Board Game Cafe Analysed

Analysis on Sales and Games at Dhadhu Cafe

Rafael Agusto Prabawaseputra
Study Program in Information Engineering, Faculty of Computer Science
Universitas Dian Nuswantoro
Semarang, Indonesia
rafaelagusto9c34@gmail.com

Ardiawan Bagus Harisa
Department of Computer Science and Information Engineering
National Taiwan University of Science and Technology
Taipei, Taiwan
ardiawanbagusharisa@gmail.com

Indra Gamayanto
Study Program in Information System, Faculty of Computer Science
Universitas Dian Nuswantoro
Semarang, Indonesia
indra.gamayanto@dsn.dinus.ac.id

Abstract*— This study aims to enhance understanding of the factors influencing sales and game favoured by players at Dhadhu Cafe, a themed board game cafe. Key factors such as operational hours, promotional activities, game preferences, and menus are analysed using data visualization and multiple linear regression to inform strategic decision-making for the business. Despite their classification nature, decision trees and random forests are also applied in mining sales data, with random forest mitigating decision tree overfitting. Findings aligned between multiple linear regression and data visualization, revealing increasing sales on post-COVID-19. Sales peak on Saturdays, with the most effective sales hours observed from 3 to 8 p.m. (mode = 6 p.m.) daily. Promotions significantly impact sales, while other events have minimal effects. Drinks such as tea, yakult lychee, and matcha are dominating the sales as well as French fries. Light and party games, which typically last 15-45 minutes and accommodate 2-4, 2-6, and 2-12 players, are preferred for teaching players. These games often feature abstract and light themes, including mechanics such as dexterity, abstraction, tactics, and dice rolling.

***Keywords*— sales, board games cafe, analysis, multiple linear regression**

I. INTRODUCTION

Cafes have become increasingly popular gathering places, not only for relaxation but also for meetings and events [1]. In Semarang, Central Java, there were 115 registered cafes in 2019, 145 in 2020, and 169 in 2022 [2], indicating a continuous increase. With the rising attractiveness of cafes, their numbers are also growing. As competition among cafes intensifies, unique selling points are needed to attract customers. Amenities such as board games can enhance customer satisfaction and draw more visitors [3]. Cafes that offer board games as their main feature are known as board game cafes. Dhadhu board game cafe is one example among many in Indonesia, but the only one in Central Java, despite the presence of three known board game libraries around.

Promotions play a crucial role in boosting sales [4]. However, not all promotions are equally effective. Data analysis can identify which promotions yield better results. Events can also be significant attractions if executed well [5]. Like promotions, not all events are equally effective. Analysing data can help determine which events succeed for Dhadhu cafe. Various factors such as time, date, and day can influence cafe transactions [6]. Data mining facilitates the extraction of valuable insights from sales data, aiding in the identification of patterns and rules [7].

According to [1], there are three effective strategies for attracting customers: good service, unique products, and competitive pricing. From the simple random sampling research conducted by Santoso [6], it is stated that price and menu variety have little influence on revenue, but opening hours and the duration of business operations have a greater impact on revenue. A study on the application of data mining was conducted by Zou and Li for E-commerce using the K-nearest neighbours (KNN) algorithm, yielding an accuracy rate of 94.2%. The marketing recommendation from the experiment is then more effective, and reduces marketing costs [8]. Dhadhu Board Game Café was founded in Semarang with the objective of integrating research and business into a cohesive initiative. Furthermore, it provides a permanent space for a community committed to the development and advancement of board games.

This research aims to identify emergent patterns and factors influencing Dhadhu cafe's sales, as well as determine the types of board games preferred by its customers. This will enable Dhadhu cafe to take appropriate and effective steps to increase sales. In this study, we conducted data analysis at Dhadhu cafe and test the applicability of multiple linear regression, decision tree, and random forest methods to discover the key information that can be used to maximize revenue from Dhadhu cafe. While primarily designed for classification tasks, decision tree and random forest are also effectively utilized for mining sales data, with random forest being particularly adept at mitigating decision tree overfitting.

II. LITERATURE REVIEW

A. Board Game Cafe

A cafe is a place with a simple concept, usually serving drinks, light snacks [9], and tends to be fast-paced [10]. In the past, cafes were mainly favoured by young people, but now they are used by people of all ages, even as places for discussing work and business such as meetings [1]. This has led to the increasing development of cafes in Semarang city according to data [2]. To attract customers amidst competition, one way is through providing facilities [3]. One facility that can be provided in a cafe is board games, and a cafe that offers this facility as its main attraction is called a board game cafe [11]. Another way to attract customers is through marketing or promotion. Marketing or promotion is a way or action to disseminate information among the public [9], providing promotions, events are some forms of marketing. Cafes that rarely hold events will hinder their

marketing efforts [5]. Fig. 1 shows two players from Dhadhu gamer community compete in game titled *Catan*.

Fig. 1. Two players (customers) playing a *Catan* board game in Dhadhu Board Game Cafe.

B. Data Mining

In business, decision-making must be based on a solid foundation, as wrong decisions can result in losses [12]. Conversely, accurate decision-making can make marketing effective and maximize profits. Information that can aid in decision-making comes in the form of patterns or rules [8]. These patterns or rules can be discovered using data mining [7]. Data mining is part of Knowledge Discovery in Databases (KDD) used to identify the potential and utility of data to uncover understandable data patterns [13]. The first step in data mining is to select from the available data, followed by data cleansing to correct missing, erroneous, and duplicate data. Subsequently, data transformation can be performed to reshape the data for easier processing, followed by data mining to discover patterns or useful information. Finally, interpretation can be done to make the discovered patterns or information easily understandable [14].

C. Multiple Linear Regression (MLR)

Multiple linear regression (MLR) is a powerful statistical technique used to analyse the relationship between the dependent and independent variables [15]. It extends the concept of simple linear regression, which deals with only one independent variable, to accommodate multiple predictors. In MLR, the relationship between the dependent variable Y and multiple independent variables $X_1, X_2, \dots, X_p$ is represented by (1).

$$X = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon \quad (1)$$

Where Y is the dependent variable (the variable we want to predict or explain), $X_1, X_2, \dots, X_p$ are the independent variables (also known as predictors or explanatory variables). The intercept term β_0 representing the value of Y when all independent variables are zero, and $\beta_1, \beta_2, \dots, \beta_p$ are the coefficients (also known as regression coefficients or parameters), indicating the change in Y associated with a one-unit change in the corresponding independent variable, holding other variables constant. The error term ϵ representing the difference between the observed values of Y and the values predicted by the model. It captures the variability in Y that is not explained by the independent variables.

The goal of MLR is to estimate the regression coefficients $\beta_0, \beta_1, \beta_2, \dots, \beta_p$ that minimize the sum of squared differences between the observed values of the dependent variable and the values predicted by the model. This is typically done using the method of least squares. MLR provides valuable insights into the relationships between multiple variables and allows us to make predictions or understand the impact of changes in the independent variables on the dependent variable. It is widely used in various fields such as economics, finance, social sciences, and engineering for predictive modelling, hypothesis testing, and understanding complex relationships among variables.

D. Decision Tree

Decision tree (DT) is a visual model used to predict and build classification and regression models in the form of a tree [16]. Although decision trees can be used for regression, this method is more commonly used for classification. The Gini Index can be used to find suitable locations for features in the tree [16], [17]. The formula for the Gini Index is shown in (2).

$$G = 1 - \sum_{k=1}^{c} (P_k)^2 \quad (2)$$

Where G is the Gini Index, c represents the number of classes or categories., and P_k represents the proportion of instances in class k at a specific node.

The Gini Index measures the impurity or uncertainty of a node in a decision tree. A lower Gini Index indicates a more homogeneous node, meaning that it is purer in terms of class distribution, while a higher Gini Index indicates a more mixed or impure node. Decision tree algorithms use the Gini Index to determine the best split points for features, aiming to minimize impurity and maximize information gain at each node. This helps in creating a tree structure that efficiently classifies or predicts the target variable.

E. Random Forest

Random forest (RF) is one of the ensemble algorithms performed by combining multiple predictions from decision trees, whose subsets are obtained randomly [18]. For additional coefficients used in this study, R-squared (R^2) and Mean Absolute Error (MAE) are employed. Mean Absolute Error $MAE = \frac{\sum_{i=1}^{n} |y_i - \hat{y}_i|}{n}$ is the average absolute difference between the actual and predicted values [20].While R^2 is a coefficient indicating how much the dependent variable is explained by the independent variable [19] and can be formulated as presented in (3):

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} \quad (3)$$

Where n represents the number of observations, y_i represents the actual value of the dependent variable for observation i, $\hat{y}_i$ represents the predicted value of the dependent variable for observation i, and $\bar{y}$ represents the mean of the actual values of the dependent variable. The coefficients R^2 and MAE, provide valuable insights into the performance of the regression model, including how well the model fits the data and the accuracy of its predictions. They are commonly used to evaluate the effectiveness of regression models and to compare different models.

III. Methodology

This research aimed to conduct data mining to identify factors influencing sales and also analyse game data. Data mining is a part of Knowledge Discovery in Datasets used to

discover specific patterns or rules from the dataset of Dhadhu cafe. By conducting KDD on the sales data of Dhadhu cafe, factors influencing sales can be identified. This information can then be used by Dhadhu cafe as a reference to determine the next steps, especially for marketing purposes.

A. Research Instruments

In this study, we use sheet file as our initial dataset and Google Colab as processing tool. The hardware setup consists of a PC with the following specifications: an AMD Ryzen 5 2400G CPU and 8GB of RAM.

B. Data Collection Method

The primary data in this study refers to the promotions data of Dhadhu obtained directly by the author from the Instagram account of Dhadhu cafe between 2021 and 2023, totaling 19 promotions recorded. Meanwhile, the secondary data consists of a numeric time series dataset in Excel (spreadsheet) format containing information about the sales and rentals of Dhadhu cafe from other sources. The sales data consists of 25,506 entries from February 14, 2021, to December 31, 2023, while the game rental data consists of 2,966 entries from June 14, 2021, to June 14, 2022. There are 70 board game titles in 2020, and around 90 in 2023 in Dhadhu cafe. The sales and game rental data from Dhadhu cafe include the following information:

- **Sales Time**: Indicates when the purchase was made.
- **Ordered Items**: Items ordered by customers.
- **Promotions Held**: Includes promotions and events held by Dhadhu cafe, classified as Roll the dice, Playday, Tournaments, and discounts.
- **Point of Sale**: The location of the transaction, whether in-store or online.
- **Payment Method**: Methods used in payment, divided into Cash, Card, and E-Money.
- **Board Game Rental Time**: Indicates when the board game rental was made.
- **Title of Board Game Rented**: The title of the board game rented.
- **Board Game Specifications**: Specifications such as theme, number of players, and game mechanics of the board game rented.

C. Research Steps

The initial stage involves 1) data selection, where dataset features are analysed to determine which ones will be utilized in the research. Subsequently, the 2) preprocessing stage is divided into two parts: the first is conducted in Excel to eliminate irrelevant features and cleanse the data, while the second takes place in Google Colab after data transformation. 3) Data transformation encompasses the addition of significant features and the merging of event data. After transforming the data, we preprocess the sales data to create two new datasets: one for daily gross sales and another for monthly gross sales. Next, we use One Hot Encoding to convert categorical features into new ones that show whether a particular category is present or not in the data. Then, we 4) visualize the data to better understand its patterns.

Prediction data is segregated into training and test datasets. The daily training data only uses 10% of the total daily data because of its large volume. From the author's experiments, using too many daily training data can make the model overly complex. This occurs due to the presence of outliers and fluctuations in the data, such as changes in menu prices or the removal of menu items. Meanwhile, for monthly training data, it consists of 75% of the total monthly data. Following the retrieval of results from 4) data mining, analysis and 5) interpretation are performed to analyse the three methods and evaluate model performance.

As we mentioned earlier, the data mining is conducted on sales and game rents data utilizing Multiple linear regression. Although, Decision tree and Random Forest method are also applied. It is important to note that the author will compare the results of the three methods with information obtained from

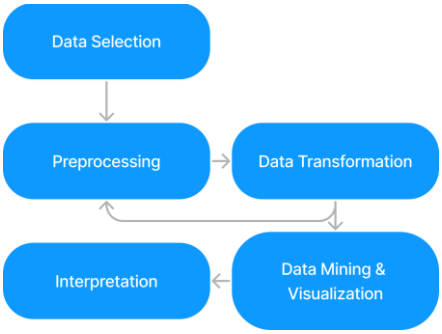


Fig. 2. Research steps used in this study.

data visualization rather than directly comparing the methods. This is because decision tree and random forest methods, while capable of regression, are more commonly used for classification. They do not produce coefficients in terms of currency but only provide feature importance rankings. The flowchart employed in this research is depicted in Fig. 2.

1) Data Selection

The research begins by selecting the data to be used. Sales data includes information such as order time, purchased products, payment method, and other details. Some features deemed irrelevant to sales, such as invoice numbers and outlet locations, are not used in the study. Board game rental data contains information on borrowing time, borrower details, and borrowed board game specifics. Certain features, such as borrower's personal data and information about the assigned game master, are not utilized in the study due to their lack of relevance and the privacy issue. The number of actual players is also omitted as it may not always align with the required number of players for some board games. Table 1 and 2 show the sales data and board game rent data.

2) Data Preprocessing

After determining the data to be used, the removal of features irrelevant to sales is conducted. Data with empty or anomalous values, such as those lacking item details or payment methods, are eliminated. Takeaway charge is also excluded as it only appears in the initial months of sales data. Additionally, data with empty payment methods are replaced with cash payment. Board game rental data is sorted based on borrowing time. Preprocessing steps continue in Google Colab, including data transformation and the separation of features with more than one entry, such as game mechanics and titles. Data visualization is performed before returning to preprocessing. The sales data is divided into daily and monthly. Certain months are removed from the monthly dataset due to incompleteness of daily transaction records. After cleaning up, the combined dataset is used for data mining. One Hot Encoder is employed to create separate features for each entry. This allows for the examination of the

weight of each feature after the data mining process is completed. Refer to Table 3 and 4 for data after splits.

TABLE I. THE ORIGINAL SALES DATA

Invoice ID	Order time	Payment time	Outlet (location)	Cashier name	Product	Sales (transaction	Payment method
IU901A0A	2021-09-01; 13:35	2021-09-01; 13:35	Semarang	Faiz	Salmon mentai, Yakult lychee	Rp. 50.000	Bank transfer
IU901AAA	2021-09-01; 15:35	2021-09-01; 17:00	Semarang	Faiz	Hazelnut choco, Tea	Rp. 45.000	Cash
…	…	…	…	…	…	…	…

TABLE II. THE ORIGINAL BOARD GAME RENT DATA

Timestamp	Email/ contact	Name/ID	Game	Categories	Mechanics	Duration	Players & actual players	Theme
2021-09-01; 15:35	…	…	Santorini	Light game	Abstract, puzzle	15-30 min	2-4; 4	Abstract, greek
…	…	…	…	…	…	…	…	

TABLE III. SALES DATA AFTER SPLIT

Date	Month	Year	Week	Day	Hour	Items	POS	Payment method	Sales
15	2	2021	2	1	20	Banana float	1	1	25000
31	12	2023	1	7	19	Black avocado	1	2	29000
…	…	…	…	…	…	…	…	…	…

TABLE IV. BOARD GAME RENTAL DATA AFTER SPLIT

Time	Game title	Category	Mechanics	Duration	Player numbers	Theme
2021-09-01; 15:35	Quoridor	Light	Abstract, tactic	15	2-4	Abstract
…	…	…	…	…	…	…

TABLE V. TRANSFORMED SALES DATA.

No	Date	Month	Year	Week	Day	Hour	Item	POS	Payment method	Sales	Type	Event
1	15	2	2021	2	1	20	Banana float	1	1	25000	3	1
2	31	12	2023	1	7	19	Black avocado	1	2	29000	3	1
…	…	…	…	…	…	…	…	…	…	…	…	…

3) Data Transformation

The first step in this stage is to add transaction numbers to each month in the sales data. Timestamps are also separated into hours, dates, months, and years of purchase. Day and week features are added to provide sales time details. The order type indicates transaction locations, with code 1 for in-store transactions and 2 for online (delivery) transactions. Payment methods are categorized into 3 categories: cash (1), card (2), transfer (3), and voucher (4). The Type feature indicates the type of item ordered, divided into main course (1), side dish (2), and drinks (3). The Event feature indicates the events of the day, divided into several categories: (1) indicating no event, (2) Roll the Dice, (3) Playday, (4) Tournaments, (5) Workshops, and (6) Promotions. After this transformation, the sales data will be more structured, while board game data remains same, see Table 5. The next step is to upload files and perform data transformation in Google Colab. In the sales data, daily and monthly gross features are added as dependent variables. Additionally, the game borrowing time feature is also converted to datetime format.

Roll the Dice (2) is an attractive mini-event designed to engage customers, allowing them to roll a dice and receive discounts or promotional menu items based on the outcome. Playday (3) offers a friendly playtime experience, facilitating interactions among players who were previously unknown to each other, fostering new friendships on the spot. Tournaments (4) provide an opportunity for customers to participate in competitive activities and earn rewards. Workshops (5) cover thematic topics of interest, encouraging participation and learning, while promotions (6) offer straightforward discounts or free menu to customers.

4) Data Mining & Visualization

In this stage, data visualization and data mining itself are performed. After the preprocessing and data split stages, data visualization is conducted to comprehend the patterns and trends present in sales. The visualization of daily and monthly gross over 3 years, spanning from February 2021 to December 2023 are presented. The results of data mining and produced graphs are presented on the following Results section.

5) Intepretation

In this step, we analyse the results or patterns discovered through data mining techniques to gain insights and understanding. This step is crucial for making informed decisions and taking actionable steps based on the mined data. Overall, interpretation is a critical step as it bridges the gap between raw data and actionable insights, enabling organizations to make informed decisions and drive meaningful outcomes. We analysed various factors influencing daily and monthly gross sales, including the impact of specific days, months, promotional events, and games, as detailed in the results section.

IV. Results and Discussion

A. Data Mining & Visualization

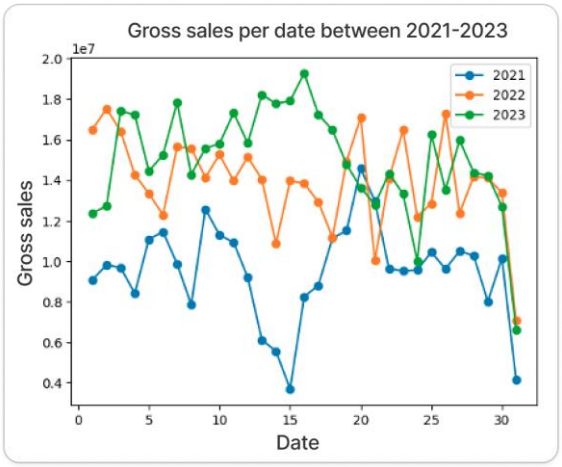


Fig. 3. The comparison of monthly gross sales between 2021 to 2023.

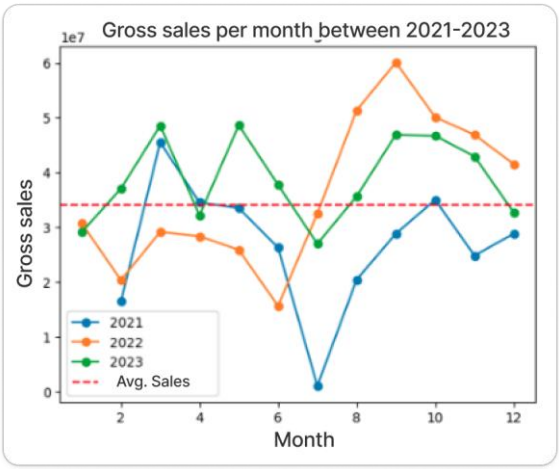


Fig. 4. The comparison of daily gross sales between 2021 to 2023.

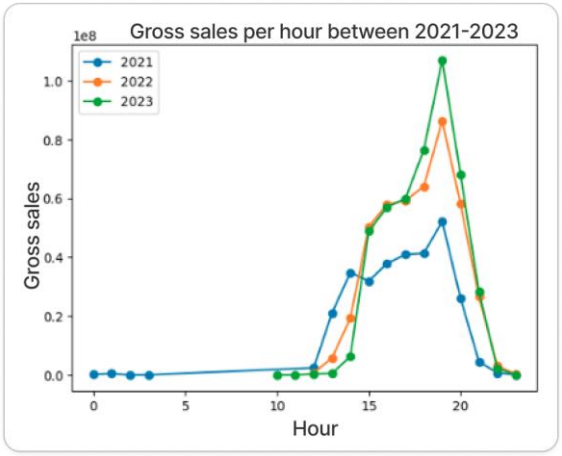


Fig. 5. The distribution of hourly gross sales between 2021 to 2023.

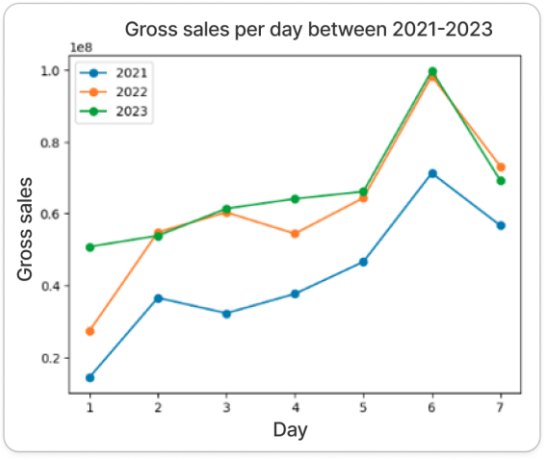


Fig. 6. The comparison of daily gross between 2021 to 2023.

As mentioned on the previous section, the data visualization and data mining are conducted to uncover the patterns and trends present in sales. Graph comparisons of daily and monthly gross over 3 years (February 2021 to December 2023) are generated. The graphical representation of daily and monthly gross sales patterns over 3 years is depicted in Fig. 3 and 4. The sales graph of Dhadhu cafe can be considered relatively stable in 2023, with declines in months with extended holidays. Some months were still disrupted by COVID-19, such as July 2021 to the beginning of 2022. It regained its customers from the second half of 2022 (post COVID-19 impact).

The distribution of gross sales by hour can be observed in Fig. 5, indicating that effective sales occur between 15:00 (3 p.m.) and 20:00 (8 p.m.), with the mode = 18:00 (6 p.m.). Additionally, according to Fig. 6, it is evident that Dhadhu cafe sales has fairly stable growths across 3 years, with the peak sales on Saturdays.

Meanwhile, the graph depicting the types of items purchased can be seen in Fig. 6, where 1 indicates Monday, and 7 indicates Sunday. From Fig. 7, it can be seen that 65% of items purchased in 2021 were drinks (Type 3). According to our data, this trend persists in 2022 and 2023, with the percentage of drink purchases exceeding 65%. This underscores the significance of drinks as the backbone of a cafe like Dhadhu. In Fig. 8, Tea, Zombie Matcha, Yakult Lychee, and Meeple Fries are the preferred menu items. This trend remains consistent in 2022 and 2023 as well. Other than customer being familiar with those items, the price of the top items sold are affordable. It might be one of the reason the items are preferred by customers.

In Fig. 9, it can be observed that the most frequently used payment method in 2021 is Cash. The most commonly used payment method in 2022 remains the same, but there is a change in 2023. Upon examining Fig. 10, the most frequently used payment method shifts from Cash to Card.

According to Dhadhu Cafe, of all sales transactions at Dhadhu cafe, over 99% occur in-store, and less than 1% is online. As we expected, interviews with customers indicated that they visit Dhadhu cafe, or similar places, specifically for its unique selling point: board games. In other words, it's the board games bundled with the menu that attract customers the most, not just the menu items. This suggests that themed-cafe are more appealing to customers compared to regular ones.

According to Fig. 11 and 12, abstract and light-themed board games like *Rhino Hero*, *Dixit*, *Quoridor* and *Santorini* were popular in 2020. In 2021 and 2022, *Quridor* and *Santorini* take the first and second places, and these board games continued to be the most played by customers. The specifications of the board games have also remained largely unchanged from year to year. That is also a reason we only show charts from 2020. Those games applied mechanics such as dexterity, abstract, tactic, and rolling dice, please refer to Fig. 13. While board games with a duration of 15-45 minutes and player counts of 2-4, 2-6, and 2-12 are the most frequently played as shown by Fig. 14 and 15. Fig. 16 shows that the abstract theme is the most played theme. This observation may contain bias, because during interviews with the game master staff, it was revealed that they tend to introduce new players to the easiest-to-understand games, which happen to be those particular ones. However, now we understand which games the game master prefers to teach to new players given the limited time and staff availability.

Once data are analysed through visualization, proceeds to the data mining stage. The results of data mining for the daily dataset are presented in Table 6. From the table, we can conclude that MLR remains the preferred method for analysing this type of data, as it exhibits the lowest error (8.5%). It's important to note that the currency used is in Indonesian Rupiah (IDR).

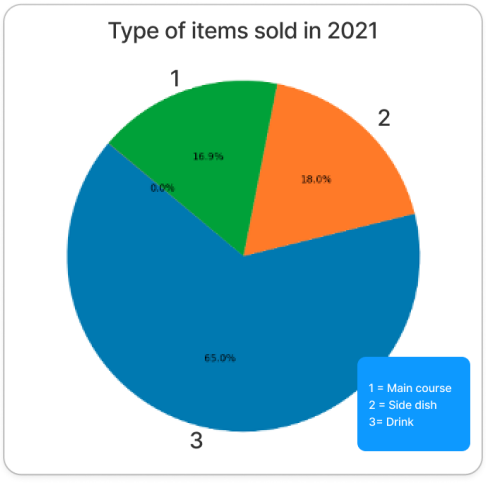


Fig. 7. Type of items sold in 2021.

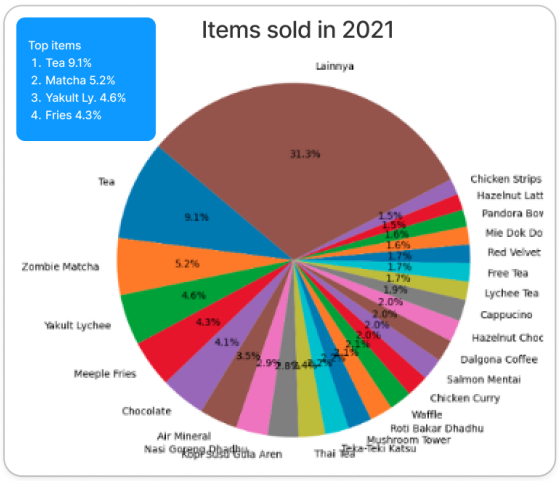


Fig. 8. Pie Chart of menu items in 2021.

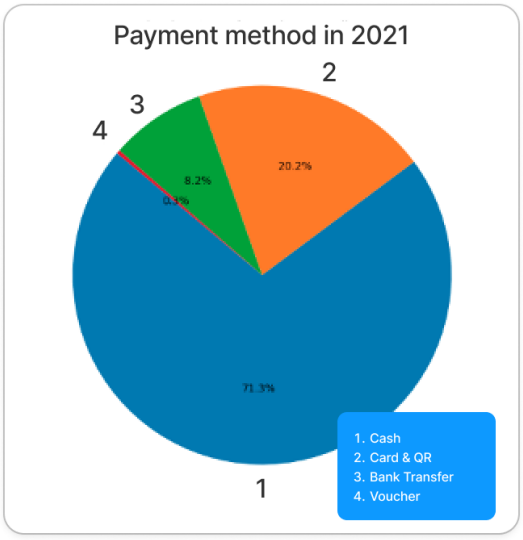


Fig. 9. Payment methods in 2021.

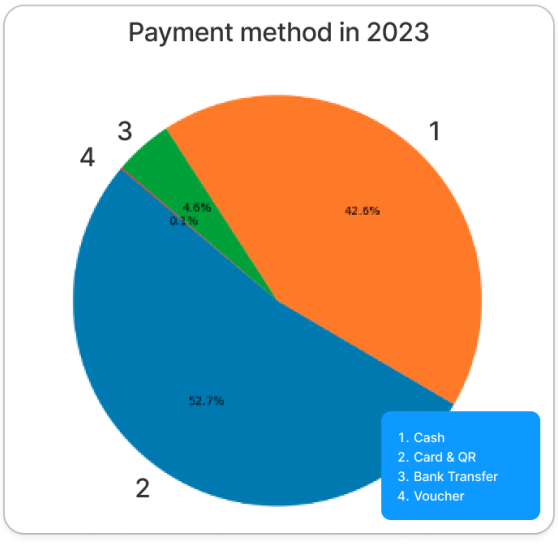


Fig. 10. Payment methods in 2023.

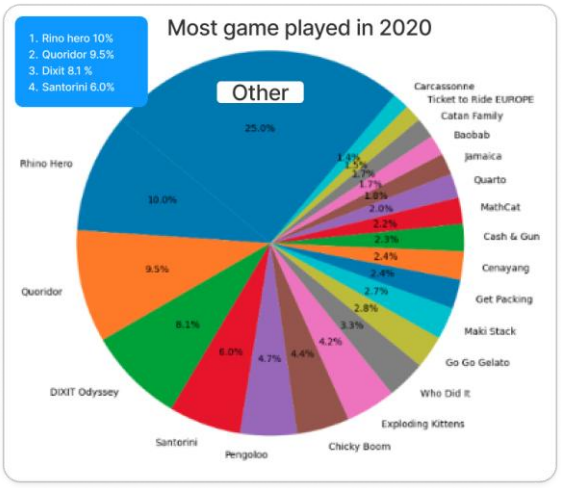


Fig. 11. The most played games in Dhadhu cafe in 2020.

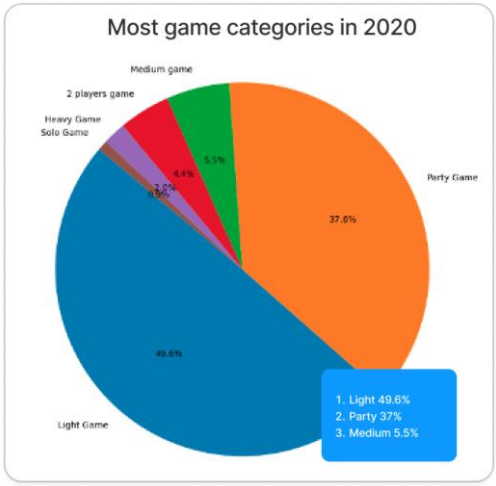


Fig. 12. The most played game categories in Dhadhu cafe in 2020.

B. Intepretation

The results indicate that the multiple linear regression model on daily data has a coefficient of determination R^2 of 94.4% (Table 6), implying that a significant portion of the variation in the data can be explained by the model. The mean absolute error (MAE) value is 109,187 Indonesian Rupiah (IDR), equivalent to 8.5% of the daily gross average (1,284,544). While the confidence interval values reach 95%. Based on the results obtained from multiple linear regression on daily data, the Halloween Event promotion weighs 5,450 Rupiah, meaning it has positive impact on sales. Meanwhile,

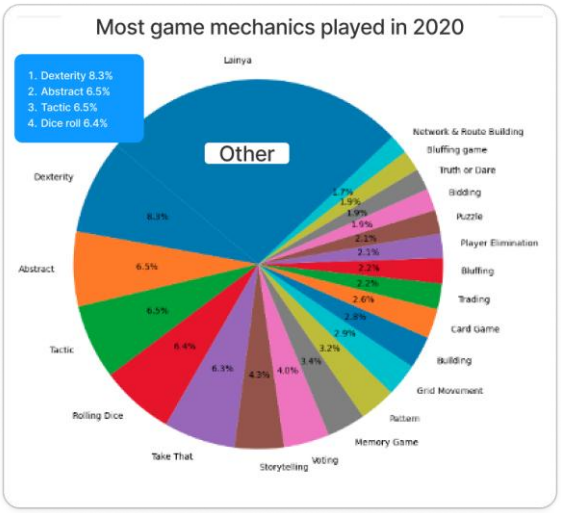


Fig. 13. The most played game mechanics in Dhadhu cafe in 2020.

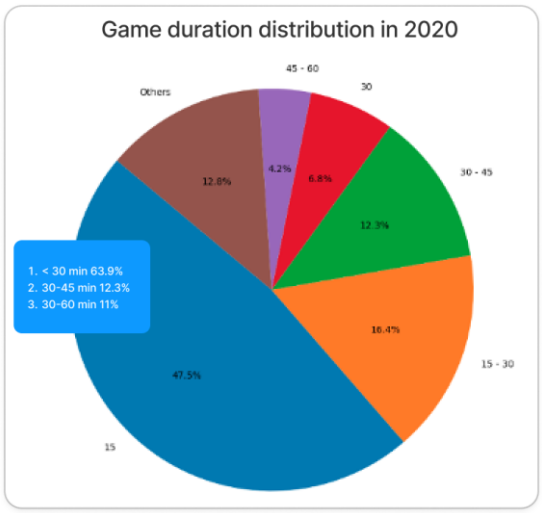


Fig. 14. The distribution of gameplay duration in Dhadhu cafe in 2020.

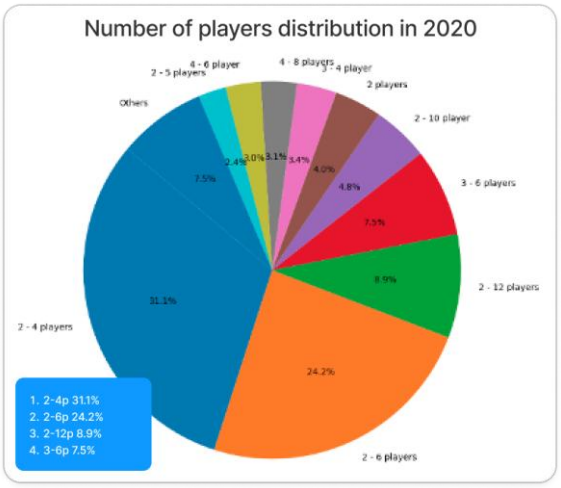


Fig. 15. The most common player numbers in Dhadhu cafe in 2020.

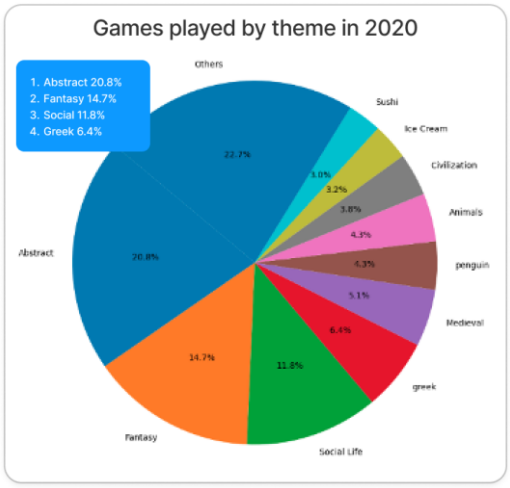


Fig. 16. The most played game themes in Dhadhu cafe in 2020.

all Ramadhan promotions have a weight of 37,142 Rupiah, though with different confidence interval ranges. However, other package promotions do not affect daily gross. The hours that positively impact daily gross are 14:00-22:00, except at 17:00. The days that positively impact daily gross are Tuesday, Thursday, and Saturday. Meanwhile, the months that positively influence daily gross are March, April, May, July, October, and November. Table 7 presents some positive-impact features on the daily sales.

TABLE VI. DAILY GROSS AND MONTHLY GROSS PREDICTION RESULT

Data	**Daily**			**Monthly**		
Methods	MLR	DT	RF	MLR	DT	RF
MAE	109,187.14	227,201.53	142,797.89	1,457,131.41	7,578,426.25	4,363,835.06
Error %	8.5%	17.68 %	11.11 %	4.01%	20.90%	12.03%
R^2	94.42 %	73.79 %	89.47 %	97.75%	46.95%	77.56%

TABLE VII. SOME NOTABLE POSITIVE-IMPACT FEATURES ON DAILY SALES (HOUR, DAY, MONTH)

Feature	**Values**
Hour_18	10025.37, 95% CI: [-4812.57, 15943.54]
Hour_20	10650.92, 95% CI: [-3803.77, 11856.13]
Hour_22	13211.45, 95% CI: [-3113.10, 32201.80]
Day_2	24612.98, 95% CI: [-8363.25, 20763.28]
Day_4	21762.44, 95% CI: [-4923.75, 20031.65]
Day_6	1911.06, 95% CI: [-10931.99, 5786.72]
Month_5	22925.16, 95% CI: [-9075.42, 18556.84]
Month_10	5253.51, 95% CI: [-12757.72, 9580.47]
Month_11	35218.56, 95% CI: [-1866.01, 24276.10]

TABLE VIII. SOME NOTABLE POSITIVE FEATURES ON MONTHLY SALES (EVENT, WEEK, HOUR, MONTH, DAY).

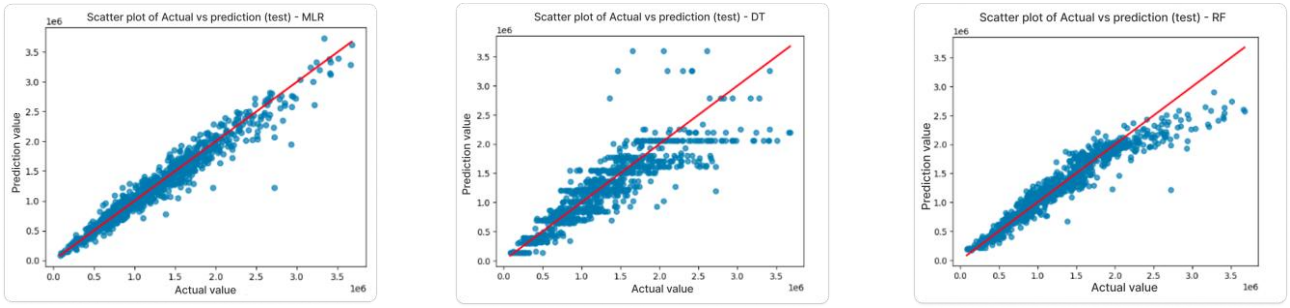
Feature	**Values**
Event_1 (No event)	1403.29, 95% CI: [-1463.13, 5871.12]
Event_4 (Tournament)	3130.19, 95% CI: [-6011.83, 6743.04]
Event_6 (Promotions)	5595.08, 95% CI: [461.50, 8523.34]
Week_2	3283.08, 95% CI: [-1507.73, 12577.62]
Week_3	5228.35, 95% CI: [-6510.03, 7649.83]
Hour_18	3613.04, 95% CI: [-1055.31, 7193.24]
Hour_20	3144.21, 95% CI: [-1271.72, 7333.88]
Hour_21	6573.27, 95% CI: [3282.37, 11008.30]
Month_9	61.33, 95% CI: [-71.88, 110.24]
Month_11	48.92, 95% CI: [-54.00, 83.38]
Month_12	35.91, 95% CI: [-189.32, 95.71]

The results from analysing monthly data through multiplelinear regression provide valuable insights, explaining a significant chunk—97.75%—of the test data and averaging an error of 1,457,131 Rupiah (4.01%). Despite

having fewer data compared to the daily dataset, our analysis suggests that the predictions in scatterplots closely match the actual values. Unlike the daily data, we found that the Ramadan package positively impacts monthly gross. Events like promotions and tournaments also play significant roles, despite the varied popularity of different playday events hosted by Dhadhu cafe. Drinks and side dishes strongly influence monthly gross, align with the visualized sales data Additionally, our examination time patterns showed that positive coefficients start appearing in the late afternoon 5 p.m. – 10 p.m., peaking in the evening but dipping slightly around 7 p.m.. Days like Monday, Thursday, Saturday, and Sunday tend to show positive impacts on sales. While Saturday and Sunday have similar coefficients, the confidence interval's upper bound for Saturday is notably higher, suggesting that Saturday's influence might be greater.

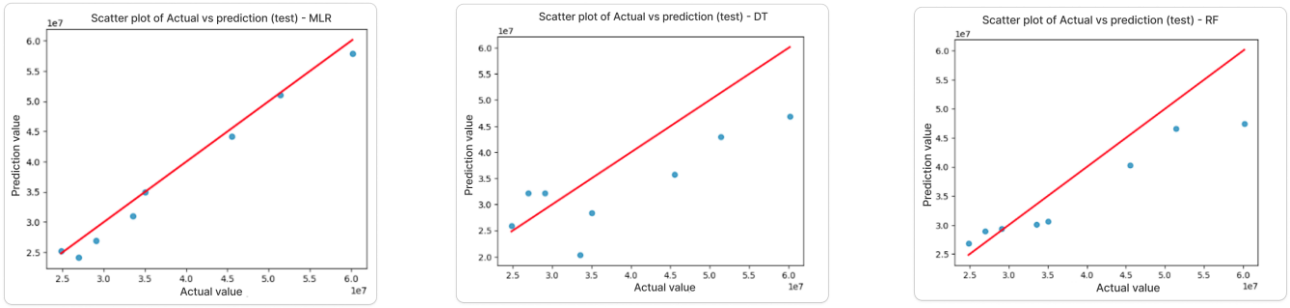
stronger performance compared to previous years. It reflects a consistent upward trend in sales performance over the years. Specifically, in 2021, we had 3 months above the average, in 2022 we had 5, and in 2023, we observe 8 months exceeding the average. This indicates a clear path of growth and improvement in sales over time. Refer to Table 8 for notable features on monthly sales.

We also want to test the usage of decision tree in analysing Dhadhu's sales data, applied to daily data showed an R^2 of 73% on the test set, with a 17% error rate. As we expected, this contrasts with regression, which typically shows errors below 10% and R^2 values exceeding 90%. Suggesting decision trees may not be the best fit for regression analysis of Dhadhu's sales. The features significant by the decision tree for daily data are outlined in Table 9, with importance scores



(a) (b) (c)

Fig. 17. Scatterplot of daily gross sales prediction vs actual between methods-multiple linear regression (a), decision tree (b), random forest (c).



(a) (b) (c)

Fig. 18. Scatterplot of monthly gross sales prediction vs actual between methods-multiple linear regression (a), decision tree (b), random forest (c).

This observation aligns with graphical representations. When examining the coefficients for dates, a trend emerges: sales typically start strong at the beginning of the month, experience a slight decline around the middle, then bounce back to positivity. However, during the mid-month period, the confidence intervals expand considerably, indicating sales instability. This trend is especially pronounced in 2021 and 2022. In contrast, 2023 shows an upward trend in sales during this period. As a result of this variability, the confidence intervals widen further toward the end of the month, leading to many coefficients turning negative. Despite fluctuations, certain months consistently contribute positively to monthly gross. September, November, and December stand out, with September showing the highest value, reflecting an upward trend observed in the graphs. Notably, in 2023, the lower bound has the smallest value, while the upper bound is highest, hinting at a potentially stronger performance compared to the previous years. This suggests a possibly

ranging from 0 to 1, derived from the reduction in Gini impurity for each feature divided by the total reduction. However, due to the multitude of features in both daily and monthly sales data, the importance values are quite small, indicating that this method may be less suitable for numerical data. Additionally, scatter plots generated by the decision tree method show greater point dispersion compared to multiple linear regression, implying less precise predictions (see Fig. 17 and 18). The results from the decision tree applied to monthly data were even less favorable, with an R^2 of only 46.9% on the test set, despite an 80% R^2 on the training data. However, the average error rate increased to 20%, indicating potential overfitting. This discrepancy between high R^2 on the training data and low R^2 on the test data suggests that the model may not generalize well to unseen data. Consequently, the decision tree model may be less effective when extrapolated beyond the dataset. Key features identified by the decision tree are detailed in Table 10.

TABLE IX. MOST IMPORTANT FEATURES IN DECISION TREE ON DAILY DATA

Feature	Importance
Type 3 (Drink)	0.74
Item_Chocolate	0.15
Type 1 (Main Course)	0.04
Type 2 (Side Dish)	0.015
Item_Lychee Tea	0.007

TABLE X. MOST IMPORTANT FEATURES IN DECISION TREE ON MONTHLY DATA

Feature	Importance
Day_4 (Thursday)	0.75
Item_Mineral water	0.13
Week_2	0.05
Item_Kopi Susu Gula Aren	0.025
POS_1 (in-store)	0.007

TABLE XI. IMPORTANT FEATURES IN RANDOM FOREST ON DAILY SALES DATA

Feature	Importance
Type_ 3 (Drinks)	0.68
POS_1 (In-store)	0.082
Type_1 (Main course)	0.055
Type_2 (side dish)	0.033
Item_Red Velvet	0.022
Item_Chocolate	0.016
Item_Mineral water	0.013
Item_Mie Dok Dok	0.006
Payment_Method_1(Cash)	0.0049
Item_Meeple Fries	0.0043

TABLE XII. IMPORTANT FEATURES IN RANDOM FOREST ON MONTHLY SALES DATA

Feature	Importance
Hour_18	0.093
***Event*_1 (no event)**	0.080
POS_1 (in-store)	0.078
Tipe_3 (drinks)	0.072
Day 1	0.062
Item_Americano	0.054
Hour 20	0.044
Day_4	0.039
Day_2	0.037
Item_Mineral water	0.031

Random forest, a combination of multiple decision trees, tackles issues like overfitting and handling large feature sets. When applied to daily sales data, it achieves an R^2 of 90% on training data and 89% on test data, a significant improvement over the decision tree's 73% R^2. The error rate also decreased to 11%. Important features for daily data, as identified by random forest, are listed in Table 11. The analysis of feature importance shows a more balanced distribution compared to decision trees, with Type 3 (beverages) and POS 1 (in-store) emerging as crucial features, similarly identified in the other methods.

Furthermore, the scatter plot produced by random forest demonstrates improved clustering around actual values, indicating enhanced predictive accuracy compared to the decision tree. The R^2 on test data rises from 40% to 70%, and the error rate decreases to 12%. Important features for monthly data, according to random forest analysis, are listed in Table 12. Several important features identified also exhibit high values in multiple linear regression, indicating that random forest is better suited for such tasks than the decision tree.

Interestingly, mineral water emerges as a key feature across all three methods. Surprisingly, tea (a drink) is not considered a top feature in all methods, despite its significantly high sales as observed in data visualization. Upon investigation with management, it was revealed that the business often provides complimentary tea to customers during promotional activities or events. This could introduce a bias that disrupts the mining process.

Moreover, it is not appropriate to compare numerical data analysis between decision trees, random forest, and regression. Based on the results obtained, multiple linear regression with monthly dataset visualization appears to be the most suitable data mining approach. Despite minor differences between daily and monthly data, the results underscore their impact on the final outcome.

V. CONCLUSION

In this study, we have demonstrated the use of data visualization and data mining methods to analyse board game cafe sales and game preferences. Multiple linear regression is the most suitable method for the numerical dataset we used in this study; however, we also applied decision tree and random forest to test the key features as findings. The findings might be beneficial for cafe management (not limited to Dhadhu cafe) to support strategic moves aimed at enhancing sales and brand.

The research findings in terms of the game data analysis indicate that board games with a play duration of 15-45 minutes, accommodating 2-4, 2-6, and 2-12 players, are most preferred. Game masters prefer teaching lighter, party-style games due to their shorter playtime and ease of understanding. Thus, players gravitate towards games with mechanics like dexterity, abstract elements, tactics, and rolling dice. These games often feature abstract and light themes, such as titles like *Quoridor* and *Santorini*.

Sales data analysis reveals a consistent uptrend in Dhadhu cafe's sales, with monthly stability despite slight declines during extended holidays, particularly post-COVID-19. Peak sales occur on Saturdays, with the most effective sales hours between 3 p.m. and 8 p.m. (mode = 6 p.m.). Most transactions take place in-store, with cash and card being the preferred payment methods. Drink purchases dominate, accounting for 65% of total sales, while packaged deals are less favored. Top-selling items include tea, fries, and matcha. As observed, months such as April, September, and November yield highly positive profits.

Obviously, for our dataset, Multiple linear regression emerges as the most suitable method, compared to Decision trees and random forests which are suitable for classification tasks. In such case, decision trees often face overfitting issues, especially with monthly data and small sample sizes. Regression analysis highlights the significant influence of main courses and drinks on daily sales, with packages having a larger impact on overall sales. Events have a minor effect

on revenue. Monthly regressions indicate that drinks and side dishes have the most influence on sales, while packages have a lesser impact. These findings align well with visualization data. Decision trees and random forests also identify important features such as drinks, main courses, side dishes, and days of the week. Random forests mitigate overfitting and perform better on test data compared to decision tree.

While we acknowledge that decision trees and random forests may not be the suitable methods for our case, we included them in the study to explore the boundaries of their application in this context. Despite being primarily used for classification tasks, decision trees and random forests can also be applied to sales data mining, with random forests offering a solution to decision tree overfitting. It's important to recognize that the limitations of our dataset could contribute to the less satisfactory outcomes observed with the random forest method. Further research needs to consider the complimentary items as gifts such as free drink (tea for instance; it is also often included in a bundle/ package) because it may affect the data mining process.

ACKNOWLEDGMENT

Thanks to Dhadhu Board Game Cafe and its game community, GAI Lab, and IntSys Lab.

REFERENCES

[1] R. Fauzan, R. Supryanita, and R. Rahmatika, 'ANALISA STRATEGI PEMASARAN UNTUK PENINGKATAN DAYA SAING PADA BISNIS KAFE DI KOTA BUKITTINGGI (STUDI KASUS KAFE TERAS KOTA)', *Jurnal Manajemen Bisnis Syariah*, vol. 1, no. 1, Art. no. 1, Apr. 2021, doi: 10.31958/mabis.v1i1.2835.

[2] Dinas Komunikasi, Informatika, Statistik dan Persandian Kota Semarang, 'Portal Semarang Satu Data'. Accessed: Apr. 28, 2024. [Online]. Available: https://data.semarangkota.go.id/data/list/4

[3] A. Ballang, A. K. Pongtuluran, and R. Tangdialla, 'Pengaruh Lokasi Dan Fasilitas Terhadap Kepuasan Pelanggan di Cafe Mie Setang Kecamatan Rantepao Kabupaten Toraja Utara', *MENAWAN : Jurnal Riset dan Publikasi Ilmu Ekonomi*, vol. 1, no. 6, pp. 113–125, Oct. 2023, doi: 10.61132/menawan.v1i6.74.

[4] A. A. Makhbuby and R. Himmati, 'PROMOTION MIX UNTUK MENINGKATKAN VOLUME PENJUALAN COFFE BERBASIS TAKE AWAY PADA CAFE PESEN KOPI KOTA BLITAR', *Jurnal Cakrawala Ilmiah*, vol. 1, no. 12, Art. no. 12, Aug. 2022, doi: 10.53625/jcijurnalcakrawalailmiah.v1i12.3216.

[5] M. A. Gama, 'STRATEGI KOMUNIKASI PEMASARAN KOPI TJANGKIR 13', *Jurnal Ilmu Sosial dan Ilmu Politik (JISIP)*, vol. 7, no. 2, 2018, doi: 10.33366/jisip.v7i2.1588.

[6] S. A. Santoso, 'Pengaruh Variasi Menu, Harga, Jam Kerja, Dan Lama Usaha Terhadap Pendapatan Warung Tegal Di Kecamatan Ciputat Timur', bachelorThesis, Fakultas Ekonomi dan Bisnis Uin Jakarta, 2019. Accessed: Apr. 28, 2024. [Online]. Available: https://repository.uinjkt.ac.id/dspace/handle/123456789/47938

[7] R. Ridwan, H. Lubis, and P. Kustanto, 'Implementasi Algoritma Neural Network dalam Memprediksi Tingkat Kelulusan Mahasiswa', *JURNAL MEDIA INFORMATIKA BUDIDARMA*, vol. 4, no. 2, Art. no. 2, Apr. 2020, doi: 10.30865/mib.v4i2.2035.

[8] J. Zou and H. Li, 'Precise Marketing of E-Commerce Products Based on KNN Algorithm', *Computational Intelligence and Neuroscience*, vol. 2022, pp. 1–12, Aug. 2022, doi: 10.1155/2022/4966439.

[9] Badan Pengembangan dan Pembinaan Bahasa, 'Hasil Pencarian - KBBI VI Daring'. Accessed: Apr. 28, 2024. [Online]. Available: https://kbbi.kemdikbud.go.id/entri/kafe

[10] Michael and A. Rahman, 'Kafe dan Gaya Hidup: Studi pada Pengunjung Kafe di Wilayah Barombong Kota Makassar', *Jurnal Multidisiplin Madani*, vol. 2, no. 10, Art. no. 10, Oct. 2022, doi: 10.55927/mudima.v2i10.1548.

[11] Isa, 'Tips Buka Usaha Board Game Cafe #3: Karyawan'. Accessed: Apr. 28, 2024. [Online]. Available: https://boardgame.id/?p=56265

[12] A. R. Riszky and M. Sadikin, 'Data Mining Menggunakan Algoritma Apriori untuk Rekomendasi Produk bagi Pelanggan', *Jurnal Teknologi dan Sistem Komputer*, vol. 7, no. 3, pp. 103–108, Jul. 2019, doi: 10.14710/jtsiskom.7.3.2019.103-108.

[13] T. Taslim and F. Fajrizal, 'Penerapan algorithma k-mean untuk clustering data obat pada puskesmas rumbai', *Digital Zone: Jurnal Teknologi Informasi dan Komunikasi*, vol. 7, no. 2, pp. 108–114, Nov. 2016, doi: 10.31849/digitalzone.v7i2.602.

[14] H. Kurniawan, S. Defit, and Sumijan, 'Data Mining Menggunakan Metode K-Means Clustering Untuk Menentukan Besaran Uang Kuliah Tunggal', *Journal of Applied Computer Science and Technology*, vol. 1, no. 2, Art. no. 2, Dec. 2020, doi: 10.52158/jacost.v1i2.102.

[15] - Ahyani Junia Karlina, 'Estimasi Hasil Panen Ayam Pedaging Menggunakan Algoritma Regresi Linear Berganda', *Estimasi Hasil Panen Ayam Pedaging Menggunakan Algoritma Regresi Linear Berganda*, vol. 3, no. 6, Art. no. 6, Jun. 2023.

[16] A. T. Nurani, A. Setiawan, and B. Susanto, 'Perbandingan Kinerja Regresi Decision Tree dan Regresi Linear Berganda untuk Prediksi BMI pada Dataset Asthma', *Jurnal Sains dan Edukasi Sains*, vol. 6, no. 1, Art. no. 1, May 2023, doi: 10.24246/juses.v6i1p34-43.

[17] H. M. Asmara and S. T. Fatah Yasin Al Irsyadi, 'Analisa Perbandingan Hasil Pohon Keputusan Dengan Gain Ratio, Information Gain, Dan Gini Index Pada Pemasaran Produk Herbal di CV. Al-Ghuroba', diploma, Universitas Muhammadiyah Surakarta, 2016. Accessed: Apr. 28, 2024. [Online]. Available: https://eprints.ums.ac.id/47859/

[18] S. J. Rigatti, 'Random Forest', *Journal of Insurance Medicine*, vol. 47, no. 1, pp. 31–39, Jan. 2017, doi: 10.17849/insm-47-01-31-39.1.

[19] E. Kasuya, 'On the use of r and r squared in correlation and regression', *Ecological Research*, vol. 34, no. 1, pp. 235–236, 2019, doi: 10.1111/1440-1703.1011.

[20] C. J. Willmott and K. Matsuura, 'Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance', *Climate Research*, vol. 30, no. 1, pp. 79–82, Dec. 2005, doi: 10.3354/cr030079.