LONDON
METROPOLITAN
UNIVERSITY

islington college
(इस्लिङ्टन कलेज)

Module Code & Module Title

CU6051NI Artificial Intelligence

25% Individual Coursework

Submission: Proposal

Academic Semester: Autumn Semester 2025

Credit: 15 credit semester long module

Student Name: Ajoob Sagar Kansakar

London Met ID: 23056153

College ID: NP01CP4S240080

Assignment Due Date: 17/12/2025

Assignment Submission Date: 17/12/2025

Submitted To: Er. Roshan Shrestha

GitHub Link	*https://github.com/AjoobKansakar/AI_Coursework.git*

I confirm that I understand my coursework needs to be submitted online via MST Classroom under the relevant module page before the deadline for my assignment to be accepted and marked. I am fully aware that late submissions will be treated as non-submission and a mark of zero will be awarded.

Table of Contents

Table of Figures

1. Introduction

1.1 Topic / AI Concepts used

Artificial Intelligence (AI) refers to computer systems that can perform complex tasks normally done by human reasoning, decision making, creating, etc. With Artificial intelligence one of the most widely used branches of AI is Machine Learning (ML), that enables systems to learn patterns from data and make predictions without being explicitly programmed (May, 2024).

This project focuses on Supervised Machine Learning, specifically Regression techniques. Supervised learning is a technique that uses labelled data sets to train AI models to identify the underlying patterns and relationships with the goal of creating a model that can predict correct outputs on new real-world data (Belcic, 2025).

Regression refers to a supervised learning technique that predicts a continuous numerical value based on one or more independent features by which it finds relationships between input variables (features) and a target variable, enabling the prediction of unseen data (geeksforgeeks, 2025).

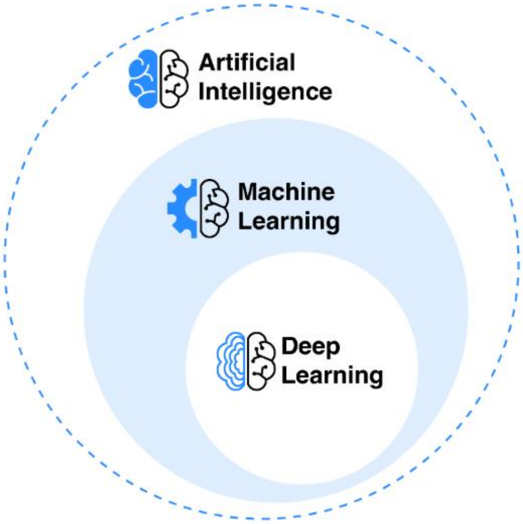


Figure 1: Artificial Intelligence, Machine Learning, and Deep Learning

1.2 Chosen Problem Domain

Nutrition plays a vital role in maintaining a healthy lifestyle and calorie intake is one of the most important nutritional indicators which directly affects the body weight, energy levels, and overall health. Calories are primarily derived from factors like macronutrients such as proteins, carbohydrates, and fats, each contributing a specific amount of energy to the body.

Calories are traditionally calculated with a technique known as the 4-9-4 system which refers that proteins and carbohydrates each contain about 4 calories per gram and fats have 9 calories per gram, and also alcohol has 7 calories per gram (Ailsa Harvey & Joanna Fantozzi, 2022).

The problem domain of this project is to predict the calorie content of food items using nutritional information such as protein, carbohydrates, fats, fibre, and sugar. By applying Machine Learning regression models, this project aims to automate calorie estimation and analyse how different nutrients contribute to total energy content in a food.

2. Background

2.1 Research on the Problem Domain

Calorie is a measure of energy. The calorie number we see on food labels refers to a kilocalorie (Kcal), also referred to as a food calorie. A kilocalorie means 1000 calories and 1kcal is the amount of energy it takes to heat one kilogram of water one degree Celsius at sea level. For foods, calorie count is measured by how much energy that food stores in its chemical bonds. Several studies have explored the relationship between nutritional components and calorie content which shows that:

- Protein provides approximately 4kcal per gram
- Carbohydrates provide approximately 4kcal per gram
- Fat provides approximately 9kcal per gram
- Other factors such as sugar, fibre, cholesterol, calcium and many more factors provide the remaining kcal count (LetsTalkScience, n.d.).

Many existing systems rely on predefined formulas to calculate calories. However, these formulas do not always account for complex interactions between nutrients such as fibre, sugar, and fat composition.

Recent research in machine learning has shown hat regression-based models can effectively predict nutritional values using food composition datasets. Algorithms such as Linear Regression, Decision Trees, and Random Forests have been widely used in food science, health analytics, and dietary recommendation systems.

2.2 Review and Analysis of Existing Work

Existing projects in this domain include:

1. Using machine learning for food caloric and health risk assessment

This study investigates the application of machine learning methods to estimate energy content and classify the health risks of foods based on USDA National Nutrient Database using nutritional composition such as carbohydrates, protein, total lipid, and total sugar content (NIH, 2024).

2. Machine Learning Regression Approach for Estimating Energy Consumption of Appliances in Smart Home

This study uses machine learning algorithms such as Decision Tree (DF), Linear Regression (LR) to estimate the energy consumption of several appliances accurately (Abdulsattar, 2023).

3. Machine Learning in Nutrition Research

This study uses ML to research on nutrition providing a resource which nutrition researchers can refer to guide the traditional techniques (Kirk, 2022).

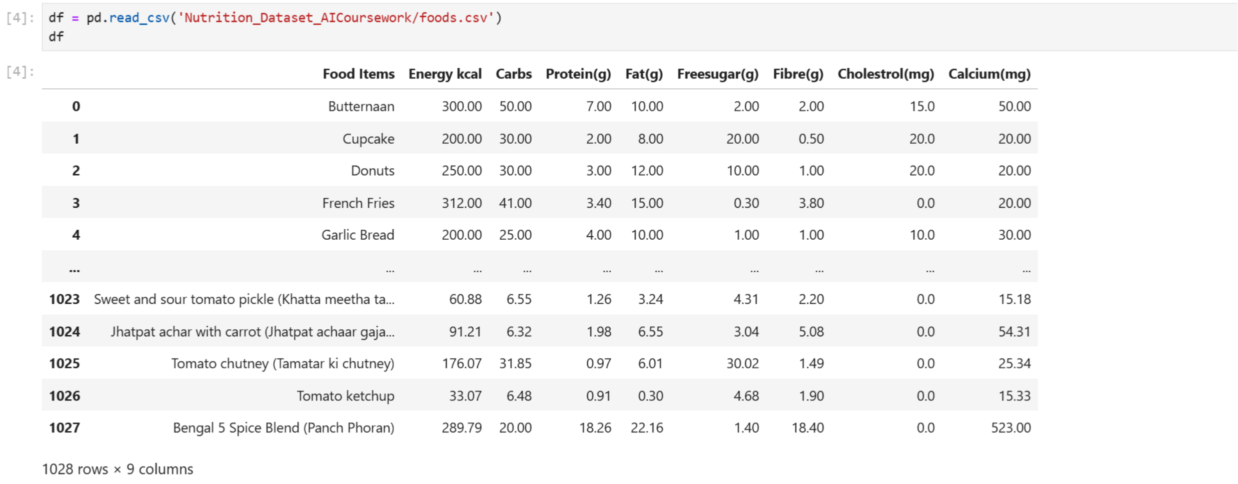
2.3 Dataset Background

The dataset used for this project is obtained from Hugging Face which showcases the Foods Nutritional Contents.

https://huggingface.co/datasets/adarshzolekar/foods-nutrition-dataset

The dataset contains 1,028 food items with the following attributes:

- Food item name
- Energy (Kcal)
- Carbohydrates
- Proteins
- Fats
- Free sugar
- Fiber
- Cholesterol
- Calcium



```
[4]: df = pd.read_csv('Nutrition_Dataset_AICoursework/foods.csv')
     df
```

[4]:

	Food Items	Energy kcal	Carbs	Protein(g)	Fat(g)	Freesugar(g)	Fibre(g)	Cholestrol(mg)	Calcium(mg)
0	Butternaan	300.00	50.00	7.00	10.00	2.00	2.00	15.0	50.00
1	Cupcake	200.00	30.00	2.00	8.00	20.00	0.50	20.0	20.00
2	Donuts	250.00	30.00	3.00	12.00	10.00	1.00	20.0	20.00
3	French Fries	312.00	41.00	3.40	15.00	0.30	3.80	0.0	20.00
4	Garlic Bread	200.00	25.00	4.00	10.00	1.00	1.00	10.0	30.00
...	...	...	...	...	...	...	...	...	...
1023	Sweet and sour tomato pickle (Khatta meetha ta...	60.88	6.55	1.26	3.24	4.31	2.20	0.0	15.18
1024	Jhatpat achar with carrot (Jhatpat achaar gaja...	91.21	6.32	1.98	6.55	3.04	5.08	0.0	54.31
1025	Tomato chutney (Tamatar ki chutney)	176.07	31.85	0.97	6.01	30.02	1.49	0.0	25.34
1026	Tomato ketchup	33.07	6.48	0.91	0.30	4.68	1.90	0.0	15.33
1027	Bengal 5 Spice Blend (Panch Phoran)	289.79	20.00	18.26	22.16	1.40	18.40	0.0	523.00

1028 rows × 9 columns

Figure 2: Dataset Overview

This Dataset is suitable for supervised regression because it includes both input features (X) and a target variable (y) which is Energy kcal.

3. Proposed Solution

3.1 Proposed Approach to Solve the Problem

The objective of this project is to predict the calorie content (Energy in kcal) of food items using their nutritional composition such as protein, carbohydrates, fats, fiber, and sugars. Instead of relying solely on fixed calorie formulas, this project adopts a data-driven Machine Learning approach to learn the relationship between nutritional values and total calories.

The proposed solution treats calorie prediction as a supervised regression problem, where nutritional attributes act as input features and calorie value is the target variable. The solution uses Machine Learning regression models to predict calorie content based on nutritional values.

Steps involved are:

1. Load and explore the dataset from Hugging face
2. Perform data pre-processing (Normalization)
3. Split the dataset into training and testing sets (train_test_split)
4. Train regression models
5. Evaluate and compare model performance
6. Analyse the predictions and results after training the model.

3.2 Algorithms Used

3.2.1 Linear Regression

Analysis used to predict the value of a variable based on the value of another variable where the variable to be predicted is known as dependent variable and the variable used to predict the other variable's value is known as independent variable (IBM, n.d.).

It is a baseline model that assumes a linear relationship between nutritional values and calorie content.

3.2.2 Decision Tree Regressor

Supervised learning algorithm used for both classification and regression tasks which consisting a hierarchical tree structure which consist of a root node, branches, internal nodes and leaf nodes (geeksforgeeks, 2025).

Decision Tree Regressor model helps identify how different nutrients influence calorie values under various conditions by splitting data into branches on feature conditions.

3.2.3 Random Forest Regressor

A powerful tool in data science, enabling accurate predictions and analysis of complex datasets using an advanced machine learning algorithm. It combines multiple decision trees into a single ensemble. Each tree is built from a different subset of the data and makes an independent prediction and averaging all the tree's predictions (AnalytixLabs, 2023).

Random Forest Regressor combines multiple decision trees to improve prediction accuracy and reduce overfitting.

3.3 Pseudocode of the Solution

Pseudocode refers to a step-by-step description of an algorithm which do not use any programming language in its representation instead provides simple English language text for human understanding rather than machine reading which helps developers and researchers plan its logic and structure before writing the actual code (geeksforgeeks, 2025).

```
START
IMPORT required libraries
LOAD nutrition dataset
SELECT INPUT features (protein, carbohydrates, fats, etc.)
SELECT target variable (Energy kcal)
PREPROCESS data
    -   Handle missing values
    -   Normalize numerical features
SPLIT dataset INTO training AND testing sets
FOR EACH algorithm IN [Linear Regression, Decision Tree, Random Forest]:
        TRAIN model ON training data
        PREDICT calories ON test data
        EVALUATE model USING MAE, RMSE, R2 score
COMPARE model performances
SELECT best preforming model
END
```

3.3.1 Pseudocode for Linear Regression

```
START
        IMPORT required libraries
        LOAD food nutrition dataset

        SELECT input features (Protein, Carbohydrates, Fat, Fibre, Sugar, Cholesterol, Calcium)
        SELECT target variables (Energy kcal)

        PREPROCESS dataset
    -   Handle missing values
    -   Normalize numerical features

    SPLIT dataset into training and testing sets
    INTIALIZE Linear Regression model
    TRAIN model using training data
    PREDICT calorie values for test data
    EVALUATE model using MAE, RMSE and R2 score
END
```

3.3.2 Pseudocode for Decision Tree Regressor

```
START
    IMPORT required libraries
    LOAD food nutrition dataset

    SELECT input features (Protein, Carbohydrates, Fat, Fibre, Sugar, Cholesterol, Calcium)
    SELECT target variables (Energy kcal)

    PREPROCESS dataset
  - Handle missing values

    SPLIT dataset into training and testing sets
    INTIALIZE Decision Tree Regressor
    TRAIN model using training data
    PREDICT calorie values for test data
    COMPUTE evaluation metrics MAE, RMSE and R2 score
END
```

3.3.3 Pseudocode for Random Forest Regressor

```
START
    IMPORT required libraries
    LOAD food nutrition dataset

    SELECT input features (Protein, Carbohydrates, Fat, Fibre, Sugar, Cholesterol, Calcium)
    SELECT target variables (Energy kcal)

    PREPROCESS dataset
  - Handle missing values
  - Normalize numerical features

    SPLIT dataset into training and testing sets
    INTIALIZE Random Forest Regressor with multiple decision trees
    TRAIN model using training data
    PREDICT calorie values for test data
    EVALUATE model using MAE, RMSE and R2 score
END
```

3.4 Diagrammatical Representation

Diagrammatical representation is the use of visual tools such as flowchart, charts, graphs, to present complex numerical or conceptual data in simplified, and visually appealing format.

Flowchart is a diagram that depicts a process, system or computer algorithm which are widely used in multiple fields to document, study, and plan complex processes with the help of diagrams (Lucidchart, n.d.).

3.4.1 Combined Flowchart

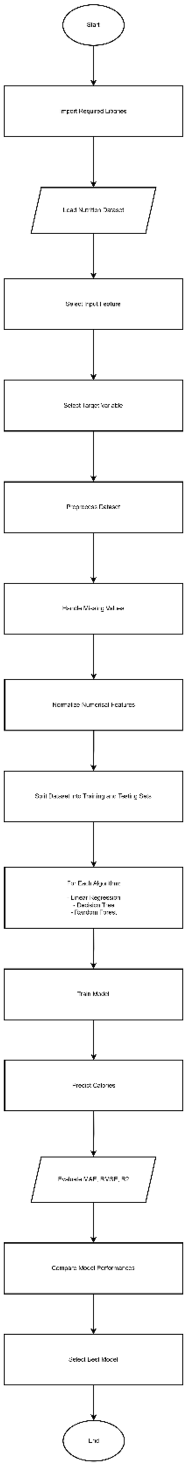


Figure 3: Flowchart

3.4.2 Linear Regression Flowchart

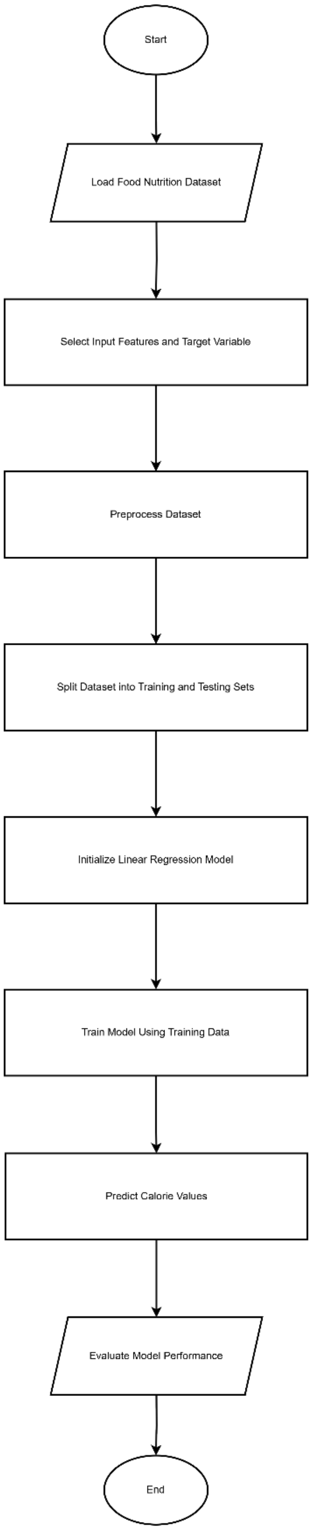

Start
Load Food Nutrition Dataset
Select Input Features and Target Variable
Preprocess Dataset
Split Dataset into Training and Testing Sets
Initialize Linear Regression Model
Train Model Using Training Data
Predict Calorie Values
Evaluate Model Performance
End

Figure 4: Linear Regression Flowchart

3.4.3 Decision Tree Regressor Flowchart

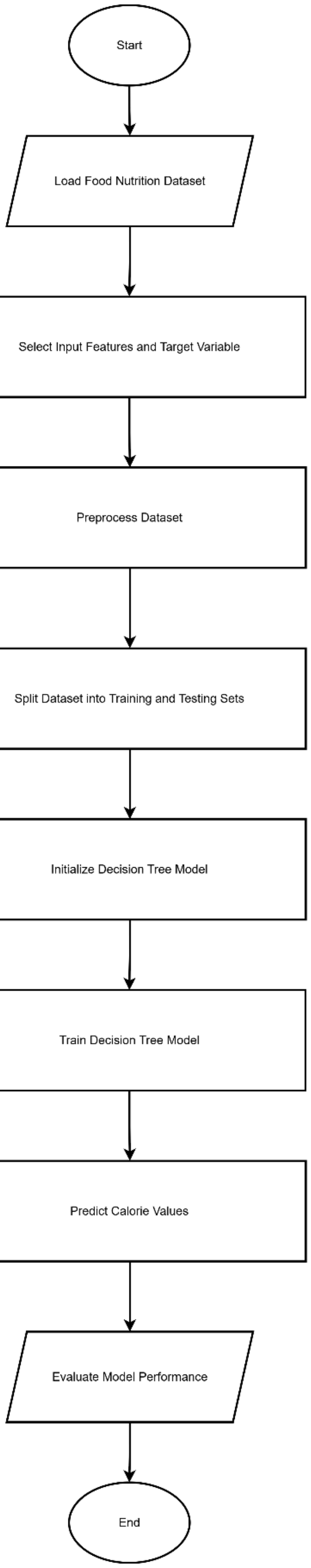

Start
Load Food Nutrition Dataset
Select Input Features and Target Variable
Preprocess Dataset
Split Dataset into Training and Testing Sets
Initialize Decision Tree Model
Train Decision Tree Model
Predict Calorie Values
Evaluate Model Performance
End

Figure 5: Decision Tree Flowchart

3.4.4 Random Forest Regressor Flowchart

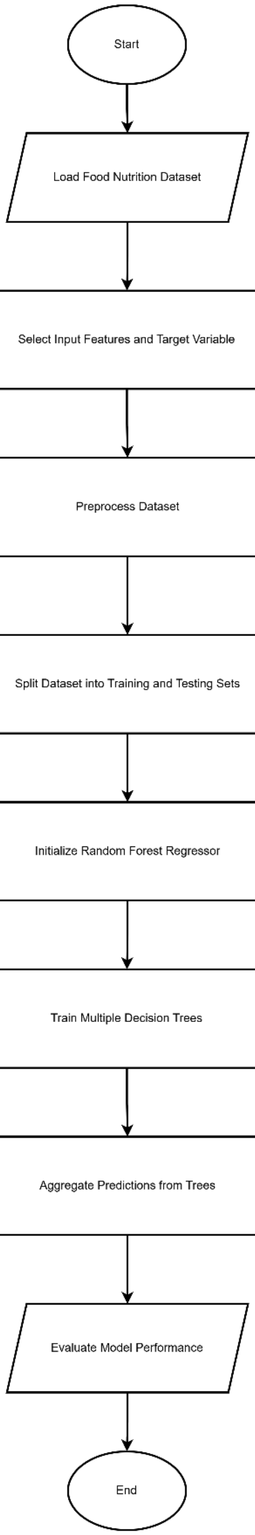

Start
Load Food Nutrition Dataset
Select Input Features and Target Variable
Preprocess Dataset
Split Dataset into Training and Testing Sets
Initialize Random Forest Regressor
Train Multiple Decision Trees
Aggregate Predictions from Trees
Evaluate Model Performance
End

Figure 6: Random Forest Flowchart

4. Conclusion

4.1 Analysis of Work Done

This project successfully demonstrates how Machine Learning regression techniques can be used to predict calorie content based on nutritional values. Linear Regressions provides a clear baseline, while Decision Tree and Random Forest models helps to capture complex non-linear relationships.

The comparison of multiple models allows performance evaluation and better understanding of how nutrients influence calories.

4.2 Real-World Application

The proposed solution can be applied in the following real-world sectors:

1. Nutritional tracking applications
2. Fitness and diet planning systems
3. Health analytics platforms
4. Food recommendation systems

Additionally, it can also help users estimate calorie intake even when calorie values are missing from food labels.

4.3 Further Work

Future improvements may include the following features:

1. Adding more nutritional features to the dataset for more accurate calculations
2. Increasing dataset size
3. Applying deep learning techniques
4. Integrating the model into a real-time web or mobile applications

5. References

1) Abdulsattar, N. F., 2023. *Machine Learning Regression Approach for Estimating Energy Consumption of Appliances in Smart Home.* [Online]
Available at: https://ieeexplore.ieee.org/document/10217991
[Accessed 14 December 2025].

2) Ailsa Harvey & Joanna Fantozzi, 2022. *How calories are calculated: The science behind your food..* [Online]
Available at: https://www.livescience.com/62808-how-calories-are-calculated.html
[Accessed 14 December 2025].

3) AnalytixLabs, 2023. *Random Forest Regression — How it Helps in Predictive Analytics?.* [Online]
Available at: https://medium.com/@byanalytixlabs/random-forest-regression-how-it-helps-in-predictive-analytics-01c31897c1d4
[Accessed 15 December 2025].

4) Belcic, I., 2025. *What is Supervised Learning?.* [Online]
Available at: https://www.ibm.com/think/topics/supervised-learning
[Accessed 14 December 2025].

5) geeksforgeeks, 2025. *Decision Tree in Machine Learning.* [Online]
Available at: https://www.geeksforgeeks.org/machine-learning/decision-tree-introduction-example/
[Accessed 15 december 2025].

6) geeksforgeeks, 2025. *Regression in Machine Learning.* [Online]
Available at: https://www.geeksforgeeks.org/machine-learning/regression-in-machine-learning/
[Accessed 14 December 2025].

7) geeksforgeeks, 2025. *What is PseudoCode.* [Online]
Available at: https://www.geeksforgeeks.org/dsa/what-is-pseudocode-a-complete-tutorial/
[Accessed 15 december 2025].

8) IBM, n.d. *What is linear regression?.* [Online]
Available at: https://www.ibm.com/think/topics/linear-regression
[Accessed 15 december 2025].

9) Kirk, D., 2022. *Machine Learning in Nutrition Research.* [Online]
Available at:
https://www.sciencedirect.com/science/article/pii/S2161831323000923
[Accessed 14 december 2025].

10) LetsTalkScience, n.d. *Science Behind Calories and Nutrition Facts label.* [Online]
Available at: https://letstalkscience.ca/educational-resources/stem-explained/science-behind-calories-and-nutrition-facts-labels
[Accessed 14 December 2025].

11) Lucidchart, n.d. *What is a Flowchart?.* [Online]
Available at: https://www.lucidchart.com/pages/what-is-a-flowchart-tutorial
[Accessed 16 December 2025].

12) May, K., 2024. *What is Artificial Intelligence?.* [Online]
Available at: https://www.nasa.gov/what-is-artificial-intelligence/
[Accessed 14 December 2025].

13) NIH, 2024. *Enhancing dietary analysis: Using machine learning for food caloric and health risk assessment.* [Online]
Available at: https://pubmed.ncbi.nlm.nih.gov/39366774/
[Accessed 14 December 2025].