CONVERSATIONAL AI: ACCELERATED DATA SCIENCE [ADVANCED]-UCS622-2324EVESEM

Nutritional Values for Common food analysis

Submitted by:

102103227 Anchal

BE Third Year-

3COE 15

Submitted to:

Dr. Arun Singh Pundir

THAPAR INSTITUTE
OF ENGINEERING & TECHNOLOGY
(Deemed to be University)

Computer Science and Engineering Department

Thapar Institute of Engineering and Technology, Patiala

April 2024

TABLE OF CONTENTS

Abstract

This project aims to address the public health concern of poor dietary choices in India by creating a comprehensive database of nutritional values for common Indian foods. The database will serve as a centralized repository of accurate and reliable data, providing transparency regarding the nutritional value of foods and helping to promote balanced diets and encourage healthier food choices. The project utilizes machine learning algorithms, including linear regression, random forest, and decision tree regression, to predict missing nutritional values based on available features. The algorithms provide insights into the relationship between input features and the target variable, and can handle both linear and non-linear relationships. The performance of each algorithm is evaluated using metrics such as mean squared error (MSE) or R-squared on a separate test set.

This centralized and user-friendly resource empowers individuals to make informed dietary decisions. Furthermore, the project utilizes a Python library, PuLP, to implement a Simplex optimization algorithm. This allows us to create personalized meal plans that minimize calorie intake while maintaining a balanced diet and meeting specific protein, carbohydrate, and fat requirements. By promoting healthier food choices and improved nutrition, this project has the potential to contribute significantly to better public health outcomes in India.

Introduction

2.1 Problem description

Undoubtedly, diet plays a critical role in overall health and well-being. However, according to the Times of India, India has the second highest number of deaths globally caused by poor dietary choices [India is No. 2 among countries with most deaths caused by poor food choices]. This alarming statistic highlights a significant public health concern in India. Fluctuations in nutrient intake can significantly increase the risk of developing chronic diseases.

2.2 Problem Challenges

Several challenges contribute to the inadequate dietary patterns observed in India:

- **Lack of awareness regarding balanced diet:** A large portion of the population may not possess a thorough understanding of the principles of a balanced diet. This knowledge gap can hinder individuals from making informed choices about the foods they consume.

- **Difficulty in accessing accurate and reliable information on nutritional value of common foods:** In some cases, readily available information on the nutritional content of common Indian foods may be scarce or unreliable. This lack of transparency makes it difficult for individuals to assess their dietary intake.

- **Limited diversity in dietary patterns:** Restricted dietary patterns can lead to deficiencies in essential nutrients. This may be due to factors such as limited access to a variety of fresh fruits and vegetables, or cultural preferences that emphasize certain food groups over others.

- **Influence of marketing and advertising promoting unhealthy food choices:** Marketing and advertising campaigns can significantly influence food choices. The aggressive promotion of unhealthy processed foods can overshadow healthier alternatives.

2.3 Novelty in Work

This project has the potential to address these challenges through the creation of a comprehensive database of nutritional values for common Indian foods. This database can offer several unique contributions:

- **Centralized resource:** The database will serve as a centralized repository of accurate and reliable data on the nutritional content of commonly consumed foods in India.

- **Accessibility:** The data will be made readily accessible through a user-friendly platform. This ease of access will empower individuals to make informed decisions about their dietary choices.

- **Promoting balanced diets:** By providing transparency regarding the nutritional value of foods, the database can help educate the population about balanced diets and encourage healthier food choices.

- **Improved public health outcomes:** The project has the potential to contribute significantly to improved public health outcomes in India by promoting better nutrition and reducing the prevalence of chronic diseases.

3. Literature Survey

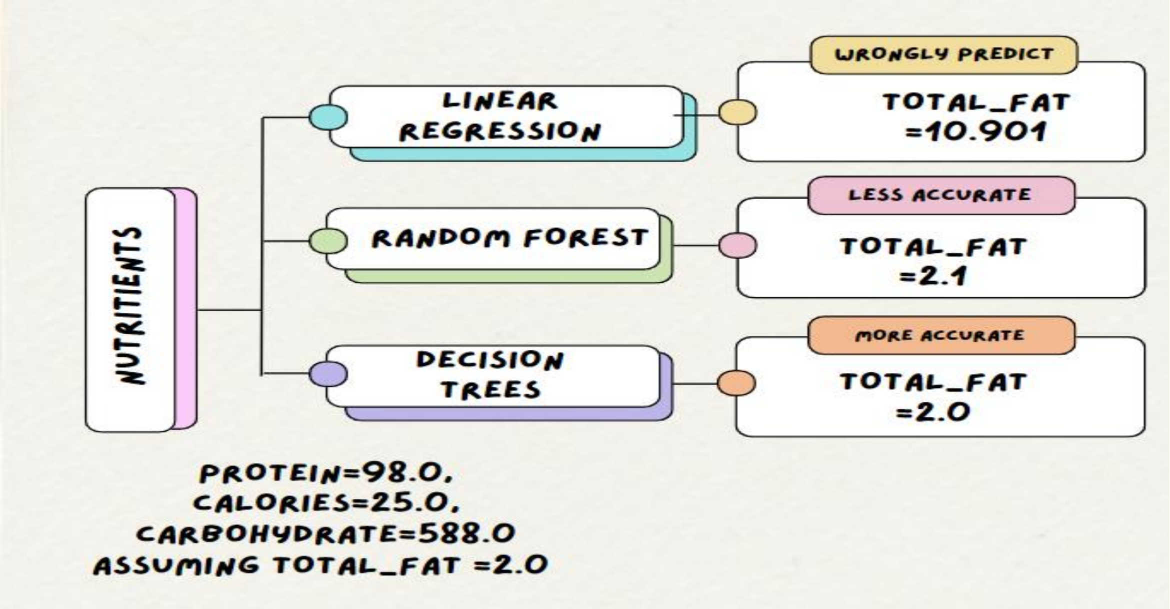
S.NO	YEAR	Author(s)	Focus of the paper	Key point in converge	Technique used	Parameters analysed
1.	2007	M. Burke	This paper discusses the nutritional goals and dietary strategies for athletes.	The paper highlights the importance of carbohydrate intake for athletes, particularly for those involved in endurance sports. It also emphasizes the need for adequate protein intake to support muscle repair and growth.	The paper is a review of existing literature on sports nutrition.	The paper analyzes the role of macronutrients (carbohydrates, proteins, and fats) and micronutrients (vitamins and minerals) in athletic performance.
2.	2021	S. M. Soltanizadeh, M. R. Kordi, M. H. Pak, and M. A. Pak	This paper investigates the nutritional risks among adolescent athletes with disordered eating.	The paper finds that adolescent athletes with disordered eating are at risk of nutritional deficiencies, menstrual abnormalities, and decreased bone mass density.	The paper is a review of existing literature on adolescent athletes and disordered eating.	The paper analyzes the relationship between disordered eating, nutritional deficiencies, menstrual abnormalities, and bone mass density in adolescent athletes.
3.	2016	Li, M., Li, X., & Yang, Y.	Deep learning and reinforcement learning convergence.	Deep neural networks as function approximators in reinforcement learning.	DQNs, DDPG, A3C, PPO.	Network architecture, exploration-exploitation, reward shaping, memory replay.

4.	1982	Montgomery, D. C., & Peck, E. A	Linear regression models and their analysis.	Estimation, testing, and prediction in linear regression.	Ordinary least squares, maximum likelihood estimation, hypothesis testing.	Coefficients, residuals, multicollinearity, heteroscedasticity.
5.	2004	Kutner, M. H., Nachtsheim, C. J., & Neter, J.	Linear regression models and their analysis.	Model building, specification, and validation.	Hypothesis testing, confidence intervals, model selection criteria.	Coefficients, multicollinearity, influential observations, residuals.
6.	2022	Xiaotong Yuan, Xin Wang, Xiaolin Hu	The paper proposes an adaptive gradient method with dynamic bounds on the learning rate for deep learning.	The dynamic bounds on the learning rate help achieve faster convergence and better generalization performance.	AdaBoost, an optimization algorithm that uses a cosine annealing schedule to adjust the learning rate.	Learning rate, batch size, number of epochs, and comparison with other optimization algorithms.

4. Methodology

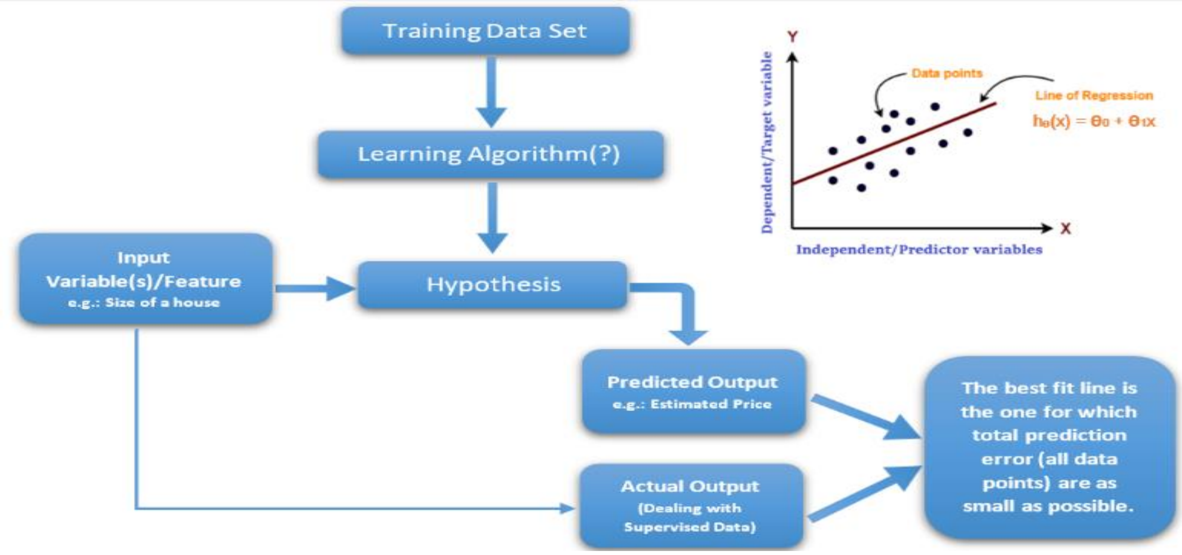
4.1 Machine Learning

Data Flow Diagram



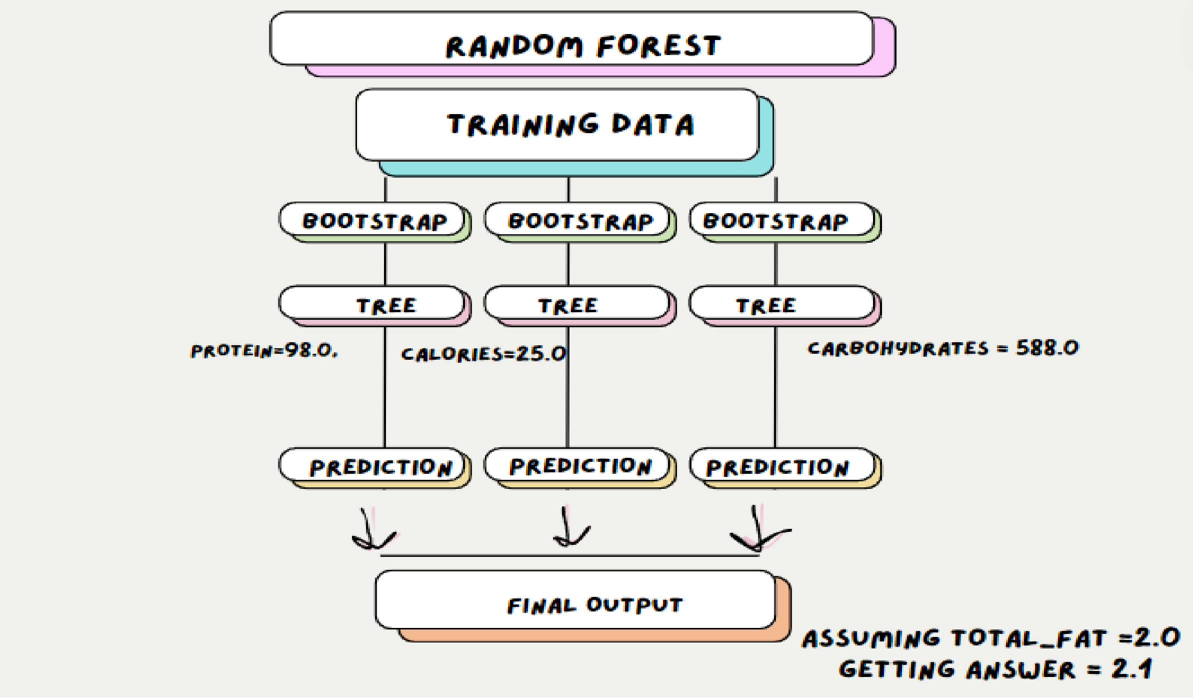
Here's a brief explanation of why these three algorithms are used:

Linear Regression


Training Data Set
Learning Algorithm(?)
Input Variable(s)/Feature e.g.: Size of a house
Hypothesis
Predicted Output e.g.: Estimated Price
Actual Output (Dealing with Supervised Data)
The best fit line is the one for which total prediction error (all data points) are as small as possible.
Y
X
Data points
Line of Regression
hθ(x) = θ0 + θ1x
Dependent/Target variable
Independent/Predictor variables

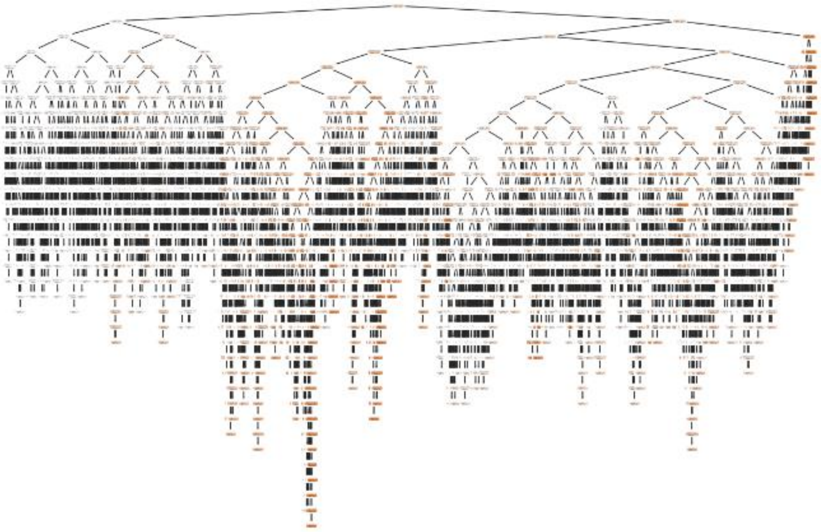
Linear Regression: This is a simple and interpretable algorithm that can provide insights into the relationship between the input features and the target variable. In this case, it can help in understanding how the protein, calories, and carbohydrate content are related to the total fat content.

Random Forest



Random Forest: This is an ensemble algorithm that combines multiple decision trees to provide a more accurate and robust prediction. It can handle non-linear relationships and interactions between the features, and is less prone to overfitting.

Decision Trees



This is a simple and interpretable algorithm that can provide insights into the decision-making process of the model. It can handle both numerical and categorical features, and is easy to visualize and interpret.

In this project, linear regression, random forest, and decision tree regression are used to predict the total fat content in a given set of features (protein, calories, and carbohydrate) for common Indian foods. The goal is to create a comprehensive database of nutritional values for these foods, and these machine learning algorithms can help in predicting the missing nutritional values based on the available features.

As for why the predicted output for the given input features is different for each algorithm, it's because each algorithm uses a different approach to model the relationship between the input features and the target variable.

Linear regression assumes a linear relationship between the features and the target variable, and may not capture the complex relationships that exist in real-world data. **Random forest**, on the other hand, uses multiple decision trees to capture these complex relationships, but may overfit the data if the number of trees is too high. **Decision tree**, being a simple algorithm, may not capture the nuances of the data as well as random forest, but can still provide a reasonable prediction.

As for which algorithm gives the correct output, it's important to note that the true total fat content for the given input features is not known. Therefore, we cannot determine which algorithm is the most accurate. However, we can evaluate the performance of each algorithm using metrics such as mean squared error (MSE) or R-squared on a separate test set.

Based on the provided code, it seems that the true total fat content for the given input features is 2.0. However, without evaluating the performance of each algorithm on a separate test set, we cannot determine which algorithm is the most accurate. It's possible that all three algorithms may have similar performance, or that one algorithm may perform better than the others. Further analysis is needed to determine this.

4.1 Mathematical

We should get:

1. **One gram of protein per day per kg** of the person (e.g. 70g of proteins per day if you’re a 70 kg person)
2. The **carbohydrates** should be **half of the calories** you assume per day.
3. The remaining part should be **fat**

So the proportion of **fat, carbohydrates,** and **proteins** for a 70 kg person with 1500 calories should be something like this:

Now actually, rather than fix the number of calories, we are more interested in keeping the calories as small as possible because we have a certain number of **grams** of protein, carbohydrates, and fat.

This is where the **optimization algorithm** comes in. We will have an indicative number of calories that we want to achieve, but we will rather try to **minimize it**, keeping in mind that we **strictly** have to satisfy the number of proteins, carbohydrates, and fat that we have in our plan.

Let’s formalize it.

2. About Optimization

The following is a technical definition of optimization:

An algorithm is an **optimization** if it finds the minimum or maximum of a function in a given domain.

For example, let’s take the following function:

In this case, the **domain** of the function is the interval [-1, 1]. The minimum of the function in that domain is x = 0 and y = 0. We want an algorithm that can **find the minimum.**

our goal at this point is to optimize the number of calories. In particular, we want to **minimize them**. At the same time, we do want our nutrition plan **to be varied,** and we want it to satisfy the requirements of protein, carbohydrates, and fat that we want to achieve.

As we said, we are trying to **minimize** a cost-**loss** function. To solve this task using the following method, the loss function has to be **linear**.
This means that given a set of variables **x,**

$$\mathbf{x} = \{x_1, x_2, ..., x_n\}$$

Image by author

Our loss function will be something like that:

$$L(\mathbf{x}) = \mathbf{c}^T \cdot \mathbf{x} = \sum_{i=1}^{N} c_i x_i$$

Image by author

where **c** is a fixed vector of costs.
What is **x** in our case? And what is **c**? Well, as we have to **minimize the calories,** it is pretty evident that **c** has to be the vector of calories that each food has, let's say, in 100 grams. And as we said, **it is fixed***.

*I'd love for it not to be fixed, and I'd love to change it to 0.5 calories per kg of chocolate, but unfortunately, it's not like this.

Fantastic. Now, what is **x**? If you followed the idea, I think it is clear that now, let's say, x_1 has to be the **units** (relative to 100 g) of "food number 1" that we decide to eat.

Great. So let's say that I tell you to minimize L(x), and I tell you **nothing else.** Well, in that case, the answer is pretty simple: **x = 0** and L(0) = 0. But I heard that it is challenging to survive without eating.

This means that we have to minimize the function, **provided** that we respect **certain constraints**. These constraints are, as we said, the number of grams of **protein, carbohydrates,** and **fat** that we need to put in our bellies.

In a similar way to what we have done before, we will now describe **three** other vectors, which we will call:

- **f**, that is the vector of **fat** (e.g. 100 grams of food number 1 have f_1 grams of fat)

$$\mathbf{f} = \{f_1, \ldots, f_n\}$$

Image by author

- **e,** that is the vector of **carbohydrates** (e.g. 100 grams of food number 1 have c_1 grams of carbohydrates)

$\mathbf{e} = \{e_1, ..., e_n\}$I didn’t use **c** because I already have it in my loss function. I know. Poor business move. **p,** that is the vector of **proteins** (e.g. 100 grams of food number 1 have p_1 grams of proteins)

$$\mathbf{p} = \{p_1, ..., p_n\}$$

Image by author

And the **constraints** look like this:

$$\begin{cases} \sum_{i=1}^{N} e_i x_i \geq E \\ \sum_{i=1}^{N} p_i x_i \geq P \\ \sum_{i=1}^{N} f_i x_i \geq F \end{cases}$$

Image by author

where E, P, and F are the grams of carbohydrates, proteins, and fat that we need to eat.

So in a few words, **yes**, we want to eat the fewest calories possible, **but we also** want them to be enough to provide the right amount of energy for our bodies.

As we have a lot of food, we try all the possible combinations that satisfy the constraints above and pick the one with the minimum amount of calories (the so-called minimum is not a possibility). This is rather a case of **continuous linear optimization,** which has a well-known algorithm that can be used to solve it, and it is called Simplex.

Specifying exactly how these algorithm works is beyond the scope of this article. Let’s just say that the idea is that we are **inside a domain,** and the minimum or maximum of the function inside that domain is located at the extreme points.

We will use the simplex algorithm, in particular, with **Python** and a library that is known as **PuLP**. Let’s dive in.

Results and Analysis

Observed Result:-

`	**Linear Regression:**	**Random Forest**	**Decision Tree**
1. R-squared:	0.2314354827302032	0.6938772422530355	0.49168807750954113
2. Mean squared error:	527.0092090699603	209.9101751556789	348.55247440273035
3. Mean absolute error:	18.091192104847305	7.307920950936309	8.05887372013652
4. cross-validation R-squared:	0.2255792589205587	0.6557314972641132	0.40136831242911947

The CUDA/cuDF code is approximately **4.2x faster** than the original code!

S.no	**Execution Speed**	
1.	**Orignal Code**	Execution time: 2.34 seconds
2.	**Cuda/cuml**	Execution time: 0.56 seconds

5.1 Dataset Description

DATASET COLLECTED Open ML(available in json format)
No of rows -8789(8.8 K food types) Columns 77
Common food
Link of dataset collection- https://www.v7labs.com/blog/best-free-datasets-for-machine-learning#public-government-datasets-for-machine-learning
—open ml
https://www.openml.org/search?type=data&status=active&id=43825
Converted dataset link(csv format)
-https://drive.google.com/file/d/1RRhXmnl67JWk-j8i4-5Xp8M8Bix4kcYv/view?usp=sharing

Target variables

- serving_size
- carbohydrates
- total_fat
- calories
- protein

5.2 tabular format of results

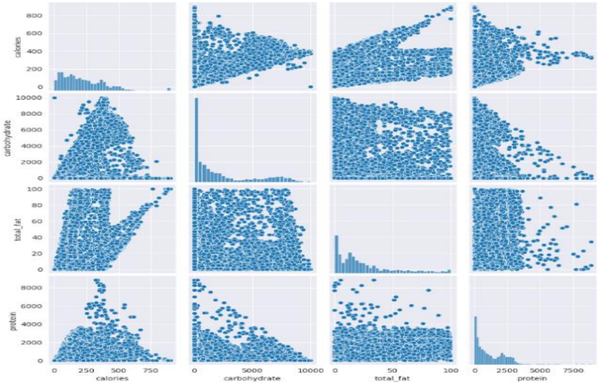
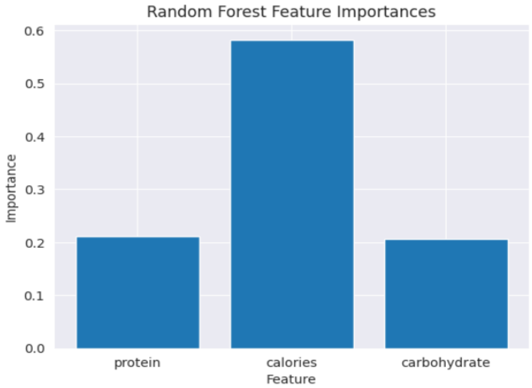
S.no	Image	description
Fig 5.2.1	 	**Linear regression** assumes a linear relationship between the features and the target variable, and may not capture the complex relationships that exist in real-world data.
Fig 5.2.2	 	**Random forest**, on the other hand, uses multiple decision trees to capture these complex relationships, but may overfit the data if the number of trees is too high.

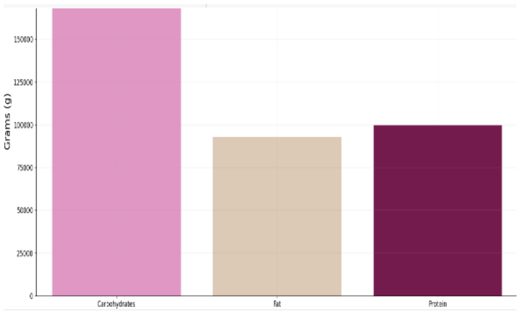
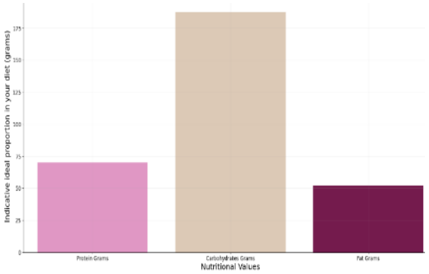
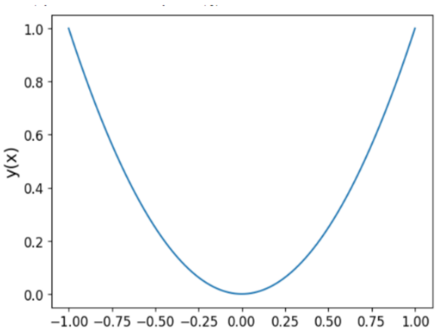
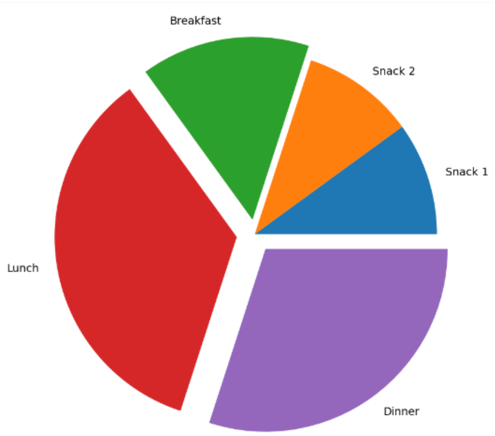
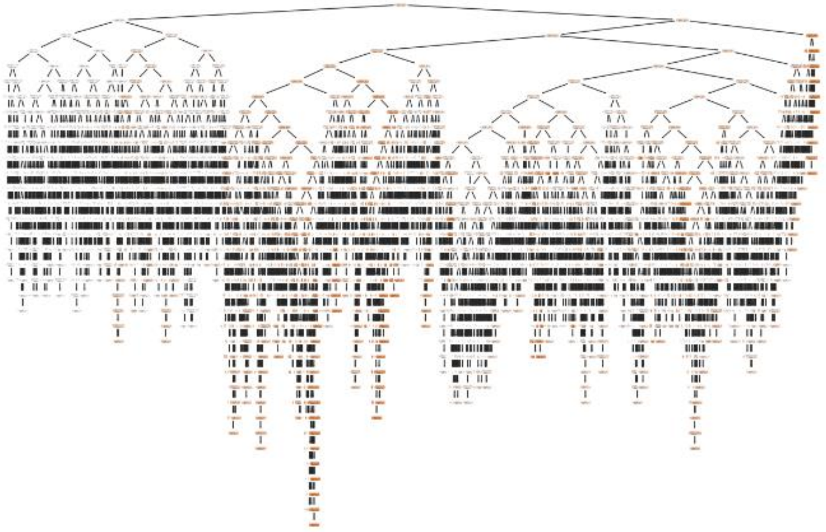
Fig 5.2.3		The proportion of **fat, carbohydrates,** and **proteins** for a 70 kg person with 1500 calories should be something like this.
Fig 5.2.4		
Fig 5.2.5		we **strictly** have to satisfy the number of proteins, carbohydrates, and fat that we have in our plan.

Fig 5.2.5		**Random split** per day data in 5 (one per meal). d(Breakfast) = 1/0.15 d(Lunch) = 1/0.35 d(Dinner) = 1/0.30 d(Snack 1)= d(Snack 2) = 1/0.10
Fig 5.2.5		**Decision tree**, being a simple algorithm, may not capture the nuances of the data as well as random forest, but can still provide a reasonable prediction.

Conclusion and Future Scope

Conclusion: The creation of a comprehensive database of nutritional values for common Indian foods can significantly contribute to improved public health outcomes in India by promoting better nutrition and reducing the prevalence of chronic diseases. Machine learning algorithms, including linear regression, random forest, and decision tree regression, can provide accurate predictions of missing nutritional values based on available features. The algorithms can handle both linear and non-linear relationships and provide insights into the decision-making process of the model.

Future Scope: The database can be further expanded to include a wider variety of Indian foods and nutritional values. The machine learning algorithms can be fine-tuned and optimized to improve their performance. The user-friendly platform can be developed to provide easy access to the database for individuals, enabling them to make informed decisions about their dietary choices. Additionally, the database can be integrated with existing health and wellness platforms to provide a more comprehensive solution for promoting healthy eating habits in India.

References

1.-Times of india (April 2024 2019).
https://timesofindia.indiatimes.com/life-style/health-fitness/health-news/india-is-2.no-2-among-countries-with-most-deaths-caused-by-poor-food-choices/articleshow/69019345.cms
3.-optimisation algorithm
https://towardsdatascience.com/understanding-optimization-algorithms-in-machine-learning-edfdb4df766b
4.Simplex method
https://www.geeksforgeeks.org/simplex-algorithm-tabular-method/
5.Chronic disease
https://en.wikipedia.org/wiki/Chronic_condition
6.calories
How to Calculate Calories per Day: 7 Steps (with Pictures)
7.Linear Regression
https://www.codespeedy.com/fitting-dataset-into-linear-regression-python-model/
8.Random Forest
https://www.geeksforgeeks.org/random-forest-algorithm-in-machine-learning/
9.Theory Content help
https://www.blackbox.ai/
10.Decision tree
https://www.geeksforgeeks.org/decision-tree-implementation-python/
11.Opt
Optimization with PuLP — PuLP 2.8.0 documentation
Research Papers
12.[1]The athlete’s diet: nutritional goals and dietary strategies
https://www.cambridge.org/core/journals/proceedings-of-the-nutrition-society/article/athletes-diet-nutritional-goals-and-dietary-strategies/E7C61FAB06177AB6E1A299B91879E902
Published online by Cambridge University Press: 28 February 2007
13.[2]Nutritional Risks among Adolescent Athletes with Disordered Eating
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8394476/#:~:text=Most%20studies%20refer%20to%20adult%20elite%20athletes%2C%20however,to%20menstrual%20abnormalities%2C%20and%20decreased%20bone%20mass%20density.
Researches
14.https://books.google.co.in/books?hl=en&lr=&id=tCIgEAAAQBAJ&oi=fnd&pg=PR13&dq=linear+regression+analysis&ots=lgsbVvd3So&sig=z0kHrl2wUmmgh

OWYAypsz7PwJVQ&rediresc=y#v=onepage&q=linear%20regression%20analysis&f=false
15.https://www.tandfonline.com/doi/abs/10.1080/01431160412331269698,,,https://onlinelibrary.wiley.com/doi/abs/10.1046/j.1365-277X.2002.00343.x
16.https://ieeexplore.ieee.org/abstract/document/9776710