregations. Grouped by INVOICE ISSUE DATE and STATE, using the sum aggregation function for the QUANTITY attribute.
3.3	Data Preprocessing	Pivot table	Isolating the states as columns.
3.4	Data Preprocessing	Select Columns	The chosen columns were INVOICE ISSUE DATE and STATE (MG and SP).
3.5	Data Preprocessing	Moving average	Compute a mean aggregation over a sliding window, using months as the aggregate time period of the time series.
4	Model Building	Time Series Forecasting	Time series prediction using Vector Autoregression (VAR) Model. Using 36 forecasting steps. We evaluate the model with Bayesian Information Criterion (BIC) as index.
5	Visualization	Line Chart	Visualize the time series' sequence and progression.

We got the following result (Figure 5):



Revista científica - ISSN 1807-5908
Avaliado pelo sistema Double Blind Review
Disponivel em https://periodicos.uff.br/sbijournal

Sales with high quantities justified the behavior of a higher total sales volume associated with SP at the beginning of the historical series (which may represent, for example, repressed market demand). However, maintaining the current management strategies, the historical behavior regarding the sales volume of these two states suggests that the company will lose the consumer market in SP and increase sales in MG in the coming years.

5.2 Analysis of customer behavior

By analyzing customers' lifetime and retention rates, the company can better understand sales behavior over time and make strategic decisions to increase loyalty and purchase satisfaction. This analysis can assist in identifying additional upsell opportunities, personalizing offers, improving the customer experience, reducing costs and increasing revenue. With a deeper understanding of customer behavior from lifetime and retention rate, the company can improve its marketing strategy, create effective loyalty programs and establish lasting customer relationships.

5.2.1 Lifetime

The customer's lifetime represents the duration of the relationship between him and the company, that is, the period in which the customer remains active as a buyer. This metric was calculated by subtracting the last purchase from each customer's first purchase, as shown in Table 6.

Table 6 Description of the main steps related to generating the customer lifetime.

Nº	Stage	Process	Description
1	Business Understanding	Customer Behavior	-
2	Data Acquisition	Load Data	-
3	Data Preprocessing	Select Columns	Manual selection of data attributes and composition of the data domain. The chosen columns were INVOICE ISSUE DATE and CUSTOMER CODE.
4	Feature Engineering	Feature construction	Grouping by CUSTOMER CODE applying the maximum and minimum aggregation functions to create, respectively, the LAST ISSUE and FIRST ISSUE attributes, related to each customer. Creation of the CUSTOMER BINDING TIME (CBT) column, using the formula: CBT= (LAST ISSUE - FIRST ISSUE)/ 365
5	Visualization	Histogram	Bar chart showing the distribution of the number of customers by years of service.

SUSTAINABLE BUSINESS
INTERNATIONAL JOURNAL
Revista científica - ISSN 1807-5908
Avaliado pelo sistema Double Blind Review
Disponivel em https://periodicos.uff.br/sbijournal
PPGAd uff
Programa de Pós-Graduação em Administração - UFF

We thus obtain the result presented in Figure 6.

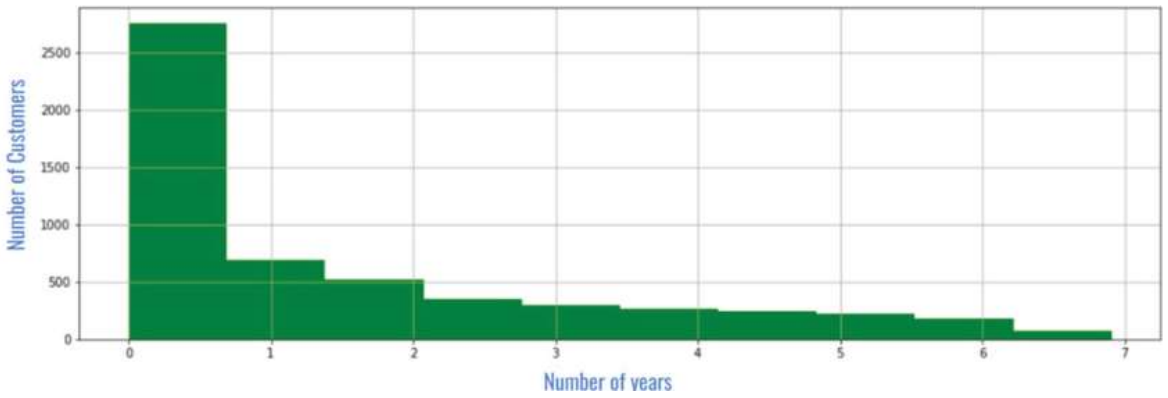


Fig. 6 Histogram showing the number of customers grouped by lifetime.

Most customers generally have a "purchase relationship" of less than one year.

We isolated the two states under study using the same previous DST (Table 6). Thus, when comparing MG and SP states, we found that the first has a longer-lasting customer loyalty than the second (Figure 7).

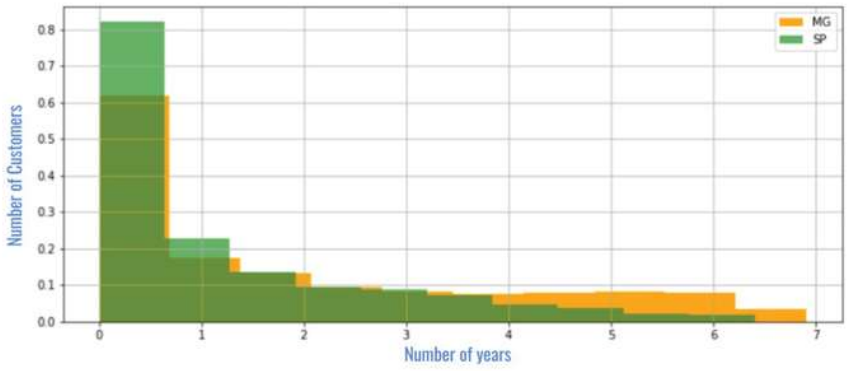


Fig. 7 Histogram showing the number of customers grouped by lifetime in the states of MG and SP.

5.2.2 Retention Fee

The retention rate measures customer loyalty, i.e., the percentage of customers who continue to buy from the company concerning the total number of customers. By analyzing this rate, businesses can gain

Revista científica - ISSN 1807-5908
Avaliado pelo sistema Double Blind Review
Disponivel em https://periodicos.uff.br/sbijournal

insight into customer retention and identify areas where they need to improve their services or products to retain customers. This process is shown in Table 7.

Finally, the sales volume time series (Figure 5) can be justified by comparing the Retention Rate of these two states (Figure 8). This metric was calculated considering the purchase in the 12-month window of each customer.

We noticed that in SP state, from 2020, it started to lose customer loyalty, while MG increased loyalty in this same year.

Table 7 Description of the main steps related to generating the customer retention rate.

N°	Stage	Process	Description
1	Business Understanding	Customer Behavior	-
2	Data Acquisition	Load Data	-
3.1	Data Preprocessing	Select Rows	Filter by states under analysis: SP and MG. Also, maintaining a column with the
3.2	Data Preprocessing	Select Columns	Manual selection of data attributes and composition of the data domain. The chosen columns were INVOICE ISSUE DATE, CUSTOMER CODE and STATE
4	Feature Engineering	Feature construction	Creation of columns RETENTION RATE GENERAL (without filters), RETENTION RATE MG (filtered by MG state) and RETENTION RATE SP (filtered by SP state), indexed by the date. The values were returned by a function designed to calculate the churn rate over 12 months using a sliding window approach. Subsequently converting these values into retention rate. This function takes as parameters a DataFrame (containing INVOICE ISSUE DATE and CUSTOMER CODE) and the window size in months. Within this function, a loop is made that runs through a certain number of months of the period. For each iteration, the start and end dates of the time window are defined and the customers who abandoned during this window are identified. The churn rate is calculated by dividing the number of customers who abandoned by the total number of unique customers. Churn rate values and corresponding dates are stored in lists. Finally, a conversion to retention rate is made.
5	Visualization	Line Chart	Visualize the time series' sequence and progression for comparing the results.

SUSTAINABLE BUSINESS
INTERNATIONAL JOURNAL
Revista científica - ISSN 1807-5908
Avaliado pelo sistema Double Blind Review
Disponivel em https://periodicos.uff.br/sbijournal
PPGAd uff
Programa de Pós-Graduação em Administração - UFF

5.3 Targeted advertising

Customer segmentation and sales recommendation are important strategies for targeted advertising, as they allow the company to optimize its advertising efforts by reaching the right people with the right messages, increasing the effectiveness and efficiency of advertising.

5.3.1 Customer segmentation

Customer segmentation divides the company's target audience into smaller groups with similar characteristics and needs, allowing it to customize its marketing campaigns for each group. This makes it possible to create more relevant and personalized messages, increasing the effectiveness of targeted advertising. The segmentation DST can be seen in Table 8.

After the process, three groups of customers were highlighted (Figure 9):

- C1 - With high code payment method and low purchase price.
- C2 - With low code payment method and also low purchase value.
- C3 - With low code payment method and high purchase value.

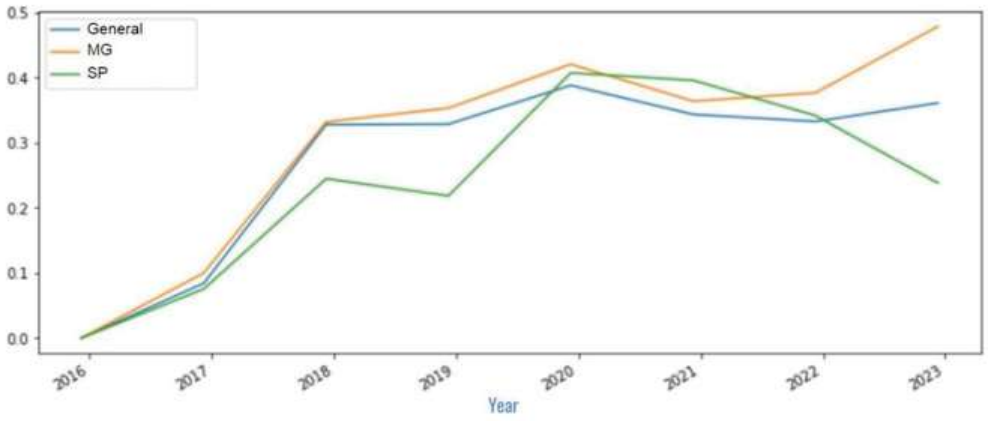


Fig. 8 Comparative line graph of the retention rate.

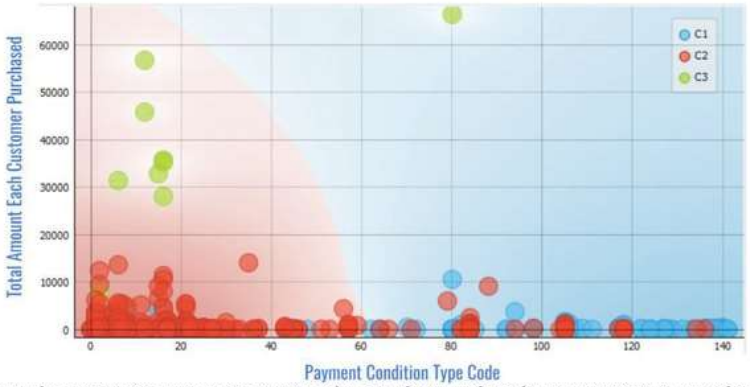


Fig. 9 Scatter chart representing customer segmentation. The most relevant attributes for representation in it were selected by the algorithms.

Revista científica - ISSN 1807-5908
Avaliado pelo sistema Double Blind Review
Disponivel em https://periodicos.uff.br/sbijournal

5.3 2 Sale Recommendation

The sales recommendation, in turn, uses algorithms and data analysis technologies to suggest specific products or services for each customer based on their purchase history and browsing behavior. This strategy increases marketing efficiency by offering personalized and targeted offers to each customer, increasing the chances of conversion. This process is presented in **9** and an excerpt from the generated decision tree can be seen in Figure 10.

Important results were extracted, such as the inference of buying product code 1 followed by buying products 1446 and 1506 in 8% of all occurrences.

Table 8 Description of the main steps related to customer segmentation.

N°	Stage	Process	Description
1	Business Understanding	Targeted Advertising	-
2	Data Acquisition	Load Data	-
3.1	Data Preprocessing	Select Columns	Manual selection of data attributes and composition of the data domain. The chosen columns were STATE, CUSTOMER CODE, PAYMENT CONDITION TYPE, SALE ORIGIN, ITEM CODE, QUANTITY, SITUATION TYPE
3.2	Data Preprocessing	Grouping	Grouped by CUSTOMER CODE, returning the aggregation of SALES COUNT (of purchases made by each customer), SALES SUM (of the total quantity of items purchased) and mode aggregation function for the attributes STATE, CUSTOMER CODE, PAYMENT CONDITION TYPE, SALE ORIGIN, ITEM CODE, SITUATION TYPE.
4	Feature Engineering	Correlation Analyzes	Compute all pairwise attribute correlations using the Pearson algorithm for multivariate analysis. It showed strong relationships between the attributes SALES SUM and PAYMENT CONDITION TYPE (relationship chosen for viewing in this article), SALES COUNT and SALES SUM, CUSTOMER CODE and PAYMENT CONDITION TYPE and PAYMENT CONDITION TYPE and ITEM CODE
5	Model Building	Clustering	Groups items using the k-Means clustering algorithm. It clusters the data and generates a new dataset with an added meta attribute representing the cluster label. Regarding the initialization method, the first center is randomly selected and subsequent centers are chosen from the remaining points with a probability that is proportional to the squared distance from the closest center. We enable silhouette scores functionality to calculate silhouette scores, which allows for automatic selection of the optimal number of clusters. The higher the silhouette score, the better the quality of clustering.
6	Visualization	Scatterplot	Scatter plot visualization with exploratory analysis for the attributes with the highest correlation factor.

Revista científica - ISSN 1807-5908
Avaliado pelo sistema Double Blind Review
Disponivel em https://periodicos.uff.br/sbijournal

Table 9 Description of the main steps related to client recommendation.

Nº	Stage	Process	Description
1	Business Understanding	Targeted Advertising	-
2	Data Acquisition	Load Data	-
3.1	Data Preprocessing	Transformation	Manual selection of data attributes and composition of the data domain. The chosen columns were ITEM CODE and CUSTOMER CODE. Type conversion of attributes from numerical to categorical to enable processing in the next steps.
3.2	Data Preprocessing	Reshaping Table	Reshape data table based on column values using One-hot encoding transaction data process. Transforming each item code into a column of the table and the lines being the customer codes.
4	Model Building	Association Rule Learning	Finds frequent itemsets in the data using Apriori Algorithm. The minimal support parameter was set to 3%.

6. Conclusions and Future Work

This research presents a Data Science Trajectories model to be applied in commercial sales transactions. We limit the problem to the furniture segment with real data from the ERP of a company in the pole of Uba´, Minas Gerais, Brazil.

From the modeling of the DST implemented with machine learning and data visualization techniques, we obtained the following result:

- Problem discovery: we discovered that July and September have the highest sales, which was proven by extracting the seasonal and trend components. Data exploration and the application of Time Series Forecasting with Vector Autoregression demonstrated that the company might lose the consumer market in SP and increase sales in MG in the coming years.
- Understanding the problem: Through a process based on the customer binding time metric and histogram visualization, we found that the loss of customers in the state of SP is associated with low customer loyalty. By calculating the retention rate, we found that SP lost loyalty from 2020 onwards.
- Proposition of interventions: We present a customer segmentation, with the k-Means clustering algorithm and visualization with scatter plot with exploratory analysis for the attributes with the highest correlation factor. With the association rule, we find frequent itemsets in the data using Apriori Algorithm for sales recommendation.

We emphasize that it was not the focus of this work to evaluate all the possibilities of data analysis about the case study company but rather to present the possibility of applying the DST to support the management strategy.

Moreover, we understand that an enterprise-level data-driven culture must be embraced for data analytics to provide a competitive advantage to an organization. This organization must consider that analytics must be seen and used as a strategic asset, management must support analytics throughout the enterprise and the insights must be available to those who need them.

Finally, as future work, we intend to:

- Propose a workflow supported by DST that can help the Data Science process implemented with Code Interpreter[28]. This powerful plugin provides ChatGPT with a working Python interpreter in a sandbox, enabling, for example, the AI to handle uploads and downloads, solve mathematical problems, perform data analysis and convert files between formats. In addition, it allows data scientists to perform complex programming tasks by simply asking a question in regular language, helping with prototyping and allowing them to optimize the process.
- Investigate the influence of fog-computing architecture on the data science process. This concept has been closely related to organizational data management[29][30]. For example, in a geographically distributed ERP, edge devices at each branch can collect locally generated data such as transactions, logs, performance metrics and more. These devices can perform initial data pre-processing, filtering, normalizing and aggregating relevant information before sending it to the central headquarters. This would reduce the amount of data to be transferred, save bandwidth and speed up the analysis process. This research will try to answer how this process would be mapped to a generic DST.

Acknowledgments

The authors would like to thank the Ubá-based technology company Tek-System, which, through an inter-institutional partnership with IFSudeste-MG and UFJF, made this research possible by providing data and technical support. Additionally, the authors would like to acknowledge the financial support from CNPq (313568/2023-5 and 310266/2021-1) and INESC P&D Brasil for funding this research.

References

McAfee, A., & Brynjolfsson, E. (2012). Big data: The management revolution. Harvard Business Review.

Duan, L., & Daxu, L. (2021). Data analytics in industry 4.0: A survey. Information Systems Frontiers, 1–17.

Xiong, X., Li, J., & Li, M. (2017). Current state and development trend of Chinese furniture industry. Journal of Wood Science, 63(5), 433–444.

Eybers, S., & Mayet, R. (2020). From data to insight: A case study on data analytics in the furniture manufacturing industry. In International Conference on Integrated Science (pp. 55–64). Springer.

Pioli, L., Dorneles, C. F., Macedo, D. D. J., & Dantas, M. A. R. (2022). An overview of data reduction solutions at the edge of IoT systems: A systematic mapping of the literature. Computing, 104(8), 1867–1889. https://doi.org/10.1007/s00607-022-01073-6

Suman, S., & Pogarcic, I. (2017). Development of ERP and other large business systems in the context of new trends and technologies. International Journal Vallis Aurea, 3(2), 79–92.

Schoenherr, T., & Speier-Pero, C. (2015). Data science, predictive analytics, and big data in supply chain management: Current state and future potential. Journal of Business Logistics, 36(1), 120–132.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). CRISP-DM 1.0 step-by-step data mining guide.

Martínez-Plumed, F., et al. (2019). CRISP-DM twenty years later: From data mining processes to data science trajectories. IEEE Transactions on Knowledge and Data Engineering, 33(8), 3048–3061.

Ko, S., et al. (2016). A survey on visual analysis approaches for financial data. Computer Graphics Forum, 599–617.

Pavlyshenko, B. M. (2019). Machine-learning models for sales time series forecasting. Data, 4(1), 15.

Zhang, G. P., & Iq, M. (2005). Neural network forecasting for seasonal and trend time series. European Journal of Operational Research, 160(2), 501–514.

Revista científica - ISSN 1807-5908
Avaliado pelo sistema Double Blind Review
Disponivel em https://periodicos.uff.br/sbijournal

Catal, C., et al. (2019). Benchmarking of regression algorithms and time series analysis techniques for sales forecasting. Balkan Journal of Electrical and Computer Engineering, 7(1), 20–26.

Singh, R., & Khan, I. A. (2012). An approach to increase customer retention and loyalty in B2C world. International Journal of Scientific and Research Publications, 2(6), 1–5.

Sabbeh, S. F. (2018). Machine-learning techniques for customer retention: A comparative study. International Journal of Advanced Computer Science and Applications, 9(2).

Mutanen, T. (2006). Customer churn analysis–A case study. Journal of Product and Brand Management, 14(1), 4–13.

Turkmen, B., et al. (2022). Customer segmentation with machine learning for online retail industry. The European Journal of Social & Behavioral Sciences.

Ziafat, H., & Shakeri, M. (2014). Using data mining techniques in customer segmentation. Journal of Engineering Research and Applications, 4(9), 70–79.

Tabianan, K., Velu, S., & Ravi, V. (2022). K-means clustering approach for intelligent customer segmentation using customer purchase behavior data. Sustainability, 14(12), 7243.

Jooa, J., Bangb, S., & Parka, G. (2016). Implementation of a recommendation system using association rules and collaborative filtering. Proceeding Computer Science, 91, 944–952.

Wanaskar, U., Vij, S., & Mukhopadhyay, D. (2013). A hybrid web recommendation system based on the improved association rule mining algorithm. arXiv preprint arXiv:1311.7204.

Cakir, O., & Aras, M. E. (2012). A recommendation engine by using association rules. Procedural Social and Behavioral Sciences, 62, 452–456.

Demsar, J., et al. (2013). Orange: Data mining toolbox in Python. The Journal of Machine Learning Research, 14(1), 2349–2353.

VanderPlas, J. (2016). Python data science handbook: Essential tools for working with data. O'Reilly Media, Inc.

Carneiro, T., et al. (2018). Performance analysis of Google Colaboratory as a tool for accelerating deep learning applications. IEEE Access, 6, 61677–61685.

Abimovel. (2006). Overview of the Furniture Sector. General information. Abmovable. www.abimovel.com.

Presidência da República. Secretaria Geral. Subchefia para Assuntos Jurídicos. (2018). Lei nº 13.709, de 14 de agosto de 2018. Lei Geral de Proteção de Dados Pessoais (LGPD).

OpenAI. (2023). ChatGPT: Code Interpreter Feature. https://openai.com/blog/chatgpt-plugins.

Dantas, M. A. R., Bogoni, P. E., & Filho, P. J. F. (2020). An application study case tradeoff between throughput and latency on fog-cloud cooperation. International Journal of Networking and Virtual Organisations, 23, 247–260.

Larcher, L., Stroele, V., & Dantas, M. (2021). A cloud-based system for distance learning supported by fog-cloud cooperation. International Journal of Grid and Utility Computing, 12, 618.

Submetido pelos autores em: 25/07/2024.
1a rodada de avaliação concluída em: 18/10/2024.
Aprovação em: 06/11/2024.