



FACULTY OF INFORMATION TECHNOLOGY AND ELECTRICAL ENGINEERING

Mahdi Akbari

**Recipe Popularity Prediction in Finnish Social Media by
Machine Learning Models**

Master's Thesis

Degree Programme in Computer Science and Engineering

September 2023

Akbari M. (2023) Recipe Popularity Prediction in Finnish Social Media by Machine Learning Models. University of Oulu, Degree Programme in Computer Science and Engineering. Master's Thesis, 55 p.

ABSTRACT

In recent times, the internet has emerged as a primary source of cooking inspiration, eating experiences and food social gathering with a majority of individuals turning to online recipes, surpassing the usage of traditional cookbooks. However, there is a growing concern about the healthiness of online recipes. This thesis focuses on unraveling the determinants of online recipe popularity by analyzing a dataset comprising more than 5000 recipes from Valio, one of Finland's leading corporations. Valio's website serves as a representation of diverse cooking preferences among users in Finland. Through examination of recipe attributes such as nutritional content (energy, fat, salt, etc.), food preparation complexity (cooking time, number of steps, required ingredients, etc.), and user engagement (the number of comments, ratings, sentiment of comments, etc.), we aim to pinpoint the critical elements influencing the popularity of online recipes. Our predictive model-Logistic Regression (classification accuracy and F1 score are 0.93 and 0.9 respectively)- substantiates the existence of pertinent recipe characteristics that significantly influence their rates. The dataset we employ is notably influenced by user engagement features, particularly the number of received ratings and comments. In other words, recipes that garner more attention in terms of comments and ratings tend to have higher rates values (i.e., more popular). Additionally, our findings reveal that a substantial portion of Valio's recipes falls within the medium health Food Standards Agency (FSA) score range, and intriguingly, recipes deemed less healthy tend to receive higher average ratings from users. This study advances our comprehension of the factors contributing to the popularity of online recipes, providing valuable insights into contemporary cooking preferences in Finland as well as guiding future dietary policy shift.

Keywords: Online Recipes Popularity, User Engagement, Recipe Rating, Classification, Logistic Regression, Finnish Food Social Media.

Akbari M. (2023) Reseptin suosion ennustaminen suomalaisessa sosiaalisessa mediassa koneoppimismalleilla. Oulun yliopisto, Tietojenkäsittelytieteen ja tekniikan koulutusohjelma. Diplomityö, 55 s.

TIIVISTELMÄ

Internet on viime aikoina noussut ensisijaiseksi inspiraation lähteeksi ruoanlaitossa, ja suurin osa ihmisistä on siirtynyt käyttämään verkkoreseptejä perinteisten keittokirjojen sijaan. Huoli verkkoreseptien terveellisyydestä on kuitenkin kasvava. Tämä opinnäytetyö keskittyy verkkoreseptien suosioon vaikuttavien tekijöiden selvittämiseen analysoimalla yli 5000 reseptistä koostuvaa aineistoa Suomen johtavalta maitotuoteyritykseltä, Valiolta. Valion verkkosivujen reseptit edustavat monipuolisesti suomalaisten käyttäjien ruoanlaittotottumuksia. Tarkastelemalla reseptin ominaisuuksia, kuten ravintoarvoa (energia, rasva, suola, jne.), valmistuksen monimutkaisuutta (keittoaika, vaiheiden määrä, tarvittavat ainesosat, jne.) ja käyttäjien sitoutumista (kommenttien määrä, arviot, kommenttien mieliala, jne.), pyrimme paikantamaan kriittiset tekijät, jotka vaikuttavat verkkoreseptien suosioon. Ennustava mallimme - Logistic Regression (luokituksen tarkkuus 0,93 ja F1-pisteet 0,9) - osoitti merkitsevien reseptiominaisuuksien olemassaolon. Ne vaikuttivat merkittävästi reseptien suosioon. Käyttämiimme tietojoukkoihin vaikuttivat merkittävästi käyttäjien sitoutumisominaisuudet, erityisesti vastaanotettujen arvioiden ja kommenttien määrä. Toisin sanoen reseptit, jotka saivat enemmän huomiota kommenteissa ja arvioissa, olivat yleensä suosituimpia. Lisäksi selvisi, että huomattava osa Valion resepteistä kuuluu keskitason terveystieteiden alueelle (arvioituna FSA Scorella), ja mielenkiintoisesti, vähemmän terveellisiksi katsotut reseptit saavat käyttäjiltä yleensä korkeamman keskiarvon. Tämä tutkimus edistää ymmärrystä verkkoreseptien suosioon vaikuttavista tekijöistä ja tarjoaa arvokasta näkemystä nykypäivän ruoanlaittotottumuksista Suomessa.

Avainsanat: verkkoreseptien suosio, käyttäjien sitoutuminen, reseptin arvosana, luokittelu, logistinen regressio, suomalainen ruoka sosiaalisessa mediassa.

TABLE OF CONTENTS

ABSTRACT

TIIVISTELMÄ

TABLE OF CONTENTS

FOREWORD

ABBREVIATIONS

| | | |
|--------|--------------------------------------|----|
| 1. | INTRODUCTION | 10 |
| 1.1. | Background of the Research..... | 10 |
| 1.2. | Related Works | 11 |
| 1.3. | Research Questions | 17 |
| 1.4. | Contributions | 17 |
| 1.5. | Structure of the Thesis | 18 |
| 2. | IMPLEMENTATION | 20 |
| 2.1. | Dataset and Preprocessing | 20 |
| 2.2. | Feature Engineering..... | 22 |
| 2.2.1. | Feature Selection Scenarios..... | 22 |
| 2.2.2. | Sentiment Analysis of Comments | 23 |
| 2.2.3. | Feature Importance by SHAP | 24 |
| 2.3. | Recipe Ratings Annotation..... | 25 |
| 2.4. | Classifiers | 26 |
| 2.4.1. | Logistic Regression | 26 |
| 2.4.2. | Multilayer Perceptron..... | 26 |
| 2.4.3. | Random Forest..... | 27 |
| 2.4.4. | Support Vector Machine..... | 27 |
| 2.4.5. | Gradient Boosting..... | 28 |

| | | |
|--------|--|----|
| 2.5. | Model Selection..... | 29 |
| 2.5.1. | Cross-Validation..... | 29 |
| 2.5.2. | Evaluation Metrics..... | 30 |
| 2.6. | Evaluation of Recipe Healthiness..... | 31 |
| 3. | RESULT AND DISCUSSION..... | 32 |
| 3.1. | Insight into Data | 32 |
| 3.2. | Classification Results (RQ1) | 32 |
| 3.2.1. | Recipes year of publication | 32 |
| 3.2.2. | Model Selection..... | 35 |
| 3.2.3. | Feature Selection Scenarios..... | 35 |
| 3.2.4. | Confusion Matrix of Classifiers and Analysis of Errors | 37 |
| 3.3. | Feature Importance (RQ2)..... | 40 |
| 3.4. | Comparison to Similar Studies (RQ3)..... | 41 |
| 3.5. | Healthiness of Recipes (RQ4) | 42 |
| 4. | CONCLUSION | 45 |
| 4.1. | Limitations and Future Works | 46 |
| 5. | REFERENCES | 49 |
| 6. | APPENDICES | 55 |
| 6.1. | Logistic Regression Classifier SHAP Values..... | 55 |
| 6.2. | Random Forest Classifier SHAP Values..... | 56 |
| 6.3. | Gradient Boosting Classifier SHAP Values | 57 |

Figures

| | |
|--|----|
| Figure 1- Visual representation depicting the steps for forecasting the popularity of Valio's online recipes. | 19 |
| Figure 2- High-level diagram of the various phases of our study. | 20 |
| Figure 3- A sample of recipe profile collected from https://www.valio.fi/ | 21 |
| Figure 4- FSA health score classes. | 31 |
| Figure 5. Correlation heat map of the retrieved dataset from Valio website. | 33 |
| Figure 6- a) Frequency of recipes based on their year of publication, and b) boxplot of given rates to recipes based on the recipes' year of publication. | 34 |
| Figure 7- The defined classes for each recipe based on given rates by users. | 34 |
| Figure 8- a) Acc., and b) F1 values for different classifiers in each scenario. | 36 |
| Figure 9- Based on Scenario 3 and best hyperparameters of classifiers: a) Confusion matrix of SVM, Random Forest, Gradient boost, and MLP, and b) Confusion matrix of the Logistic Regression with the distribution of errors (89+3) on the year of recipes' publication. | 38 |
| Figure 10- Results of the trained Logistic Regression for: a) before and b) after 2016. | 39 |
| Figure 11- SHAP values for a) Gradient Boosting, b) Logistic Regression, and c) Random Forest. | 40 |
| Figure 12- a) The distribution of FSA Score for recipes published in the Valio website, and b) the relation between given rates by users to recipes and the recipes' FSA score. | 44 |

Tables

| | |
|---|----|
| Table 1- Basic statistics of the crawled dataset. | 22 |
| Table 2- Hyperparameters for grid search model selection scheme. | 30 |
| Table 3- Best hyperparameters detected by grid search. | 35 |
| Table 4- Comparison between this thesis and similar study. | 42 |
| Table 5- Most frequent tags in Valio recipes. | 48 |

FOREWORD

I am immensely grateful for the invaluable support and guidance that I have received throughout the journey of my master's thesis. This work would not have been possible without the contributions and encouragement of numerous individuals who have been instrumental in shaping the outcome of this research.

First, I extend my gratitude to my supervisor, Professor Mourad Oussalah. I am fortunate to have the opportunity to be part of his research team. I would also like to express my heartfelt appreciation to MSc. Mehrdad Rostami. Mehrdad's willingness to share knowledge and support were invaluable, and I am grateful for the camaraderie we shared in this academic journey. Finally, I must thank my kind friend Lauri Ikkala who helped me translate the abstract of the thesis into Finnish language.

I owe a debt of gratitude to my father, mother and brother for their unyielding encouragement and belief in my abilities. Their unwavering support has been my driving force, and I am profoundly thankful for their sacrifices and love.

Lastly, I extend my deepest appreciation to my wife, Aie. Her unwavering support, patience, and understanding during this demanding period since we moved to Finland in 2019 have been my constant motivation. Her love and encouragement have been my anchor, and I am forever grateful for her presence in my life. After a long journey, now we are opening a new season of our life.

I hope this thesis serves as a small contribution to the world of knowledge.

Oulu, 29.09.2023

Mahdi Akbari

ABBREVIATIONS

| | |
|--------|-------------------------------|
| Acc. | Accuracy evaluation metric |
| C | Regularization strength |
| FSA | Food Standards Agency |
| L1 | Ridge Regularization |
| L2 | Lasso Regularization |
| LDA | Latent Dirichlet Allocation |
| MLP | Multilayer Perceptron |
| NLP | Natural Language Processing |
| ReLU | Rectified Linear Unit |
| RQ | Research Question |
| SHAP | SHapley Additive exPlanations |
| SVM | Support Vector Machine |
| WHO | World Health Organization |
| XAI | Explainable AI |
| Φ | SHAP Value |

1. INTRODUCTION

1.1. Background of the Research

Over the past few years, the internet has emerged as a go-to source for cooking inspiration and innovative cooking ideas. As an illustration, Valio Ltd, a Finnish dairy product manufacturer and one of Finland's largest corporations, operates an online platform dedicated to food recipes. This platform has witnessed a consistent increase in its user base year after year, reaching over 25 million visits annually. Similar digital recipe hubs are gaining momentum globally. One of the most widely recognized food websites in the United States is Allrecipes.com, boasting an impressive user base of around 7 million subscribers. The platform hosts a vast collection of culinary knowledge, with approximately 750,000 recipes uploaded and a staggering 180 million recipe views. In addition to Allrecipes.com, there exists other food-related websites, including Kochbar.de, Ichkoche.at, and Chefkoch.de (primarily serving German-speaking audiences).

Moreover, based on a survey, it is already revealed that more than half of the participants relied on online sources for their cooking needs [1]. These statistics suggest that the contemporary trend in searching for recipes online has potentially eclipsed the traditional use of cookbooks. This transition is hardly surprising, given the multitude of advantages offered by online cooking communities. These platforms are easily accessible via computers and smartphones, housing extensive databases of recipes spanning various cuisines. Also, they integrate social networking features, allowing users to connect, engage in discussions, establish cooking communities, and share their cooking experiences.

However, there is a downside to the reliance on online recipes, as highlighted by recent research [2]. Many online recipes tend to be less healthy, and users often struggle to differentiate between nutritious and less nutritious options. Ironically, it is the less healthy recipes that often garner more attention and popularity. Another study [3] hinted at possible connections between the popularity of specific online recipes and higher obesity rates in the United States [3]. Therefore, gaining a thorough understanding of the factors driving the popularity of online recipes is crucial for promoting healthier dietary choices.

Each recipe featured on online platforms boasts a unique set of characteristics, encompassing nutritional details, complexity of preparation, publication date, user ratings, comments, and more. The interplay of these attributes collectively determines whether a recipe gains prominence or remains relatively obscure. Nevertheless, the pivotal question of which of these elements significantly influences a recipe's popularity (i.e., rates by users) has remained unanswered. Consequently, the primary objective of this thesis is to illuminate the factors that underpin the popularity of online recipes, with the aim of deepening our comprehension of why certain recipes attain widespread recognition while others remain overlooked. The main objectives, contributions, and structure of this thesis are explained in Sections 1.3, 1.4, and 1.5,

respectively. First, related works in the domain of popularity prediction in social media are discussed below.

1.2. Related Works

In this section, we provide an overview of specific research that leverages food platforms to gain insights into dietary behavior, social and cultural influences, as well as health-related issues within these online communities. We also delve into studies that explore similar themes but utilize alternative online data sources such as Twitter.com. Several of these papers have played a substantial role in shaping the direction and relevance of this thesis. Research utilizing online data to investigate various facets of human nutrition represents a relatively recent and emerging field. When compared to traditional approaches within the realm of nutrition science, such as questionnaires, these studies offer several advantages. They are less intrusive and less prone to biases inherent in self-reported data collection methods. Additionally, online studies often benefit from larger sample sizes and the potential for global scalability [1]. However, it is essential to acknowledge potential drawbacks, such as the assumption that searching for a specific recipe equates to consumption of that dish. Nevertheless, this section will showcase successful studies in this domain, emphasizing the significance and potential of this field of research.

A study by De Choudhury et al. [4] using Twitter data was conducted to understand dietary behavior. They argued that social media, like Twitter, is well-suited for this purpose because users share many aspects of their daily lives, including what they eat. They collected 892,000 tweets containing food-related keywords and linked this data to user demographics, interests, and social connections. To estimate calorie content, they averaged nutritional information from online sources based on specific keywords. Their first study found a strong correlation (77%) between the calorie content of tweeted food and obesity rates across 50 U.S. states. They used this data to build models predicting obesity rates using demographic features and food mentions in tweets. They connected this data with societal factors like education and income, revealing that higher-educated individuals tend to tweet and eat less calorie-dense foods. They also explored the social aspects of obesity using two Twitter networks (friendship and mention networks) and found that friends are more likely to share similar food interests.

Abbar et al. [5] emphasized the growing importance of social media as a valuable resource for public health research, particularly in examining disparities related to food access and health. Their study focused on desserts, which are defined regions marked by limited access to affordable, healthy food. These areas are known to be linked to poor dietary choices and health issues like obesity, diabetes, and heart disease. Precisely identifying these regions and their challenges is a matter of significant public interest. The authors noted that previous studies on desserts often relied on surveys and self-reported data, which lacked robust research methods and sufficient sample sizes. To address these limitations, this study leveraged Instagram, a rapidly growing social media platform where users share photos. The researchers collected data through

Instagram's official API, allowing them to access public images and associated metadata, including food-related hashtags.

Similarly, Fried et al. [6] undertook a study to explore the potential of Twitter posts as a source of information for predicting population characteristics related to food habits and behaviors. Their analysis focused on tweets that were associated with specific food-related hashtags. This data collection effort spanned a period from October 2013 to May 2014, resulting in a substantial dataset comprising 3.5 million tweets. One of the notable achievements of their research was the successful implementation of predictive tasks based on the textual content of these tweets. Their predictive models demonstrated superior performance compared to most baseline studies. These models harnessed the power of approximately 30 million words extracted from the tweets to make predictions. The predictive tasks encompassed a range of population characteristics, including the geographical locations of authors (city, region, state) and state-level attributes such as the prevalence of overweight individuals, diabetes rates, and even political leanings (party preferences). To build these predictive models, the researchers leveraged two sets of features. First, they utilized lexical features, which involved identifying food-related keywords sourced from food glossaries. Second, they employed topic modeling, a technique used to uncover hidden thematic structures within textual data. In particular, their experiments on predicting diabetes yielded valuable insights. Furthermore, the researchers harnessed this data to create various visualizations, including geo-referenced heatmaps, word clouds, and temporal histograms. These visualizations allowed for the discovery of intricate global patterns in terms of food consumption, providing a deeper understanding of the relationships between Twitter discussions, population characteristics, and dietary habits.

Wagner and Aiello [7] conducted a quantitative study using data from the social media platform Flickr to explore gender-based differences in food-related content and associated stereotypes in media. They considered food not only as a basic necessity but also as a means of expressing one's identity in modern societies. Their research aimed to determine if gender-specific uploading patterns existed and what factors drove these patterns. They collected a substantial dataset comprising approximately 15 million Flickr images from 1 million users spanning from 2005 to 2014, with 41% of the users being female. They filtered out non-food-related posts using an online vocabulary of common food-related terms and retained posts with at least one food-related tag and publicly available user gender information. Their findings revealed statistical evidence of specific food types being predominantly posted by one gender. For instance, they observed that beer was 41% more popular among men, with 24% of men posting at least one beer picture compared to 17% of women. In a second study, they analyzed the top 100 images from search engine results for terms like "eating meat" or "eating fish." Crowd workers determined which gender group was more likely to consume the foods depicted in the images (male or female, adults or younger individuals). This survey uncovered intriguing trends, including the popularity of alcohol among men despite frequent promotion with women. Conversely, foods like milk and fast-food were perceived as gender-neutral, while sweets, milk, and coffee were more associated with females in the media. In summary, Wagner and Aiello

argued that their approach could complement traditional surveys on dietary preferences and food consumption, showcasing the potential of this technique.

Chunara et al. [8] conducted a cross-sectional study to investigate the relationship between social networks and obesity prevalence. Their research aimed to determine if the interests expressed by users on Facebook could predict obesity rates in the United States. The authors selected users based on their interests, categorizing them as either positively or negatively related to obesity. For example, activities like "Watching television" were considered sedentary and linked to obesity, while "outdoor fitness activities" indicated an active and healthy lifestyle. They employed linear regression and k-fold cross-validation to model user activity levels, with k-fold cross-validation being particularly useful when dealing with small datasets. It helps improve prediction accuracy and ensures statistical relevance by repeatedly splitting the data into training and test sets. The findings indicated that Facebook users with activity-related interests had a predicted obesity prevalence that was 12% lower across the United States and approximately 7.2% lower in New York areas. In contrast, neighborhoods in New York City with interests such as watching TV showed a 27.5% obesity prevalence. In summary, the study revealed a significant association between non-active interests and obesity, but the authors noted the need for further research to fully comprehend these connections.

In their research, Said and Bellogin [9] investigated the role of social interaction in enhancing potential food recommendation systems within the context of online recipes. They emphasize that, unlike other real-world product recommendations (such as videos or music), food recommendations have a critical health aspect. Any system influencing a user's health must be cautious, irrespective of the business implications for the provider. Additionally, they stress the importance of considering the user's geographical location, as certain regions face a higher risk of food-related health issues. The study utilized a dataset collected from Allrecipes.com in October 2013, containing information from 170 thousand users, 54 thousand recipes, 8400 ingredients, and 17 million ratings. Health-related data, primarily focusing on obesity rates, was sourced from County Health Rankings, including data from over 3400 U.S. counties. Mapping users to their respective counties presented challenges due to users providing location information in freeform text. Manual matching and dataset refinement were employed to address this issue. The initial investigation involved ten counties, specifically those with the lowest and highest obesity rates, and analyzed ingredient frequencies based on user ratings. The study successfully identified these county groups based on ingredient usage frequencies. The authors suggest that this information could be valuable for future personalized food recommendation systems, allowing for the promotion of healthier recipes or the creation of personalized recipes based on individual obesity risks. While acknowledging limitations, such as uncertainty regarding ingredient quantities used in meals and the challenges of identifying ingredients from user-generated freeform text, the authors view this early work as a promising foundation for future research.

Trattner, Elswiler, and Howard [2] conducted a study comparing the healthiness of online-sourced recipes, ready meals, and recipes from cooking books, considering the growing concern about the nutritional aspects of dietary choices and their impact on

health. They noted that health issues have been associated with poor nutritional behavior, and programs like ChooseMyPlate in the US and Change4Life in the UK aim to promote home cooking as a healthier option. However, they argued that the healthiness of a meal depends on what is cooked and how it is prepared. In their study, they performed a statistical analysis of three types of meals commonly consumed in modern societies: recipes sourced from the internet, ready-made meals, and recipes from popular cooking books. They compared 100 recipes from books, 100 ready-made meals, and online recipes from the food platform Allrecipes.com. They collected 5,237 online recipes spanning from 2000 to 2010. For the comparison, they selected main dishes with sufficient nutritional information, such as carbohydrates, sodium, energy content, and fat. To assess the healthiness of these recipes, they applied two international standards: the World Health Organization (WHO) guidelines and the "traffic light" system of the UK's Food Standards Agency (FSA). The WHO score considered the content of seven essential nutrients, while the FSA score categorized recipes based on four major macronutrients (green for healthy and red for unhealthy). In their initial study, they found that only six online recipes fully met the WHO guidelines. Overall, recipes from Allrecipes.com tended to be less healthy, often failing to meet the standards for fat, saturated fat, and fiber content, though they generally met protein requirements. Additionally, they observed that recipes from cooking books were lowest in sodium, followed by internet recipes and ready meals. When it came to sugar content, cooking books and internet recipes often met the criteria equally, while ready meals generally performed the best. In a second study examining the temporal aspect, they found that the results remained consistent over time. The authors concluded that while internet-sourced recipes may not be as healthy as expected, there are limitations to their approach. Variations in actual consumption behavior, such as not following exact ingredients and instructions, as well as potential variations in nutrient values on food labels and in the nutrient calculation approach, need to be considered.

Kusmierczyk and Nørvåg [10] conducted a study to uncover patterns in online recipe titles and explore practical applications based on their findings. Their primary objective was to understand the relationships between the words used in recipe titles and the nutritional values of those recipes. They argued that despite users primarily communicating through text, there is limited information available on the connections between textual content and health-related factors. To achieve this, they analyzed a dataset comprising 204,000 online recipes from Allrecipes.com. During preprocessing, they filtered out recipes lacking sufficient nutrient information, resulting in a dataset of approximately 58,000 recipes. Recipe titles, being short and free-form text, required preprocessing as well. They removed special characters, numbers, and stopwords. Given the potential for ambiguities and misspellings in titles, they applied stemming and retained words that appeared at least twice. This process yielded 4,679 unique words. In their initial experiment, they conducted a statistical analysis of nutrient value distributions for each word found in recipe titles. They used information gained to measure the influence of individual food words on nutrient content. This analysis revealed correlations between specific food words and nutrients, as well as correlations among different nutrients. In their second experiment, they employed a novel approach combining Latent Dirichlet Allocation (LDA) and linear regression to create a low-level and interpretable model based on the findings of the

first experiment. LDA was used to model recipe ingredients, while linear regression linked these variables to nutritional values. Their validation process demonstrated that this model produced the best results. Finally, in their third experiment, they attempted to predict the nutritional values of recipes solely based on the words found in their titles.

Kusmierczyk, Trattner, et al. [11] conducted research in the field of food and recipe innovation within online food communities. They emphasized the importance of innovation for the long-term success of restaurants and chefs but noted a lack of research focused on the virtual dimension of this domain. For their study, they utilized a dataset collected from Kochbar.de, which included over 400,000 recipes spanning from 2008 to 2014. These recipes were accompanied by various metadata, including preparation instructions and categories. They also had data on 230 distinct recipe categories, 200,000 different users, and approximately 7 million recipe ratings. However, only around 5,000 users regularly uploaded recipes (more than 10 recipes each). They chose Kochbar.de for its rich metadata and additional recipe information, such as ingredients and nutrient values, which were crucial for their approach, heavily reliant on ingredient combinations. One major challenge was that the ingredients were only available as free-form text, necessitating preprocessing and filtering. They employed a simple statistical filtering approach, retaining ingredient names that occurred more than 100 times while replacing those occurring less than 200 times with more common alternatives to eliminate ambiguities. This process resulted in 2,208 distinct ingredients (from an initial 334,000 before filtering). In their first study, they focused on exploring community patterns and measured innovation and complexity using three different features, two of which employed entropy and conditional entropy, while the third was an innovation factor metric based on Jaccard similarity, all considering the ingredients of recipes for comparison. They found that although the number of ingredients remained constant, innovation within the community continuously increased. They hypothesized that users combined known ingredients to create new and novel recipes, although the rate of growth was gradually declining, suggesting that innovation might plateau in the future. Innovation in the community also displayed seasonal and time-dependent patterns, with slight variations throughout the year, particularly with peaks at the beginning of the year and after the summer, possibly reflecting increased creativity at these times. In the second study, they examined innovation patterns at the user level, filtering out users with fewer than 10 recipes to enhance the reliability of results. This resulted in a user set of approximately 5,000 users. They observed two distinct types of users: those with lower innovation factors and more innovative users. Analyzing the innovation factors of each user over time using linear regression, they found that for most users, innovation remained relatively consistent over the years. Furthermore, they explored the factors influencing innovation, with user location being the most explanatory feature (measured with information gain). This finding was unexpected and suggested the need for further investigation.

Ahn et al. [12] conducted a comprehensive study aimed at uncovering the fundamental patterns governing ingredient combinations in cuisines worldwide. Their primary goal was to address the question of whether there exist quantifiable and reproducible principles that guide our selection of certain ingredient pairings while

avoiding others in culinary practices. The researchers based their investigation on the intriguing hypothesis of "shared flavor compounds" within combined ingredients. This hypothesis posits that ingredients sharing common flavor components are more likely to complement each other in taste. To illustrate this concept, they cited examples like restaurants pairing white chocolate and caviar due to the presence of the organic compound trimethylamine in both ingredients. They relied on constructing a bipartite graph that connected ingredients and their associated flavor components. On average, most ingredients were found to have 51 such flavor compounds. This ingredient-compound network served as the foundation for formulating and testing their hypothesis through topological properties. Their research journey involved gathering data from three distinct online recipe websites: Allrecipes.com, Epicurious.com (both from the United States), and Menupan.com (Korean). The inclusion of the Korean website was strategic, aiming to avoid a potential Western bias in the results. Their dataset encompassed 1021 unique flavor compounds and 381 distinct ingredients, with an average of eight ingredients assigned to each recipe. In their initial experiment, statistical evidence emerged indicating that North American and Western European cuisines tend to combine ingredients that share more flavor compounds. In stark contrast, Asian cuisine predominantly embraces ingredient pairings characterized by greater contrast in flavor profiles. This observation was subsequently validated in a second experiment, which examined the likelihood of ingredients sharing more compounds being prevalent in specific cuisines. To delve deeper into their findings, the researchers sought to identify the key ingredients responsible for these patterns. They discovered that only a handful of ingredients, though frequently used in specific cuisines, significantly contributed to these effects. For instance, North American cuisine frequently featured ingredients like eggs, cream, cacao, butter, and milk, while East Asian cuisine relied on ingredients such as onions, ginger, pork, and chicken. In a third experiment, Ahn et al. compared various cuisines, revealing that South European and Latin cuisines exhibit greater similarities to Asian cuisine than to Western European cuisine. Notably, these cuisines incorporate ingredients that do not share as many flavors' compounds. Overall, Ahn and his team's research sheds light on the intriguing dynamics of ingredient combinations in different cuisines, providing valuable insights into the complex interplay of flavors, ingredients, and culinary traditions across the globe.

A holistic research by Dominik [1] employs a statistical analysis of datasets from two popular food community websites: Allrecipes.com and Kochbar.de, representing distinct Western food cultures. By analyzing both platforms comparatively, the study seeks to uncover more general insights. The analysis focuses on recipe characteristics and the underlying social networks of these websites, referred to as "features." These features are derived from previous studies on online content popularity and food preferences. To assess the predictive power of these features and test the statistical findings, predictive modeling experiments were conducted. The results indicate the presence of generally valid recipe characteristics that strongly influence the future popularity of online recipes. Notably, user activity features, such as ratings, comments, and the number of uploaded recipes, appear to be significant predictors for the Kochbar.de dataset. In contrast, innovation features like recipe novelty, ingredient popularity rank, and image features (e.g., saturation and image entropy) seem to exert more influence on the popularity of Allrecipes.com recipes.

1.3. Research Questions

The central aim of this master's thesis is to uncover applicable factors that contribute to the popularity of online recipes in Valio website (<https://www.valio.fi/reseptihak u/>). Throughout the progression of this thesis, we will explore the following research inquiries (RQ):

- RQ1: To what degree can we identify trends in popularity within Valio as a Finnish online food social network?
- RQ2: What are the primary factors that influence recipe ratings and popularity in Valio?
- RQ3: How does our study of the Valio as a Finnish online food website compare to other studies on similar platforms in other countries?
- RQ4: What is the health quality of recipes, and how do users' ratings correlate with healthiness in Valio?

1.4. Contributions

The objective of this thesis is to uncover previously undisclosed facets of online recipe popularity within Valio's platform. To conduct feature analysis, we have developed more than 15 distinct features. Additionally, we have introduced novel features related to nutrition, innovation, and healthiness, and preparation difficulty expanding the scope of recipe popularity analysis. These features have been organized into sets, such as nutritional features (e.g., energy, fat, and salt content), preparation difficulty features (e.g., preparation time, number of required ingredients, etc.), and user engagement features (e.g., comments, number of rates, etc.).

Utilizing these features, we have employed a machine learning algorithm to predict the popularity of recipes. This approach is grounded in classification rather than regression analysis. Our experiment evaluates each feature set both individually and in combination, considering various time periods within a recipe's lifecycle. We employed state-of-the-art classifiers, including Logistic Regression, Random Forest, Support Vector Machine, Gradient Boosting, and Multilayer Perceptron neural network. We selected these classifiers based on their popularity in the context this research. Most related literature (explained above), used these classifiers. To assess the performance of these classification models, we adopted a cross-validation protocol, and the results for each feature set are documented. Figure 1 provides a more detailed visualization of this process. It is important to highlight that this research for the first time (i) explores the popularity of the Finnish recipe platform and (ii) investigates the degree to which Finnish recipes achieve popularity in the online domain.

1.5. Structure of the Thesis

This master's thesis is organized into four chapters. After this introductory chapter, Chapter 2 introduces the Valio datasets, which serve as the foundation for this thesis. Additionally, this chapter outlines the chosen methodology for addressing our research questions. It also offers insights into the feature engineering process and the selection of models (classifiers) before any analysis takes place. Moreover, it provides a detailed account of the development of various features designed to capture the popularity of online recipes. Moving forward to Chapter 3, we present and discuss the results of our studies, which aim to uncover potential correlations between recipe and user-related factors and recipe popularity. Finally, in Chapter 4, we draw conclusions based on our research findings and propose directions for future studies in this domain.

Valio Recipes

Rating Analysis
★★★★★

Comments Sentiment:
😊 😐 😊 😄 😁

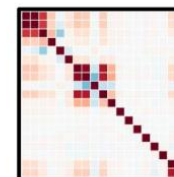
Recipe Healthiness by FSA score



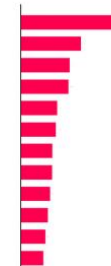
Feature Matrices

| | |
|--|--|
| | |
| | |
| | |
| | |
| | |
| | |

Correlation Heatmap



Feature Importance by SHAP



Best Trained Classifier

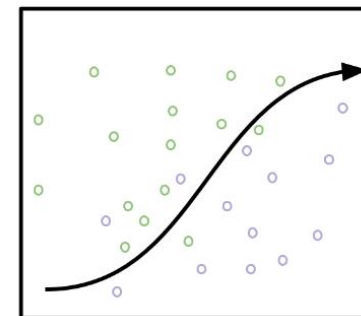


Figure 1- Visual representation depicting the steps for forecasting the popularity of Valio's online recipes.

2. IMPLEMENTATION

The comprehensive diagram illustrating the fundamental stages of our research is displayed in Figure 2. As portrayed in this representation, our study encompasses six principal phases: i) the retrieval and preprocessing of recipes, ii) the extraction of recipe attributes and normalizing them between 0-1, iii) the classification of recipes based on identified features and user-provided ratings, iv) the utilization of the SHAP (SHapley Additive exPlanations) method to assess the importance of each feature, v) a comparison of our results with existing literature, and vi) the determination of health factors relation with ratings. The first three steps address RQ1, while steps four, five, and six correspond to RQ2, RQ3, and RQ4, respectively.

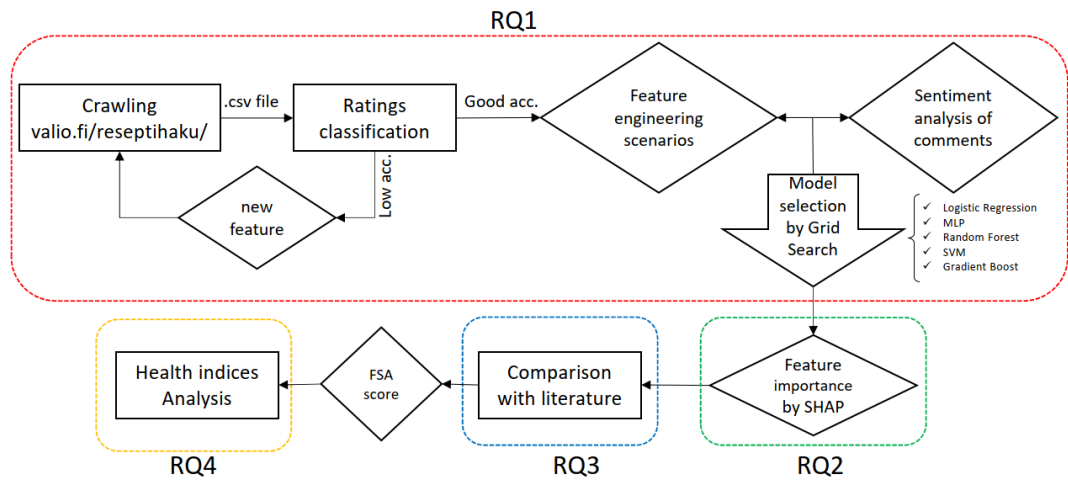



Figure 2- High-level diagram of the various phases of our study.


2.1. Dataset and Preprocessing

We compiled our dataset by collecting data from www.valio.fi (Figure 3), which was selected due to its status as one of the most frequented and largest food-oriented social media platforms, attracting over 25 million visits annually. Our web crawling efforts resulted in the acquisition of 5,472 recipes that had been posted between 2010 and 2022. For each recipe, we gathered various attributes including the Recipe name, Published Time, Ingredients, Preparation Time, Difficulty Level, Tags, Users' Ratings, Users' Comments, and the nutritional content of the recipe per 100 grams (such as Energy, Protein, Carbohydrates, Fat, Saturated Fat, Dietary Fiber, and Salt). Table 1 presents basic statistics regarding this dataset and offers an overview of the extracted entities. To align with the objectives of our study, we only retained recipes that included available ingredients and nutritional information. Specifically, during our crawling phase, we discarded 663 recipes lacking nutritional data while retaining 4,833 recipes that met our criteria. The recipe profile view, illustrated in Figure 3, offers a comprehensive glimpse into the recipe's content.

☰
KOTIKEITTIÖ
AMMATTILAISET
ARKIRUOKA



🔍
📖



Recipe image

Hunajabroilerivuoka

Recipe name and descriptions

Helppo ja edullinen arkiruoka syntyy yhdessä vuossa. Riisit ja juurekset kypsyvät ja maustuvat vuon pohjalla ja pitävät samalla broilerinkoivet mehukkaina.

Uunibroilerit
Uuniruokat
Broilerireseptit

Kananmunaton
Gluteeniton
Laktoositon

Valmistusaika Preparation time

Työaika 15 min

Uunissa 1 h


Yhteensä 1 h 15 min

Vaikeustaso

Aloitteleva kokkaaja 🔥🔥🔥

Difficulty level

★★★★★ 3 arvostelua



Siirry samankaltaisiin resepteihin

AINEKSET

−
4 annosta
+

Ingredients

4 dl pitkäjyväistä riisiä

2 pientä punasipulia

150 g palsternakkaa

150 g porkkanaa

6 dl kanalientä

1 prk (2 dl) **Valio Hyvä suomalainen Arki® ruokakermaa**

2 tl sinappia

2 rkl hunajaa

n. 1 kg marinoituja broilerinkoipia

Cocking direction

OHJE

- Mittaa riisi isoon uunivuokaan. Kuori ja paloittele sipulit, palsternakka ja porkkana. Lisää vuokaan.
- Sekoita kanaliemeen ruokakerma, sinappi ja hunaja. Kaada vuokaan. Tarvittavan liemen määrään vaikuttaa se kuinka laakeassa astiassa vuon valmistat. Lisää hieman vettä jos valmistat ruoan laakeassa astiassa esim. pellillä.
- Nostele broilerinkoivet päällimmäiseksi. Kaada myös marinadi vuokaan.
- Paista uunin keskiosassa 200 asteessa n. 1 h, riippuen koipien koosta.

VINKKI

Tuunaa ruoasta omaan makuun sopiva annos lisäämällä esim. kukkakaalia, parsakaalia, tuoreita tomaatteja, maustekastikkeita tai sitruunaa.

📄 LISÄÄ KAUPPALISTALLE

Nutrition factors

| Energia | Proteiini | Hiilihydraatit | Rasva | Tyydyttynyt rasva | Ravintokuitu | Suola |
|----------|-----------|----------------|-------|-------------------|--------------|-------|
| 123 kcal | 7 g | 7 g | 7 g | 2 g | 1 g | 0,9 g |

Figure 3- A sample of recipe profile collected from <https://www.valio.fi/>.

Table 1- Basic statistics of the crawled dataset.

| Item | Description |
|-----------------------------|-------------|
| Number recipes | 5,472 |
| Years of publication | 2010-2022 |
| Preprocessed recipes | 4833 |
| Recipes containing rating | 3197 |
| Recipes containing comments | 3060 |

2.2. Feature Engineering

2.2.1. Feature Selection Scenarios

There are numerous possibilities for generating features to facilitate the prediction of food and recipe popularity. In their research on food choices, Scheibehenne et al. [13] highlight the diverse array of factors that influence our decisions, including taste, texture, nutritional content, physical environment, attitudes, motives, individual preferences, and information. Consequently, several predictive features have been employed, drawing from insights in cognitive psychology, features utilized by prior researchers (such as Elsweiler et al. [14] or Rokicki, Herder, and Trattner [15]), as well as knowledge gleaned from the analysis of popularity. This thesis capitalizes on these earlier efforts and defines three scenarios encompassing various features designed to capture the popularity dynamics of online recipes. *Scenario 1* incorporates nutritional features, *Scenario 2* emphasizes more features related to preparation difficulty, and *Scenario 3* introduces additional user related features beyond those in the first two scenarios. More specifically:

- In Scenario 1, we primarily focus on the nutritional features of the recipes. We extract the nutritional content of each recipe, provided in terms of numerical distribution of the main nutrient components per 100 grams. This scenario considers seven nutrition components per 100 grams, which include: Energy, Proteins, Carbohydrates, Fat, Saturated Fats, Dietary Fiber, and Salt. These values are obtained from the recipe webpage and used as features in our analysis.
- In Scenario 2, we focus on preparation-related features. Preparation Time indicates the average time required to make a specific food item. Additionally, the Difficulty Level assesses how challenging it is to prepare the given food. This attribute has three different values: 1, 2, and 3, where 1 represents the simplest recipes, and 3 signifies the most complex recipes in terms of food preparation complexity. The ingredients in a recipe are the fundamental substances used to create a particular food item. Moreover, the website provides the amount of each required ingredient for each recipe (Number of Ingredient), along with a preparation guide. This scenario also incorporates the Number of Steps involved in cooking each recipe as presented on the website.

- Scenario 3 goes further by considering user comments, which enable reviews with user perspectives added to a food's content. This provides additional information for other users, potentially enhancing the advisory process. In total, 14,081 ratings and 24,630 comments were gathered for all foods. Therefore, this scenario includes additionally the total number of ratings (Rating Count) for each food, the total number of comments (Comment Count) on each recipe, the sentiment of comments, and the number of tags for each food.

2.2.2. *Sentiment Analysis of Comments*

Sentiment analysis, also known as opinion mining, is a branch of Natural Language Processing (NLP) dedicated to identifying and understanding the emotions conveyed in a sequence of words. It aims to decode the underlying attitudes, sentiments, and emotions. This approach is pivotal in social media analysis, where countless pieces of user-generated content, from tweets to status updates, offer a rich reservoir of public opinions on various subjects. By utilizing sentiment analysis, entities like businesses and researchers can capture public perceptions—whether positive, negative, or neutral—on products, events, campaigns, and trending topics. Such insights grant a profound grasp of consumer patterns, preferences, and market tendencies that can be challenging to glean from raw data alone.

The contemporary digital landscape is defined by the widespread presence of social networks. Here, users aren't mere recipients of content but also its creators. The enormous amounts of daily-generated unstructured data on these platforms are teeming with insights about user inclinations, preferences, and societal interactions. Social media mining, which involves techniques to extract and interpret this data, leans heavily on sentiment analysis for classifying and understanding user feedback. Recognizing the emotions behind user content enables companies to refine their marketing approaches, respond to customer feedback efficiently, and forecast upcoming market shifts. Furthermore, sentiment analysis is instrumental in pinpointing potential brand champions or influencers, evaluating promotional campaign success, and even forecasting stock market trends based on public sentiment. Fundamentally, it creates a bridge connecting the vast data streams on social media to practical intelligence for businesses and researchers.

To gauge the sentiment expressed in recipe reviews, we employed *SentiStrength* [16]. Sentiment Strength is a pivotal and popular concept within the domain of sentiment analysis, offering nuanced insights [17]–[19]. In the realm of natural language processing and sentiment analysis, Sentiment Strength refers to the measurement of the intensity or magnitude of sentiment expressed within textual data. Unlike simple sentiment classification, which categorizes text into broad categories like positive, negative, or neutral, Sentiment Strength provides a fine-grained evaluation of sentiment intensity, allowing for a deeper understanding of the emotional nuances conveyed in textual content. Sentiment Strength analysis involves the assignment of numerical values to sentiment expressions, typically on a continuous

scale, that capture the degree of sentiment polarity. This scale enables the differentiation between sentiments that vary in intensity, ranging from mildly positive or negative to strongly positive or negative. Crucially, Sentiment Strength assessment accounts for various linguistic elements that influence the perception of sentiment, such as degree modifiers like adverbs and adjectives. These modifiers can significantly impact the strength of sentiment expressions. For instance, "very happy" conveys a stronger positive sentiment than "somewhat happy." Incorporating context is another fundamental aspect of Sentiment Strength analysis. The interpretation of sentiment often depends on the surrounding context. A sentiment expression may have different strengths in different contexts. For instance, the sentiment associated with the term "good" may vary in strength when describing a product's quality versus evaluating the weather. Sentiment Strength analysis holds considerable applicability in diverse research domains. It is instrumental in gauging the emotional impact of textual data in fields ranging from customer feedback analysis to social media sentiment tracking.

2.2.3. Feature Importance by SHAP

Feature importance illuminates the significance of each input variable in a machine learning model concerning its impact on the target outcome. In predictive modeling, features do not uniformly influence the model's accuracy or performance. Some variables may be intrinsically linked to the target, thus essential for reliable predictions, while others may be superfluous or even harmful, inducing noise and hindering model efficacy. By assessing the relevance of each feature, data scientists can unravel the intricate data patterns and correlations, thereby refining the model to retain only indispensable features. This refinement often yields models that are more streamlined, efficient, and interpretable — a crucial aspect in applications where the rationale behind a prediction is as vital as the prediction itself.

Feature importance holds paramount significance in machine learning and data analytics for several reasons. First, it fosters transparency, shedding light on the model's decision-making process. This clarity is indispensable in sectors like healthcare, finance, and law, where the ramifications of model decisions can be profound. Second, it facilitates model streamlining; by excluding inconsequential or extraneous features, we can craft models that are both computationally efficient and robust against overfitting. Moreover, delving into feature importance can unearth surprising insights and data relationships, offering domain experts and decision-makers a treasure trove of actionable intelligence based on the model's identified key factors. In a nutshell, feature importance seamlessly connects raw data with insightful interpretation, transforming predictive models from inscrutable black boxes into enlightening instruments for judicious decision-making.

SHAP (SHapley Additive exPlanations) [20] is a widely used method to explain feature importance in machine learning models [21]–[24]. It provides insights into why a model made a particular prediction for a given instance. In our research, we used SHAP to calculate the importance of each feature in predicting the Rate Groups of recipes. SHAP values are based on cooperative game theory and provide a way to distribute the contribution of each feature to the prediction. In the context of our work,

SHAP allows us to answer questions like, "How much did each feature influence a recipe being classified as 'Good,' 'Normal,' or 'Bad'?"

The Base Value (Φ_0) represents the expected model output when no features are considered. In other words, it is the prediction that the model would make if it had no information about any features. SHAP values assign an importance score to each feature, denoted as Φ_i , which indicates the impact of the i -th feature on the model's prediction. These feature importance scores sum up the difference between the model's prediction and the base value. For a specific prediction x , the SHAP value for feature i ($\Phi_i(x)$) can be calculated as the difference between the model's prediction for the input x and the expected prediction (base value) when feature i is not included.

$$\Phi_i(x) = f(x) - \Phi_0 - \Sigma(\Phi_i'(x)), \quad (1)$$

where $\Phi_i(x)$ is the SHAP value of feature i in the absence of feature i , and $f(x)$ refers to the prediction made by a machine learning model for a specific input data point.

The SHAP values for feature i in the entire dataset X are often summarized using statistics like mean or median to understand the average impact of that feature across all predictions:

$$\Phi_i(X) = \Sigma(\Phi_i(x))/n, \quad (2)$$

where n is the number of predictions in the dataset.

The key idea behind SHAP values is to consider all possible feature combinations and their contributions to the model's output, thus providing a comprehensive explanation of why a particular prediction was made. These values can be used to gain insights into feature importance, understand model behavior, and interpret individual predictions.

2.3. Recipe Ratings Annotation

We categorized recipes into different Rate Groups based on their average ratings. These groups help in understanding the quality of recipes as perceived by users. The Rate Groups were defined as follows:

- ✓ Class 'bad': Recipes with ratings below 1.
- ✓ Class 'normal': Recipes with ratings between 1 and 3.5.
- ✓ Class 'good': Recipes with ratings over 3.5.

2.4. Classifiers

To address our research inquiries, we assessed several classifiers (explained in more detail below), determining their accuracy. We then chose the most precise one as our baseline classifier to fulfill the research objectives. Section 2.5.2 elaborates on the evaluation metrics used to identify the baseline classifiers. Additionally, we utilized feature engineering techniques in constructing the optimal baseline classifier, with further information on the feature engineering procedures provided in Section 2.5.

2.4.1. Logistic Regression

Logistic Regression [25] is a fundamental statistical technique widely used in various fields, including machine learning, epidemiology, and social sciences, for binary classification tasks [26]–[29]. The core idea behind Logistic Regression is to model the probability that an input data point belongs to one of classes. Unlike linear regression, which aims to predict continuous numerical values, Logistic Regression employs the logistic or sigmoid function to map a linear combination of input features to a probability value between 0 and 1. This logistic curve, characterized by its distinctive S-shaped form, serves as the foundation for the model. The logistic function's output represents the estimated probability of the positive class (usually denoted as "1" in binary classification) given the input features. The logistic regression model calculates a weighted sum of the input features, known as the linear combination, and applies the logistic function to this combination to produce the probability estimate. The model is trained using a dataset with known outcomes, and its parameters (weights and intercept) are optimized to maximize the likelihood of the observed data. One of the notable advantages of Logistic Regression is its interpretability. We can readily analyze the impact of each input feature on the probability of belonging to the positive class by examining the coefficients assigned to them. Additionally, Logistic Regression provides insights into the odds ratio, which quantifies how much the odds of the positive outcome change for a one-unit increase in a particular feature, making it particularly useful for explanatory modeling and hypothesis testing.

In logistic regression, "C" represents the regularization parameter, also known as the "inverse of regularization strength." It is a hyperparameter that helps control the trade-off between fitting the model to the training data as closely as possible and preventing overfitting.

2.4.2. Multilayer Perceptron

A Feed Forward Multilayer Perceptron (MLP) [30] is a type of artificial neural network widely used in the context of social media analysis [31]–[34]. It consists of an input layer, one or more hidden layers, and an output layer. The activation function, commonly ReLU (Rectified Linear Unit), introduces non-linearity to the model. MLPs

are known for their ability to capture complex relationships in data. In our case, it is used for multiclass classification to predict the Rate Groups. Cross-Entropy Loss, also known as log loss, is a loss function used to measure the dissimilarity between predicted class probabilities and actual class labels. It is often used in classification tasks, and the goal is to minimize this loss during training. Cross-entropy loss is suitable for multiclass classification, making it an appropriate choice for our Rate Group classification. Adam [35] is an optimization algorithm commonly used for training neural networks. Adam adjusts learning rates for each parameter during training, which can lead to faster convergence and better performance in many cases.

2.4.3. Random Forest

Random Forest [36] is a highly versatile and effective ensemble learning technique employed in machine learning studies [37]–[40]. This technique is employed for both classification and regression tasks, making it exceptionally adaptable to various research domains. At its core, Random Forest is a robust ensemble of decision trees, and it excels in handling intricate datasets while delivering reliable and accurate predictions. The essence of Random Forest lies in its ability to mitigate the common shortcomings of individual decision trees, such as overfitting. Decision trees, as standalone models, tend to capture noise and specific patterns in the training data, which can lead to poor generalization to new, unseen data. To address this limitation, Random Forest employs a technique known as "bagging" or "Bootstrap Aggregating." Bagging involves creating multiple decision trees, each trained on a distinct subset of the data. These subsets are generated by randomly selecting samples from the original dataset, with replacement, which introduces diversity into the training process. Additionally, at each node of the decision tree, a random subset of features is considered for splitting. This further adds randomness and robustness to the model. The power of Random Forest lies in the ensemble nature of the model. By combining the predictions of multiple decision trees, it effectively reduces overfitting while improving predictive accuracy. During prediction, each tree in the ensemble provides its output (e.g., class prediction in classification tasks or a numerical value in regression tasks), and the final prediction is determined through a majority vote (for classification) or an average (for regression). Random Forest's adaptability, interpretability, and ability to handle high-dimensional data and complex relationships between variables make it a valuable tool for various research tasks.

2.4.4. Support Vector Machine

Support Vector Machine (SVM) [41] is a pivotal machine learning algorithm widely used in social media data analysis studies [42]–[44]. It is a versatile and powerful tool designed for both classification and regression tasks. SVM's strength lies in its effectiveness in scenarios where data is not linearly separable, making it a valuable asset for addressing complex research questions. At its core, SVM endeavors to find an optimal hyperplane—a multidimensional decision boundary—within the feature space that segregates data points into distinct classes. What sets SVM apart is its quest

to find the hyperplane that maximizes the margin, which represents the distance between the decision boundary and the closest data points from each class, known as "support vectors." This approach instills confidence in classification by aiming for a broader margin, as it is less susceptible to the influence of noisy or outlier data points. In situations where linear separation is not feasible, SVM introduces the "kernel trick." This ingenious mathematical technique empowers SVM to transform data into higher-dimensional spaces where linear separation becomes viable. Popular kernel functions, including the linear, polynomial, and radial basis function (RBF) kernels, enable SVM to tackle a wide spectrum of nonlinear problems inherent to many research domains. This method serves as a robust and adaptable machine learning tool, capable of handling high-dimensional data and intricate relationships between variables. Its ability to accommodate both binary and multiclass classification problems make it applicable to various research contexts. Moreover, SVM's regularization parameter (C) offers the flexibility to fine-tune the balance between margin maximization and error minimization, ensuring that the model aligns with specific research objectives.

2.4.5. Gradient Boosting

Gradient Boosting [45] is a powerful ensemble machine learning technique. It is renowned for its exceptional predictive capabilities and versatility in handling a wide range of data-driven challenges [46]–[48]. Gradient Boosting is particularly well-suited for regression and classification tasks, making it a valuable tool for extracting insights and generating accurate predictions. At its core, Gradient Boosting is an ensemble method that assembles multiple weak learners, typically decision trees, into a strong predictive model. The key innovation behind Gradient Boosting is the sequential nature of its training process. It builds a series of decision trees iteratively, each one focusing on the errors or residuals of the previous tree. The process begins with the creation of a simple decision tree, often referred to as a "shallow tree" or a "stump." This initial tree is used to make predictions, which are then compared to the actual target values. The differences between these predictions and the true values represent the errors or residuals. Subsequent decision trees are constructed to specifically target these errors. Each new tree is designed to correct the mistakes made by the ensemble of trees built so far. It places higher importance on the data points that were previously misclassified or for which predictions were the farthest from the actual values. As more and more trees are added to the ensemble, Gradient Boosting adapts and refines its predictions, gradually reducing the errors and improving accuracy. The final prediction is achieved by combining the outputs of all the individual trees, with each tree having a weighted say in the final result. One of the notable advantages of Gradient Boosting is its robustness to various data types and its ability to handle both regression and classification tasks. Additionally, it provides insights into feature importance.

2.5. Model Selection

Grid search [49] is a technique used to systematically search for the best combination of hyperparameters for a machine learning model. Hyperparameters are parameters that are set before training a model and cannot be learned from the data. Examples include learning rates, the number of hidden layers in a neural network, or the depth of a decision tree.

The grid search process involves defining a range of possible values for each hyperparameter and then evaluating the model's performance for all combinations of these values. This allows us to identify the set of hyperparameters that yield the best results based on a predefined evaluation metric, such as accuracy (Acc.), F1-score, or another appropriate measure for our specific task. Hyperparameters tuned in grid search model selection scheme are shown in Table 2.

2.5.1. Cross-Validation

We divided our dataset into training and validation subsets. For each hyperparameter combination, we used cross-validation to train the model on the training data and evaluate its performance on the validation data. Cross-validation helps prevent overfitting and provides a more robust estimate of model performance. Then, we identified the combination of hyperparameters that yields the highest performance based on the evaluation metric. This combination is considered the "best" for our classification task. Grid search helps us find hyperparameters that lead to the best model performance, ensuring that our model is well-tuned for our specific problem. It systematically explores the hyperparameter space, ensuring that no promising configurations are overlooked. By using cross-validation, grid search helps prevent overfitting by assessing model performance on unseen data. In our research, we applied grid search as part of our model selection process to determine the best classifier among several options. By employing grid search, we ensured that the chosen classification algorithm was optimized for the specific problem we were addressing, leading to more accurate and reliable results in our research. Combination of grid search and cross validation is popular among scholars to train classifiers and derive the best model [50]–[52].

Table 2- Hyperparameters for grid search model selection scheme.

| Method | Hyperparameters | Range of change |
|---------------------|--|---------------------------------|
| Logistic Regression | C | [1.0, 0.5, 0.1] |
| | Regularization | [Ridge (L1), Lasso (L2)] |
| MLP | Learning rate | [0.001, 0.01, 0.1] |
| | Number of hidden layers | [1, 2, 3] |
| | Number of neurons in each hidden layer | [64, 128, 256] |
| Random Forest | Min samples leaf | [1, 2, 3, 4, 5, 6] |
| | Max depth | [1, 2, 3, 4, 5, 6] |
| | Min samples split | [2, 3, 4, 5, 6] |
| SVM | Kernel | [Linear, Radial Basis Function] |
| | C | [1, 2, 3, 4, 5, 6] |
| Gradient Boosting | Learning rate | [0.1, 0.2, 0.3] |
| | Max depth | [1, 2, 3, 4, 5, 6] |
| | Min child weight | [3, 4, 5, 6] |
| | Subsample | [1.0, 0.5, 0.1] |
| | n estimators | [50,100,150] |

2.5.2. Evaluation Metrics

We used an appropriate evaluation metric as accuracy (Acc.) and F1-score to assess the model's performance for each combination of hyperparameters. Both F1 score and Acc. are common metrics used to evaluate the performance of classification models, but they capture different aspects of model performance.

2.5.2.1. Accuracy

Accuracy (Acc.) is a straightforward metric that measures the overall correctness of a classification model. It is calculated as the ratio of correctly predicted instances (both true positives and true negatives) to the total number of instances in the dataset.

$$Acc. = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}}, \quad (3)$$

This metric is easy to understand and interpret, but it may not always be the most appropriate metric, especially when dealing with imbalanced datasets. In cases where one class dominates the dataset, a high accuracy score can be misleading, as the model may simply predict the majority class most of the time and perform poorly on the minority class.

2.5.2.2. F1 Score

The F1 score is a metric that balances both precision and recall. Precision measures how many of the positive predictions made by the model were actually correct, while recall measures how many of the actual positive instances were correctly predicted by

the model. The F1 score combines these two metrics to provide a single value that summarizes the model's performance.

$$F1 = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{(\text{Precision} + \text{Recall})}, \quad (4)$$

where:

$$\text{Precision} = \frac{\text{True Positives}}{(\text{True Positives} + \text{False Positives})}, \quad (5)$$

and

$$\text{Recall} = \frac{\text{True Positives}}{(\text{True Positives} + \text{False Negative})}. \quad (6)$$

The F1 score is particularly useful when dealing with imbalanced datasets because it considers both false positives and false negatives, making it a more robust metric in such scenarios. A higher F1 score indicates better model performance in terms of balancing precision and recall.

2.6. Evaluation of Recipe Healthiness

We employed the UK Food Standards Agency's "traffic light" system (Figure 4) to calculate the healthiness of recipes, which is referred to as the FSA score. This score was determined by evaluating macronutrients: sugar, fat, saturated fat, and salt content. The score spans a spectrum from green (indicating healthiness) to red (indicating unhealthiness). To put it simply, a lower FSA score signifies greater nutritional health, whereas higher scores are indicative of pronounced unhealthiness. This factor allows us to understand how users' preferences align with the perceived healthiness of the recipes. It provides insights into whether healthier recipes receive higher ratings.

| | Fat | Saturates | Sugars | Salt |
|------------------------|---|---|---|---|
| What is HIGH? | Over 17.5g / 100g Or more than 21g / portion | Over 5g / 100g Or more than 6g / portion | Over 22.5g / 100g Or more than 27g / portion | Over 1.5g / 100g Or more than 1.8g / portion |
| What is MEDIUM? | Between 3g / 100g and 17.5g / 100g | Between 1.5g / 100g and 5g / 100g | Between 5g / 100g and 22.5g / 100g | Between 0.3g / 100g and 1.5g / 100g |
| What is LOW? | 3g / 100g or less | 1.5g / 100g or less | 5g / 100g or less | 0.3g / 100g or less |

ref: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8102049/figure/F1/>

Figure 4- FSA health score classes.

3. RESULT AND DISCUSSION

This section is the culmination of our research journey, where we present and analyze the empirical findings that address the core questions driving this study. Divided into five distinct sub-sections, each dedicated to answering a specific research question, we delve into the data-driven insights that illuminate our research objectives. These findings not only provide clarity on the matters at hand but also contribute to the broader discourse within popularity prediction of recipes. In the following pages, we navigate through our results, unravel their implications, and offer a comprehensive discussion that underscores the significance of our research endeavors.

3.1. Insight into Data

The primary objective of this research revolves around the prediction of a recipe's Rating Group through the utilization of various features. The Rating Group is determined based on user-provided ratings. Consequently, a stronger correlation between these additional features and the rating can offer valuable insights into the factors influencing it. As depicted in Figure 5, the features most strongly correlated (approximately 30%) with the Rating include the sentiment of comments, the total number of comments, and the total number of rates, all of which exhibit a positive correlation. This suggests that as the sentiment of comments, the total number of comments, or the total number of rates increases, the recipe's ratings are expected to increase in a linear fashion. Additionally, the number of steps involved in cooking the recipe and the number of ingredients used display a relatively high correlation (approximately 20%). The highest correlation observed (approximately 90%) is between the fat content and energy of a recipe, indicating that recipes with higher fat content are expected to possess greater energy levels. Furthermore, the number of steps, the number of ingredients, and the recipe's difficulty level are highly correlated (approximately 45%). For more detailed information, please refer to Figure 5.

3.2. Classification Results (RQ1)

3.2.1. *Recipes year of publication*

Most of the recipes available on the Valio website were posted prior to 2014, as depicted in Figure 6a. Also, Figure 6b reveals that, on average, recipe ratings have been on the rise in recent years. However, the year 2022 stands out as an exception. This is due to the fact that the median rating for recipes published in 2022 is zero. Consequently, we decided to exclude the year 2022 from our modeling efforts because it didn't have sufficient time to garner attention or receive ratings. Therefore, for the purposes of this research, we focused on recipes that were published between 2010 and 2021.

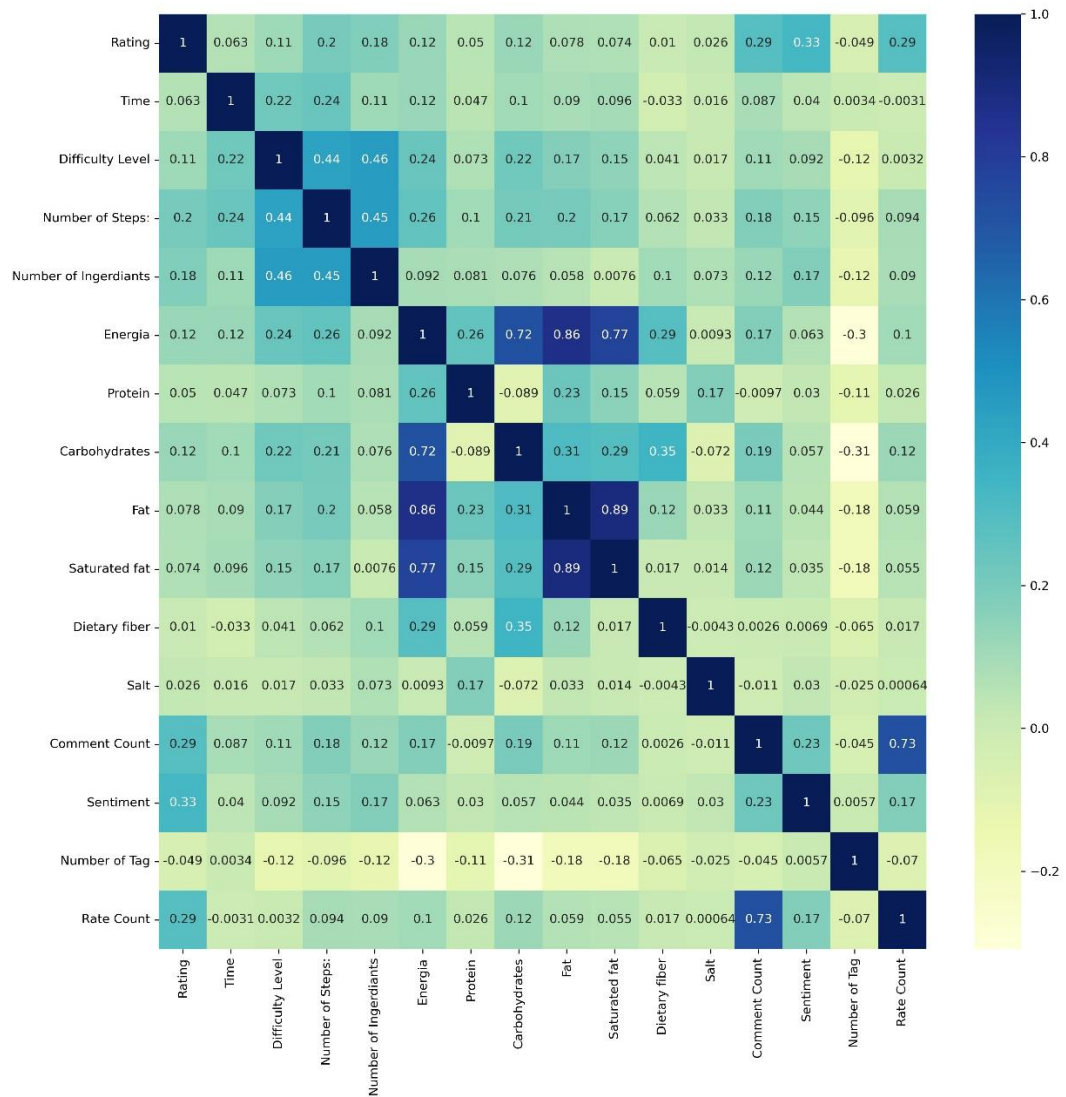


Figure 5. Correlation heat map of the retrieved dataset from Valio website.

By considering the average user ratings for each recipe, which fall on a continuous scale ranging from 0 to 5, we established (as explained in Section 2.3) what we call the "Rating Group." Figure 7 illustrates this, revealing that the majority of recipes belong to the "good" Rating Group, with ratings exceeding 3.5, making up 55% of the total. Meanwhile, approximately 30% of the recipes fall into the "bad" Rating Group. In the subsequent section, we delve into the classification outcomes that are derived from these three defined rating classes.

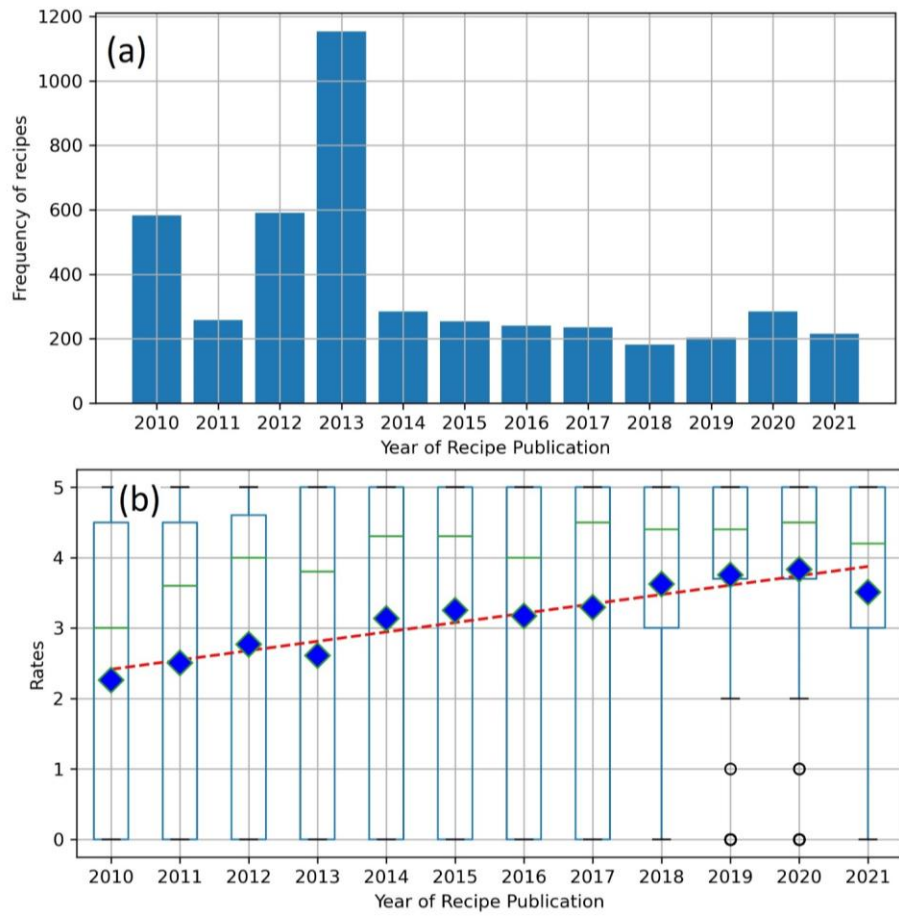


Figure 6- a) Frequency of recipes based on their year of publication, and b) boxplot of given rates to recipes based on the recipes' year of publication.

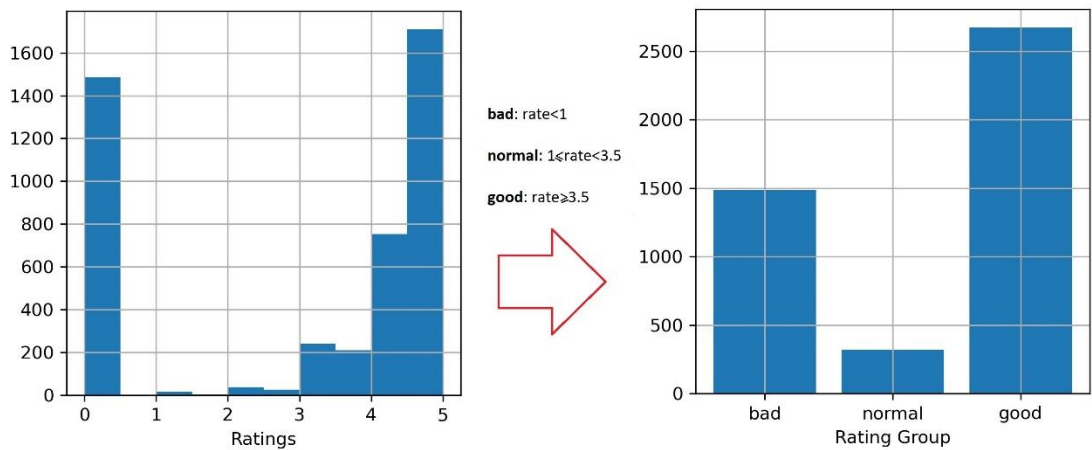


Figure 7- The defined classes for each recipe based on given rates by users.

3.2.2. Model Selection

The evaluation of models in this study was conducted based on a specific set of hyperparameters, as outlined in Section 2.5 (Table 2), and across various scenarios, as discussed in Section 2.2.1. The results of these evaluations, in terms of performance metrics, are presented in Table 3. Remarkably, in all combinations tested, Scenario 3 consistently emerged as the top-performing scenario, as evidenced by its high accuracy (Acc.) and F1 score (shown in Figure 8).

Furthermore, among the classifiers examined, Logistic Regression, Random Forest, and Gradient Boosting stood out as the most effective ones (Acc. and F1 around 90%). Given these results, we opted to proceed with these three top classifiers using the hyperparameters specified in Table 3 for more in-depth investigations in the upcoming sections.

Table 3- Best hyperparameters detected by grid search.

| Method | Hyperparameters | Best option | Best Scenario | F1 | Acc. |
|---------------------|-------------------------|-----------------------|---------------|------|------|
| Logistic Regression | C | 1 | 3 | 0.90 | 0.93 |
| | Regularization | L1 | | | |
| MLP | Learning rate | 0.001 | 3 | 0.69 | 0.72 |
| | Number of hidden layers | 3 | | | |
| | Number of neurons | 128 | | | |
| RandomForest | Min samples leaf | 5 | 3 | 0.90 | 0.93 |
| | Max depth | 5 | | | |
| | Min samples split | 4 | | | |
| SVM | Kernel | Radial Basis Function | 3 | 0.87 | 0.90 |
| | C | 1 | | | |
| Gradient Boosting | Learning rate | 0.1 | 3 | 0.89 | 0.92 |
| | Max depth | 5 | | | |
| | Min child weight | 4 | | | |
| | Subsample | 0.1 | | | |
| | n estimators | 150 | | | |

3.2.3. Feature Selection Scenarios

The study explored three distinct scenarios, each involving different sets of features for classification. In the first scenario, only nutritional features were considered, while the second scenario included both nutritional and preparation difficulty features. Surprisingly, the first and second scenarios resulted in similar levels of classification precision, although scenario 2 exhibited a slight advantage.

However, the most accurate classification approach was observed in the third scenario (Figure 8). In this scenario, user engagement features were introduced in addition to the nutritional and preparation difficulty features, which collectively yielded the best classification scheme.

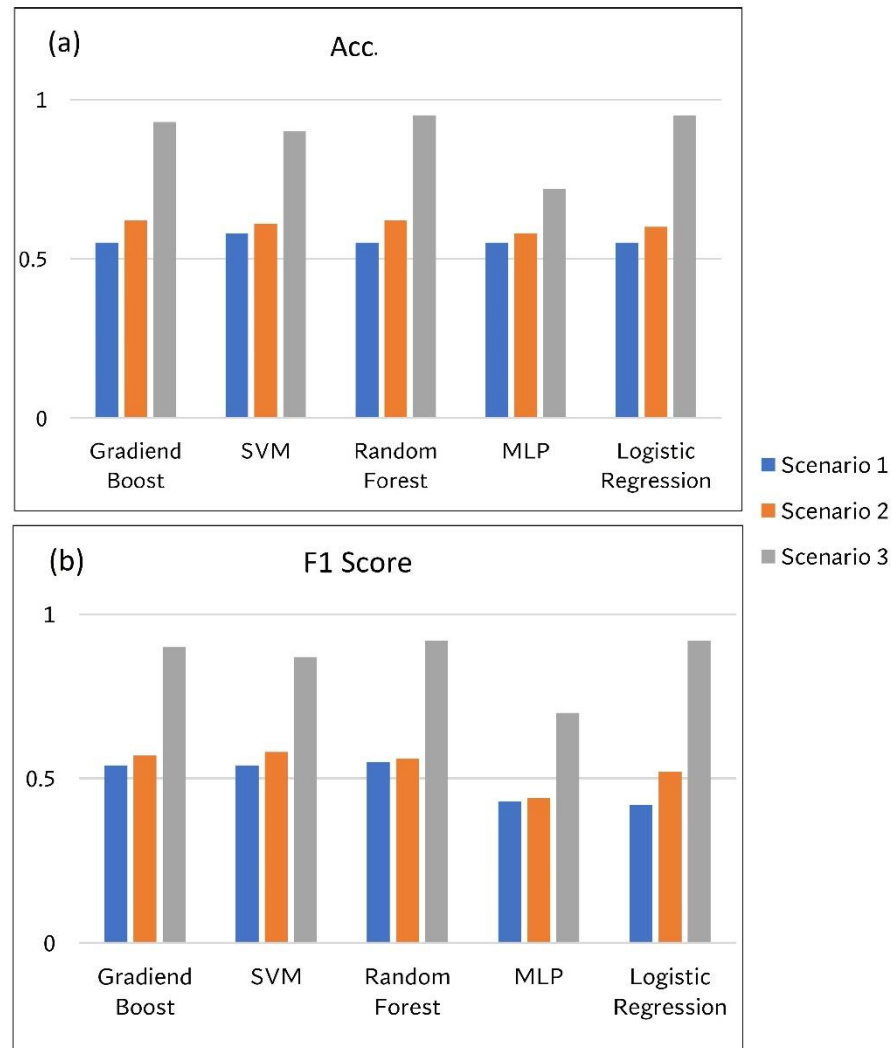


Figure 8- a) Acc., and b) F1 values for different classifiers in each scenario.

In the context of social media, user engagement features, encompassing likes, comments, shares, and reactions, are of paramount importance due to their profound influence on user experience and platform vitality [53], [54]. These features not only enhance user interaction and satisfaction but also promote content visibility and virality. They facilitate community building, fostering connections among users, and establishing a sense of belonging [55]. Moreover, user engagement metrics serve as valuable indicators for platform operators and marketers, offering insights into user preferences and content effectiveness. As a result, a comprehensive understanding and effective utilization of user engagement features are integral to the success and sustainability of social media ecosystems [56]. The significance of user engagement features is underscored in this research by defining Scenario 3, showcasing their potential to greatly enhance classification performance. As an illustration, consider the case of Logistic Regression Gradient Boosting, and Random Forest which emerged as the top-performing classifiers. Their Acc. notably improved from about 50% (in Scenario 1) and 60% (in Scenario 2) to above 90%. Similarly, the F1 score for these

classifiers exhibited a substantial increase, about 90% from a previous level of about 50%.

3.2.4. Confusion Matrix of Classifiers and Analysis of Errors

Figure 9a and b presents the confusion matrix of all studied classifiers. The visual representation of this data highlights that Random Forest and Gradient Boosting outperform MLP and SVM in terms of classification performance. Notably, Logistic Regression, as depicted in Figure 9b, emerges as the top-performing classifier in this context. Using Logistic Regression, we encountered 92 misclassified recipes, as indicated in Figure 9b. Among these misclassifications, the majority, comprising 89 samples, were erroneously labeled as 'good' by the model when they should have been categorized as 'normal.' Additionally, 3 samples were incorrectly classified as 'good' when they were actually 'bad.' It is notable that a substantial portion of these misclassified samples is concentrated in the period preceding the year 2016 (Figure 9b).

To address this issue, we opted to partition the dataset into two distinct sets based on the publication year: one set encompassing data from 2016 and another containing data from 2016 onwards. This partitioning resulted in a more uniform distribution of errors across the year of publication, although it didn't lead to significant changes in Acc. and F1 scores (Figure 10). This further underscores our earlier emphasis on the significance of the publication year as a pivotal factor, given that user preferences and classification patterns are expected to evolve over time.

The Logistic Regression model that was trained prior to 2016, as shown in Figure 10a, exhibits slightly lower Acc. and F1 scores, measuring at 90% and 85%, respectively. However, it is worth noting that the Acc. and F1 values remained consistent for the model trained after 2016, as illustrated in Figure 10b, with no discernible change.

Furthermore, when considering the mean error rates for each year, before and after 2016, there were 8 misclassified samples on average before 2016 and a slight improvement to 6 misclassified samples on average after 2016. This indicates a modest enhancement in model performance following the 2016 timeframe.

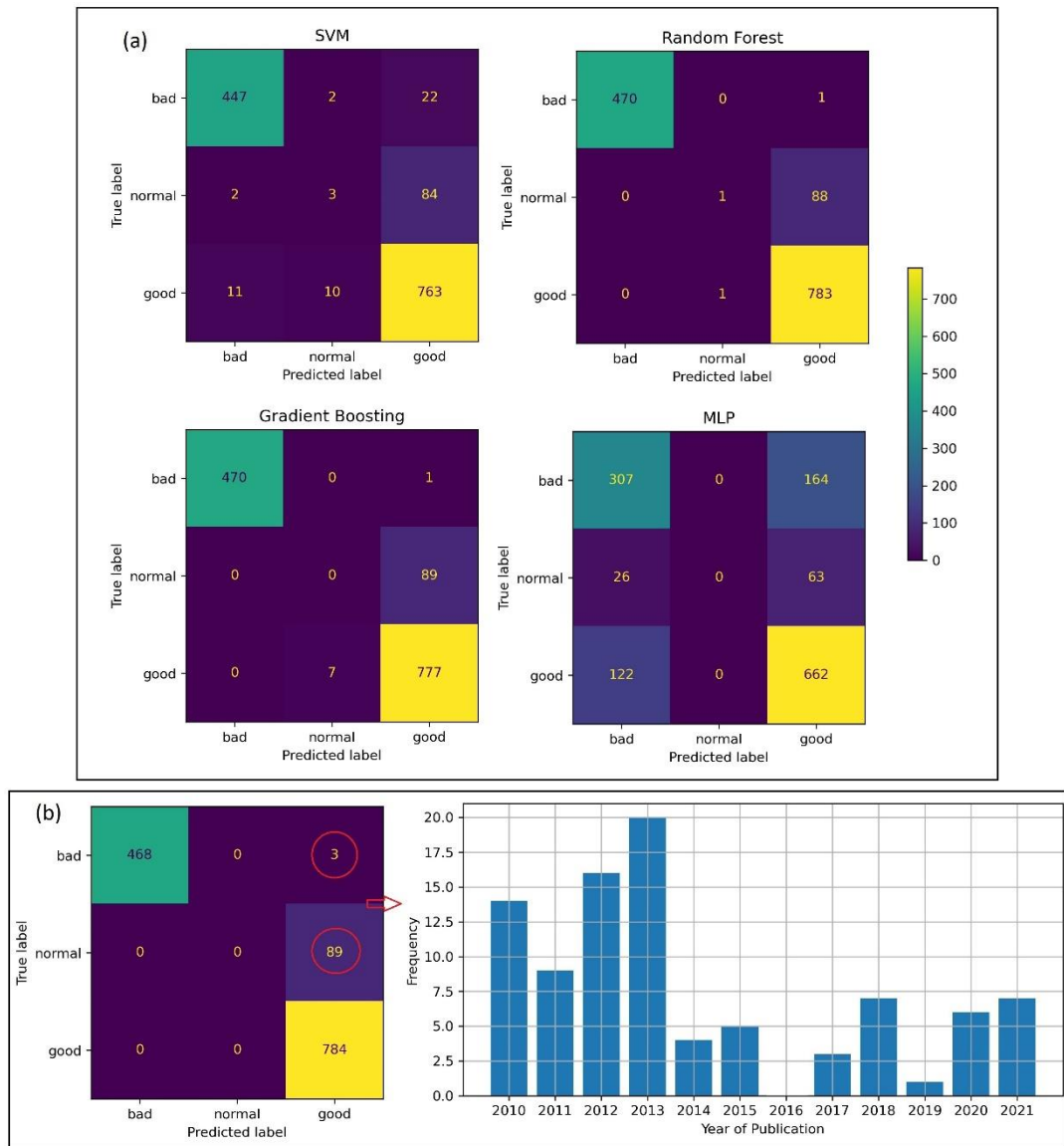


Figure 9- Based on Scenario 3 and best hyperparameters of classifiers: a) Confusion matrix of SVM, Random Forest, Gradient boost, and MLP, and b) Confusion matrix of the Logistic Regression with the distribution of errors (89+3) on the year of recipes' publication.

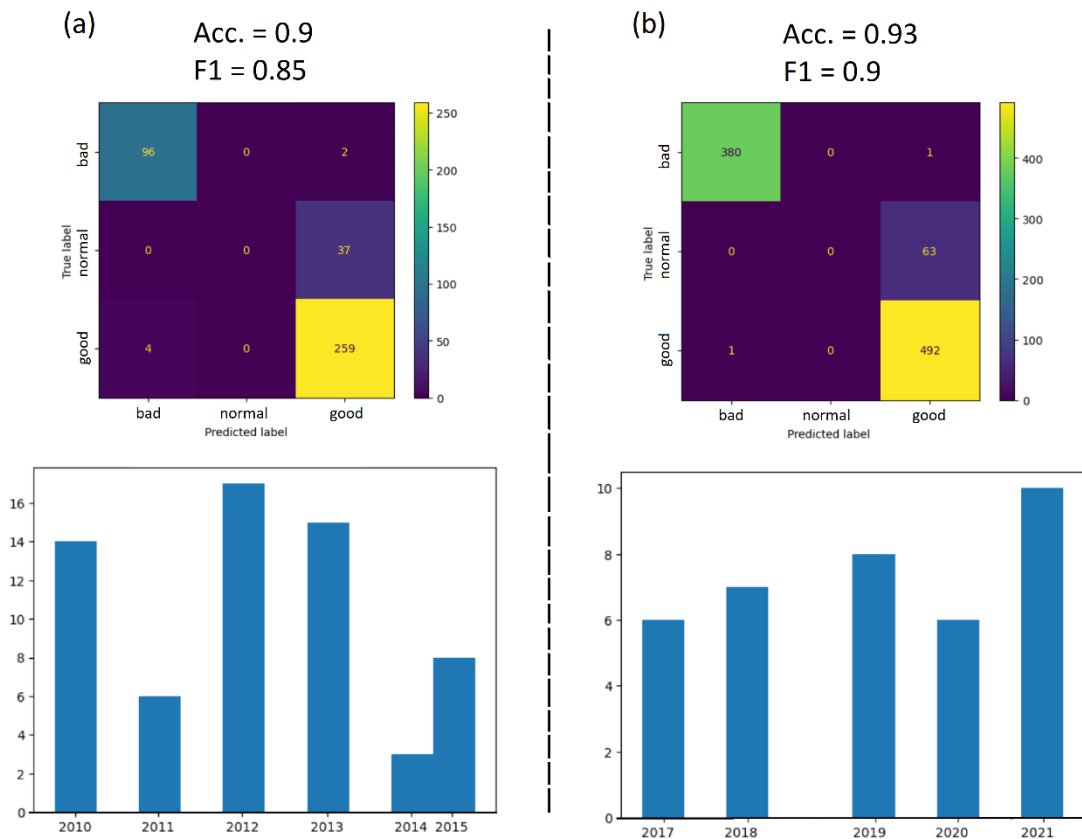


Figure 10- Results of the trained Logistic Regression for: a) before and b) after 2016.

A uniform distribution of error across independent variables is of paramount importance in classification tasks. When errors are evenly spread across variables, it signifies a balanced and reliable model performance. This uniformity suggests that the classifier is making consistent and well-informed decisions across features, ensuring that it does not disproportionately influence the classification outcome. In contrast, an imbalanced distribution of errors can lead to biased predictions, where certain values of a feature are favored over others, potentially causing misclassification and reduced overall accuracy. A uniform distribution of error not only enhances the model's robustness but also provides a more comprehensive understanding of the relationships between features and the classification outcome, ultimately contributing to more effective decision-making and better-informed insights.

The dynamics of user taste in social media platforms undergo continuous evolution over time, reflecting a multifaceted interplay of factors [57]. These temporal shifts are driven by a multitude of influences, including cultural trends, societal changes, technological advancements, and evolving user demographics. As users engage with a diverse array of content, their preferences, interests, and consumption patterns naturally adapt and transform. Social media platforms, with their constant influx of new content and interactions, act as dynamic ecosystems where users are exposed to an ever-changing landscape of information and ideas. This exposure, coupled with the network effects inherent to social media, contributes to the diffusion and adoption of emerging trends and concepts, further shaping the collective taste of the user

community. Consequently, the study of user taste variation over time in social media is not only instrumental for understanding evolving user behavior but also holds significant implications for content creators, marketers, and platform operators seeking to navigate and harness the dynamics of these online environments.

3.3. Feature Importance (RQ2)

Using Scenario 3, we employed the three best classifiers, namely Random Forest, Gradient Boosting, and Logistic Regression, to calculate the SHAP values for the features, as illustrated in Figure 11. As anticipated based on the scenarios' results, user engagement features take the lead as the most significant drivers in the classification, specifically Rating Count and Comment Count. While Rating Count remains the most crucial feature across all classifiers, the hierarchy of other features' importance can vary depending on the classifier chosen. For instance, in Logistic Regression and Random Forest, Comment Count secures the second position, whereas Gradient Boosting places Energy in the second position.

In the case of Gradient Boosting (as shown in Figure 11a), nutritional features such as Energy, Fat, and Salt hold significant importance. However, in Logistic Regression (as seen in Figure 11b), the number of Tags and the Number of Ingredients take higher ranks compared to Energy and Fat. On the other hand, Random Forest (depicted in Figure 11c) exhibits a lower degree of reliance on other features and predominantly hinges on Rating Count.

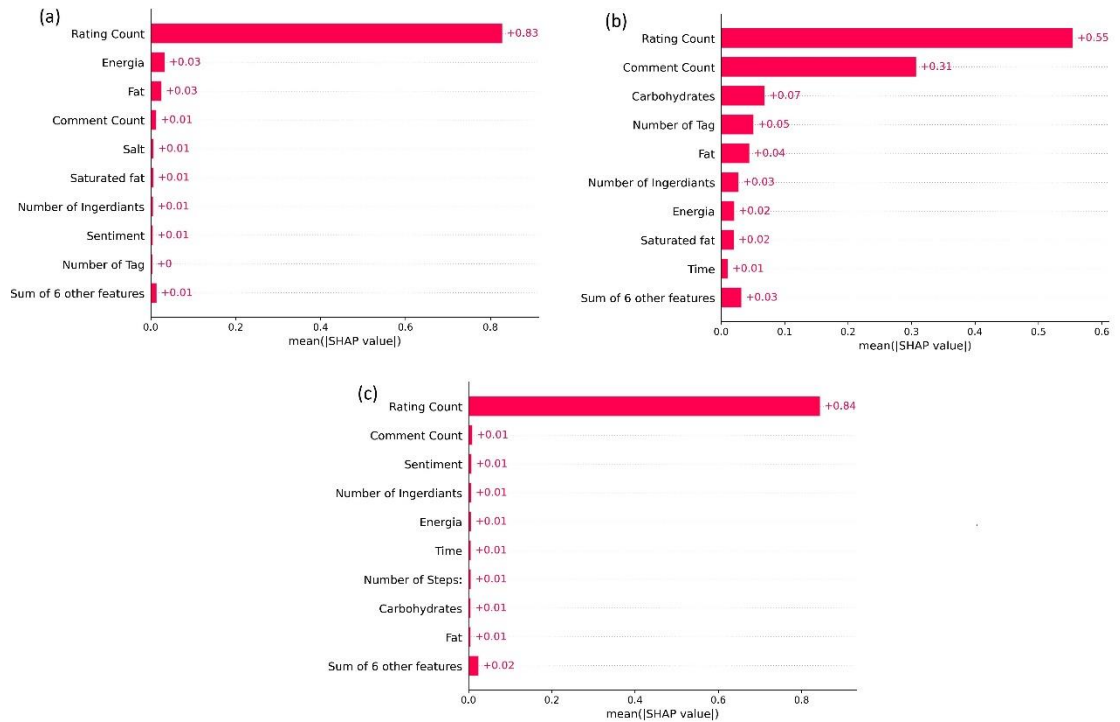


Figure 11- SHAP values for a) Gradient Boosting, b) Logistic Regression, and c) Random Forest.

To investigate the significance of other features in the absence of Rating Count and Comment Count, we deliberately excluded these two influential factors and conducted a detailed analysis of the SHAP values. Further information and a comprehensive breakdown of these findings can be found in the Appendices section, specifically in Figure S1 to Figure S6.

3.4. Comparison to Similar Studies (RQ3)

A comparative study was conducted involving Valio, a Finnish online food social network, and similar platforms from other countries: Allrecipes.com, representing American food culture, and Kochbar.de, representing German food culture. This comparison is found on Mößlang study research [1], sought to assess the predictive capability of various features and validate statistical insights through predictive modeling experiments. The outcomes of these experiments point to the existence of recipe characteristics that have broad applicability and significantly impact the future popularity of online recipes across all platforms. Nevertheless, notable distinctions in popularity patterns among these three websites were observed.

According to Mößlang study [1], which utilized recipes from Kochbar and Allrecipes, the predictive modeling experiments demonstrated that the developed features possess strong predictive capabilities. The primary objective was to predict whether a recipe would become more popular than the average within a specified time frame. To achieve this, three distinct classifiers—Random Forest, Naive Bayes, and Generalized Linear Models—were employed. The study achieved notable success, with accuracies of up to 89% observed for certain configurations. Notably, the models trained on Kochbar.de data consistently exhibited the highest accuracy values. Among the classifiers, Random Forest appeared to be the most effective, although none of them consistently outperformed the others.

These modeling results validated previous experimental assumptions, highlighting the presence of universally applicable characteristics that significantly influence the future popularity of online recipes. These factors included the upload user's prior activity, the presentation quality, and the novelty of a recipe's concept. However, the impact of user activity features, such as written and received ratings/comments or the number of uploaded recipes, was more pronounced on Kochbar.de's recipe popularity. On the other hand, innovation-related features, such as recipe novelty and ingredient popularity rank, as well as image features like saturation and image entropy, seemed to have a greater influence on the popularity of Allrecipes.com and its recipes. As previously mentioned, in the case of Valio, a Finnish platform, the Number of Rating and Number of Comment from users emerged as the most influential factors in determining a recipe's rating. Additionally, depending on the choice of classifier, nutritional factors like Energy or Fat were also found to be important drivers.

These findings underscore the significance of cultural nuances and variations in user behavior across different online food social networks. They emphasize the necessity of tailoring recommendation systems and analytical approaches to the specific characteristics and preferences of each platform.

Table 4- Comparison between this thesis and similar study.

| Item | Data | Attributes | Best Methods | Accuracy | Key findings |
|--------------------|------------|--|--|----------|---|
| This thesis | Valio | <ul style="list-style-type: none"> Nutritional Difficulty User engagement | <ul style="list-style-type: none"> ✓ Logistic Regression ✓ Random Forest | 0.93 | User engagement features improve prediction substantially. |
| Möbblang study [1] | Allrecipes | <ul style="list-style-type: none"> Nutritional Difficulty User engagement | <ul style="list-style-type: none"> ✓ Random Forest | 0.89 | Innovation-related features and image features have a greater influence on the popularity |
| | Kochbar | <ul style="list-style-type: none"> Recipe novelty Recipe image | | | User engagement features were more pronounced on popularity. |

3.5. Healthiness of Recipes (RQ4)

The recipes published on the Valio website, as depicted in Figure 12a, are predominantly not classified as very healthy. A majority of them receive an FSA score of 6, indicating a medium health rating. Interestingly, recipes with low FSA scores, below 4, are relatively scarce on the platform, suggesting that less healthy recipes are more prevalent on the website. This biased distribution of FSA scores, with an abundance of less healthy options, potentially contributes to the increased popularity of unhealthy foods.

Conversely, as illustrated in Figure 12b, there is a discernible trend in the average ratings based on FSA scores. Generally, higher FSA scores are associated with higher average ratings, with ratings expected to increase from 2 to over 3 as FSA score improves. However, there is a slight drop in the mean ratings for very unhealthy recipes (FSA = 9). Interestingly, the median rating for very healthy recipes (FSA = 3) is zero, indicating that the majority of users may not be as interested in these highly healthy options. Nevertheless, the median rating substantially increases to over 4 for recipes with higher FSA scores (6 or 7), suggesting that these moderately and low healthy recipes tend to be more appealing to users. This nuanced relationship between FSA scores and user ratings underscores the tendency of user preferences to less healthy foods.

Researchers have shown that people often have a preference for unhealthy foods due to a combination of evolutionary, psychological, and sensory factors [58], [59]. Evolutionarily, our ancestors had a survival advantage by seeking out calorie-dense foods, which are often high in fat and sugar, as these provided the energy needed for survival. This preference for calorie-rich foods has been ingrained in our biology over time. Psychologically, unhealthy foods can trigger pleasure and reward centers in the brain, releasing feel-good neurotransmitters like dopamine, which reinforce the desire for such foods. Additionally, the sensory appeal of unhealthy foods, with their often sweet, salty, or fatty flavors and pleasing textures, can make them more attractive.

While these factors contribute to the appeal of unhealthy foods, it is essential to balance these indulgences with a healthy diet.

Furthermore, advertisements have a significant effect on the promotion of unhealthy foods in social medias. A study [60] reveals that on social media platforms, young people's responses to unhealthy food advertising posts significantly outweigh their responses to both healthy and non-food-related posts. This heightened response is evident in their increased attention, better memory retention, more favorable peer assessment, and greater likelihood to share such content. Considering the extensive usage of social media among adolescents, these findings carry significant implications for the regulation of marketing practices targeted at this demographic. Currently, many countries restrict advertising to children up to the age of 13, and some extend this limit to 15, as seen in Ireland and the UK.

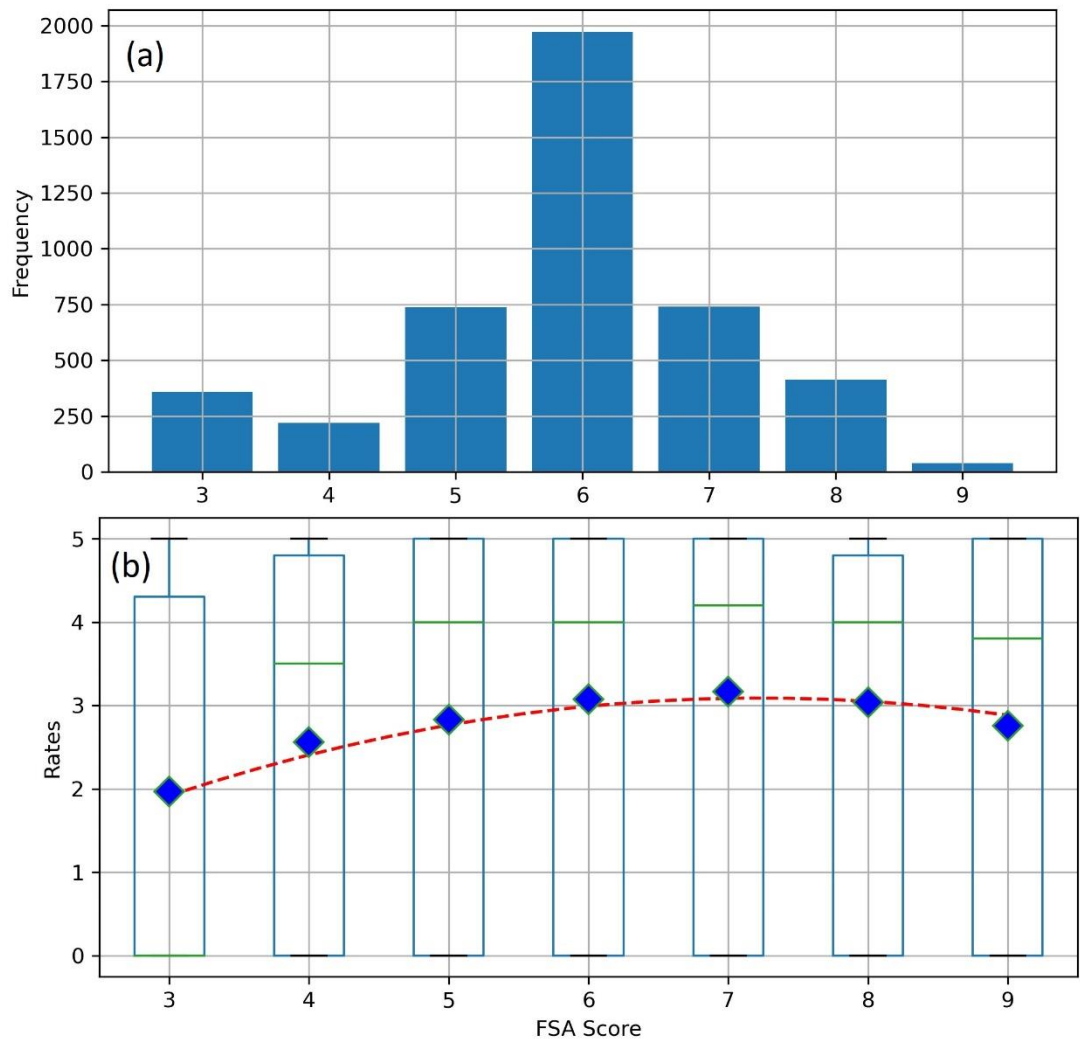


Figure 12- a) The distribution of FSA Score for recipes published in the Valio website, and b) the relation between given rates by users to recipes and the recipes' FSA score.

4. CONCLUSION

This chapter serves as the conclusion of the thesis, offering a concise summary of the findings and a discussion of potential limitations in the approach. Additionally, it provides an outlook on opportunities for improvement and future studies. The core objective of this master's thesis revolves around providing enhanced insights into the concealed patterns and mechanisms governing the popularity of online recipes. A comprehensive understanding of these sociodynamic processes has the potential to facilitate the development of advanced, health-conscious recommender systems. Such systems hold the promise of addressing the escalating food-related health challenges prevalent in contemporary societies. The approach employed in this research entailed a statistical analysis of datasets sourced from Valio website. This platform represents distinct cooking culture in Finland. Through comparative analysis, a more generalized understanding of the processes governing recipe popularity emerged. These analyses were grounded in recipe characteristics (nutritional and preparation complexity) as well as user engagement features (e.g., rates, and comments). To assess the predictive capacity of these features and validate the statistical findings, predictive modeling experiments were conducted. The key findings of this master's thesis can be summarized as follows:

- RQ1) Through the application of various classifiers, including SVM, MLP, Random Forest, Logistic Regression, and Gradient Boosting, alongside a grid search model selection approach, we have effectively identified trends in popularity within the Finnish online food social network. Among these classifiers, Logistic Regression stands out with a notable accuracy of 0.93 and an impressive F1 score of 0.9 during testing. This high level of accuracy achieved using machine learning techniques demonstrates that recipe popularity can be reliably determined, underscoring the utility of such approaches for trend analysis within this online food social network.
- RQ2) The primary factors that influence recipe ratings and popularity, as indicated by our findings, revolve around user engagement and interaction with the recipes. Notably, recipes that receive more attention in terms of the number of comments and ratings tend to be more popular (get higher rates). Therefore, the key influencing factors on the popularity of each recipe are user engagement features, specifically number of comments, ratings, and sentiment of comments. The sentiment or tone of the comments left by users can impact a recipe's popularity. These findings highlight the pivotal role of user engagement and feedback in determining the popularity and ratings of online recipes, underlining the importance of community interaction in the context of food-related content.
- RQ3) Valio, as a Finnish online food social network, was compared to similar platforms in other nations, specifically Allrecipes.com (representing American food culture) and Kochbar.de (representing German food culture). For Kochbar, user activity features, such as written and obtained ratings/comments, as well as the number of recipes uploaded, show reliable results in predicting recipe popularity. In contrast, Allrecipes demonstrates a greater influence of

innovation features, including recipe innovation and ingredient popularity, as well as visual features like image saturation and entropy, on the popularity of its recipes. In the case of Valio, the number of ratings and comments received by users emerges as the most influential factor in determining a recipe's rating.

- RQ4) The health quality of the recipes on the Valio website generally falls within the medium range based on the FSA score. Interestingly, our analysis indicates that users' average ratings tend to be higher for recipes that are categorized as less healthy. This finding suggests a potential unhealthy tendency of Finnish users' taste, highlighting an intriguing area for further investigation.

4.1. Limitations and Future Works

Recognizing certain constraints arising from the varying approaches and paradigms within the examined online recipe communities is crucial. The findings presented in this context are specific to the datasets available from the Valio website. Therefore, further investigations are imperative to ascertain the broader applicability of these findings. The chosen features for predicting recipe popularity align with established best practices in the field and are anticipated to generate dependable results. Another limitation arises from the temporal aspect, specifically concerning recipe publication dates and received ratings. Notably, the most accurate classifiers tend to exhibit lower accuracy between 2010 and 2014, possibly due to evolving user preferences over the past decade in Finland.

Potential future applications encompass the development of health-conscious recommendation systems or tools aiding users in recipe uploads. Such systems could provide recommendations for adjusting recipe attributes to maximize their future popularity. Moreover, the inclusion of visual features derived from the images associated with each recipe could enhance the classification process. Additionally, large language models can be harnessed to extract valuable insights from user comments on each recipe, beyond the current utilization of sentiment score averages.

Moreover, Explainable prediction models can be considered as another research direction for the future works. Explainable AI (XAI) has emerged as an indispensable tool in the realm of recipe popularity prediction within Finnish social media. In a space where cultural nuances and evolving food trends significantly influence users' preferences, understanding the 'why' behind a prediction is as vital as the prediction itself. With XAI, data scientists and researchers can decipher the intricate features- from ingredients, preparation techniques, to seasonality- that drive a recipe's virality. This not only enables content creators to tailor their offerings more effectively but also instills trust among users, as they gain transparency into how certain recipes gain traction. In the dynamic landscape of Finnish food social media, the clarity provided by XAI stands as a beacon for stakeholders at every level. XAI demystifies complex algorithms and sheds light on the intricate relationships between various data points. By offering transparent and interpretable predictions, it empowers content creators, influencers, and marketers to fine-tune their strategies, ensuring their offerings

resonate with the Finnish palate. As consumers become more discerning and curious about the content they consume, the ability of XAI to provide clear rationales behind recipe popularity predictions further enhances credibility and trust. Moreover, it facilitates proactive adjustments in response to real-time feedback, fostering a more responsive and interactive digital ecosystem. In essence, as Finnish culinary trends evolve, Explainable AI serves as both a compass and a map, guiding stakeholders through the intricate maze of audience preferences and ensuring continued relevance in a competitive space.

Furthermore, an analysis of recipe tags (Table 5), of which 171 unique tags were identified among the retrieved recipes on the Valio website, offers a promising avenue. Tags in social media posts serve as pivotal elements that play a multifaceted role in shaping the online landscape. These small, but impactful, textual markers, often referred to as hashtags or simply tags, serve several crucial purposes. Firstly, they enhance content discoverability by categorizing posts into relevant topics or themes, facilitating the process of content search and exploration for users. Tags also foster engagement and participation by enabling users to join conversations and communities centered around shared interests. Moreover, they amplify the reach of posts by increasing their visibility to a broader audience, potentially extending the impact of the content. For content creators and marketers, strategic tag usage can significantly enhance the effectiveness of social media campaigns by targeting specific demographics and trends. In essence, tags in social media posts bridge the gap between content creators and consumers, promoting interaction, discoverability, and the dissemination of ideas, making them an indispensable component of the social media ecosystem. Exploring the individual impact of each tag on recipe ratings presents an intriguing research opportunity.

Table 5- Most frequent tags in Valio recipes.

| Tag | Frequency |
|--------------------|------------------|
| Vegetable | 12% |
| Gluten free | 9% |
| Plenty of protein | 7% |
| Low-lactose | 6% |
| Lactose free | 6% |
| Christmas recipes | 4% |
| Harvesting | 3% |
| Oven dishes | 2% |
| Less salt | 2% |
| Lighter | 2% |
| Salads | 2% |
| New Year | 1% |
| Soups | 1% |
| Cheesecakes | 1% |
| Savory pies | 1% |
| Eggless | 1% |
| Breads | 1% |
| Nut recipes | 1% |
| Cocktail pieces | 1% |
| Christmas pastries | 1% |
| Casserole | 1% |
| Rice recipes | 1% |
| Tuna recipes | 1% |

5. REFERENCES

- [1] D. Mößlang, “Predicting the Popularity of Online Recipes,” 2017. doi: 10.1145/1787234.1787254.
- [2] C. Trattner, D. Moesslang, and D. Elsweiler, “On the predictability of the popularity of online recipes,” *EPJ Data Sci.*, vol. 7, no. 1, 2018, doi: 10.1140/epjds/s13688-018-0149-5.
- [3] C. Trattner, D. Parra, and D. Elsweiler, “Monitoring obesity prevalence in the United States through bookmarking activities in online food portals,” *PLoS One*, vol. 12, no. 6, p. e0179144, 2017.
- [4] M. De Choudhury, S. Sharma, and E. Kiciman, “Characterizing dietary choices, nutrition, and language in food deserts via social media,” in *Proceedings of the 19th acm conference on computer-supported cooperative work & social computing*, 2016, pp. 1157–1170.
- [5] S. Abbar, Y. Mejova, and I. Weber, “You tweet what you eat: Studying food consumption through twitter,” in *Proceedings of the 33rd annual acm conference on human factors in computing systems*, 2015, pp. 3197–3206.
- [6] D. Fried, M. Surdeanu, S. Kobourov, M. Hingle, and D. Bell, “Analyzing the language of food on social media,” in *2014 IEEE International Conference on Big Data (Big Data)*, 2014, pp. 778–783.
- [7] C. Wagner and L. M. Aiello, “Men eat on mars, women on venus? an empirical study of food-images,” in *Proceedings of the ACM Web Science Conference*, 2015, pp. 1–3.
- [8] R. Chunara, L. Bouton, J. W. Ayers, and J. S. Brownstein, “Assessing the online social environment for surveillance of obesity prevalence,” *PLoS One*, vol. 8, no. 4, p. e61373, 2013.
- [9] A. Said and A. Bellog\’in, “You are What You Eat! Tracking Health Through Recipe Interactions.,” 2014.
- [10] T. Kusmierczyk and K. Nørnvåg, “Online food recipe title semantics: Combining nutrient facts and topics,” in *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, 2016, pp. 2013–2016.
- [11] T. Kusmierczyk, C. Trattner, and K. Nørnvåg, “Temporal patterns in online food innovation,” in *Proceedings of the 24th international conference on world wide web*, 2015, pp. 1345–1350.
- [12] Y.-Y. Ahn, S. E. Ahnert, J. P. Bagrow, and A.-L. Barabási, “Flavor network and the principles of food pairing,” *Sci. Rep.*, vol. 1, no. 1, p. 196, 2011.
- [13] B. Scheibehenne, L. Miesler, and P. M. Todd, “Fast and frugal food choices:

Uncovering individual decision heuristics,” *Appetite*, vol. 49, no. 3, pp. 578–589, 2007.

- [14] D. Elswailer, C. Trattner, and M. Harvey, “Exploiting food choice biases for healthier recipe recommendation,” in *Proceedings of the 40th international acm sigir conference on research and development in information retrieval*, 2017, pp. 575–584.
- [15] M. Rokicki, E. Herder, T. Kuśmierczyk, and C. Trattner, “Plate and prejudice: Gender differences in online cooking,” in *Proceedings of the 2016 conference on user modeling adaptation and personalization*, 2016, pp. 207–215.
- [16] M. Thelwall, K. Buckley, G. Paltoglou, D. Cai, and A. Kappas, “Sentiment strength detection in short informal text,” *J. Am. Soc. Inf. Sci. Technol.*, vol. 61, no. 12, pp. 2544–2558, 2010, [Online]. Available: <http://sentistrength.wlv.ac.uk/>
- [17] R. W. Hardian, P. E. Prasetyo, U. Khaira, and T. Suratno, “Analisis Sentiment Kuliah Daring Di Media Sosial Twitter Selama Pandemi Covid-19 Menggunakan Algoritma Sentistrength: Online Lecture Sentiment Analisis On Twitter Social Media During The Covid-19 Pandemic Using Sentistrength Algorithm,” *MALCOM Indones. J. Mach. Learn. Comput. Sci.*, vol. 1, no. 2, pp. 138–143, 2021.
- [18] A. M. Rabab’Ah, M. Al-Ayyoub, Y. Jararweh, and M. N. Al-Kabi, “Evaluating sentistrength for arabic sentiment analysis,” in *2016 7th International Conference on Computer Science and Information Technology (CSIT)*, 2016, pp. 1–6.
- [19] M. Thelwall, “The Heart and soul of the web? Sentiment strength detection in the social web with SentiStrength,” *Cyberemotions Collect. Emot. Cybersp.*, pp. 119–134, 2017.
- [20] S. M. Lundberg and S.-I. Lee, “A Unified Approach to Interpreting Model Predictions,” in *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds. Curran Associates, Inc., 2017, pp. 4765–4774. [Online]. Available: <http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf>
- [21] K. El Mokhtari, B. P. Higdon, and A. Baccar, “Interpreting financial time series with SHAP values,” in *Proceedings of the 29th annual international conference on computer science and software engineering*, 2019, pp. 166–172.
- [22] Y. Meng, N. Yang, Z. Qian, and G. Zhang, “What makes an online review more helpful: an interpretation framework using XGBoost and SHAP values,” *J. Theor. Appl. Electron. Commer. Res.*, vol. 16, no. 3, pp. 466–490, 2020.
- [23] W. E. Markovits and D. M. Eler, “From explanations to feature selection:

assessing SHAP values as feature selection mechanism,” in *2020 33rd SIBGRAPI conference on Graphics, Patterns and Images (SIBGRAPI)*, 2020, pp. 340–347.

- [24] W. Zhao, T. Joshi, V. N. Nair, and A. Sudjianto, “Shap values for explaining cnn-based text classification models,” *arXiv Prepr. arXiv2008.11825*, 2020.
- [25] D. R. Cox, “Regression models and life-tables,” *J. R. Stat. Soc. Ser. B*, vol. 34, no. 2, pp. 187–202, 1972.
- [26] J. Ahn, “Digital divides and social network sites: Which students participate in social media?,” *J. Educ. Comput. Res.*, vol. 45, no. 2, pp. 147–163, 2011.
- [27] S. T. Indra, L. Wikarsa, and R. Turang, “Using logistic regression method to classify tweets into the selected topics,” in *2016 international conference on advanced computer science and information systems (icacsis)*, 2016, pp. 385–390.
- [28] A. Vannucci, K. M. Flannery, and C. M. Ohannessian, “Social media use and anxiety in emerging adults,” *J. Affect. Disord.*, vol. 207, pp. 163–166, 2017.
- [29] J. C. Levenson, A. Shensa, J. E. Sidani, J. B. Colditz, and B. A. Primack, “The association between social media use and sleep disturbance among young adults,” *Prev. Med. (Baltim.)*, vol. 85, pp. 36–41, 2016.
- [30] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, “Learning representations by back-propagating errors,” *Nature*, vol. 323, no. 6088, pp. 533–536, 1986.
- [31] K. Patil and N. Jadhav, “Multi-layer perceptron classifier and paillier encryption scheme for friend recommendation system,” in *2017 International Conference on Computing, Communication, Control and Automation (ICCCUBEA)*, 2017, pp. 1–5.
- [32] A. H. Danesh and H. Shirgahi, “Predicting trust in a social network based on structural similarities using a multi-layered perceptron neural network,” *Iium Eng. J.*, vol. 22, no. 1, pp. 103–117, 2021.
- [33] P. Garg and S. N. Singh, “Multilayer Perceptron Optimization Approaches for Detecting Spam on Social Media Based on Recursive Feature Elimination,” in *Applications of Artificial Intelligence and Machine Learning: Select Proceedings of ICAAIML 2021*, Springer, 2022, pp. 501–510.
- [34] S. Ghosal and A. Jain, “Analysis of Misogynistic and Aggressive Text in Social Media with Multilayer Perceptron,” in *Emerging Technologies in Data Mining and Information Security: Proceedings of IEMIS 2022, Volume 3*, Springer, 2022, pp. 589–596.
- [35] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” *arXiv Prepr. arXiv1412.6980*, 2014.

- [36] L. Breiman, “Random Forests,” vol. 45, pp. 5–32, 2001.
- [37] M. Aufar, R. Andreswari, and D. Pramesti, “Sentiment analysis on youtube social media using decision tree and random forest algorithm: A case study,” in *2020 International Conference on Data Science and Its Applications (ICoDSA)*, 2020, pp. 1–7.
- [38] F. Huang, J. Chen, Z. Lin, P. Kang, and Z. Yang, “Random forest exploiting post-related and user-related features for social media popularity prediction,” in *Proceedings of the 26th ACM international conference on Multimedia*, 2018, pp. 2013–2017.
- [39] P. Karthika, R. Murugeswari, and R. Manoranjithem, “Sentiment analysis of social media network using random forest algorithm,” in *2019 IEEE international conference on intelligent techniques in control, optimization and signal processing (INCOS)*, 2019, pp. 1–5.
- [40] C.-C. Hsu *et al.*, “Social media prediction based on residual learning and random forest,” in *Proceedings of the 25th ACM international conference on Multimedia*, 2017, pp. 1865–1870.
- [41] C. Cortes and V. Vapnik, “Support-vector networks,” *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, 1995.
- [42] A.-Z. Ala’M, H. Faris, J. Alqatawna, and M. A. Hassonah, “Evolving support vector machines using whale optimization algorithm for spam profiles detection on online social networks in different lingual contexts,” *Knowledge-Based Syst.*, vol. 153, pp. 91–104, 2018.
- [43] M. N. Murty and R. Raghava, “Support vector machines and perceptrons: Learning, optimization, classification, and application to social networks,” 2016.
- [44] A. M. U. D. Khanday, Q. R. Khan, and S. T. Rabani, “SVMBPI: support vector machine-based propaganda identification,” in *Cognitive Informatics and Soft Computing: Proceeding of CISC 2020*, 2021, pp. 445–455.
- [45] J. H. Friedman, “Greedy function approximation: a gradient boosting machine,” *Ann. Stat.*, pp. 1189–1232, 2001.
- [46] V. Athanasiou and M. Maragoudakis, “A novel, gradient boosting framework for sentiment analysis in languages where NLP resources are not plentiful: A case study for modern Greek,” *Algorithms*, vol. 10, no. 1, p. 34, 2017.
- [47] S. Neelakandan and D. Paulraj, “A gradient boosted decision tree-based sentiment classification of twitter data,” *Int. J. Wavelets, Multiresolution Inf. Process.*, vol. 18, no. 04, p. 2050027, 2020.
- [48] M. H. Abdurrahman, B. Irawan, and C. Setianingsih, “A review of light gradient boosting machine method for hate speech classification on twitter,” in *2020 2nd*

International Conference on Electrical, Control and Instrumentation Engineering (ICECIE), 2020, pp. 1–6.

- [49] J. Bergstra and Y. Bengio, “Random search for hyper-parameter optimization.,” *J. Mach. Learn. Res.*, vol. 13, no. 2, 2012.
- [50] A. Chadha and B. Kaushik, “A hybrid deep learning model using grid search and cross-validation for effective classification and prediction of suicidal ideation from social network data,” *New Gener. Comput.*, vol. 40, no. 4, pp. 889–914, 2022.
- [51] M. Adnan, A. A. S. Alarood, M. I. Uddin, and I. ur Rehman, “Utilizing grid search cross-validation with adaptive boosting for augmenting performance of machine learning models,” *PeerJ Comput. Sci.*, vol. 8, p. e803, 2022.
- [52] T. Yan, S.-L. Shen, A. Zhou, and X. Chen, “Prediction of geological characteristics from shield operational parameters by integrating grid search and K-fold cross validation into stacking classification algorithm,” *J. Rock Mech. Geotech. Eng.*, vol. 14, no. 4, pp. 1292–1303, 2022.
- [53] A. Geissinger and C. Laurell, “User engagement in social media--an explorative study of Swedish fashion brands,” *J. Fash. Mark. Manag.*, vol. 20, no. 2, pp. 177–190, 2016.
- [54] H. Shahbaznezhad, R. Dolan, and M. Rashidirad, “The role of social media content format and platform in users’ engagement behavior,” *J. Interact. Mark.*, vol. 53, no. 1, pp. 47–65, 2021.
- [55] R. Jaakonmäki, O. Müller, and J. Vom Brocke, “The impact of content, context, and creator on user engagement in social media marketing,” in *Proceedings of the Annual Hawaii International Conference on System Sciences*, 2017, vol. 50, pp. 1152–1160.
- [56] P. M. Di Gangi and M. M. Wasko, “Social media engagement theory: Exploring the influence of user engagement on social media usage,” *J. Organ. End User Comput.*, vol. 28, no. 2, pp. 53–73, 2016.
- [57] K. Lewis, J. Kaufman, M. Gonzalez, A. Wimmer, and N. Christakis, “Tastes, ties, and time: A new social network dataset using Facebook. com,” *Soc. Networks*, vol. 30, no. 4, pp. 330–342, 2008.
- [58] B. Naderer, “Advertising unhealthy food to children: on the importance of regulations, parenting styles, and media literacy,” *Curr. Addict. Reports*, vol. 8, pp. 12–18, 2021.
- [59] J. A. Mennella and N. K. Bobowski, “The sweetness and bitterness of childhood: Insights from basic research on taste preferences,” *Physiol. & Behav.*, vol. 152, pp. 502–507, 2015.
- [60] G. Murphy, C. Corcoran, M. Tatlow-Golden, E. Boyland, and B. Rooney, “See,

like, share, remember: adolescents' responses to unhealthy-, healthy-and non-food advertising in social media," *Int. J. Environ. Res. Public Health*, vol. 17, no. 7, p. 2181, 2020.

6. APPENDICES

6.1. Logistic Regression Classifier SHAP Values

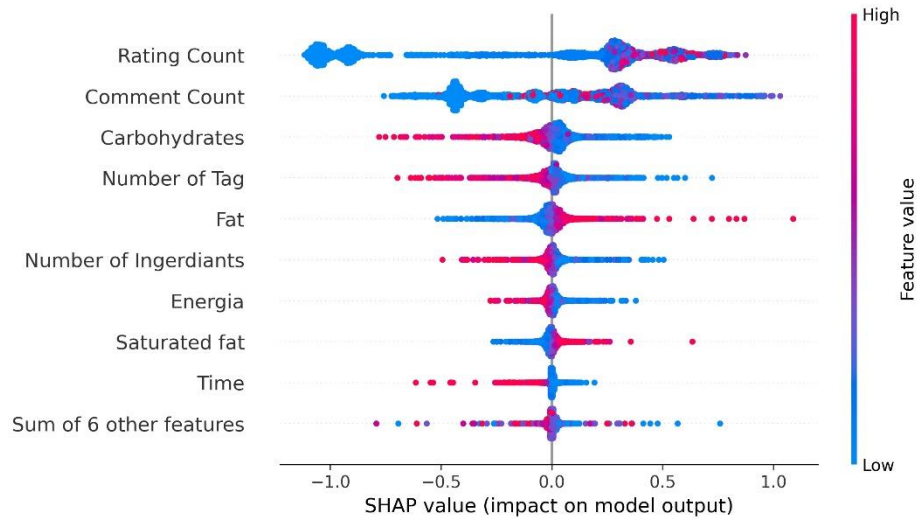


Figure S1- The importance of futures in Logistic Regression classifier.

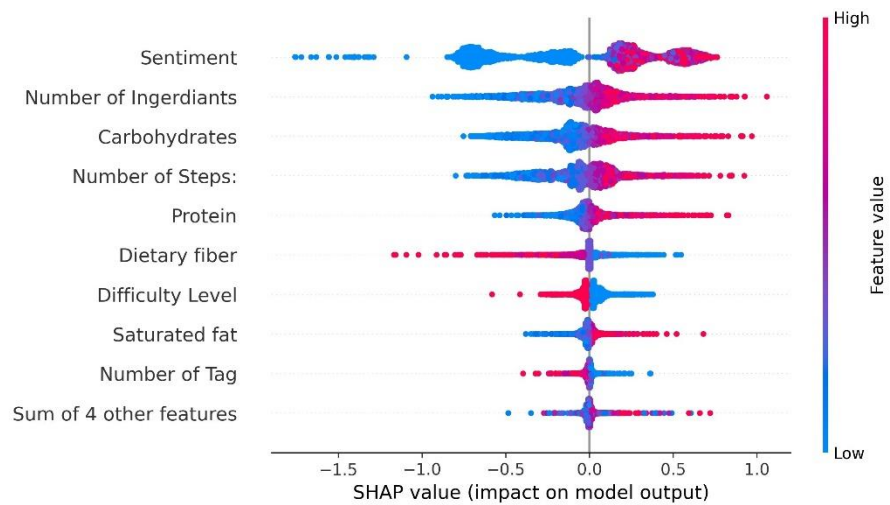


Figure S2- The importance of futures in Logistic Regression classifier when rating count and comment count are excluded.

6.2. Random Forest Classifier SHAP Values

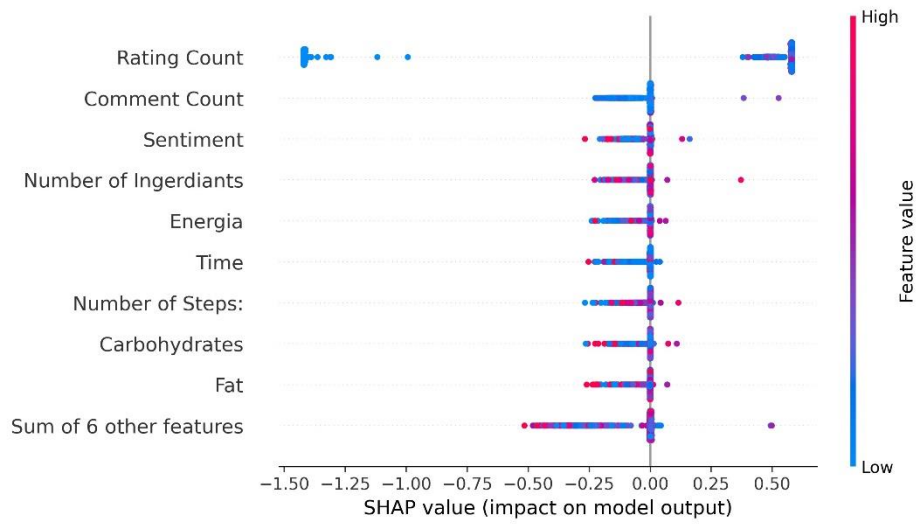


Figure S3- The importance of futures in Random Forest classifier.

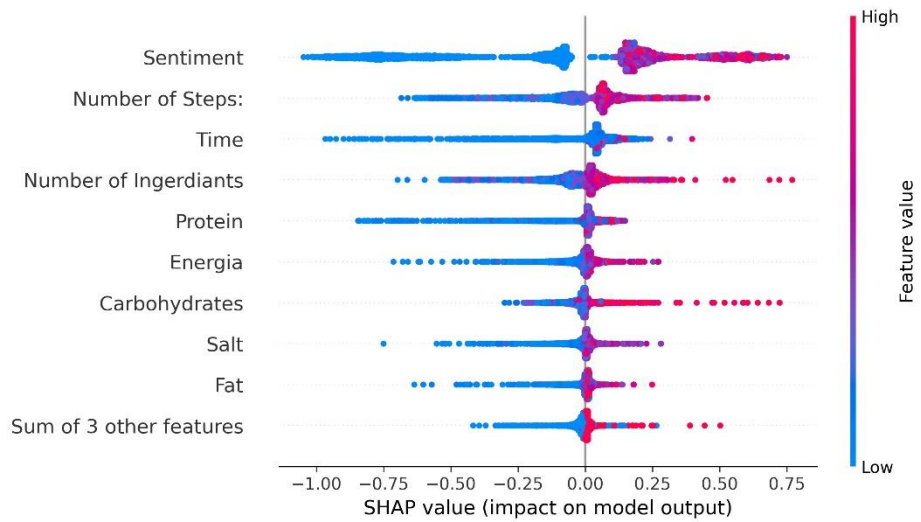


Figure S4- The importance of futures in Random Forest classifier when rating count and comment count are excluded.

6.3. Gradient Boosting Classifier SHAP Values

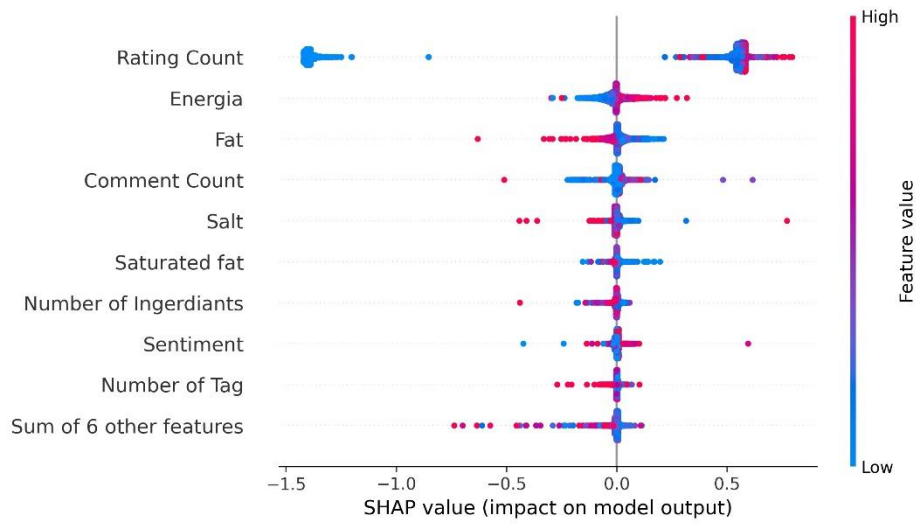


Figure S5- The importance of futures in Gradient Boosting classifier.

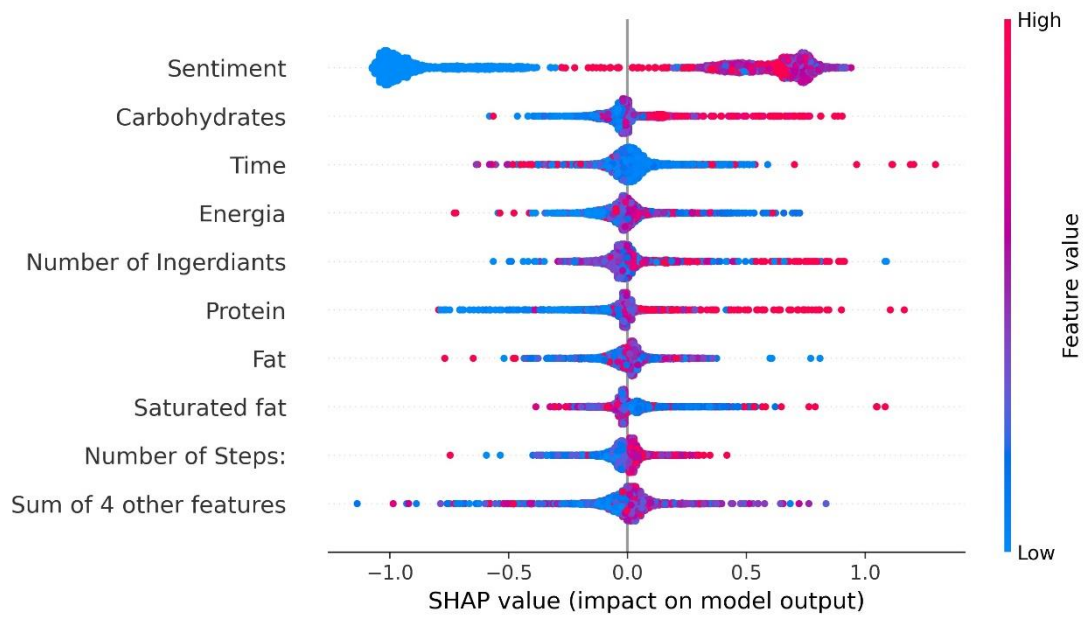


Figure S6- The importance of futures in Gradient Boosting classifier when rating count and comment count are excluded.