International Journal of Information Management Data Insights 5 (2025) 100303

Contents lists available at ScienceDirect

International Journal of Information Management Data Insights

journal homepage: www.elsevier.com/locate/jjimei

# A machine learning algorithm for personalized healthy and sustainable grocery product recommendations

Laura Z.H. Jansen [a,b,*], Kwabena E. Bennin [b]

[a] Marketing and Consumer Behavior Group, Wageningen University & Research, Hollandseweg 1, 6706 KN Wageningen, the Netherlands
[b] Information Technology Group, Wageningen University & Research, Hollandseweg 1, 6706 KN Wageningen, the Netherlands

ARTICLE INFO

*Keywords:*
Food recommender system
Product suggestions
Groceries
Sustainability
Healthiness

ABSTRACT

Nowadays, retailers try to optimize the shopping experience for consumers by offering personalized services. Recommending food options, i.e. providing consumers suggestions on what products to buy, is one of such services. Food recommender systems for grocery shopping are typically preference-based, using consumers' shopping history to determine what products they would like. These systems can predict well what a consumer would potentially like to buy, however, they do not stimulate consumers to buy healthier or more sustainable food options. In response to increasing global concerns about public health and sustainability, this paper aims to integrate healthiness and sustainability levels of food options in recommender systems to encourage consumers to buy better food options. To assess the impact of integrating healthiness and sustainability information of food choices in predicting an item to buy, we employ three food recommendation models: a Baseline popularity-based model, Restricted Boltzmann Machine (RBM), and Variational Bayesian Context-Aware Representation (VBCAR) based on (1) preferences, (2) preferences and health, (3) preferences and sustainability, and (4) all combined attributes. Models were trained and tested using two different datasets: Instacart and a Dutch supermarket dataset. The experimental results indicate improved performance for VBCAR compared to Baseline and RBM. Models that emphasize healthiness and/or sustainability of food choices do not significantly alter model performance compared to preference-based models. The results of the health and sustainability-based recommender systems demonstrate the potential of recommender systems to assist people in finding healthier and more sustainable products that are also suited to their preferences.

## 1. Introduction

Recommender systems navigate consumers through the overwhelming amount of product options by suggesting alternative products that match the needs of a specific consumer (Alamdari et al., 2020, Ricci et al., 2011). Leveraging such technology for digital transformation can increase customer satisfaction in the digital landscape (Abbu et al., 2021), which is beneficial now that more and more consumers are shifting toward the online environment for grocery shopping. To illustrate, Dutch online grocery shopping sales with home delivery quadrupled from 2018 to 2023 (Rabobank, 2023), specifically accelerated by the COVID-19 pandemic (Baarsma and Groenewegen, 2021). Expectations are that the Dutch online grocery market revenue will surpass 6 billion euros, doubling its value compared to 2022 (Statista, 2024). In this online grocery environment, consumers typically get recommended products based on their preferences that reflect what they have liked in the past (Trattner and Elsweiler, 2017). Many algorithms have been used to provide accurate items to consumers (see for example the article by Portugal et al. (2018) for an overview).

Although providing suggestions based on preferences can be helpful to consumers in finding suitable product alternatives in an online grocery store, these systems are typically designed from a retailer's perspective to drive sales consumption (Smith and Linden, 2017) and suggestions merely reinforce existing eating habits (Starke, 2019). The need to develop recommender systems that go beyond consumer preferences and earlier purchases is important in the context of healthy and sustainable eating behavior, given the rise in obesity levels and the urgent need to combat climate change (Willett et al., 2019). Nowadays, information is available on the attributes that determine whether a product is healthier or more sustainable such as low sugar levels and organic farming. Still, a bottleneck prevails in altering consumer behavior toward lasting change. Various factors contribute to this

* Corresponding author.
*E-mail address:* laura1.jansen@wur.nl (L.Z.H. Jansen).

https://doi.org/10.1016/j.jjimei.2024.100303

Available online 20 November 2024

bottleneck, such as cultural and social contexts, convenience, costs, and the food environment (Munt et al., 2017). Consequently, consumers often lack the knowledge and motivation to choose healthily or sustainable (Hollywood et al., 2013, Munt et al., 2017) and would benefit from guidance in finding more healthy and sustainable alternatives to change their eating behavior (Jansen et al., 2023).

It is crucial to guide consumers in achieving broader goals of healthy and sustainable eating, which are eating behaviors that "promote all dimensions of individuals' health and wellbeing; have low environmental pressure and impact; are accessible, affordable, safe and equitable; and are culturally acceptable" (FAO, WHO 2019, p.9). Nevertheless, preference-based recommender systems primarily focus on taste and personal preferences, without necessarily considering the nutritional value of foods (Chen et al., 2023). This often reinforces unhealthy eating habits, as the systems primarily strengthen existing preferences without encouraging more balanced choices (Chen et al., 2023). By focusing on past behavior, these systems tend to offer a limited variety of recommendations, restricting consumers to a narrow selection of familiar products and missing opportunities to educate individuals about the importance of balanced nutrition (Starke and Trattner, 2021). Education and diversity in offerings might increase consumer's adherence to dietary changes over time (Hauptmann et al., 2022, Starke and Trattner, 2021).

The inclusion of health is becoming more of a hot topic in food recommender systems (FRS) research, for instance in generating meal plans or by substituting ingredients within recipes (Trattner and Elsweiler, 2017). Nonetheless, little research has been conducted on FRS tailored for grocery shoppers, particularly in terms of including health and sustainability considerations in these information systems. In the few studies conducted on sustainable grocery recommendations, Asikis (2021) proposed multi-objective optimization algorithms to provide sustainable recommendations but only evaluated the outcomes based on the costs, environmental impact, and nutritional quantities of the basket rather than using the more common offline evaluation metrics such as Precision and Recall (Ghannadrad et al., 2022). Another previous work identified sustainability-minded grocery shoppers and learned from their purchasing patterns to identify sustainable products but did not evaluate a recommender system (Tomkins et al., 2018). Regarding health, Bodike et al. (2020) used a simple collaborative filtering approach to recommend products and improved grocery recommendations by including the nutritional value of food products using natural language processing. Hafez et al. (2021) focused on item-to-item recommendations for healthier grocery shopping without including information on consumer buying behavior. All in all, although some research has focused on integrating health or sustainability attributes of products, developing Grocery Recommender Systems (GRS) that incorporate all attributes - preferences, health, and sustainability - and evaluating these systems with traditional machine learning metrics remains unexplored.

Accordingly, the main objective of this study is to utilize implicit purchase data to examine models that consider the health and sustainability value of food options in addition to consumer preferences when suggesting grocery products to the consumer and to evaluate these models with offline metrics. It is assumed that continuously recommending products that are more sustainable or healthier than products initially chosen by the consumer will improve diet and environmental footprint. We aim to empirically validate various machine learning-based recommendation models while also studying whether including healthiness and/or sustainability information about food choices in information systems impacts model performance compared to preference-based models. Specifically, we aim to answer the following research question: *"How do the proposed models that include healthiness and/or sustainability information of food choices perform compared to existing preference-based models?"*

This research tests three recommendation algorithms for preference-based healthy, sustainable food choices based on real-life shopping basket data. The outcomes of preference-based recommendations are evaluated in comparison to other objectives, including preference and health, preference and sustainability, and a combination of these factors. As a first model, we take a simple Baseline model that recommends the most popular items.[1] Second, a Restricted Boltzmann Machine (RBM) model opted by Salakhutdinov et al. (2007) was used.[2] RBMs learn the probability distribution over input data using two layers: a visible (input) and hidden (learned features) layer, while continuously updating the input data with backward reconstruction (Salakhutdinov et al., 2007). Third, a more up-to-date and state-of-the-art algorithm specifically designed for grocery shopping was modified for our research (VBCAR by Meng et al. (2021)). The VBCAR[3] model predominantly employs Bayesian Skip-gram and Variational Autoencoder techniques to learn low-dimensional embeddings of users and items, while also accounting for contextual factors in purchase behavior, such as the healthiness of products.

In summary, this study contributes to the field of recommender systems by utilizing implicit consumer data to examine health and sustainability-focused recommendation models by evaluating these models with offline metrics. We propose to consider the healthiness and sustainability level of food choices for product recommendations and adjust various models for item prediction by integrating health and sustainability information. Through experiments, we show that results from models incorporating the healthiness and/or sustainability level of food choices do not significantly deter from preference-based model results. Moreover, experiments verify that advanced techniques that capture underlying relationships offer more accurate recommendations than baseline popularity models. The following part of the paper is organized as follows: Section 2 provides a background on FRS with a specific focus on grocery recommendations. Section 3 describes the materials and methods used. Section 4 provides detailed results of our experiments on two grocery datasets. We compare three models (Baseline, RBM, and VBCAR) for preference-based product suggestions and recommendations based on additional product information. Finally, we provide conclusions and recommendations for future work.

## 2. Background

### 2.1. Food recommender systems

Recommender systems that leverage previous buying behavior (customer-product interaction) to suggest useful product options to consumers have gained much attention from both academia and industry for many years (Wei et al., 2007). In recent years, the use of recommender systems has extended to specific fields such as food and, even more specifically, grocery shopping (e.g., Yuan et al., 2016). Within grocery recommendations, these systems aim to predict which grocery products a consumer might buy in the future based on previous transactions. These systems have been implemented on grocery shopping platforms, including major retailers like Kroger (www.kroger.com) and Albert Heijn (www.ah.nl). By implementing these systems, retailers seek to enhance the shopping experience for consumers (Elahi et al., 2021). In doing this, recommender systems influence what information consumers see in the shopping environment and consequently influence decision-making by selection and ranking of choices (Jesse and Jannach, 2021). Benefits for consumers are manifold, such as receiving a narrow consideration set specifically tailored to an individual (Sarwar et al., 2002) or exploring new or additional products (Fayyaz et al., 2020).

Various methods for delivering personalized food recommendations have been researched. The three most common types of filtering methods are content-based, collaborative, and hybrid filtering (Portugal et al., 2018, Ricci et al., 2011). In a content-based system, items will be

[1] https://github.com/microsoft/recommenders/tree/main
[2] https://github.com/microsoft/recommenders/tree/main
[3] https://github.com/mengzaiqiao/VBCAR

recommended to a user based on the content features of items that a user has previously interacted with. Unlike content-based recommendations, collaborative filtering (CF) approaches are not dependent on item data but collect a huge amount of user data to find alternatives. Collaborative filtering is an approach that relies on user-item interactions to identify patterns in behavior such as purchase history and consequently recommend products that are popular among similar users. The rationale is to find products that similar users have bought in the past because similar users are a good indicator of relevancy for the current user. Hybrid filtering methods combine the advantages of content and collaborative filtering methods by processing different data sources. In other words, a hybrid system 'tries to use the advantages of A to fix the disadvantages of B' (Ricci et al., 2011, p.13).

These traditional recommender systems capture static preferences (isolated preferences), whereas providing food recommendations in the context of grocery shopping requires capturing preferences in a sequence over time (Wang et al., 2019). As such, sequential recommender systems (SRS) are often used to model user-item interactions as a dynamic sequence in the grocery shopping context (Wang et al., 2019). Typical examples of these are next item and next basket recommendations, where a model recommends additional items to add to an existing basket or for the next basket, respectively (Katz et al., 2022). Another possibility is recommending options to switch a product in the basket (i.e. recommendations at the basket/checkout page) or at the moment of product selection (i.e. recommendations at the product page), so-called food swaps (Jansen et al., 2021).

Typically, recommender systems create suggestions based on what consumers have bought or liked in the past and thus present information within the bubble of consumer preferences (Jansen et al., 2024, Trattner and Elsweiler, 2017). While preference-based food recommendations can increase consumer satisfaction (Pecune et al., 2020), there are also opportunities to achieve specific goals with recommender systems, such as trying to get people to eat healthier or more sustainably (Jesse and Jannach, 2021). For instance, a recommender system might suggest options that a user will probably like, while, at the same time, highlighting healthier options (Starke et al., 2021).

Rather than a market-driven focus, models can take into account the health and sustainability level of products. In doing this, FRS are not just a tool for matching preferences but can serve as persuasive technologies that influence food choices and behaviors while allowing freedom of choice (Jesse and Jannach, 2021). Incorporating health and sustainability factors into recommender systems can transform those systems into persuasive tools to subtly nudge consumers toward healthier and more sustainable food choices. As posited in Nudge Theory (Thaler and Sunstein, 2009), recommender systems then act as a digital nudge to influence behavior in a way that supports better food choices without restricting consumer choice. Such digital interventions are expected to enhance behavioral change toward healthier and more sustainable food choices (Jansen et al., 2023).

By integrating machine learning, recommender systems not only personalize interventions to better match individual preferences but also offer a powerful tool for supporting consumers in making healthier or more sustainable food choices. Machine learning empowers recommender systems to continuously learn from consumer interactions such as purchase or browsing behavior (Schafer et al., 1999). This allows the systems to identify behavioral patterns important for accurate predictions on what products consumers would like to buy. In this regard, machine learning can shed light on the features of recommendations that are most effective in driving behavioral change.

### 2.2. Grocery recommender systems

This section starts by discussing two types of recommender systems that are closely related to the grocery recommendation problem: (1) next item recommendation, and (2) next basket recommendation. Next item recommendations focus on predicting the next item that a user is likely to buy based on the previous shopping history (Ilyas et al., 2022). For instance, the system might recommend adding butter if a user has bread, milk, and eggs in the basket. Next basket recommendations go one step further and focus on predicting an entire set of items that a user might be interested in buying (Shao et al., 2022). As such, next basket recommendations identify combinations of products that are likely to be bought together by a user.

A typical technique used to develop next item or next basket recommendations is a sequential model based on Markov Chains (Ilyas et al., 2022). Markov Chains (MC) models are an improvement over the more standard models such as Collaborative Filtering because MC models can include sequential behavior rather than simply predicting based on the complete purchase history of a consumer (Chen and li, 2021). An example is Factorizing Personalized Markov Chains (FPMC), which is a model that leverages matrix factorization and MC to model users' general interests and basket transition relations (Rendle et al., 2010). The downside of MC techniques is that they predict future behavior based on only the last or few last behaviors (items or baskets bought) (Ilyas et al., 2022, Wang et al., 2020).

For both next item and next basket recommendations, there is a growing emphasis on deep learning techniques because of the ability to learn longer sequential and temporal information (Ilyas et al., 2022). Improved performance based on deep learning techniques has been shown (e.g., Yu et al., 2016), but a recent previous study also showed that deep learning techniques might not be more effective than simple frequency-based models (Li et al., 2023). Deep learning models are useful for exploring and finding hidden patterns in data, but not so much for finding repeated items that play a substantial role in grocery shopping recommendations (Li et al., 2023).

Aside from typical techniques underlying recommender systems, item and basket recommendations often revolve around a widely accepted objective: consumer preferences (e.g., Bai et al., 2019; Fouad, Hussein, Rady, Yu, & Gharib, 2022; Hoang & Le, 2021). This is often also reflected in the definition of these recommendations. For example, Shao et al. (2022) define a next basket recommender system as a system that aims to predict a user's next basket by modeling the user's preferences based on the shopping history. As a result, recommender systems often promote unhealthy and unsustainable alternatives that align with consumer's previous buying behavior. Previous work in the FRS domain showed some positive results when recommending healthier recipes (e.g., Starke and Trattner, 2021) or diets (e.g., Yang et al., 2017), but only to a limited extent. To our knowledge, none of the existing GRS has investigated leveraging health and sustainability information of food choices to improve model performance compared to the typical preference-based models.

While the challenge of integrating health and sustainability of food choices as additional information in recommender systems is specific to the food domain, various other, more generic, challenges should also be considered. Most important are the amount and the sparseness of the data (Bodike et al., 2020). Supermarkets often have a huge product database, of which some products are never or hardly bought by consumers, which makes it hard to find users who purchased the same items (Bodike et al., 2020). Many approaches cannot handle such large and sparse datasets, but Restricted Boltzmann Machines (RBM) present a solution (Salakhutdinov et al., 2007). RBM models learn compressed representations of features in training data, which reduces the dimensionality of data. Regarding the health and sustainability of food options, RBM models offer the possibility to integrate as much item feature information as one would want, as the models can take low-level features from the items as input.

Bayesian techniques can also be leveraged to integrate prior information based on large datasets (Abdar et al., 2021, Ansari et al., 2018), while incorporating different types of information such as healthiness and sustainability of food choices in one coherent model (Condliff et al., 1999). Such deep learning methods can consider semantic information well and this semantic information leads to model improvement with

more input (Dong et al., 2017). In this context, skip-gram-based models have shown promise to optimally understand relationships between items and to improve predictions, using word embedding techniques (Meng et al., 2021, Mikolov et al., 2013). However, skip-gram-based methods have yet to be explored for healthier and more sustainable grocery shopping.

In this paper, three models are used to address the challenges of incorporating preferences, health, and sustainability factors into grocery recommendations. A Baseline model provides a simple reference point for comparison, while two other models offer increasingly sophisticated approaches for capturing the multifaceted nature of grocery decision-making and generating personalized recommendations that promote health and sustainability in addition to preferences. An RBM model is chosen for its ability to capture complex patterns and interactions within high-dimensional data. VBCAR is selected because it combines collaborative filtering techniques and variational Bayesian inference and regularization mechanisms to leverage user preferences and auxiliary information regarding health and sustainability information in a principled and probabilistic framework.

## 3. Methodology

### 3.1. Data collection and preprocessing

#### 3.1.1. The datasets

Two datasets were selected for our analysis: the first one is a publicly available benchmark dataset[4], and the second one is a recent real-life grocery shopping dataset that contains all the required information on health and sustainability features. This combination allowed us to validate the models using comprehensive real-world data while also comparing the results with an established benchmark dataset.

The first dataset used was the Instacart dataset,[4] which is publicly available and often used in research on recommender systems (e.g., Ariannezhad et al., 2022, Faggioli et al., 2020, Li et al., 2023). Instacart is a web service that provides grocery delivery in the US. The anonymized data file comprised six files containing information on a consumer's order, such as order ID, product ID, aisles, and departments. Data was collected in 2017 from more than 200,000 customers, containing more than 3 million grocery orders of 50,000 items. The specific date of each order was missing, but the order of transactions was provided for each user. The dataset also gave information on the day of the week and the hour of the day the order was placed.

The second dataset was provided by a large Dutch supermarket chain, of which data from September until December 2021 was used in this paper. Purchase data included all purchases from customers who made, on average, one order per week in the period of data collection. Moreover, separate datafiles for each week with pricing data were provided, as well as information on the active price promotions each week and lists of products with their corresponding healthiness (3740 products) and sustainability (1293 products) information. The healthiness of products was based on the United Kingdom Food Standard Agency (FSA) score and the corresponding Nutri-Score label (Santé Publique France, 2023), which has been used in previous work on food recommendations (e.g., Starke et al., 2021). The Nutri-Score label was shown on the products during data collection (September – December 2021). The sustainability of products was defined by the European Union organic label, a certificate indicating that EU rules on organic farming in terms of production, processing, transportation, and storage were respected (European Commission, 2023).

#### 3.1.2. Pretreatment of data

*3.1.2.1. Instacart data. Instacart - integration of multiple data sources:* The different files of Instacart data (i.e., order_products_prior, orders, products, departments, aisles) were merged into one file based on a unique identifier (e.g., order_id in order_products_prior and orders), leading to one dataset with information on among others order, product, user, product description, aisle, department, and whether the dataset belonged to prior, train, or test set. The final purchase dataset contained over 32 million rows of data.

*Instacart - data transformation:* By default, the Instacart data lacked health and sustainability features. Thus, two variables for the health and sustainability level of food choices were generated and added by randomly assigning values. For NutriScore, values 0 (no NutriScore label) and values 1 to 5 (NutriScore A to E) were assigned to products based on the distribution in the Dutch supermarket dataset (see Section 3.1.3). Organic labels were assigned to 50% of the products (organic=1 versus non-organic=0). Moreover, a column for quantity bought was created by randomly assigning values between 1 and 4 products bought, with 50% assigned 1, 20% assigned either 2 or 3 products, and 10% assigned 4 products bought. A variable for ratings was added by equally assigning values 1 to 5 to each product randomly. For easy data processing, all cells and headers were transformed into capital letters.

*Instacart - data cleaning:* Three data cleaning steps were followed (Bauer et al., 2023): (1) remove users with too few baskets, (2) remove items that appear in too few baskets, or (3) remove too small baskets. Instacart data was cleaned by setting the threshold at 3 for the number of purchases a user should have made, how often an item should have been bought, and how many products a basket should contain.

*Instacart - data reduction:* Following Meng et al. (2021), data was selected by randomly taking 25% of the data to allow the merging of data files. To overcome computational problems, a random subset of 30,000 rows (referred to as a small dataset) or 143,000 (referred to as a large dataset) was used for model training and testing. Table A1 explains the different variables and one example of the merged Instacart dataset.

*3.1.2.2. Dutch supermarket data. Dutch supermarket - integration of multiple data sources:* Six different types of datasets were provided by the Dutch supermarket: (1) Purchase Data, (2) Price Data, (3) NutriScore, (4) Organic Label, (5) Category Structure, and (6) Promotions.[5] Purchase Data was provided in 39 separate datasets (September to December 2021), which were merged for data analysis. Price Data was supplied in separate weekly datasets, which required merging for data analysis. Separate files were provided for NutriScore, Organic Label, Category Structure, and Promotions. To enable merging, week numbers and year (2021) were added to the Purchase Data based on the day of purchase. With a left join, NutriScore, Organic Label, and Category Structure were merged into the Price Data. After that, a left join merged this data into the Purchase Data, while also adding a column to indicate whether a product was on promotion using the Promotions data. The purchase data size for September to December 2021 was around 23 million rows.

*Dutch supermarket - data transformation:* NutriScore and Organic labels were replaced with numerical values (1 to 5 for NutriScore A to E, and a value of 0 for missing labels. For organic, the presence of label was indicated with a 1 and no label with a 0). A variable for product ratings, required for running the Baseline and RBM models, was added by equally assigning values 1 to 5 to each product randomly.

*Dutch supermarket – data cleaning:* Anomalies in the data were removed. For instance, headers of the purchase files were repeated every 10,000 rows and had to be deleted. Moreover, cells and headers were

[4] http://www.instacart.com/datasets/grocery-shopping-2017

[5] Note that information about price, category structure, or promotions were not used in the current paper.

converted into capital letters for easy processing, leading and trailing spaces were stripped, and postal codes in a different format than Dutch standard (e.g., NNNNLL) were removed. Duplicates were removed and missing data was replaced with NaN for textual data (e.g., NutriScore label), the mean value for numerical data (e.g., price), or zero ('0') for numerical data that was not measurable (e.g., PE ART. NR. Number). Rows were deleted if the article number (CE ART. NR.) was not present. The full dataset was cleaned for infrequent users, items, and baskets using the same procedure as Instacart data with a threshold of 3.

*Dutch supermarket – data reduction:* To allow the merging of datasets, the purchase dataset was reduced to 25%, in line with the data reduction procedure for Instacart. In line with Instacart, a random subset of 30,000 rows (referred to as small dataset) and 143,000 rows (referred to as large dataset) was used for model training and testing. Table A2 explains the different variables in the Dutch supermarket data and one example of the merged dataset.

#### 3.1.3. Description of data

Table 1 displays the statistics of both datasets after pre-processing, including the number of transactions, items, consumers, main categories, the average number of items in a basket (transaction), and the average number of baskets for a user. The Instacart dataset (25%) included 89,685 customers, 603,677 orders, and 41,476 products divided into 21 main categories. The average basket contained 11 products per basket and each user bought on average 7 baskets. Most products were ordered from the categories 'Produce', 'Dairy eggs', 'Snacks', and 'Beverages', and the most ordered products were bananas, strawberries, spinach, and avocado. NutriScore and Organic information were divided among the sample using the division of the Dutch supermarket data for NutriScore and Organic. The Dutch supermarket data was utilized as a reference point for external extrapolation of NutriScore and organic information, which provides a more credible basis than random assignment.

The Dutch supermarket data (25%) had around 14 thousand customers who ordered 47,397 times in total. The 12,818 products were divided into 18 main categories such as 'potatoes, vegetables and fruits' and 'candy, cookies, chips, and nuts'. Customers shopped around 3 times on average and had an average of 32 products in their baskets during the data collection period (September to December 2021). NutriScore information was available for 2852 products. The division of NutriScore labels A to E was as follows: A (weight 0.07), B (weight 0.03), C (weight 0.03), D (weight 0.05), and E (weight 0.03). Originally, organic information was present on 581 out of 12,818 products. For analysis, 50% of the products were randomly given an organic label.

**Table 1**
Statistics of the datasets.

| | Instacart | Dutch supermarket |
|---|---|---|
| **Transactions (total orders)** | 603,677 | 47,397 |
| **Products** | 41,476 | 12,818 |
| **Customers** | 89,685 | 14,005 |
| **Main categories** | 21 | 18 |
| **Average size of basket (products per order)** | 11.21 | 32.65 |
| **Average baskets per user** | 6.73 | 3.38 |
| **Products with NutriScore (health)** *Weight* | | |
| - **NutriScore: no label** (0.78) | 32,330 | 9966 |
| - **NutriScore A** (0.08) | 3321 | 955 |
| - **NutriScore B** (0.03) | 1273 | 352 |
| - **NutriScore C** (0.03) | 1225 | 432 |
| - **NutriScore D** (0.05) | 2073 | 714 |
| - **NutriScore E** (0.03) | 1254 | 399 |
| **Products with Organic label (sustainability)** | | |
| - **Organic: no label** | 20,702 | 6473 |
| - **Organic: label** | 20,774 | 6345 |

### 3.2. Experimental design

#### 3.2.1. Algorithm construction

Algorithm construction consisted of three phases (see Fig. 1): (1) identify the user preferences, (2) classify the products according to their health and/or sustainability level, and (3) recommend suitable products. In other words, the algorithm aimed to provide food recommendations based on user preferences and product attributes (health and/or sustainability level).

***Phase 1: User preference classifier -*** User preferences were retrieved by analyzing the two datasets encompassing previous buying behavior of consumers. To determine consumers' preferences, a target item that was bought by a certain consumer was taken to identify the products that were most frequently bought together with this target item using the behavior of other users who have interacted with the target item.

***Phase 2: Product healthiness and sustainability classifier -*** After understanding user preferences, the algorithm must assess a product's sustainability or health level. The Dutch supermarket data measured the healthiness score of products with the NutriScore label. The NutriScore label is based on the nutrient profiling (NP) system of the UK Food Standards Agency (FSA) (Food Standards Agency, 2011). The FSA score is a scoring system to determine the nutrient content per 100 g of food or drinks (Hagmann & Siegrist, 2020). Negative and positive points are given to components of the food product, leading to a 5-level score from dark green (healthy) to red (unhealthy) with a value ranging from -15 (healthy) to +40 (unhealthy). To assess whether a product was organic, the EU organic farming classification was used. These products comply with strict regulations from the EU on production, transportation, and storage (European Commission, n.d.).

***Phase 3: Recommender system*** – User input (previous buying behavior) and product input (healthiness and sustainability score) were used to recommend products that are tailored to the user. In the case of retail, there is a set of users U and a set of items I and each user purchases a subset of these items, i.e., the basket. A basket is defined as a set of items such that $B = \{x_1, x_2, \ldots x_n\}$ where $x_i \in I$ denotes an item from itemset I. Given the purchase history, the goal of the recommender system is to predict other products to buy based on the consumer's baskets, whilst constraining the prediction for the healthiness or sustainability level of the product. This is a multi-objective recommender system that tries to optimize the algorithm by simultaneously ranking a set of potential items according to several criteria: user preferences, health, and/or sustainability.

#### 3.2.2. The recommendation algorithms

Three machine learning models that identify recommendable products based on the product's match with consumer preferences and the healthiness or sustainability level of the product were empirically assessed (Ricci et al., 2011). Identifying (a set of) products to recommend is a collaborative filtering problem, aimed at finding similar users or items based on the purchase history of a consumer, that is, the user interaction with products. Aside from integrating preference information, the models were also tested for considering the health and/or sustainability information of products. As such, each of the three models had four versions: a basic version that only included preferences and three other versions that added healthiness information, sustainability information, and both health and sustainability information of products. Two machine learning models, RBM and VBCAR, were compared to the Baseline performance of recommending the most popular item.

***Restricted Boltzmann Machine (RBM)*** – Restricted Boltzmann Machines are generative neural network models that consist of visible (input) and hidden (feature learning) layers (Ghojogh et al., 2022). The visible layer is the input data, whereas the hidden layer represents features or embeddings of the visible data. In RBM models, the visible layers are connected to hidden layers, and vice versa, but a visible (hidden) layer cannot be connected to another visible (hidden) layer. This means that there is a connection between input data and learned

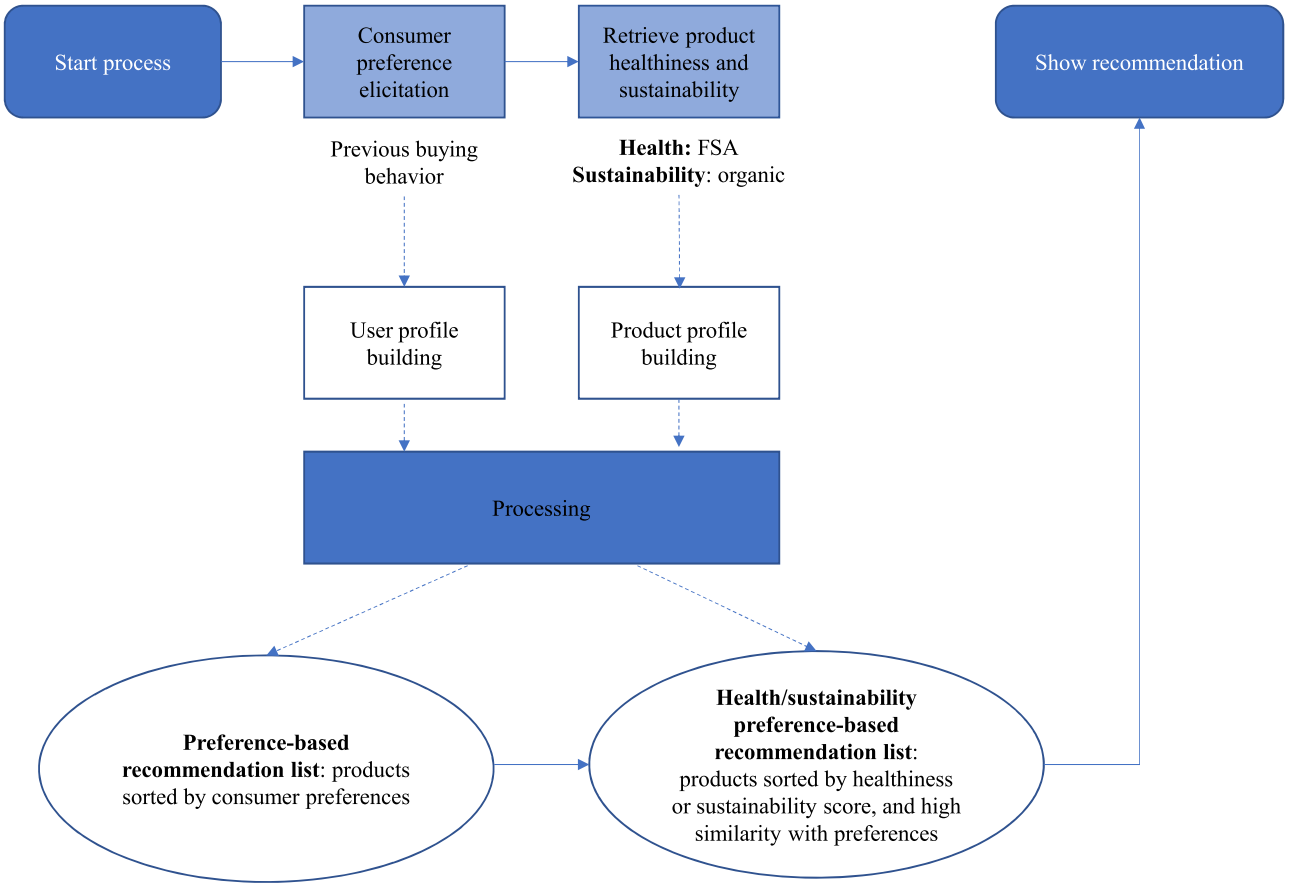


**Fig. 1.** Algorithm construction.

features, but not between the elements of visible and hidden layers themselves. An RBM model's task is to learn the joint probability distribution over input data using visible (input) and hidden (learned features) layers (P(*visible, hidden*)). It does this in two phases: (1) a feed-forward pass in which input is forwarded to feature learning layers, which then adds a bias and sends the value to an activation function, and (2) a feed-backward pass in which the input layers are back trained using activated hidden units, leading to biased reconstructed inputs as output (note that both visible and hidden layers have a bias). Such a model is defined by the following function, where the first term captures the correlation between visible and hidden units and the other terms capture the bias of the units.

$$H = -\sum_{i\ j\ \in G} v_i\ w_{ij}\ h_j - \sum_{i=1}^{m} v_i\ a_i - \sum_{j=1}^{n} h_i\ b_i$$

The model is designed to handle large sparse datasets because it focuses on typical behavior rather than particular instances (Salakhutdinov et al., 2007), which is validated for movie recommendations and is also considered useful for the case of grocery shopping where people usually buy only a few of the available options. For training, the model uses a variant of stochastic gradient descent called contrastive divergence, which means that the weight of connections between neurons is adjusted. To integrate health and/or sustainability in the model, item embeddings were updated with information on the healthiness and sustainability levels of food items during data training. The data are represented using multinomial units, i.e. discrete probability distributions (vectors of probabilities).

***Variational Bayesian Context-Aware Representation (VBCAR)*** - The VBCAR model is specifically designed for grocery recommendations to predict items a user will buy next. The model is optimized for including side information from products, such as the product description or brand (Meng et al., 2021). It uses a Bayesian model to learn low-dimensional representations of users and grocery items, while it can also include additional side information such as product category, or, in our case, the healthiness and sustainability information of products. The underlying model is the Triple2Vec model (Wan et al., 2018), which samples triples of item1, item2, and a user (co-purchase of two items by the same user) to learn the latent embeddings for users and items to predict the occurrence probability of those triples, in combination with a Bayesian Skip-gram model to represent users and items as Gaussian distributions, i.e., as probabilistic quantities (Barkan, 2017):

$$P(\boldsymbol{z}^u) = N(0, \alpha^2 \boldsymbol{I}),\ p(\boldsymbol{z}^i) = N(0, \alpha^2 I),$$

where $\alpha = 1$

To perform an exact inference on the posterior density of variables in the model, Meng et al. (2021) proposed to use a Variational Bayes approach that maximizes a lower bound of the logarithm marginal likelihood (i.e., the probability of observing a particular triple of item, item, user in the data) to infer the embeddings of users and items. In the end, product recommendations are a ranked list of possible items. This model can handle big datasets containing high-dimensional data and is specifically designed to model complex relationships between users, items, and features. To integrate health and/or sustainability information of products, item embeddings were concatenated with the information on healthiness and/or organic aspects of food items in data training, similar to the RBM model. One-hot encoding was used to represent side information.

### 3.2.3. Split technique

All models, Baseline, RBM, and VBCAR, employed a 70/30 split. RBM model used a stratified splitter to randomly select 70% for training

and 30% for testing from the user/affinity matrix. A temporal split strategy was used in the VBCAR model, where baskets were split into training (70%), testing (30%), and validation (the last 20% of training) based on the temporal order of the baskets.

#### 3.2.4. Algorithm performance assessment

In line with Meng et al. (2021), four metrics are reported for the ranking task of product recommendations: precision@k, recall@k, NDCG@k, and MAP@k. These metrics, focused on relevance and ranking of models, are often used for recommender system evaluation and allow for comparison of quality across papers (Bauer et al., 2023, Valcarce et al., 2020). Adopting various metrics to evaluate systems can provide insights into different aspects of the system (Bauer et al., 2023).

*Precision* is the number of correctly recommended items compared to the total number of items. It measures the accuracy of predictions made by a model.

$$\text{Precision} = \frac{|L_k^n \cap R_u|}{n}$$

*Recall* is the number of correctly recommended items compared to the total number of relevant recommended items. It measures the ability of a model to capture all the relevant items.

$$\text{Recall} = \frac{|L_k^n \cap R_u|}{|R_u|}$$

*NDCG* (Normalized Discounted Cumulative Gain) assesses the ranking quality by sorting results on their relative relevance, meaning it assumes that items higher in the list are more relevant. By considering both the relevance and the position of items in a ranked list, NDCG measures the effectiveness of a model in predicting relevant items.

$$\text{NDCG} = \frac{\sum_{k=1}^{n} G(u,n,k)D(k)}{\sum_{k=1}^{n} G^*(u,n,k)D(k)}$$

*MAP* (Mean Average Precision) assesses the ranking quality by checking whether all relevant items are ranked high in the list, meaning it is an average precision at the positions where a relevant item is found. The function below (Average Precision) can be averaged over the whole dataset to get the mean value.

$$\text{AP} = \frac{1}{|R_u|} \sum_{k=1}^{n} \| \left(L_u^n[k] \in R_u\right) P_u @k$$

with $\|$ as indicator function.

Experiments were run on a Windows 11 Pro desktop computer featuring a 64-bit operating system, equipped with a 12-core processor and 32GB of RAM. Results were obtained using Visual Studio Code with Python 3.11.2, including statistical testing in Rstudio. Table 2 gives an overview of the different algorithms and the hyperparameters used. For the full functional architecture of this study, see Fig. 2.

### 3.3. Statistical tests

The current research investigated differences between models (preference-based version of each model), differences when adding additional information (compare preference-based models to models that integrate health, sustainability, or both), and differences between datasets (Instacart and Dutch supermarket datasets and the small and large version of datasets).

First, to analyze the statistical significance of the performance of each algorithm, the Two Samples T-Test was adopted to compare the means of a combination of two models. Independence of data was assumed because, despite 30K and 143K datasets containing identical data, the distinctions between the datasets were not critical when assessing model performance. Second, the same procedure was used to assess whether additional health or sustainability information significantly alters results. The normality of data was confirmed based on the Shapiro-Wilk test and Q-Q plots (Thode, 2002). The F-test for homogeneity of variances was used to assess whether there were significant differences in variance among the models. If non-significant, the Two samples T-test was adopted, whereas the Welch *t*-test was used if variances between the models were significantly different. Due to the limited number of samples, Hedges g was preferred over Cohen's d to calculate effect sizes (Goulet-Pelletier and Cousineau, 2018).

Lastly, for dataset comparison, dependency of data was assumed when comparing the small or the large datasets to each other. To compare datasets from Instacart and the Dutch supermarket to each other, independence of data was assumed. Data was not normally distributed according to the Shapiro-Wilk test (Thode, 2002). Consequently, the homogeneity of variances was checked and confirmed with Levene's test (Nordstokke et al., 2011). Wilcoxon Rank Test or Wilcoxon Rank Sum Test, the non-parametric equivalent of the *t*-test, was performed depending on whether data was paired or not, respectively (Xia, 2020).

**Table 2**
Three models and the hyperparameters.

| Algorithm | Hyperparameters | Values tested |
|---|---|---|
| Baseline | K (items to recommend) | 10 |
| RBM | Number of hidden units | 600 |
| | Number of training epochs | 18 |
| | Minibatch size | 200 |
| | Probability of keeping a connection to hidden unit active | 0.99 |
| | K (items to recommend) | 10 |
| VBCAR | Model | VAE (Variational Auto Encoder) |
| | Epoch | 100 |
| | Latent dimensions | 256 |
| | Embedding dimensions | 64 |
| | Learning rate | 0.001 |
| | Batch size | 256 |
| | Alpha | 0.01 |
| | Negative items for each user | 100 |

## 4. Results and analysis

This section presents the results for assessing overall model performance when integrating previous buying history, performance over different datasets, and whether models can accurately capture health and sustainability information of food choices.

### 4.1. Model comparison

Tables 3 and 4 present the results of the comparison of the Dutch supermarket and Instacart datasets for N=30,000 and N=143,000, respectively. When assessing the performance of the different models for the preference-based version, we observe that the VBCAR model achieved better performance than the Baseline and RBM models across all metrics (i.e., MAP, NDCG, Precision, and Recall). Two Sample T-tests were conducted to compare the means of the various models to each other for each metric. Mean values were 0.0578 (SD = 0.0347), 0.2640 (SD = 0.1940), and 0.6540 (SD = 0.0862) for Baseline, RBM, and VBCAR respectively. Statistical significance testing (see Table 5) reveals that, except for Precision between the Baseline and RBM model, all preference-based models exhibited significant differences across each metric. Table 5 also shows significant effect sizes for these models across all metrics.

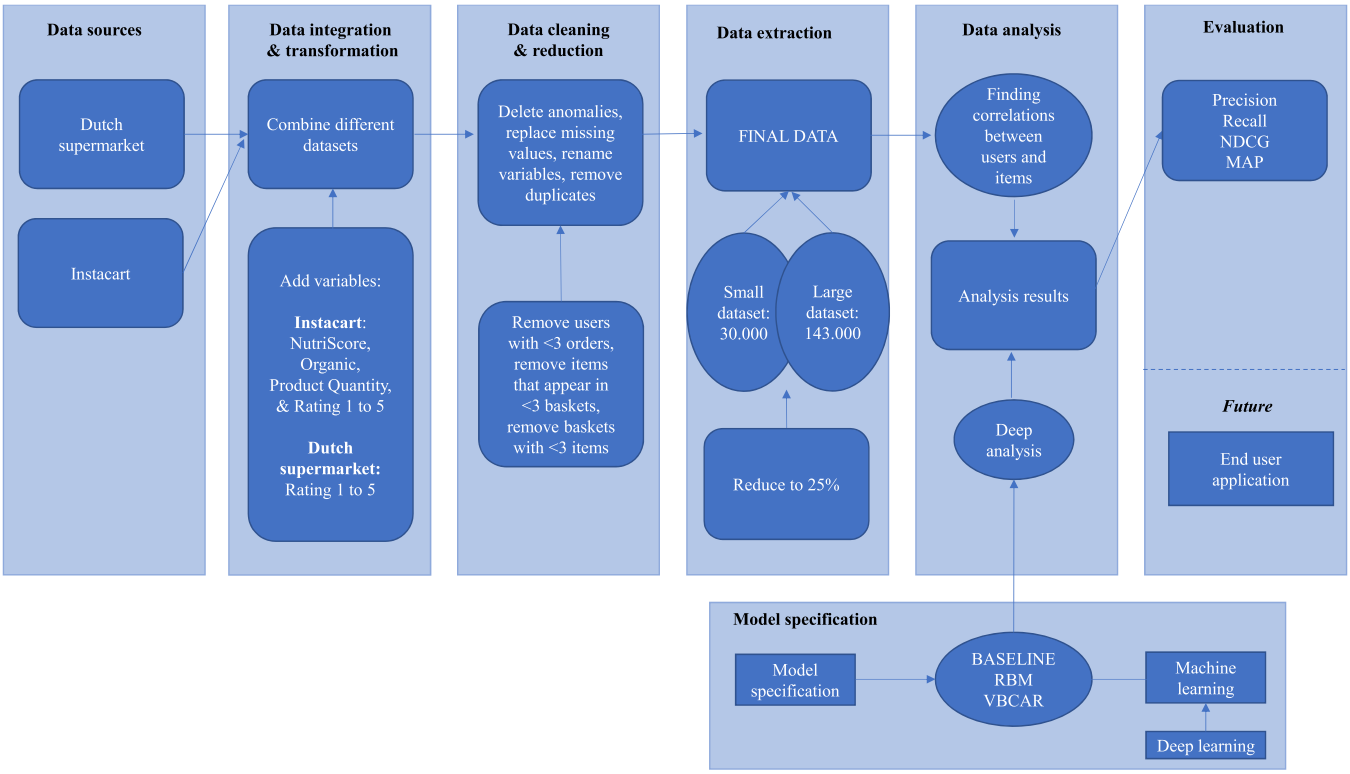


**Fig. 2.** Functional architecture.

### 4.2. Effect of adding health or sustainability information

To answer the main research question on the influence of including healthiness and/or sustainability information of food choices on model performance, we compared four different versions for all models (Baseline, RBM, VBCAR): (1) preference-based, (2) preference and health based, (3) preference and sustainability based, and (4) all combined. In line with the overall model performance, VBCAR outperformed the Baseline and RBM models across all metrics (i.e., MAP, NDCG, Precision, and Recall), regardless of the additional information included. Table 3 and Table 4 show no difference in performance of the Baseline model, and only minor changes for RBM and VBCAR. Significance testing (see Table 6) demonstrated that the inclusion of health information, sustainability information, or both, did not yield a statistically significant change in outcome measures compared to the preference-based model across all three algorithms.

A boxplot, as shown in Fig. 3, was generated to visualize the distribution of evaluation outcomes for all models. The boxplot shows that variability happened between the models rather than within each variation based on additional health or sustainability information. The vertical spread indicates that median values for VBCAR models were higher compared to both RBM and Baseline models (as confirmed by significant results presented in Section 4.1), whereas the horizontal alignment of the boxplots suggests similar distributions (as confirmed by non-significant results presented in this section).

### 4.3. Dataset comparison

The current research used two different datasets (Instacart and Dutch supermarket data) and two different data sizes (30K and 143K). The Dutch supermarket data reflected consumer shopping behavior from 2021, with shoppers being shown health and sustainability information about products. Instacart data represented buying behavior from 2017 with constructed variables for health and sustainability. The two datasets were distinct from each other, implying independence. However, when comparing the 30K and 143K datasets, the analysis was conducted within the dataset (Instacart or Dutch supermarket), assuming dependency within the small and large datasets.

#### 4.3.1. Compare Instacart to Dutch supermarket data

Although Tables 3 and 4 suggest better Precision and NDCG for the Dutch supermarket model (except VBCAR NDCG for small datasets) and improved Recall and MAP for the Instacart data (except VBCAR MAP for large datasets) for both small datasets (30K) and large datasets (143K), Wilcoxon Rank Sum Test did not yield statistically significant differences between the Dutch supermarket and Instacart dataset for both small (30K) and large (143K) datasets. Significance testing results are shown in Table 7.

#### 4.3.1. Compare small to large datasets for Instacart and Dutch supermarket data separately

Wilcoxon Rank Test assessed the differences between small and large datasets as presented in Tables 3 (small datasets: 30K) and 4 (large datasets: 143K). Results in Table 7 suggest that the declined performance for Baseline and RBM models and the improved performance of the VBCAR model with more data was not significant, except for the Precision metric. The precision of models significantly improved with more data for both Dutch supermarket and Instacart data.

## 5. Discussion

### 5.1. Discussion of findings

This research has investigated the performance of a multi-objective GRS focused on the health and sustainability level of food choices. A Variational Bayesian Context-Aware Representation (VBCAR) and a Restricted Boltzmann Machine (RBM) model were used and compared to a Baseline model to explore the effect of integrating health and sustainability information of products alongside user preferences in a GRS.

Regarding overall model performance, experiments on two real-life

**Table 3**
Overall performance on the two datasets ($n = 30.000$).

| | INSTACART | | | |
|---|---|---|---|---|
| | Precision | Recall | NDCG | MAP |
| BASELINE | | | | |
| preferences | 0.0287 | 0.0784 | 0.0642 | 0.0349 |
| preferences and health | 0.0287 | 0.0784 | 0.0642 | 0.0349 |
| preferences and organic | 0.0287 | 0.0784 | 0.0642 | 0.0349 |
| all | 0.0287 | 0.0784 | 0.0642 | 0.0349 |
| RBM | | | | |
| preferences | 0.1095 | 0.3406 | 0.3462 | 0.2507 |
| preferences and health | 0.1093 | 0.3403 | 0.3465 | 0.2510 |
| preferences and organic | 0.1092 | 0.3316 | 0.3435 | 0.2490 |
| all | *0.1096* | *0.3413* | *0.3474* | *0.2522* |
| VBCAR | | | | |
| preferences | 0.6855 | 0.7996 | 0.8940 | 0.7008 |
| preferences and health | **0.6873** | **0.8027** | 0.8940 | 0.7083 |
| preferences and organic | 0.6773 | 0.7967 | 0.8945 | **0.7125** |
| all | 0.6855 | 0.8006 | **0.8959** | 0.7105 |
| | DUTCH SUPERMARKET | | | |
| | Precision | Recall | NDCG | MAP |
| BASELINE | | | | |
| preferences | 0.0937 | 0.0305 | 0.1328 | 0.0158 |
| preferences and health | 0.0937 | 0.0305 | 0.1328 | 0.0158 |
| preferences and organic | 0.0937 | 0.0305 | 0.1328 | 0.0158 |
| all | 0.0937 | 0.0305 | 0.1328 | 0.0158 |
| RBM | | | | |
| preferences | *0.4552* | *0.2018* | *0.5634* | *0.1706* |
| preferences and health | 0.4507 | 0.2011 | 0.5602 | 0.1694 |
| preferences and organic | 0.4437 | 0.1992 | 0.5531 | 0.1660 |
| all | 0.4489 | 0.2001 | 0.5593 | 0.1684 |
| VBCAR | | | | |
| preferences | 0.7458 | **0.5587** | **0.8628** | **0.4933** |
| preferences and health | **0.8207** | 0.4144 | 0.8442 | 0.3669 |
| preferences and organic | 0.7322 | 0.5520 | 0.8565 | 0.4897 |
| all | 0.7288 | 0.5511 | 0.8498 | 0.4805 |

Note. The best-performing result is highlighted in bold and underlined (note that this is always the VBCAR model). The best result for the RBM model is given in italics (for BASELINE there is no difference between models).

**Table 4**
Overall performance on the two datasets ($n = 143.000$).

| | INSTACART | | | |
|---|---|---|---|---|
| | Precision | Recall | NDCG | MAP |
| BASELINE | | | | |
| preferences | 0.0291 | 0.0753 | 0.0625 | 0.0328 |
| preferences and health | 0.0291 | 0.0753 | 0.0625 | 0.0328 |
| preferences and organic | 0.0291 | 0.0753 | 0.0625 | 0.0328 |
| all | 0.0291 | 0.0753 | 0.0625 | 0.0328 |
| RBM | | | | |
| preferences | 0.0844 | 0.2452 | 0.2731 | 0.1956 |
| preferences and health | 0.0846 | 0.2457 | 0.2733 | 0.1959 |
| preferences and organic | *0.0877* | *0.2548* | *0.2913* | *0.2121* |
| all | 0.0846 | 0.2457 | 0.2733 | 0.1959 |
| VBCAR | | | | |
| preferences | 0.6430 | 0.8485 | 0.9011 | 0.7565 |
| preferences and health | 0.6428 | 0.8479 | 0.9007 | 0.7550 |
| preferences and organic | 0.6451 | **0.8491** | 0.9010 | 0.7556 |
| all | **0.6479** | 0.8486 | **0.9042** | **0.7593** |
| | DUTCH SUPERMARKET | | | |
| | Precision | Recall | NDCG | MAP |
| BASELINE | | | | |
| preferences | 0.0832 | 0.0238 | 0.1242 | 0.0123 |
| preferences and health | 0.0832 | 0.0238 | 0.1242 | 0.0123 |
| preferences and organic | 0.0832 | 0.0238 | 0.1242 | 0.0123 |
| all | 0.0832 | 0.0238 | 0.1242 | 0.0123 |
| RBM | | | | |
| preferences | 0.4077 | 0.1850 | 0.5294 | 0.1590 |
| preferences and health | 0.4086 | 0.1853 | 0.5298 | 0.1591 |
| preferences and organic | *0.4178* | *0.1891* | *0.5413* | *0.1644* |
| all | 0.4064 | 0.1845 | 0.5291 | 0.1590 |
| VBCAR | | | | |
| preferences | 0.9354 | 0.5411 | 0.9375 | 0.8413 |
| preferences and health | **0.9367** | **0.5427** | 0.9348 | 0.8393 |
| preferences and organic | 0.9355 | 0.5424 | **0.9412** | **0.8481** |
| all | 0.9364 | 0.5426 | 0.9363 | 0.8414 |

Note. The experiment was run 10 times and the average value of the results is computed. The best-performing result is highlighted in bold and underlined (note that this is always the VBCAR model). The best result for the RBM model is given in italics (for BASELINE there is no difference between models).

datasets showed that VBCAR and RBM significantly achieved better performance than the Baseline model for each of the variations: preferences, preferences and health, preferences and sustainability, and all combined. The experiment also showed that VBCAR outperformed RBM, potentially caused by the superiority of the domain-specific approach of VBCAR compared to the general deep-learning approach of RBM. The accuracy of the RBM model was previously verified using a movie dataset, where distinct factors play a role compared to grocery shopping data. Groceries encompass a wide variety of products that have different characteristics (e.g., freshness, seasonality) and purchase factors (e.g., nutritional preferences and allergies, expiration date, availability) that influence choices (Steptoe et al., 1996), which can be better captured with a model that is context-aware and able to learn complex patterns and to model uncertainties by estimating probabilities rather than making predictions. While recommender systems share common principles, the multifaceted nature and specific characteristics of the grocery shopping domain might require a more context-dependent algorithm to provide relevant recommendations to grocery shoppers (Ricci et al., 2022). Yet, results should be approached with caution due to the wide confidence intervals indicating uncertainty surrounding the magnitude of the observed effects (Thompson, 2007). To increase certainty of the effect sizes, future work could use a more extensive dataset that covers a longer time period. While the 4-month period from September to December captures a lot of variability in consumer behavior, purchasing patterns, and product availability due to seasonal factors, a longer period might provide more meaningful insights into consumer shopping behavior and allow for more precise estimates.

Concerning the effect of adding a product's health or sustainability information, T-test results revealed that adding this information did not introduce a significant difference to the resulting metrics MAP, NDCG, Precision, and Recall compared to preference-based models, which suggests that all variations of models could be used indistinctly. The congruence between consumer preferences and health and sustainability considerations is advantageous as it shows that health and sustainability information of food choices can be promoted without a significant drop in performance. Potentially, highlighting this information will not deter consumers from making these purchases. This insight is valuable for retailers and store managers who frequently perceive a misalignment between decisions that drive sales and those that improve health (Gravlee et al., 2014).

The insignificant performance differences between the models can be interpreted from several angles. First, the datasets likely represent a diverse range of consumer behaviors, preferences, and priorities. The lack of significant performance differences could indicate that the proportion of individuals who prioritize health and sustainability is relatively small compared to those who prioritize other factors such as taste, convenience, or affordability. Second, the results of these datasets might reflect a limited availability and accessibility of healthier and sustainable food options, decreasing the possibility of incorporating these factors into purchasing decisions. Future work can investigate the value of introducing and saliently presenting health and sustainability information to consumers in shifting preferences and purchase behavior over time. It can be investigated whether consumers make healthier and/or more sustainable food choices by comparing historical data where consumers were not exposed to health and sustainability information versus transactional data where health and sustainability information was present in the (online) store.

**Table 5**
Two Sample T-Test Model Comparison (Baseline, RBM, VBCAR).

| | | | | | | 95% CI | | | 95% CI for Hedges g | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Metric | Means[1] | t | df | p-value[2] | Lower | Upper | Hedges' g | Lower | Upper |
| **Baseline** | MAP | 0.0240 | -8.0222 | 6 | **0.0002** | -0.2219 | -0.1182 | -4.9327 | -7.9573 | -1.9080 |
| **vs** | NDCG | 0.0959 | -4.5623 | 6 | **0.0038** | -0.5102 | -0.1550 | -2.8052 | -4.9243 | -0.6862 |
| **RBM** | Precision * | 0.0587 | -2.0817 | 3.1907 | 0.1233 | -0.5093 | 0.0983 | -1.2800 | -2.9314 | 0.3714 |
| | Recall | 0.0520 | -5.0653 | 6 | **0.0023** | -0.2835 | -0.0988 | -3.1145 | -5.3525 | -0.8766 |
| **Baseline** | MAP * | 0.6980 | -9.0702 | 3.0365 | **0.0027** | -0.9089 | -0.4391 | -5.577 | -8.9034 | -2.2507 |
| **vs** | NDCG | 0.8989 | -32.998 | 6 | **5.154 exp-8** | -0.8625 | -0.7434 | -20.2895 | -31.1867 | -9.3924 |
| **VBCAR** | Precision | 0.6539 | -12.812 | 6 | **1.389 exp-5** | -0.7088 | -0.4815 | -7.8780 | -12.3304 | -3.4255 |
| | Recall * | 0.7856 | -8.9521 | 3.1917 | **0.0023** | -0.9857 | -0.4814 | -5.5044 | -8.7964 | -2.2125 |
| **RBM** | MAP | 0.1940 | -6.5589 | 6 | **0.0006** | -0.6920 | -0.3160 | -4.0329 | -6.6531 | -1.4126 |
| **vs** | NDCG * | 0.4280 | -6.5437 | 3.2849 | **0.0055** | -0.6890 | -0.2527 | -4.0236 | -6.6398 | -1.4074 |
| **VBCAR** | Precision | 0.2642 | -3.6652 | 6 | **0.0105** | -0.6498 | -0.1295 | -2.2537 | -4.1774 | -0.3300 |
| | Recall | 0.2432 | -6.1723 | 6 | **0.0008** | -0.7574 | -0.3274 | -3.7952 | -6.3130 | -1.2774 |

* Equal variances could not be assumed, so the Welch T-Test was performed.
[1] Means are values for Baseline (Baseline vs RBM), VBCAR (Baseline vs VBCAR), and RBM (RBM vs VBCAR), respectively
[2] Significant p-values are bold and underlined

**Table 6**
Two Sample T-Test Model variation comparison (preference, health, and/or sustainability).

| | | | | | | | 95% CI | | | 95% CI Hedges g | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | Variations[1] | Metric | Means[2] | t | Df | p-value | Lower | Upper | Hedges' g | Lower | Upper |
| **RBM** | **Preference** | MAP | 0.1939 | -0.0043 | 6 | 0.9967 | -0.0710 | 0.0707 | 0.0027 | -1.5019 | 1.5072 |
| | **vs** | NDCG | 0.4275 | -0.0058 | 6 | 0.9956 | -0.2429 | 0.2418 | 0.0036 | -1.5010 | 1.5081 |
| | **Health** | Precision | 0.2633 | -0.0066 | 6 | 0.9950 | -0.3361 | 0.3343 | 0.0040 | -1.5005 | 1.5085 |
| | | Recall | 0.2431 | -0.0010 | 6 | 0.9992 | -0.1207 | 0.1206 | 0.0006 | -1.5039 | 1.5052 |
| | **Preference** | MAP | 0.1979 | 0.1355 | 6 | 0.8967 | -0.0665 | 0.0743 | -0.0833 | -1.5885 | 1.4219 |
| | **vs** | NDCG | 0.4323 | 0.0440 | 6 | 0.9664 | -0.2337 | 0.2423 | -0.0270 | -1.5316 | 1.4776 |
| | **Sustainable** | Precision | 0.2646 | 0.0029 | 6 | 0.9978 | -0.3342 | 0.3350 | -0.0018 | -1.5063 | 1.5027 |
| | | Recall | 0.2437 | 0.0110 | 6 | 0.9916 | -0.1164 | 0.1175 | -0.0068 | -1.5113 | 1.4978 |
| | **Preference** | MAP | 0.1939 | -0.0034 | 6 | 0.9974 | -0.0717 | 0.0715 | 0.0021 | -1.5024 | 1.5066 |
| | **vs** | NDCG | 0.4273 | -0.0076 | 6 | 0.9942 | -0.2426 | 0.2411 | 0.0047 | -1.4999 | 1.5092 |
| | **All** | Precision | 0.2624 | -0.0134 | 6 | 0.9898 | -0.3360 | 0.3323 | 0.0082 | -1.4963 | 1.5128 |
| | | Recall | 0.2429 | -0.0050 | 6 | 0.9961 | -0.1216 | 0.1211 | 0.0031 | -1.5014 | 1.5076 |
| **VBCAR** | **Preference** | MAP | 0.6674 | -0.24001 | 6 | 0.8183 | -0.3426 | 0.2814 | 0.1476 | -1.359 | 1.6542 |
| | **vs** | NDCG | 0.8934 | -0.22445 | 6 | 0.8299 | -0.0646 | 0.0537 | 0.1380 | -1.3683 | 1.6443 |
| | **Health** | Precision | 0.6734 | 0.27121 | 6 | 0.7953 | -0.1566 | 0.1957 | -0.1668 | -1.6739 | 1.3404 |
| | | Recall | 0.7504 | -0.2495 | 6 | 0.8113 | -0.3797 | 0.3094 | 0.1534 | -1.3534 | 1.6601 |
| | **Preference** | MAP | 0.7015 | 0.0330 | 6 | 0.9748 | -0.2563 | 0.2632 | -0.0203 | -1.5249 | 1.4843 |
| | **vs** | NDCG | 0.8983 | -0.0235 | 6 | 0.9818 | -0.0572 | 0.0561 | 0.0146 | -1.4900 | 1.5192 |
| | **Sustainable** | Precision | 0.6493 | -0.0783 | 6 | 0.9401 | -0.1483 | 0.1391 | 0.04816 | -1.4566 | 1.5529 |
| | | Recall | 0.7833 | -0.0193 | 6 | 0.9852 | -0.2841 | 0.2797 | 0.0119 | -1.4927 | 1.5164 |
| | **Preference** | MAP | 0.6979 | -0.0005 | 6 | 0.9996 | 0.2621 | 0.2620 | 0.0003 | -1.5043 | 1.5048 |
| | **vs** | NDCG | 0.8966 | -0.0977 | 6 | 0.9253 | -0.0599 | 0.0553 | 0.0601 | -1.4448 | 1.5650 |
| | **All** | Precision | 0.6512 | -0.0450 | 6 | 0.9654 | -0.1462 | 0.1409 | 0.0278 | -1.4768 | 1.5324 |
| | | Recall | 0.7842 | -0.0119 | 6 | 0.9909 | -0.2839 | 0.2812 | 0.0073 | -1.4972 | 1.5119 |

[1] Note that 'Preference' refers to the presence-based model, 'Heath' to the preference and health model, 'Sustainable' to the preference and sustainability model, and 'All' to the combined model based on preferences, health, and sustainability information.
[2] The Mean values given are from the health, sustainable, or combined model. Mean values of the preference-based RBM and VBCAR models are 0.1940 & 0.6980 (MAP), 0.4280 & 0.8989 (NDCG), 0.2642 & 0.6539 (PRECISION), and 0.2432 & 0. 7856 (RECALL), respectively.

Furthermore, significance testing showed no evidence of improved performance for different datasets (Instacart versus Dutch supermarket data) and data sizes (30K versus 143K). These evaluation results on various datasets show the generalizability and robustness of the models (Bauer et al., 2023). Food choices in the Dutch Supermarket data were expected to be influenced by the presence of additional product labeling to indicate the health (NutriScore label) and sustainability (organic label) level of food products, as labeling could raise awareness among consumers about the health and sustainability aspects of food products, potentially influencing choices. Despite this difference with the Instacart dataset, in which constructed variables were employed to indicate the health and sustainability features of products, no significant difference in performance was found. It is important to consider that labeling is expected to be most effective when all products in the store are displayed with the label (Hagmann and Siegrist, 2020), a condition that was not reflected in the current data. Future work can address this by employing transaction data in which interventions such as labeling are present on all products.

While increased data typically enhances recommender system outcomes (Schafer et al., 1999), the current results did not align with this expectation. This deviation from expectations may be caused by the relatively small size of all datasets. Given the premise that deep learning models perform best if they can learn from more data, future researchers are advised to explore the dynamics between data volume and the efficacy of multi-objective GRS. An interesting line of research is to examine the effectiveness of domain-specific algorithms on a larger dataset capturing long-term shopping behavior including seasonality effects. Furthermore, integrating consumer information could greatly improve personalization results, but privacy regulations should be in place to assure consumers that their data will be protected (Schafer et al., 1999).

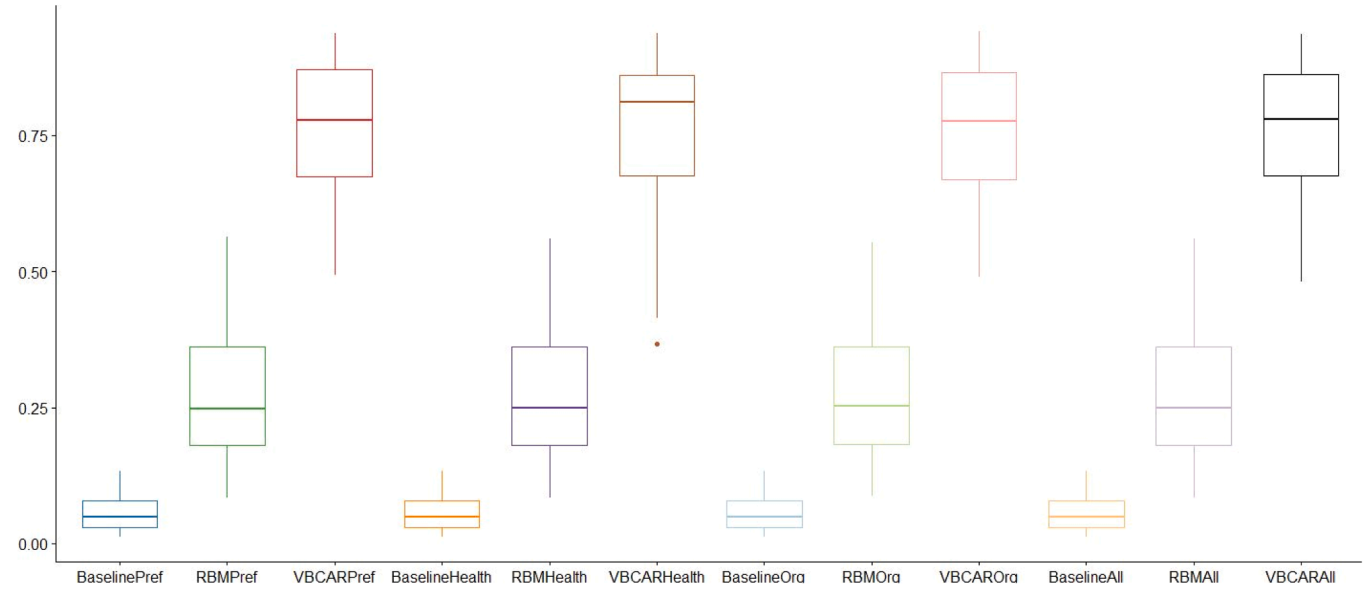


**Fig. 3.** Boxplot comparison of performance across all models
Note. Pref refers to preference-based models, Health also includes health information, Org also includes sustainability information, and All is the combined model based on preferences, health, and sustainability information.

**Table 7**
Dataset comparison.

| Dataset comparison | Metric | Test statistic | p-value[1] |
|---|---|---|---|
| **30K Instacart** | MAP | W = 48 | 0.1730 |
| **vs** | NDCG | W = 80 | 0.6636 |
| **30K Dutch supermarket** | Precision | W = 96 | 0.1729 |
| | Recall | W = 48 | 0.1730 |
| **143K Instacart** | MAP | W = 64 | 0.6635 |
| **vs** | NDCG | W = 96 | 0.1729 |
| **143K Dutch supermarket** | Precision | W = 80 | 0.6636 |
| | Recall | W = 64 | 0.6636 |
| **30K Dutch supermarket** | MAP | V = 42 | 0.8439 |
| **vs** | NDCG | V = 42 | 0.8439 |
| **143K Dutch supermarket** | Precision | V = 0 | **0.0024** |
| | Recall | V = 42 | 0.8439 |
| **30K Instacart** | MAP | V = 32 | 0.6087 |
| **vs** | NDCG | V = 26 | 0.3249 |
| **143K Instacart** | Precision | V = 10 | **0.0248** |
| | Recall | V = 26 | 0.3249 |

[1] Significant p-values are bold and underlined

### 5.2. Limitations

A limitation of the current work is the use of deep learning approaches to test model performance, as these predictive models often lack interpretability about the underlying mechanisms for model prediction. The current research specifically focused on integrating the health (FSA) and sustainability (organic) scores of products, but the exact interpretability of the models is lacking. Interpretability is important for both retailers, to refine marketing strategies, and for consumers, to understand why suggestions are given. Nonetheless, while interpretability is valuable in understanding how models arrive at their output, the lack of it does not diminish the importance of the study's conclusions. Not only does the current work underscore the need for continued research into explainable AI, but it also shows that models integrating health and sustainability features are equally efficient in product recommendations as a preference-based model.

Another limitation is the underlying data used for determining recommender system accuracy. First, a four-month period of data might not fully capture long-term behavior, but we believe that this period still provides meaningful insights into consumer shopping behavior, particularly during a diverse time of the year (September to December). A second possible threat to the validity of the findings is how the data was preprocessed. For instance, removing users with too few items or baskets might skew results, and random sampling may lead to a loss of significant information (Famili et al., 1997). However, careful documentation of all the preprocessing steps mitigates this threat by ensuring transparency and reproducibility. Third, since the datasets were obtained in 2017 and 2021, the preference relationships and potentially the health and sustainability information of products might have changed along with product modifications. For instance, the increase in the introduction and consumption of plant-based products would require new data on such products. In line with this, the publicly available dataset employed did not contain health and sustainability features of products and randomly assigning values for these two variables might have affected the results. Nonetheless, semi-synthetic data is often used as a comparison and considered valuable (e.g., Karatzoglou et al., 2010), especially when compared with a recent, real-world dataset containing all required information.

Moreover, differences in shopping habits between online and offline environments highlight a need for testing algorithms with online store data rather than offline channel data (Ariannezhad et al., 2021). A drawback is that publicly available datasets oftentimes do not represent recent purchase behavior and private datasets inhibit replicability. This research provides valuable insights, but the field would greatly benefit from anonymized, up-to-date, publicly available transaction data from an online store. By making research more reproducible, the validity of the results increases.

Relatedly, future work can examine the implementation of a recommender system algorithm in a (simulated) online grocery store to explore the potential of a recommender system that integrates nutritional and environmental criteria as a digital marketing tool for public health- and sustainability purposes. An extension of this work could involve online grocery stores offering recommendations to consumers, assessing their influence on purchasing behavior and examining consumer reactions to these suggestions. User experience, which is crucial for the success of recommendations, is not solely determined by the accuracy of recommendations but also looks at aspects such as privacy or situational factors that are best explored in a real-life grocery shopping scenario (Knijnenburg et al., 2012). Furthermore, real-life applications of recommender systems could generate insights into the effect

on sales volume and revenue, further improving insights into the effectiveness of product recommendations.

### 5.3. Theoretical and practical implications

An examination of a GRS that includes healthiness and sustainability information of products yields several contributions. Theoretically, our study shows that a context-dependent algorithm can handle additional criteria around healthiness and sustainability better than the more general baseline and RBM algorithm. Findings on the superior performance of the context-dependent algorithm demonstrate that machine learning can effectively integrate complex side information and show the importance of investigating the more sophisticated techniques in GRS research. Existing theoretical models could be expanded by incorporating broader contextual factors, leading to a more holistic understanding of consumer food choices. For instance, while traditional theoretical models such as Nudge theory mainly focus on personal preferences, they can also account for global issues such as health and sustainability factors of food choices. Machine learning's ability to analyze and learn from large datasets enables the integration of complex behavioral and contextual data, supporting the integration of broader social considerations into the consumer decision-making process.

Furthermore, the insignificant results between preference-based models and models incorporating health and sustainability information suggest that incorporating such factors into the models does not notably affect their performance. This implies that adding health and sustainability considerations may not interfere with the model's ability to predict consumer choices. From a behavioral theory perspective, this could align with concepts like Nudge Theory, which posits that subtle changes in how choices are framed can influence decisions without limiting options. In this context, machine learning models could provide insights into how integrating health and sustainability information into recommendations might subtly guide consumer decisions to more sustainable or healthier food choices.

This study additionally demonstrates the need for future research on how to actively shape consumer decision-making toward healthier and more sustainable grocery choices. Educating consumers on the importance of health and sustainability might eventually influence their preferences, making this information more impactful in future recommender systems. As such, our findings can inform the development of more advanced and responsible recommender systems that better educate consumers and contribute to broader sustainability goals.

In turn, machine learning allows to analyze how consumers respond to specific digital interventions that highlight health benefits or the environmental impact of products. Analyzing such responses provides the opportunity to refine theoretical models of consumer decision-making and behavioral change, with a specific focus on the role of education in influencing consumer choices. By examining how different recommendation strategies impact consumer food choices, researchers can gain a deeper understanding of the effectiveness of various digital interventions. To illustrate, machine learning can reveal which types of product information are most persuasive in encouraging healthier eating habits. Such evidence can contribute to the refinement of existing theories by providing a more nuanced view of how digital interventions influence consumer food choice behavior.

This study also has various practical implications. Even though there is a growing awareness of health and sustainability issues among consumers (Jansen et al., 2023), integrating these issues into a recommender system for grocery shopping does not significantly improve performance compared to preference-based models. This suggests that awareness alone might not translate into improved behavior, highlighting the need for better education on the importance of health and sustainability, potentially influencing future consumer behavior and the effectiveness of multigoal recommender systems. Furthermore, significant differences between models indicate that advanced recommender systems can lead to more accurate and relevant recommendations, enhancing the overall shopping experience for consumers. These advanced models make better use of the vast amount of data that retailers are collecting these days, allowing for more tailored recommendations. Potentially, retailers can leverage additional data sources related to product healthiness and sustainability to enhance the relevance of recommendations for niche groups motivated to buy healthy and sustainable. In doing this, retailers can support broader sustainability goals (United Nations, 2023) by guiding consumers toward healthier and more sustainable products.

## 6. Conclusion

This paper investigates the impact of integrating health and sustainability information of grocery products for improving the performance of a GRS. Employing three models with four versions each (preferences, health, sustainability, combination of all) and two datasets, results show that a recommender system that integrates health and sustainability information of products in addition to consumer preferences is equally efficient in product recommendation as a model based on consumer preferences alone. These results demonstrate that recommender systems have the potential to assist consumers in finding healthier and more sustainable choices that are also suited to their preference, but also highlight the need for increasing consumer awareness of the healthiness and sustainability level of food choices. In conclusion, these insights show that information systems such as GRS hold promise in advancing sustainable development. Over time, health and sustainability-focused recommender systems can improve consumer buying behavior as they provide a variety of food products including healthier and more sustainable options. Our findings position grocery recommendations as an effective strategy to be implemented in the online shopping environment.

## Declarations

### *Availability of data and materials*

The Instacart dataset and codes (Baseline, RBM, VBCAR) used are publicly available, links are provided in the manuscript. Data from the Dutch supermarket is not publicly available. Notebook files to merge the Instacart data and perform exploratory data analysis are available on Github (https://github.com/laura1jansen/algorithmhealthsust/tree/main). Similar procedures were used for the Dutch supermarket dataset.

### *Funding*

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

### CRediT authorship contribution statement

**Laura Z.H. Jansen:** Writing – original draft, Visualization, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Kwabena E. Bennin:** Writing – review & editing, Supervision, Conceptualization.

### Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Appendices

### *Appendix A. Variables in the datasets*

**Table A1**
Variables in the Instacart dataset.

| | Meaning of variable | Example |
|---|---|---|
| CUSTOMER_ID | Number corresponding to a customer | 205,970 |
| ORDER_ID | Number corresponding to an order | 3 |
| PRODUCT_ID | Number corresponding to product | 21,903 |
| PRODUCT_QUANTITY | Amount of product bought | 1 |
| ORDER_DOW | Day of the week | 5 |
| ORDER_HOUR_OF_DAY | Hour of the day | 17 |
| DAYS_SINCE_PRIOR_ORDER | Number of days since last order | 12 |
| ADD_TO_CART_ORDER | Sequence in which products in the same order were added to the cart | 4 |
| REORDERED | Whether a product was ordered before (1) or not (0) | 1 |
| ORDER_NUMBER | Order sequence number of specific consumer | 16 |
| AISLE_ID | Number corresponding to an aisle | 123 |
| DEPARTMENT_ID | Number corresponding to a department | 4 |
| HOOFDCATEGORIE | The main category of the product | PRODUCE |
| SUBCATEGORIE | The sub category of the product | PACKAGED VEGETABLE FRUITS |
| OMSCHRIJVING | Description of the product including the brand, product name and content | ORGANIC BABY SPINACH |
| NUTRISCORE | Healthiness score of the product based on the NutriScore label (values 0 (no label) and 1 to 5) | 0 |
| ORGANIC | Sustainability label indicating whether a product is Organic (1) or not (0) | 1 |
| RATINGS | The rating assigned to the product (values 1 to 5) | 4 |
| EVAL_SET | Evaluation set the transaction belongs to | PRIOR |

* Note that the variables NUTRISCORE, ORGANIC, and RATINGS are constructed variables

**Table A2**
Variables in the Dutch supermarket dataset.

| | Meaning of variable | Example |
|---|---|---|
| CUSTOMER_ID | Number corresponding to a customer | 100XXX |
| ORDER_ID | Number corresponding to an order | 538XXXX |
| STORE_ID | Number corresponding to a store | 5XX |
| POSTALCODE | Postal code of customer | 1234AB |
| ORDERS_TIME | Time of order (yyyy-mm-dd) | 15/9/21 |
| PRODUCT_ID | Number corresponding to product (same as CE ART. NR.) | 559,572 |
| PRODUCT_QUANTITY | Number of products bought | 1 |
| PRODUCT_PRIJS | Price of product bought | 4.09 |
| WEEKNR | Number of the week | 37 |
| JAARNR | Number of the year | 2021 |
| PE ART. NR. | Number uniquely identifying each product | 559,573 |
| CE ART. NR. | Number identifying each product (same as PRODUCT_ID) | 559,572 |
| MERK | Brand of the product | BIO+ |
| OMSCHRIJVING | Description of the product including the brand, product name and content | BIO+ PIJNBOOMPITTEN BIOLOGISCH BK 75GR |
| VERPAK | Measuring unit (e.g., KG, FLS, PAK) | BK |
| INHOUD_X | Content | 75 |
| EENHEID | Measuring unit (e.g., KG, ML, GR, L, ST) | GR |
| HOOFDCATEGORIE | The main category of the product | AARDAPPELEN, GROENTE, FRUIT |
| SUBCATEGORIE | The sub category of the product | FRUIT |
| SUBSUBCATEGORIE | The sub category of the product | NOTEN EN GEDROOGDE VRUCHTEN |
| SUBGR. OMSCHR. | Description of subgroup of the product (most specific category information) | NOTEN EN GEDROOGDE VRUCHTEN |
| NUTRISCORE | Healthiness score of the product based on the NutriScore label (values 0 (no label) and 1 to 5) | 3 |
| ORGANIC | Sustainability label indicating whether a product is Organic (1) or not (0)Note that the column ORGANIC50 has 50/50 ratio of label or non-label. | 1 |
| RATINGS | The rating assigned to the product (values 1 to 5) | 3 |
| PROMO | Whether product was on promotion in a specific week (yes or no) | NO |
| EVAL_SET | Evaluation set the transaction belongs to | TRAIN |

* Note that the variables EVAL_SET and RATINGS are constructed variables
* Other variables present in the data but irrelevant for the current research are PRODNR, SUBGR, ST, EAN CE, EAN PE, VAR, VOLGNUMMER, LAST_ORDERED, and all WEB variables that refer to the description in the online store.

## References

Abbu, H., Fleischmann, D., & Gopalakrishna, P. (2021). The digital transformation of the grocery business - driven by consumers, powered by technology, and accelerated by the COVID-19 pandemic. In Á. Rocha, H. Adeli, G. Dzemyda, F. Moreira, & A. M Ramalho Correia (Eds.), *Trends and applications in information systems and technologies, worldcist 2021. advances in intelligent systems and computing* (Eds., pp. 329–339). Springer International Publishing. Cham, Springer International Publishing. https://doi.org/10.1007/978-3-030-72660-7_32.

Abdar, M., Pourpanah, F., Hussain, S., Rezazadegan, D., Liu, L., Ghavamzadeh, M., et al. (2021). A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Information Fusion, 76*, 243–297. https://doi.org/10.1016/j.inffus.2021.05.008

Alamdari, P. M., Navimipour, N. J., Hosseinzadeh, M., Safaei, A. A., & Darwesh, A. (2020). A systematic study on the recommender systems in the E-commerce. *IEEE*

*Access : Practical Innovations, Open Solutions, 8*, 115694–115716. https://doi.org/10.1109/ACCESS.2020.3002803
Ansari, A., Li, Y., & Zhang, J. Z. (2018). Probabilistic topic model for hybrid recommender systems: a stochastic variational Bayesian approach. *Mark. Science, 37*, 987–1008. https://doi.org/10.1287/mksc.2018.1113
Ariannezhad, M., Jullien, S., Li, M., Fang, M., Schelter, S., & Rijke, M. (2022). ReCANet: a repeat consumption-aware neural network for next basket recommendation in grocery shopping. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 1240–1250). https://doi.org/10.1145/3477495.3531708
Ariannezhad, M., Jullien, S., Nauts, P., Fang, M., Schelter, S., & de Rijke, M. (2021). Understanding multi-channel customer behavior in retail. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management, Assoc. Comput. Mach* (pp. 2867–2871). https://doi.org/10.1145/3459637.3482208
T. Asikis, Multi-objective optimization for value-sensitive and sustainable basket recommendations, 2021. https://doi.org/10.48550/arXiv.2111.05944.
Baarsma, B., & Groenewegen, J. (2021). COVID-19 and the demand for online grocery shopping: Empirical evidence from the Netherlands. *De Economist, 169*, 407–421. https://doi.org/10.1007/s10645-021-09389-y
Bai, T., Zou, L., Zhao, W. X., Du, P., Liu, W., Nie, J.-Y., et al. (2019). CTRec: a long-short demands evolution model for continuous-time recommendation. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, Assoc. Comput. Mach* (pp. 675–684). https://doi.org/10.1145/3331184.3331199
Barkan, O. (2017). Bayesian neural word embedding. In *Proceedings of the AAAI Conference on Artificial Intelligence* (pp. 3135–3143). https://doi.org/10.1609/aaai.v31i1.10987
Bauer, C., Zangerle, E., & Said, A. (2023). Exploring the landscape of recommender systems evaluation: practices and perspectives. *ACM Transactions on Information Systems*. https://doi.org/10.1145/3629170
Bodike, Y., Heu, D., Kadari, B., Kiser, B., & Pirouz, M. (2020). A novel recommender system for healthy grocery shopping. In K. Arai, S. Kapoor, & R. Bhatia (Eds.), *Advances in Information and Communication: Proceedings of the 2020 Future of Information and Communication Conference (FICC)* (pp. 133–146). Springer International Publishing, Cham. https://doi.org/10.1007/978-3-030-39442-4_12.
Chen, G., & li, Z. (2021). A new method combining pattern prediction and preference prediction for next basket recommendation. *Entropy, 23*, 1430. https://doi.org/10.3390/e23111430
Chen, Y., Guo, Y., Fan, Q., Zhang, Q., & Dong, Y. (2023). *Health-Aware food recommendation based on knowledge graph and multi-task learning* (p. 12). Foods. https://doi.org/10.3390/foods12102079
Condliff, M., Lewis, D., Madigan, D., & Inc, T. (1999). *Bayesian mixed-effects models for recommender systems* (pp. 23–30). ACM SIGIR.
Dong, Y., Su, H., Zhu, J., & Zhang, B. (2017). Improving interpretability of deep neural networks with semantic information. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 4306–4314). https://doi.org/10.1109/CVPR.2017.110
Elahi, M., Beheshti, A., & Goluguri, S. (2021). *Recommender systems: challenges and opportunities in the age of big data and artificial intelligence* (pp. 15–39). Data Science and Its Applications. https://doi.org/10.1201/9781003102380-2
European Commission, Organics at a glance. https://agriculture.ec.europa.eu/farming/organic-farming/organics-glance_en, n.d. (accessed December 14, 2023).
Faggioli, G., Polato, M., & Aiolli, F. (2020). Recency aware collaborative filtering for next basket recommendation. In *Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization, Assoc. Comput. Mach* (pp. 80–87). https://doi.org/10.1145/3340631.3394850
Famili, A., Shen, W.-M., Weber, R., & Simoudis, E. (1997). Data preprocessing and intelligent data analysis. *Intelligent Data Analysis*, 3–23. https://doi.org/10.3233/IDA-1997-1102
FAO, WHO, Sustainable healthy diets – Guiding principles. https://www.who.int/publications/i/item/9789241516648, 2019 (accessed January 9, 2024).
Fayyaz, Z., Ebrahimian, M., Nawara, D., Ibrahim, A. F., & Kashef, R. F. (2020). Recommendation systems: Algorithms, challenges, metrics, and business opportunities. *Applied Science, 10*, 7748. https://doi.org/10.3390/app10217748
Food Standards Agency, The Nutrient Profiling Model, (2011). https://www.gov.uk/government/publications/the-nutrient-profiling-model (accessed March 27, 2023).
Fouad, M. A., Hussein, W., Rady, S., Yu, P. S., & Gharib, T. F. (2022). An efficient approach for rational next-basket recommendation. *IEEE Access : Practical Innovations, Open Solutions, 10*, 75657–75671. https://doi.org/10.1109/ACCESS.2022.3192396
Ghannadrad, A., Arezoumandan, M., Candela, L., & Castelli, D. (2022). Recommender systems for science: A basic taxonomy. In *IRCDL 2022: 18th Italian Research Conference on Digital Libraries, CEUR Workshop Proceedings (CEUR-WS.org)*.
B. Ghojogh, A. Ghodsi, F. Karray, M. Crowley, Restricted Boltzmann machine and deep belief network: Tutorial and survey, 2022. https://doi.org/10.48550/arXiv.2107.12521.
Goulet-Pelletier, J. C., & Cousineau, D. (2018). A review of effect sizes and their confidence intervals, Part {I}: The Cohen's d family. *The quantitative methods for psychology, 14*, 242–265. https://doi.org/10.20982/tqmp.14.4.p242
Gravlee, C. C., Boston, P. Q., Mitchell, M. M., Schultz, A. F., & Betterley, C. (2014). Food store owners' and managers' perspectives on the food environment: an exploratory mixed-methods study. *BMC Public Health, 14*, 1031. https://doi.org/10.1186/1471-2458-14-1031
Hafez, M. M., Redondo, R. P. D., Vilas, A. F., & Pazó, H. O. (2021). Multi-criteria recommendation systems to foster online grocery. *Sensors, 21*, 3747. https://doi.org/10.3390/s21113747
Hagmann, D., & Siegrist, M. (2020). Nutri-Score, multiple traffic light and incomplete nutrition labelling on food packages: Effects on consumers' accuracy in identifying healthier snack options. *Food quality and preference, 83*, Article 103894. https://doi.org/10.1016/j.foodqual.2020.103894
Hauptmann, H., Leipold, N., Madenach, M., Wintergerst, M., Lurz, M., Groh, G., et al. (2022). Effects and challenges of using a nutrition assistance system: results of a long-term mixed-method study. *User modeling and user-adapted interaction, 32*, 923–975. https://doi.org/10.1007/s11257-021-09301-y
Hoang, Q.-V. P., & Le, D.-T. (2021). Modeling multi-intent basket sequences for next-basket recommendation. In *13th International Conference on Knowledge and Systems Engineering (KSE)* (pp. 1–6). https://doi.org/10.1109/KSE53942.2021.9648773
Hollywood, L. E., Cuskelly, G. J., O'Brien, M., McConnon, A., Barnett, J., Raats, M. M., et al. (2013). Healthful grocery shopping. Perceptions and barriers. *Appetite, 70*, 119–126. https://doi.org/10.1016/j.appet.2013.06.090
Ilyas, Q., Mehmood, A., Ahmad, A., & Ahmad, M. (2022). A systematic study on a customer's next-items recommendation techniques. *Sustain, 14*, 7175. https://doi.org/10.3390/su14127175
Jansen, L., van Kleef, E., & Van Loo, E. J. (2021). The use of food swaps to encourage healthier online food choices: a randomized controlled trial. *The international journal of behavioral nutrition and physical activity, 18*, 156. https://doi.org/10.1186/s12966-021-01222-8
Jansen, L. Z. H., Bennin, K. E., van Kleef, E., & Van Loo, E. J. (2024). Online grocery shopping recommender systems: Common approaches and practices. *Computers in human behavior, 159*, Article 108336. https://doi.org/10.1016/j.chb.2024.108336
Jansen, L. Z. H., Van Loo, E. J., Bennin, K. E., & van Kleef, E. (2023). Exploring the role of decision support systems in promoting healthier and more sustainable online food shopping: A card sorting study. *Appetite, 188*, Article 106638. https://doi.org/10.1016/j.appet.2023.106638
Jesse, M., & Jannach, D. (2021). Digital nudging with recommender systems: Survey and future directions. *Computers in human behavior reports, 3*, Article 100052. https://doi.org/10.1016/j.chbr.2020.100052
Karatzoglou, A., Amatriain, X., Baltrunas, L., & Oliver, N. (2010). Multiverse recommendation: n-dimensional tensor factorization for context-aware collaborative filtering. In *Proceedings of the Fourth ACM Conference on Recommender Systems, Assoc. Comput. Mach* (pp. 79–86). https://doi.org/10.1145/1864708.1864727
Katz, O., Barkan, O., Koenigstein, N., & Zabari, N. (2022). Learning to ride a buy-cycle: A hyper-convolutional model for next basket repurchase recommendation. In *Proceedings of the 16th ACM Conference on Recommender Systems, Assoc. Comput. Mach* (pp. 316–326). https://doi.org/10.1145/3523227.3546763
Knijnenburg, B. P., Willemsen, M. C., Gantner, Z., Soncu, H., & Newell, C. (2012). Explaining the user experience of recommender systems. *User Modeling and User-Adapted Interaction, 22*, 441–504. https://doi.org/10.1007/s11257-011-9118-4
Li, M., Jullien, S., Ariannezhad, M., & de Rijke, M. (2023). A next basket recommendation reality check. *ACM Transactions on Information Systems, 41*, 1–29. https://doi.org/10.1145/3587153
Meng, Z., McCreadie, R., Macdonald, C., & Ounis, I. (2021). Variational Bayesian representation learning for grocery recommendation. *International Journal of Information Retrieval Research, 24*, 347–369. https://doi.org/10.1007/s10791-021-09397-1
Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). *Distributed representations of words and phrases and their compositionality* (pp. 3111–3119). Advances in Neural Information Processing Systems 26 (NIPS 2013).
Munt, A. E., Partridge, S. R., & Allman-Farinelli, M. (2017). The barriers and enablers of healthy eating among young adults: a missing piece of the obesity puzzle: A scoping review. *Obesity Reviews : An Official Journal of the International Association for the Study of Obesity, 18*, 1–17. https://doi.org/10.1111/obr.12472
Nordstokke, D., Zumbo, B., Cairns, S. L., & Saklofske, D. (2011). The operating characteristics of the nonparametric Levene test for equal variances with assessment and evaluation data. *Pract. Assess. Res. Eval., 16*, 1–8. https://doi.org/10.7275/5t99-zv93
Pecune, F., Callebert, L., & Marsella, S. (2020). *A recommender system for healthy and personalized recipe recommendations* (pp. 15–20). HealthRecSys@RecSys.
Portugal, I., Alencar, P., & Cowan, D. (2018). The use of machine learning algorithms in recommender systems: A systematic review. *Expert systems with applications, 97*, 205–227. https://doi.org/10.1016/j.eswa.2017.12.020
Rabobank, Onderzoek: Online boodschappen doen steeds meer ingeburgerd: Rabobank online food retail index. https://www.rabobank.nl/kennis/d011397204-online-boodschappen-doen-steeds-meer-ingeburgerd-rabobank-online-food-retail-index, 2023 (accessed June 4, 2024).
Rendle, S., Freudenthaler, C., & Schmidt-Thieme, L. (2010). Factorizing personalized Markov chains for next-basket recommendation. In *Proceedings of the 19th International Conference on World Wide Web, Assoc. Comput. Mach.* (pp. 811–820). https://doi.org/10.1145/1772690.1772773
Ricci, F., Rokach, L., & Shapira, B. (2011). *Recommender systems handbook*. Springer. https://doi.org/10.1007/978-0-387-85820-3
Ricci, F., Rokach, L., & Shapira, B. (2022). *Recommender systems handbook: third edition*. Springer US. https://doi.org/10.1007/978-1-0716-2197-4
Salakhutdinov, R., Mnih, A., & Hinton, G. (2007). Restricted Boltzmann machines for collaborative filtering. In *Proceedings of the 24th International Conference on Machine Learning, Assoc. Comput. Mach* (pp. 791–798). https://doi.org/10.1145/1273496.1273596
Santé Publique France, Nutri-Score. https://www.santepubliquefrance.fr/determinants-de-sante/nutrition-et-activite-physique/articles/nutri-score, 2023 (accessed March 25, 2023).
Sarwar, B., Karypis, G., Konstan, J., & Riedl, J. (2002). Recommender systems for large-scale E-Commerce scalable neighborhood formation using clustering. In *Proceedings*

*of the 5th International Conference on Computer and Information Technology* (pp. 291–324).

Schafer, J., Konstan, J., & Riedl, J. (1999). Recommender Systems in E-Commerce. In *Proceedings of the 1st ACM Conference on Electronic Commerce, Assoc. Comput. Mach* (pp. 158–166). https://doi.org/10.1145/336992.337035

Shao, Z., Wang, S., Zhang, Q., Lu, W., Li, Z., & Peng, X. (2022). A systematical evaluation for next-basket recommendation algorithms. In *2022 IEEE 9th International Conference on Data Science and Advanced Analytics (DSAA)* (pp. 1–10). https://doi.org/10.1109/DSAA54385.2022.10032359

Smith, B., & Linden, G. (2017). Two decades of recommender systems at Amazon.com. *IEEE internet computing, 21*, 12–18. https://doi.org/10.1109/MIC.2017.72

Starke, A. (2019). *RecSys challenges in achieving sustainable eating habits* (pp. 29–30). HealthRecSys@RecSys.

Starke, A. D., & Trattner, C. (2021). Promoting healthy food choices online: A case for multi-list recommender systems. In D. Glowacka, & V. Krishnamurthy (Eds.), *Proceedings of the ACM IUI 2021 Workshops, CEUR Workshop Proceedings (CEUR-WS.org)* (Eds.) https://edepot.wur.nl/551689.

Starke, A. D., Willemsen, M. C., & Trattner, C. (2021). Nudging healthy choices in food search through visual attractiveness, front. *Artificial intelligence, 4*, Article 621743. https://www.frontiersin.org/articles/10.3389/frai.2021.621743.

Statista, Revenue of the online food delivery market in the Netherlands from 2018 to 2028, by segment. https://www.statista.com/forecasts/1265498/revenue-segment-online-food-delivery-netherlands, 2024 (accessed June 4, 2024).

Steptoe, A., Pollard, T., & Wardle, J. (1996). Development of a Measure of the Motives Underlying the Selection of Food: The food choice questionnaire. *Appetite, 25*, 267–284. https://doi.org/10.1006/appe.1995.0061

Thaler, R., & Sunstein, C. (2009). *NUDGE: improving decisions about health.* Wealth, and Happiness.

Thode, H. C. (2002). *Testing for normality* (1st ed.). CRC Press.. https://doi.org/10.1201/9780203910894

Thompson, B. (2007). Effect sizes, confidence intervals, and confidence intervals for effect sizes. *Psychology in the schools, 44*, 423–432. https://doi.org/10.1002/pits.20234

Tomkins, S., Isley, S., London, B., & Getoor, L. (2018). Sustainability at scale: towards bridging the intention-behavior gap with sustainable recommendations. In *Proceedings of the 12th ACM Conference on Recommender Systems, Assoc. Comput. Mach* (pp. 214–218). https://doi.org/10.1145/3240323.3240411

Trattner, C., & Elsweiler, D. (2017). *Food recommender systems: important contributions.* Challenges and Future Research Directions. https://doi.org/10.48550/arXiv.1711.02760

United Nations, The 17 goals | Sustainable development. https://sdgs.un.org/goals, n.d. (accessed June 6, 2023).

Valcarce, D., Bellogín, A., Parapar, J., & Castells, P. (2020). Assessing ranking metrics in top-N recommendation. *International Journal of Information Retrieval Research, 23*, 411–448. https://doi.org/10.1007/s10791-020-09377-x

Wan, M., Wang, D., Liu, J., Bennett, P., & McAuley, J. (2018). Representing and Recommending shopping baskets with complementarity, compatibility and loyalty. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management* (pp. 1133–1142). https://doi.org/10.1145/3269206.3271786

Wang, S., Hu, L., Wang, Y., Cao, L., Sheng, Q. Z., & Orgun, M. (2019). Sequential recommender systems: Challenges, progress and prospects. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, International Joint Conferences on Artificial Intelligence Organization.* https://doi.org/10.24963/ijcai.2019/883

Wang, S., Hu, L., Wang, Y., Sheng, Q., Orgun, M., & Cao, L. (2020). Intention nets: Psychology-inspired user choice behavior modeling for next-basket prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence* (pp. 6259–6266). https://doi.org/10.1609/aaai.v34i04.6093

Wei, K., Huang, J., & Fu, S. (2007). A survey of E-commerce recommender systems. In *2007 International Conference on Service Systems and Service Management* (pp. 1–5). https://doi.org/10.1109/ICSSSM.2007.4280214

Willett, W., Rockström, J., Loken, B., Springmann, M., Lang, T., Vermeulen, S., et al. (2019). Food in the anthropocene: the EAT-lancet commission on healthy diets from sustainable food systems. *The Lancet, 393*, 447–492. https://doi.org/10.1016/S0140-6736(18)31788-4

Xia, Y. (2020). Correlation and association analyses in microbiome study integrating multiomics in health and disease. In J. Sun (Ed.), *Prog. mol. biol. transl. sci* (pp. 309–491). Academic Press. https://doi.org/10.1016/bs.pmbts.2020.04.003.

Yang, L., Hsieh, C.-K., Yang, H., Pollak, J. P., Dell, N., Belongie, S., et al. (2017). Yum-Me: A personalized nutrient-based meal recommender system. *ACM Transactions on Information Systems, 36*, 1–31. https://doi.org/10.1145/3072614

F. Yu, Q. Liu, S. Wu, L. Wang, T. Tan, A dynamic recurrent model for next basket recommendation, 2016. https://doi.org/10.1145/2911451.2914683.

Yuan, M., Pavlidis, Y., Jain, M., & Caster, K. (2016). *Walmart online grocery personalization: behavioral insights and basket recommendations.* Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-47717-6_5