TOWARDS TRANSPARENT RECOMMENDERS: AN EXPLANATION-BASED NEGOTIATION APPROACH

A Thesis

by

Berk Buzcu

Submitted to the
Graduate School of Sciences and Engineering
in Partial Fulfillment of the Requirements for
the Degree of

Master of Science

in
Department of Computer Science

Özyeğin University
October 2023

TOWARDS TRANSPARENT RECOMMENDERS: AN EXPLANATION-BASED NEGOTIATION APPROACH

Approved by:

Assistant Professor Reyhan Aydoğan
Department of Computer Science
Özyeğin University

Assistant Professor Furkan Kıraç
Department of Computer Science
Özyeğin University

Assistant Professor Fatma Başak Aydemir
Department of Computer Science
Boğaziçi University

Date Approved: Aug 11, 2023

For the love of learning, and the joy of sharing knowledge

ABSTRACT

As more and more recommendation systems are used in different areas and they are exposed to more ethical concerns, there is a growing demand for transparent and persuasive interactions with these systems. Toward this end, incorporating explainability in recommendation systems has emerged as a promising approach to enhance sociability and user trust. This thesis focuses on recommendation systems that utilize explainability techniques to foster sociability by providing precise and understandable explanations for their recommendations. The proposed recommendation system utilizes a combination of data-driven transparent mechanisms and human-agent negotiation approaches. The system generates personalized recommendations based on individual preferences and other similar user-tailored factors and engages in a negotiation with the users via discussions through explanations and real-time feedback mechanisms. The system reacts to user responses online, tailoring subsequent recommendations and explanations to convince the user.

This thesis encompasses Nutrition Virtual Coach (NVC) agents that generate personalized food recommendations based on individual factors like allergies, eating habits, lifestyles, and ingredient preferences. It mainly focuses on explanation generation techniques to enhance the transparency and trustworthiness of the system by improving the NVC agent's sociability in multiple steps. Ultimately, we incrementally conducted multiple experiments with participants from various backgrounds to evaluate the acceptability and effectiveness of the system.

The findings from the experiments generally indicate that most participants appreciate the opportunity to provide feedback and receive explanations for the given

recommendations. The participants prefer receiving information tailored to their specific needs and expectations. Additionally, the participants expressed their thoughts on various forms of explanations. The findings indicate that comparative explanations are not appreciated as much as informative explanations. The users seem to prefer direct and simple explanations that explain items respectively.

ÖZETÇE

Farklı sektörlerde giderek artan öneri sistemlerinin kullanımıyla ve yükselen etik kaygılarla birlikte, bu sistemlerle etkileşimde bulunurken artan bir şeffaflık ve açıklayıcılık ihtiyacı bulunmaktadır. Bu nedenle, yapay zekanın açıklanabilirliği, öneri sistemlerine entegre edilerek, etkileşimi teşvik etme ve kullanıcı güvenini artırma amacıyla umut verici bir yaklaşım olarak öne çıkmıştır. Bu tez, açıklanabilirlik tekniklerini kullanarak öneri sistemlerinde sosyal etkileşimleri teşvik etmeye odaklanmaktadır.

Tasarlanan öneri sistemi, şeffaf veri tabanı mekanizmaları ve insan-merkezli müzakere yaklaşımlarının bir kombinasyonunu kullanmaktadır. Bu sistem, bireysel tercihler ve benzeri kişiselleştirilmiş faktörlere dayalı olarak kişiye özel öneriler sunmakta ve açıklamalar ile gerçek zamanlı geri bildirim mekanizmaları aracılığıyla kullanıcılarla müzakere etmektedir. Sistem, kullanıcıların çevrimiçi yanıtlarına tepki vererek sonraki önerileri ve açıklamaları, kullanıcıları ikna edecek şekilde uyarlamaktadır.

Bu tez, yiyecek alerjileri, yeme alışkanlıkları, yaşam tarzları ve yemek içeriği tercihleri gibi kişisel faktörlere dayalı olarak kişiselleştirilmiş gıda önerileri sunan Sanal Beslenme Koç'larını ele almaktadır. Temel olarak, Sanal Beslenme Koç'unun şeffaflığını ve güvenilirliğini artırmak amacıyla sistemin etkileşimini birden fazla adımda geliştirmeye yönelik açıklama uyarlama tekniklerine odaklanmaktadır.

Sonuç olarak, sistem kabul edilebilirliğini ve etkililiğini değerlendirmek için farklı geçmişe sahip katılımcılarla aşamalı olarak çok sayıda deney gerçekleştirmiştir. Deneylerden elde edilen sonuçlar genel olarak katılımcıların geri bildirim sağlama ve verilen önerilere ilişkin açıklamalar alma fırsatını takdir ettiğini göstermektedir.

Katılımcılar, kendi özel ihtiyaçları ve beklentilerine uygun bilgi almayı tercih

etmektedirler. Ayrıca, katılımcılar çeşitli açıklamalarla ilgili düşüncelerini dile getirmişlerdir. Bulgular, karşılaştırmalı nitelikteki açıklamaların bilgilendirici nitelikteki açıklamalara göre tercih edilmediğini göstermektedir. Kullanıcıların, önerileri neden seçildiğini direk olarak belirten, doğrudan ve basit açıklamaları tercih ettikleri görülmektedir.

ACKNOWLEDGEMENTS

First and foremost, I would like to extend my gratitude to Assistant Professor Reyhan Aydoğan for her continuous support during my studies, both academically and personally. Her guidance, judgment and perseverance against my stubbornness has advanced me greatly in the field of Computer Science.

This work has been supported by the Scientific and Research Council of Turkey (TÜBİTAK, G.A. 120N680). We extend our gratitude towards TUBITAK for their financial aid program allowed this thesis to grow beautifully.

I thank my dear friends, Mehmet Onur Keskin and Gevher Yesevi as they were always there for me during my journey. Also, I thank my mentors and teaching assistants from the AI center, Furkan Kınlı and Barış Özcan, for all the wisdom and time they shared with me. Additionally, I must thank our hardworking interns at IISL, Berkecan Koçyiğit and Emre Kuru for being great mentees and working diligently during my time at IISL.

Also, I share this gratitude with Doğa Yılmaz, whom has spent countless hours in the lab with me throughout our master's degree journey.

Lastly, I thank my beloved, my family and all the names not mentioned here. I could not have come this far without their unconditional support.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

INTRODUCTION

With the rapid growth of data and the increasing complexity of algorithms, traditional recommender systems often need more transparency and provide adequate explanations for their recommendations. As a result, users are left in the dark, needing help comprehending the rationale behind the suggested items, leading to a lack of trust and satisfaction. Thus, incorporating explainability in recommender systems has emerged as a critical area of research and development in recommendation systems [2]. Researchers of the field address this issue by offering transparent and interpretable justifications for their recommendations, empowering users to make more informed decisions. This enhanced transparency builds user confidence and enables users to provide feedback and fine-tune their preferences on-line, ultimately building a more personalized and engaging recommendation experience.

This thesis focuses on Nutrition Virtual Coach (NVC) systems, specifically in the context of health-oriented food recommendation systems, aiming to enhance their reliability and trustworthiness via explanations.

Non-communicable diseases, including cardiovascular diseases, chronic respiratory diseases, and diabetes, contribute to approximately 63% of all global deaths [1]. The World Health Organization underscores the preventable nature of these diseases by addressing common risk factors, particularly unhealthy nutritional habits and diets. Although numerous recipes are freely accessible (i.e., via many online collectors), determining the "best" recipe for a *specific* individual in a *given* situation can be remarkably complex. Indeed, it involves managing a wide range of possibilities while

[1] https://www.who.int/news-room/fact-sheets/detail/noncommunicable-diseases

considering bounding variables such as allergens, nutritional values, personal requirements, calorie intake, historical data, and momentary preferences. Consequently, there is a need for a personalized support system. However, individuals' dietary choices are influenced by personal preferences, cultural norms, and taste, often leading to the consumption of tasty yet potentially unhealthy components concealed within processed food items. As a result, society requires expert guidance to make appropriate and sustainable dietary decisions [3, 4, 5]. With this intuition, nutrition virtual coaches (NVCs) are systems that aim to recommend recipes that align with users' specific needs and preferences while considering their health and long-term needs [6]. However, they often lack transparency and clarity, leading to a lack of trust and effectiveness [5]. In this context, a transparent and convincing recommendation system is important, as it can convince individuals to make healthier food choices while considering their unique preferences and constraints, thereby promoting better health outcomes and overall well-being.

In pursuit of this objective, the popularity of food recommender systems, designed to aid individuals in choosing recipes, has surged [4, 5]. The demand for such systems arises from the growing impact of globalization, which has expanded the accessibility and diversity of food options. Additionally, the widespread prevalence of processed food has been linked to metabolic and overweight problems [7], emphasizing the necessity for these recommender systems to guide users towards healthier and more balanced food choices. Moreover, some studies have proposed semantic models [8] and incorporated negotiation techniques to guide users towards desired quality of life goals [9] via convincing explanations accompanying recommendations. While these efforts have contributed to the field of recommender systems, to the best of our knowledge, there is currently no existing system that fully qualifies as an explainable Nutrition Virtual Coach (NVC). Such a coach would provide recommendations, explain them to the user, and engage in interactive discussions to foster desired behavioral changes.

Engaging the user in a back-and-forth communication is crucial as it allows the user to dive into the concept and build a more solid and backed-up knowledge/awareness that undoubtedly boosts information retention. Such mechanisms can assume a rather simplistic – yet effective – form of feedback [10]. Building on that, verifying/fixing misunderstandings and elaborating on follow-up questions becomes more feasible (from a designer/developer perspective) and easy to handle (from a user perspective). Additionally, we consider that there is a conflict of interest between the human-user of the system and the NVCs given the humans tendency to make their food selection based on their taste preferences rather than the healthiness of a recipe.

Therefore, this thesis first presents an end-to-end negotiation framework for an interactive nutrition virtual coach system that can generate recommendations and explanations while receiving feedback and leveraging [11]. In order to experiment with this framework, we designed three key components: a straightforward health score calculation module, a multi-criteria additive utility function to prioritize recipe rankings, and an OWL ontology for classifying and linking users and ingredients most suitably. Additionally, we utilize a simple explanation generation module. Later on, we incorporate simpler forms of explanations and gradually improve upon the explanation generation techniques of the system derived from the different forms of explanations in the literature. Accordingly, we examine the existing approaches, and adopt them in our Nutrition Virtual Coach. Finally we conduct user experiments to compare various approaches for explanation generation in terms of effectiveness and user preferences.

This thesis is organized as follows. Chapter 2 provides essential background information relevant to the scope of this thesis. Chapter 3 examines existing works and research within the field, shedding light on the relevant literature. Chapter 4 introduces the explainable interactive negotiation framework. Chapter 5 incorporates post-hoc explanation techniques from existing literature in the Explainable AI domain to further enhance the system. Moreover, Chapter 6 explores alternative explanation techniques, by integrating them with recommendation generation for a more cohesive and integrated approach. Finally, Chapter 7 summarizes our findings according to our experimental results, and potential future work directions. The key findings indicate a strong vein to explore within the domain of goal oriented recommendation systems.

CHAPTER II

BACKGROUND

In this chapter, we aim to lay the foundations of the techniques used within the scope of this thesis. Section 2.1 explains the concept of recommender systems, specifically the techniques this thesis draws inspiration from, Section 2.2 elaborates on the Explanation generation techniques available in the literature.

2.1 Recommender Systems

Recommender systems are software applications or algorithms that provide personalized suggestions to users, recommending items, products, content, or actions based on their historical interactions and preferences. These systems collect and analyze data on user profiles, item descriptions, and user-item interactions, utilizing recommendation algorithms such as Collaborative Filtering (CF) and Content-Based Filtering (CBF) to generate tailored recommendations. Recommender systems find extensive use in various domains, including e-commerce, music and food recommendation, enhancing user experiences by facilitating the discovery of relevant and engaging items, ultimately benefiting the system stakeholders through increased user engagement and satisfaction. The main intuition of such systems is approximating items either based on their properties (CBF) or user related properties (CF).

Collaborative Filtering is one of recommender systems' oldest and most widely used techniques. It is based on the concept that users who have previously shown similar preferences will likely agree on future preferences. This idea originated from "word-of-mouth" recommendations, where people trust suggestions from those with similar tastes [12]. CF's core idea is to find user similarities based on their interactions. This can be done using various similarity metrics, such as Cosine similarity [13] or

Pearson correlation coefficient [14]. These metrics measure how closely one user's preferences align with those of another user.

Content-Based Filtering emerged from the need to recommend items based on their content attributes, such as textual descriptions, metadata, or features. It found its roots in information retrieval and natural language processing [12].

2.2 Explanation Generation

The most suitable strategies presented in the literature to generate explanations could be classified as follows:

- **User-Centered Explanation:** The generated explanations are meant to assist users in achieving their goals. Sovrana and Vitali [15] emphasize that users are satisfied with the explanations if they are guided in answering the questions about the process of fulfilling their goals. An explanation such as "*we recommend you the following food recipe to lose weight since it has low fat and rich fiber*" could be considered an instance of this explanation type. It implicitly answers what is necessary to lose weight, which aligns with the hypothetical user's goals.

- **Knowledge-Based Explanations:** These explanations are generated by inferring some formal rules and facts in the knowledge base. For instance, a recommendation engine can offer a camera with less memory and resolution by referring to the rule that states "Less memory $\wedge$ lower resolution $\rightarrow$ cheap" [16]. Such rules need to be given to the system, and they can be derived from a decision tree modeling the system's or user's behavior. In other words, decision trees could be utilized to learn why the underlying decision is made from the data, and the rules extracted from the constructed decision tree can give insights on how the system works to the user as an explanation [17].

- **Example-Based Explanations:** Based on historical data or previous experiences, a system can generate some explanations by generalizing past behaviors/-patterns for a given new situation [18]. For example, assume that a food recipe consisting of sugar-free ingredients was recommended to a diabetic person by a recommender system that recommends food to ill people, and the results were satisfactory. If a new diabetic person joins the system, it might generate the following explanation alongside its recommendation "Diabetic people are often satisfied with this food recipe with sugar-free ingredients.".

- **Content-Based Explanations:** Inspired by the content-based recommendation approach, the system can analyze the features of the items appreciated by a particular user and extract the preferred values for those features to explain the recommended item to that user [16]. For instance, the system can generate an explanation such as "This food recipe contains mozzarella, so you might like it." if the user previously liked the food recipes that contain mozzarella specifically.

- **Contextual Explanations:** External factors affecting the decision could be used to generate such explanations. For instance, "Today fish is fresh. It has just arrived. Therefore, I recommend creamy salmon pasta." [8].

- **Contrastive Explanations:** A recent review by [19] provides empirical evidence supporting the practical utility of everyday contrastive explanations, "comparing a certain phenomenon with a hypothetical one" [18]. While asking about a certain choice, someone may think of alternatives and wonder why those were not recommended with respect to the given one. Contrastive explanations focus on the difference between the current choice and alternative ones. For instance, "We were going to recommend you a healthier option, which is Turkish Salad instead of American Salad that contains a substantially higher amount of fats.”.

- **Counterfactual Explanations:** On the one hand, they are similar to contrastive explanations and focus on the differences between alternative options. On the other hand, these explanations rely on hypothetical factors instead of factual factors [8]. For instance, “If you did not have an allergy to seafood, I would recommend you a salmon salad. However, I have to recommend you turkey salad.”.

CHAPTER III

LITERATURE REVIEW

This section briefly overviews the literature on recommender systems, particularly focusing on food recommendation systems and their evolution to embrace explainable and interactive recommendations.

3.1 Post-Hoc Explanation Generation Mechanisms

In recent years, there has been ample research focusing on developing techniques for post-hoc explanation generation in various domains within the Machine Learning literature. These techniques are designed to explain the predictions made by complex black-box models. They operate "post-hoc", meaning they generate explanations after the main model has made its predictions, without requiring modifications to the underlying model's architecture or training process. The goal is to improve transparency and interpretability by providing human-understandable justifications for the model's decisions. Post-Hoc Explanation Generation models often leverage feature importance analysis [20], rule-based reasoning [21], gradient-based attribution [22], or surrogate models [23] to generate meaningful explanations that can shed light on the factors influencing the model's predictions. These explanations help stakeholders gain insights into how the given model arrives at its decisions, builds trust, and facilitates error analysis, making them valuable tools for practical applications and model understanding [24].

3.2 Explainable Recommendation Systems

Recent studies have emphasized incorporating explanations into recommendations to enhance transparency, trust, and acceptability. Although explanations in recommender systems still need to be fully widespread, some approaches (or combinations of them) are gaining attraction. Tintarev and Masthoff [16] investigate various aspects of explanations' impact, such as transparency, scrutability, trustworthiness, persuasiveness, effectiveness, efficiency, and satisfaction. All of these attributes contribute to enhancing the reliability of the system as supported by Gedikli *et. al.* [25] in which they assessed varying explanation attributes, including user satisfaction, efficiency, effectiveness, and trust, by evaluating different explanation styles in recommender systems. Additionally, Herlocker *et. al.* [26] investigates the effects of varying techniques used to explain collaborative-filtering recommendation methods where they follow the principle of collaborate filtering; they show ratings of other users similar to their interests to a user in the form of explanations. They could not reach indicate no concrete outcome according to their assumption within the groups of ordinary users, however, they find out their system makes it more convenient for an expert to sympathize with a recommendation (e.g., the group of experts are more fond of the system with higher success in predictions of user acceptance). Sharma and Cosley [27] devised a framework to investigate the influence of social explanations on decision-making processes within a music recommendation setting that supports this by concluding that varying form of explanations might have different effects depending on the person. Similarly, Milliecamp *et. al.* [28] explores the visual explanations within the music domain. The study shows that users react to explanations depending on their need for cognition, confidence, musical sophistication, and visualisation literacy. Explanations boost confidence for those with low need for cognition and speed up song judgments for those with higher musical sophistication. Users with lower visualisation literacy tend to judge songs more quickly and precisely. Furthermore, Pu and

Chen [29] develops an explanation interface to investigate the user experience advantages of using explanations for building trust and to assess whether system features can contribute to trust-related benefits. Their study shows that users prefer to re-use systems that offer explanations more often than those that do not. A major finding shows that users prefer a comparative explanation style where they get a broader view of available items and respective differences. Symeonidis *et. al.* [30] introduces a prototype for a movie recommender system designed to gauge user satisfaction via various explanation styles. They point out that providing an explanation along movie recommendations will increase the likelihood of a user estimating it's ranking of the movie while also increasing the amount of correct estimations to predicting an user's favorite movie by boosting the user's confidence in providing information to the system. These assumptions are shown to be correct by user studies.

3.3 Food Recommendation Systems

Recommendation systems have been employed in the nutrition domain for some time, with objectives ranging from promoting health, sustainability, and taste-based combinations. In 1986, Hammond et. al. [31] develops one of the earliest food recommender systems. It is named "CHEF" and leverages case-based planning to replace or improve food items within recipes. It requires a substantial initial knowledge base, extensive pre-processing, and the creation of (backup) plans for each recipe. More recently, in 2010, Freyne and Berkovsky [32] implemented recommender algorithms, such as collaborative filtering (CF) and content-based (CB) approaches, to recommend recipes. The study concludes that incorporating ingredient weights within CF and CB improved prediction accuracy. In turn, Ge et. al. [33] introduces the concept of personalization in food recommendations, prioritizing health over taste. Chi et. al. [34] focuses on recommending food for individuals with chronic conditions (i.e., kidney diseases) using an Ontology Web Language (OWL) ontology integrating health-relevant aspects. Chen et. al. [35] proposes a generalized framework for healthy recommendations, explicitly targeting the modification of unhealthy recipes. They introduce a deep learning-based method called IP-embedding to match recipes with desired ingredients, creating a pseudo recipe that meets the requirements and then matching it with healthy ingredients and real recipes using the mean squared error (MSE) metric. Similarly, Teng e ak. [36] developes a point-wise comparison metric to understand how to transform recipes using ingredient substitutions for healthier alternatives, and Elsweiler et. al. [37] addresses ingredient and food substitution, metricizing nutritional values to encourage users to prefer healthier options. Overall, the food recommendations approaches often rely on factors such as recipe content (e.g., ingredients) [38, 39, 40], user behavior history (e.g., eating history) [41, 42], and/or dietary preferences [42, 43]. Conventional food recommendation approaches so far are mostly "one-shot", offering to the user minimal (if any) possibilities to interact with the recommender engine/agent. Nevertheless, with the advent of explainable technologies pursuing predictors and classifiers transparency, understandability, and inspectability to boost trust [44], recommender systems (among the others) are expected to provide explanations for their recommendations [45, 38, 41, 39], allowing users to justify, control, and discover new aspects of the suggested outcomes [46, 42, 43]. Along this line, Padhiar et. al. [8] proposes a food recommender system that generates explanations based on a knowledge-based ontology. However, the explanatory system only attempts to explain a given recommendation via different methods, with no dialogue option (e.g., no way for the user to reply/interact). Samih et. al. [47] further explores this concept by developing a knowledge-based explainable recommender system that makes use of a probabilistic soft-logic framework to generate explanations. Lawo et. al. [9] aimed to enhance the interaction between users and virtual assistants by incorporating a cluster of consumers with ethical and social priorities into the recommendation process and considering their feedback and preferences.

CHAPTER IV

INTERACTIVE EXPLAINABLE RECOMMENDER

Goal-oriented recommender systems offer tailored recommendations to users based on their specific objectives and individual needs. These systems are designed to guide users toward achieving their desired outcomes, whether finding a product to purchase, discovering relevant content, or completing a specific task. They find applications in various domains, such as e-commerce, travel planning, and food recommendations. However, these systems can sometimes overlook specific user preferences as they prioritize recommendations aligned with the system's predefined goal. This misalignment can lead to conflicts between users and the system when attempting to reach a good outcome by the system's goals. For instance, consider a healthy lifestyle-oriented food recommender with a user preferring food recipes that are not necessarily healthy but delicious for them. To address these conflicts, designers of such systems must strike a delicate balance between meeting users' immediate goals and providing a comprehensive, personalized user experience. Incorporating mechanisms for user feedback and offering explanations for recommendations play vital roles in mitigating these conflicts and enhancing the overall user experience. Therefore, we model the conflict resolution scenario as a negotiation in a dialogical setting, where the system and the user strive to find a consensus while considering their respective perspectives on the user's goals and needs. This way, the system may utilize the sociability of the system as a means to convince the user to utilize it's goal-oriented recommendation. In this study, we employ a food recommendation case study to validate our methods. This choice is driven by the apparent conflict arising from divergent user preferences concerning taste and the healthiness of food. Such systems are often called Nutrition Virtual Coaches (NVC), which seek to improve lifestyles of system users through recommendations.

Accordingly, this chapter proposes a framework leveraging ontology-based reasoning to interact with the user in a turn-based fashion for a recommender system. The goal is to understand the users' needs and interests while making recommendations, as it considers the user's preferences during the recommendation process. Section 4.1 describes the explanation-based negotiation protocol that governs the interaction between the user and the recommender system. Section 4.2 presents our case study and it's components. Section 4.4 elaborates on the acceptability and appreciation of our proposed framework.

4.1 Explanation-based Negotiation Framework

Since we formulate our solution in the form of negotiation, the development of a protocol is essential to establish a structured dialogue that defines how the users and the agent will exchange offers, explanations and feedback, and reach agreements during the recommendation process. Toward this end, we utilize the alternating offers protocol commonly used in negotiation [48] and alter it to accommodate explanations and feedback options as illustrated in Figure 1. In the context of food recommendation, the user first specifies their constraints (C), which may consist of the ingredients the user may be allergic (e.g., milk, peanuts) to; the (dis)liked ingredients (e.g., specific meat/vegetables); and the desired type of cuisine (e.g., Middle Eastern, Italian, French). After receiving the user's constraints, the agent makes a recommendation (R) along with its explanation (ϵ). The user can *accept* R, *leave* without an agreement, *criticize* R, ϵ, or both. When the user makes a critique, the agent can revise its recommendation/explanation, generating (R') or (ϵ'), or both. This interaction

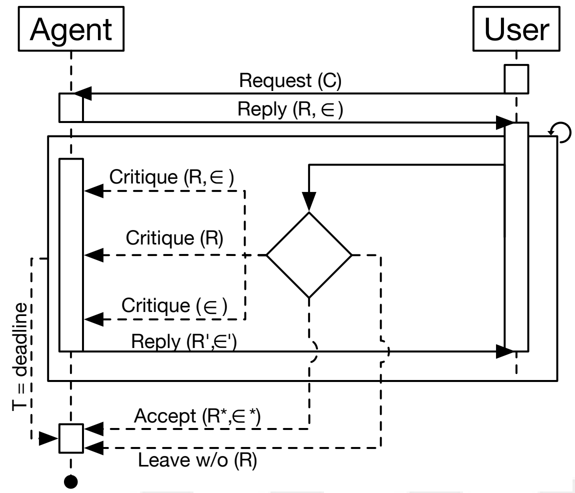


Figure 1: FIPA Ddescription of the Negotiation Protocol

continues in a turn-taking fashion until reaching a termination condition (i.e., Accept or Leave w/o Recommendation) or the time deadline is reached. An user can criticize the given recommendation by referring to pre-structured critiques as follows, where Y denotes one of the ingredients chosen by the user. (i) "I ate Y recently", (ii) "I'm allergic to Y", (iii) "I don't like Y", and (iv) "I want to give custom feedback". Similarly, the user can criticize the explanations communicated alongside the recommendations with the pre-defined statements such as (i) "The explanation is not convincing", (ii) "The explanation does not fit my case", (iii) "The explanation is incomplete", (iv) "The explanation is not clear enough", and (v) "I disagree with the explanation".

4.2 Case Study: Nutrition Virtual Coach

The popularity of unhealthy food choices is made worse by the abundance of popular internet recipes that promote unhealthy eating habits [49]. The goal of the NVCs is to help users make better dietary choices and reduce health risks through personalized recommendations. To achieve this, it is important to address the conflict between user preferences and the healthiness of recipes. Additionally, the system well-founded explanations for why a particular recipe is recommended could go a long way to persuade users to improve their eating habits. For such an NVC, we design an OWL-based Ontology database that includes ontological concepts to represent *users* and *food ingredients* crucial to form relations between the entities of the system. The *User* concept characterizes the individuals and their eating habits, including any allergy, religious, and lifestyle restrictions. The food concept is characterized by recipes and ingredients that are grouped in classes (e.g., cow-hearts, cherry tomatoes, etc. are grouped under the category of *Tomatoes*). A comprehensive view from *Food* concept in the Protege is shown in Figure 2.

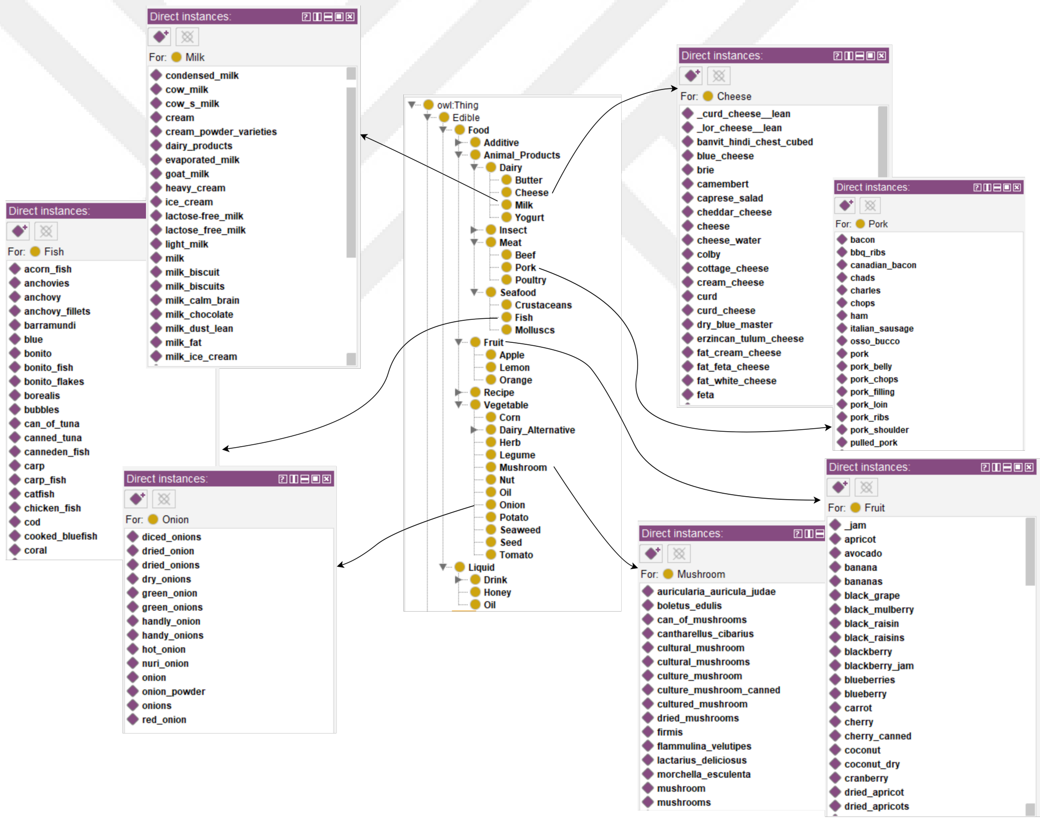


Figure 2: Protege View of the Food Class

We establish the object property of *doesNotEat* to identify which food ingredients the user would/should avoid as seen in Figure 3. The limitations, such as the prohibition of pork for Muslims, are represented by linking object properties (depicted as diamonds) to both the "User" and "Food" concepts. The system verifies whether a particular user class would/could consume a given ingredient class by the *doesNotEat* relation between users and food ingredients. We utilized a compact and localized recipe dataset [50] to build the ontology instances by fitting the ingredients into the respective concept structure manually. We annotated the recipe ingredients by the classes of ingredients within the ontology. A final filter on recipes with incomplete information leaves 1.3K recipes to recommend.

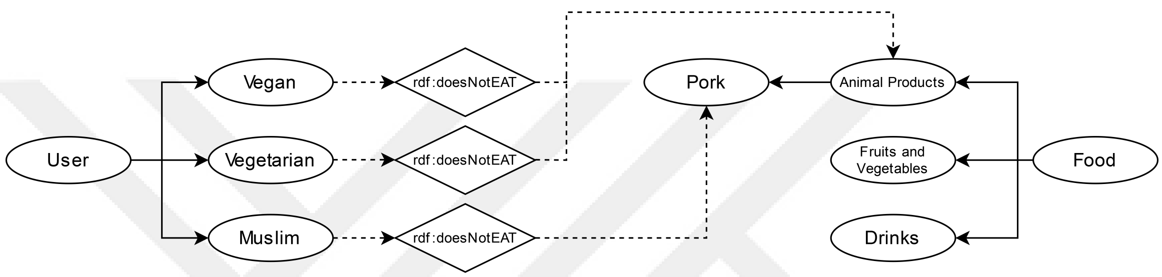


Figure 3: Broad Overview of the Ontological Structure for Food Concept

4.3 The Baseline Recommendation Strategy

In this section, we explain the main recommendation strategy of the food recommender system under the following outline. Section 4.3.1 explains the initial filtering and scoring of the food recipes under various modules. Then, Section 4.3.2 elaborates the utility function used in determining which recipes to recommend from a healthiness perspective. Finally, 4.3.3 outlines the calculation of the user satisfaction score used in the utility estimation of the recipes.

4.3.1 Filtering and Scoring Recipes

To analyze the applicability of the designed protocol, we have developed a basic recommendation strategy relying on filtering and scoring the recipes concerning the user's constraints and healthiness (see Algorithm 1). First, the NVC agent filters the recipes according to the user's eating habits/constraints via ontology reasoning on what (classes of) ingredients the user would not consume (Lines 1–3). Assuming that the user is vegan, the NVC agent first filters the recipes containing animal-related products. Then, if the same user specifies that they do not like "zucchini", the NVC agent removes the recipes containing zucchini from the remaining candidate list, R_u. In turn, the utilities of the remaining candidate are calculated by considering both healthiness and their alignment with the user preferences. Then, the recipes are sorted according to the calculated utilities (Lines 4–5)[1]. The recipe with the highest utility is taken as a candidate recipe, and the system retroactively generates an explanation in line with the recipe's properties (Lines 6–7). This candidate recipe and its corresponding explanation are given to the user. When the NVC agent receives feedback from the user regarding the recipe, F_r, it filters the candidate recipes according to the updated constraints given by the feedback and selects the highest-ranked recipe similarly (Lines 10-15). When the NVC agent receives feedback from the user regarding the explanation, F_ϵ, it simply generates a new explanation with the underlying recipe (Lines 16–18).

4.3.2 Utility Estimation

To select the suitable recipe, this study relies on multi-criteria decision-making [51]. Multi-criteria decision analysis allows decisions among multiple alternatives evaluated by several conflicting criteria [52]. The adopted multi-criteria decision analysis is done by ranking recipes through a multi-criteria function. The multi-criteria function gives each recipe a score in the dataset. One of the main advantages of using a mathematical function is the transparency of the function and its outcomes. This feature is well suited for our proposed NVC due to the explainability of the generated behavior. The overall utility of the recipes, based on the multi-criteria, is computed by considering three criteria: Active Metabolic Rate (AMR) score, nutrition value score, and users' Satisfaction score. The final score of the recipes is the weighted sum

[1]The details of the utility calculation are explained below.

Algorithm 1 Baseline Recommendation Strategy & Explanation Generation

Require:

R: Recipes;
U: User;
$R_u \subset R$: Recipe dataset tailored for the user;
H_u: Eating habits of the user;
P_u: User Constraints/Preferences;
r_c: Candidate recipe;
ϵ: Explanation for candidate recipe;
F_r: Feedback to the recipes;
F_ϵ: Feedback to the explanation;

Ensure: r_c, ϵ

if firstRecommendation **then**
    $R_u \leftarrow$ filterRecipesByCondition(R, H_u)
    $R_u \leftarrow$ filterRecipesByCondition(R_u, P_u)
    $U_{R_u} \leftarrow$ calculateUtilities(R_u)
    $R_u \leftarrow$ rankRecipes(R_u, U_{R_u})
    $r_c \leftarrow$ getHighestRankRecipe(R_u)
    $\epsilon \leftarrow$ generateExplanation(r_c)
else
    if F_r exists **then**
        $R_u \leftarrow$ filterRecipesByCondition(R_u, F_r)
        $U_{R_u} \leftarrow$ calculateUtilities(R_u)
        $R_u \leftarrow$ rankRecipes(R_u, U_{R_u})
        $R_c \leftarrow$ getHighestRankRecipe(R_u)
    end if
    if F_ϵ exists **then**
        $\epsilon \leftarrow$ generateExplanation(r_c)
    end if
end if
return (r_c, ϵ)

of the score provided by each module as presented by Equation 1 where w_a, w_n, w_u denote the weights of each AMR score, nutrition value score, and users' satisfaction score, respectively. Note that each score is normalized to ensure that the overall score is ranged within [0,1].

$$recipeScore = w_n * nutrientsScore + w_a * amrScore + w_u * UsersScore \quad (1)$$

The nutrient-based score is calculated according to the nutritional information of the recipes, such as proteins, lipids, carbohydrates, cholesterol, sodium, and saturated fats. These nutrients have respective recommended amounts for a healthy life [1]. In this work, we take into account the nutrition intake limits specified by the WHO organization [2]. Accordingly, the nutrition-based score is calculated as seen in Equation 2, where each nutrition score is calculated according to Equation 3. We assume that consuming less than each nutrient's minimum amount (min_n) is better than its maximum amount (max_n). By following this heuristic, the individual score of each nutrient is calculated.

$$\begin{split} nutrientScore(recipe) = score(pro) + score(lip) + score(cb) + \\ score(ch) + score(sod) + score(sat) \end{split} \tag{2}$$

$$score(n) = \begin{cases} 5 & \text{if } n \in [min_n, max_n] \\ 3 & \text{if } n < min_n \\ 1 & \text{else} \end{cases} \tag{3}$$

AMR is the number of calories a person must consume daily depending on height, sex, age, weight, and activity level. Such preliminary information is taken during the registration of the users. The value of AMR is based on the value of Basal Metabolic Rate (BMR), the number of calories required to keep a body functioning at rest, the person's activity level, and the person's desire to maintain or reduce his current weight. Table 1 presents the values to keep the current weight. To compute the AMR score based on the minimum and maximum amount of calories required for a given user available in literature [1], we rely on the same assumption of Equation 3 that is consuming fewer calories than required ($score = 3$) is better than consuming more calories than required ($score = 1$). In addition, when the amount of calories computed is between the minimum and maximum amount of calories, the score is set to 5. Conventionally, the most used formula to compute BMR is the Harris equation [53] with Equation 4 and 5, for men and women, respectively. The authors

[2] https://www.who.int/news-room/fact-sheets/detail/healthy-diet

Table 1: Daily Recommended Kilo-calories (kcal) Intake to Maintain Weight [1]

Activity level	Daily calories
Too little exercise	$calories = BMR * 1.2$
Light exercise	$calories = BMR * 1.375$
Moderate exercise	$calories = BMR * 1.55$
Strong exercise	$calories = BMR * 1.725$
Very strong exercise	$calories = BMR * 1.9$

estimated the constants of Equation 4 and 5 by several statistical experiments [53].

$$BMR = 10 * weight + 6.25 * height - 5 * age + 5 \tag{4}$$

$$BMR = 10 * weight + 6.25 * height - 5 * age + 161 \tag{5}$$

4.3.3 User Satisfaction Score

The user satisfaction score is calculated by considering the recipe's popularity among all users and the current user's preferences equally. For the recipe's popularity, we use the ratings the other users gave between [1, 5]. These values are normalized to [0, 1]. Meanwhile, regarding the user's preferences, we check how many ingredient classes are considered to be liked by the user. Here, to determine whether an ingredient is liked or not, we can use explicit feedback from the user as well as rely on user profiling to predict whether the given ingredient is likely to be preferred to be consumed. Here, we use Jaccard Similarity [54] to estimate individual user satisfaction (the rate of the preferred ingredients over the number of all the ingredients of a given recipe). Let us assume the user-submitted his preference for ingredients (i_1, i_2, i_3) and we have a recipe such that $R_1 = i_1, i_2, i_5, i_6$. Each ingredient that exists with the liked constraint is considered to be 1 and 0 otherwise. The mean of this operation is 0.5, which is effectively the score of R_1 for this user. For all the recipes, the scores are then max-normalized to place the values between [0, 1], resulting in a relative level of importance for the given recipe. For instance, let us assume that the system knows that the user likes the ingredients i_1, i_2, and i_3 and calculate the score of a recipe consisting of the following ingredients:i_1, i_2, i_5, i_6. The individual user satisfaction would be 2/4, according to Jaccard similarity. If the overall user rating of that recipe is equal to 4 out of 5, then the overall score would be equal to 0.65 ((0.5+0.8)/2).

4.3.4 Explanation Generation

We present a straightforward explanation generation approach to demonstrate how the NVC agent interacts with its users. Two types of explanations are considered: health-related and preference-related. Six promising health-related explanations are generated from our dataset by taking the nutritional values into consideration, such as protein, vitamin coverage, and cholesterol counts. If the recommended recipe satisfies the given health conditions, we need to identify the main ingredient contributing to this condition, as seen in Table 2, where X is the amount of the nutrition. Note that more sophisticated explanations could be produced by consulting some nutritionists and enhancing ontology reasoning. Three different explanations are presented regarding the user preferences: chosen cuisine, chosen ingredient, and overall popularity of the recipe (i.e., community score S) as seen in Table 2. To generate such explanations, we match the ingredients of the recipe (i.e., Y, Z) with the corresponding explanation. Finally, the agent generates its complete explanations by combining a fixed starting statement and randomly chosen health and preference-related explanations.

4.4 Evaluation

The acceptability and appreciation of the proposed framework are evaluated via a user study involving 53 participants. This section presents the experimental setup and participants, and discusses the results of the user experiments.

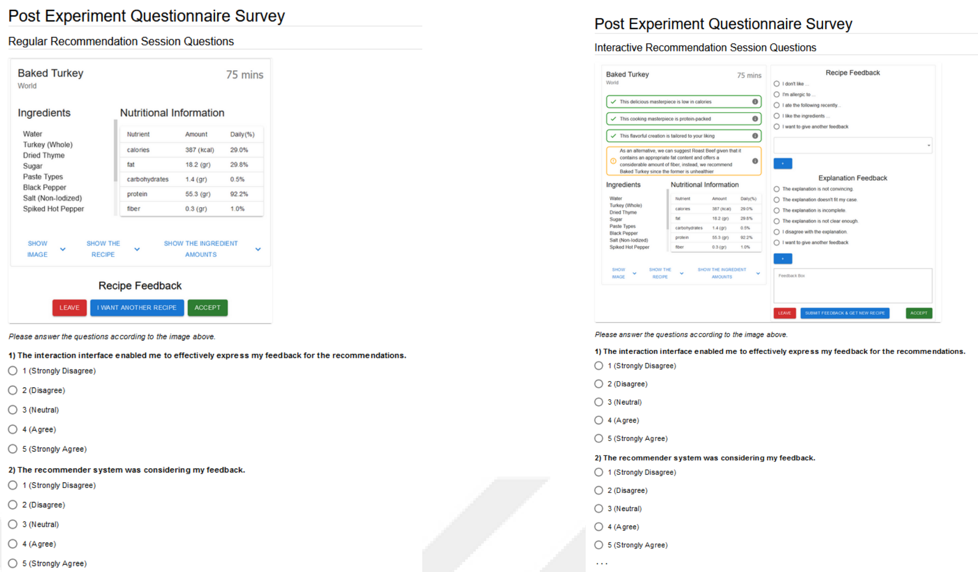
4.4.1 Experimental Setup

To investigate the effect of interactive explanations and critiques introduced by the proposed protocol, we leverage a Web-based platform allowing the users to experience both the explanation-based negotiation protocol (i.e., the interactive recommender) and its replica without the explanations and critiques component (i.e., a regular recommender) — see Figure 5. Prior to the experiments, each user is asked to fill out a pre-survey and registration form to specify their gender, age, height, weight, sports activity level, eating habits, and allergies[3]. This information is used to estimate the healthiness score of recipes recommended to the user (see Section 4.3). Seeking to instill different experiences, the two environments (i.e., regular and interactive e-coach/food recommenders) are proposed randomly to each user with a 5-minute break between the two sessions. At the end of the test, the users are asked to fill a questionnaire consisting of mostly 5-point Likert scale questions regarding their experience in both sessions, employing a within-subject design [55].

Table 2: Explanation Variables and Examples

Criteria	Example
Health related explanations	
Protein amount covers user needs for a meal	This recipe contains X grams of protein, which is about X% of your daily requirement. Your body needs proteins from your organs to your muscles and consuming the necessary amount of it is important!
Calorie count is above 30 % of BMR	You should eat this food because it covers a good portion of your necessary daily calorie intake by X. You should eat X for a balanced diet and maintain a healthy weight!
Vitamin B amount covers the user needs for a meal	B vitamins have a direct impact on your energy levels, brain function, and cell metabolism, and you should consume X gram for a healthy vitamin B level in a meal!
Vitamin C amount covers the user needs for a meal	Vitamin C is needed for the growth and repair of tissues in all parts of your body. This recipe contains: X% of it and eating it will make you feel more energetic!
The amount of iron must be enough for a meal	Lack of Iron in your system could be critical, causing a disease that is known as iron deficiency anemia. This recipe supplies you with X of it.
The cholesterol must be below a threshold	This recipe has a very low cholesterol count. Cholesterol is linked with a higher risk of cardiovascular disease, and this food is great for low amounts of it.
Preference related explanations	
Users' chosen cuisine matches the recipe cuisine	This recipe is also a part of the cuisine you like: Z
Users' chosen ingredient matches the recipe	This recipe contains the ingredient you wanted: Y_1 and Y_2.
The recipes community score is above a threshold	This recipe was rated S stars by the community, and you might like it too!

[3] We would like to state that the experiment protocol adopted in this study was approved by the Ethics Committee of Özyeğin university, and informed consent was obtained from all participants.



(a) Regular Recommender Session **(b)** Interactive Recommender Session

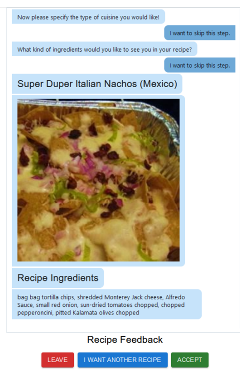
Figure 4: Regular and Interactive Recommendation Session Questionnaires

4.4.2 Participants

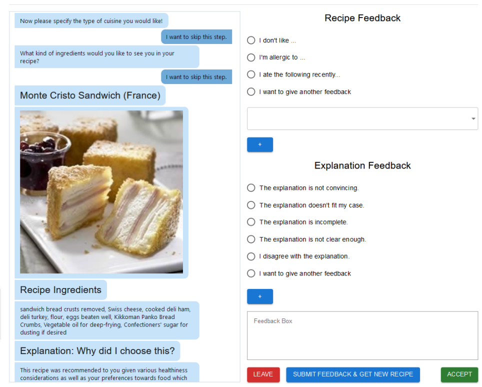
We have recruited 53 users (i.e., 33 men and 20 women with various backgrounds). The mean age is 25.9 with a max of 51 and the min is 19 years old. The participants were asked to rank the importance of five criteria: "Nutritional factors", "Past experience with taste", "How it looks", "Price of the ingredients", and "Cooking style" while making their decisions on a food recommendation. Figure 6 shows the histogram analysis of this ranking. Twenty-eight users seem to prefer the recipes they knew already. Ten users have chosen nutritional factors among those that they prefer or consider healthier.

4.4.3 Experimental Results

This section analyzes the findings of the tests and the users' responses to the post-test survey. Since the experiment is composed of two sessions (i.e., Interactive and Regular) and the comparative questions are the same for both, we performed a within-analysis statistical tests. The data is not normally distributed which is one of the main assumptions made by the pairwise T-test. Thus, we apply its corresponding non-parametric test called the Wilcoxon sign rank test [55]. For all tests, the Confidence Interval (CI) is set to 0.95, $\alpha = 1 - CI = 0.05$.



(a) Regular Recommender Session



(b) Interactive Recommender Session

Figure 5: Regular and Interactive Recommendation Session Interfaces

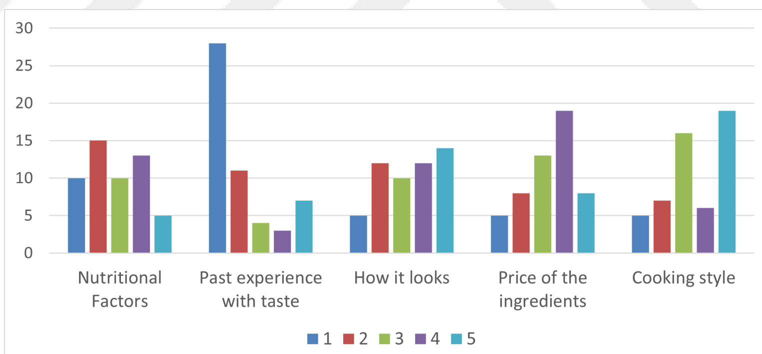
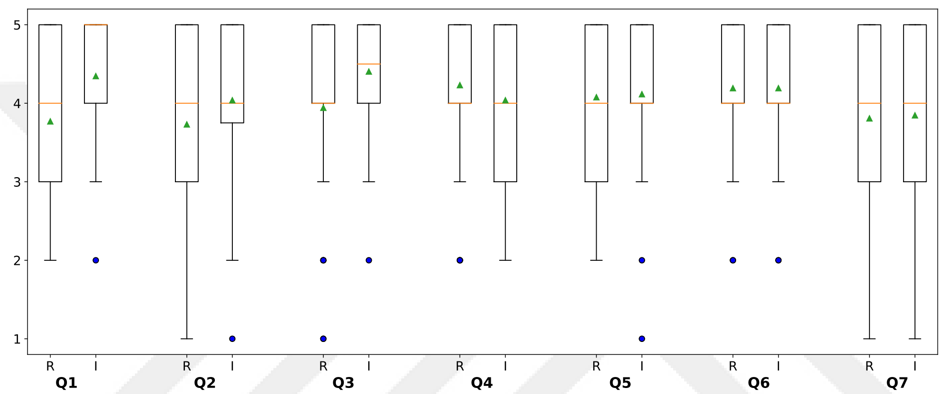


Figure 6: Pre-Survey Questionnaire Ranking Question Histogram

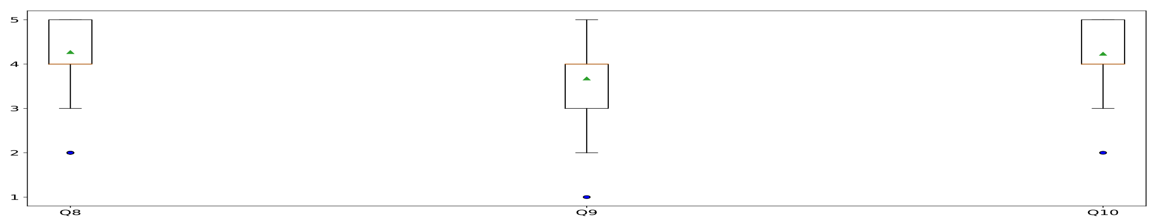
Figure 7 shows the box plot (including the mean) and the p-values of the first set of questions in the experiments between the regular and interactive sessions respectively. The yellow lines represent the median, the triangles in green the means, and the small blue circles the outliers. The results show that for Q1 (p=0.001) and Q2 (p=0.011) the interactive session is significantly better than the regular session. Indeed, these two questions aim at the sociability of the system, and the difference may stem from the engagement provided by the interactive session. Q3 (p=0.002) is about the completeness of the system. The users answered that the interactive session provided a better set of information than the regular session to make an informed decision. Q4 (p=0.931), Q5 (p=0.431), and Q6 (p=0.506) qualify the usability of the system. Assessing them, we can convene that adding an interactive dimension to the system can still be effective and efficient. Nevertheless, only a minor part of the users (12 out of 53 users or 23% of them) still prefer the regular one over the interactive system. Furthermore, we questioned the users regarding their experiences with the



Id	Questions label	Metric	W	p
Q1	The interaction interface enabled me to effectively express my feedback for the recommendations.	Sociability	52.0	**0.001**
Q2	The recommender system was receptive to my feedback.	Sociability	52.5	**0.011**
Q3	The recommender system gave me enough information to decide.	Completeness	40	**0.002**
Q4	The recommender system was easy to use.	Usability	170.5	0.931
Q5	I found a recipe without consuming too much time.	Usability	157.0	0.436
Q6	The interaction with the system was smooth and effortless.	Usability	138.5	0.506
Q7	I would like to use such a system in my daily life.	Acceptability	34.5	0.356

Figure 7: Box Plot and p-values of the First Set of Questions

explanations, as illustrated in Figure 8. Concerning Q8, mean, median, and mode are greater or equal to 4 ("Agree"). Moreover, three quartiles of Q8 are equal or greater than 4 ("Agree") which means most of the users (about 75%) were satisfied with the explanations. Regarding question Q9 related to the usefulness of the explanations, two quartiles are greater or equal to 4 ("Agree") and two quartiles are between 2 ("Disagree") and 3 ("Neutral"). That means at least half of the participants agree with the usefulness of the explanation while others either disagree or are neutral. Finally, concerning Q10, three quartiles are equal or greater than 4 ("Agree"), which means most of the users appreciate receiving explanations in addition to recommendations. Moreover, since the mode is 5, users who strongly agree are greater than those who agree to receive explanations.



Id	Questions label	Metrics	Mode	Median	Mean	std
Q8	The explanations exchanged during the interaction were satisfactory.	Satisfaction	4	4	4.26	0.79
Q9	The explanations for recommendations have impacted my choice.	Usefulness	4	4	3.67	1.02
Q10	Rate your appreciation of the idea of receiving explanation in addition to recommendations.	Usefulness	5	4	4.23	0.80

Figure 8: Average Ratings of The Explanation Related Questionnaire

Finally, the experiment revealed that while users scored the looks of an image to be less important than other factors contributing to their decision, at the end of the experiment, a considerable 55% of them gave the highest score (Strongly Agree) to whether the images influenced their decision or not.

CHAPTER V

POST-HOC EXPLANATION GENERATION TECHNIQUES

This chapter proposes a "Post-Hoc" explanation generation technique to improve the transparency and the sociability of the food recommender system to nudge the users to consume healthier food while utilizing the explainable negotiation-based recommendation framework as described in Chapter 4. We explore the effectiveness and convincability of such techniques toward users of the system. Such techniques often explain the "why" an outcome has been reached by a model for the system developer to understand the reasoning behind a model that is otherwise uninterpretable. We map these techniques to explain the reasoning behind a recommendation, making it an end-user-facing process that requires generating a text that a human can understand. Section 5.1 elaborates how we utilize the concept of post-hoc explanations in our recommendation framework, Section 5.2 explains the text generation method that is made to represent explanations to the user and Section 5.3 details our experimental results from an online user study.

5.1 Post-Hoc Explanation Generation Strategies

The post-hoc explanation generation techniques are commonly adapted by explainability related literature [56] given the need to explain black-box models. These techniques follow the process illustrated by Figure 9. The system has a black box model that makes predictions from the given input, which is then explained by another module often called the "*explainer module*", commonly consistent of a simpler, more interpretable machine learning method such as decision trees [57].

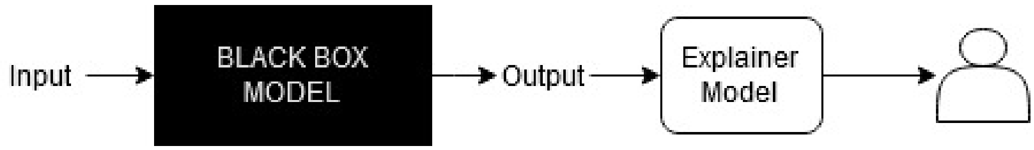


Figure 9: Sample Tree for Item-Based Explanations

Section 5.1.1 elaborates on our use of decision trees to explain given food recommendations. Section 5.1.2 explains the contrastive food recommendations, where we offer an alternative and explain the differences between the recommended item and the contrasting item. Finally, Section 5.2 explains how all these approaches are combined and turned to human-readable texts.

5.1.1 Item & User-Based Explanations

We consider explanations from two perspectives: historical data gathered from other users' outcomes (User-based Explanations) and the recommended item's properties (Item-based Explanations). Combining these two dimensions offers a broader set of explanations from both a system-global perspective and a user-specific item-related perspective. Towards the generation of such explanations, decision trees can be used because they are simple and intuitive models that can be easily understood and visualized. They can explain the reasoning behind AI predictions or decisions in a more straightforward form than an otherwise black-box model [44]. Decision trees determine feature importance by measuring how much each feature reduces the impurity or disorder in the data when used for splitting, with features that consistently reduce impurity more being considered more important for making predictions, which is exploited by the proposed recommendation system. Thus, exposing an exploitable outcome to understand which features are important to utilize for the explanation generation. Accordingly, we propose a technique to generate explanations as shown in Algorithm 2. In order to discover the important features significantly influencing users' decisions (e.g. carbohydrates, protein, etc.), a decision tree is constructed from a labelled dataset (see Line 1). When we employ the user-based explanation generation method, the decision tree is constructed from historical data in which the data is labelled with **all users' decisions** (i.e., accept or reject). Conversely, the item-based explanation generation approach utilizes the decision tree constructed from data labelled according to **the current user's constraints and feedback**. For that tree, filtered and low-scoring items are are negatively labelled (-1), the items that aligned with the user's constraints are positively labelled (+1) and the rest is labeled neutrally (0). After sorting features with respect to their importance (Line 2), we choose three of them to generate an explanation for the given item (Lines 3–4).

Algorithm 2 Item-Based/User-Based Explanation Generation

Require:
  ϵ_T: Selected features for explanation templates
  R_l: Set of labelled items;
Ensure: ϵ: Explanations of a given item
1: tree $\leftarrow$ DecisionTreeClassifier(R_l);
2: sortedFeatures $\leftarrow$ tree.getExplanationTags()
3: **for** i = 0; i < 3; i + 1 **do**
4: $\quad \epsilon_T \leftarrow \epsilon_T \cup sortedFeatures.nextImportantFeature()$
5: **end for**
6: return ϵ_T

In our Nutrition Virtual Coach use case, we utilize the proposed algorithms with a feature set comprised of recipe's nutritional information (Carbohydrates, Protein, Fiber, Fat and Calorie scores), preparation and cooking times of the recipe, and the preference scores. Figure 10 illustrates a sample item-based tree from one of the live experiment participants' data. For this participant, one could observe that the protein is the most important decision factor for the constructed tree, as it is also visible on Figure 11 as well.

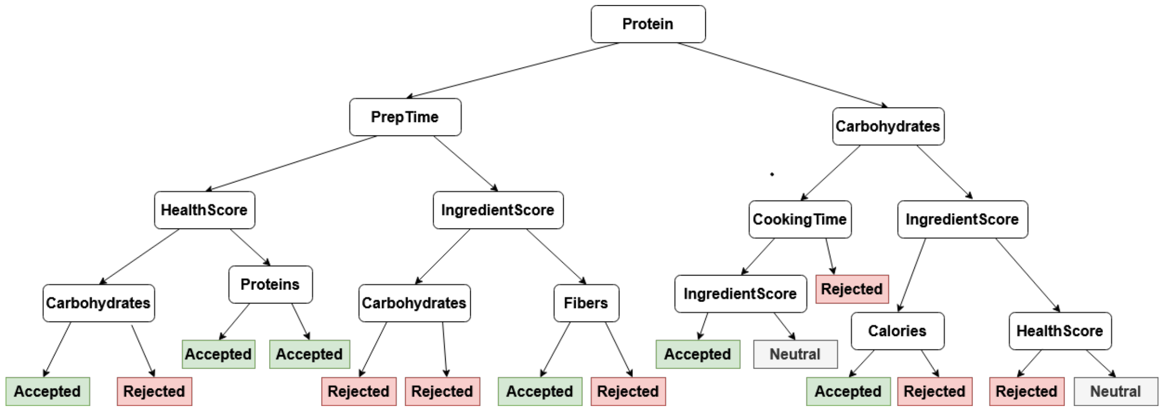


Figure 10: Sample Tree for Item-Based Explanations

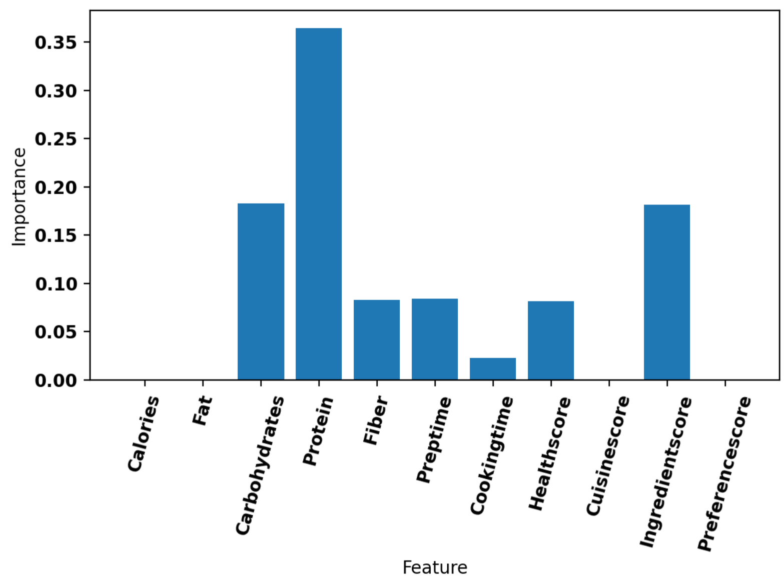


Figure 11: Corresponding Feature Importances for the Figure 10

5.1.2 Contrastive Explanations

Contrastive explanations are explanations that clarify why a particular recommendation has been made by highlighting the crucial distinctions between the recommended item and an alternative, in counterfactual scenarios. Rather than explaining why something happened in isolation, contrastive explanations provide insights by comparing the chosen recommendation to what might have happened if a different choice or circumstance had prevailed, shedding light on the factors that influenced the specific recommendation.

In our use case of Nutrition Virtual Coaches, we generate such contrastive explanations as outlined in Algorithm 3. First, we select an alternative that is similar to the recommended item but it's *recipeScore* is less than the recommended one. To do so, we utilize a pool of filtered (i.e., eliminated from the recommendation pool due to the user constraints/preferences) and/or low-scoring (i.e., not healthy or not tasty for the given user) recipes. We employ the Jaccard Similarity metric [58], where $J(A, B) = |A \cap B|/|A \cup B|$, to determine the recipe similarity based on their ingredients. From this candidate set of recipes, we choose the one whose similarity with the current recommendation is maximum (Line 1).

Algorithm 3 Contrastive Explanation Generation

Require:

R: List of recommended recipes;
R': List of filtered and/or low-recipeScore recipes;
$HScore()$: Scoring function to measure healthiness;
$PScore()$: Scoring function to measure user satisfaction;
r: The recommended recipe;
F: set of recipe features

Ensure:

ϵ^+: Positive Counter Example Explanations of a given recipe
ϵ^-: Negative Counter Example Explanations of a given recipe

1: r' $\leftarrow \underset{maxP_u(r)}{\text{argmax}} \|JaccardSimilarity(r, R')\|$;
2: **for each** f in F **do**
3:     **if** (HScore(r', f) < HScore(r, f)) || (PScore(r') > PScore(r)) **then**
4:         $\epsilon^- \leftarrow$ f ;
5:     **else**
6:         $\epsilon^+ \leftarrow$ f ;
7:     **end if**
8: **end for**
9: return ϵ^+, ϵ^-

Afterwards, we compare features of the chosen counter recipe with those of the recommended recipe one by one. If the feature of the chosen recipe has a lower score for some decision criteria (e.g., healthiness or user satisfaction), we added them into negative feature set, ϵ^-, (Lines 2–4); otherwise, inserted into positive feature set, ϵ^+, (Lines 5–7). Those features will be used to build a contrastive explanation sentence highlighting the positive side of the recommended recipe while sending away the contrastive recipe by emphasizing its negative sides.

5.2 Explanation Grammar Structure & Visual Components

From the features acquired by the methods explained in the previous sections, we generate a sentence using the pre-defined grammar-based structure. The structures are composed of two variants: one for the user / item-based explanations is shown in Figure 12 and the other one for contrastive explanation in Figure 13. The phrase repository of the system consists a set of phrases for each decision factor (e.g., for protein: "...provides sufficient protein..."), and other types of phrases such as subject and noun (e.g., "...this recipe..."). The user/item-based explanations are alluring sentences about each positive feature. They are intended to be brief and pithy, whereas contrastive explanations aim to create a comparative explanation with a worse alternative (which can be longer).

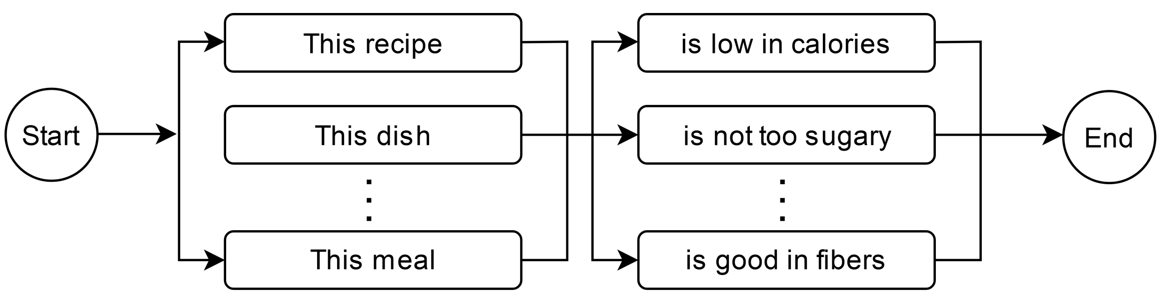


Figure 12: Grammar Structure of the Item / User Based Explanations

Figure 14 shows the novel interface developed to display these explanations. We added visual aspects of explainable recommendations given the success of "graphics" in explaining recommendations [28]. The health-oriented explanations are shown in a green box. Contrastive explanations are outlined in yellow. Additionally, we present nutritional factors related to food to the user. Since our goal is to measure the performance of explanations, the food's picture and details of the recipe are

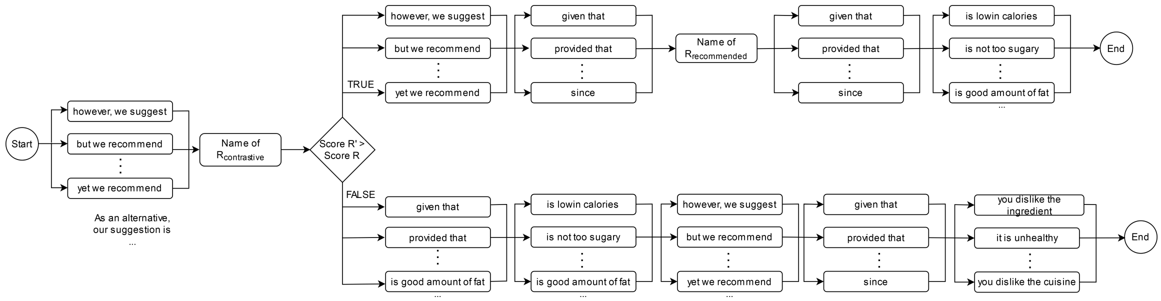


Figure 13: Grammar Structure of the Contrastive Explanations

not immediately visible given that the users are likely to ignore the explanations and decide barely on recipe information. However, the user can access the hidden information if they click respective buttons while coming to a decision.

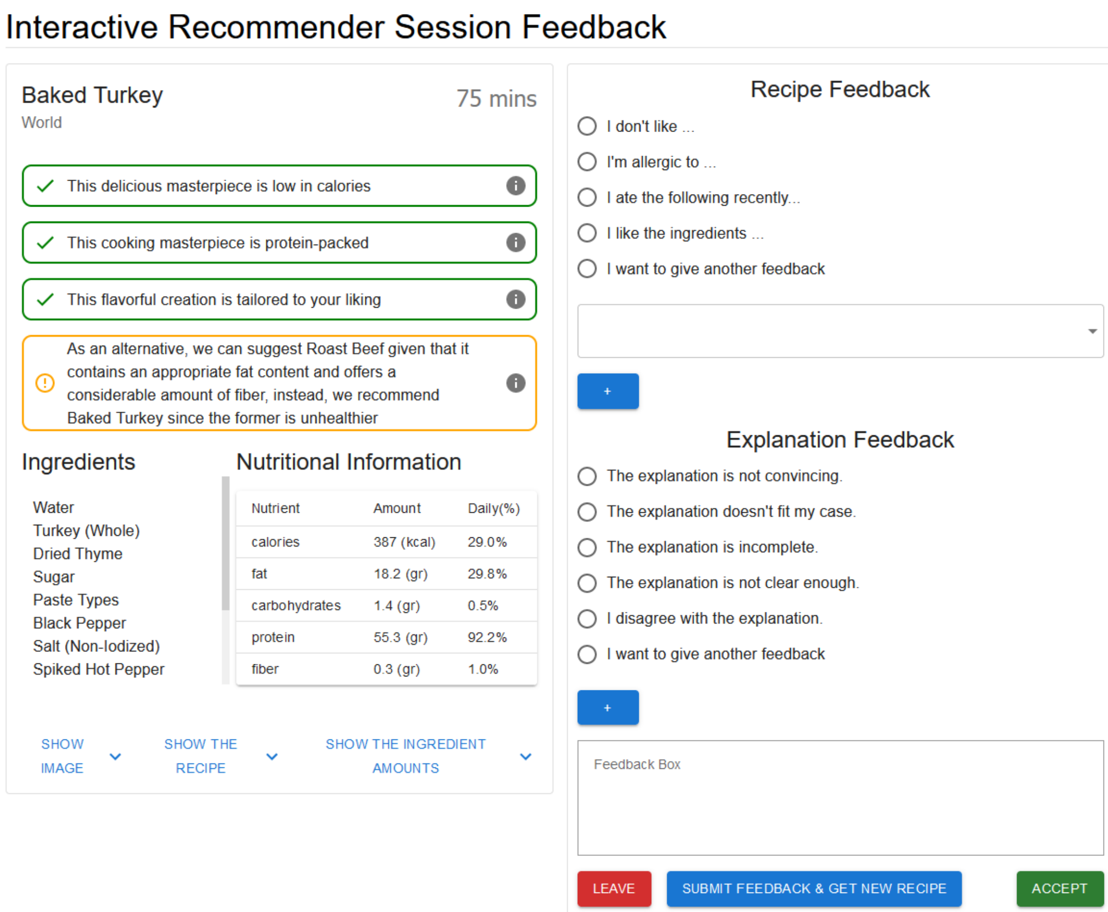


Figure 14: A Sample of the Food Recommendation Interface with Explanations

5.3 Evaluation

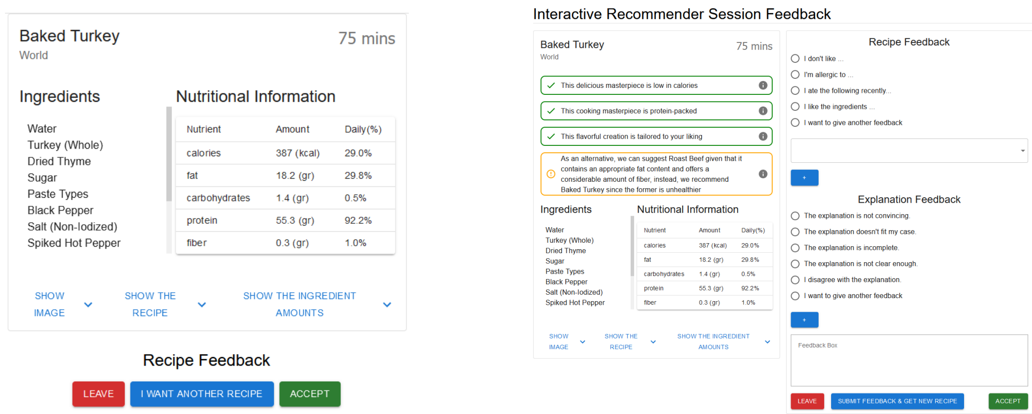
To evaluate the performance of the proposed negotiation framework equipped with enhanced explanations, we conducted a new online experiment via a Web-based interface for food recommendations. The experimental setup and participants are presented in Section 5.3.1 and Section 5.3.2, respectively. Consecutively, Section 5.3.3 reports and discusses the experimental results elaborately.

5.3.1 Experimental Setup

To assess the acceptability and effectiveness of the proposed negotiation-based recommendation framework, we asked participants to experience two variants of food recommender systems, namely *regular recommender*, and *interactive recommender*, where the explanation-based negotiation approach is adjusted to an interactive setting, providing explanations and allowing feedback (i.e., approvals and critiques). It is worth noticing that we revised the Web participant interfaces in both conditions based on the feedback received in the earlier study presented in [11]. We improved how the food recipes and their supportive explanations are displayed to communicate the explanations more effectively and diminish the effect of factors irrelevant to the quality of explanations, such as pictures. Nutritional information and main ingredients are shown directly alongside several types of explanations. Conversely, as visible in Figure 15, a picture of the food as well as the details of the recipes are not directly displayed, but available only via an additional click.

We follow the following steps in our experiments[1]. Before conducting the experiments, every participant completed a pre-survey and registration form to provide information about their gender, age, height, weight, level of physical activity, dietary preferences, and any allergies they might have. This information concurs to estimate

[1] We would like to highlight that the experiment protocol adopted in this study was approved by the Ethics Committee of Özyeğin University, and informed consent was obtained from all participants.

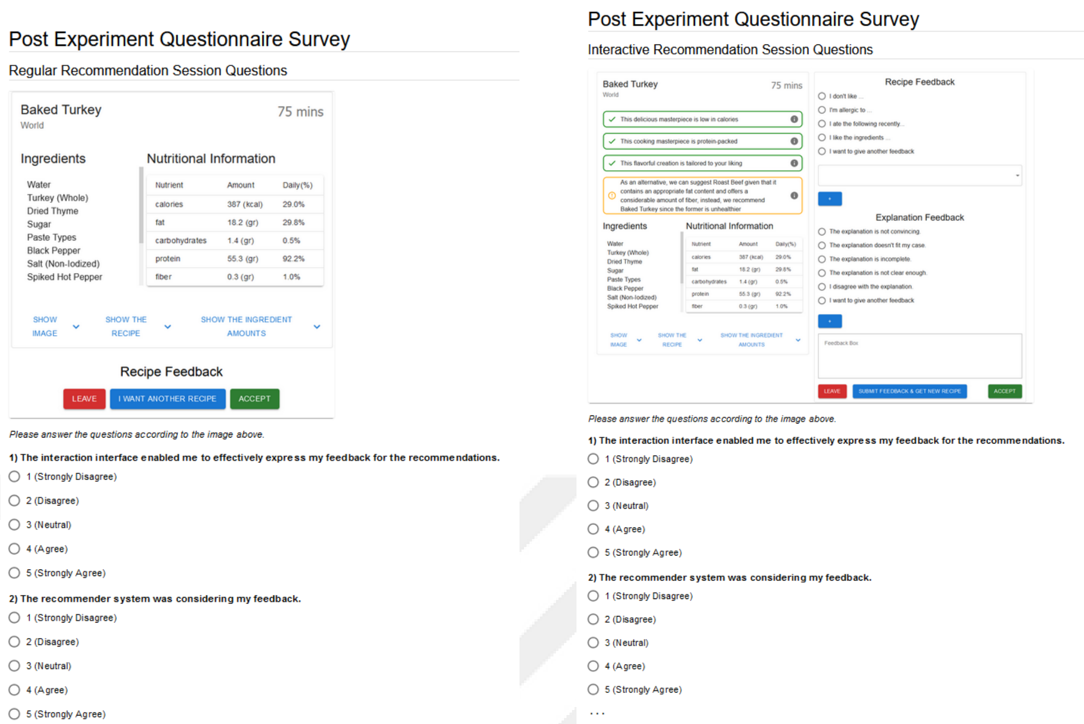


(a) Regular Recommender System **(b)** Interactive Recommender Session

Figure 15: Regular and Interactive Recommendation Sessions

the healthiness score of recipes recommended to the participant (see Section 4.3). To reduce the learning effect among the sessions, the participants were split into two "groups", inverting the starting settings order. Therefore, a three-minute break was given between the two sessions. Initially, we scheduled a longer break. However, in our pilot experiment, we received negative feedback about the too-long waiting interval.

Following the completion of the experiment, the participants are asked to fill in a questionnaire that primarily comprises 5-point Likert scale questions to assess their experiences in both sessions (one questionnaire per session). The questionnaire follows a within-subject design [55] to gather participants' insights regarding the system's success. To facilitate recalling their experiences and differentiate the sessions, a picture (screen capture) of the given system's setting is displayed at the beginning of the questionnaire page (see Figure 16). Finally, additional 5-point Likert scale questions were asked to the participants about their perceptions of the received explanations during the interactive system.



(a) Regular Recommender Session (b) Interactive Recommender Session

Figure 16: Regular and Interactive Recommendation Sessions Questions

5.3.2 Participants

In total, there 54 participants (19 female, 35 male) with diverse backgrounds and age groups took part in the test. The mean age of the attendees is 26.31 years old (with a minimum of 19 and a max of 58 years old). The participants were requested to order the importance of five criteria, relative to a given food recommendation: “Nutritional factors”, “Past experience with taste”, “How it looks”, “Price of the ingredients”, and “Cooking style”. Figure 17 shows the histogram analysis of the questionnaire. The participants ranked these factors on a scale of 1 to 5, with 1 being the most important factor. One could observe that the majority of the participants (i.e., 69 % of the participants) ranked past experience with the taste of such food to be the most crucial factor in deciding their food recipes to cook, whereas 21% of the participants marked nutritional factors to be the most important. Conversely, 39% of the participants marked cooking style as the least important factor, whereas the food's appearance was rated the least important by 26% of the participants.

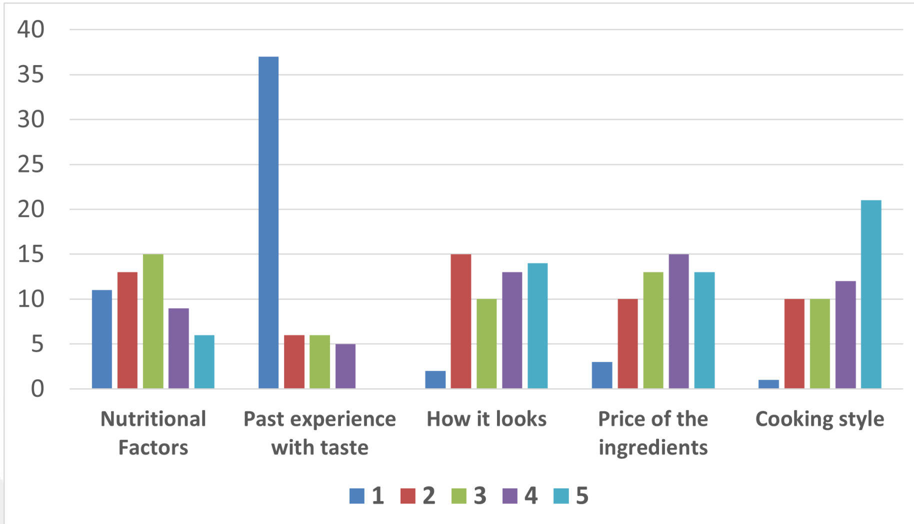


Figure 17: Histogram Analysis of the Pre-Survey Questionnaire

5.3.3 Experimental Results

The success of the self-explanatory systems is usually measured under two categories of metrics; subjective and objective metrics [59, 60, 18]. Objective metrics are metrics derived from the participant actions within the experimental setup, such as success rate (i.e., percentage of sessions ending with an agreement), number of rounds per session, healthiness level of the accepted food recipe. Subjective metrics denote the participant scores for the post-experiment questionnaire (see Figure 21 below). The subjective evaluation questions are about perceived effectiveness, level of detail, user satisfaction, understandability, informativeness, and convenience, meaning that the explanations are appropriate relative to the stated user preferences and constraints. In addition, we asked about the general idea of receiving explanations in addition to recommendations.

We first analyzed the number of sessions that ended successfully. Out of 54 sessions, only two regular and one interactive session ended without any agreements. It is worth noting that the participant who failed to find agreement with the interactive system also couldn't find one with the regular system.

Moreover, participants reached an agreement in the third round on average when they engaged in the interactive session (i.e., average=3.1 standard deviation= 3.5). In contrast, they accepted the given offer in the forth round on average for the regular session (i.e., average=3.6 standard deviation= 3.9). Total number of rounds per each participant in each session can be seen in Figure 18 where the red and blue bars denote the total number of rounds for the regular and interactive sessions, respectively. Compared to regular recommender session, 18 participants had more interaction in the interactive session, whereas 14 participants finished the interaction in the same round for both sessions. In contrast, 22 participants required more rounds to find an agreement in the regular sessions. That shows that interactive recommender systems do not necessarily take more rounds to reach an agreement.

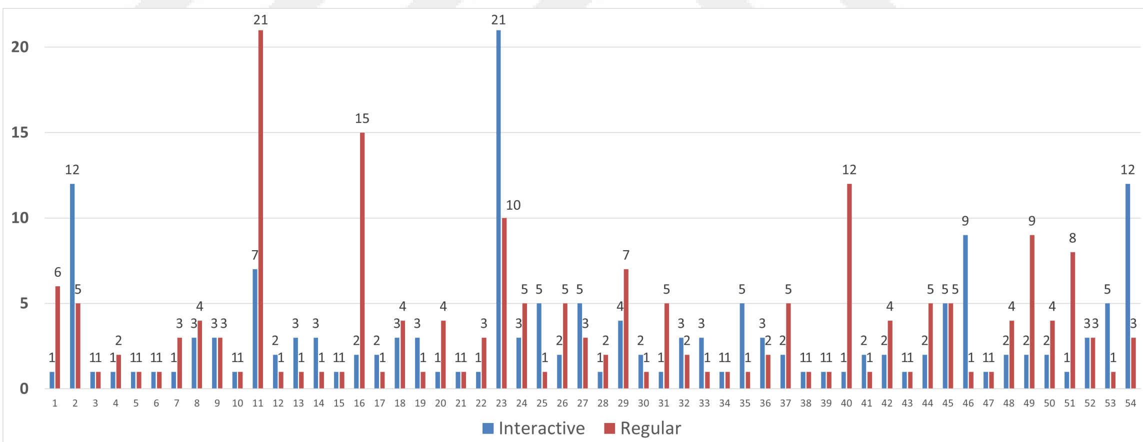
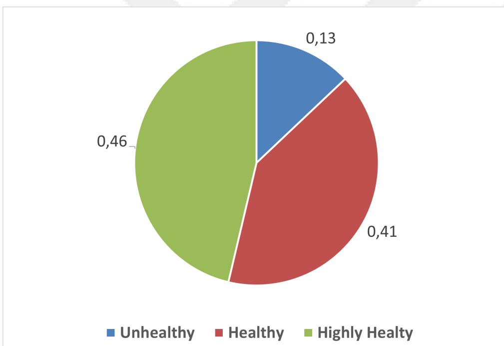


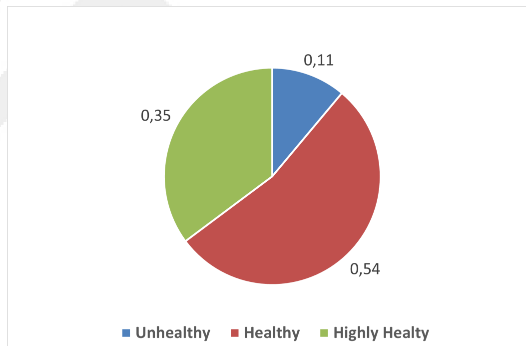
Figure 18: Total Number of Rounds Per Participant, for Each Interaction Type

For the interactive session, 19 participants accepted recipes of – what we classified as – highly healthy foods; 29 participants preferred healthy foods, and six accepted unhealthy food recipes. For the regular session, on the other hand, the participants accepted 25 highly healthy options and 22 healthy options; in contrast, seven participants went for unhealthy options. These results are illustrated in Figure 19. That shows that the interactive and regular sessions are similarly effective in meeting the objective of recommending healthy foods. Recall that the recommendation strategy itself is the same in both sessions.

The aforementioned results concerning the total number of rounds per session indicate that 18 participants ended the session in the regular session earlier than the interactive one. It is possible that they enjoyed exploring the system more in the interactive system. Since the recommendation strategy employs a time-based concession strategy, the longer it endures, it may offer less healthy food relative to its previous offers. As a result, regular sessions may end up with healthier food recipes compared to the interactive system in some cases. On the other hand, there are less unhealthy food recipes agreed by the participants in the interactive session.



(a) Regular Session



(b) Interactive Session

Figure 19: Percentage of Healthiness Level of the Agreement

Out of 54 participants, the system received the following evaluative feedback for interactive sessions:

- “The explanation doesn’t fit my case”, from 4 participants,
- “The explanation is not convincing”, from 4 participants,
- “The explanation is not clear enough”, from 2 participants,
- “The explanation is incomplete”, from 1 participants,
- “I disagree with the explanation”, from 1 participants.

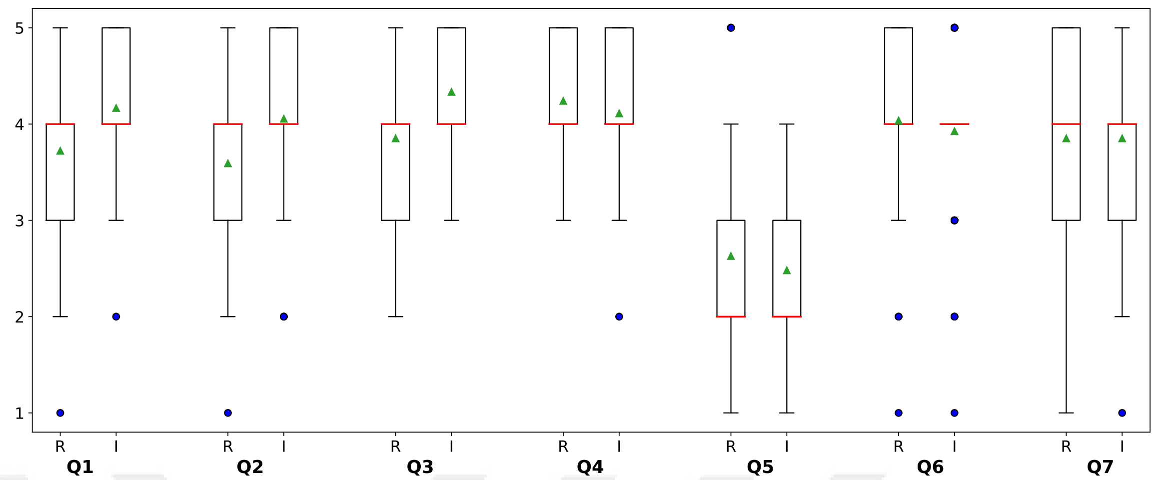
Additionally to our given feedback options, which were all negative, participants utilized the custom feedback option to compliment the explanations: "The explanation is acceptable" or "The explanations are enough for me".

Furthermore, we analyzed the users' responses to the post-experiment survey to examine how they perceived the regular and interactive recommendation system. Since each participant experienced both sessions and the questions are the same for both, we performed a within-analysis statistical comparison test. The data is not normally distributed which is one of the main assumptions made by the pairwise T-test. Thus, we apply the corresponding non-parametric test called the Wilcoxon sign rank test [55]. For all tests, the Confidence Interval (CI) is set to 0.95, $\alpha = 1 - CI = 0.05$.

Figure 20 shows the box plot of the comparative questionnaire between the regular (R) and the interactive (I) session, respectively. The orange lines represent the median, the triangles in green the means, and the small blue circles the outliers.

The analysis in the box plot shows that there is a significant improvement for the interactive sessions, especially for questions $Q1(p = .002)$ and $Q2(p = .001)$. These two questions measure the system's sociability. That is, the results show that the interactive session is statistically significantly better than the regular session in terms of sociability. This improvement is reasonable given the additional dialogue options, such as feedback mechanisms, in the interactive session. Q3 measures the amount of information the participants perceived to be fruitful. The added explanations were recognized by the participants to be effective, hence, here too a significant improvement has been reported ($p = .0009$). In other words, the participants perceived that the interactive session provided better information than the regular session to make an informed decision.

Moreover, questions Q4 ($p = .874$), Q5 ($p = .838$), and Q6 ($p = .910$) qualify the usability of the system. These values show that there is no significant difference. That

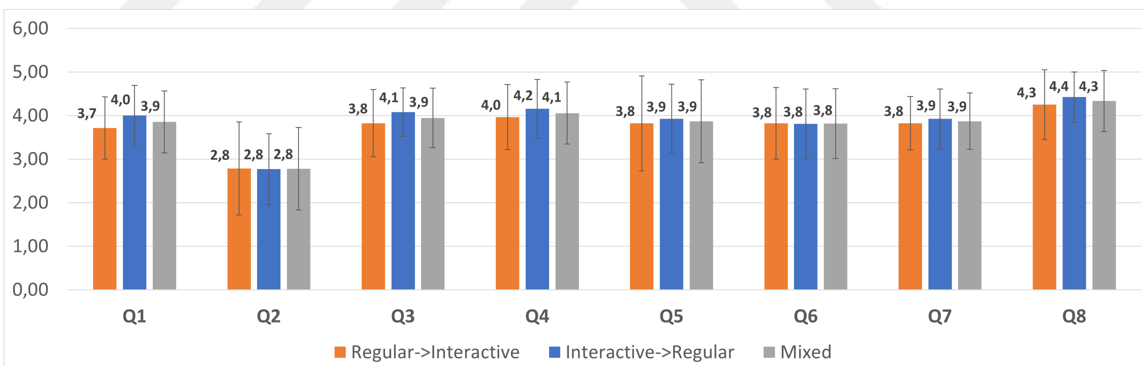


Id	Questions label	Metric	W	p
Q1	The interaction interface enabled me to effectively express my feedback for the recommendations.	Sociability	102.0	**.002**
Q2	The recommender system was receptive to my feedback.	Sociability	33.0	**.001**
Q3	The recommender system gave me enough information to decide.	Completeness	58.5	**.0009**
Q4	The recommender system was easy to use.	Usability	121.5	.874
Q5	It was time consuming to find the best recipe in the system.	Usability	155.5	.838
Q6	The interaction with the system was smooth and effortless.	Usability	102.0	.910
Q7	I would like to use such a system in my daily life.	Acceptability	33.0	.5

Figure 20: Box Plot and p-values of Comparative Analysis of Ssubjective Questions Between Regular and Interactive Sessions

means that adding an interactive dimension to the system, can still be effective and efficient. This is in line with what we found earlier about the similar number of turns. Lastly, we measured the acceptability of the two versions of the system. According to the statistical test, there is no significant difference between regular and interactive systems for Q7 ($p = .5$). The average acceptability score for the interactive session is approximately 3.85, where 3 is neutral and 4 denotes "agree". Furthermore, we asked all participants which systems they prefer. Only a minor part of the participants (3 out of 54 participants or 6% of them) prefer the regular one over the interactive system. In other words, the majority favors the interactive system (45 participants). The rest is indifferent.

Apart from the comparative analysis, we also ask questions to assess the perceived quality of the explanations in our system. Hoffman et al. provide a list of so called goodness criteria for explanations [60]. Inspired by those statements, we created corresponding statements for the food recommendation system and asked each participant to what extent they agreed. Figure 21 shows the questions and the respective average scores. To examine whether a learning effect may have influenced the results, we report the average scores with respect to (i) participants who started with the regular sessions (i.e., Regular → Interactive), (ii) participants who started with the interactive session (i.e., Interactive → Regular), and (iii) all participants irrespective of the order of sessions (i.e., Mixed). It is clearly seen that the counter-balancing technique works. The results for both orderings are similar. In general, participants are satisfied with the given explanations and appreciated the idea of receiving explanations in addition to the given recommendations. They do not agree that the explanations were too detailed. In addition, they found the explanations help them choose the most convenient recipe.



Id	Questions label
Q1	The explanations for recommendations has helped me choose the most convenient recipe.
Q2	The explanations for recommendations were too detailed.
Q3	The explanations displayed during the interaction were satisfactory.
Q4	The explanations for recommendations were clear and easy to understand.
Q5	The explanations were sufficient to make an informed decision for healthiness.
Q6	The explanations were realistic in terms of healthiness of given recipes.
Q7	The explanation let me know how convenient the recipe is.
Q8	Rate your appreciation of the idea of receiving explanation in addition to recommendations.

Figure 21: Evaluation Questionnaire Results, Shown Per Order of the Sessions

Lastly, we categorized participants based on their responses to the pre-survey question – the importance of the factors on their decision making (See Figure 17. Since there are a few participants who found the most important factor as how the food looks, price of the ingredients, and cooking style, we only categorized the participants who voted the most important factor in choosing a recipe to be the past experience with taste and the nutritional factors of a given food. This categorization is also in line with our objectives. Figure 22 shows the score of the aforementioned explanation related questions and responses of the participants in each category. Note that since order of session does not influence the results, we only show the average scores for all participants who fit in the given category. We could not find any significant differences in their responses.

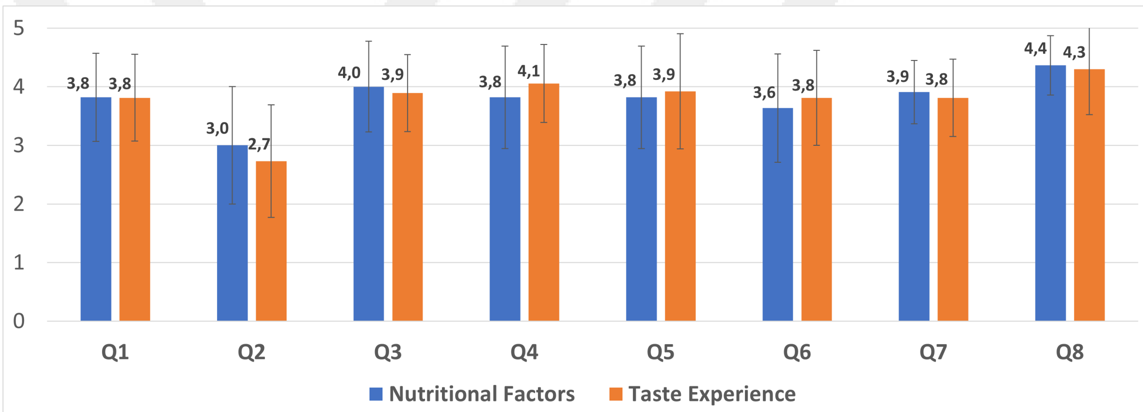


Figure 22: Questionnaire Results Per Pre-Survey Answers

CHAPTER VI

RECOMMENDATIONS WITH ANTE-HOC AND POST-HOC EXPLANATIONS

In the preceding chapters, our approach to explainable AI adhered to the general principle of generating the explanations post-hoc, meaning that the systems generated the explanations after the recommendations were created. While this method served its purpose, it lacked a direct connection between the recommendation process and the explanation generator. On the other hand, we recognize the intuitive appeal of generating explanations alongside recommendations, which can be achieved by utilizing inherently explainable machine learning techniques to generate these recommendations. In this chapter, we present two new methods that incorporate machine learning to generate explanations: post-hoc clusters-based explanations with clustering methods (Section 6.1) and ante-hoc popularity-based explanations via Random Forests classification (Section 6.2).

6.1 Cluster-based Post-Hoc Explanations

This section presents a novel explanation generation idea based on clustering. Clustering is a crucial technique in recommendation systems as it enables grouping similar items based on shared attributes [61]. This organization enhances recommendation quality by suggesting items within the same cluster when a user engages with a particular item, improving relevance and user satisfaction. While generating post-hoc explanations, it might be intuitive to consider important descriptive features of the items similar to the current recommendation. Consider that the system recommends a recipe. Why does the system recommend that recipe but not another one? There should be some distinguishable features that the user prefers. How can we detect those features? We can cluster the items with respect to the user's estimated preferences and desires. As usual, it is expected to have similar behavior or pattern in the same cluster, and someone can inquire which features distinguish those items in the same cluster with the current recommendation. Ultimately, we can utilize those features in our explanation generation approach. We employ random forest classifiers to distinguish the cluster features in our approach.

To achieve this, we propose the steps shown in Figure 23. It is worth noting that we assume the system has a user model estimating the users' preferences and needs/interests. We first determine how an item can be represented by means of a feature vector capturing those preferences. For instance, a food recipe could be defined as a collection of features such as user-relative taste preference score or recipe healthiness score. Based on those feature sets, our approach suggests applying a clustering algorithm to determine distinguishable recipe groups concerning users' preferences and needs. We suggest using a classifier such as Random Forest for each cluster to detect the important features differentiating the items from other groups (i.e., clusters). We train a classifier to determine whether the given item belongs to the first cluster, another classifier for the second, and so forth. Random Forest could yield the features importance values which could be utilized to select the features to be used in the explanation. The most important feature in the proposed explanation generation approach is chosen to generate a post-hoc explanation. We call this approach Cluster-based Explanation.

Finally, we could also utilize the cluster mechanism to generate a contrastive example, and thus, contrastive explanations too. Algorithm 4 elaborates on this process. First, we choose the most similar item with the recommended item from another cluster (Line 1). Then, we compare the values of each feature of contrastive item with the recommended item's. When the score of the chosen recommendation is

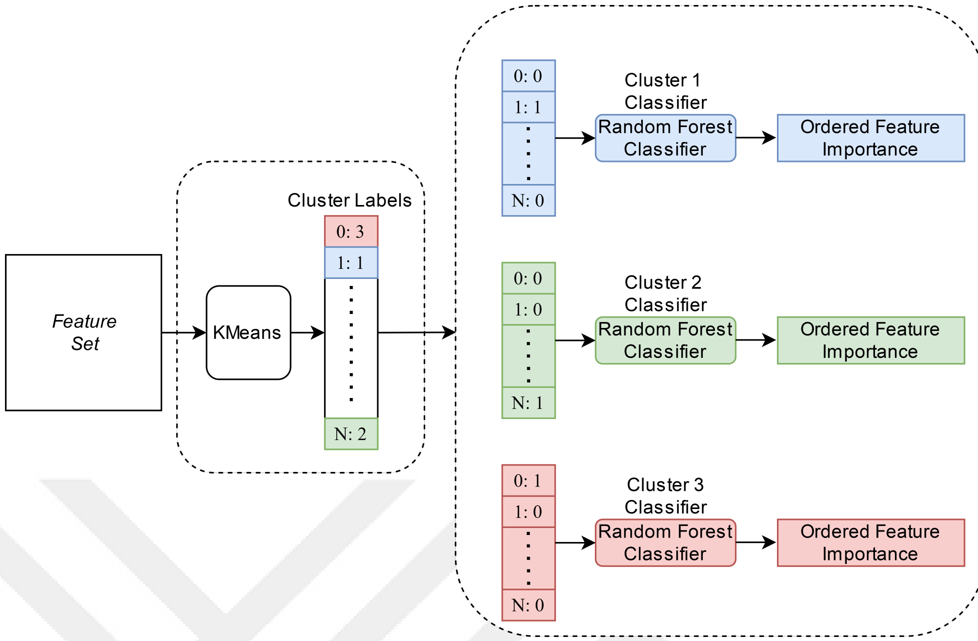


Figure 23: The K-Means Feature Importance Analysis Process

higher for a given feature, we consider it to be a positive contrastive feature, whereas the vice-versa is applied for the negatively contrastive features (Line 3 to 9). While generating an explanation, we promote the recommended item with the positively contrastive features, whereas negative features indicating why the contrastive example is not suggested by the system.

Algorithm 4 Cluster-Based Contrastive Explanation Approach

Require:

  ϵ^+, ϵ^-: Positively and negatively contrastive features;
  F: Feature set;
  R'_c: Set of scored recipes in the same cluster as r;
  r, r': Recommended and contrastive items;

1: $r' \leftarrow \text{argmin}_{c \in R'_c}\, distance(c, r)$
2: **for each** feature f in F **do**
3:     **if** score(r[f]) $>$ score(r'[f]) **then**
4:         $\epsilon^+ \leftarrow \epsilon^+ \cup f$
5:     **else**
6:         $\epsilon^- \leftarrow \epsilon^+ \cup f$
7:     **end if**
8: **end for**
9: return ϵ^+, ϵ^-, r, r'

Recall that we consider an Nutrition Virtual Coach as a use case and in our experimental evaluations. For this case study, we represent each food recipe as a vector of *preference score*, *health score*, *price score*, *time score*, and *taste score* in terms of user preferences. In the following part, we explain how these scores are calculated to suit the clustering technique used in our user experiments as features.

Preference Score: To calculate the preference score of a user for the recipe dataset, we utilize a novel Active Learning framework [62] that is proven effective within our research project. Briefly, the system first generates a diverse sample of recipes from the dataset. It asks the user to specify whether they like or dislike a given recipe. Afterwards, the system shows a set of recipes to the user. The participants are asked to indicate the correct labels for the predictions made by the system via adjusting whether they like it's respective features or not. These user feedback are utilized to generate synthetic data to enrich the user's preference data and increase the system's accuracy with a small dataset. Ultimately, the labeling generated from this process is used to train a supervised learning model, Logistic Regression. We use the positive class probability by the model as the indicator for the user preference score for a given recipe, which corresponds to the likelihood of user's acceptance of a recipe.

Health Score: In order to calculate the health score, we follow the previous intuition presented where we take the personal information of the users to calculate their nutritional needs in a day (see Section 4.3.2). The final *healthScore* is the mean of *calorieScore* and respective nutrient scores for each nutrition value of the recipe. To calculate the *calorieScore* we use the daily active metabolic rate (AMR) described in the Section 4.3.1 and the final score is derived as described in Equation 6 where R corresponds to a recipe within the dataset. Essentially, we constrain the calorie score within the range of [0, 1], and we assume that a higher amount of calories is better as long as it is less than the active metabolic rate. For the nutrient scores, we use the nutrient density score [63] as described in equation 7. To simplify the healthiness decision, each nutrition is scored higher if they have a higher density per calories except for the amount of fat, where it is reversed ($1 - fatScore$). Finally, all of the scores are clamped to the region of $[0, 1]$ then averaged to derive the final $healthScore$.

$$calorieScore_R = \begin{Bmatrix} \frac{R_{calories}}{max_{calories}}, & \text{if } calories_R \leq AMR \\ 0, & \text{else} \end{Bmatrix} \tag{6}$$

$$nutrientScore_R = \frac{amount_{nutrient}(gr)}{calories_R(kcal)} \tag{7}$$

$$healthScore_R = \sum_{nutrient} w_{nutrient} * score_{nutrient} \tag{8}$$

Price Score: We first label the recipes within three classes; cheap, standard and expensive (labelled as $, $$, $$$ in order). We assume that the cheaper is better for a given food recipe, therefore, we assign these classes the scores of 1, 0.67, 0.33 respectively.

Time Score: The time scores are a summation of the recipes preparation and cooking time. For the calculation of this score, we assume that the quicker is better. Therefore, we apply max-normalization on the time of preperation in terms of minutes and reverse the order of scores for all the recipes. Thus, the quickest recipe is scored as "1".

Taste Score: The taste score corresponds to how well the flavor preferences of the user matches the flavor profile of a recipe. The flavor profile is comprised of the following tastes: *Savory*, *Bitter*, *Sour*, *Salty*, *Sweet* and *Spicy*, which are recognized as the main tastes the humans can distinguish [64]. Each food recipe holds boolean fields for the dimensions of a flavor profile. Table 3 shows an example recipe where each taste is labelled "1" if it is a part of the profile, or "0" if it is not. Finally, we ask user to elicitate their desired tastes in the same form (e.g., they mark whether

Table 3: Flavors of a Tomato Soup

Recipe	Savory	Spicy	Sour	Salty	Sweet	Bitter
User Profile	1	1	1	1	0	0
Tomato Soup	1	0	1	1	0	0

or not they want it in a boolean fashion). For each recipe, we apply the Jaccard Similarity [58] as described in the Equation 9 to calculate the final score within the range of [0, 1].

$$tasteScore_R = \frac{flavors_R \cap flavors_{user}}{flavors_R \cup flavors_{user}}, \tag{9}$$

In determining the amount of clusters, we utilized elbow analysis. Ultimately, we determined the number of clusters to be 4 from the testing data. An example elbow analysis generated by the live data of 10 users can be seen in Figure 24.

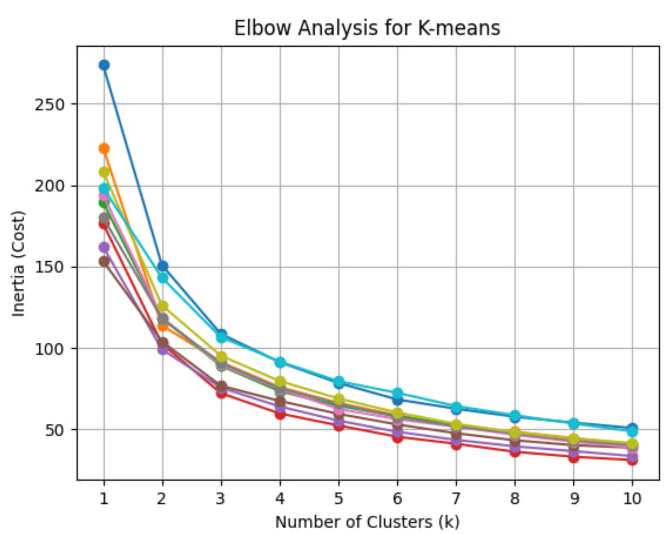


Figure 24: Elbow Analysis for the K-Means Cluster from 10 Participants

Figure 25 shows the clusters for all recipes for a. random experiment participant. Each line on the chart represents an instance from the dataset, and each color represents the membership to a cluster.

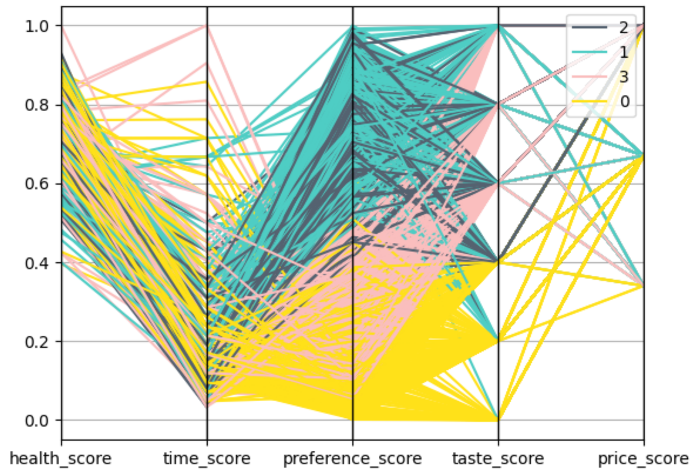


Figure 25: Parallel Coordinates Representation of Multi-Dimensional Clusters

Accordingly, in our experimental setup we follow the Algorithm 5 indicating the steps to calculate a recommendation and a respective explanation for that recommendation. First, we calculate the *finalScore*, which is the summation of all scores multiplied with their corresponding weights derived from the recipe decision method (Lines 1-3). These weights are sum normalized between [0, 1], so the resulting range is between [0, 1] for the total calculation where the score of "1" represents the most valuable recipe for the user. Then, we select the recipe with the highest *finalScore* (Line 4), we acquire the respective cluster from the clustering model (Line 5). Finally, we get the most important feature for that cluster using the respective Random Forest classifier for the selected cluster (Line 6).

6.2 Explainable Popularity-based Recommendation

The explanation generation techniques employed thus far have been post-hoc, which could result in explanations that do not accurately reflect the actual decision-making process. This discrepancy arises because the system's overall outcomes may need to align better with the individual recommendations made by the system. In this

Algorithm 5 Enhanced Baseline Recommendation Strategy

Require:
ϵ: Selected feature
F: Feature set of K-Means;
P_u: Users feature preference weights;
R: Set of scored items;
r: Recommended item;
C: KMeans clusters;
Ensure: ϵ, r
for each r of R **do**
  $R_{finalScore} = \sum_{f}^{F_R} Score_f * P_u(f)$
end for
$r \leftarrow \text{argmax}_{R_{finalScore}}$ (R)
$c \leftarrow C(r)$
$\epsilon \leftarrow$feature with highest feature importance for c
return ϵ, r

approach, we utilize historical data-set indicating past recipe acceptance and a machine learning approach to classify recipe acceptance. In particular, a Random Forest Classifier is utilized to generate recommendations and explanations. Therefore, the proposed structure is inherently connected to the recommendation process, as both are generated from the same model. The Random Forest classifier is a valuable choice for popularity-based recommendation systems because it handles large-scale datasets and provides robust predictions while being explainable in nature. While making a recommendation, we calculate the probability estimates of each recipe and sort them based on the probability of acceptance (labelled as "1"). The recommendation algorithm combines this knowledge with personalization derived from previous sections and recommends the recipe with the highest level of probability of acceptance. On the other hand, we utilize the feature importance generated by the same Random Forest model in generating an explanation.

Algorithm 6 details the recommendation selection and explanation generation process. First, we train a Random Forest Classifier using our popularity labeled data (Line 1). Then, we predict the class selection probability for each class (Line 2). Our classes can either be acceptable (1) or not acceptable (0). We select the item with the maximum probability of acceptance (Line 2). This item's features are then compared according to the Random Forest model's feature importance vector, then the most important feature is selected to be the explanation (Line 3-4).

Algorithm 6 Popularity-based Recommendation Strategy

Require:
ϵ: Selected feature
F_r: Ordered Features of with respect to an item;
R: All recipes,
R_p: Subset of popular recipes that were recommended in previous experiments; labeled "1" if they were ever accepted or "0" if they were recommended but not accepted;
Ensure: ϵ, r
1: $randomForest \leftarrow RandomForestClassifier(R_p)$
2: $R' \leftarrow randomForest.predictproba(R)$
3: $r \leftarrow \text{argmax}\, R'[1]$
4: $\epsilon \leftarrow max(randomForest.featureImportance(F_r))$
5: return ϵ, r

Additionally, we utilize the Popularity-based Recommendation to generate contrastive explanations befitting the context. Algorithm 7 explains the process of contrastive explanation generation for popularity recommendation process. First, we extract the subset of recipes that are labelled as not recommended by the algorithm (Line 1). Then, we select the recipe with the minimum distance in the feature space from the recommendation (Line 2). Finally, we select the most important feature according to the feature importance of the Random Forest Classifier (Line 2).

In our former user studies we have collected the results of our recommended recipes and labelled them "1" if they are ever accepted by any user and "0" if they are recommended to an user but they were never accepted, and ignored the recipes that were never recommended. We applied one-hot encoding to the features (i.e., *flavors of food*, *meal type*, *price* and *cooking style*) to classify them according to our labelling strategy. We utilized this data to train a Random forest classifier to predict whether

Algorithm 7 Popularity-based Contrastive Explanation Selection

Require:

F: Feature set of Popularity;

P_u: Users feature preference weights;

R: Set of scored recipes;

r: Recommended recipe;

Ensure: ϵ: Explanation feature; r': Contrastive recipe;

$C \leftarrow R[0]$

$r' \leftarrow \text{argmin}_{c \in C}\, distance(c, r)$

$\epsilon \leftarrow \text{argmax}_{f \in F}$ featureImportance(r)

return ϵ, r'

a recommendation will be accepted or rejected by the user based on its popularity label in past experiments.

6.3 Explanation Generation Grammar Structure

For the cluster-based recommendation process, we utilize the grammar structure described in Section 5.2 for the Item-User based trees. However, we developed a new grammar structure for the popularity based recommendation explanations given their contextual difference from the other methods. Figure 26 and Figure 27 show the standard grammar and the contrastive grammar structures respectively. These grammars construct a human-enjoyable sentence from the given tags of explanations, mainly generated by the system's explanation generation methods.

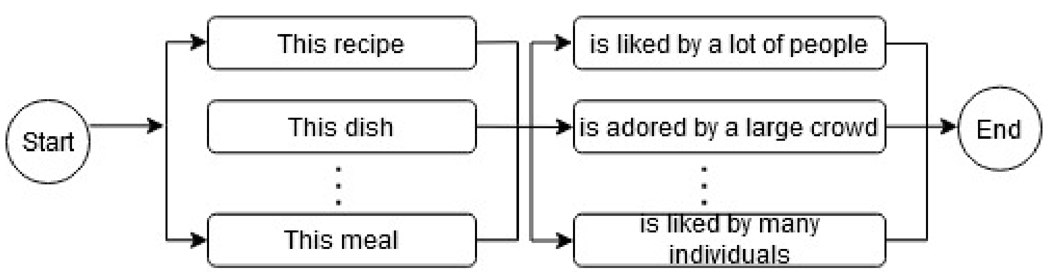


Figure 26: Popularity-Based Explanation Generation Grammar

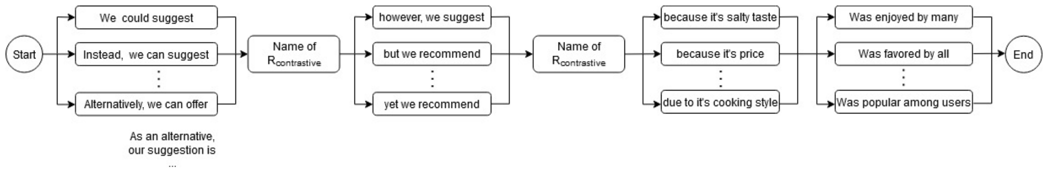


Figure 27: Popularity-Based Contrastive Explanation Generation Grammar

6.4 Evaluation

In order to thoroughly evaluate our latest iteration of explanations, we conducted experiments via a Web-based interface for food recommendations similar to previous studies. However, our focus in this chapter is on explanation generation rather than the interactions. Therefore, this experiment did not incorporate the negotiation protocol developed in Chapter 4. The experimental setup is presented in Section 6.4.1, Section 6.4.2 explains the experiment participants and provides statistical insights. Consecutively, Section 6.4.3 reports and discusses the experimental results elaborately.

6.4.1 Experimental Setup

Prior to commencing the experiments, each participant was required to fill out a pre-survey and registration form, wherein they provided details about their gender, age, height, weight, level of physical activity, dietary preferences, and any allergies they might have. Additionally, they were asked to rank food related factors, and their taste preferences. The system utilizes this information to score the recipes respective to each participant's healthiness and preferences (refer to Section 6.1). To evaluate the acceptability and effectiveness of the explanation-generation techniques proposed, we conducted a study involving participants experiencing three iterations of food recipes, each accompanied by three explanations. The system presents a recipe each time in the following order of recommendation strategies: (i) Popularity-Based Recommendation, (ii) Baseline Recommendation (Chapter 5), (iii) Enhanced

Baseline Recommendation and respective explanations compatible with the recommendation strategy. Additionally, the participants were asked to provide feedback on the perceived performance (effectiveness and convincability) of these explanations using a 5-point Likert scale in the form of star ratings (see Figure 28).

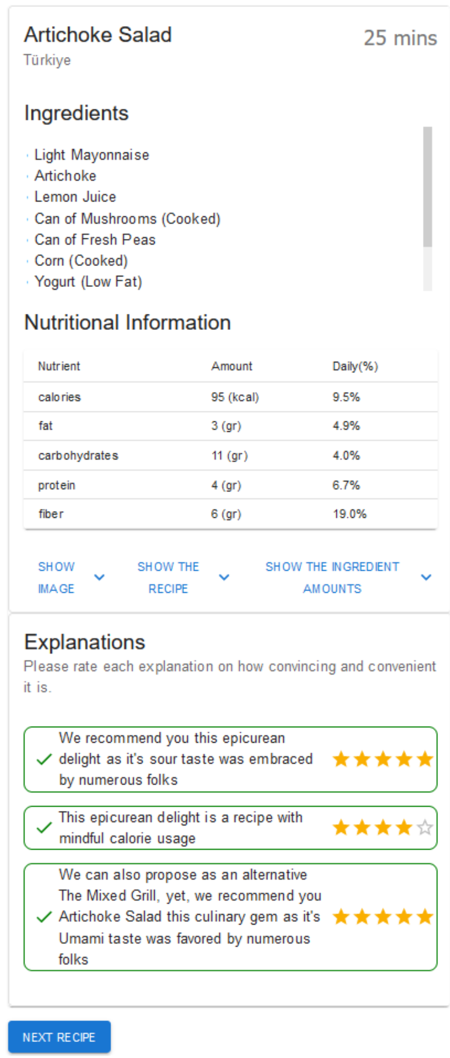


Figure 28: User Interface for Explanation Rating Step of the Experiment

After completing the explanation rating step, the system asks the user to choose their favorite recipe among the shown recipes with respective explanations. This design choice was based on research suggesting that users make better-informed decisions without experiencing excessive cognitive load when selecting from three items simultaneously [65]. Figure 29 illustrates the novel interface developed for this step.

Notably, the user can easily view nutritional information, recipe ingredients, and various explanations. However, the food's picture and detailed recipe information are not immediately visible, but the user can still access them by clicking the respective buttons, similar to earlier studies. After concluding the experiment, the participants are requested to complete a questionnaire consisting mainly of 5-point Likert scale questions. The questionnaire aims to assess their experiences with the explanations provided by the system. Participants are shown an explanation generated by the system and asked to respond to seven questions designed to gauge the effectiveness and success of the explanations they received.

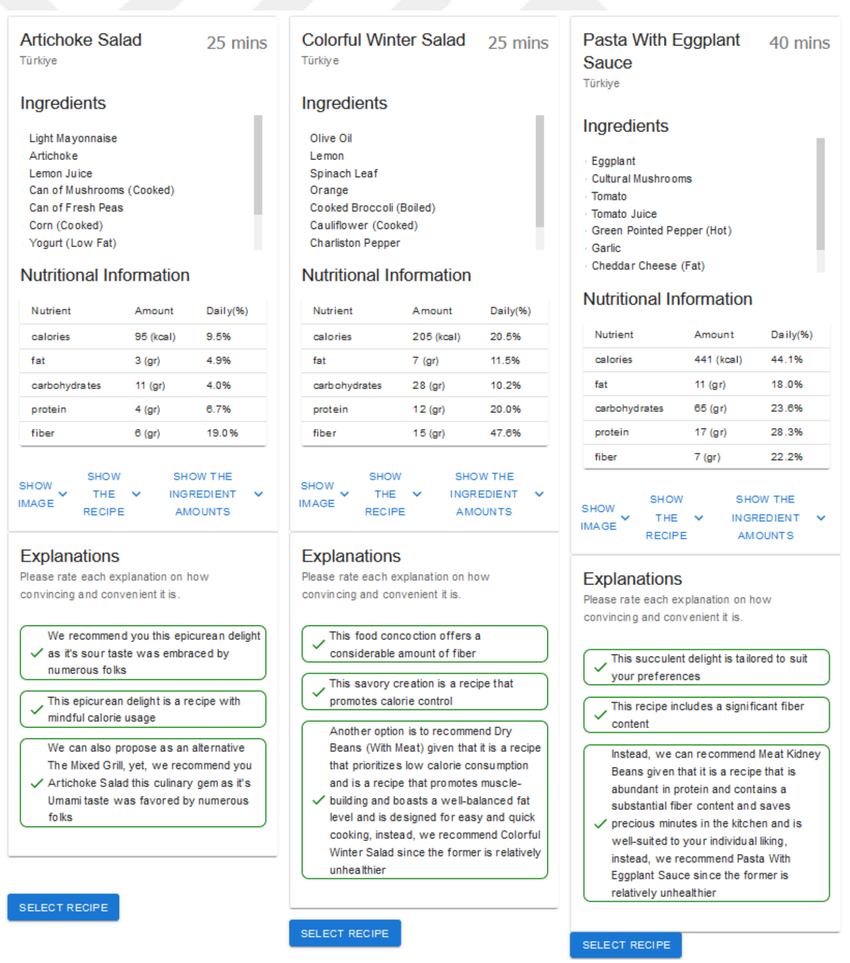
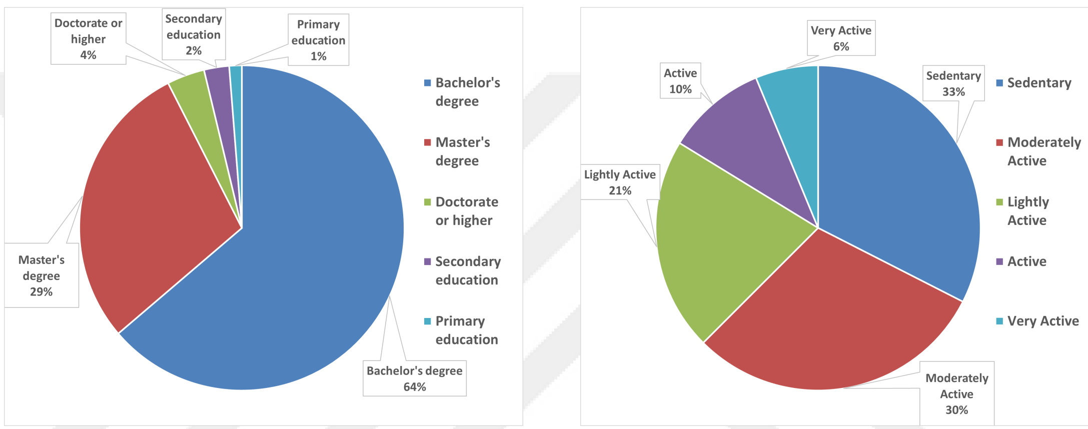


Figure 29: User Interface for Recipe Selection Step of the Experiment

6.4.2 Participants

In total, there are 80 participants (25 female, 55 male) with diverse backgrounds and age groups took part in the test. Figure 30 shows the distribution of the participants in terms of their educational background and sports level. Meanwhile, the mean age of the attendees is 24.70 years old (with a minimum of 18 and a max of 59 years old).



(a) Educational Level **(b)** Sports Level

Figure 30: Distributions of the Participants

The participants were requested to order the importance of five criteria, relative to a given food recommendation: “Nutritional factors”, “Past experience with taste”, “How it looks”, “Price of the ingredients”, and “Cooking style”. In Figure 31, you can see the histogram analysis of the questionnaire. Participants were asked to rate various factors on a scale from 1 to 5, with 1 indicating the highest importance. The results show that a significant portion of the participants (specifically, 46%) considered their experience with the taste of such food to be the most critical factor influencing their cooking recipes. Additionally, 35% of the participants prioritized the healthiness of the food as their top concern.

Conversely, 41% of the participants considered the time required to prepare the recipe the least important factor. In contrast, 28% of the participants rated the food price as the least significant factor in their decision-making process.

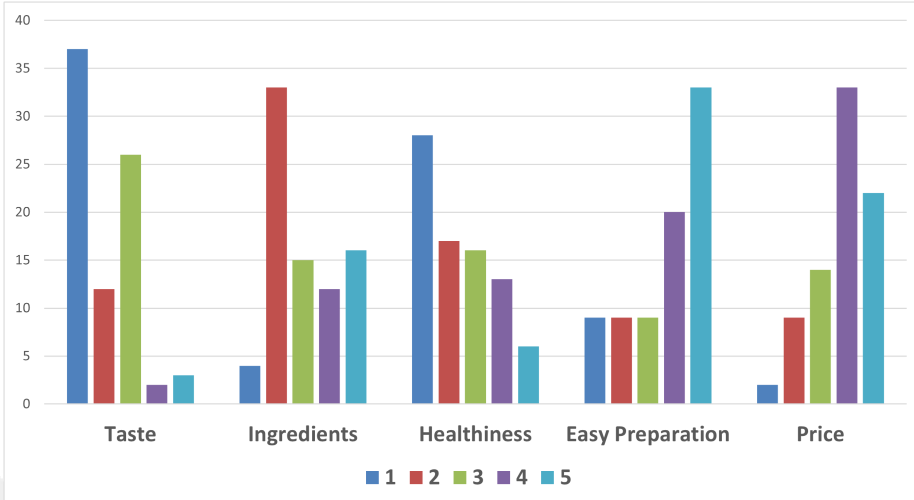


Figure 31: Histogram Analysis of the Pre-Purvey Questionnaire

6.4.3 Experimental Results

The experimental setup is mainly comprised of experiment participants providing subjective input on explanations offered to recommendations. Figure 32 shows the broadest view of the means of answers the participants gave for each type of explanation. We applied the Repeated Measures ANOVA statistical test rejected the null hypotheses, which revealed a significant difference among the types of explanations (F=3.71, p=0.0003).

The data, as determined by the Kolmogorov normality test, does not conform to a normal distribution, a crucial assumption for conducting pairwise T-tests. Consequently, we opted for the appropriate non-parametric alternative, the Wilcoxon signed-rank test [55], for our statistical tests. In all our analyses, we set the Confidence Interval (CI) to 0.95, corresponding to a significance level of $\alpha = 0.05$. Our tests involved organizing explanations into groups based on the recommendation method used for their generation. These groups include:

- Baseline Recommendation: Explanations generated with the Baseline recommendation method in Chapter 5. Here, we use the following explanation methods: Item-based Tree, User-based Tree, Contrastive Tree explanations.

- Enhanced Baseline Recommendation: Explanations that are generated with the Enhanced Baseline recommendation method in Section 6.1. Here, we employ the following explanation methods: Basic Cluster (Section 6.1), Tree-based Cluster (Section 5.1), and Contrastive Cluster explanations.

- Popularity Recommendation: Consisting of explanations related to popularity-based recommendation method in Section 6.2. Here, we apply the following explanation methods: Basic Popularity(Section 6.2), Tree-based Popularity (Section 5.1), and Contrastive Popularity explanations.

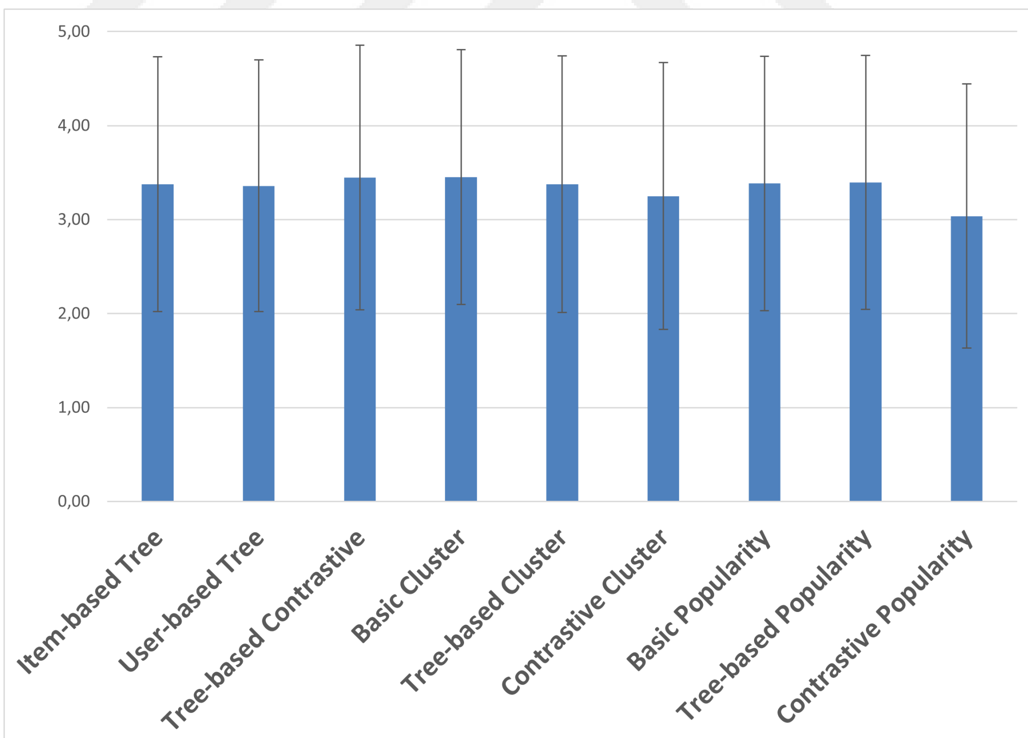
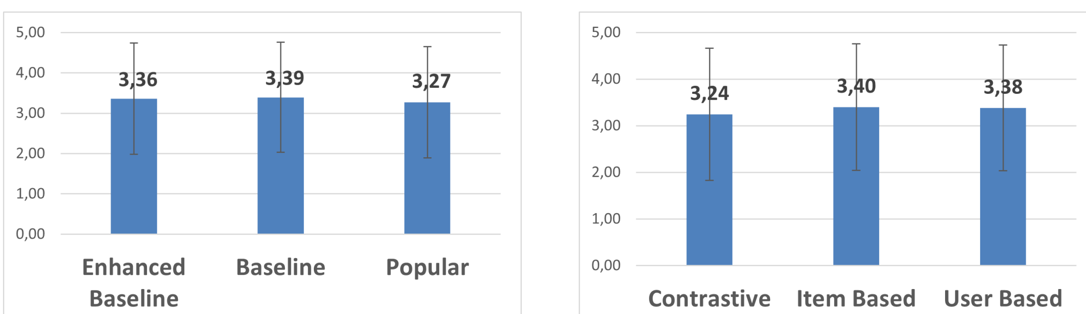


Figure 32: Average Rating of Each Type of Explanation

We conducted pairwise tests between these groups, yielding the following results: Enhanced Baseline vs. Popularity Recommendation ($p = 0.13$), Enhanced Baseline vs. Baseline Recommendation (p = 0.44), and Popularity vs. Baseline Recommendation (**p = 0.04**). One could notice that the explanations generated along the Popularity-based recommendation have underperformed compared to the ones generated with the baseline recommendation strategy, as seen in Figure 33a. However, before drawing such conclusions, we must look into further analysis.

Additionally, we categorized explanation generation techniques based on their underlying differences and the form of labelling strategy they utilized.

- Item Based: Methods that utilize an item's attributes on-line (Basic Cluster, Tree-based Cluster, and Item-based Tree)
- User Based: Techniques that use historical data (user acceptance) as labeling (Basic Popularity, Tree-based Popularity, and User-Based Tree)
- Contrastive: Explanations that are in the contrastive form. (Contrastive Cluster, Contrastive Popularity, and Tree-based Contrastive Techniques)



(a) Recommendation Technique **(b)** Explanation Generation Technique

Figure 33: Mean Scores of Explanations Grouped per Method

We conducted pairwise tests between these groups, leading to the following outcomes: Contrastive vs. Item-based (**p = 0.02**), User-based vs. Item-based (p = 0.84), and Contrastive vs. User-based (**p = 0.02**). Observing Figure 33b, we note that the contrastive explanations under-performed compared to the other methods, slightly more significantly. We notice a trend where the participants prefer explanations drawing power from the attributes of the recommended item more. The participants did not appreciate both contrastive explanations and popularity-based metrics, potentially pointing to the fact that users care more about facts on their recommendation than comparative explanations.

Subsequently, we conducted pairwise tests to compare different types of explanation generation techniques within the same recommendation strategy individually as follows:

Popularity-Based Recommendation:

- Basic Popularity vs. Tree-Based Popularity ($p = 0.70$)
- Basic Popularity vs. Contrastive Popularity ($p \leq \mathbf{0.0001}$)
- Contrastive Popularity vs. Tree-Based Popularity ($p \leq \mathbf{0.0001}$)

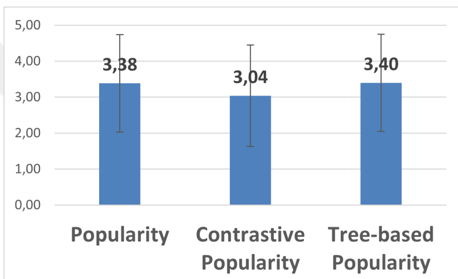
Enhanced Baseline Recommendation:

- Basic Cluster vs. Tree-Based Cluster ($p = 0.78$)
- Basic Cluster vs. Contrastive Cluster ($p = 0.13$)
- Contrastive Cluster vs. Tree-Based Cluster ($p = 0.16$)

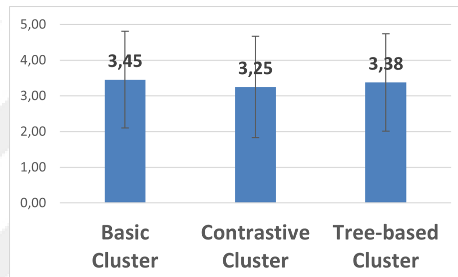
Baseline Recommendation:

- Item-Based Tree vs. User-Based Tree ($p = 0.92$)
- Contrastive Tree vs. Item-Based Tree ($p = 0.44$)
- Contrastive Tree vs. User-Based Tree ($p = 0.31$)

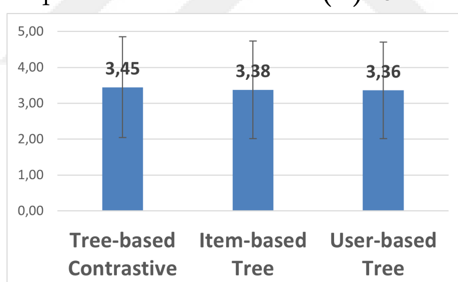
These results provide insights into the comparative performance of explanation techniques within each recommendation strategy. The significant p-values (highlighted) indicate noteworthy differences deserving further investigation. We note that only in popularity group that the contrastive explanations significantly underperform. In other recommendation groups, there is no significant results observed. Additionally, Figure 35 shows the percentage of acceptance among the recommendation engines. Our take-aways may be further supported by this outcome given the simplest method of recommendation seems to be favored more than the others. However, this may be just a result of combination of explanations and the food recipe being more fitting to the users.



(a) Popularity Based Explanations



(b) Cluster Based Explanations



(c) Tree Based Explanations

Figure 34: Mean Scores of Explanations per Recommendation Engine Type

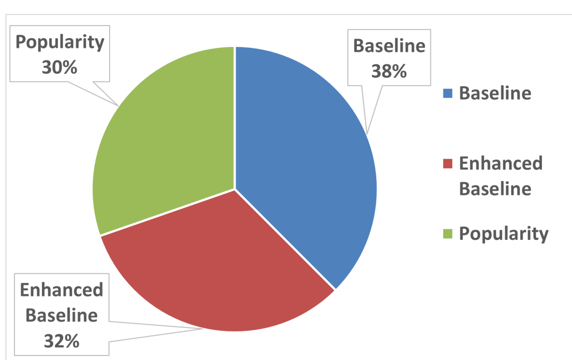
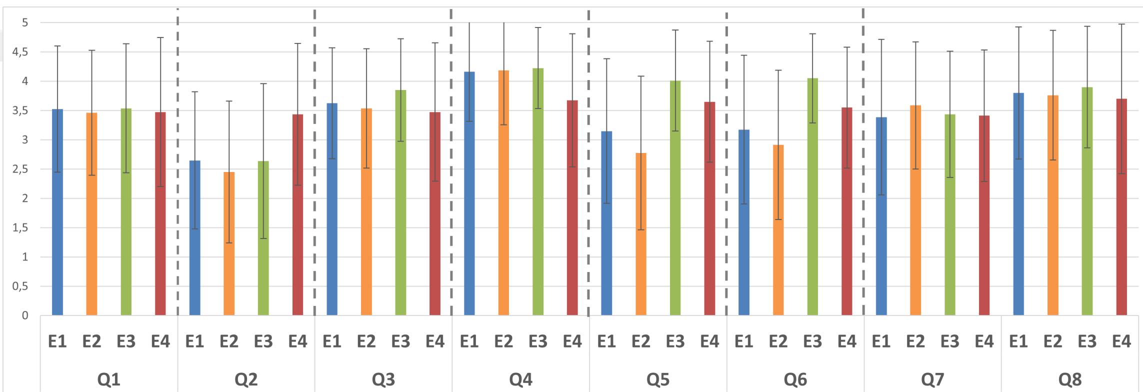


Figure 35: Acceptance Rates of Each Type of Recommendation Engine

Moreover, we conducted an analysis of user responses to the post-experiment survey to assess their perceptions of the explanations offered during the recommendations. Since each participants were given each form of explanations during the experiment, and the survey questions as well as the provided examples of the types were identical for all participants, we employed a within-subjects statistical comparison test. Figure 36 shows the average responses to the questionnaire questions, as well as the questions and respective explanations.



Id	Label
Q1	This type of explanation for recommendations has helped me choose the most convenient recipe.
Q2	This type of explanation for recommendations were too detailed.
Q3	This type of explanation displayed during the interaction were satisfactory.
Q4	This type of explanation for recommendations were clear and easy to understand.
Q5	This type of explanation were sufficient to make an informed decision for healthiness.
Q6	This type of explanation were realistic in terms of healthiness of given recipes.
Q7	This type of explanation let me know how convenient the recipe is.
Q8	Rate your appreciation of the idea of receiving this type of explanations in addition to recommendations.
E1	Popularity-based explanation sample
E2	Cluster-based explanation sample
E3	Tree-based explanation sample
E4	Contrastive explanation sample

Figure 36: Evaluation of the Explanations Questionnaire Results, Shown per Type of Explanation

Table 4 shows the Wilcoxon paired test results for each type of explanation for each question. We observe significant differences between pairs of explanations on Q2, Q3, Q4, Q5 and Q6 whereas there is no significance within explanation pairs for Q1, Q7 and Q8. The Q3 tells us that tree based explanations were the most satisfactory explanations, it is also the simplest explanation generation method. One could draw the conclusion that the participants favor simplistic methods over complicated ones.

This finding is further supported by Q4, where contrastive explanations were found too complicated and the fact that they were rated lowest among the explanation types. The simpler explanations were found to be more effective in coming to healthy decisions, as it is seen from Q5. Finally, Q8 shows us that the users would still use this system despite it's short-comings with no significant difference among pairs of explanations.

Table 4: Pairwise Wilcoxon Test Result p-values to Show Significance

Pairwise P-Value	**Q1**	**Q2**	**Q3**	**Q4**	**Q5**	**Q6**	**Q7**	**Q8**
E1 vs E2	0.565	0.045	0.549	0.679	**0.007**	0.063	0.294	0.909
E1 vs E3	0.986	0.683	**0.010**	0.573	$\leq$ **0.000**	$\leq$ **0.001**	0.902	0.579
E1 vs E4	0.878	$\leq$ **0.001**	0.426	**0.001**	**0.003**	**0.012**	0.859	0.635
E2 vs E3	0.524	0.053	**0.034**	0.621	$\leq$ **0.001**	$\leq$ **0.001**	0.498	0.475
E2 vs E4	0.855	$\leq$ **0.001**	0.638	$\leq$ **0.001**	$\leq$ **0.001**	$\leq$ **0.001**	0.284	0.751
E3 vs E4	0.656	$\leq$ **0.001**	**0.017**	$\leq$ **0.001**	**0.023**	**0.001**	0.641	0.228

CHAPTER VII

CONCLUSION

The advent of complex machine learning models has led to remarkable improvements in recommendation accuracy within the domain of recommender systems but at the cost of a diminished ability to comprehend and interpret the recommendations. The skepticism increases exponentially when the decisions are safety-critical (i.e., AI outcomes can significantly influence people's life and health — like nutrition). This thesis first presents an interactive explainable recommendation framework where the system seamlessly negotiates with its users by making offers and explaining why the recommendation is good for them. In return, the user can criticize the recommendation and/or associated explanation. The ultimate goal of the framework is to improve the system's transparency via the incorporation of interactive explanations. Therefore, we incrementally implement and experiment with various techniques from different vantage points. First, we utilize less complex methods to generate explanations. Then, we advance in post-hoc explanation generation techniques by drawing inspiration from the explainable AI community. Finally, we re-link the recommendation and explanation generation processes, forming a more harmonious process.

User experiments have been conducted to evaluate the proposed systems at each system iteration. First, a comparative analysis was held where the participants were asked to experience an interactive recommender system, incorporating our developments so far, and a regular one, a system version without the dialogical additions of explanation and feedback mechanism. In all our results combined, one aspect is certain: The users appreciate the transparency provided by the system. Moreover, they perceived that the information and process for choosing their food recipe were more informative and complete in the proposed interactive recommender setting. They felt the system was more sociable and reactive to their feedback. Additionally, the users reported an insignificant difference in cognitive load between sessions with and without explanations. Although both types of recommenders might have recommended the same food item (in the same conditions, utilizing the same strategy to decide recommendations), experimental results showed that the participants were more satisfied generally with the idea of explanations and appreciated the generated explanations. Furthermore, interactive sessions performed slightly better in terms of effectiveness regarding the number of agreements and rounds. Thus, we can draw the conclusion that the introduction of explanations had an overall positive effect over user decisions without costing much in terms of user's attention and happiness.

Furthermore, we held a final experiment to assess the performance of each type of explanation generation methods according to the user's perceptions where we discovered simpler and more digestable forms of explanations serve better for user experience and convincability. A key lesson from this study is that the user experience and interface of the system strongly impacts the recommendation success. Users can get frustrated or lose focus if the explanations aren't clear or if other aspects of the recommendation grab their attention first. An interesting vein to explore would be the utilization of large language models to further enhance the sociability of a recommendation system. A system that can actually chat with a user on and off topic might be more intriguing for the user since it would be more sociable and the explanations would be conveyed smoothly.

We mostly utilized the Decision Tree and Random Forest Machine Learning models to foster explainability via feature importance analysis. On the other hand, recent research has sought to introduce explainability through counter-factual explanations. Accordingly, we plan to explore the following questions: (i) How could we use counter-factuality to explain recommender systems? (ii) In what other ways can we use the

Machine Learning methods used in our proposed approach? There is undoubtedly an opportunity for enhancing the recommendation algorithm, such as refining its ability to learn user preferences over time and adapting its behavior accordingly in a multiple-shot setting. In the future, we aim to explore the impact of the timing of explanation display on the user interaction and decision-making process. It is important to note that the current system generates explanations each time a recommendation is provided. Additionally, an intriguing topic to investigate involves assessing the efficacy of on-demand explanations and identifying the circumstances under which such demand arises.

REFERENCES

[1] R. Y. Toledo, A. A. Alzahrani, and L. Martinez, "A food recommender system considering nutritional information and user preferences," *IEEE Access*, vol. 7, pp. 96695–96711, 2019.

[2] P. Langley, B. Meadows, M. Sridharan, and D. Choi, "Explainable agency for intelligent autonomous systems," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 31, pp. 4762–4763, 02 2017.

[3] S. Corrado, G. Luzzani, M. Trevisan, and L. Lamastra, "Contribution of different life cycle stages to the greenhouse gas emissions associated with three balanced dietary patterns," *Science of the Total Environment*, vol. 660, pp. 622–630, 2019.

[4] T. N. T. Tran, A. Felfernig, C. Trattner, and A. Holzinger, "Recommender systems in the healthcare domain: state-of-the-art and research issues," *Journal of Intelligent Information Systems*, vol. 57, no. 1, pp. 171–201, 2021.

[5] D. Calvaresi, R. Carli, J.-G. Piguet, V. H. Contreras, G. Luzzani, A. Najjar, J.-P. Calbimonte, and M. Schumacher, "Ethical and legal considerations for nutrition virtual coaches," *AI and Ethics*, vol. 3, pp. 1–28, 2022.

[6] T. N. T. Tran, M. Atas, A. Felfernig, and M. Stettinger, "An overview of recommender systems in the healthy food domain," *Journal of Intelligent Information Systems*, vol. 50, pp. 501–526, 06 2018.

[7] M. J. Gibney, C. G. Forde, D. Mullally, and E. R. Gibney, "Ultra-processed foods in human health: a critical appraisal," *The American journal of clinical nutrition*, vol. 106, no. 3, pp. 717–724, 2017.

[8] I. Padhiar, O. Seneviratne, S. Chari, D. Gruen, and D. L. McGuinness, "Semantic modeling for food recommendation explanations," in *2021 IEEE 37th International Conference on Data Engineering Workshops (ICDEW)*, pp. 13–19, 2021.

[9] D. Lawo, T. Neifer, M. Esau, and G. Stevens, "Buying the 'right' thing: Designing food recommender systems with critical consumers," in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, (New York, NY, USA), pp. 1–13, Association for Computing Machinery, 2021.

[10] D. Jannach, P. Pu, F. Ricci, and M. Zanker, "Recommender systems: Past, present, future," *AI Magazine*, vol. 42, pp. 3–6, 2021.

[11] B. Buzcu, V. Varadhajaran, I. Tchappi, A. Najjar, D. Calvaresi, and R. Aydoğan, "Explanation-based negotiation protocol for nutrition virtual coaching," in *Principles and Practice of Multi-Agent Systems (PRIMA) 2022*, (Valencia , Spain), p. 20–36, Springer-Verlag, 2022.

[12] F. Ricci, L. Rokach, and B. Shapira, *Introduction to Recommender Systems Handbook*, pp. 1–35. Boston, MA: Springer US, 2011.

[13] A. Singhal and I. Google, "Modern information retrieval: A brief overview," *IEEE Data Engineering Bulletin*, vol. 24, pp. 35–43, 01 2001.

[14] D. Freedman, R. Pisani, and R. Purves, "Statistics (international student edition)," *Pisani, R. Purves, 4th edn. WW Norton & Company, New York*, 2007.

[15] F. Sovrano and F. Vitali, "Generating user-centred explanations via illocutionary question answering: From philosophy to interfaces," *ACM Transactions on Interactive Intelligent Systems*, vol. 12, no. 4, pp. 1–32, 2022.

[16] N. Tintarev and J. Masthoff, *Explaining Recommendations: Design and Evaluation*, pp. 353–382. Boston, MA: Springer US, 2015.

[17] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi, "A survey of methods for explaining black box models," *ACM computing surveys (CSUR)*, vol. 51, no. 5, pp. 1–42, 2018.

[18] J. van der Waa, E. Nieuwburg, A. Cremers, and M. Neerincx, "Evaluating xai: A comparison of rule-based and example-based explanations," *Artificial Intelligence*, vol. 291, p. 103404, 2021.

[19] T. Miller, "Explanation in artificial intelligence: Insights from the social sciences," *Artificial intelligence*, vol. 267, pp. 1–38, 2019.

[20] M. Saarela and S. Jauhiainen, "Comparison of feature importance measures as explanations for classification models," *SN Applied Sciences*, vol. 3, p. 272, 02 2021.

[21] Y. Xu, J. Collenette, L. Dennis, and C. Dixon, "Dialogue-based explanations of reasoning in rule-based systems," in *3rd Workshop on Explainable Logic-Based Knowledge Representation*, June 2022.

[22] M. Ancona, E. Ceolini, A. C. Öztireli, and M. H. Gross, "A unified view of gradient-based attribution methods for deep neural networks," 2017.

[23] X. Zhu, D. Wang, W. Pedrycz, and Z. Li, "Fuzzy rule-based local surrogate models for black-box model explanation," *IEEE Transactions on Fuzzy Systems*, vol. 31, no. 6, pp. 2056–2064, 2023.

[24] D. Cemiloglu, M. Catania, and R. Ali, "Explainable persuasion in interactive design," in *2021 IEEE 29th International Requirements Engineering Conference Workshops (REW)*, pp. 377–382, 2021.

[25] F. Gedikli, M. Ge, and D. Jannach, "Understanding recommendations by reading the clouds," in *E-Commerce and Web Technologies*, (Berlin, Heidelberg), pp. 196–208, Springer Berlin Heidelberg, 2011.

[26] J. Herlocker, J. Konstan, and J. Riedl, "Explaining collaborative filtering recommendations," *Proceedings of the ACM Conference on Computer Supported Cooperative Work*, pp. 241–250, 01 2001.

[27] A. Sharma and D. Cosley, "Do social explanations work? studying and modeling the effects of social explanations in recommender systems," *WWW 2013 - Proceedings of the 22nd International Conference on World Wide Web*, pp. 1133––1144, 04 2013.

[28] M. Millecamp, N. N. Htun, C. Conati, and K. Verbert, "To explain or not to explain: The effects of personal characteristics when explaining music recommendations," in *Proceedings of the 24th International Conference on Intelligent User Interfaces*, p. 397–407, 2019.

[29] P. Pu and L. Chen, "Trust building with explanation interfaces," in *International Conference on Intelligent User Interfaces, Proceedings IUI*, vol. 2006, pp. 93–100, 01 2006.

[30] P. Symeonidis, A. Nanopoulos, and Y. Manolopoulos, "Moviexplain: A recommender system with explanations," in *Proceedings of the Third ACM Conference on Recommender Systems*, RecSys '09, (New York, NY, USA), p. 317–320, Association for Computing Machinery, 2009.

[31] K. J. Hammond, "Chef: A model of case-based planning," in *AAAI*, pp. 267–271, 1986.

[32] J. Freyne and S. Berkovsky, "Recommending food: Reasoning on recipes and ingredients," in *User Modeling, Adaptation, and Personalization*, vol. 6075, pp. 381–386, 06 2010.

[33] M. Ge, F. Ricci, and D. Massimo, "Health-aware food recommender system," in *Proceedings of the 9th ACM Conference on Recommender Systems*, RecSys '15, (New York, NY, USA), p. 333–334, Association for Computing Machinery, 2015.

[34] Y.-L. Chi, T.-Y. Chen, and W.-T. Tsai, "A chronic disease dietary consultation system using owl-based ontologies and semantic rules," *Journal of Biomedical Informatics*, vol. 53, pp. 208–219, 2015.

[35] M. Chen, X. Jia, E. Gorbonos, C. T. Hoang, X. Yu, and Y. Liu, "Eating healthier: Exploring nutrition information for healthier recipe recommendation," *Information Processing & Management*, vol. 57, no. 6, p. 102051, 2020.

[36] C.-Y. Teng, Y.-R. Lin, and L. A. Adamic, "Recipe recommendation using ingredient networks," in *Proceedings of the 4th Annual ACM Web Science Conference*, WebSci '12, p. 298–307, 2012.

[37] "Sigir '17: Proceedings of the 40th international acm sigir conference on research and development in information retrieval," (New York, NY, USA), Association for Computing Machinery, 2017.

[38] P. Forbes and M. Zhu, "Content-boosted matrix factorization for recommender systems: experiments with recipe recommendation," in *Proceedings of the fifth ACM conference on Recommender systems*, pp. 261–264, 2011.

[39] M. El-Dosuky, M. Z. Rashad, T. Hamza, and A. El-Bassiouny, "Food recommendation using ontology and heuristics," in *Advanced Machine Learning Technologies and Applications: First International Conference, AMLTA 2012, Cairo, Egypt, December 8-10, 2012. Proceedings 1*, pp. 423–429, Springer, 2012.

[40] C.-Y. Teng, Y.-R. Lin, and L. A. Adamic, "Recipe recommendation using ingredient networks," in *Proceedings of the 4th annual ACM web science conference*, pp. 298–307, 2012.

[41] M. Ueda, S. Asanuma, Y. Miyawaki, and S. Nakajima, "Recipe recommendation method by considering the users preference and ingredient quantity of target recipe," in *Proceedings of the international multiconference of engineers and computer scientists*, vol. 1, pp. 12–14, 2014.

[42] T. Mokdara, P. Pusawiro, and J. Harnsomburana, "Personalized food recommendation using deep neural network," in *2018 Seventh ICT International Student Project Conference (ICT-ISPC)*, pp. 1–4, IEEE, 2018.

[43] M. Ueda, M. Takahata, and S. Nakajima, "User's food preference extraction for personalized cooking recipe recommendation," in *Workshop of ISWC*, pp. 98–105, Citeseer, 2011.

[44] S. Anjomshoae, A. Najjar, D. Calvaresi, and K. Främling, "Explainable agents and robots: Results from a systematic literature review," in *AAMAS, Montreal, Canada, May 13–17, 2019*, pp. 1078–1088, 2019.

[45] D. Gunning and D. Aha, "Darpa's explainable artificial intelligence (xai) program," *AI Magazine*, vol. 40, pp. 44–58, 6 2019.

[46] A. Vultureanu-Albişi and C. Bădică, "Recommender systems: An explainable ai perspective," in *2021 International Conference on INnovations in Intelligent SysTems and Applications (INISTA)*, pp. 1–6, 2021.

[47] A. Samih, A. Adadi, and M. Berrada, "Towards a knowledge based explainable recommender systems," in *Proceedings of the 4th International Conference on Big Data and Internet of Things*, BDIoT'19, (New York, NY, USA), pp. 1–5, Association for Computing Machinery, 2019.

[48] R. Aydoğan, D. Festen, K. V. Hindriks, and C. M. Jonker, "Alternating offers protocols for multilateral negotiation," in *Modern Approaches to Agent-based Complex Automated Negotiation*, pp. 153–167, Springer, 2017.

[49] A. Starke, C. Trattner, H. Bakken, M. Johannessen, and V. Solberg, "The cholesterol factor: Balancing accuracy and health in recipe recommendation through a nutrient-specific metric," in *CEUR Workshop Proceedings*, vol. 2959, 01 2021.

[50] “Yemek tarifleri.” https://www.diyetkolik.com/yemek-tarifleri/, 2022. Accessed: 2022-01-01.

[51] A. Ishizaka and S. Siraj, “Are multi-criteria decision-making tools useful? an experimental comparative study of three methods,” *European Journal of Operational Research*, vol. 264, no. 2, pp. 462–471, 2018.

[52] L. Wang, R. M. Rodríguez-Domínguez, Y.-M. Wang, *et al.*, “A dynamic multi-attribute group emergency decision making method considering experts’ hesitation,” *International Journal of Computational Intelligence Systems*, vol. 11, pp. 163–182, 01 2018.

[53] J. A. Harris and F. G. Benedict, “A biometric study of human basal metabolism,” *Proceedings of the National Academy of Sciences*, vol. 4, no. 12, pp. 370–373, 1918.

[54] R. Ayub, M. a. Ghazanfar, M. Maqsood, and A. Saleem, “A jaccard base similarity measure to improve performance of cf based recommender systems,” in *2018 International Conference on Information Networking (ICOIN)*, pp. 1–6, 01 2018.

[55] J. Lazar, J. Feng, and H. Hochheiser, *Research Methods in Human-Computer Interaction*. 04 2017.

[56] M. T. Ribeiro, S. Singh, and C. Guestrin, “”why should i trust you?” explaining the predictions of any classifier,” in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1135–1144, 2016.

[57] R. Confalonieri and T. R. Besold, “Trepan reloaded: A knowledge-driven approach to explaining black-box models,” in *European Conference on Artificial Intelligence*, pp. 2457–2464, 2020.

[58] P. Jaccard, “The distribution of the flora in the alpine zone.1,” *New Phytologist*, vol. 11, no. 2, pp. 37–50, 1912.

[59] J. Hulstijn, I. Tchappi, A. Najjar, and R. Aydoğan, “Metrics for evaluating explainable recommender systems,” in *AAMAS, EXTRAAMAS 2023, London, England, May 29, 2023*, pp. 212—-230, Springer, 2023.

[60] R. R. Hoffman, S. T. Mueller, G. Klein, and o. Litman, “Metrics for explainable ai: Challenges and prospects,” *arXiv:1812.04608 [cs.AI]*, 2018.

[61] C.-L. Liao and S.-J. Lee, “A clustering based approach to improving the efficiency of collaborative filtering recommendation,” *Electronic Commerce Research and Applications*, vol. 18, pp. 1–9, 2016.

[62] F. Cantürk and R. Aydoğan, “Explainable active learning for preference elicitation,” p. 25, 08 2023.

[63] A. Drewnowski and V. L. Fulgoni, "Nutrient density: principles and evaluation tools123," *The American Journal of Clinical Nutrition*, vol. 99, no. 5, pp. 1223S–1228S, 2014.

[64] S. A. Gravina, G. L. Yep, and M. Khan, "Human biology of taste," *Annals of Saudi Medicine*, vol. 33, no. 3, pp. 217–222, 2013.

[65] H. Shimazu, "Expertclerk: Navigating shoppers' buying process with the combination of asking and proposing," in *Proceedings of the 17th International Joint Conference on Artificial Intelligence - Volume 2*, IJCAI'01, (San Francisco, CA, USA), p. 1443–1448, Morgan Kaufmann Publishers Inc., 2001.

VITA

Berk Buzcu earned his Bachelor of Science in Computer Science from Özyeğin University in 2020. Subsequently, he pursued further academic education as a Master's Student at the same institution. During this time, he actively participated in the CHIST-ERA (locally Tübitak) funded project, "EXPECTATION: Personalized Explainable AI for decentralized agents with heterogeneous knowledge," His focus within the scope of the project was the application of explainability techniques in recommender systems which aligned with his research interests. The main goal was to enhance the success of recommender systems in achieving user goals via sociability. In addition to his thesis research, Berk developed a human-robot negotiation environment while also touching other multi-agent related topics that involve Blockchain and drone co-operation.