UNIVERSITI
MALAYA

WQD7006 Machine Learning For Data Science

Semester 1, Session 2023/2024

INDIAN FOOD RECIPE RECOMMENDER USING MACHINE LEARNING

Group Project

Matric Number	Name
S2180377	Ying Ming Tang
S2164604	Elaine Li
22058059	Ze Ying Tan
S2197999	Hun Yee Chong
17193844	Tarsvini A/P Ravinther

TABLE OF CONTENT

CHAPTER 1 INTRODUCTION

1.1 Background of Study

In this era of modern globalization, people are very conscious about their health where they will include at least some healthy food in their daily meals (Shafaat et al., 2022). In addition, Lei et al. (2020) also mentioned that food is a daily essential aspect for human beings to maintain good health. Nowadays, with the increasing standard of living, it also leads to the increasing desire for good and quality food. Consequently, more and more people come out with online food recipes recommendation systems in order to help in making decisions on what food to eat or cook. When it comes to cooking, factors such as the diet, nutrients, ingredients as well as serving sizes have to be taken into consideration. Besides, there is not much time to waste, especially for the working adults and mothers who prepare meals for children and family. This is because there are so many aspects to be taken care of to just cook a meal. Based on Xie and Lou (2022), with the help of the recommendation system, it will allow them to save time to search and cook the meals they want based on their own preferences and also improve the efficiency of information acquisition. Although there are many researchers working on food recipe recommender systems, none of them have explored Indian cuisine. Therefore, this research project will be focusing on Indian food recipe recommendations. With this recommendation system, it will provide recipes to users by considering the quantity, calories and nutrition of ingredients which will be used. In addition to that, the relationship between recipes which share similar ingredients will be identified. Besides, the recommendation system in this study aims to balance between personalized food preference and personalized health requirement. Lastly, it is targeted to evaluate models based on performance and accuracy which align with users' preferences and nutritional needs. Data is created and obtained from Archana's Kitchen website through crawling, scraping, cleaning and transforming content. On the other hand, in this study, supervised machine learning algorithms such as Support Vector Machine (SVM), Logistics Regression, Naive Bayes, Random Forest and Decision Tree will be used to tackle problems like classification and clustering.

1.2 Problem Statement

In the past, several research studies on food recipe recommendation have been proposed with the consideration of various goals such as to suggest healthy recipes or to create a personalized recommendation based on users' tastes. This is because food waste can be influenced by users' preferences (Jain, H., 2018). A good food recipe recommender system shall be able to encourage or attract users to start with the recipe preparation. With this, the problem of food wastage will be reduced as well. However, recommending food recipes is still an open issue as most researchers have overlooked the importance of nutritional aspects (Chen et al., 2020) such as food quantity, food quality, calories and nutrition of ingredients. This oversight leads to the problem of incomplete understanding of the users' dietary needs. Moreover, not many researchers have studied Indian food recipes. With all of the above, this implies the need for new studies and exploration in this field.

1.3 Research Objectives

The research objectives are stated as follows:

i. To identify ingredient combinations suitable for different diets based on food nutrition.
ii. To determine the relationship between recipes that share similar ingredients.
iii. To build a personalized food recommendation system based on user preference on food and health requirements.
iv. To evaluate the model based on the performance and accuracy of the suggested diet to the users' preferences and nutritional needs.

CHAPTER 2 LITERATURE REVIEW

2.1 Overview

Table 2.1: Critical Analysis Table For Recommender System Using Machine Learning

No.	Citation	Model	Finding	Limitation	Future Research
1.	Alshanketi, F. (2023)	K-Nearest Neighbors, Naive Bayes, Logistic Regression, Random Forest, Decision Tree, SVM, Stochastic Gradient Descent, Gradient Boosting, XGBoost and AdaBoost classifiers.	- The study explored the relationship between ingredients and cuisines. - SVM algorithm outperformed the other algorithms by achieving the highest accuracy of 80%.	- The study overlooked factors like cooking methods and cultural influences in cuisine exploration. It also did not discuss challenges or limitations in using machine learning for real-world cuisine category prediction.	- May investigate the impact of cooking methods, cultural influences, and regional variations on cuisine prediction in order to create a more comprehensive model. - May extend the research to consider practical applications of machine learning in culinary settings, such as recipe recommendation systems or food pairing analyses.
2.	A. Banerjee, a. Noor, N. Siddiqua and M. N. Uddin. (2019)	Naïve Bayes, Random Forest, SVM, Fuzzy Lookup, Fuzzy Match	- The study sought improved food recommendations for those with chronic kidney disease. Random Forest outperformed Naïve Bayes and SVM, achieving an accuracy of 99.75%.	- The study used a small size of dataset which only contains 25 medical attributes, with 400 patients and 61 foods.	- May obtain larger dataset that include more health attributes, observations, and choice of food to provide a more comprehensive output.
3.	S. Jayaraman, T. Choudhury & P. Kumar (2017)	Naïve Bayes, Random Forest, Linear Support Vector Classification, Multinomial regression	- The study aimed to analyze the correlation between cuisines and the ingredients. - Linear SVC outperformed the other models with a 79% accuracy and demonstrated quicker processing compared to logistic regression and random forest.	- The accuracy is less than 80% for all four methods. - The study only used accuracy and running time as the performance metrics.	- May explore different methods on modeling to achieve higher accuracy. - May increase the performance metrics, such as Recall, Precision, F-measure, ROC, to have a comprehensive result.
4.	Elsweiler, D., Trattner, C., & Harvey, M. (2017)	Random Forest, Logistic Regression, Naïve Bayes	- The study aimed to provide recommendations to users for healthier dishes, based on the feasibility of substituting meals. - Random Forest has the best performance among the three	- The recommendation was not based on individual user preference. - For some circumstances, the recommendation will require user to choose	- May focus on user preference to enhance the recommendation system. -May increase nutritional elements such as fiber, protein in the recommendation. - May explore eye-tracking studies to investigate the user behavior on different recipe information.

			models, with 84.78% of accuracy. -10-fold cross validation is being used in this study. - Feature selection methods are being used to ensure the models were robust and interpretable.	non-preference items to proceed, for example vegetarians were required to choose from two meat-based dishes.	
5.	Kardam, S. S., Yadav, P., Thakkar, R., & Ingle, A. (2021)	K-Means, Random Forest	- The study aimed to build a website to provide recipe recommendations based on various health factors. - The study used two Machine Learning models, which L-means is used to cluster the food according to the calories, and Random Forest is used to classify the food items and provide prediction based on the input.	- The study solely examined K-Means and Random Forest as base machine learning models without assessing their accuracy. - The recipe recommendation relies on only 7 user inputs, potentially limiting its ability to cater to individual needs.	- May explore more Machine Learning methods to gain a good performance of models. - May evaluate the machine learning models based on different performance metrics, such as Recall, Precision, F-measure and ROC.
6.	Habibi, M., & Cahyo, P. W. (2020)	Cosine similarity, Support Vector Machine	- The study aimed to classify journals by using a classification model. - Support Vector Machine has higher accuracy by comparison to Cosine Similarity, with accuracy of 75%.	- The study only used two models which are the support vector machine and cosine similarity. - The accuracy is only 75%.	- May explore other machine learning methods to have a better accuracy of the result.
7.	Khatter et al. (2021)	Cosine similarity	- Cosine similarity is able to work efficiently and effectively using lesser computational time compared to other available similarities for the recommendation system.	- This study is limited to only using Cosine similarities. - The authors stated that Cosine similarity performs the best, but they did not state how well it performs.	- May further improve the proposed recommendation system so that users can rate and comment on the movies. - This proposed recommendation system may be integrated into a variety of ecommerce websites and may be used as a base for other recommendation systems.

The above critical analysis table shows the research studies conducted on various machine learning (ML) techniques applied in deploying recommender systems. Other than using the supervised ML models, Cosine similarity technique is also commonly used for recommender systems which can be seen in the papers from Habibi and Cahyo (2020) as well as Khatter et al. (2021). In summary, Habibi and Cahyo (2020) classified papers and proposed to explore new ML techniques to improve accuracy over the currently used SVM and cosine similarity models. Besides, Khatter et al. (2021) mentioned that this technique had better computational time and was able to work efficiently compared to other similarities. To add on, cosine similarity is used as a metric in various machine learning algorithms such as KNN in order to determine the distance between neighbors. This helps texture data to detect texture similarity in the recommendation process (M, D., 2022, July 7).

To further elaborate, cosine similarity is a metric that is used to measure the similarity between two vectors in orientation and not in size. It is determined by the dot product of two non-zero vectors divided by the magnitude of each vector. The range of cosine similarity is from -1 to 1, while 1 indicates greater similarity, 0 indicates no similarity found and -1 indicates the two vectors are in opposite directions. Consider two vectors a and b , the cosine similarity is defined by the below formula. (Mana, S. C., & Sasipraba, T., 2021).

$$\text{Cosine similarity (a,b)} = \frac{(a.b)}{||a||.||b||}$$

Therefore, this study will employ cosine similarity techniques to assist in building the Indian food recipe recommender system.

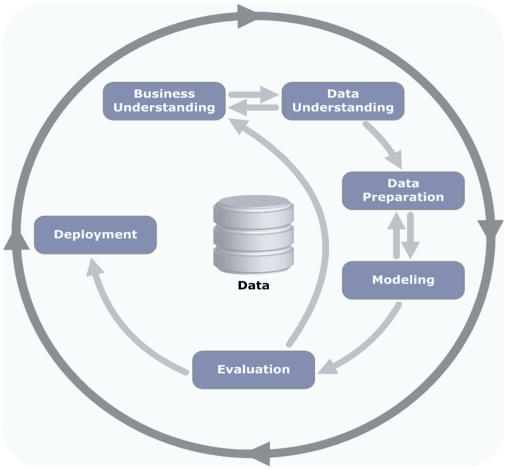
CHAPTER 3 METHODOLOGY

3.1 Dataset

The dataset is created from a website called https://www.archanaskitchen.com/ using web scraping. It consists of 6,871 rows and 15 columns which are ‘Srno’, ‘RecipeName’, ‘TranslatedRecipeName’, ‘Ingredients’, ‘TranslatedIngredients’, ‘PrepTimeInMins’, ‘CookTimeInMins’, ‘TotalTimeInMins’, ‘Servings’, ‘Cuisines’, ‘Course’, ‘Instructions’, ‘TranslatedInstructions’ and ‘URL’.

3.2 Framework

Cross-Industry Standard Process for Data Mining (CRISP-DM) framework will be applied throughout this project. This framework consists of a total of 6 phases which are business understanding, data understanding, data preparation, modeling, evaluation and deployment. These phases will be shown clearly in the diagram below.



3.2.1 Data Preparation

After checking, columns ‘‘Ingredients’ and ‘TranslatedIngredients' have 6 missing values respectively. Hence they are removed. Unnecessary columns such as ‘Srno’, ‘RecipeName’, ‘Ingredients’, ‘Instructions’, ‘TranslatedInstructions’, ‘URL’ are also removed. Since the focus of this project is on Indian food only, only Indian cuisines are retained which reduces the dataset to 4,871 rows.

3.2.2 String to Word Vector

StringToWordVector in Weka is used to filter ingredients from strings into vectors using word tokenizer class. Besides that, stopwords are also removed and TF-IDF is also utilized to weight words which helps in emphasizing important words while de-emphasizing common ones.

3.2.3 Machine Learning Algorithms

Supervised learning such as SVM, Logistics Regression, Naive Bayes, Random Forest and Decision Tree will be employed.

3.2.3.1 Logistics Regression

The logistic regression model is a widely used algorithm for prediction tasks. It offers several advantages in prediction, including its interpretability and ease of implementation. Logistic regression provides interpretable coefficients that allow us to understand the relationship between input variables and the predicted outcome. It also provides probability estimates, which can be useful for decision-making. Logistic regression handles both continuous and categorical variables, making it applicable to various prediction scenarios. Additionally, logistic regression is computationally efficient and can handle large datasets and high-dimensional feature spaces.

3.2.3.2 SVM

Support Vector Machines (SVMs) are powerful algorithms used for classification and prediction tasks. SVMs offer several advantages in prediction. They perform well in high-dimensional spaces, making them suitable for prediction tasks with a large number of predictors. SVMs can capture complex, non-linear relationships in the data by using kernel functions to transform the data into higher-dimensional spaces. They are robust to outliers, reducing their impact on the prediction process. SVMs generate flexible decision boundaries, allowing for accurate prediction even when classes or patterns are not easily separable.

3.2.3.3 Naive Bayes

Naive Bayes is a probabilistic classification algorithm that is widely used for prediction tasks. It offers several advantages in prediction. Naive Bayes models are computationally efficient and can handle large datasets with high-dimensional feature spaces. They are particularly effective when dealing with categorical and text data. Naive Bayes is robust to irrelevant features, allowing for efficient prediction even when there are many predictors. It also handles sparse data well, making it useful for prediction tasks with limited available data.

3.2.3.4 Random Forest

The random forest model is an ensemble learning algorithm that combines multiple decision trees to make predictions. It offers several advantages in prediction tasks. Random forests are robust to outliers and noise in the data, making them suitable for handling uncertainties. They can handle high-dimensional data without the need for feature selection or extensive preprocessing. Random forests capture complex relationships and interactions between variables without assuming a specific data distribution. They also provide estimates of feature importance, allowing for insights into the relative importance of different variables in the prediction process.

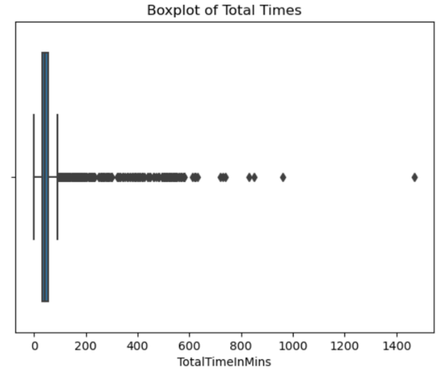
3.2.3.4 Decision Tree

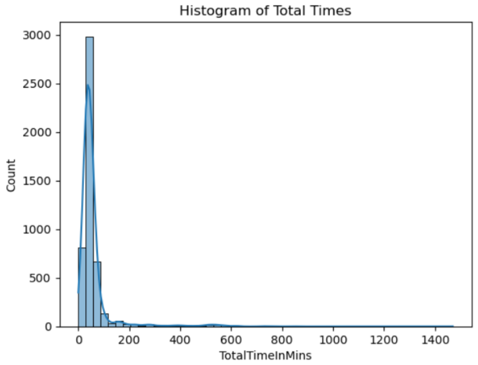
It builds a model like the structure of the tree. The method uses the if-then rule of mathematics to create subcategories that fit within broader categories and allow for precise and organized categorization.

CHAPTER 4 RESULTS AND DISCUSSION

4.1 Exploratory Data Analysis

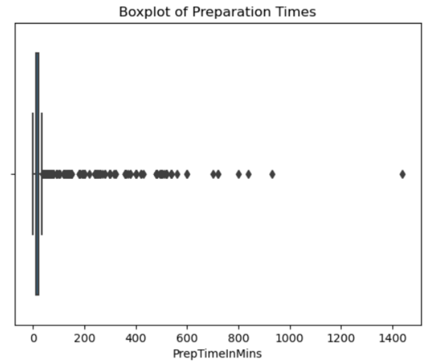
TotalTimeInMins

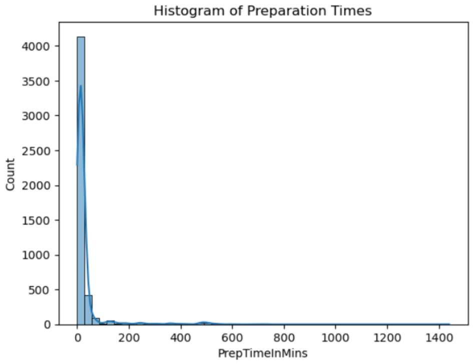




TotalTimeInMins is skewed to the right with a mean of 57.59 minutes and a median of 40 minutes. The boxplot shows that 75% of the data lie within 80 minutes . The time above 100 seems to be the outliers in this data.

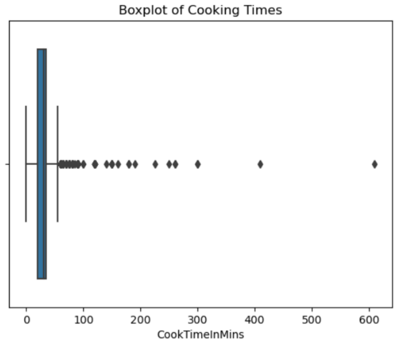
PrepTimeInMins

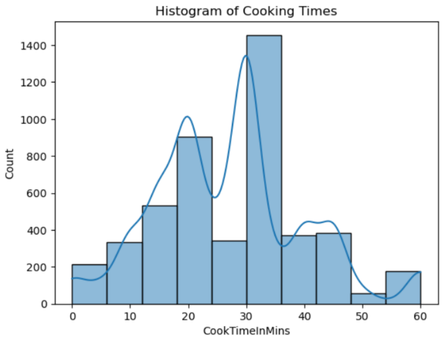




PrepTimeInMins is skewed to the right with a mean of 28.80 minutes and a median of 10 minutes. The boxplot shows that 75% of the data have preparation time less than 40 minutes and the data also comes with extreme outliers.

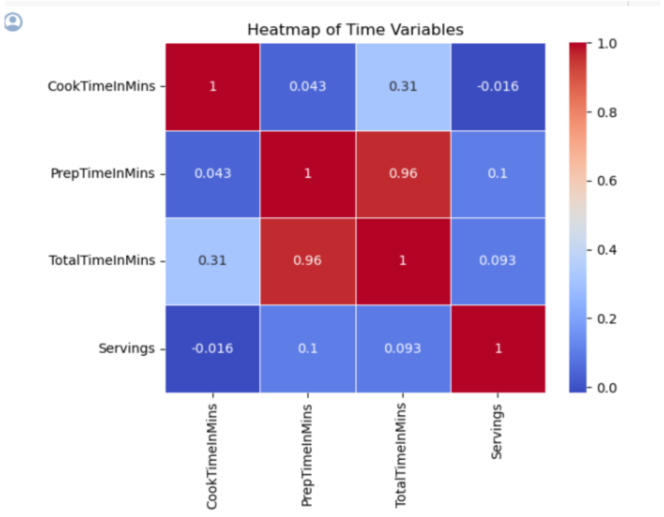
CookTimeInMins

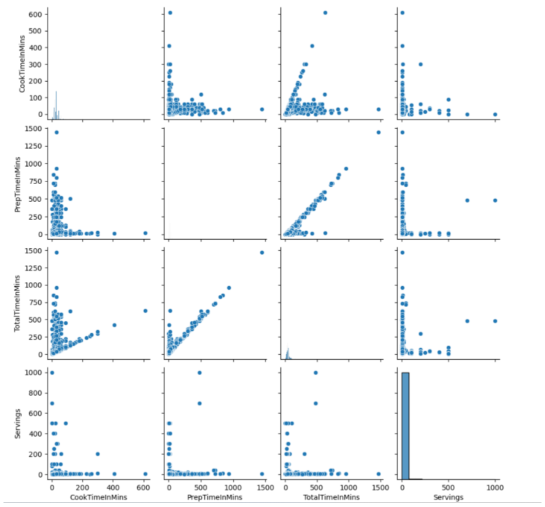




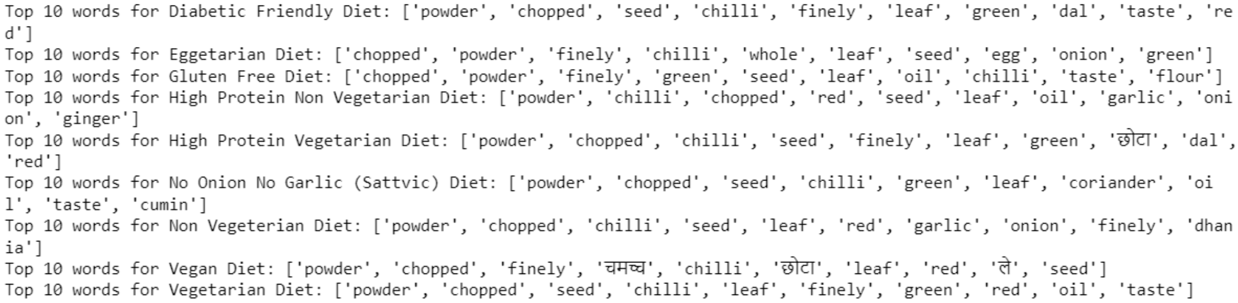
CookTimeInMins is quite normally distributed with a mean of 28.78 minutes and a median of 30 minutes. According to the histogram, the most frequent CookTimeInMins falls within the range of 30 to 35 minutes, indicating that this duration is common for the majority of recipes.

Relationship



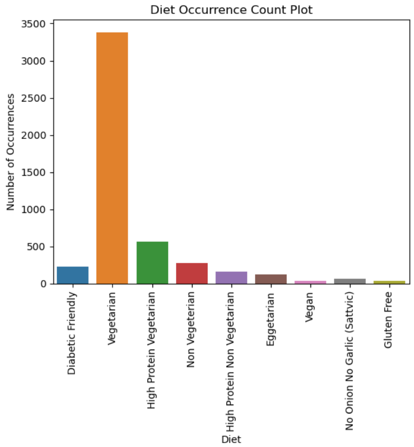


The heatmap and pairplot visualizations highlight an unexpected low positive correlation between servings and total time. The pairplot also suggests an absence of direct proportionality between serving size and total time. It can be concluded that an increase in serving size does not directly lead to an increase in time taken. In fact, there is a negative correlation between servings and cooking time, indicating that as serving size increases, the CookTimeInmins tend to decrease. This phenomenon could be attributed to cooking methods that may be more efficient for larger quantities, resulting in shorter overall cook times. The charts also show that total time is highly correlated with preparation time and total time increases linearly with preparation time, indicating that preparation time contributes more to total time taken compared to cooking time.



```
Top 10 words for Diabetic Friendly Diet: ['powder', 'chopped', 'seed', 'chilli', 'finely', 'leaf', 'green', 'dal', 'taste', 're
d']
Top 10 words for Eggetarian Diet: ['chopped', 'powder', 'finely', 'chilli', 'whole', 'leaf', 'seed', 'egg', 'onion', 'green']
Top 10 words for Gluten Free Diet: ['chopped', 'powder', 'finely', 'green', 'seed', 'leaf', 'oil', 'chilli', 'taste', 'flour']
Top 10 words for High Protein Non Vegetarian Diet: ['powder', 'chilli', 'chopped', 'red', 'seed', 'leaf', 'oil', 'garlic', 'oni
on', 'ginger']
Top 10 words for High Protein Vegetarian Diet: ['powder', 'chopped', 'chilli', 'seed', 'finely', 'leaf', 'green', 'छोटा', 'dal',
'red']
Top 10 words for No Onion No Garlic (Sattvic) Diet: ['powder', 'chopped', 'seed', 'chilli', 'green', 'leaf', 'coriander', 'oi
l', 'taste', 'cumin']
Top 10 words for Non Vegeterian Diet: ['powder', 'chopped', 'chilli', 'seed', 'leaf', 'red', 'garlic', 'onion', 'finely', 'dhan
ia']
Top 10 words for Vegan Diet: ['powder', 'chopped', 'finely', 'चमच्च', 'chilli', 'छोटा', 'leaf', 'red', 'ले', 'seed']
Top 10 words for Vegetarian Diet: ['powder', 'chopped', 'seed', 'chilli', 'leaf', 'finely', 'green', 'red', 'oil', 'taste']
```

The word frequency of individual ingredients within each diet was retrieved as shown above. It is concluded that for almost all diets in Indian cuisines use similar ingredients and cooking methods.



Diet is used as the response variable. However the data is highly unbalanced which causes the models to produce inaccurate results. To address this, we introduced a new column called "Diet_new," which simplifies the categories into two options: "Vegetarian or Vegan" and "Non-Vegetarian." Within the "Non-Vegetarian" category, we included diets such as High Protein Non-Vegetarian, Non-Vegetarian, Gluten-Free, and Diabetic Friendly. The "Vegetarian" category comprises diets like Vegan, Eggetarian, High Protein Vegetarian, No Onion No Garlic (Sattvic) and Vegetarian.

4.2 Results

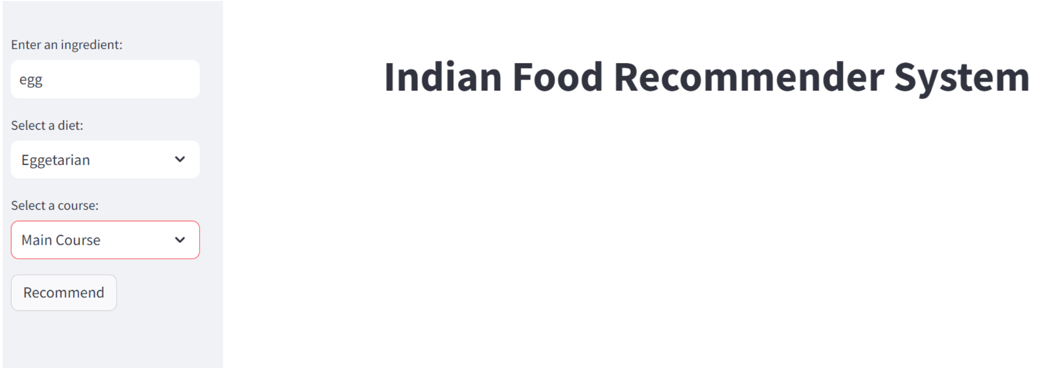
We started the evaluation using 10-fold cross-validation. Moreover, performance was tested through testing datasets and confusion matrix measures.

No	Evaluation Metrics	Support Vector Machine	Logistics Regression	Naïve Bayes	Random Forest	Decision Tree
1	Accuracy	90.49%	81.89%	70.26%	90.29%	**91.07%**
2	Precision	90.30%	83.10%	84.50%	**91.30%**	**91.30%**
3	Recall	90.50%	81.90%	84.80%	90.30%	**91.10%**
4	F-Measure	90.40%	82.30%	84.60%	89.50%	**90.60%**
5	ROC Area	86.70%	83.80%	83.90%	**95.40%**	88.50%

According to the outcomes, the decision tree has performed well in terms of accuracy, precision, recall, and F-Measure. Although the random forest method achieves the highest Area under the ROC Curve, the decision tree still secures the second-highest position in this metric. So, it can be inferred that the Decision Tree model outperforms others in predicting the diet class within this dataset.

4.3 Deployment

The last step is to deploy the food recommender in streamlit where users can enter an ingredient, select their preferred diet and course. The app will show the top three recipes based on your options.
https://indianfoodrecipe-recommendersystem.streamlit.app/



Enter an ingredient: egg

Select a diet: Eggetarian

Select a course: Main Course

CHAPTER 5 CONCLUSION

This study uses a dataset from an Indian recipe website that includes the ingredients of various cuisines and diets with approximately 8,000 recipes. The exploration of this dataset through exploratory data analysis (EDA) applying data visualization has been done before moving on to machine learning. Then, a preprocessing step was performed to clean and prepare the dataset to fit machine learning models. Since the original ingredients column was in textual format, a TF-IDF vector design matrix approach was used to extract features and convert string to vectors before fitting the model. The study observed that Decision Tree yields the best prediction for this problem. This is because decision trees provide a measure of variable importance, helping identify key ingredients that contribute significantly to a recommended recipe or diet. This study can be further utilized to improve the taste of food items, improve their quality attributes and develop more recipes.

REFERENCE

A. Banerjee, a. Noor, N. Siddiqua & M. N. Uddin. (2019). Food Recommendation using Machine Learning for Chronic Kidney Disease Patients. *2019 International Conference on Computer Communication and Informatics (ICCCI)*. https://doi.org/10.1109/ICCCI.2019.8821871

Alshanketi, F. (2023). Machine Learning Model for Predicting the Cuisine Category from a Dish Ingredients. *2023 International Conference on Smart Computing and Application (ICSCA)*, 1-6. https://doi.org/10.1109/icsca57840.2023.10087436

Chen et al. (2020). Eating healthier: Exploring nutrition information for healthier recipe recommendation. *Information Processing and Management, 57*(6), 102051. https://doi.org/10.1016/j.ipm.2019.05.012

Elsweiler, D., Trattner, C., & Harvey, M. (2017). Exploiting Food Choice Biases for Healthier Recipe Recommendation. *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*. https://doi.org/10.1145/3077136.3080826

Habibi, M., & Cahyo, P. W. (2020). Journal Classification Based on Abstract Using Cosine Similarity and Support Vector Machine. *JISKa*, *Vol.4,No.3*(2527–5836).

Jain, H. (2018). *CAPRECIPES: a context-aware personalized recipes recommender for healthy and smart living* [Master's thesis, Uttar Pradesh Technical University]. University of Victoria Libraries. https://dspace.library.uvic.ca/handle/1828/9583

Kardam, S. S., Yadav, P., Thakkar, R., & Ingle, A. (2021). Website on Diet Recommendation Using Machine Learning. *International Research Journal of Engineering and Technology (IRJET)*, *08*(2395–0056).

Khatter et al. (2021). Movie Recommendation System using Cosine Similarity with Sentiment Analysis. 2021 Third International Conference on Inventive Research in Computing Applications (ICIRCA). https://doi.org/10.1109/icirca51532.2021.9544794

Lei et al. (2020). Composing recipes based on nutrients in food in a machine learning context. *Neurocomputing*, 415, 382–396. https://doi.org/10.1016/j.neucom.2020.08.071

M, D. (2022, July 7). *What is cosine similarity and how is it used in machine learning?* Analytics India Magazine. https://analyticsindiamag.com/cosine-similarity-in-machine-learning/

Mana, S. C., & Sasipraba, T. (2021). Research on Cosine similarity and Pearson correlation based recommendation models. *Journal of Physics: Conference Series*, *1770*(1), 012014. https://doi.org/10.1088/1742-6596/1770/1/012014

S. Jayaraman, T. Choudhury & P. Kumar (2017). Analysis of classification models based on cuisine prediction using machine learning. *2017 International Conference on Smart Technologies for Smart Nation (SmartTechCon)*. https://doi.org/10.1109/SmartTechCon.2017.8358611

Shafaat et al. (2022). Food Recipe Recommendation Based on Ingredients Detection Using Deep Learning. *2022 Association for Computing Machinery*, 191-198. https://doi.org/10.1145/3542954.3542983

Xie, W., & Lou, H. (2022). Implementation of key technologies for a Healthy Food Culture Recommendation System using Internet of things. *Mobile Information Systems*, 2022, 1–12. https://doi.org/10.1155/2022/9675452