Advanced Machine Learning Group Project

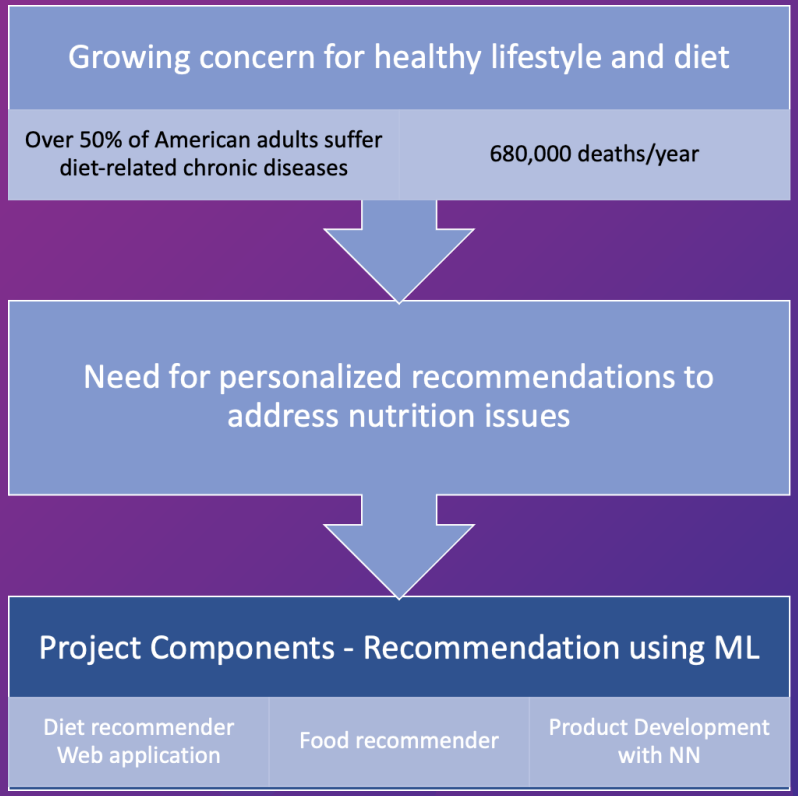
Amey, Evin, Jyotis, Biyun, Gaytri

TABLE OF CONTENTS

01

PURPOSE

Project Purpose



02

DATA INFORMATION

Data Info and Cleaning

- Data Description
  - Two sets from Kaggle: Recipes and Reviews
  - Recipes
    - Over 500,000 recipes from 312 different categories
    - 28 columns
      - Cooking time, ingredients/instructions, servings, nutritional content, etc.
  - Reviews
    - Over 1.4 million reviews from 270,000+ users
    - Info:
      - 8 columns
      - Review (text), rating, recipe ID, author, etc.

- Data Cleaning
  - Not much (Woohoo!)
  - Drop NAs
  - Formatting
    - Date column

Data Info

```
data_recipes
```

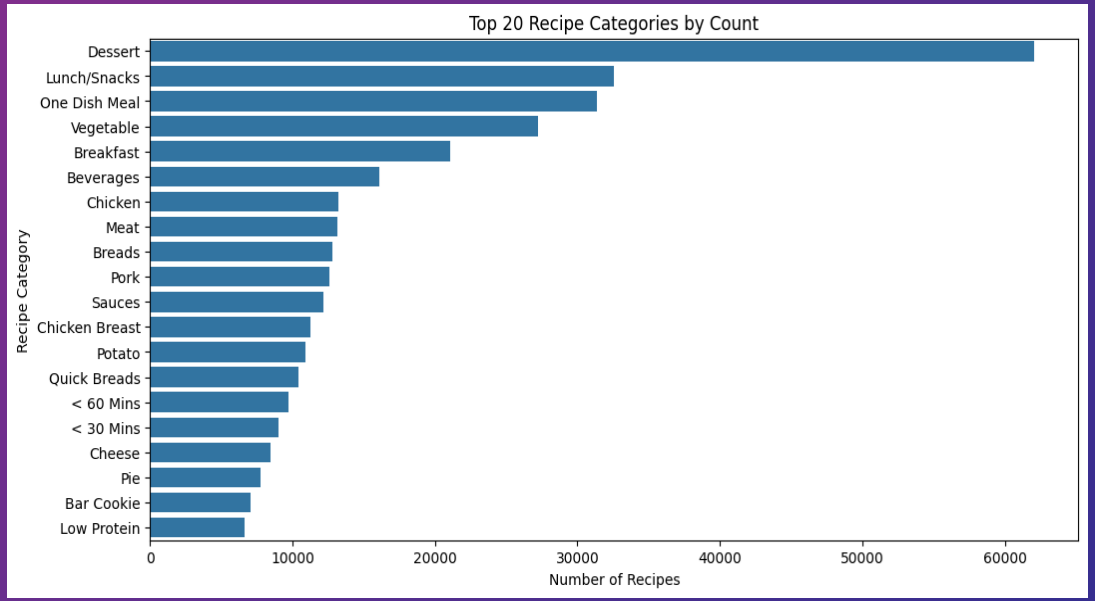
	RecipeId	Name	AuthorId	AuthorName	CookTime	PrepTime	TotalTime	DatePublished	Description
3	41	Carina's Tofu-Vegetable Kebabs	1586	Cyclopz	PT20M	PT24H	PT24H20M	1999-09-03 14:54:00	This dish is best prepared a day in advance to...
5	43	Best Blackbottom Pie	34879	Barefoot Beachcomber	PT2H	PT20M	PT2H20M	1999-08-21 10:35:00	Make and share this Best Blackbottom Pie recip...
16	54	Carrot Cake	1535	Marg CaymanDesigns	PT50M	PT45M	PT1H35M	1999-09-13 15:20:00	This is one of the few recipes my husband ever...
26	64	Almond Pound Cake	125579	GrandmaIsCooking	PT1H	PT15M	PT1H15M	1999-08-07 16:33:00	Make and share this Almond Pound Cake recipe f...
54	94	Blueberry Buttertarts	1556	Strawberry Girl	PT25M	PT15M	PT40M	1999-09-12 05:46:00	Make and share this Blueberry Buttertarts reci...
...	...	...	...	...	...	...	...	...	...
		Pork							Saw this on

03

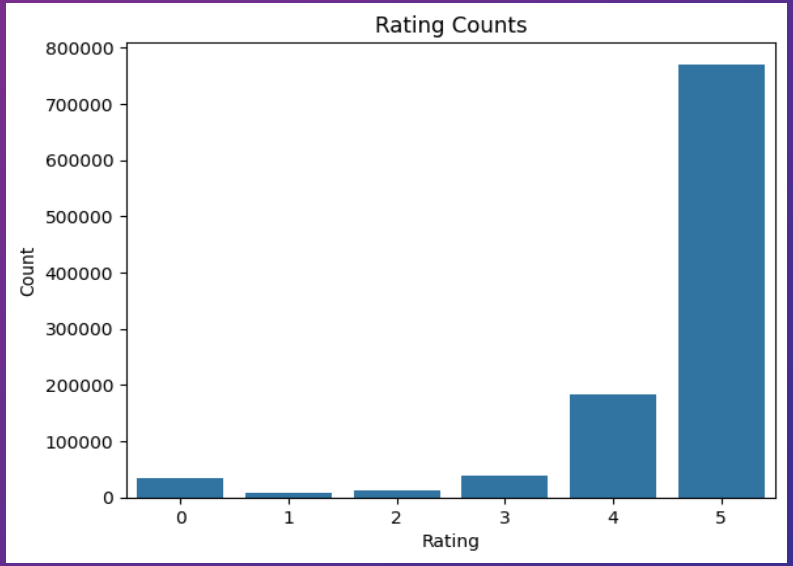
EDA

EDA

- Top 2 categories:
  - Dessert
  - Lunch Snacks
- People struggle with stopping unhealthy choices and consistency
  - “65% of dieters return to pre-diet weight within three years.” - LiveStrong
- Healthier options for unhealthy cravings



EDA

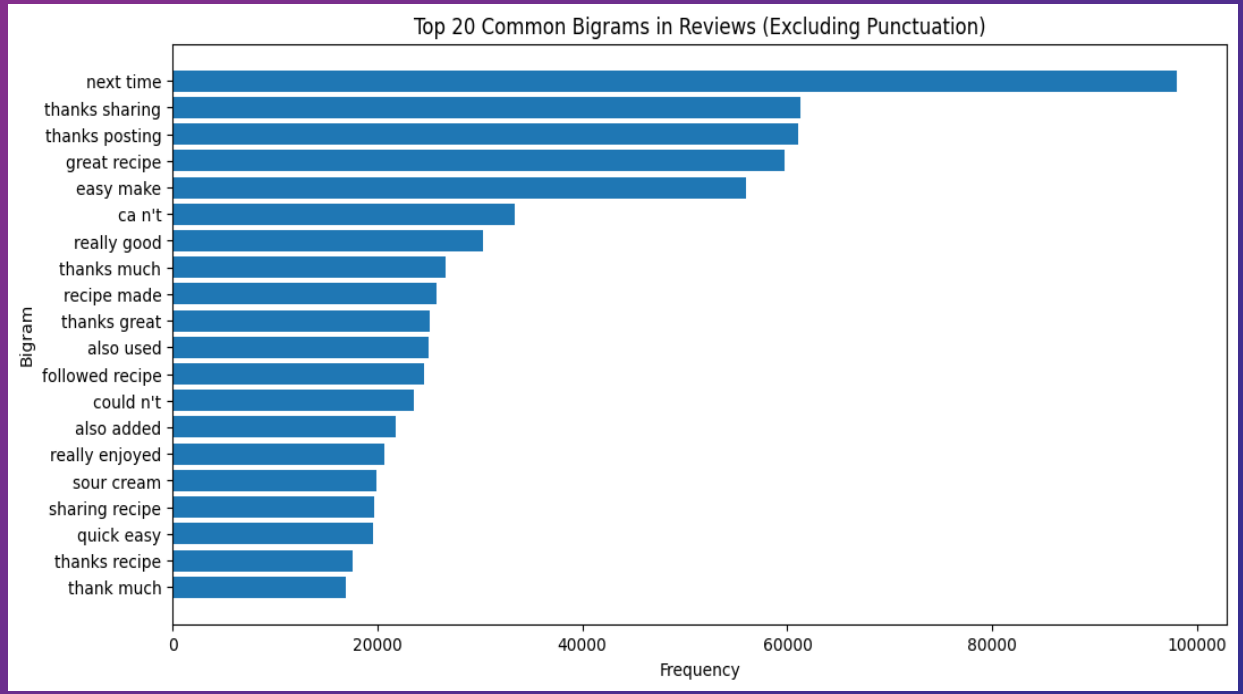


- Recipes are designed for limited cooking experience, but for everyone
- Easy to cook
- Non-expensive ingredients

- People clearly like them!
- 80%+ success rate with over 1 million reviews

EDA

- Top 20 most common bigrams
  - Not very telling
- Confirms previous EDA
  - People want easy recipes
  - Overall quality reviews



04

RECOMMENDER

What is a food recommendation engine?

- Personalized Nutritional Recommendations: Utilizes nutritional content and ingredients of foods to provide personalized recommendations, catering to individual dietary needs and preferences.
- Consideration of Dietary Restrictions: Takes into account specific dietary restrictions and preferences, such as allergies or personal food choices, ensuring safe and suitable options.
- Discovery of New and Nutritious Foods: Helps users explore a variety of healthy food options, expanding their dietary choices and combating food boredom.
- Promotion of Healthy Eating Habits and Positive Impact on Long-term Health.

What is content-based recommendation engine?

- A content-based recommendation system leverages the intrinsic properties of items, such as text, images, or audio, to suggest similar items to users. It identifies and utilizes patterns or attributes related to the items for making recommendations.
- Advantages of a Content-Based Approach:
  - Independent Recommendations: It can generate suggestions without needing other users' data, making it ideal for new users or items.
  - Personalized Relevance: The recommendations are closely aligned with the user's individual preferences and interests.
- Challenges in Content-Based Systems:
  - Limited Variety: Tends to recommend items similar to those already liked by the user, potentially leading to a lack of diversity in suggestions.
  - Scalability Issues: Handling large and diverse datasets can be challenging.
  - Dependency on Accurate Content: The effectiveness of the system relies on the correctness and consistency of item attributes.

Comparison of Tree-Based Algorithms in Nearest Neighbors

Feature	BallTree	KDTree	Brute-Force
Data Structure	Organizes data in a tree based on 'balls'.	Organizes data in a tree using 'k-dimensional' splits.	Computes distances between all pairs, no additional structure.
Efficiency in High-Dimensional Data	Less efficient compared to KDTree.	Efficiency decreases significantly as dimensionality increases.	Not affected by dimensionality, but computationally intensive.
Speed with Large Datasets	Generally faster than brute-force for large datasets.	Faster than BallTree for low-dimensional data.	Slowest, especially as the size of the dataset increases.
Best Use Cases	Better for datasets with moderate to high dimensions.	Best for datasets with lower dimensions (fewer features).	Effective for small datasets with any dimensionality.
Complexity	Moderately complex, involves choosing appropriate metrics.	Similar to BallTree in terms of setup complexity.	Simplest, no additional setup required.

Comparison of kNN and FAISS

Feature	kNN (k-Nearest Neighbors)	FAISS (Facebook AI Similarity Search)
Algorithm Type	Traditional machine learning algorithm for neighbor searches.	Advanced library for efficient similarity search in high-dimensional data.
Scalability	Less scalable, performance decreases with larger datasets.	Highly scalable, designed for large-scale data.
Speed	Slower on large datasets due to computational intensity.	Faster, uses optimized algorithms and data structures.
Complexity	Simpler to implement and understand.	More complex to set up, requires understanding of indexing and quantization.
Suitability for High-Dimensional Data	Less suitable, as distance calculations become complex.	Highly suitable, optimized for high-dimensional vector search.
Best Use Cases	Ideal for small to medium datasets and simpler applications.	Best suited for large datasets and performance-critical applications.
Customizability	Limited to distance metrics and k value.	Offers advanced customization options like different indexing strategies.

FAISS

- The class is initialized by taking a DataFrame (`nutritional_df`) and a list of columns (`cols_to_divide`) that represent the nutritional attributes of recipes.
- Then it is normalized using nutritional attributes using `MinMaxScaler` to ensure that each feature contributes equally to the distance computations.
- These normalized features are then converted into a NumPy array of type `float32`, which is a requirement for compatibility with FAISS.
- The FAISS index (`IndexFlatL2`) is created and trained with these vectors. This index is used for efficient nearest neighbor searches using the L2 (Euclidean) distance.
- The `find_closest_recipes` method allows finding the `k` closest recipes to a given recipe based on their nutritional content, excluding the recipe itself from the recommendations.
- Note: The `FAISSRecipeRecommender` class as written is a straightforward implementation of a nearest neighbor search using FAISS, focusing on accuracy and simplicity with `IndexFlatL2`. It does not involve the complexity of quantization-based indexing methods, which are generally used for larger scale or more performance-critical applications.

Closest recipes for RecipeId 41.0 from the original nutritional_df:

	RecipeId	Name	Calories	FatContent	SaturatedFatContent	CholesterolContent	SodiumContent	CarbohydrateContent	FiberContent	SugarContent	ProteinContent	Reci
515719	534346.0	Lower Carb Bread/Hogie Roll	434.500000	11.800	1.550	0.000	594.550000	70.000	8.700	0.850	14.900000	
373563	387075.0	Banana Almond Date Soy Smoothie	340.400000	12.600	1.300	0.000	184.200000	48.000	8.600	21.600	15.600000	
447139	463666.0	David Lynch's Quinoa	383.500000	10.000	1.300	0.000	622.900000	60.600	8.300	1.600	14.600000	
388519	402507.0	Middle Eastern Tahini Oatmeal (Vegan)	452.100000	11.800	1.600	0.000	134.400000	74.300	8.400	36.200	15.600000	
196841	205476.0	Muesli With Yogurt and Cashews	413.300000	12.900	2.300	0.000	72.700000	64.300	9.200	22.600	15.500000	

Automatic Diet Recommendation

Modify the values and click the Generate button to use

Age

25 − +

Height(cm)

187 − +

Weight(kg)

73 − +

Gender

- (•) Male
- ( ) Female

Activity

Light exercise

Little/no exercise — Extra active (very active & physical job)

Choose your weight loss plan:

Maintain weight

BMI CALCULATOR

Body Mass Index (BMI)

20.88 kg/m^2

Normal

Healthy BMI range: 18.5 kg/m^2 - 25 kg/m^2.

CALORIES CALCULATOR

The results show a number of daily calorie estimates that can be used as a guideline for how many calories to consume each day to maintain, lose, or gain weight at a chosen rate.

Maintain weight	Mild weight loss	Weight loss	Extreme weight loss
2446 Calories/da	2201 Calories/da	1957 Calories/da	1467 Calories/da
↓ -0 kg/week	↓ -0.25 kg/week	↓ -0.5 kg/week	↓ -1 kg/week

Generating recommendations...

DIET RECOMMENDATOR

Recommended recipes:

BREAKFAST	LUNCH	DINNER
Arroz C'atum	Crunchy Tuna Cat Treats	Venison Soup
Baked Tuna and Zucchini Risotto	Pinot Noir Beef Stew	Modern Venison Roast
Chicken & Pasta in Peanut Sauce	Modern Venison Roast	Grilled Tuna With White Bean and Charred Onion Salad
Tuna Potato Spicy Scratch	Halibut and Bean Stew	Easy Thai Chicken Wraps
Linguine with Salmon and Mushrooms	Poisson En Papillote, Victorian Style	Tuscan Halibut for Two

Recommendation Generated Successfully !

Choose your meal composition:

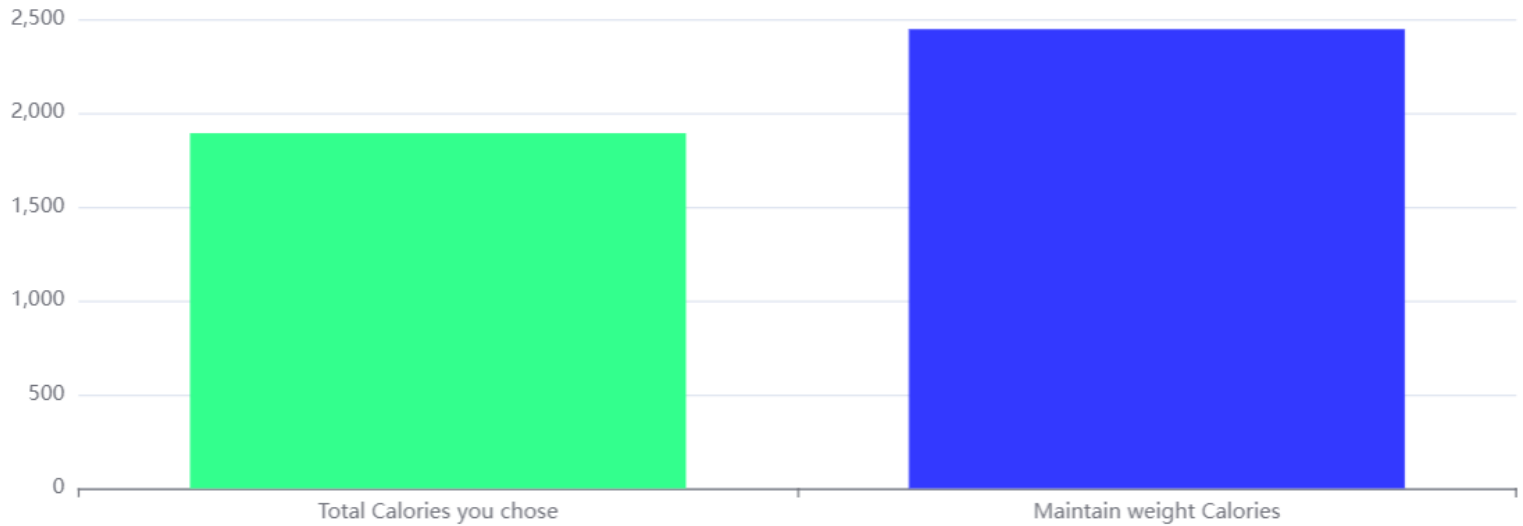
Choose your meal composition:

Choose your breakfast: Arroz C'atum

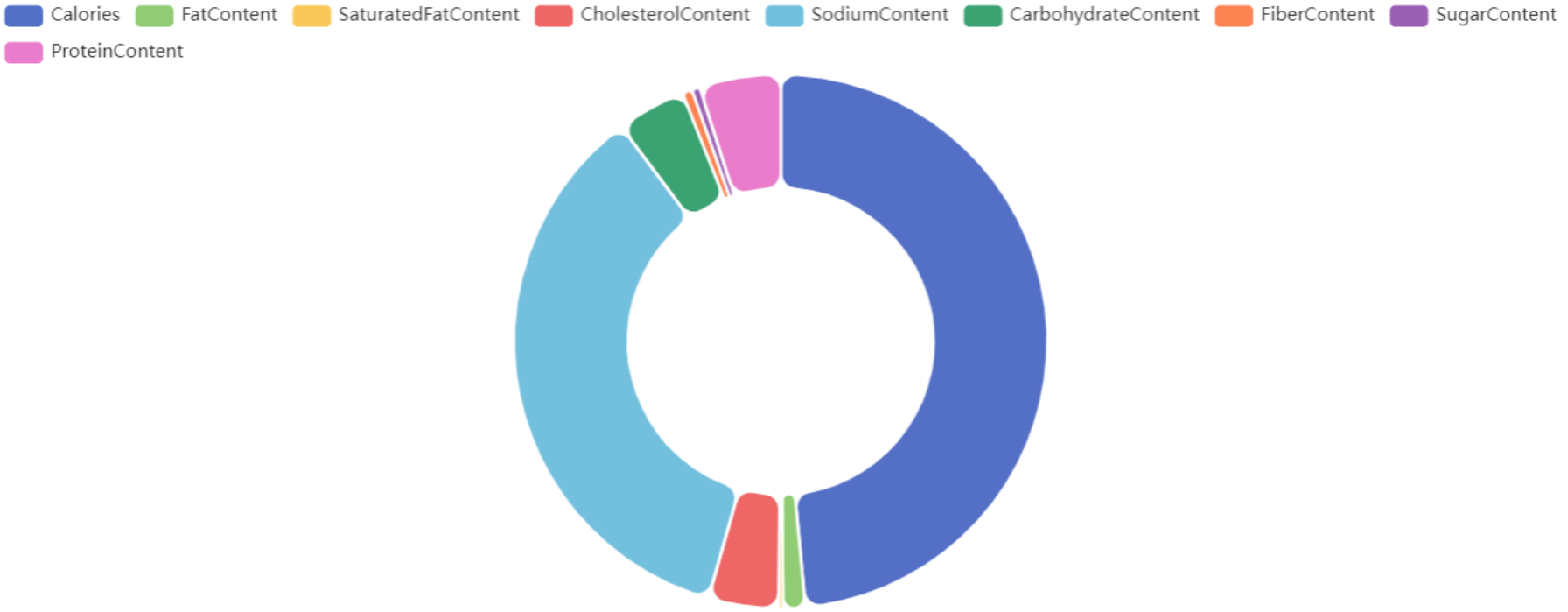
Choose your launch: Crunchy Tuna Cat Treats

Choose your dinner: Venison Soup

Total Calories in Recipes vs Maintain weight Calories:



Nutritional Values:



05

NEURAL NETWORK

Motivation Behind Using Neural Network

- Enables more sophisticated, personalized, and comprehensive approach to dietary recommendations
- Offers several advantages:
  - Captures **complex relationships** present in data (dietary preferences impacted by variety of factors)
  - Can be tailored to individuals (whole different level of **personalization)**
  - **Adapts** to new data to ensure relevant recommendations that resonate with changing user needs
  - **Models non-linear functions** and handles intricate nature of data well
  - Efficiently integrates and processes multi-modal data, since diet systems often include diverse forms of data

Background

- Goal to develop an innovative neural network model to accurately **predict the protein content of food products**
  - This model will analyze:
    - Nutritional values, including calories, fat, saturated fat, cholesterol, sodium, carbohydrates, fiber, sugar content, etc.

- Leverage nutritional data for **product development**
  - Can strategically use protein content information from model in marketing techniques and introducing new, protein-rich products to the market, catering to health-conscious consumers

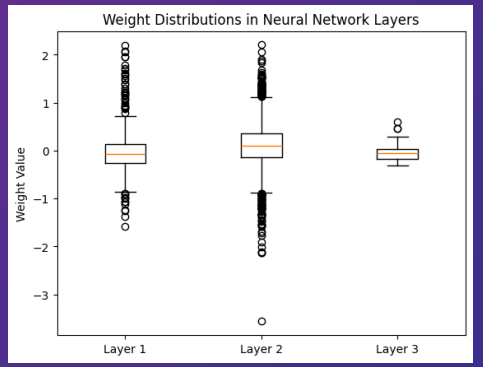
- Emphasis on **predicting protein content,** enabling us to innovate and reformulate products to meet specific nutritional goals

- Introduction of high-protein products aligns with the broader goal of aiding in **obesity reduction**
  - Protein is crucial for fat loss and maintaining a healthy weight, making these products particularly appealing to those on calorie-restricted diets

  - *Clinical Evidence Supporting High-Protein Diets:* Research, such as a 12-month study involving 130 overweight individuals on calorie-restricted diets, shows that high-protein groups can lose significantly more body fat compared to groups consuming normal protein levels. This scientific backing reinforces the market potential for our new protein-rich products. (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2129146/)

Code for Implementation of Neural Network

- Defines model with **two hidden layers** and an **output layer for regression**
- Compiles model with **Adam optimizer** and **mean squared error loss function**
- Trains model on the standardized training data for *50 epochs* and a *batch size of 32*
- Evaluates model on test set
- Computes **mean squared error** as a measure of predictive accuracy



Results of Code

- The training loss decreases over epochs, which is expected as the model learns from the training data.
- Around epoch 4, validation loss sharply decreases, but improvement is not sustained (loss increases in subsequent epochs)
    - Suggests that the model may be overfitting to the training data.
        - Can add dropout layers
        - Reduce model complexity
        - Adjust hyperparameters to improve generalization to unseen data
- **Final MSE** of 0**.0643** on test set indicates strong ability to generalize on unseen data, although there is room for improvement

Product Innovation with Higher Protein Content

- New Product Development
  - Scenario: Developing a new cookie.
    - Current Cookie: 1.73g protein, high sugar, and fat.
    - Our Innovation: New cookie with 10g protein, balancing sugar and fat content
- Based on our analysis, we suggest a higher protein content for a new product, enhancing its nutritional value.
- We aim to transform a popular dessert associated with obesity into a healthier version, catering to consumer preferences for both taste and health.
- Leveraging our neural network predictive capabilities for developing a groundbreaking product, meeting market demands for higher protein content and addressing health concerns.

Cookie use case

- Market potential (m): 50,000 potential customers
- Coefficient of innovation (p): 0.02 (2% of the market potential will try the product due to innovation each period)
- Coefficient of imitation (q): 0.15 (15% of the potential market will try the product due to imitation effects each period)

Year	Innovators (p×m)	Previous Cumulative Adopters (N(t-1))	Imitators [(q - p) × N(t-1) / m × (1 - N(t-1) / m)]	New Adopters (Sales)	Cumulative Adopters
1	1,000	0	0	1,000	1,000
2	1,000	1,000	1,350	2,350	3,350
3	1,000	3,350	2,810	3,810	7,160
4	1,000	7,160	3,443	4,443	11,603
5	1,000	11,603	3,032	4,032	15,635

06

ANALYSIS AND FINAL CONCLUSIONS

Analysis and Final Conclusions

Content-based diet recommender using algorithms

KNN, NN, FAISS to tailor to diverse dietary needs

Leveraging FAISS scalability & efficiency

User-centric, health-oriented solutions

High satisfaction rate of recipe database

Potential impact of our system in promoting healthier dietary choices

Scope for enhancing the recommender system by exploring advanced indexing methods in architectures

Diversity in recommendations

Refine the recommendation accuracy

User feedback

Valuable tool in guiding users towards healthier and more personalized dietary choices

Recommended diet

Develop comparable food with more protein

THANKS!

DO YOU HAVE ANY QUESTIONS?

07

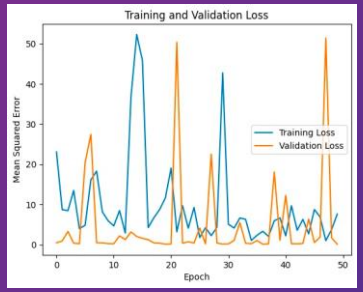
Appendix

References

- https://www.kaggle.com/datasets/irkaal/foodcom-recipes-and-reviews?select=recipes.csv
- https://www.cs.cmu.edu/~ckingsf/bioinfo-lectures/kdtrees.pdf
- https://en.wikipedia.org/wiki/K-d_tree
- https://en.wikipedia.org/wiki/Ball_tree#:~:text=The%20ball%20tree%20nearest%2Dneighbor,near
- est%20points%20encountered%20so%20far.
- https://www.kdnuggets.com/2020/10/exploring-brute-force-nearest-neighbors-algorithm.html
- https://oneapi-src.github.io/oneDAL/daal/algorithms/k_nearest_neighbors/k-nearest-neighbors-knn
- -classifier.html
- https://www.kaggle.com/datasets/taniaj/cryptocurrency-market-history-coinmarketcap
- https://github.com/zakaria-narjis/Diet-Recommendation-System/tree/main
- https://www.kaggle.com/code/agnishdutta/recipe-recommender
- https://www.livestrong.com/article/13764583-diet-statistics/
- CHATGPT

Models

1. BallTree :- $$d(x, y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$
1. KDTree:- KDTree also typically uses Euclidean distance, although other metrics can be applied. The distance calculation is the same as for BallTree.
1. Brute-Force Algorithm:- $$\text{cosine similarity}(x, y) = \frac{x \cdot y}{\|x\|\|y\|} = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2}\sqrt{\sum_{i=1}^{n} y_i^2}}$$
1. kNN:- Typically uses Euclidean distance, although other metrics can be applied. The decision is based on a majority vote or averaging of the k closest points.
1. FAISS
1. Neural Network:-



Findings from Neural Network

- Loss on validation set surpasses that of training set in certain epochs
  - Implies **overfitting** and less generalization to new, unseen data
- Loss on training set consistently decreases during training process
  - Demonstrates that model is indeed learning from data provided
- Essential to **monitor additional metrics** other than loss values
- **Visualizing training/validation loss trends over epochs** will help in analyzing model performance
- **Investigating regularization techniques** or **adjusting hyperparameters** may improve generalization