**EVALUATION OF CUSTOMER SPENDING ON FOOD PRODUCTS USING MACHINE LEARNING**

***NAME:***

***ID:***

**Table of Contents**

[Chapter 1: Introduction 3](#_Toc182310276)

[1.1 Context of the Research 3](#_Toc182310277)

[1.2 Problem Statement 4](#_Toc182310278)

[1.3 Purpose of this project and its application 4](#_Toc182310279)

[1.4 Research Questions 4](#_Toc182310280)

[1.5 Research aim and objectives 5](#_Toc182310281)

[1.5.1 Aim 5](#_Toc182310282)

[1.5.2 Objectives 5](#_Toc182310283)

[1.6 Research Novelty 5](#_Toc182310284)

[Chapter 2: Background 6](#_Toc182310285)

[2.1 Introduction to the Literature 6](#_Toc182310286)

[2.2 Selection of the Papers 6](#_Toc182310287)

[2.3 Critical analysis of crucial papers 8](#_Toc182310288)

[2.4 Comparison of the Papers 11](#_Toc182310289)

[2.5 Conceptual Framework 11](#_Toc182310290)

[2.6 Literature Gaps 12](#_Toc182310291)

[2.7 Summary 12](#_Toc182310292)

[Chapter 3: Data 13](#_Toc182310293)

[3.1 Data Collection Process 13](#_Toc182310294)

[3.2 Dataset Description 13](#_Toc182310295)

[3.3 Rationale for choosing the dataset 13](#_Toc182310296)

[3.4 Data preprocessing 14](#_Toc182310297)

[3.5 EDA 14](#_Toc182310298)

[Chapter 4: Ethical Issues 15](#_Toc182310299)

[4.1 Dataset Ethics 15](#_Toc182310300)

[4.2 Research Ethics 15](#_Toc182310301)

[Chapter 5: Methodology 16](#_Toc182310302)

[5.1 Introduction 16](#_Toc182310303)

[5.2 Proposed Methodological Architecture 16](#_Toc182310304)

[5.3 Justification of the Methodological Architecture 16](#_Toc182310305)

[5.4 Software and Libraries used 17](#_Toc182310306)

[5.5 Machine Learning models used and its justification 17](#_Toc182310307)

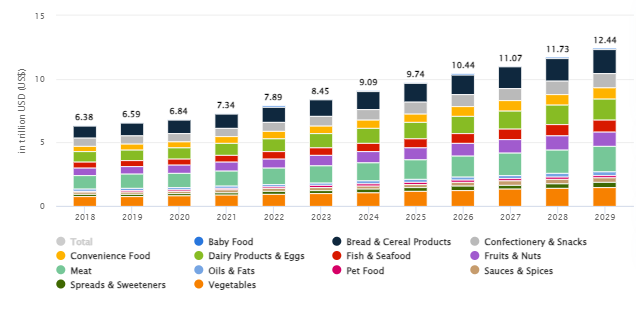
[5.6 Conclusion 19](#_Toc182310308)

[References 20](#_Toc182310309)

# Chapter 1: Introduction

## 1.1 Context of the research

The term 'Customer Spending' can be defined as the total amount of money consumers spend on goods and services, primarily including induced consumption and autonomous consumption. According to Cai (2023) and Daroch, Nagrath and Gupta (2021), the factors influencing consumer spending include economic factors such as income, tax rates, consumer confidence, product pricing, market conditions and customer preferences. Therefore, changes in customer product pricing and income levels can lead to variations in customer spending on food products, as it is linked with the change in market conditions. Global customer expenditure on foods and beverages will reach $9.09 trillion in 2024 (approximately 10.7% of total customer expenditure) (Cognitive Market Research, 2023). Additionally, with the increasing population and income, market revenue in the food industry (globally) is expected to reach $12.44 trillion by 2030 (Statista, 2024) (***Refer to Figure 1***). Therefore, identifying spending patterns within this increasing market trend enables personalised marketing and ensures efficient resource allocation for better profitability and customer retention.

******

***Figure 1: Market revenue in the food sector across the world from 2018 to 2030***

(Source: Statista, 2024)

## 1.2 Problem Statement

Evaluating customer spending behaviour within the highly competitive, growing market is crucial for retail food businesses to optimise their operations and increase operational sustainability. Traditional methods (such as Recency, Frequency and Monetary (RFM) analysis and Basket analysis) of analysing customer spending are often limited due to their lack of ability to uncover complex relationships between customer spending and external attributes like spending preferences (Heldt, Silveira and Luce, 2019; Christy *et al.*, 2018). With the growing availability of large datasets capturing demographic information and purchasing habits, there is an opportunity to leverage machine learning techniques for more accurate and insightful predictions of customer spending on food products.

## 1.3 Purpose of this project and its application

This project aims to leverage Machine Learning (ML) techniques (such as Clustering and Classification) to predict customer spending on food products based on customer attributes and spending behaviour. This can enable businesses to make data-driven decisions about consumer spending on food products and their buying behaviour, allowing food businesses within the retail landscape to optimise resource allocation and streamline their productions. The application of this research can be extended to fields like retail management, customer spending prediction and supply chain optimisation, as this study can predict customer spending on food products and their spending behaviour, which can be beneficial for businesses to manage their material resources and optimise their supply chain networks.

## 1.4 Research questions

* What factors influence customer spending on food products and their spending behaviour?
* What is the significance of statistical analysis and Machine Learning techniques in the prediction of customer spending?
* What is the implication of ML algorithms (Clustering and Classification) in predicting customer spending on food products based on historical spending and demographic attributes?

## 1.5 Research aim and objectives

### 1.5.1 Aim

This study aims to evaluate customer spending on food products using Machine Learning (ML) techniques.

### 1.5.2 Objectives

* To evaluate different factors influencing customer spending on food products
* To assess the significance of Machine Learning (ML) techniques for evaluating customer spending on food products
* To develop ML models (Clustering and Classification) for predicting customer spending on food products based on historical spending data and demographic attributes
* To recommend effective strategies to retail food businesses for enhancing customer spending on food products and implying effective market strategies

## 1.6 Research Novelty

The novelty of this study lies in integrating advanced ML techniques with customer spending prediction models tailored explicitly to the retail food sector. Unlike conventional methods (RFM analysis and Basket analysis) followed in most past studies based on linear relationships and simplistic assumptions, this project intends to explore non-linear, high-dimensional relationships in large datasets to unleash underlying patterns in customers' spending behaviour. Additionally, this study has incorporated diverse features like region and customer historical spending, and it provided more accurate and actionable predictions for estimating customer spending on food products. Additionally, clustering techniques can allow this study to perform customer segmentation based on purchasing habits and spending.

# Chapter 2: Background

## 2.1 Introduction to the literature

Literature reviews refer to evaluating methodological choices followed in past studies and their respective findings within the context of a specific research field. The literature review in a study is based on the evaluation of purpose, methodological choices and findings of five selected peer-reviewed papers related to customer spending on food products and the implication of ML algorithms in predicting customer spending and customer behaviour. The prediction of customer spending behaviour based on historical spending data is linked with the subject area of data science and machine learning, which integrates concepts of data distribution, exploratory data analysis, and ML modelling.

## 2.2 Selection of the Papers

The papers were selected based on multiple factors, such as their relevance to this study, publication details, recency (published after 2018), uniqueness of methodological choices, and depth of findings. The table below provides the fulfilment of these selection criteria for each selected paper.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Title and Author** | **Relevance to Project** | **Machine Learning Focus** | **Customer Spending Focus** | **Food Products Context** | **Prediction/Forecasting** |
| “Applicability of machine learning techniques in food intake assessment: A systematic review” (Chaves *et al.*, 2021) | Explores machine learning techniques applicable to food assessment and relevant to predicting customer spending. | Strong focus on various machine learning methods using tools like WEKA and R. | Indirect:  Focus on food intake rather than spending. | Direct:  Applicable to food products. | Partial:  Focus on intake assessment rather than sales prediction using Supervised Learning methods. |
| “Predicting future consumer purchases in grocery retailing with the condensed Poisson lognormal model” (Trinh and Wright, 2022) | It is highly relevant as it directly predicts consumer purchases in grocery retailing, similar to food product spending using ML techniques like Condensed Negative Binomial Distribution (CNBD) and Condensed Poisson lognormal model (CPLN) | It uses a Poisson lognormal model, which can complement ML approaches. | Direct focus on predicting consumer spending. | Direct:  Focus on grocery purchases, similar to food products. | Strong:  Predicts future purchases, aligning with spending. |
| “Customer Value Types Predicting Consumer Behavior at Dutch Grocery Retailers”(Janssens *et al.*, 2020) | Relevant for understanding how customer value types impact spending behaviour in grocery retailing. | Uses statistical and behavioural approaches, not primarily ML-based. | Direct focus on consumer behaviour and spending. | Direct: Focus on grocery retailing. | Strong: Predicts consumer behaviour and spending in retail. |
| "From Intention to Action: Predicting Purchase behaviour with Consumers' product expectations and perceptions" (Kytö, Virtanen and Mustonen, 2019) | Relevant for understanding the link between consumer perceptions and purchase behaviour. | Focus on consumer behaviour prediction, applicable to ML frameworks. | Direct focus on predicting consumer behaviour. | Indirect: More focused on general purchase behaviour, less on food. | Strong: Predicts purchasing behaviour based on perceptions. |

***Table 1: Selection of the papers***

## 2.3 Critical analysis of key papers

**Paper 1:** “Applicability of machine learning techniques in food intake assessment: A systematic review” (Chaves *et al.*, 2021)

The research by Chaves *et al.* (2021) focused on evaluating the applicability of the ML techniques in the estimation of the food intake of customers. This work has utilised 36 secondary research studies on ML modelling to predict food consumption using systematic literature reviews and meta-analyses. The obtained results of this study revealed that the implication of Decision tree (DT), Support Vector Machine (SVM), Naive Bayes Regressor, k-nearest Neighbour (KNN) Regressor and Artificial Neural Networks (ANN) are highly reliable in the prediction of customer food intake. Moreover, the obtained predictive accuracy of the DT and ANN models is considerably higher than that of other models due to their capability of capturing non-linear relationships between food intake customer demographics and spending habits. Moreover, the findings by Chaves *et al.* (2021) emphasised that the capability of the DT model to express possible results of a series of choices related to attributes (customer demographics and historical spending habits) through pre-defined criteria makes it a preferable choice for predicting food intake of customers.

The paper can be partially linked with the motive of this project as it fails to provide detailed insights about customer spending on food products; instead, it emphasises aspects of the food intake preferences of customers. The good part of the paper includes a detailed discussion of ML algorithms and computational tools for predicting food consumption. On the other hand, limitations include the non-existence of the predictive power of the ML for predicting the food intake of customers and their possible applications in predicting the food preferences of customers.

**Paper 2:** “Predicting future consumer purchases in grocery retailing with the condensed Poisson lognormal model” (Trinh and Wright, 2022)

The aim of this study was to predict future consumer purchases in grocery retailing based on attributes like previous purchase behaviour and customer demographics. The study by Trinh and Wright (2022) utilised two grocery retailing datasets from the UK. It focused on developing new models (a mixture of Erlang-2, Poisson distributions or a condensed Poisson lognormal model (CPLN)) for predicting future consumer purchase amounts in grocery retailing. The CPLN and CNBD models imply that they have reduced the errors in the estimation of customer spending on grocery retailing by approximately 50% (7%) and 67% (8%), respectively, compared to benchmark models such as the Decision Tree (14%) or Linear regression (16%) (Trinh and Wright, 2022). The study's strengths included a detailed discussion of the methodological framework followed and the implications of the new approach for predicting customer spending on grocery retailing. Additionally, theoretical and practical implications for retailers have been discussed in this paper, which makes this study suitable for evaluating the implications of the models within the field of retail management. On the other hand, the limitations included a need for more focus on traditional models such as DT, KNN, and SVM, which made the comparative analysis inadequate for this study.

**Paper 3:** “Customer Value Types Predicting Consumer Behavior at Dutch Grocery Retailers”(Janssens *et al.*, 2020)

This paper aimed to predict customer value types (based on three different aspects: Satisfaction, repurchase intention, and word-of-mouth) within the grocery retail sector in the Netherlands. The study by Janssens *et al.* (2020) used survey questionnaires as the research instrument. It developed 'Partial least squares structural equation modelling (PLS-SEM)' for predicting factors affecting the value types of customers using SPSS Statistical Software. The results obtained from this study revealed that factors like low prices, attractive bargains, and easy access to products for customers have a substantial positive influence on high customer spending. This study is relevant to this research as it provided a theoretical base for evaluating factors affecting customer spending on food products. The limitations include focusing more on customer spending on food products. Instead, this study has focused on customer value types. Moreover, the implication of ML algorithms for predicting customer spending must be included, which can show the approach's limitations in predicting customer behaviour based on large volumes of customer spending data. On the contrary, the strengths of this study include a detailed discussion of different aspects of customer value types within grocery retail formats (on-discounter, soft-discounter, and hard-discounter).

**Paper 4:** “From intention to action: Predicting purchase behaviour with consumers’ product expectations and perceptions” (Kytö, Virtanen and Mustonen, 2019)

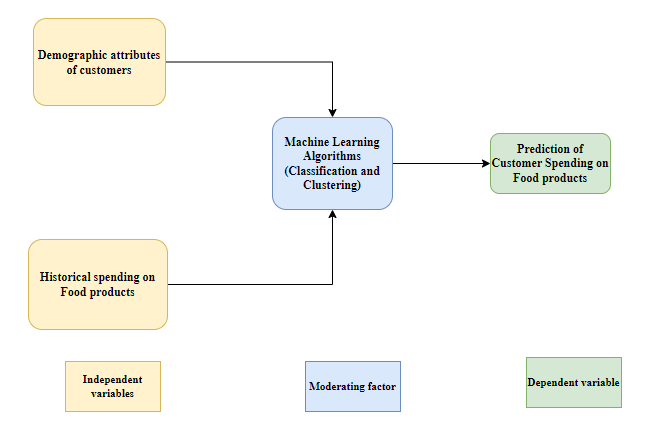
The study by Kytö, Virtanen and Mustonen (2019) explored the relationship between purchase intentions and actual purchase behaviour for two spoonable dairy snack products: natural yoghurt and flavoured protein quark. This paper aimed to evaluate two critical phases of the buying process: expectations based on brand and package picture and perception after tasting the product at home. This study collected data through surveys involving two groups of consumers: users of natural yoghurt (n = 105) and users of flavoured protein quark (n = 107) from Finland. Based on the collected data from a survey of customers from the Food sector in Finland, predictive modelling (Logistic Regression) was performed to assess how well purchase intention ratings (both expectation and perception) could forecast actual purchase behaviour. Logistic regression models were developed to classify buyers and non-buyers for both product types, and an accuracy of 67.3% for yoghurt and 72.3% for quark was obtained from the model. The findings of this can be relatable to this study as it has explored consumer purchase behaviour in the context of food products, providing insights into how intention and perception influence actual purchases.

The main strengths of this study include the incorporation of multiple surveys and follow-ups with surveyed customers, ensuring a longitudinal view of consumer behaviour, which strengthens the predictive accuracy of the models. On the other hand, the limitations of this study include the need for more focus on diverse food products, which can limit the generalisability of the model in predicting food consumption.

## 2.4 Comparison of the papers

The paper by Chaves et al. (2021) has provided a detailed view of the implication of Ml algorithms on the prediction of food intake of customers based on the SLR approach. On the other hand, the study by Trinh and Wright (2022) and Kytö et al. (2019) have developed different state-of-the-art ML algorithms such as DT, KNN, and SVM along with advanced techniques like condensed Poisson lognormal model (CPLN)) to predict customer behaviour. The results revealed that integrating a large volume of customer demographic data and historical consumer spending allows businesses to predict customer spending on food products, enabling them to streamline their inventory level. Based on the comparative analysis, it can be stated that the study by Trinh and Wright (2022) has provided better insights related to customers' spending on retail food products.

## 2.5 Conceptual Framework

******

***Figure 2: Conceptual Framework***

## 2.6 Literature Gaps

Findings of past studies have primarily focused on evaluating customer purchase intention and value types within the retail food sector. The studies by Trinh and Wright (2022), Kytö et al. (2019), and Chaves et al. (2021) have shown a lack of focus on the prediction of customer spending on food products. Moreover, these studies have not adequately provided a discussion of previous spending on food products and their impact on future spending. In previous studies, the segmentation of customers based on purchasing power and demographics has not been adequately explored, reflecting a potential gap in past literature. This study has focused explicitly on the implication of ML (Classification and clustering) techniques for evaluating customer spending on food products.

## 2.7 Summary

This chapter has helped me understand the critical research on customer spending on food products using machine learning. Machine learning, which includes both classification and clustering, has not been done before, which has helped motivate the research.

# Chapter 3: Data

## 3.1 Data collection process

The research data has been sourced through secondary data collection from the UCI Machine Learning Repository. The ***“Wholesale Customers”*** dataset involves March 30 expenditure data across a variety of product classes by wholesale distributors' customers (UCI Machine Learning Repository, 2024). Such a dataset, having historical expenditure data in several attributes, allows machine learning techniques to be modelled to predict the spending behaviour of consumers in retail food.

## 3.2 Dataset description

The "Wholesale Customers" dataset consists of 440 instances, with each instance representing the annual spend of a customer on six categories: Fresh, Milk, Grocery, Frozen, Detergents\_Paper, and Delicassen. Additionally, the dataset comprises two categorical variables: Channel, which includes Horeca or Retail, and Region, which is the dependent variable, classified into Lisbon, Oporto, or Other (UCI Machine Learning Repository, 2024). Since each spending attribute is measured as an integer in monetary units, the dataset gives an excellent resolution for customers' purchasing patterns across different types of products, making the dataset suitable for classification and clustering tasks in machine learning.

## 3.3 Rationale for choosing the dataset

I have chosen the dataset because it is relevant and valuable in analysing customer spending within the retail food sector, aligning with the research objectives to evaluate consumer spending patterns. Prior studies have proved that economic status and consumer expenditure trends significantly impact the production of food products (Ali and Ali, 2020). The features are categorical and continuous in the dataset, which are helpful in the analysis, and the customer data collected from the real world are genuine to building up machine learning models. Due to its nature, quality, and capacity to deliver on the research objectives, it is well suited to predictive modelling for customer expenditure studies.

## 3.4 Data preprocessing

## 3.5 EDA

# 

# Chapter 4: Ethical Issues

## 4.1 Dataset ethics

The dataset has maintained ethical standards by encrypting data anonymisation as it contains no Personally Identifiable Information (PII) about individual customers. This is important for concealing the data privacies or the customers' identities, and this aligns well with the ethics involved in data processing or analyses. This data is fully available to the public through the UCI Machine Learning repository to preserve its relevance and proper usage for scientific purposes.

## 4.2 Research ethics

Adhering to ethical standards in any research has been one of the most significant factors in any research endeavour. This research followed the rules of the Copyright, Patents and Designs Act (1988) in the sense that all the sources used in obtaining the current information were duly acknowledged so that no traces of plagiarism were seen (Government of UK, 1988). Besides, the research has concluded the fair use principle because this dataset has been used mainly for research and not for commercial purposes. The research has not infringed on copyright or patent laws because it respects the rights of data owners and creators, hence keeping the integrity of the dataset intact.

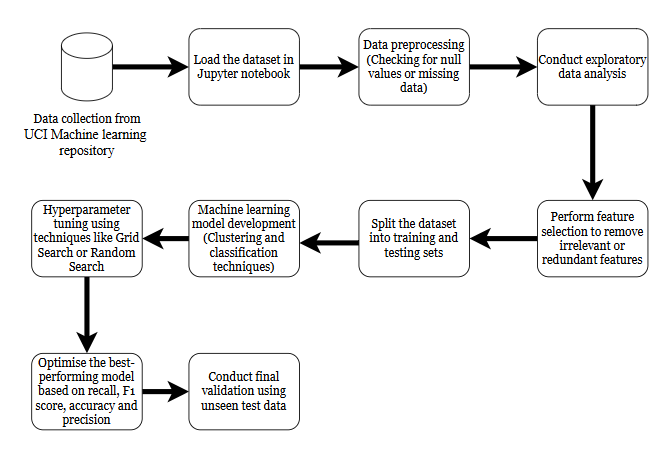
# 

# Chapter 5: Methodology

## 5.1 Introduction

The methodology chapter gives an overview of the research approach and further describes the techniques and tools that have been used to evaluate customers' spending on food products using machine learning. It ensures that the overall methodology covers data preprocessing, feature selection, model development, and performance evaluation to correctly identify customers' spending patterns across different regions.

## 5.2 Proposed Methodological Architecture



***Figure 3: Proposed methodological architecture***

## 5.3 Justification of the Methodological Architecture

The approach I developed for this research in the form of a methodological structure has been created to provide a systematic approach to analysing customer spending data through machine learning techniques. The approach begins with data collection obtained from the UCI machine learning repository as it allows the use of a well-validated and preprocessed dataset that is available for public use and commonly used within the field of data science ***(Refer to Figure 3)***. Data preprocessing follows to handle null values and missing data, which is necessary for ensuring data quality and avoiding biases while training the ML models (Karrar, 2022). Exploratory data analysis (EDA) is done to understand the relationship between the features and identify potential outliers, which is significant in discovering data patterns (Rao, Vardhan and Shaik, 2021). The dataset is then split into training and testing datasets where cross-validation could be performed. Feature selection is one of the ways of including it in the model to select only the most relevant features to improve the performance and interpretability of the model, as shown by Pudjihartono *et al.* (2022). This structured methodology aligns with the best machine learning model development practices and has robust, reliable results.

## 5.4 Software and Libraries Used

Python is the primary programming language to analyse the collected data within the Jupyter Notebook Interface. Python has been chosen for its capability, widely available resources to handle big data sets and integration of machine learning algorithms. According to Samuel and Mietchen (2024), the Jupyter Notebook provides an interactive platform wherein a user mainly writes the code and displays the output and also provides a way of explaining the work done and then building models, modifying and testing them. In addition, I have used several important libraries for data analysis and machine learning model development based on the secondary data. These include NumPy and Pandas, which have effectively manipulated and handled numerical and tabular data. Matplotlib and Seaborn have been used for the data visualisation to check for any correlation or trend between the variables being analysed. In sci-kit-learn, functions exist for encoding labels and splitting data into train and test sets, and they are used to preprocess the given data (Hao and Ho, 2019). The same has been used in implementing machine learning algorithms like K-means clustering and classification models.

## 5.5 Machine Learning models used and its justification

The research uses both cluster-based and classification analysis, which provide a strong understanding of customer expenditure. Clustering focuses on defining a particular set of customers whose behaviour is similar in terms of spending patterns. It requires a proper approach to marketing and customer segmentations in specific regions. Classification involves using the model to predict the target variable Region, which would help classify the customers depending on this difference in their spending behaviour across regions. On the one hand, the clustering analysis has helped me to determine what kind of spending behaviour towards various products across the regions belonging to a particular cluster. At the same time, classification enables me to predict customer types, thus helping to decide on targeting and resource allocation in regions based on data.

|  |  |
| --- | --- |
| **ML models** | **Justification for selection** |
| K-means Clustering | It is efficient for large datasets and clearly segments customers based on spending attributes. |
| Hierarchical Clustering | It facilitates cluster hierarchy exploration and does not necessarily require pre-specifying the number of clusters. |
| Random Forest Classifier | It is highly accurate and robust in dealing with noisy and complex data, making it valuable for classification tasks that generate many trees. |
| Bagging Classifier | It reduces variance; hence, the model has more stability and reduces overfitting in prediction. |
| AdaBoost Classifier | This is suitable for handling imbalanced data and enhancing prediction accuracy by focusing on errors in previous rounds. |
| XGBoost Classifier | They are known for speed and performance. Because of regularisation and hyperparameter tuning, they are often more efficient than other classifiers, which also addresses the concern of overfitting. |
| CatBoost Classifier | It can handle categorical data and is effective against overfitting, making it fit for complex, real-world datasets. |

***Table 1: Rationale for the selection of ML models***

## 5.6 Conclusion

This chapter, therefore, combines the approach and process undertaken to execute the data for further analysis using Python programming. The rationale for selecting the ML models again strengthens the foundation for data analysis and machine learning model formation. By utilising classification and clustering methods and hyperparameter fine-tuning, the approach is targeted to deliver insightful information regarding customer behaviour to be a stand for data-driven retail food business decisions.

# References

Ali, T. and Ali, J. (2020). 'Factors affecting consumers' willingness to pay for health and wellness food products', *Journal of Agriculture and Food Research*, 2(100076), p. 100076. doi: https://doi.org/10.1016/j.jafr.2020.100076.

Cai, J. (2023). 'Research on the Influencing Factors of Consumer Buying Behavior’, *Highlights in Science Engineering and Technology*, 61, pp. 119–127. doi https://doi.org/10.54097/hset.v61i.10281.

Chaves, L. O. *et al.* (2021). 'Applicability of machine learning techniques in food intake assessment: A systematic review', *Critical Reviews in Food Science and Nutrition*, 63(7), pp. 1–18. doi: https://doi.org/10.1080/10408398.2021.1956425.

Christy, A. J. *et al.* (2018). 'RFM Ranking – an Effective Approach to Customer Segmentation', *Journal of King Saud University - Computer and Information Sciences*, 33(10). doi https://doi.org/10.1016/j.jksuci.2018.09.004.

Cognitive Market Research (2023) *Consumer Food and Beverages Market Research Reports*, *Cognitive Market Research*. www.cognitivemarketresearch.com. Available at: https://www.cognitivemarketresearch.com/list/food-%26-beverages/consumer-food-and-beverages (Accessed: October 23 2024).

Daroch, October 23, G. and Gupta, A. (2021) 'A study on factors limiting online shopping behaviour of consumers', *Rajagiri Management Journal*, 15(1), pp. 39–52. doi: https://doi.org/10.1108/ramj-07-2020-0038.

Government of UK (1988) *Copyright, Designs and Patents Act 1988*, *Legislation.gov.uk*. www.legislation.gov.uk. Available at: https://www.legislation.gov.uk/ukpga/1988/48/contents.

Hao, J. & Ho, T. K. (2019). 'Machine Learning Made Easy: A Review of Scikit-learn Package in Python Programming Language', *Journal of Educational and Behavioral Statistics*, 44(3), p. 107699861983224. doi: https://doi.org/10.3102/1076998619832248.

Heldt, R., Silveira, C. S. and Luce, F. B. (2019). 'Predicting customer value per product: From RFM to RFM/P', *Journal of Business Research*, 127, pp. 444–453. doi https://doi.org/10.1016/j.jbusres.2019.05.001.

Janssens, K. *et al.* (2020). 'Customer Value Types Predicting Consumer Behavior at Dutch Grocery Retailers’, *Behavioral Sciences*, 10(8), p. 127. doi: https://doi.org/10.3390/bs10080127.

Karrar, A. E. (2022). 'The Effect of Using Data Preprocessing by Imputations in Handling Missing Values', *Indonesian Journal of Electrical Engineering and Informatics (IJEEI)*, 10(2). doi https://doi.org/10.52549/ijeei.v10i2.3730.

Kytö, E., Virtanen, M. and Mustonen, S. (2019). 'From intention to action: Predicting purchase behaviour with consumers’ product expectations and perceptions, and their individual properties’, *Food Quality and Preference*, 75, pp. 1–9. doi https://doi.org/10.1016/j.foodqual.2019.02.002.

Pudjihartono, N. *et al.* (2022). 'A Review of Feature Selection Methods for Machine Learning-Based Disease Risk Prediction', *Frontiers in Bioinformatics*, 2. doi: https://doi.org/10.3389/fbinf.2022.927312.

Rao, A. S., Vardhan, B. V. & Shaik, H. (2021). *Role of Exploratory Data Analysis in Data Science*, *IEEE Xplore*. doi https://doi.org/10.1109/ICCES51350.2021.9488986.

Samuel, S. and Mietchen, D. (2024). 'Computational reproducibility of Jupyter notebooks from biomedical publications’, *GigaScience*, 13. doi: https://doi.org/10.1093/gigascience/giad113.

Statista (2024) *Food - Worldwide | Statista Market Forecast*, *Statista*. www.statista.com. Available at: https://www.statista.com/outlook/cmo/food/worldwide (Accessed: October 23 2024).

Trinh, October 23ght, M. J. (2022) 'Predicting future consumer purchases in grocery retailing with the condensed Poisson lognormal model', *Journal of Retailing and Consumer Services*, 64, p. 102809. doi <https://doi.org/10.1016/j.jretconser.2021.102809>.

UCI Machine Learning Repository (2024). *UCI Machine Learning Repository*, *archive.ics.uci.edu*. Available at: https://archive.ics.uci.edu/dataset/292/wholesale+customers (Accessed: November 12 2024).