Institut Riset dan Publikasi Indonesia (IRPI)
MALCOM: Indonesian Journal of Machine Learning and Computer Science
Journal Homepage: https://journal.irpi.or.id/index.php/malcom
Vol. 5 Iss. 3 July 2025, pp: 1000-1011
ISSN(P): 2797-2313 | ISSN(E): 2775-8575

Predictive Sales Analysis in Coffee Shops Using the Random Forest Algorithm

Shella Norma Windrasari[1*], Hendro Margono[2],
Yudistira Ardi Nugraha Setyawan Putra[3]

[1,3]Departemen of Human Resource Development, Faculty Graduate School,
Airlangga University, Indonesia
[2]Departemen of Information and Library Science, Faculty of Social and Political Sciences,
Airlangga University, Indonesia

E-Mail: [1] shella.norma.win-2024@pasca.unair.ac.id, [2]hendro.margono@fisip.unair.ac.id,
[3]yudistira.ardi.nugraha-2023@pasca.unair.ac.id

Received Apr 19th 2025; Revised Jun 22th 2025; Accepted Jul 22th 2025; Available Online Jul 31th 2025, Published Jul 31th 2025
Corresponding Author: Shella Norma Windrasari
Copyright © 2025 by Authors, Published by Institut Riset dan Publikasi Indonesia (IRPI)

Abstract

The coffee shop industry has experienced significant growth, evolving into a highly competitive marketplace demanding specialty coffee and personalized experiences. While data-driven strategies are crucial for optimizing operations, many owners still struggle to effectively leverage their sales data to understand dynamic customer behavior and enhance decision-making. Addressing this gap, this study explores the application of machine learning (ML) techniques, specifically the Random Forest Regressor model, to predict sales performance within the coffee shop business environment. By analyzing factors such as transaction timing, store location, product type, and day of the week, this research aims to uncover patterns that can enhance inventory management and customer engagement. The Random Forest model was evaluated through cross-validation, yielding a mean Mean Squared Error (MSE) of 80.97, which indicates moderate predictive accuracy and represents an improvement over traditional forecasting methods commonly employed in the industry. Feature importance analysis revealed that Premium Beans is the most influential predictor, followed by seasonal trends (month), time of day, and weekend sales patterns. These findings underscore the importance of incorporating temporal and contextual factors into forecasting models. Despite these promising results, the model's performance exhibited variability, suggesting room for further refinement through better feature selection, the inclusion of external variables (e.g., weather, local events), and advanced hyperparameter tuning. This study highlights the potential of ML in enhancing operational efficiency and decision-making in the coffee shop industry, while also pointing to opportunities for future research to optimize prediction models and drive profitability.

Keyword: Data Analysis, Coffee Shop, Machine Learning, Random Forest, Sales Prediction.

1. INTRODUCTION

In recent years, the coffee shop industry has experienced significant growth, evolving into a highly competitive marketplace driven by the popularity of specialty coffee and the demand for unique, personalized experiences. For example in particular regions like Hongkong, where keen competition has emerged due to the increasing importance of this segment in the foodservice industry [1]. This expansion has transformed coffee shops from mere caffeine stops into vibrant community hubs where people gather to work and socialize. However, this rapid growth has intensified competition, making it crucial for business owners to adopt effective, data-driven strategies to navigate various operational decisions, from product selection to promotional timing and stock placement, all while adapting to shifting consumer preferences and market trends [2]. Consequently, coffee shop operators increasingly recognize the need to leverage data analytics to gain insights into customer behavior, which can drive sales and enhance loyalty.

Despite the clear advantages of data-driven strategies, many coffee shop owners still struggle to deeply understand their customers' purchasing patterns [3][4]. The primary challenge lies in analyzing data across dimensions such as time, location, and product category. Customer preferences can vary widely depending on the time of day, with certain items being more popular during morning rush hours and others in the afternoon [5]. Geographical factors also play a role, as urban customers may have different tastes than those in suburban or rural areas. Additionally, the diverse product offerings in coffee shops complicate the analysis, making it difficult to determine which items resonate most with customers at different times and locations [6]. This ambiguity can lead to suboptimal decisions regarding inventory management and promotional strategies, ultimately impacting profitability and customer satisfaction. The rapid pace of change in consumer trends, fueled by social media and online reviews, further complicates matters, requiring coffee shop owners not only to keep up with these shifts but also to anticipate future trends [7].

A robust analytical framework is essential for processing and interpreting data in real-time, enabling agile decision-making in this dynamic environment. While traditional analytical methods and rule-based systems provide foundational insights, they often struggle to capture the intricate, non-linear patterns inherent in complex customer behaviors and rapidly shifting market trends [8]. To overcome these limitations and further enhance analytical capabilities, the integration of machine learning (ML) techniques has become increasingly relevant in the retail and food & beverage (F&B) sectors [9].

By leveraging advanced algorithms such as Random Forest and K-Nearest Neighbors (KNN), businesses can gain deeper, predictive insights into customer behavior and market trends that are beyond the scope of simpler, predefined rules [10]. These algorithms are particularly effective for tasks like precise sales prediction and nuanced transaction time segmentation, allowing companies to make truly informed decisions that dynamically align with evolving consumer demands [11]. Previous studies have demonstrated the efficacy of these ML approaches in accurately forecasting sales and optimizing inventory management [7], [10], [12], often outperforming traditional methods when compared. Moreover, the implementation of ML in physical store operations has shown promising results, streamlining processes and enhancing customer experiences, ultimately driving profitability in an increasingly competitive landscape [2].

Machine learning, a subset of artificial intelligence, involves the use of algorithms and statistical models to enable computers to perform specific tasks without explicit instructions [13], [14], making it ideally suited for identifying complex, unforeseen correlations within vast datasets. In the context of retail and F&B, ML can analyze vast amounts of data to identify patterns, predict outcomes, and make recommendations [15]. This capability is particularly valuable in an industry characterized by rapidly changing consumer preferences and market dynamics, where static, rule-based approaches quickly become obsolete. By employing ML, businesses can enhance their operational efficiency, improve customer engagement, and ultimately increase their bottom line [16]. However, despite its proven advantages in broader retail contexts, the application of ML specifically in the context of coffee shops—which are characterized by an extensive variety of products and highly dynamic visiting hours that challenge conventional analysis—remains relatively limited and presents a significant research gap.

The purpose of this study is to explore the use of ML in predicting sales performance within the dynamic coffee shop business environment. Specifically, we aim to examine how factors such as transaction timing, store location, product type, and day of the week influence customer purchasing behavior. As the coffee shop industry continues its rapid growth, leveraging data-driven insights becomes essential for improving decision-making processes related to product offerings and inventory planning. While many coffee shop owners still rely on intuition or past experiences, these methods often fall short in dynamic market conditions where customer preferences can shift throughout the day or week. This study seeks to bridge that gap by addressing the following research questions: How can ML be effectively applied within the coffee shop business environment to enhance sales performance?; What specific factors (such as transaction timing, store location, product type, and day of the week) most significantly influence customer purchasing behavior in coffee shops when analyzed with ML techniques?; How accurately does a predictive model, specifically Random Forest, perform in forecasting coffee shop sales based on historical transaction data, considering the inherent variability in customer preferences and operational complexities?

Despite the recognized advantages of machine learning, its application in coffee shops remains limited. This research acknowledges the potential challenges in applying ML, such as data sparsity from diverse product offerings and the complexities of real-time implementation in a fast-paced retail setting. However, by addressing these nuances, this study presents a unique opportunity to yield valuable insights into customer behavior and preferences, ultimately offering data-driven strategies that can enhance profitability and customer satisfaction in this evolving industry.

2. MATERIALS AND METHOD

This study adopts a quantitative approach using supervised ML techniques to predict sales performance in a coffee shop environment as shown in Figure 1. The research is based on a dataset consisting of historical transaction records collected over a six-month period. The dataset includes key features such as transaction date, transaction time, product category, store location, unit price, and transaction quantity. The methodological process begins with data preprocessing, where data types are standardized and categorical string values are cleaned for consistency and accuracy. As part of feature engineering, additional time-related features such as hour of transaction, day of the week, and weekend indicators are derived. A new column for total sales is also calculated by multiplying unit price with transaction quantity. These steps enhance the dataset to better capture time-based purchasing patterns.

Exploratory Data Analysis (EDA) is conducted to identify preliminary trends [17], such as product popularity across different times and days, as well as peak hours for transactions. These insights are crucial for informing the predictive modeling process and aligning business strategies with customer behavior. For the predictive modeling phase, the study utilizes the Random Forest Regressor algorithm to forecast sales values. The model is trained on input features including time of day, day of the week, store location, and product type, while the target variable is total sales. Model performance is evaluated using 5-fold cross-validation, and Mean Squared Error (MSE) is used as the primary evaluation metric to assess the accuracy of the predictions [18]. A lower MSE indicates better model performance . This methodology provides a robust analytical framework for understanding customer purchasing behavior in coffee shops and aims to generate actionable insights to support data-driven decision-making, particularly in areas such as inventory management, product planning, and promotional timing.

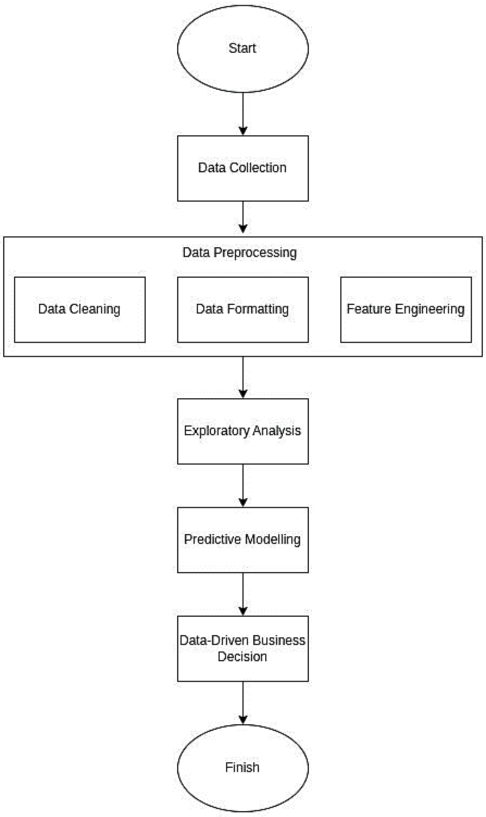

Figure 1. Research Methodology

2.1. Type and Research Approach

This study adopts a quantitative research approach with an emphasis on applied experimentation in the field of data analysis and machine learning. The primary objective is to develop a sales prediction model based on historical transaction data from a coffee shop. Utilizing the Random Forest Regressor algorithm, the research aims to identify purchasing behavior patterns influenced by transaction timing, store location, and product types. The resulting model is expected to provide valuable insights to support data-driven decision-making, particularly in areas such as product planning, inventory management, and the strategic timing of promotions.

2.2. Data Collection

The dataset utilized in this study comprises a comprehensive record of transactions from a coffee shop, encompassing a total of 149,116 entries collected over a six-month period from January to June 2023. This dataset serves as a rich source of information, capturing various attributes critical for analysis, including transaction ID, transaction date, transaction time, transaction quantity, store ID, store location, product ID, unit price, product category, product type, and product detail. All columns in the dataset are complete with no null values, providing a robust foundation for further analysis. However, certain fields such as transaction_time, which is currently stored as an object data type require data type adjustments. Additionally, some columns exhibit high variance, which may reduce analytical efficiency. These issues will be addressed during the data preprocessing phase to ensure consistency and improve model performance. The dataset's temporal granularity allows for in-depth exploration of customer purchasing behavior in relation to time-based factors. Therefore, the selection of relevant features plays a critical role in supporting the subsequent stages of feature engineering, EDA, and predictive modeling. Careful handling of these features ensures that the insights generated are both meaningful and actionable for enhancing operational decision-making within the coffee shop business environment [19]. Table 1 provides basic information on the dataset used.

Table 1. Dataset Ground Information

No	Column	Unique Value	Non-Null Count	Data Type
1	transaction_id	149116	0	int64
2	transaction_date	181	0	datetime64
3	transaction_time	25762	0	object
4	transaction_qty	6	0	int64
5	store_id	3	0	int64
6	store_location	3	0	object
7	product_id	80	0	int64
8	unit_price	41	0	float64
9	product_category	9	0	object
10	product_type	29	0	object
11	product_detail	80	0	object

2.3. Data Preprocessing

Data preprocessing is a crucial step in preparing raw data for analysis , ensuring that the dataset is clean, well-structured, and suitable for generating reliable insights. In this study, several preprocessing techniques were applied to address different data quality issues and enhance the dataset's analytical value, as summarized in Table 2.

Table 2. Data Preprocessing Summary

No	Case	Solution	Action-to Column
1	False Data Type	Reformatting [20]	‘transaction_time’
2	Unused Column	Feature Selection [21]	‘transaction_id’, ‘store_id’, ‘product_id’, ‘product_type’, ‘product_detail’
3	Over Variance	Scalling [22]	‘transaction_qty’,’product_category’
4	Depth_Analytical Purpose	Feature Engineering [23], [24]	‘transaction_date’,’transaction_time’,’transaction_qty’,’unit_price’

The first issue identified was a false data type in the 'transaction_time' column. Originally, the time values were stored as strings, which limited their usability for time-based computations and aggregations. To resolve this, a reformatting process was carried out to convert the data type into a standardized time format, enabling more accurate time-series analysis and easier feature extraction such as hour or minute of transaction as illustrated in Figure 2.

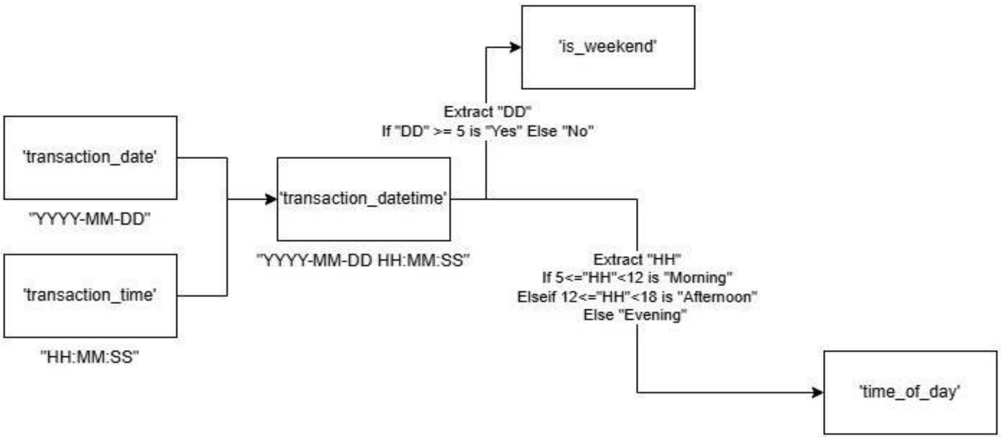

Figure 2. Illustrating Process for Cleaning and Feature Engineering From “transaction_datetime” Column

The second step involved feature selection to eliminate columns that were either redundant or not directly relevant to the analysis objectives. Columns such as 'transaction_id', 'store_id', 'product_id', 'product_type', and 'product_detail' were removed. These attributes, while useful for operational tracking or transactional identification, did not contribute meaningful patterns for the scope of this analysis, and their removal helped reduce dimensionality and potential noise in the model.

Table 3. Example from “product_category” Column Data Distribution Before Mapping

No	Value	Distribution (%)
1	Coffee	39.2
2	Tea	30.5
3	Bakery	15.3
4	Drinking Chocolate	7.7
5	Flavours	4.6
6	Coffee Beans	1.2
7	Loose Tea	0.8
8	Branded	0.5
9	Package Chocolate	0.2

Third, the dataset exhibited signs of over variance in certain features, particularly 'transaction_qty' and 'product_category' as shown in Table 3. High variance in these variables could potentially bias learning algorithms or lead to skewed model performance. To address this, scaling techniques were applied to normalize the values, ensuring that features contributed equally to the model and were measured on a comparable scale as in Table 4. This step was essential for improving the performance of algorithms sensitive to the magnitude of input values.

Table 4. Example from “product_category” Column Data Distribution After Mapping

No	Value After Mapping	Components	Distribution (%)
1	Coffee	Coffee, Coffee Beans	40.3
2	Tea	Tea, Loose Tea	31.3
3	Bakery	Bakery	15.3
4	Chocolate	Drinking Chocolate, Package Chocolate	8
5	Condiment	Flavours	4.6
6	Merchandise	Branded	0.5

Finally, to support in-depth analytical objectives, feature engineering was conducted by deriving new variables from existing ones. For instance, 'transaction_date' and 'transaction_time' were transformed into new features such as day of the week, hour of transaction, and part-of-day categories (e.g., morning, afternoon). Additionally, 'transaction_qty' and 'unit_price' were used to compute the total transaction value, which serves as a key metric in analyzing purchasing behavior. These engineered features provided richer context and improved the dataset’s ability to reveal underlying patterns and trends. The preprocessing phase not only addressed technical inconsistencies and redundancies in the dataset but also enhanced its overall analytical potential through thoughtful transformation and feature enrichment.

2.4. Exploratory Data Analysis (EDA)

EDA is a critical phase in the data analysis process that involves examining datasets to uncover patterns, trends, and insights that may not be immediately apparent. This phase is essential for understanding the underlying structure of the data and for generating hypotheses that can be tested in subsequent analyses [25]. The primary goal of EDAis to provide a comprehensive overview of the data, allowing analysts to identify relationships, anomalies, and areas of interest that warrant further investigation. During exploratory analysis, various techniques such as data visualization, summary statistics, and correlation analysis are employed to gain a deeper understanding of the dataset. This process not only helps in identifying key trends and patterns but also assists in detecting potential issues with the data, such as missing values or outliers. By thoroughly exploring the data, analysts can make informed decisions about the appropriate modeling techniques and methodologies to apply in later stages of analysis. Based on Figure 3, in this section we focus on two key aspects of exploratory analysis: sales performance and product popularity analysis. These analyses provide valuable insights into customer behavior and sales performance, ultimately guiding strategic decisions for the business.

2.5. Predictive Modelling

Predictive modeling is a statistical technique that uses historical data to forecast future outcomes. By leveraging various algorithms and ML techniques, predictive models can identify patterns and relationships within the data, enabling businesses to make informed decisions based on anticipated future trends. The primary goal of predictive modeling is to provide actionable insights that can enhance strategic planning, optimize operations, and improve customer engagement. The specific objective of predictive modeling in the context of sales analysis is to predict the product category that is likely to be sold based on various relevant features. By utilizing historical sales data, this model aims to provide insights that can assist in strategic decision-making, such as inventory management and marketing planning. The key processes for predictive modeling can be seen in Figure 4.

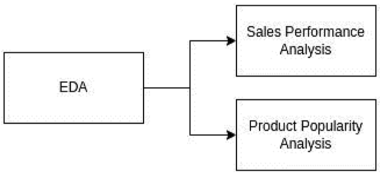

Figure 3. Focus Key Aspect for EDA

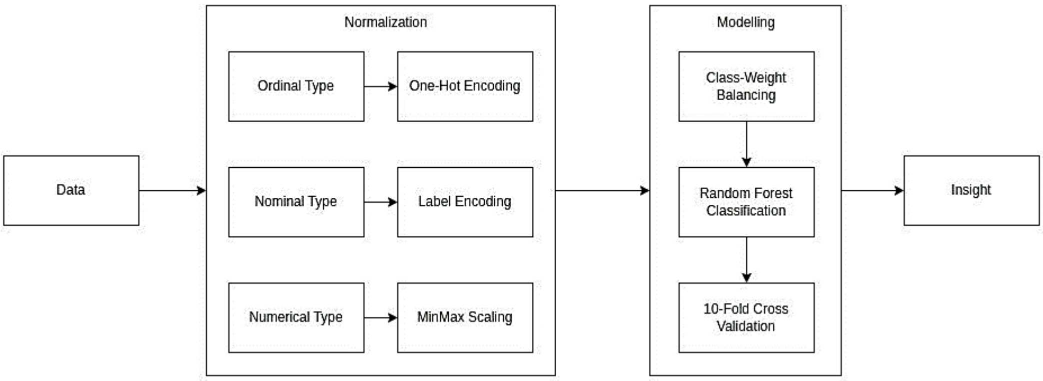

Figure 4. Key Process for Predictive Modelling

3. RESULTS AND DISCUSSION

This chapter presents the key findings derived from the data analysis process, which comprises two main components: EDA and Predictive Modeling. Each section is designed to address specific research objectives and to uncover meaningful insights that support data-driven decision-making. The first part of this chapter focuses on Exploratory Analysis, an essential step aimed at understanding the structure and dynamics of the dataset. This phase involved a thorough examination of sales data to uncover hidden patterns and trends that are not immediately apparent. By employing techniques such as data visualization, summary statistics, and correlation analysis, this study explored various aspects of sales performance [26], [27]. Key areas of focus include analyzing sales trends over time and across locations, visualizing product popularity by hour and day, and identifying peak transaction times. While these descriptive trends provide valuable initial insights into customer behavior and purchasing patterns, a deeper statistical validation will be necessary in subsequent sections to confirm whether observed patterns are significant or merely coincidental. The results are presented in a logical order to form a coherent narrative, focusing on factual data rather than extensive discussion. Tables and Figures can be used to illustrate findings, ensuring no redundant data is presented across different visual aids and text. Subtitles will be employed to further clarify the descriptions within this section. EDA By Sales Performance can view at figure 5.

Figure 5 illustrating "Total Revenue by Month" reveals significant insights into the sales performance from January to June 2023. Beginning with a modest revenue of $81,678 in January, the figures exhibited a slight decline in February, dropping to $76,145. However, March marked a pivotal turnaround with sales increasing to $98,835. This upward trend continued into April, where total revenue reached $118,941, suggesting that the market conditions or sales strategies may have begun to improve. In May, revenue surged dramatically to $156,728, reflecting an impressive growth rate that likely resulted from effective promotional efforts or seasonal demand. This momentum carried into June, where total revenue peaked at $166,486. Notably, throughout this six-month period, the average monthly revenue stood at $116,469, represented by the dashed red line on the graph. This average serves as a crucial benchmark, indicating that the sales figures in May and June not only exceeded this average but also highlighted an overall recovery and robust growth phase.

A more detailed breakdown of this growth can be observed in Figure 6, which analyzes total revenue by store location Astoria, Hell’s Kitchen, and Lower Manhattan over the same six-month period. While all three stores exhibited relatively modest performance in January and February, with revenues generally hovering between $25,000 and $30,000, a notable acceleration in sales began in March. Hell’s Kitchen, in particular, demonstrated the most pronounced growth, rapidly outpacing the other locations. This upward trajectory persisted through May and June, where Hell's Kitchen consistently recorded the highest monthly revenue, surpassing $50,000 by the end of the period. Astoria and Lower Manhattan also showed positive growth trends, with Astoria maintaining a slight edge in performance over Lower Manhattan. These findings suggest that while the general upward trend is consistent across locations, Hell's Kitchen may benefit from stronger market dynamics or more effective local strategies. Futhermore, the following analysis explores product popularity across different times of day and days of the week, offering insights into customer purchasing behavior and preferences that may underlie the observed revenue trends.

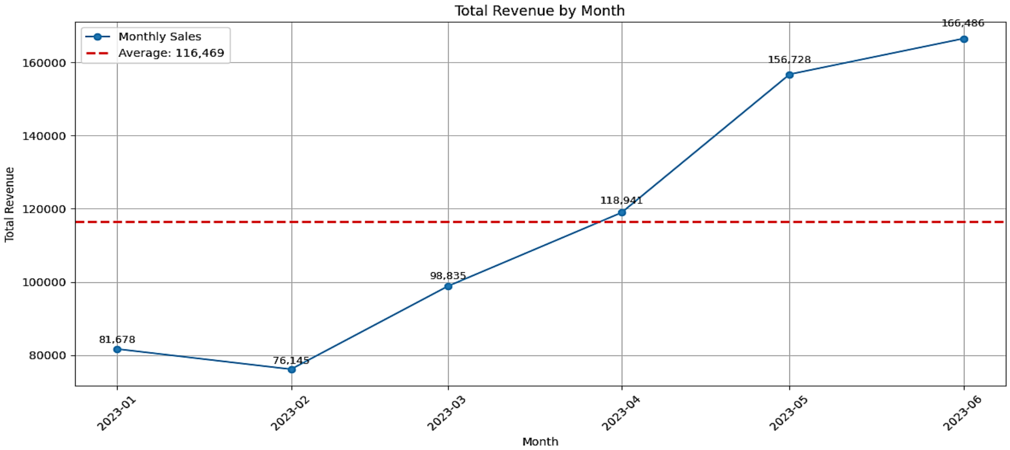

Figure 5. EDA By Sales Performance

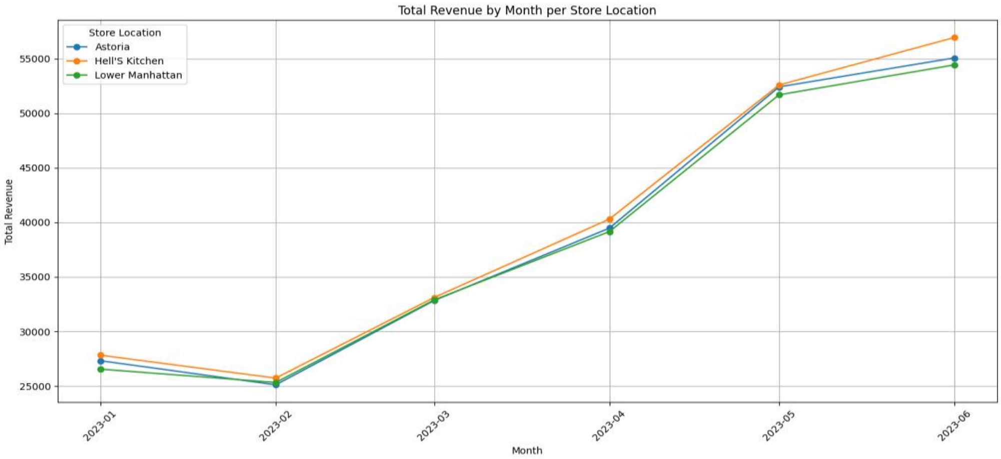

Figure 6. EDA By Sales Performance Per Store

The analysis of product category sales across the first half of 2023 reveals distinct trends in customer preferences as in Figure 7. Coffee consistently emerges as the top-selling category throughout the period, with transaction quantities steadily increasing from 10,589 in January to 21,875 in June. This notable rise highlights coffee's continued popularity and suggests it as a primary driver of revenue for the business. Similarly, tea also demonstrates strong and consistent sales, experiencing a gradual increase from 8,201 transactions in January to 16,699 in June. The growth in both coffee and tea aligns with broader consumer trends toward beverages, indicating robust demand in this segment. In contrast, categories like bakery and chocolate show more moderate increases. Bakery products, while consistently popular, grow from 2,690 transactions in January to 5,431 in June, signaling a steady but less dramatic rise. Chocolate, on the other hand, exhibits a more balanced performance, with a slight upward trend in sales, reaching 4,232 transactions in June from 2,072 in January. Condiments and merchandise, however, lag behind, with sales remaining relatively low across the months.

Our EDA follows an in-depth look at the hourly sales data reveals distinct patterns in product category demand throughout the day as shown in Figure 8. Coffee and tea stand out as the most popular products across nearly all hours, with significant peaks observed during the morning hours. Coffee, in particular, shows a steady increase in sales, reaching a high of 7,344 transactions at 10 AM, before gradually tapering off in the evening. Similarly, tea follows a similar pattern, with peak sales occurring at 10 AM (5,444 transactions) and maintaining a strong presence until mid-afternoon. These trends suggest that coffee and tea are essential products for customers during the early part of the day, likely driven by their role as morning beverages. Bakery products, though not as dominant as coffee or tea, show a clear peak in sales between 8 AM and 10 AM, with transaction quantities consistently above 2,500 units during this period. This indicates a high demand for bakery items in the morning, possibly as customers pair these products with their morning coffee or tea. On the other hand, chocolate sales show a more moderate, consistent demand throughout the day, with no significant peaks, but a steady presence from 6 AM through 3 PM. While chocolate maintains a lower volume than beverages or bakery items, it remains a steady choice for customers. Condiments and merchandise once again show the least variation across hours, with sales being consistently low throughout the day. Condiments experience a slight increase in the morning but remain relatively stable, while merchandise sales are minimal, even during peak hours, suggesting limited customer interest in this category during the day. As noted in the previous analysis, condiments and merchandise demonstrate consistently low sales throughout the day, with minimal variation across hours. Specifically, condiment sales show a slight increase in the morning but remain relatively stable thereafter, while merchandise consistently underperforms, even during peak hours. These trends highlight a potential opportunity for product strategy refinement. By leveraging predictive modeling, the goal is to investigate sales performance factors and develop targeted strategies to increase sales.

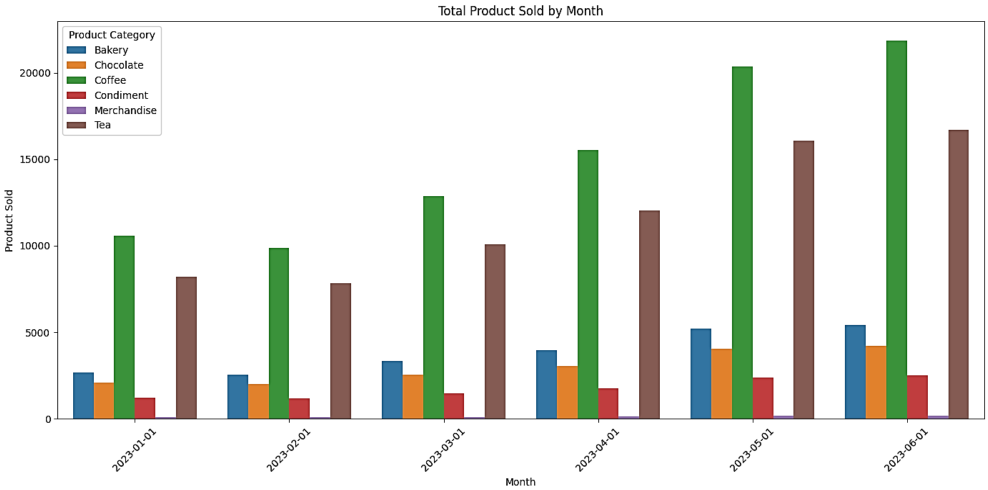

Figure 7. EDA for Product Popularity Within Months

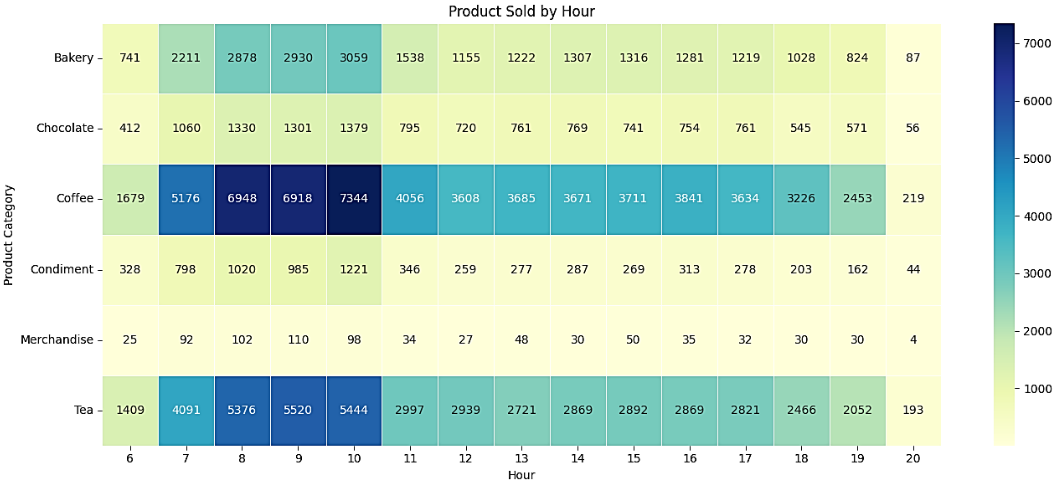

Figure 8. Product Sold Pattern

The performance of the Random Forest Regressor model was rigorously evaluated using 5-fold cross-validation, yielding a mean MSE of approximately 80.97 (as detailed in Table 5). This result suggests the model performs reasonably well, especially when considered against a simple linear regression baseline, which typically exhibits a higher MSE in this context due to its inability to capture complex, non-linear relationships inherent in sales data. While a lower MSE is always desirable, this value indicates a moderate level of predictive accuracy for the Random Forest model. However, a notable concern is the relatively high variance in MSE across the different folds (ranging from 43.83 to 132.17). This variability highlights that the model's performance can fluctuate significantly depending on the specific data subsets used for training and validation. According to research, Each fold in k-fold cross-validation represents a different combination of training and validation data, which can result in different accuracy values for each fold [28]. Such fluctuations likely stem from several factors, including pronounced seasonality (where sales patterns change throughout the year), diverse geographical influences affecting consumer preferences, or even idiosyncratic local events that impact specific store locations. This high variance suggests the model might be sensitive to these uncaptured or under-represented factors, thereby limiting its generalizability across diverse conditions [29]. To mitigate these fluctuations and improve overall model robustness, future refinements could involve incorporating additional external variables (e.g., weather data, local marketing campaigns, competitor activities), conducting more granular feature engineering to account for temporal and spatial nuances, and exploring advanced ensemble techniques or hyperparameter tuning strategies aimed at reducing variance [30].

Table 5. Random Forest Regressor Prediction Performance

Fold	MSE	Mean MSE
1	65.47	
2	43.83	
3	132.17	80.97
4	98.34	
5	65.03	

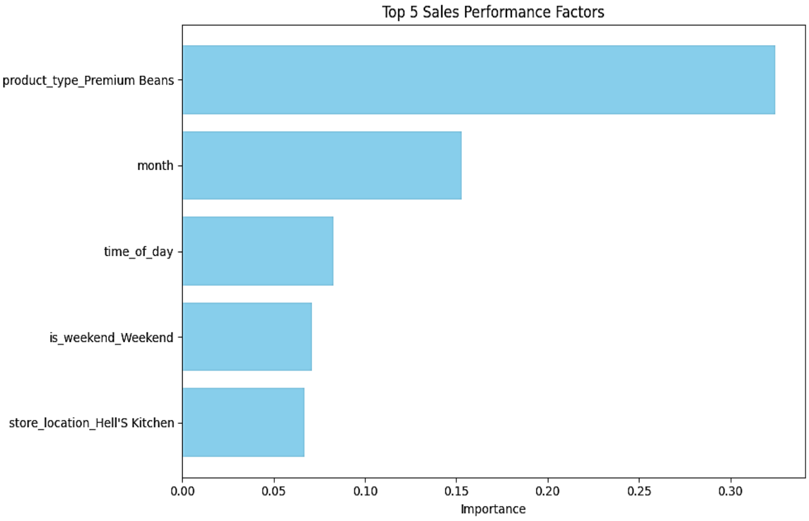

Figure 9. Sales Performance Factors Based on Random Forest Regressor

In terms of feature importance, visualized in Figure 9, the model identifies several crucial predictors for accurately forecasting sales. The most important feature is product_type_Premium Beans, with an importance score of 0.32, indicating its significant role in the model's predictions. The prominence of premium bean products is likely attributed to their higher price point and a perceived greater consumer demand for quality or exclusive items. While this aligns with general business knowledge in the coffee industry, further investigation through pricing data or customer surveys would provide empirical evidence to fully explain why these products tend to generate more revenue and exhibit distinct sales patterns compared to regular offerings. Additionally, the feature month has an importance score of 0.15, highlighting the strong role of seasonality in sales patterns. This suggests that sales are significantly influenced by the time of year, with particular months (perhaps during holidays or special promotions) seeing spikes or drops in customer demand. For instance, sales may be higher during colder months when customers are more likely to purchase warm beverages, or during certain cultural or seasonal events that affect coffee consumption. This finding underscores the importance of temporal features in forecasting, and further implies that incorporating more detailed seasonal features, such as specific holidays or regional events, could significantly enhance the model's predictive accuracy.

The time_of_day feature, with an importance score of 0.08, also plays a notable role in predicting sales, indicating that customer behavior is significantly influenced by the hour. This consistency aligns with the understanding that certain products are more popular at different times morning customers may gravitate toward coffee and pastries, while afternoon or evening customers might prefer lighter beverages or snacks. The timing of a customer's visit clearly appears to be a significant factor in shaping their purchasing decisions, and this can vary depending on store type, location, and individual customer preferences. Moreover, the is_weekend_Weekend feature, with an importance score of 0.07, suggests that sales patterns are distinctly different on weekends compared to weekdays. This aligns with common consumer behavior, where weekend shoppers often have more leisure time and higher spending tendencies, particularly in retail and food service environments. The model's sensitivity to weekends reinforces the notion that business strategies, such as promotions, inventory management, or staffing schedules, should actively consider weekend dynamics, as consumer spending habits often vary considerably from weekdays.

Despite the model's relatively strong performance in capturing these temporal patterns, several opportunities for improvement and limitations remain. The observed high variance in MSE (as discussed previously) suggests that while the Random Forest model is powerful, it might be susceptible to capturing noise in the training data, potentially leading to overfitting on specific subsets. This issue could be compounded by limitations in data granularity, particularly if the available dataset lacks detailed information on individual customer demographics or highly specific contextual events. To address these limitations and enhance the model's generalizability and reproducibility, future work could involve more sophisticated feature engineering based on deeper domain knowledge, the incorporation of additional external variables such as local events, weather conditions, or real-time social media trends, and exploring advanced regularization techniques or alternative ensemble methods to mitigate overfitting. Furthermore, conducting sensitivity analyses to understand how variations in input data affect predictions would bolster the study's credibility and provide more actionable insights for real-world application.

4. CONCLUSION

The evaluation of the Random Forest Regressor model, based on cross-validation, reveals it performs reasonably well, with a mean MSE of 80.97. While this indicates decent predictive accuracy, the model still exhibits moderate error, and there's clear room for improvement. The variance in MSE across the five folds (ranging from 43.83 to 132.17) suggests the model's performance fluctuates significantly depending on the data subsets, highlighting its sensitivity to varying conditions like seasonality and geographical factors. These factors, including sales patterns that change across different times of the year or varying consumer behavior based on store locations, contribute to this performance variability. Therefore, improving the model's generalizability across diverse scenarios will be a key area of focus in future work.

In terms of feature importance, the model identifies key predictors that strongly influence sales forecasts. The most influential feature is product_type_Premium Beans, with an importance score of 0.32, suggesting premium products significantly drive sales predictions, likely due to their higher price points and exclusive demand. Seasonality, captured by the month feature (importance score 0.15), plays a critical role, reinforcing the significance of time-based patterns in sales. This finding suggests the model could benefit from further refinement by incorporating more granular seasonal features, such as major holidays (e.g., Christmas, Eid al-Fitr) or significant local events (e.g., city festivals, university breaks), to improve forecasting accuracy.

Additionally, time_of_day (0.08) and is_weekend_Weekend (0.07) are identified as significant predictors, indicating consumer behavior is influenced by both the time of day and whether it's a weekend. These insights align with established retail patterns, where certain products are more popular during specific times (e.g., coffee in the morning, lighter snacks in the afternoon), and spending habits differ between weekdays and weekends. Thus, the model's sensitivity to these temporal factors suggests businesses should tailor their inventory management, staffing, and promotional strategies accordingly to optimize sales performance.

Despite the model's promising results, this study acknowledges several broader constraints. These include potential limitations in data quality, such as missing values or inconsistencies in transaction records, and possible sampling biases if the historical data used does not fully represent all coffee shop locations or customer segments. Furthermore, real-world implementation challenges, such as the need for robust IT infrastructure for real-time data processing and the integration of ML outputs into existing operational workflows, were not explicitly within the scope of this methodological study.

Therefore, there are several avenues for further enhancement. The observed high variance in MSE implies that better feature selection or the inclusion of additional, contextually relevant external variables—such as local weather conditions (e.g., temperature, rainfall impacting outdoor seating or beverage choice), major public holidays, nearby office working patterns, or even competitor promotions detected via social media trends—could help refine predictions and reduce fluctuations in performance. Furthermore, advanced hyperparameter tuning and exploring other modeling techniques, such as time-series specific models (e.g., ARIMA, Prophet) or deep learning approaches, could enhance the model's stability and predictive power, making it more robust for practical application in the dynamic coffee shop industry.

REFERENCES

[1] A. Y. F. Tan and A. S. Y. Lo, “A Benefit-Based Approach To Market Segmentation: A Case Study of an American Specialty Coffeehouse Chain in Hong Kong,” *J. Hosp. Tour. Res.*, vol. 32, no. 3, pp. 342–362, Aug. 2008, doi: 10.1177/1096348008317388.

[2] K. Vayadande, R. Deshpande, M. Deshpande, P. Chaudhary, A. Dhangar, and T. Dhangar, “Tracking Barista Productivity and Customer Demographics,” in *2024 5th International Conference on Data Intelligence and Cognitive Informatics (ICDICI)*, Tirunelveli, India: IEEE, Nov. 2024, pp. 891–898. doi: 10.1109/icdici62993.2024.10810948.

[3] M. Cemberci, S. Cicek Vural, C. Celik, and E. Canbaz, “The Role of Supply Chain Transparency in the Relation between Supply Chain Analytics Capabilities and Firm Performance,” *Oper. Supply Chain Manag. Int. J.*, pp. 253–263, June 2024, doi: 10.31387/oscm0570426.

[4] C. Udokwu, P. Brandtner, F. Darbanian, and T. Falatouri, “Proposals for Addressing Research Gaps at the Intersection of Data Analytics and Supply Chain Management,” *J. Adv. Inf. Technol.*, vol. 13, no. 4, 2022, doi: 10.12720/jait.13.4.338-346.

[5] W. S. Lee, J. Moon, and M. Song, “Attributes of the coffee shop business related to customer satisfaction,” *J. Foodserv. Bus. Res.*, vol. 21, no. 6, pp. 628–641, Nov. 2018, doi: 10.1080/15378020.2018.1524227.

[6] O. Putri Dahlan, S. Putri Dahlan, and M. Fahlevi, “Marketing Mix Elements on Customer Service Satisfaction at Coffee Shops in Jakarta,” *E3S Web Conf.*, vol. 448, p. 01004, 2023, doi: 10.1051/e3sconf/202344801004.

[7] P. Ruangchoengchum and P. Thatphet, “Improving Customer Service Efficiency Using Demand Forecasting with Leagile and Lean Six Sigma Concepts: A Case Study,” *Suranaree J. Soc. Sci.*, vol. 18, no. 1, June 2024, doi: 10.55766/sjss-1-2024-267502.

[8] K. Sinha, “New Trends and their Impact on Business and Society,” *J. Creat. Commun.*, vol. 3, no. 3, pp. 305–317, Nov. 2008, doi: 10.1177/097325861000300304.

[9] I. D. Sudirman and R. Rahmah, “Dynamic Pricing Optimization for Coffee Shops Using a Machine Learning Approach with Random Forest Models,” in *2025 3rd International Conference on Disruptive Technologies (ICDT)*, Greater Noida, India: IEEE, Mar. 2025, pp. 1282–1286. doi: 10.1109/icdt63985.2025.10986336.

[10] P. S. Dahake and N. Somani, “Harnessing Predictive Analytics for Accurate Consumer Behaviour Forecasting: A Comprehensive Review,” in *2024 2nd DMIHER International Conference on Artificial Intelligence in Healthcare, Education and Industry (IDICAIEI)*, Wardha, India: IEEE, Nov. 2024, pp. 1–6. doi: 10.1109/idicaiei61867.2024.10842743.

[11] P. S. Dahake, S. Chandak, R. V. Mohare, K. Wadhwani, and P. Bhadade, “The Crystal Ball of Marketing: How Predictive Analytics is Reshaping the Industry?,” in *2023 Second International Conference On Smart Technologies For Smart Nation (SmartTechCon)*, Singapore, Singapore: IEEE, Aug. 2023, pp. 304–311. doi: 10.1109/smarttechcon57526.2023.10391334.

[12] S. P. Praveen, P. Chaitanya, A. Mohan, V. Shariff, J. V. N. Ramesh, and J. Sunkavalli, “Big Mart Sales using Hybrid Learning Framework with Data Analysis,” in *2023 2nd International Conference on Automation, Computing and Renewable Systems (ICACRS)*, Pudukkottai, India: IEEE, Dec. 2023, pp. 471–477. doi: 10.1109/icacrs58579.2023.10404941.

[13] H. Pallathadka, M. Jawarneh, F. Sammy, V. Garchar, D. T. Sanchez, and M. Naved, “A Review of Using Artificial Intelligence and Machine Learning in Food and Agriculture Industry,” in *2022 2nd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE)*, Greater Noida, India: IEEE, Apr. 2022. doi: 10.1109/icacite53722.2022.9823427.

[14] H. Pallathadka *et al.*, “An investigation of various applications of machine learning in food industry,” in *AIP Conference Proceedings*, Nandyal, India: AIP Publishing, 2023, p. 090001. doi: 10.1063/5.0150516.

[15] R. Mahmoud *et al.*, “Revolutionizing Food Quality With Machine Vision and Machine Learning Techniques,” in *Food in the Metaverse and Web 3.0 Era*, IGI Global, 2025, pp. 71–124. doi: 10.4018/979-8-3693-9025-2.ch005.

[16] N. J. Watson *et al.*, “Intelligent Sensors for Sustainable Food and Drink Manufacturing,” *Front. Sustain. Food Syst.*, vol. 5, Nov. 2021, doi: 10.3389/fsufs.2021.642786.

[17] A. S. Rao, B. V. Vardhan, and H. Shaik, “Role of Exploratory Data Analysis in Data Science,” in *2021 6th International Conference on Communication and Electronics Systems (ICCES)*, Coimbatre, India: IEEE, July 2021, pp. 1457–1461. doi: 10.1109/icces51350.2021.9488986.

[18] T. O. Hodson, T. M. Over, and S. S. Foks, “Mean Squared Error, Deconstructed,” *J. Adv. Model. Earth Syst.*, vol. 13, no. 12, Dec. 2021, doi: 10.1029/2021ms002681.

[19] S.-C. Kim, S. R. Salkuti, A. M. Suresh, and M. S. Sankaran, “Data analysis and visualization on titanic and student’s performance datasets-an exploratory study,” *Int. J. Inform. Commun. Technol. IJ-ICT*, vol. 14, no. 1, p. 68, Apr. 2025, doi: 10.11591/ijict.v14i1.pp68-76.

[20] C. E. Morr *et al.,* “Data Preprocessing,” in *International Series in Operations Research & Management Science*, Cham: Springer International Publishing, 2022, pp. 117–163. doi: 10.1007/978-3-031-16990-8_4.

[21] J. A. Oribe *et al*., “Data preprocessing techniques for earth resource management,” in *Data Analytics and Artificial Intelligence for Earth Resource Management*, Elsevier, 2025, pp. 37–64. doi: 10.1016/b978-0-443-23595-5.00003-6.

[22] K. G. Samuel *et al.,* “Covid-19 Data Preprocessing Approach in Machine Learning for Prediction,” in *Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering,* Cham: Springer Nature Switzerland, 2024, pp. 328–344. doi: 10.1007/978-3-031-56396-6_21.

[23] R. Shweta *et al.,* “Preprocessing of Datasets Using Sequential and Parallel Approach: A Comparison,” in *Lecture Notes in Networks and Systems*, Singapore: Springer Singapore, 2022, pp. 311–320. doi: 10.1007/978-981-16-2126-0_27.

[24] M. J. Reena, “Preprocessing Big Data using Partitioning Method for Efficient Analysis,” in *2023 IEEE International Conference on Contemporary Computing and Communications (InC4)*, Bangalore, India: IEEE, Apr. 2023, pp. 1–6. doi: 10.1109/inc457730.2023.10262924.

[25] N. Andrienko and G. Andrienko, *Exploratory Analysis of Spatial and Temporal Data*. Berlin/Heidelberg: Springer-Verlag, 2006. doi: 10.1007/3-540-31190-4.

[26] S. A. Khan and S. S. Velan, “Application of Exploratory Data Analysis to Generate Inferences on the Occurrence of Breast Cancer using a Sample Dataset,” in *2020 International Conference on Intelligent Engineering and Management (ICIEM),* London, United Kingdom: IEEE, June 2020. doi: 10.1109/iciem48762.2020.9160290.

[27] C. Selvi G. and L. Priya G. G., Eds., “An Epidemic Analysis of COVID-19 using Exploratory Data Analysis Approach,” in *Predictive Analytics Using Statistics and Big Data: Concepts and Modeling*, BENTHAM SCIENCE PUBLISHERS, 2020, pp. 99–111. doi: 10.2174/9789811490491120010010.

[28] T. Gunasegaran and Y.-N. Cheah, “Evolutionary cross validation,” in *2017 8th International Conference on Information Technology (ICIT)*, Amman, Jordan: IEEE, May 2017, pp. 89–95. doi: 10.1109/icitech.2017.8079960.

[29] J. Smith *et al.*, “Making Early Predictions of the Accuracy of Machine Learning Classifiers,” in *Learning in Non-Stationary Environments*, New York, NY: Springer New York, 2012, pp. 125–151. doi: 10.1007/978-1-4419-8020-5_6.

[30] B. Kartal and B. B. Üstündağ, “Energy and Entropy based Intelligence Metric for Performance Estimation in DNNs,” in *2023 International Conference on Artificial Intelligence in Information and Communication (ICAIIC)*, Bali, Indonesia: IEEE, Feb. 2023, pp. 468–473. doi: 10.1109/icaiic57133.2023.10067093.