**ENERGY PREDICTION FROM INDIAN DISHES**

A Course Project report submitted

in partial fulfillment of requirement for the award of degree

**BACHELOR OF TECHNOLOGY**

in

**ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING**

by

**M.RANGA YOCHANA (2103A52097)**

**N. ANKITHA REDDY (2103A52061)**

**M. ANANYA DARSHINI (2103A52095)**

Under the guidance of

**Mr. D Ramesh**

Assistant Professor, Department of CSE.

**

**Department of Computer Science and Artificial Intelligence**

**

**Department of Computer Science and Artificial Intelligence**

**CERTIFICATE**

This is to certify that project entitled **“ENERGY PREDICTION FROM** INDIAN DISHES" is the bonafied work carried out by **N. ANKITHA REDDY, M. RANGAYOCHANA , M.ANANAYA DARSHINI** as a Course Project for the partial fulfillment to award the degree **BACHELOR OF TECHNOLOGY** in **ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING** during the academic year 2022-2023 under our guidance and Supervision.

**Mr. D Ramesh Dr. M Sheshikala**

Asst. Professor, Assoc. Prof. & HOD (CSE),

S R University, S R University,

Ananthasagar , Warangal. Ananthasagar, Warangal.

**ACKNOWLEDGEMENT**

We express our thanks to Course co-coordinator **Mr.** **D Ramesh, Asst. Prof.** for guiding us from the beginning through the end of the Course Project. We express our gratitude toHead of the department CS&AI, **Dr. M Sheshikala, Associate Professor** for encouragement, support and insightful suggestions. We truly value their consistent feedback on our progress, which was always constructive and encouraging and ultimately drove us to the right direction.

We wish to take this opportunity to express our sincere gratitude and deep sense of respect to our beloved Dean, School of Computer Science and Artificial Intelligence, **Dr C V Guru Rao**, for his continuous support and guidance to complete this project in the institute.

Finally, we express our thanks to all the teaching and non-teaching staff of the department for their suggestions and timely support.

**ABSTRACT**

Food is the key element of every human body. So, diet must be always taken into consideration. The knowledge about total intake of calories and nutrients to be consumed to maintain a fit and healthy life is needed We need a robust system which can predict the energy per serve of an food item in an restaurant without excluding the important features. But, in most of the cases, unfortunately people face difficulties during estimation and measurement of the amount of food intake due to the mainly lack of nutritional information, which may include manual process of writing down this information, and other reasons. Energy per serve in fast-food chains has become very important in this modern world and it is very difficult and subtle process, and it is not easy to calculate manually. An accurate predictive model can be valuable to businesses and consumers to determine the energy and make the necessary decision. The goal of the project is to build a model that can accurately determine the energy in food based on various features from the dataset.

**Table of Contents**

**Chapter No. Title Page No.**

1. Introduction
   1. Overview 1
   2. Problem Statement 1
   3. Existing system 1
   4. Proposed system 1
   5. Objectives
   6. Architecture
2. Literature survey 2
3. Data pre-processing
   1. Dataset description 3
   2. Data cleaning 3
   3. Data Visualization 4
4. Methodology 11
   1. Procedure to solve the given problem
   2. Model architecture
   3. Software description
5. Results and discussion 19
6. Conclusion and future scope

7.References 20

# 1. INTRODUCTION

#### 1.1 Overview

The goal of the project is to build a model that can accurately determine the energy in food based on various features from the dataset. In our dataset we 13 columns and 142 rows. Among 13 columns. We have 2 columns as strings so we do not consider them. And other 10 columns are considered as X variables. X variables are independent variables. The variables that do not rely upon any other variable within the dataset are independent variables. In our dataset Y variable is energy and it is dependent variable. The variables that get influenced by independent variables are dependent variables.

#### 1.2 Problem Statement

The energy in food is often decided by multiple factors. The energy determination, therefore,seems difficult. An accurate predictive model can be valuable to businesses and consumers to determine the energy per serve.

#### 1.3 Existing system

In the existing system we only have dataset but there is no correct prediction of energy.

#### 1.4 Proposed system

A robust automated system is required to predict the energy of every food item for which we will be implementing linear regression, KNN, Decision tree, SVM, Random forest. Linear regression comes under regression but KNN, Decision trees, SVM comes under both regression and classification. So here we are finding mean square error by using different models. By comparing all the mean square errors we got from different models we will get to know which model is best model.

#### 1.5 Objectives

* Dataset
* Data pre-processing
* Methodology
* Result
* Conclusion

### 1.6 Architecture

* Data Collection
* Training
* Evaluation
* Prediction

# 2. LITERATURE SURVEY

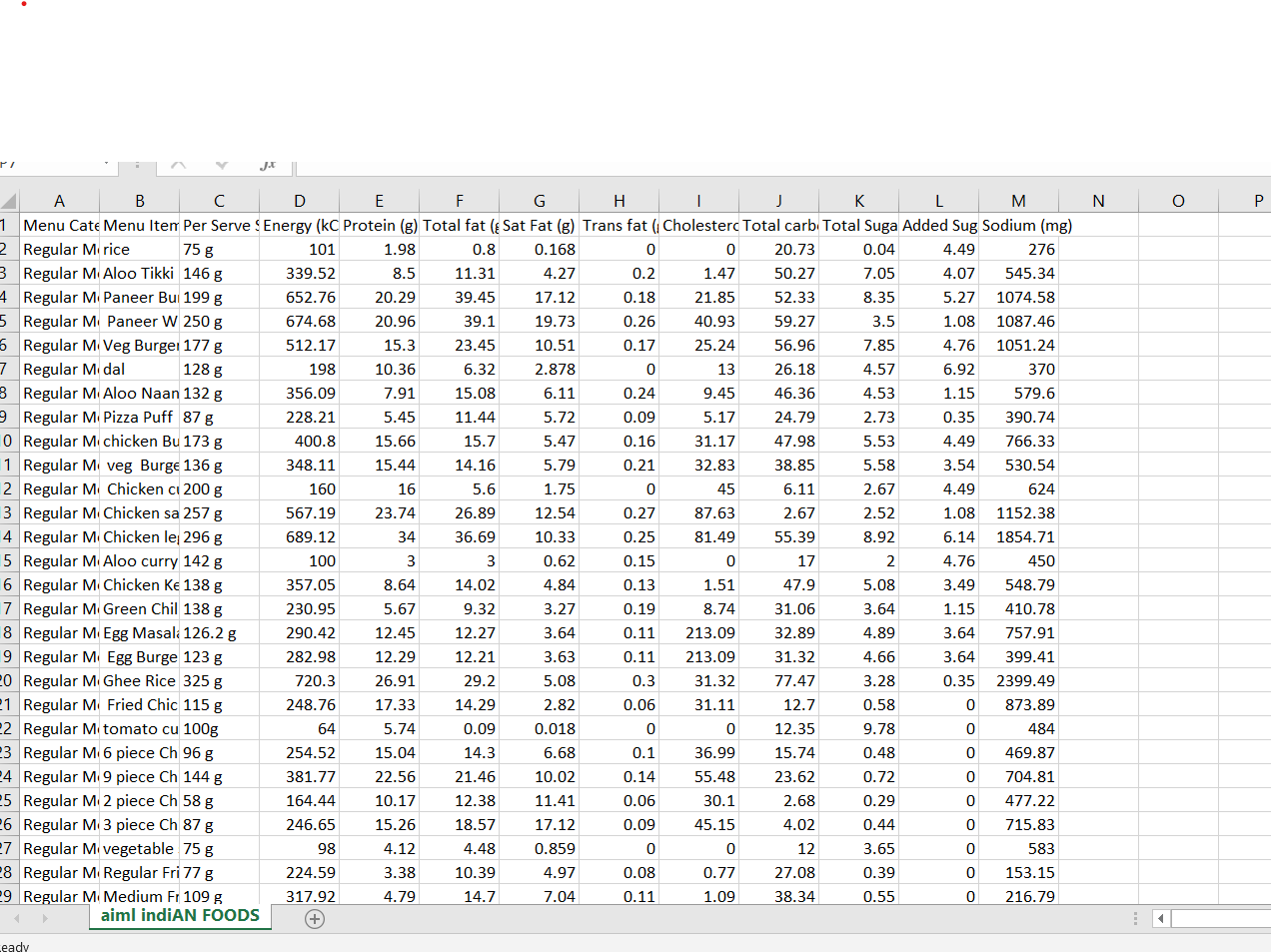
Food is the key element of every human body. So, diet must be always taken into consideration. The knowledge about total intake of calories and nutrients to be consumed to maintain a fit and healthy life is needed. But, in most of the cases, unfortunately people face difficulties during estimation and measurement of the amount of food intake due to the mainly lack of nutritional information, which may include manual process of writing down this information, and other reasons. Energy per serve in fast-food chains has become very important in this modern world and it is very difficult and subtle process, and it is not easy to calculate manually. The existing calories measurement scheme is imperfect as there is manual entry of data such as food platter weight, food platter volume etc. To do this, we have developed a fully automated calorie measurement system.

We have used various ML modules in this system to evaluate the foodstuffs. As the main objective of this research is to develop an energy consumption predictive model by using several machine learning methods in a cloud-based machine learning platform, this research focuses more on the accuracy of the methodology applied in predicting energy consumption. This dataset provides a nutrition analysis of various food items. Based on this data, system parameters will be estimated and thus, a mathematical model could be generated. Several previous studies have analysed the data-driven machine learning approach which is Support Vector Machine (SVM) to predict the energy consumption based on various nutrients present in the food item and per-serve input. Overall, SVM method managed to predict the total energy consumption with root mean square error (RMSE) of 1.4166967567682813 . Another approach used for the model is k-NN. The k-NN forecasting method was approached using a set of historical observations (daily load curves) and their successors. The k-NN method is good at classifying data but limited in forecasting future value as it only identifies similar instances in large feature space. e historical data set using Linear Regression (LR), Support Vector Regression (SVR), Random Forest (RF), Decision Tree (DT) and k-Nearest Neighbour (k-NN). The results showed that LR and SVR models had the best performance with 85.7% accuracy.

# 3. DATA PRE-PROCESSING

#### 3.1. Dataset Description

This dataset provides a nutrition analysis of every menu item on the Indian menu, including breakfast, lunch, dinner, burgers, fries, salads, soda, coffee and tea, milkshakes, and desserts. In our dataset we 13 columns and 142 rows. Among 13 columns. We have 2 columns as strings so we do not consider them. And other 10 columns are considered as X variables. X variables are independent variables. The variables that do not rely upon any other variable within the dataset are independent variables. In our dataset Y variable is energy and it is dependent variable. The variables that get influenced by independent variables are dependent variables.



#### 3.2. Dataset Cleaning

Data cleaning is the process of fixing or removing incorrect, corrupted, incorrectly formatted, duplicate, or incomplete data within a dataset. When combining multiple data sources, there are many opportunities for data to be duplicated or mislabeled. If data is incorrect, outcomes and

algorithms are unreliable, even though they may look correct. We have cleaned the string rows from the dataset to get overall energy prediction.

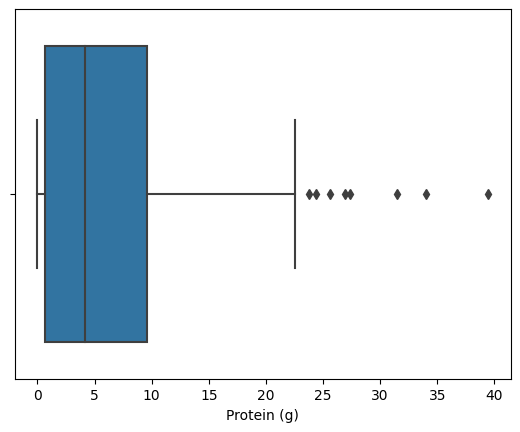
## 3.3. Data Visualization

**BOXPLOT:**

* A Box Plot is also known as Whisker plot is created to display the summary of the set of data values having properties like minimum, first quartile, median, third quartile and maximum.
* Box plots are used to show distributions of numeric data values, especially when you want to compare them between multiple groups. They are built to provide high-level information at a glance, offering general information about a group of data's symmetry, skew, variance, and outliers.

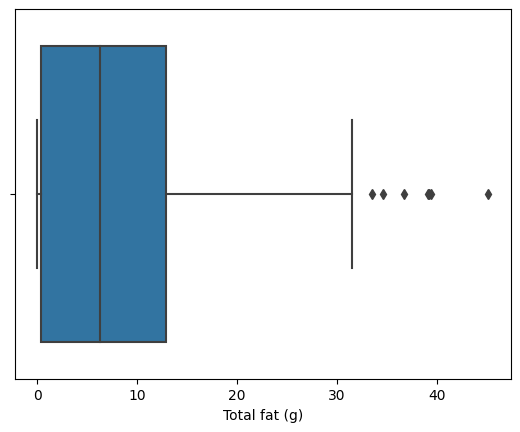
import seaborn as sns

sns.boxplot(data=d,x='Protein (g)')



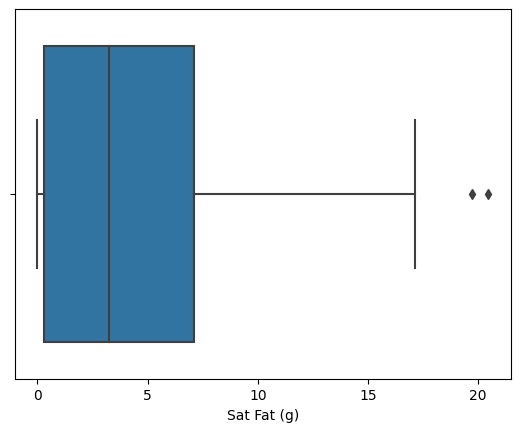
import seaborn as sns

sns.boxplot(data=d,x='Total fat (g)')

****

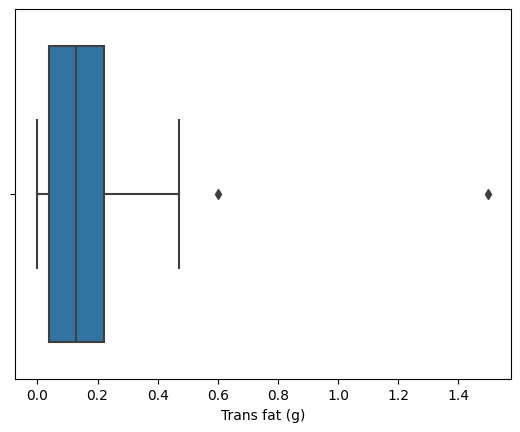
import seaborn as sns

sns.boxplot(data=d,x='Sat Fat (g)')



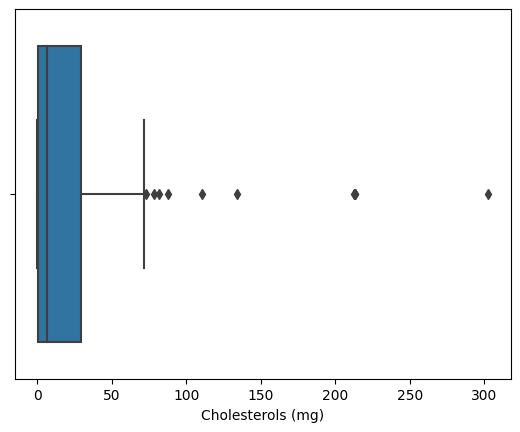
import seaborn as sns

sns.boxplot(data=d,x='Trans fat (g)')



import seaborn as sns

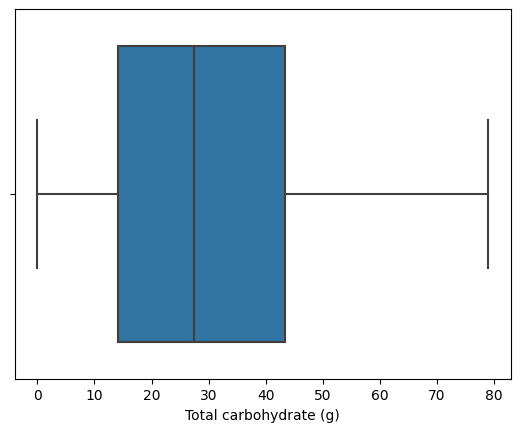
sns.boxplot(data=d,x='Cholesterols (mg)')



import seaborn as sns

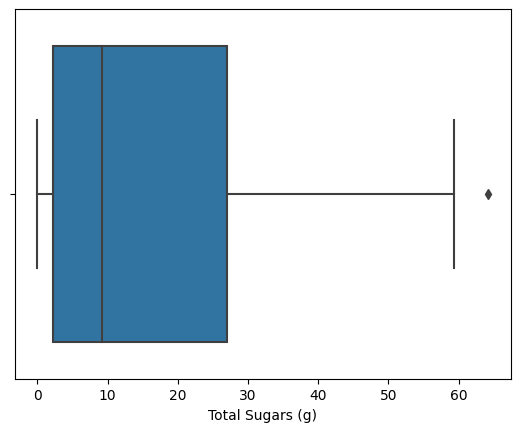
sns.boxplot(data=d,x='Total carbohydrate (g)'

)



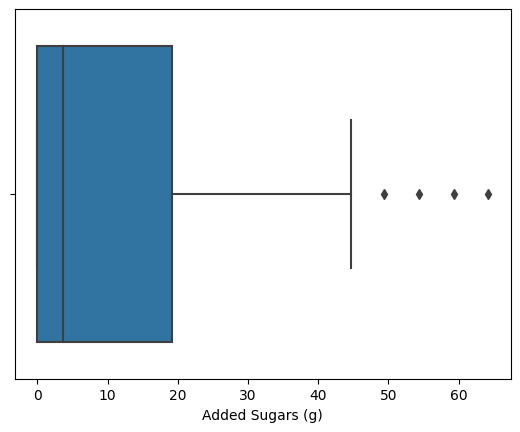
import seaborn as sns

sns.boxplot(data=d,x='Total Sugars (g)')



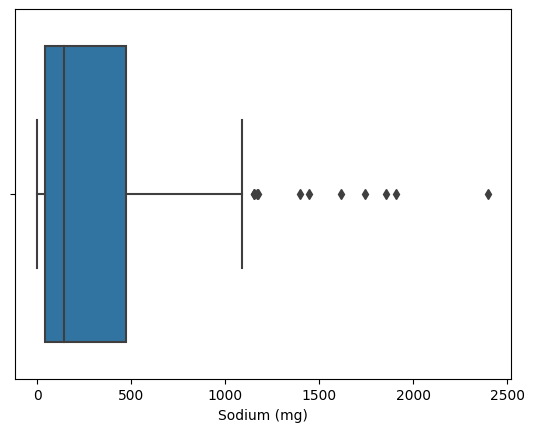
import seaborn as sns

sns.boxplot(data=d,x='Added Sugars (g)')



import seaborn as sns

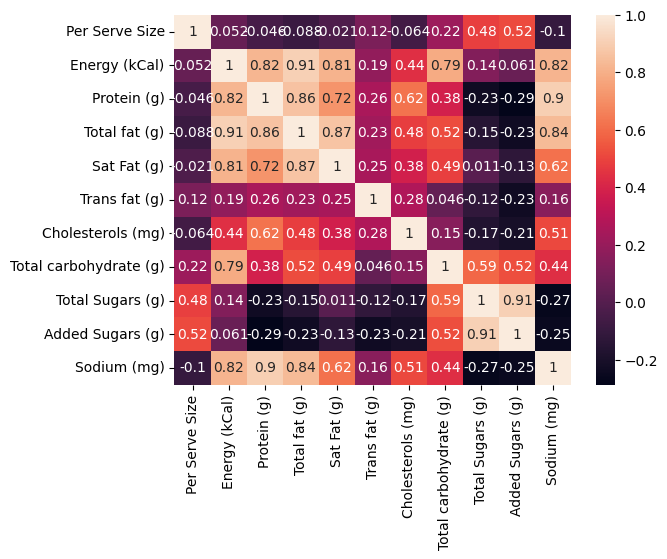
sns.boxplot(data=d,x='Sodium (mg)')



**CORELATION MATRIX:**

A correlation matrix is simply a table which displays the correlation coefficients for different variables. The matrix depicts the correlation between all the possible pairs of values in a table. It is a powerful tool to summarize a large dataset and to identify and visualize patterns in the given data.

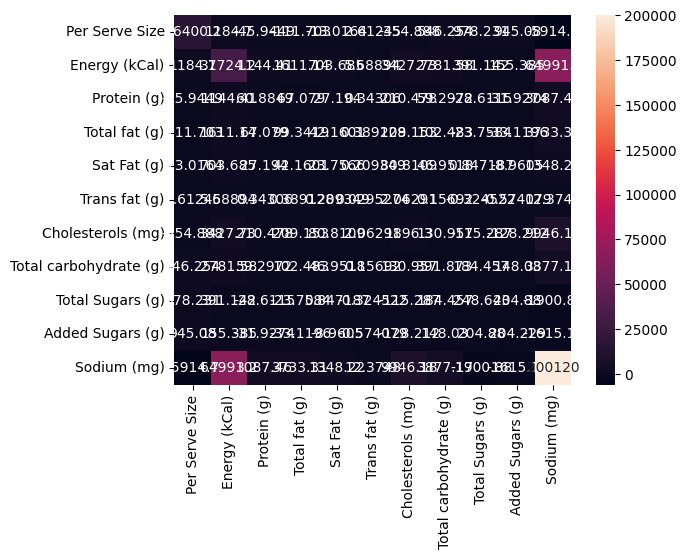
**corr(X,Y) = cov(X,Y) sXsY** where sX and sY are, respectively, the standard deviations of variables X and Y. Because a correlation is a specific form of a covariance, it has the same two properties--magnitude and sign--as a covariance. The sign indicates the direction of the relationship.



**COVARIANCE MATRIX:**

Covariance Matrix is a measure of how much two random variables gets change together. It is actually used for computing the covariance in between every column of data matrix. The Covariance Matrix is also known as dispersion matrix and variance-covariance matrix.

The covariance matrix **provides a useful tool for separating the structured relationships in a matrix of random variables**. This can be used to decorrelate variables or applied as a transform to other variables. It is a key element used in the Principal Component Analysis data reduction method, or PCA for short.

****

# 4. METHODOLOGY

#### 4.1. Procedure to Solve the given problem

The proposed solutions for the taken dataset are Linear Regression, KNN, Decision trees, Support Vector Machine (SVM).

##### a. Linear Regression:

Linear regression is used to predict the relationship between two variables by applying a linear equation to observed data. There are two types of [variable,](https://www.vedantu.com/maths/variable) one variable is called an independent variable, and the other is a dependent variable. Linear regression is commonly used for predictive analysis. The main idea of regression is to examine two things.

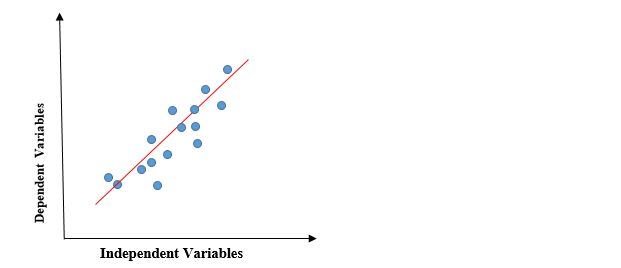
Linear Regression Equation is given below:

Y= a + b X

where X is the independent variable and it is plotted along the x-axis .

Y is the dependent variable and it is plotted along the y-axis.

If there is a single input variable (x), such linear regression is called simple linear regression. And if there is more than one input variable, such linear regression is called multiple linear regression. The linear regression model gives a sloped straight line describing the relationship within the variables.



The above graph presents the linear relationship between the dependent variable and independent variables. When the value of x (independent variable) increases, the value of y (dependent variable) is likewise increasing. The red line is referred to as the best fit straight line. Based on the given data points, we try to plot a line that models the points the best.

**b. KNN**

KNN is a simple, supervised machine learning (ML) algorithm that can be used for classification or regression tasks - and is also frequently used in missing value imputation. It is based on the idea that the observations closest to a given data point are the most "similar" observations in a data set, and we can therefore classify unforeseen points based on the values of the closest existing points. By choosing K, the user can select the number of nearby observations to use in the algorithm.

The K-NN working can be explained on the basis of the below algorithm:

* **Step-1:** Select the number K of the neighbors
* **Step-2:** Calculate the Euclidean distance of **K number of neighbors**
* **Step-3:** Take the K nearest neighbors as per the calculated Euclidean distance.
* **Step-4:** Among these k neighbors, count the number of the data points in each category.
* **Step-5:** Assign the new data points to that category for which the number of the neighbor is maximum.
* **Step-6:** Our model is ready.

## Advantages of KNN Algorithm:

* It is simple to implement.
* It is robust to the noisy training data
* It can be more effective if the training data is large.

## Disadvantages of KNN Algorithm:

* Always needs to determine the value of K which may be complex some time.
* The computation cost is high because of calculating the distance between the data points for all the training samples.

**c. Decision trees**

Decision trees are a popular [machine learning algorithm](https://www.analyticsvidhya.com/blog/2022/01/machine-learning-algorithms/) that can be used for both [regression](https://www.analyticsvidhya.com/web-stories/5-regression-techniques-you-should-know/) and [classification](https://www.analyticsvidhya.com/blog/2022/07/classification-using-pyspark-databricks-and-koalas/) tasks. They are easy to understand, interpret, and implement, making them an ideal choice for beginners in the field of [machine learning](https://www.analyticsvidhya.com/machine-learning/). In this comprehensive guide, we will cover all aspects of the decision tree algorithm, including the working principles, different types of decision trees, the process of building decision trees, and how to evaluate and optimize decision trees.

There are various algorithms in Machine learning, so choosing the best algorithm for the given dataset and problem is the main point to remember while creating a machine learning model. Below are the two reasons for using the Decision tree:

* Decision Trees usually mimic human thinking ability while making a decision, so it is easy to understand.
* The logic behind the decision tree can be easily understood because it shows a tree-like structure.

STRUCTURE:



* A decision tree which is also known as prediction tree refers a tree structure to mention the sequences of decisions as well as consequences.
* Considering the input X = (X1, X2,…Xn), the aim is to predict a response or output variable Y.
* Each element in the set (X1, X2,…Xn) is known as input variable. It is possible to achieve the prediction by the process of building a decision tree which has test points as well as branches.
* At each test point, it is decided to select a particular branch and traverse down the tree.
* Ultimately, a final point is reached, and it will be easy to make prediction.
* In a decision tree, all the test points exhibit testing specific input variables (or attributes), and the developed decision tree is represented by the branches.
* Because of flexibility as well as simple visualization, decision trees are mostly probably deployed in data mining applications for the purpose of classification.

**d. Support Vector Machine (SVM)**

Support Vector Machine or SVM is one of the most popular Supervised Learning algorithms, which is used for Classification as well as Regression problems. However, primarily, it is used for Classification problems in Machine Learning. The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future. This best decision boundary is called a hyperplane. SVM chooses the extreme points/vectors that help in creating the hyperplane. These extreme cases are called as support vectors, and hence algorithm is termed as Support Vector Machine. SVMs are used in applications like handwriting recognition, intrusion detection, face detection, email classification, gene classification, and in web pages. This is one of the reasons we use SVMs in machine learning. It can handle both classification and regression on linear and non-linear data.

The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future. This best decision boundary is called a hyperplane. SVM chooses the extreme points/vectors that help in creating the hyperplane. These extreme cases are called as support vectors, and hence algorithm is termed as Support Vector Machine. Consider the below diagram in which there are two different categories that are classified using a decision boundary or hyperplane:



e. **Random Forest:**

Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML. It is based on the concept of **ensemble learning,** which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model. **"Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset."**

Below are some points that explain why we should use the Random Forest algorithm:

* It takes less training time as compared to other algorithms.
* It predicts output with high accuracy, even for the large dataset it runs efficiently.
* It can also maintain accuracy when a large proportion of data is missing.
* Random Forest works in two-phase first is to create the random forest by combining N decision tree, and second is to make predictions for each tree created in the first phase.
* The Working process can be explained in the below steps and diagram:
* **Step-1:** Select random K data points from the training set.
* **Step-2:** Build the decision trees associated with the selected data points (Subsets).
* **Step-3:** Choose the number N for decision trees that you want to build.
* **Step-4:** Repeat Step 1 & 2.
* **Step-5:** For new data points, find the predictions of each decision tree, and assign the new data points to the category that wins the majority votes.

#### 4.2. Model architecture

The Machine Learning Architecture can be categorized on the basis of the algorithm used in training.

##### a. Supervised Learning

Classification analysis is presented when the outputs are restricted in nature and limited to a set of values. However, regression analysis defines a numerical range of values for the output. Examples of supervised learning are seen in face detection, speaker verification systems. In supervised learning, models are trained using labelled dataset, where the model learns about each type of data. Once the training process is completed, the model is tested on the basis of test data (a subset of the training set), and then it predicts the output.

## Steps Involved in Supervised Learning:

* First Determine the type of training dataset
* Collect/Gather the labelled training data.
* Split the training dataset into training **dataset, test dataset, and validation dataset.**
* Determine the input features of the training dataset, which should have enough knowledge so that the model can accurately predict the output.
* Determine the suitable algorithm for the model, such as support vector machine, decision tree, etc.
* Execute the algorithm on the training dataset. Sometimes we need validation sets as the control parameters, which are the subset of training datasets.
* Evaluate the accuracy of the model by providing the test set. If the model predicts the correct output, which means our model is accurate.

##### b. Unsupervised Learning

Unlike supervised learning, unsupervised learning uses training data that does not contain output. The unsupervised learning identifies relation input based on trends, commonalities, and the output is determined on the basis of the presence/absence of such trends in the user input. unsupervised learning is a machine learning technique in which models are not supervised using training dataset. Instead, models itself find the hidden patterns and insights from the given data. It can be compared to learning which takes place in the human brain while learning new things. It can be defined as: Unsupervised learning is a type of machine learning in which models are trained using unlabeled dataset and are allowed to act on that data without any supervision.

* Unsupervised learning is helpful for finding useful insights from the data.
* Unsupervised learning is much similar as a human learns to think by their own experiences, which makes it closer to the real AI.
* Unsupervised learning works on unlabeled and uncategorized data which make unsupervised learning more important.
* In real-world, we do not always have input data with the corresponding output so to solve such cases, we need unsupervised learning.

##### c. Reinforcement Training

This is used in training the system to decide on a particular relevance context using various algorithms to determine the correct approach in the context of the present state. These are widely used in training gaming portals to work on user inputs accordingly. Reinforcement Learning (RL) is the science of decision making. It is about learning the optimal behavior in an environment to obtain maximum reward. In RL, the data is accumulated from machine learning systems that use a trial-and-error method. Reinforcement learning uses algorithms that learn from outcomes and decide which action to take next. After each action, the algorithm receives feedback that helps it determine whether the choice it made was correct, neutral or incorrect. It is a good technique to use for automated systems that have to make a lot of small decisions without human guidance.

#### 4.3. Software description

1. **PYTHON –** Python is an interpreted, high-level, general-purposeprogramming language. Python’s design philosophy emphasizes code readability with its notable use of significant whitespace. Its language constructs and object-oriented approach aim to help programmers write clear, logical code for small and largescale projects. Python is dynamically typed and supports multiple programming paradigms, including procedural, object-oriented, and function programming.

* **GOOGLE COLAB-** Collab notebooks allow you to combine executable code and rich text in a single document, along with images, HTML, LaTeX and more. When you create your own Co-lab notebooks, they are stored in your Google Drive account. You can easily share your Co-lab notebooks with co-workers or friends,

allowing them to comment on your notebooks or even edit them. With Co-lab you can harness the full power of popular Python libraries to analyse and visualize data. The code cell below uses numpy to generate some random data and uses matplotlib to visualize it. To edit the code, just click the cell and start editing.

**5.RESULTS AND DISCUSSION**

In this project we are predicting the energy in food by using all the independent variables. As the above-mentioned dataset is supervised learning and under the supervised machine learning, it is considered as the regression. The following tables gives us the necessary information to choose the predictor with minimal error.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | LINEAR REGRESSION | KNN | DECISION TREES | SVM | RANDOM FOREST |
| ROOT MEAN SQUARE ERROR | 3.601514642  343941 | 9.56174749  82043 | 8.0995246774163 | 1.41669675  67682813 | 6.90974066  4303618 |
| MEAN SQUARE ERROR | 12.970907719  017806 | 91.427017  931035 | 65.6023 | 2.007029700  63  7767 | 47.744516  047931 |
| MEAN ABSOLUTE ERROR | 3.158474078  3651357 | 7.1560689  65517242 | 6.3762068965  51723 | 1.0284004880333  064 | 5.17400344  827586 |

**6.CONCLUSION AND FUTURE SCOPE**

#### 6.1. CONCLUSION

This dataset is used for the data visualization project. From the chosen dataset by using all the independent variables we will find the mean square error, mean absolute error and root mean square error and pick the model with minimum error so that we can determine the energy of food more accurately than the other two models.

#### 6.2. FUTURE SCOPE

There is a great need for robust systems in this area as it involves various important features, and it makes it difficult to do it manually. So, machine learning models can save a lot of time and effort while it is more accurate than the traditional methods

**7. REFERENCES**

1. <https://www.kaggle.com/code/th3c1ph3r/exhautive-eda-of-mcdonald-s-india-menu>
2. <https://www.kaggle.com/code/ankita5532/mcdonalds-india-analysis>
3. <https://www.kaggle.com/code/valchovalev/eda-mcdonald-s-india-menu-nutrition-dataset>
4. <https://www.kaggle.com/code/shivagowri19/mcdonald-s-nutrition-analysis>5. <https://www.kaggle.com/code/pantanjali/mcdonalds-diet-analysis>