Team 2: Recipe Recommendation System

Aditi Jain, Vidhya Kewale, Akshaya Perumal, Shriya Purohit, Savrina Salartash, Priyanka Singh

Introduction

In response to the growing challenge of information overload and the neglect of nutritional considerations in online recipe searches, our project seeks to develop a user-friendly web-based recipe recommendation system. This system aims to provide personalized recipe suggestions and nutritional information, catering specifically to the needs of university students who often grapple with time and budget constraints when planning meals. By integrating user-specific data such as dietary restrictions, cuisine preferences, and nutritional goals, our platform seeks to streamline the meal planning process. Leveraging Python for back-end tasks and Flask for web framework, our prototype utilizes machine learning algorithms to generate tailored recipe options. Here, we outline the objectives, methodologies, and outcomes of our project, along with results of validation and user testing. Additionally, we discuss the potential impact of our system on promoting healthier eating habits among young adults and the feasibility of its future release and development.

Literature Review

Food and Recipe Prediction

Some of the biggest components in food prediction applications include determining the ingredients in a dish, which cuisine it fits into, as well as generating recipes that cater to the users' requirements. Machine Learning algorithms such as the TF-IDF and SVM train and test a dataset consisting of the cuisine category and a list of ingredients, to extract features and identify which ingredients are common to certain cuisines, especially Indian, Mexican, British, and Brazilian cuisines [1]. Case-Based Reasoning & Naive Bayes classification were also used to enhance the generative capabilities of cooking recipe generation systems, along with NLP, to improve ingredient matching, recipe diversity, and adaptation to dietary preferences [2]. Reinforcement learning achieved a 21% improvement in the Ingredient Matching (IM) metric, generating recipes that more accurately reflect the input ingredients, based on a dataset of approximately 15,000 cooking recipes from Food.com [3].

Deep learning methods, including clustering and LSTM models, predict nutritional values and generate diet plans based on user-calorie requirements [5]. Clustering groups similar food items for nutritional prediction, and the LSTM model captures food item sequences, creating a comprehensive diet plan with a user-friendly interface. To extend this further, [8] uses similar natural processing procedures along with Convolutional Neural Networks (CNN) that help in identifying ingredients from food images with high F1 scores. This provides context for [9] , which explores recipe generation using two statistical language models. The ReProg approach, employing genetic programming, yielded 49% of recipes with good taste, while the ReComp approach, employing decomposition techniques, achieved higher results at 93%. Genetic programming was also used in AutoChef along with a Conditional Random Field (CRF) classifier to evolve new and optimal recipes by representing each recipe as a tree structure [5]. Additionally, transformer technology is used for recipe generation, processing tokenized ingredients and recipes through an encoder-decoder system, resulting in personalized recipes with a user-tested average score of 3.6 out of 5 [6].

Health and Nutrition

Home cooking has been linked to a higher quality diet, but practical considerations such as cost and time are often overlooked [4]. In recent times, popular recipe websites and their recommender systems often backfire by promoting unhealthy trendy content, contributing to societal health problems like obesity and diabetes. Furthermore, there are recipe recommenders that prioritize nutritional needs over user preferences. By simply presenting the nutritional content of a recipe, without altering the content, users can make healthier food choices consciously [12]. Robust measures such as the FSA score derived by the Food Standard Agency (FSA) in the United Kingdom have been incorporated in recipe recommendation systems recently to promote healthier food choices to combat the obesity crisis [13][17].

Methods

Innovations

In our project, we aim to mitigate bias in recommending recipes as our recommendation system is trained on a large corpus of recipes scraped from a diverse set of authors and food websites. The dataset is not biased towards any particular cuisine or food preferences, thereby suggesting a wide variety of recipes with the same user input ingredients. We have also chosen the novel approach of content-based filtering (i.e. TF-IDF (Term Frequency-Inverse Document Frequency) and K-Nearest Neighbor algorithms) using cosine similarity as a key metric to provide relevant and personalized recommendations to users. Moreover, we have implemented textual pre-processing techniques that are simple and efficient, designed to handle a large dataset of recipes demonstrating innovation in scalability and efficiency, even with a growing number of recipes.

Data Collection and Cleaning

Starting with the process of data collection, we retrieved two datasets from Kaggle (originally from Food.com) and merged them based on a common column called "Recipe ID". By combining them, we expanded our data repository with additional attributes related to each recipe. These attributes include user ratings, number of steps, and ingredients, thereby enabling a more comprehensive analysis and recommendation process. In the cleaning and preprocessing step, we removed rows with "null" or blank values, ensured consistent data types, removed incomplete or irrelevant rows, and removed duplicates using Python and OpenRefine. For instance, inconsistencies were observed in the presentation of the "Steps" column, which contains the recipe preparation instructions. Some recipes were numbered, while others lacked this numbering scheme. Utilizing General Refine Expression Language (GREL) expressions in OpenRefine, we standardized the presentation to ensure that all recipes followed a non-numbered format initially. The process of data cleaning resulted in 231,636 rows containing unique recipes.

Data Processing

To process the data further, we refined the 'tags' column in our dataset that consisted of several descriptions for each recipe. First, we pruned out descriptors such as 'course', 'occasion', 'holiday-event', and 'seasonal', which did not align with our project's objectives. Then, we developed dictionaries for cuisine types, diet types, and preparation time to extract and map the relevant tags for each recipe. The cuisine mapping dictionary consists of categories such as Asian, European, Caribbean, Latin American, and so forth. The diet type mapping dictionary refers to categories such as Vegan, Vegetarian and Low Carb. Lastly, the 'time to make' dictionary has time period mappings such as 15-minutes-or-less and 60-minutes-or-less.

Also, we removed any unnecessary words in the 'tags' column and replaced them with standardized values from the dictionaries. This made sure that all tags were consistent and easy to understand. Then, we split these tags into separate columns for cuisine, diet, and time categories, making it simpler to sort and classify recipes. To further segment recipes, we removed outliers in the 'minutes' column consisting of values that were significantly large, such as over 1440 minutes, or more than a day. We used this as the threshold cutoff so that our recipes focused on being time effective. Thus, after doing so, our processed data frame fits our goal to focus on recipes for students based on their cuisine, diet and time restrictions.

Lastly, we processed the ingredients and steps columns to seamlessly display them on the website in a numbered list format. The original dataset presented these details as lists of strings, which were not user-friendly. By parsing and formatting this data, we ensured that users could easily follow the recipe instructions and ingredient lists in a clear and structured manner. This enhancement significantly improved the user experience by providing a more intuitive and readable presentation of the recipe content.

Modeling

To refine and prepare our ingredient data for subsequent analysis and modeling, we employed text processing techniques used in natural language processing. For instance, redundant information such as weights and measures was removed from the ingredient lists, as these details do not add value to the vector encodings of the recipes and the Natural Language Toolkit library was employed to remove stop words, which are common words in the English language, and other common words that are ubiquitous among recipes, such as "oil", "sugar", "salt" etc. were filtered out.

We applied lemmatization to the ingredient data to reduce words to their base forms, thereby standardizing the vocabulary and enhancing the accuracy of our models. By lemmatizing the ingredients, we ensured that variations of the same word (e.g., "cooking" and "cooked") were represented consistently, facilitating more robust analysis and modeling outcomes. These text processing steps transformed the ingredients column into a more structured and usable format, setting the stage for subsequent analysis and modeling.

In order to develop a robust recipe recommendation system, we have explored and implemented two distinct algorithms so far; TF-IDF (Term Frequency-Inverse Document Frequency) and K Nearest Neighbors (KNN) algorithm. The TF-IDF algorithm quantifies the importance of each ingredient within a recipe by considering its frequency in the recipe relative to its frequency across all recipes in the dataset. By employing cosine similarity, we can then identify recipes with similar ingredient profiles, offering users personalized recommendations based on the ingredients they have at hand, which is used as the input on our website. We chose this approach as it captures the similarity between recipes based on ingredients and provides relevant suggestions to the user. Meanwhile, the KNN algorithm operates by finding the nearest neighbors in a high dimensional space of ingredient vectors. We chose the cosine distance as the evaluation metric for the algorithm. This means that the algorithm evaluates similarity based on the cosine distance between ingredient vectors, allowing us to recommend recipes that closely align with user preferences.

Ultimately, we chose to utilize the TF-IDF algorithm because of its effectiveness in capturing the relevance of ingredients and their significance in recipe recommendation. TF-IDF (Term Frequency-Inverse Document Frequency) considers both the frequency of ingredients in recipes and their importance across the entire dataset, allowing it to identify the most distinctive ingredients for each cuisine type, diet type, and time-to-make preference. The web application takes in four inputs from the user: ingredients, cuisine type, diet type, and time-to-make. Ingredients can be specified through an input text box, and the other three inputs are selectable via dropdown list. The program outputs the top 10 recipes, including their name, ingredients, steps to make, user rating, and nutritional information based on the algorithm. The top 10 recommended recipes will be ordered based on historical recipe ratings and the recipe complexity, determined by the number of steps, as well as the calorie content.

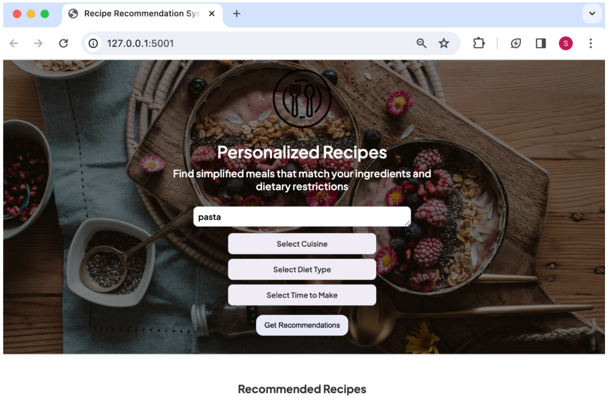

Figure 1: User Interface.

Front-End Interface

The frontend of the application, built using HTML and styled with Tailwind CSS, provides the user interface through which users interact with the system. It includes features such as dropdown menus, text boxes, and graphs for displaying recipe information, user ratings, and nutritional data. On the backend, the application utilizes Flask, a Python web framework, to handle server-side logic, route requests, and integrate with the frontend components. Flask facilitates the processing of user preferences, the retrieval of recipe data from the backend server, and the generation of personalized recipe recommendations.

The application fetches recipe data from a backend server and displays it in a visually appealing format. We initially developed a recipe recommendation application using React.js and styled it with Tailwind CSS, aiming to compare it with a Flask-based implementation to evaluate factors such as development complexity, performance, and scalability. We ultimately chose Flask for its server-side rendering, backend logic, and simpler setup, especially for implementing our Python modeling code. Using HTML, CSS, and JavaScript, we designed a user-friendly interface for the recommendation system, then integrated it into the Flask code.

To ensure personalized recipe recommendations, users are prompted to input their cuisine preferences, dietary goals, and time constraints via dropdown menus and text boxes. The application then processes this information and generates a curated list of the top 10 recommended recipes that best match the user's criteria. Each recommended recipe is accompanied by detailed information, including its name, list of ingredients, cooking time, and step-by-step instructions. This comprehensive presentation empowers users to make informed decisions about which recipes to explore further. Furthermore, the inclusion of two graphs enhances the user experience by providing additional insights. The first graph displays user ratings, allowing users to quickly assess the popularity and satisfaction levels of each recipe. The second graph presents nutritional information, such as carbohydrates, saturated fat, protein, sodium, sugar, and total fat, in a visually appealing format. This enables users to evaluate the healthfulness of each dish and make choices aligned with their dietary preferences and goals.

Ensuring accessibility across various devices, the application is responsive, dynamically adjusting its layout to fit the screen size of any device. Whether accessed from a laptop, tablet, or smartphone, users can seamlessly interact with the application and explore recipe recommendations in a visually cohesive and user-friendly manner.

Visualizations

We have incorporated two insightful visualizations to enhance user experience and facilitate informed decision-making regarding recipe choices on our website. The first visualization is a histogram showcasing the distribution of recipe ratings. On the x-axis, we have the stars, while

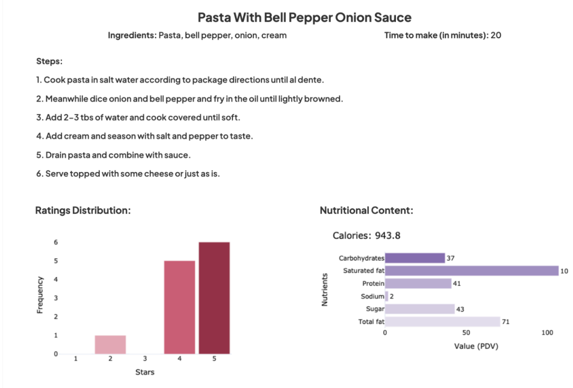

Figure 2: Example of a generated recipe and corresponding plots.

the y-axis indicates the frequency count. This histogram provides users with an overview of how recipes are rated by others, helping them gauge the popularity and quality of each recipe.

The second visualization offers nutritional information presented in terms of percent daily value (PDV). It includes key nutrients such as carbohydrates, saturated fat, protein, sodium, sugar, and total fat. This information is displayed in a bar chart format, with the number of calories prominently displayed on top of the chart. By visualizing the nutritional content of each recipe, users can make informed decisions about their dietary choices and understand the health implications of the recipes they are considering.

These visualizations are strategically placed beneath each recipe, ensuring easy access and providing users with valuable insights without overwhelming them with information. By incorporating these visual aids, our website offers an engaging and informative experience, empowering users to select recipes that align with their preferences and dietary needs.

Experiments and Evaluation

Throughout our project, we conducted extensive experimentation and evaluation to ensure the effectiveness of our model and the user-friendliness of our web application. Our experiments aimed to address two key factors:

1. Most Effective Machine Learning Model: Developed and compared two machine learning algorithms, TF-IDF and KNN, to determine which one is best suited for user-specific recipe generation.

2. User-Friendliness of the Web Application: Evaluated the usability and visual appeal of our web application through both internal testing and user feedback.

Firstly, while selecting the best recommendation algorithm, we chose the TF-IDF algorithm over KNN for several reasons, primarily due to its ability to capture the relevance and significance of ingredients within recipes. While KNN operates by finding nearest neighbors based on ingredient vectors, TF-IDF considers both the frequency of ingredients within recipes and their importance across the entire dataset. This comprehensive approach enables TF-IDF to identify the most distinctive ingredients for each cuisine type, diet preference, and time-to-make category, resulting in more tailored and accurate recipe recommendations. Additionally, TF-IDF's utilization of cosine similarity facilitates efficient comparison of ingredient profiles, making it well-suited for large-scale recommendation systems where computational efficiency is crucial. Moreover, TF-IDF's emphasis on the importance of ingredients aligns well with our goal of providing users with personalized recipe suggestions based on their specific preferences and dietary restrictions. Overall, TF-IDF emerged as the preferred choice due to its effectiveness in capturing ingredient relevance and its alignment with the objectives of our recipe recommendation system.

Initially, our focus was on ensuring the functional integrity of the application. This involved validating that the application properly displayed data, accurately processed user inputs, and provided correct recipe information. We meticulously tested each functionality, iteratively refining both the front- and back-end code until all aspects of the application performed as intended.In addition to functionality, we prioritized the aesthetic appeal of the application. Through iterative testing involving color schemes, layout designs, and figure presentation, we arrived at the final product, now hosted on Google Cloud, which offers a visually pleasing and engaging user experience. Once we were confident in the functional and visual aspects of the application, we conducted user testing with 30 volunteers and requested them to rank the page based on several factors. They were asked to evaluate the application based on several criteria, including visual appeal, usability of the recipe generation function, and overall satisfaction with the generated recipes.

First, users ranked the look of the web application without taking the quality of the output into account. We received feedback on the color scheme, the figures, and how the steps to make the recipe were displayed. With this feedback, we were able to make essential improvements to the application to make it more visually appealing to users. Next, users ranked the usability of the recipe generation function. Finally, users were asked to make a recipe of their choice from the application and rank it on a scale from 1 to 5, with 1 being the worst and 5 being the best. Overall, 98% of users had a positive experience and enjoyed the meal that the application generated for them. A key feedback that we received was that there were no quantities specified in regard to ingredients and this is a limitation that arose due to the lack of the ingredient quantities in the original dataset. In future iterations of this project, we hope to address this issue by sourcing recipes directly from recipe websites using techniques such as web scraping instead of utilizing the Kaggle dataset.

Team Member Contributions

Team members' effort has been distributed evenly, and specific task delegation is based upon member preference and skillset. Many of the tasks were completed collaboratively, ensuring that each member played a role in each step of the process.

Conclusion

In summary, our project represents a significant advancement in addressing the challenges of information overload and nutritional neglect in online recipe searches. By seamlessly integrating Python, Flask, and cutting-edge machine learning algorithms such as TF-IDF and K Nearest Neighbors, we've developed a sophisticated yet user-friendly recipe recommendation system tailored specifically for university students. Throughout our project, we have emphasized the importance of data quality, employing rigorous data collection, cleaning, and processing techniques to ensure the reliability and relevance of our dataset. By training on a diverse dataset of recipes and employing innovative textual preprocessing methods, we have effectively mitigated bias and demonstrated scalability and efficiency in handling large datasets. User feedback and usability testing have provided valuable insights into the effectiveness and user-friendliness of our application. While our initial results are promising, we recognize the potential for further enhancements. Incorporating features such as ingredient quantities and expanding the dataset's diversity are among our future considerations. In conclusion, our project marks a significant step forward in leveraging data analytics, machine learning, and front-end techniques to empower users in making informed and healthier meal choices. With its potential to foster healthier eating habits among young adults, our recipe recommendation system holds promise for broader societal impact and represents a meaningful contribution to the intersection of technology and nutrition.

References

[1] Alshanketi, F. ”Machine Learning Model for Predicting the Cuisine Category from a Dish Ingredients.” In: *2023 International Conference on Smart Computing and Application (ICSCA)*, Hail, Saudi Arabia, 2023, pp. 1-6. doi: 10.1109/ICSCA57840.2023.10087436.

[2] Galanis, Nikolaos-Ioannis, and George A. Papakostas. “An Update on Cooking Recipe Generation with Machine Learning and Natural Language Processing.” In: *2022 IEEE World Conference on Applied Intelligence and Computing (AIC)*, 17 June 2022. doi:10.1109/aic55036.2022.9848929.

[3] Fujita, Jumpei, et al. “Model for Cooking Recipe Generation Using Reinforcement Learning.” In: *2021 IEEE 37th International Conference on Data Engineering Workshops (ICDEW)*, Apr. 2021. doi:10.1109/icdew53142.2021.00007.

[4] Hamade, Hana et al. ”Associations between Cooking at Home and Nutrient and Food Group Intake among Female University Students: A Cross-Sectional Analysis on Living Arrangements.” Nutrients, vol. 15, no. 4, 2023, article 1029, https://doi.org/10.3390/nu15041029.

[5] Jabeen, Hajira, et al. “Autochef: Automated Generation of Cooking Recipes.” In: *2020 IEEE Congress on Evolutionary Computation (CEC)*, July 2020. doi:10.1109/cec48606.2020.9185605.

[6] Phirke, Nachiket, et al. ”Diet Planner Using Deep Learning.” *International Journal of Emerging Technologies and Innovative Research*, vol. 10, no. 4, April 2023, pp. f421-f425, ISSN: 2349-5162, www.jetir.org.

[7] Lam, K.N., Pham, Y.N.T., Kalita, J. ”Cooking Recipe Generation Based on Ingredients Using ViT5.” In: T.D.L. Nguyen et al. (eds.), *Intelligent Systems and Networks*, vol. 752, Springer, Singapore, 2023. https://doi.org/10.1007/978-981-99-4725-6_5.

[8] Neha, K., et al. ”Food Prediction based on Recipe using Machine Learning Algorithms.” In: *2023 Second International Conference on Augmented Intelligence and Sustainable Systems (ICAISS)*, Trichy, India, 2023, pp. 411-416, doi: 10.1109/ICAISS58487.2023.10250758.

[9] De Santos, W. A., Bezerra, J. R., Wanderley Góes, L. F., & Ferreira, F. M. F. ”Creative Culinary Recipe Generation Based on Statistical Language Models.” *IEEE Access*, vol. 8, pp. 146263-146283, 2020, doi: 10.1109/ACCESS.2020.3013436.

[10] Silveira, T., et al. “How good your recommender system is? A survey on evaluations in recommendation.” *International Journal of Machine Learning and Cybernetics*, vol. 10, no. 5, 14 Dec. 2017, pp. 813–831, https://doi.org/10.1007/s13042-017-0762-9.

[11] Ruani, M. A., et al. “Diet-Nutrition Information Seeking, Source Trustworthiness, and Eating Behavior Changes: An International Web-Based Survey.” *Nutrients*, vol. 15, no. 21, 1 Jan. 2023, p. 4515, www.mdpi.com/2072-6643/15/21/4515, doi: 10.3390/nu15214515.

[12] El Majjodi, A., et al. “Nudging towards health? Examining the merits of nutrition labels and personalization in a recipe recommender system.” In: *Proceedings of the 30th ACM Conference on User Modeling, Adaptation and Personalization*, 4 July 2022, https://doi.org/10.1145/3503252.3531312.

[13] Trattner, C., & Elsweiler, D. “An Evaluation of Recommendation Algorithms for Online Recipe Portals.” *HealthRecSys@RecSys* (2019).

[14] Ispirova, G., et al. "MsGEN: Measuring Generalization of Nutrient Value Prediction across Different Recipe Datasets." *Expert Systems with Applications*, vol. 237, 1 Mar. 2024, p. 121507, `https://doi.org/10.1016/j.eswa.2023.121507`.

[15] Chavan, P., Thoms, B., & Isaacs, J. "A recommender system for healthy food choices: building a hybrid model for recipe recommendations using big data sets." (2021).

[16] Jabeen, H., et al. "EvoChef: Show Me What to Cook! Artificial Evolution of Culinary Arts." In: *Computational Intelligence in Music, Sound, Art and Design*, 2019, pp. 156–172, `https://doi.org/10.1007/978-3-030-16667-0_11`. Accessed 13 Oct. 2022.

[17] Trattner, C., Elsweiler, D., & Howard, S. "Estimating the Healthiness of Internet Recipes: A Cross-sectional Study." *Front Public Health.* 2017 Feb 13;5:16. doi: 10.3389/fpubh.2017.00016.

[18] Patil, S., & Potdar, R. D. "Recipe Recommendation Systems: A Review." (2019).