PREDICTING CALORIC EXPENDITURE: A REGRESSION APPROACH IN MACHINE LEARNING

Project Report submitted to

MAHATMA GANDHI UNIVERSITY

In partial Fulfilment of the Requirements for the Award of the Degree of

MASTER OF SCIENCE IN STATISTICS WITH DATA SCIENCE

Submitted by

SEBASTIAN JOSE

Reg.No.220011024080

POST GRADUATE DEPARTMENT OF STATISTICS

SREE SANKARA COLLEGE, KALADY, KERALA

(Reaccredited with A Grade by NAAC

Affiliated to Mahatma Gandhi University, Kottayam)

MARCH 2024

BONAFIDE CERTIFICATE

This is to certify that the dissertation entitled "**PREDICTING CALORIC EXPENDITURE: AREGRESSION APPROACH IN MACHINE LEARNING**" is a record of original work done by **SEBASTIAN JOSE,** under the supervision of **Ms. SUMI,** a respected officer at Acutro Technologies in partial fulfilment of the requirements of the Degree of **MASTER OF SCIENCE IN STATISTICS WITH DATA SCIENCE** of Mahatma Gandhi University, Kottayam. It is further certified that the project work has not been previously formed the basis for the award of any other Degree or Diploma or other similar title to any candidate of this or any other university.

Dr. BIJU THOMAS
Head of Department of Statistics
Sree Sankara college, Kalady

DECLARATION

I, Sebastian Jose, hereby declare that the project entitled "**PREDICTING CALORIC EXPENDITURE: A REGRESSION APPROACH IN MACHINE LEARNING"** is a record of original work conducted by me under the supervision and guidance of Ms. Sumi. This project is being submitted in partial fulfilment of the requirements for the degree of Master of Science in Statistics with Data Science from Mahatma Gandhi University, Kottayam. I affirm that this project has not been previously submitted as part of any other degree, diploma, associate-ship, fellowship, or similar title, either at this university or any other educational institution. Furthermore, I acknowledge that any sources, references, or materials used in this project have been duly cited and acknowledged in accordance with academic integrity guidelines. The project represents my own findings, analysis, and conclusions based on the research conducted.

Place: Kalady

SEBASTIAN JOSE

Date: 27-03-2024

ACKNOWLEDGEMENT

The successful completion of this dissertation owes to the inspiration and constant support that I received from various sources. I take this opportunity to express my sincere gratitude to all those who help me directly or indirectly for the completion of work.

I wish to express my deep gratitude to my guide **Ms. Sumi**, Knowledge Officer, Acutro Technologies for her valuable guidance I have received throughout the period of the study. I place my sincere thanks to the authorities of Acutro Technologies, Thiruvananthapuram, for providing me with the necessary support and facilities for doing this project.

I hereby acknowledge my profound gratitude to **Dr. Preethi Nair**, Principal, Sree Sankara College, for her valuable encouragement for conducting this study.

The continuous support and advice given by **Dr. Biju Thomas**, Head of Department of Statistics, Sree Sankara College, and all the teachers in the department for their supervision, guidance, encouragement and co-operation. I sincerely express my gratitude to my teachers, my parents, my friends and all those who have encouraged me in this endeavour.

Above all, I am greatly thankful to the grace of Almighty God for the successful completion of my training programme.

SEBASTIAN JOSE

CONTENTS

CHAPTER 1
INTRODUCTION

In recent years, the field of healthcare has witnessed a significant transformation with introduction of advanced technologies and data-driven methodologies. Among these, machine learning emerged as a powerful tool. The goal of machine learning is creating a well-trained model that makes improvement over time. One of the many applications of machine learning in health care is the prediction of calories burnt during exercising.

Calorie is a unit of heat energy. In today's world, where staying fit and healthy is more important than ever, understanding how exercise affects our bodies is important. The amount of calories burned during physical activity is one indicator of our overall health and fitness level. As a human body performs some extensive activity or workout, the body temperature and heart rate start rising which leads to the production of heat energy in the body. This ultimately causes calories to burn. Accurately predicting calorie burn can help individuals set and achieve fitness goals and can also inform health coaching and wellness tracking programs. Understanding how many calories are burned throughout the day, can help one ensure that they consume the appropriate amount of food to create a caloric deficit for weight loss or an excess for muscle growth. Additionally, awareness of caloric expenditure aids in the prevention and management of chronic diseases such as obesity, diabetes, and cardiovascular conditions. By balancing caloric intake with expenditure, individuals can maintain a healthy weight, improve metabolic health, and reduce the risk of disease.

The traditional approach to calculating calorie burn through exercise primarily involved using generalized formulas that took into account an individual's weight, the duration of the activity, and the type of exercise performed. These formulas often relied on metabolic equivalents (METs), which are standardized values assigned to various physical activities based on their intensity. While this method provided a rough estimate, it did not account for individual variations such as fitness level, age, gender, or specific physiological responses to exercise. Modern day methods like heart rate monitors and wearable fitness trackers, which utilizes machine learning technologies, are used to get more accurate results.

The rationale for this project is to develop a machine learning model that can accurately predict calorie burn during physical activity. To show the same we take some input parameters such as age, gender, height and weight and apply different regression algorithms such as linear regression, decision tree regression, and random forest regression over the data to get the best and optimal results. By developing an accurate calorie burn prediction model, we can help individuals make more informed decisions about their physical activity and improve their overall health and well-being.

Despite some research efforts, predicting calorie burn using machine learning techniques still presents a significant gap in the existing literature. While existing studies have focused in this area, they often concentrate on forecasting calorie expenditure for particular types of physical activities or within specific demographic groups. Consequently, there exists a need for more generalized models capable of accurately predicting calorie burn across diverse physical activities and individuals.

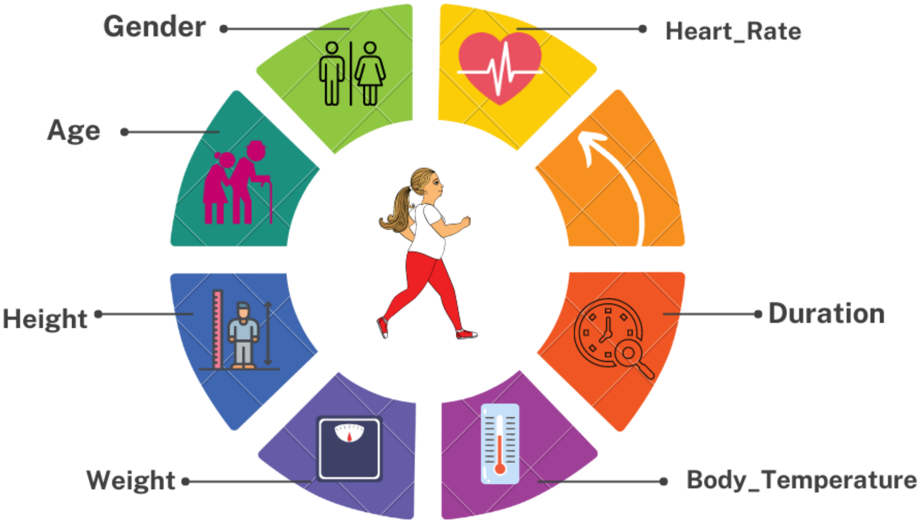


Figure representing the factors for predicting calories burnt

The objectives of this project are:

- To collect data on physiological parameters and calory burn of a number of individuals.

- Using machine learning algorithms to predict calory burn which includes linear regression, decision tree regression and random forest regression.
- To compare these models to find the best model that predicts the calory burn.

This project is divided into the following chapters: Chapter 1 gives the introduction. Here the research background is discussed. The statement of the problem, aim and objectives, are explained in detail. The literature review is also presented in this chapter. Chapter 2 discusses the materials and methods used for analysis and model building in the research. Chapter 3 highlights and explains the results of this research. The result of this project as well as the research implications for practice, conclusion and future scope are discussed towards the end of chapter 4.

1.1 LITERATURE REVIEW

Machine learning has gained widespread use in the prediction of calories burn during physical activities. These studies often collect physical activity data and other data such as heart rate, age, and gender from fitness trackers and mobile applications. This section provides an overview of some of the critical studies in this area.

[1] Sathiya T discussed to predict user's calorie and applied CNN model to classify food items from the input image. They also used image processing techniques such as deep learning model and their model provide 91.65% accuracy in predicting user's calorie from input image.

[2] Akshit Rajesh Tayade used logistics regression algorithm for diet recommendation system to support mental fitness and physical fitness and accuracy of the proposed model was 85.96.

[3] Marte Nipas discussed how to predict burned calories using a supervised learning algorithm. They used a Random forest algorithm and gained 95.77% model accuracy. They also used the iterative method to find out the appropriate output from an input. Their work is almost better than other recent work.

[4] Gunasheela B L discussed their techniques to predict calorie from input images. They used some digital image processing techniques such as image acquisition, RGB conversion, feature extraction and image enhancement so on. They segmented input images and used techniques and then combined segmented images, finally calorie predicted.

[5] KR Westerterp discussed how to determine energy expenditure by body size and body compositions and food intake and physical activity. He used body size and body compositions and some statistical techniques to evaluate calorie expenditure.

[6] The World Health Organization (WHO) discusses various factors that affect the calories burned. They say that anyone can modify their diet chart or activity level to get the desired results.

[7] Salvador Camacho states that universal obesity has been increasing day by day across the whole world and until now not a single nation has been able to resolve it. The main cause of obesity is an energy imbalance between calories eaten and calories expended. The concept of calorie imbalance cannot be sufficient to control and turn the obesity pandemic.

1.2 CALORIES AND CALORIC EXPENDITURE

A calorie is a unit of energy derived from the food and beverages we consume. It represents the amount of energy required to raise the temperature of one gram of water by one degree Celsius. In the context of nutrition, calories serve as a measure of the energy content in food, determining its potential to fuel bodily functions and activities. The concept of calories is important in understanding nutrition and managing weight, as it involves a balance between caloric intake and caloric expenditure.

Caloric intake refers to the number of calories consumed through eating and drinking. Different foods and beverages contain varying amounts of calories based on their macronutrient composition. For instance, carbohydrates and proteins each provide approximately four calories per gram, while fats provide about nine calories per gram. Alcohol, although not a macronutrient, contributes seven calories per gram. Therefore, the type and amount of food we consume directly influence our total caloric intake.

Caloric expenditure refers to the energy utilized by the body to perform various functions, including basal metabolic rate (BMR), physical activity, and the thermic effect of food (TEF). BMR accounts for the energy expended at rest to maintain essential bodily functions, such as breathing, circulating blood, and regulating body temperature. Physical activity encompasses all forms of movement, from structured exercise to daily tasks like walking and housework, each exerting a varying degree of energy expenditure. TEF represents the energy required for digestion, absorption, and metabolism of nutrients consumed.

Several factors influence caloric expenditure, including age, gender, body composition, metabolic rate, and activity level. Individuals with higher muscle mass typically have a higher BMR due to the increased metabolic demands of lean tissue. Similarly, engaging in regular physical activity not only expends calories directly but also enhances metabolic efficiency, thereby contributing to overall energy expenditure.

Calories play a pivotal role in human physiology, serving as the fundamental unit of energy that fuels our bodies. The balance between calorie intake and expenditure is crucial for maintaining a healthy weight and overall well-being.

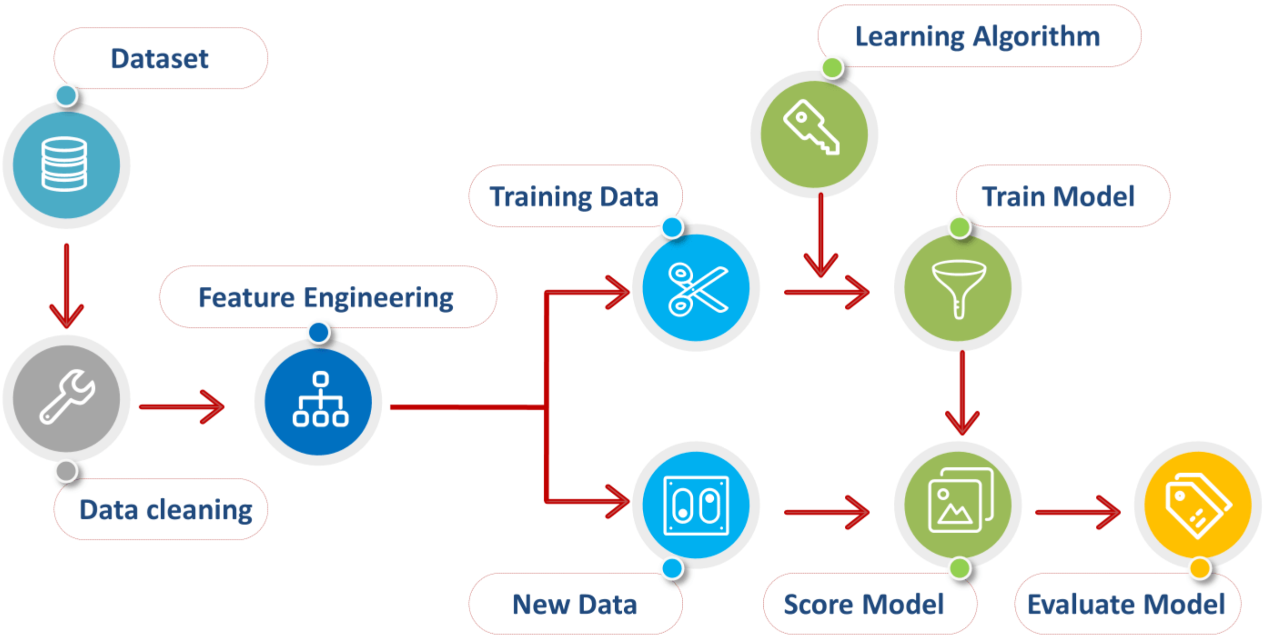
With the growth of big data, many available data can be collected and analysed for predicting caloric expenditure. A process of exploring large datasets to identify unknown patterns that are present in the data is known as data mining. Data mining and machine learning can be used interchangeably, but there are differences between the two. Data mining investigates the data in the pattern. In contrast, machine learning goes beyond what has happened in the past to predict the outcomes. The outcome is based on what the machine has learned from pre-existing data.

This project examines the prediction of calorie expenditure based on two datasets: ‘calories’ and ‘exercise’. The objective of this study is to investigate the factors influencing calory expenditure and to find the best regressor model. The performance of these machine learning algorithms is based mainly on the r square values and mean square error.

CHAPTER 2
METHODOLOGY

This section covers information regarding the dataset collected for this study as well as the data pre-processing techniques and further techniques employed to make the data ready for analysis. We also give a comparison of six machine learning algorithms used to develop our machine learning calorie expenditure model. It is good to know the objectives and essential theory behind the problem. But practically, the prediction is a totally different scenario. So, everyone needs to be aware of the practical aspect of the problem and how it’s going to be implemented practically.

We will use machine learning to make influences from the provided dataset and predict the calorie expenditure. The machine learning problem of regression involves training a model to assign input data instances to a predefined set of classes or categories. Here are the general steps involved in a regression problem using machine learning.



2.1 DATASET DESCRIPTION

Data collection is an essential process in any machine learning project as the quality of the data used has a significant impact on the performance of the resulting model. We used Kaggle as our dataset store. Kaggle is a popular platform for data scientists and machine learning practitioners to access and share datasets. We used two CSV files which hold 15000 records for 7 attributes. The attributes range from gender, age, height, weight, body temperature during exercise, heart rate, and duration of workout in the "exercise.csv" file which is used as training data. "calories.csv" contains the corresponding values of calories burnt by individuals of the exercise dataset. Once the datasets were collected, it was uploaded to Google Colab.

2.2 VARIABLE DESCRIPTION

- User_ID – the unique ID number given to every individual
- Gender – the gender of each person, ie, male or female
- Age – the age of each person
- Height – the height of each person
- Weight – weight of each person
- Duration – the number of minutes exercised by each person
- Heart_Rate – the heart rate measured in beats per minute (BPM)
- Body_Temp – the body temperature of each individual measured in Celsius
- Calories – the number of calories burned during exercising

2.3 DATA PREPROCESSING

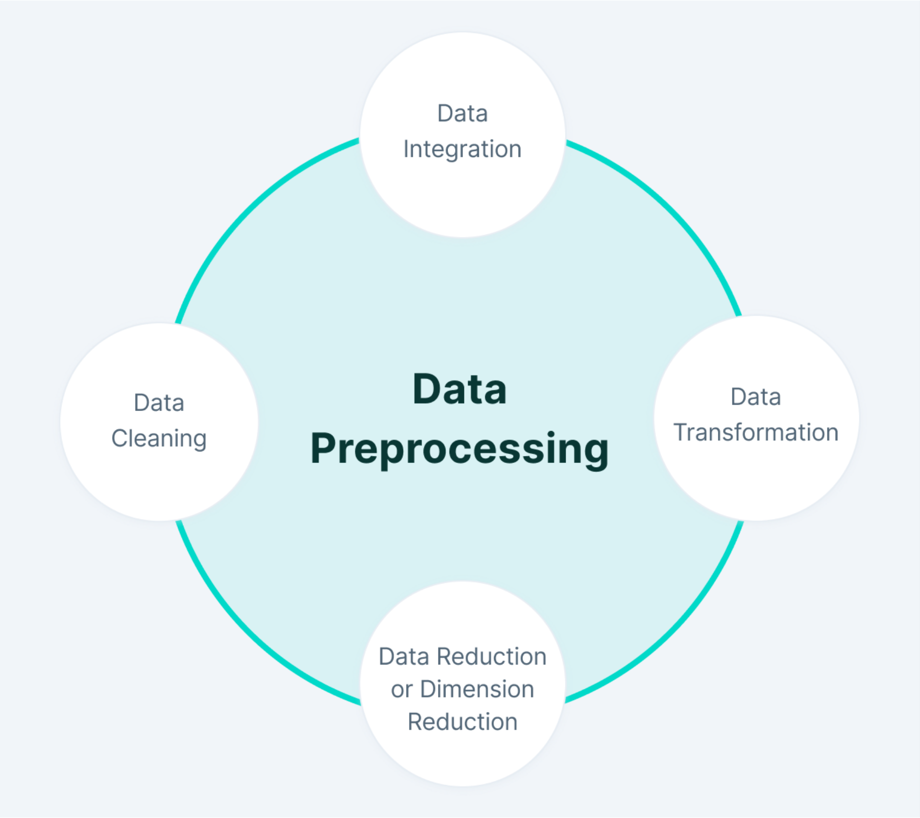
The analysis of the collected dataset starts with data preprocessing.

Data preprocessing is a critical step in the machine learning process aimed at preparing raw data for model training. It involves a series of operations to clean, transform, and improve the quality of the data, ensuring that it is suitable for analysis by machine learning algorithms.

A real-world data usually involves noises, missing values and maybe in an unusable format which cannot be used directly for machine learning models. Data preprocessing is necessary for improving the overall quality of the data and making it suitable for a machine learning model. This will increase the accuracy and efficiency of the model.

Data preprocessing involves the following steps:

- Data Collection
- Data Cleaning
- Data Transformation
- Feature Shaping
- Splitting the dataset into training and test set
- Feature scaling



Data Cleaning: Data cleaning is a crucial step in dealing with real-world data, as it often contains missing values, noise and inconsistencies. The main objective of data cleaning is to address these issues, such as filling in missing values, removing noisy data, identifying outliers and resolving data discrepancies. This will improve the accuracy and reliability of the data which is essential for training and using machine learning models.

Missing Data: Missing data is a common issue in datasets. We can fill in missing values using various methods. If an entire row or column contains a significant number of missing values, it can be dropped from the dataset. In cases where there are only a few missing values, interpolation methods like mean, median or mode imputation can be used to estimate the missing values.

Noisy data: Noisy data refers to data that contains random or irrelevant variations, errors, or inconsistencies that interfere with the analysis or interpretation of the data. This issue can be addressed through different techniques. Binning methods involves grouping continuous numerical data into discrete intervals or bins, replacing each segment with its mean or boundary. Clustering techniques group related data, allowing outliers to be identified or disregarded. Regression can also be employed to smooth out data by fitting it to a regression function.

Detecting Outliers: Outliers are then detected in the dataset to ensure data quality. Outliers are data points that deviate significantly from the rest of the dataset, potentially indicating errors, anomalies, or rare events that can distort analysis or modelling results. Identifying outliers is an important step in the data preprocessing method. For this, we draw boxplots. Boxplots are a standardized way of displaying the distribution of the data based on a five number summary. This type of plot is used to easily detect outliers.

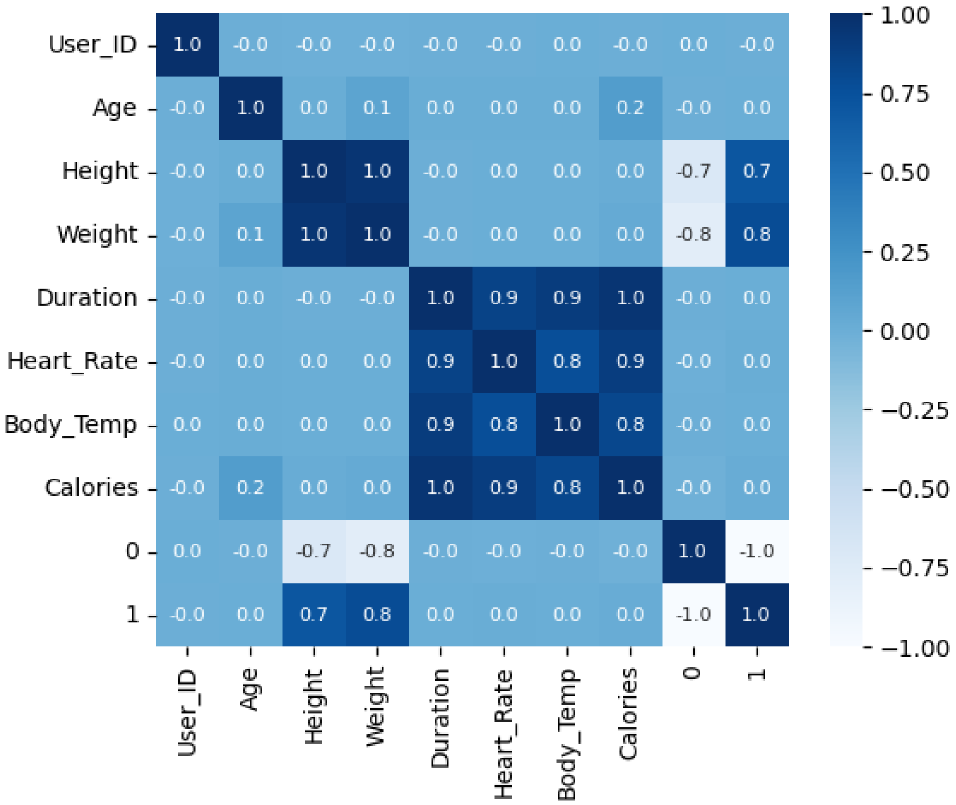
Encoding: When dealing with categorical data, encoding is a technique used in machine learning and data analysis to convert categorical data into numerical data. It is particularly useful when working with algorithms that require numerical output, as most machine learning models can only operate on numerical data. Two frequently used methods for encoding categorical variables are One-hot encoding and Label encoding.

In our data, the values under the variable ‘Gender’ indicates whether the individual is a male or female. Label encoder is used to convert these string values into numerical values ‘1’ and ‘0’ where ‘1’ represents ‘male’ and ‘0’ represents ‘female’.

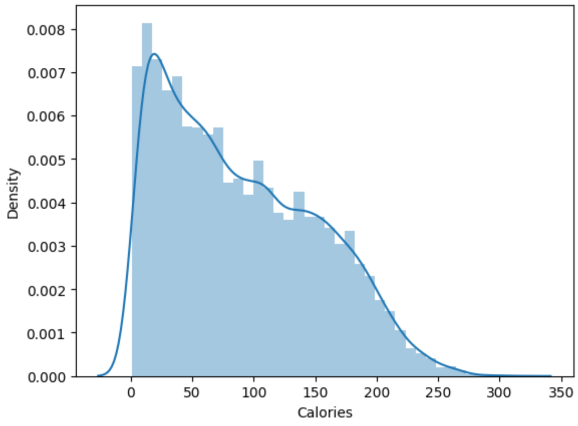
Data Reduction: In this step, irrelevant and redundant attributes are dropped from the dataset. For example, the variable ‘User_ID’ can be considered irrelevant for this research, as the focus is on analysing caloric expenditure. Hence the entire column of ‘User_ID’ is dropped from the dataset.

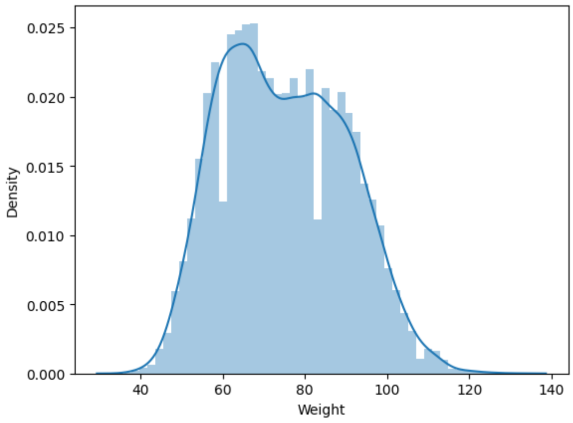
Data Visualization: Data visualization is the graphical representation of data and information to help users understand and interpret complex datasets more easily. It involves the creation of visual elements such as charts, graphs and maps to convey patterns, trends, relationships, and insights within the data. Data visualization converts large and small data sets into visuals which is easy to understand and process for humans. Here we consider distplots, heatmaps and barplots for effective understanding of the data.

Correlation in the datasets among features is illustrated in the heatmap below.

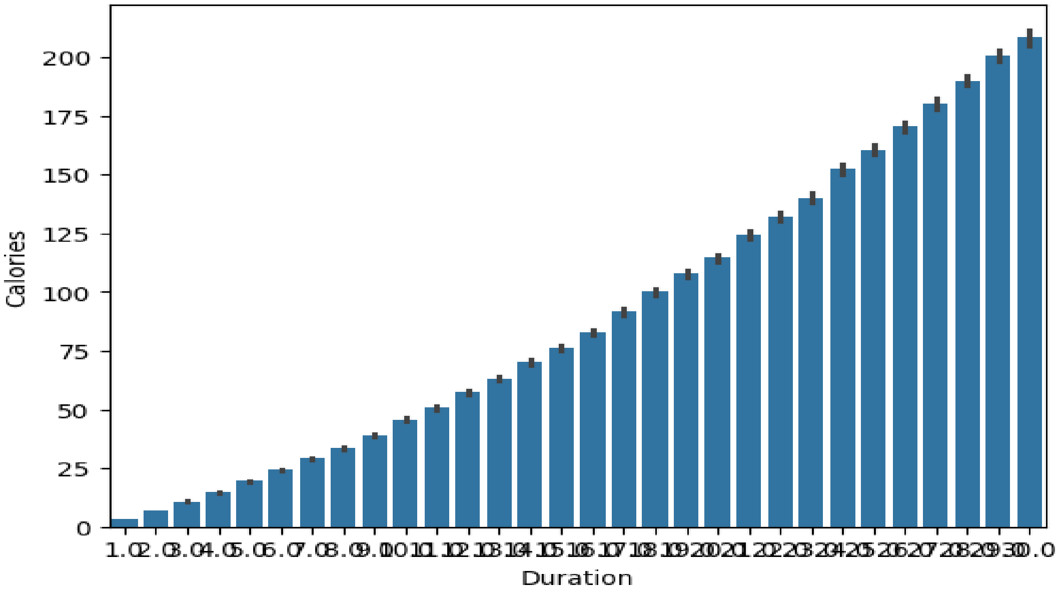


The central tendency, spread, and shape of the data is illustrated through the distplots given below.





The barplot below shows the comparison between ‘Calories’ and ‘Duration’.

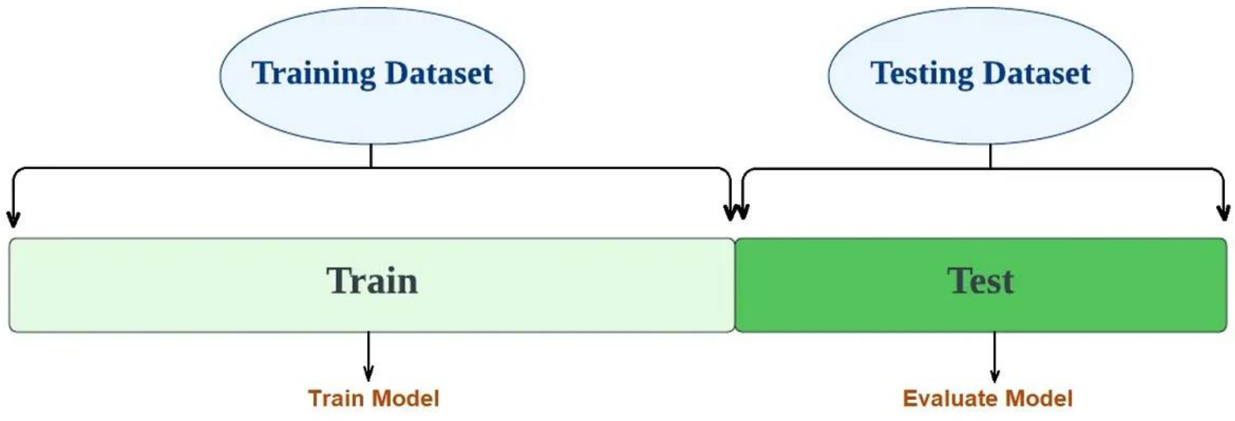


Splitting the dataset: We preprocessed the data by removing missing values and outliers. Next, we divide the dataset into a training set and testing set. This is crucial in machine learning and predictive modelling. Firstly, it allows us to assess the performance of a trained model on unseen data, which helps in evaluating its generalization ability. Suppose, we gave training to our machine learning model on a dataset and we test it using a completely different dataset, it will create difficulties for our model to understand the correlations between the models. So, we try to make a model that works well with the training set and also with the test dataset.

The training set is used to train the machine learning model on the data, allowing it to learn patterns and relationships between input features and the target variable. This set typically accounts for the majority of the dataset, usually around 70-80%.

The testing set is used to evaluate the performance of the trained model on the data.

In this project, we've used 80% of the data as training data and the remaining 20% as testing data.



Feature Scaling: Feaure scaling is the final step of the preprocessing technique.It is used to standradize or normalize the range independent variables or features in a dataset. In feature scaling, we put our variables in the same range and in the same scale so that no variable dominates the other variable.

Here, we import StandardScaler class of sklearn.preprocessing library for feature scaling in our dataset. To standardize a feature, we first calculate the mean and standard deviation of the feature across all data points. For each data point, subtract the mean from the original value of the feature. Finally, we divide the result by the standard deviation. This process will set all the variable values between 0 and 1.

2.4 ARTIFICIAL INTELLIGENCE

Artificial intelligence, or AI, is technology that enables computers and machines to simulate human intelligence and problem-solving capabilities. It aims to create systems capable of performing tasks that typically require human intelligence. This involves the development of algorithms and models that enable machines to interpret complex data, recognize patterns, and make informed decisions.

AI systems are basically software systems that use techniques such as machine learning and deep learning to solve problems in particular domains without manually coding all possibilities in software. Due to this, AI started showing promising solutions for industry and businesses as well as our daily lives.

2.5 MACHINE LEARNING

Machine learning is a branch of artificial intelligence (AI) that focuses on the development of algorithms and statistical models that enable computers to learn from and make predictions based on data, without being explicitly programmed. It involves the development of algorithms and statistical models that enable computers to automatically improve their performance on a given task through experience, without human intervention. These algorithms and models are designed to learn from data and make predictions without explicit instructions.

Arthur Samuel, an American leader in the field of machine learning introduced the term "Machine Learning" in 1959. He described machine learning as: "The field of study that gives computers the ability to learn without being explicitly programmed." Tom Mitchell gave a more modern definition: "A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E."

The ultimate goal of machine learning is to develop models that can be understood and utilized by people using an input data.

Applications Of Machine Learning:

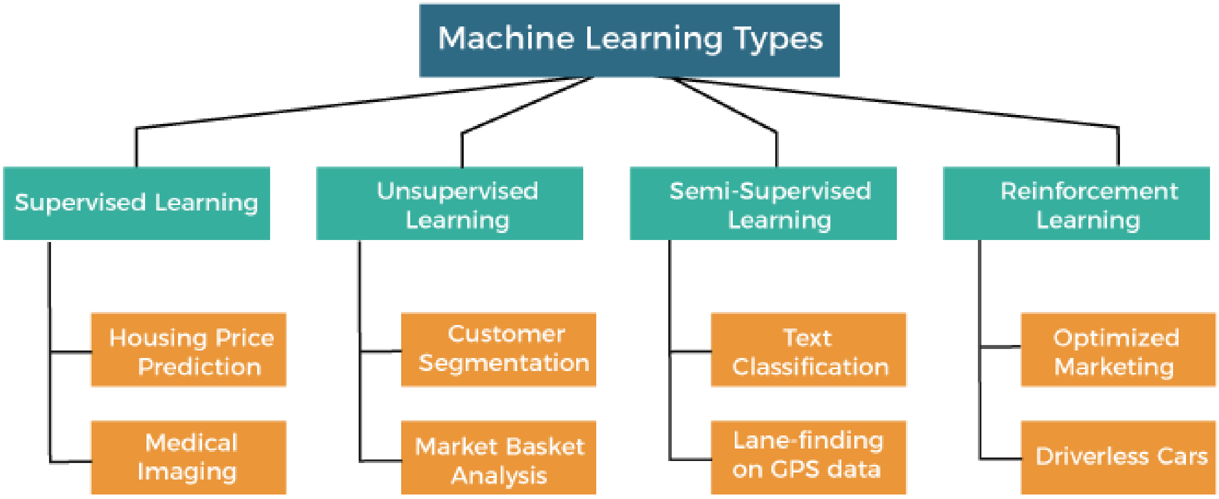
Machine learning has a wide range of applications across various fields. Some notable applications are:

1. Image Recognition
2. Speech Recognition
3. Recommender Systems
4. Fraud Detection
5. Self-Driving Cars
6. Medical Diagnosis
7. Stock Market Trading
8. Virtual Try On
9. Traffic Prediction

Types Of Machine Learning:

Machine learning algorithms can be classified into four groups. The categorisation is based on how learning is received or how feedback on the learning is given to the system developed.

1. Supervised Learning
2. Unsupervised Learning
3. Semi Supervised Learning
4. Reinforcement Learning



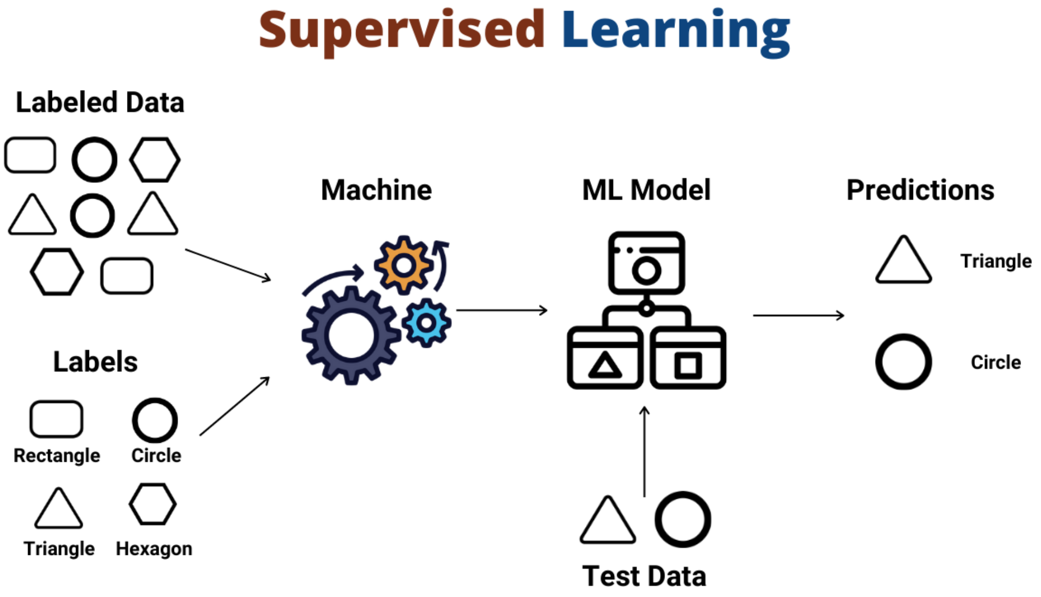
2.5.1 SUPERVISED LEARNING

In supervised learning, algorithms learn from labelled dataset. A labelled dataset means that it has both input and output parameters. The goal is to learn a mapping or relationship between the input features and the output labels, enabling the algorithm to generalize and make accurate predictions on new data. Supervised learning therefore uses patterns to predict label values on additional unlabelled data.

Supervised learning problems can be categorized into two: “classification” and “regression” problems. In classification, the output variable is categorical and the goal is to assign instances to predefined classes or categories. In regression problems, the output variable is continuous and the goal is to predict a numerical value.

Examples of supervised learning include image classification, speech recognition and predicting the price of a house based on its features.

The working of supervised learning can be understood from the diagram given below.



1. Classification

In classification problems, we try to predict results in a discrete output. Given one or more inputs, a classification model will try to predict the value of one or more outcomes. Here the target variable is a categorical value. Email spam detection can be regarded as an example of classification, where the inputs are emails and the classes are “spam” and “not spam”. Various classification algorithms include,

1. Logistic Regression
2. Decision Tree Classifier
3. Random Forest Classifier
4. Naive Bayes
5. SVM (Support Vector Machine)

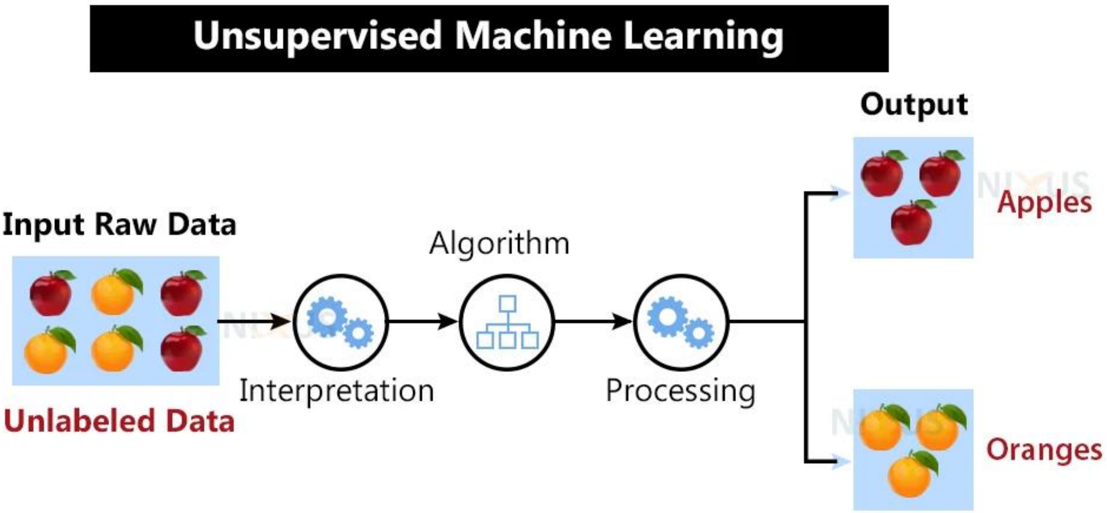
2.Regression

In regression problems, the output variables will be real or continuous values. The goal of regression is to predict the values of target variable on the basis of the input variables. The regression model captures the underlying patterns or trends in the data, allowing us to make predictions about the target variable for new or unseen data points. The various regression algorithms are,

1. Linear Regression
2. Decision Tree Regression
3. Random Forest Regression
4. Lasso Regression
5. KNN Regression
6. Stepwise Regression

2.5.2 UNSUPERVISED LEARNING

In unsupervised learning the data is unlabelled, so the learning algorithm is left to find the structures or patterns in the outputs. Unsupervised learning deals with raw, unstructured data and seeks to discover hidden patterns, structures, or relationships within the dataset. We can derive this structure by clustering the data based on relationships among the variables in the data. Although the goal maybe to discover hidden patterns within the dataset, it may also have a goal of feature learning. This allows the machine to automatically discover the representations that are needed to classify the raw data. The example given below depicts the working of unsupervised learning.



1. Clustering

Clustering algorithms aim to group similar data points together into clusters based on their intrinsic similarities. It is defined as the task of dividing the population or data pints into number of smaller groups such that the points are more similar to the other points in the same group and dissimilar to data points in another group. Basically, it is a collection of objects on the basis of similarity and dissimilarity between them. The main clustering methods are:

1. Agglomerative Clustering
2. Hierarchical Clustering
3. Divisive method
4. K-Means Clustering

2. Dimensionality Reduction

Dimensionality reduction is the task of reducing the number of features in a dataset. In machine learning tasks like regression or classification, there are often too many features to work with. The higher the number of features, the more difficult it is to model them. The process of dimensionality reduction essentially transforms the data from high-dimensional feature space to a low-dimensional feature space, thereby reducing the number of redundancies and noisy data from the dataset.

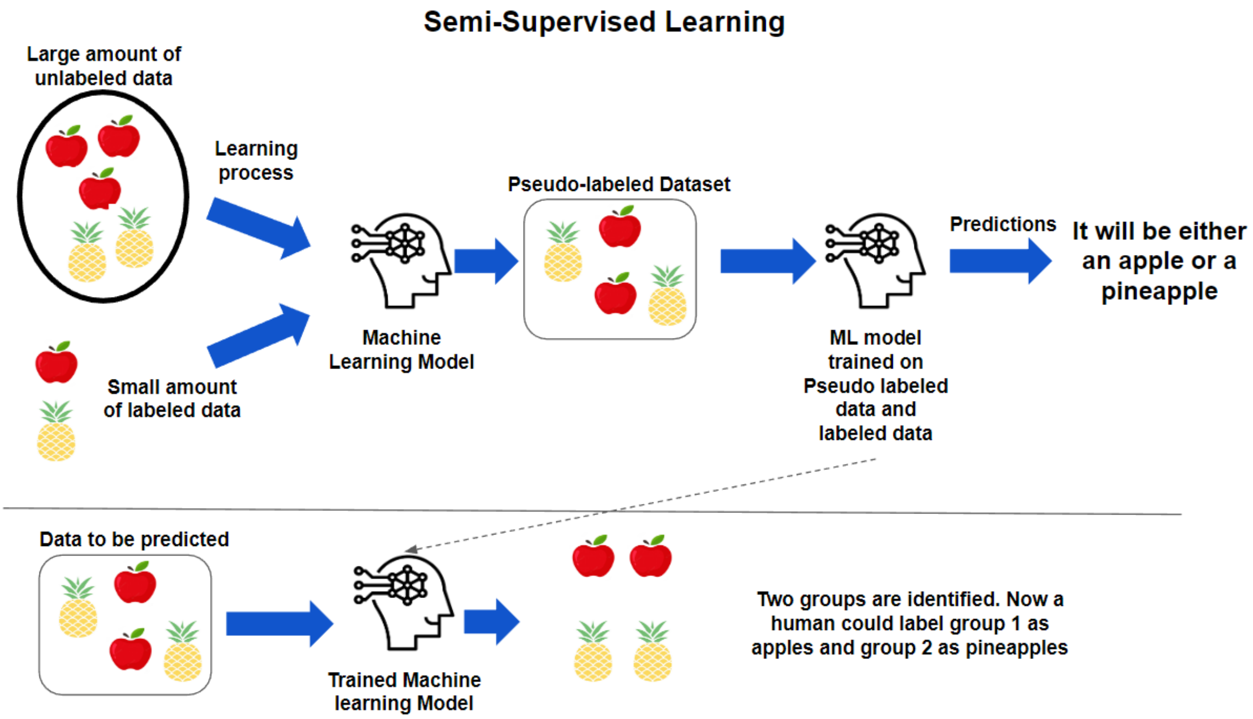
3. Association

Association rule learning is the task of discovering interesting relationships or associations between variables in large datasets. It involves identifying patterns of co-occurrence or relationships between items, features, or events within the data. Common applications of association learning include market basket analysis where we aim to discover associations between items occurring frequently.

2.5.3 SEMI-SUPERVISED LEARNING

Semi-supervised learning is a type of machine learning that falls in between supervised and unsupervised learning. It is a method that uses a small amount of labelled data and a large amount of unlabelled data to train a model. The goal of semi-supervised learning is to learn a function that can accurately predict the output variable based on the input variables, similar to supervised learning. However, unlike supervised learning, the algorithm is trained on a dataset that contains both labelled and unlabelled data.

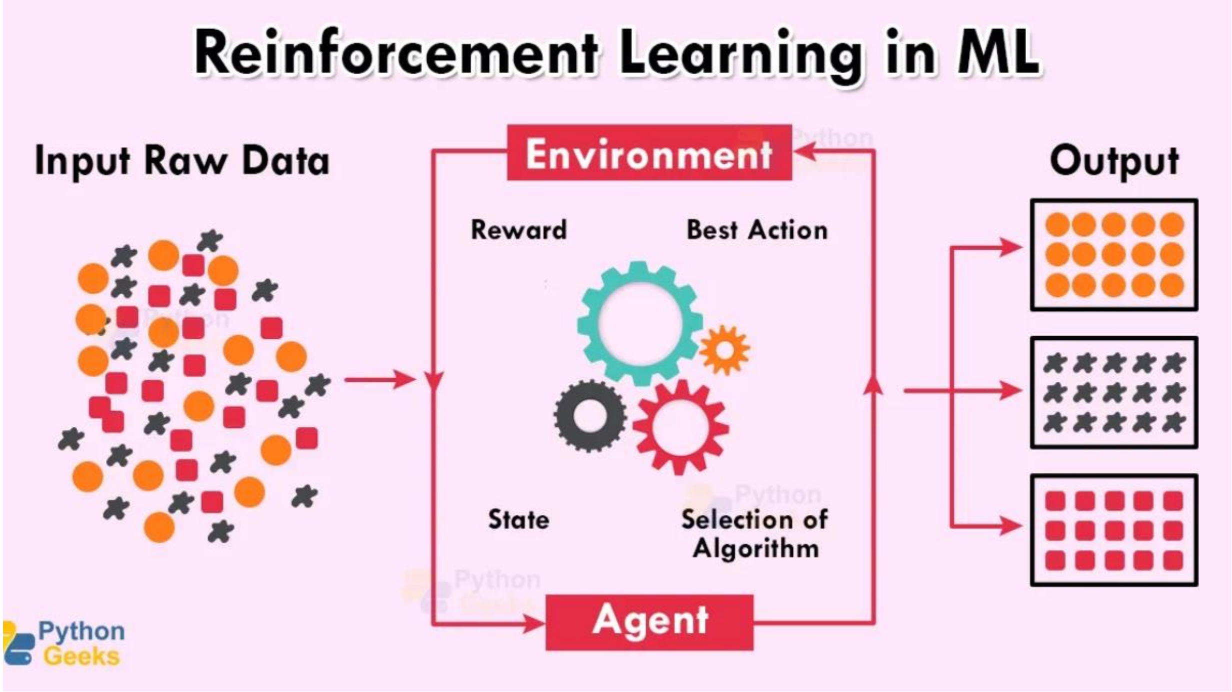
One of the key techniques in semi-supervised learning is self-training, where an initial model is trained on the labelled data, and then this model is used to predict labels for the unlabelled data. The most confident predictions are added to the labelled dataset, and the model is retrained. The following figure shows the working of a semi-supervised model:


Semi-Supervised Learning
Large amount of unlabeled data
Learning process
Small amount of labeled data
Machine Learning Model
Pseudo-labeled Dataset
ML model trained on Pseudo labeled data and labeled data
Predictions
It will be either an apple or a pineapple
Data to be predicted
Trained Machine learning Model
Two groups are identified. Now a human could label group 1 as apples and group 2 as pineapples

Semi-supervised learning is beneficial when the cost involved with labelling is too high to allow for a fully labelled training process. Some examples of semi-supervised learning include text classification, image classification, anomaly detection.

2.5.4 REINFORCEMENT LEARNING

Reinforcement Learning is a technique that enables an agent to learn in an interactive environment by trial and error using feedback from its own actions and experiences. It performs actions with the aim of maximizing rewards, or in other words, it is learning by doing in order to achieve the best outcomes. Reinforcement learning differs from supervised learning in a way that in supervised learning the model is trained with the correct answer itself whereas in reinforcement learning, there is no answer but the reinforcement agent decides what to do to perform the given task. The aim of the algorithm is to select the actions which maximises the expected reward over a period of time. The agent will find the destination much faster by utilising this policy.



2.6 REGRESSION ANALYSIS

Regression analysis is a statistical method used to understand the relationship between a dependent variable and one or more independent variables. It is commonly employed to predict or estimate the value of the dependent variable based on the values of the independent variables. The primary goal of regression analysis is to model the relationship between variables and make predictions or inferences about the dependent variable. It predicts continuous/real values such as temperature, age, salary, price, etc. Three major uses for regression analysis are determining the strength of predictors, forecasting an effect and trend forecasting.

There are times when we would like to analyse the effect of different independent features on the target or what we say dependent features. This helps us make decisions that can affect the target variable in the desired direction. Regression analysis is heavily based on statistics and hence gives quite reliable results. Due to this reason, regression models are used to find the linear as well as non-linear relation between the independent and the dependent variable.

Regression analysis is widely used in various fields, including economics, finance, social sciences, engineering, and healthcare, to understand complex relationships, make predictions, and inform decision-making processes.

Major steps involved in Regression Analysis:

The key steps involved in performing regression analysis typically include the following:

1. Data preparation: In this step, we gather the relevant data needed for the analysis, clean the data, handle missing values, and preprocess the various features.
2. Exploratory Data Analysis (EDA): Next, we try to understand the data through visualization and descriptive statistics, identify relationships between variables.
3. Model Selection: The most appropriate regression algorithm is chosen based on the problem and data characteristics.
4. Model Training: We split the data into training and testing sets and fit the selected model to the training data.
5. Model Evaluation: Assess the model's performance using suitable metrics on the testing data.

6. Model Deployment and Monitoring: Deploy the trained model in a production environment where it can make predictions on new data and monitor its performance over time.

Applications of Regression

- Sales forecasting: Predicting future sales based on historical sales data, marketing expenditure, seasonality, economic factors and other relevant variables.
- Customer lifetime value prediction: Estimating the potential value of a customer over the customer's entire relationship with the company based on past purchase history, demographics and behaviour.
- Employee performance prediction: Predicting the performance of employees based on various factors such as training, experience and demographics.
- Financial performance analysis: Understanding the relationship between financial metrics (e.g., revenue, profit) and key drivers (e.g., marketing expenses, operational costs).
- Risk analysis and fraud detection: Predicting the likelihood of events such as credit defaults, insurance claims, or fraud based on historical data and risk indicators.
- Maintenance prediction: Predicting time to failure of critical parts and machinery.

Terminologies related to Regression Analysis

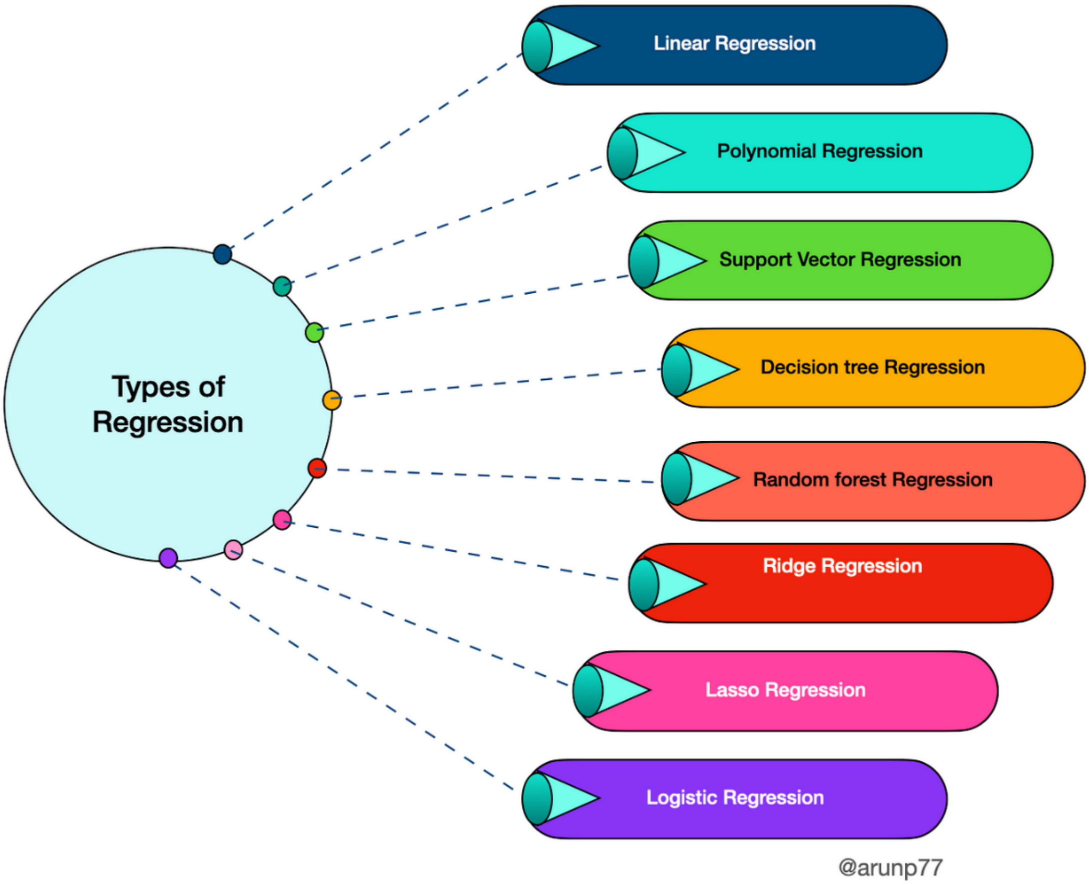
- **Dependent Variable:** It is the main factor in regression analysis in which we want to predict or understand. It is also called target variable.
- **Independent Variable:** Also called predictor variables, independent variables are the variables used to predict or explain changes in the dependent variable.
- **Outliers:** Outliers are data points that significantly deviate from the rest of the data in a dataset. An outlier may hamper the result, so it should be avoided.
- **Multicollinearity:** Multicollinearity occurs when two or more independent variables in a regression model are highly correlated with each other. It should not be present in the dataset, because it creates problems while ranking the most affecting variable.
- **Underfitting and overfitting:** Underfitting occurs when a model is too simple to capture the underlying patterns in the data, resulting in poor performance on both the

training and test datasets. Overfitting, on the other hand, happens when a model learns the noise and random fluctuations in the training data too well, leading to excellent performance on the training data but poor generalization to unseen data.

2.7 TYPES OF REGRESSION

There are various types of regressions which are used in data science and machine learning. Each type has its own importance on different scenarios, but at the core all the egression methods analyse the effect of the independent variable on dependent variables. Here we discuss some important types of regression which are given below:

- Linear Regression
- Polynomial Regression
- Support Vector Regression
- Decision Tree Regression
- Random Forest Regression
- Ridge Regression
- Lasso Regression



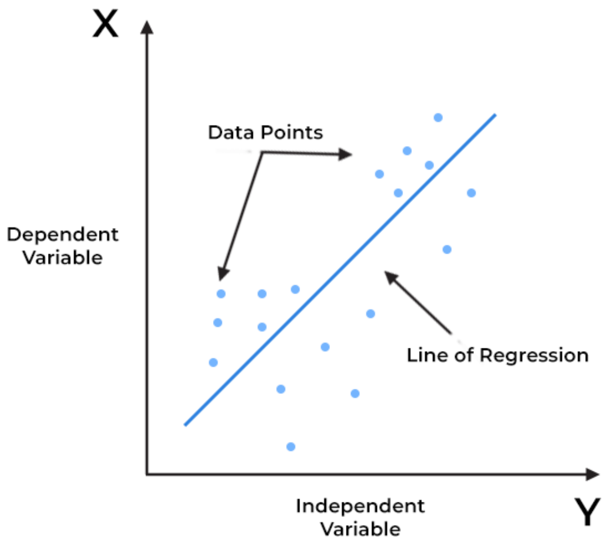
2.7.1 LINEAR REGRESSION

Linear regression is one of the easiest and most popular machine learning algorithms. It is a supervised learning algorithm used in machine learning and statistics for modelling the relationship between a dependent variable and one or more independent variables. The algorithm assumes a linear relationship between the dependent variable and the independent variables. This linearity is represented by a straight line in a two dimensional space or by a hyperplane in higher dimensions.

In linear regression, the goal is to find the best fitting line (or hyperplane) that minimizes the difference between the observed values of the dependent variable and the values predicted by the linear model. This is done by minimizing the sum of the squared differences between the observed and predicted values. Linear regression is widely used in various applications such as market analysis, financial study, making recommendations based on given data, etc. It is favoured for its simplicity, interpretability and computational efficiency.

There are mainly two types of linear regression algorithms:

1) Simple Linear Regression
2) Multiple Linear Regression



Working Of Linear Regression Algorithm

The various steps involved in the working of a linear regression algorithm is as follows:

1. Model Representation: Linear Regression models the relationship between independent variables X and a dependent variable Y using a linear equation:

   Y=β0 + β1X1 + β2X2 + ... + βnXn + ε

   Here, β0 is the intercept, β1, β2,…., βn are the coefficients of the independent variables, and ε is the error term.
2. Training: During the training phase, the goal is to find the values of the coefficients that best fit the data. This is done by minimizing the sum of squared differences between the

   observed and predicted values of the dependent variable.
3. Prediction: Once trained, the model can predict Y for new values of X using the learned coefficients and the linear equation.

4. Evaluation: The model's performance is evaluated using metrics like Mean Squared Error (MSE), Mean Absolute Error (MAE), etc, to assess its accuracy and fit to the data.
5. Interpretation: The coefficients indicate the impact of each feature on Y, providing insights into the relationship between variables.

Advantages:

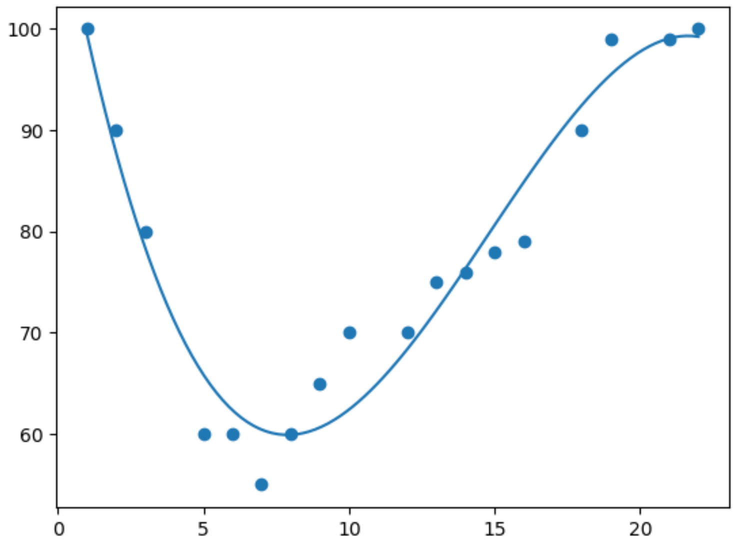
1) Linear Regression is easy to understand and interpret, making it accessible to both beginners and experts in the field.
2) Training and prediction with Linear Regression are efficient, making it suitable for large datasets and real-time applications.
3) Linear Regression can be applied to both regression and classification tasks.
4) Linear Regression does not make assumptions about the distribution of the data.

Disadvantages:

1) Linear Regression assumes a linear relationship between independent and dependent variables, which may not hold true for all datasets.
2) Without regularization techniques, Linear Regression is prone to overfitting, especially when the number of features is large relative to the number of observations.
3) Linear Regression is sensitive to outliers in the data, which can significantly affect the model's coefficients and predictions.

2.7.2 POLYNOMIAL REGRESSION

Polynomial Regression is a type of regression in which the relationship between the independent variable x and dependent variable y is modelled as an nth-degree polynomial. Unlike linear regression, which assumes a linear relationship between variables, polynomial regression can capture more complex relationships by fitting a curve to the data. In polynomial regression, the original features are transformed into polynomial features of given degree and then modelled using a linear model. This means that the datapoints are best fitted using a polynomial line.



In polynomial regression, the regression model takes the form $Y = a_0 + a_1 X + a_2 X^2 + \ldots + a_n X^n$ where, a_0, a_1,..., a_n are the coefficients and $X, X^2, \ldots, X^n$ represent the polynomial terms of the independent variable up to the nth degree. The model is still linear as the coefficients are still linear with quadratic.

Implementing polynomial regression involves selecting an appropriate degree for the polynomial, typically through techniques like cross-validation to avoid overfitting. Additionally, feature scaling and regularization techniques can help improve the performance and stability of the model.

Applications Of Polynomial Regression

The reason behind the vast use of polynomial regression is that approximately all real-world data is non-linear in nature and hence when we fit a non-linear model on the data or a curvilinear regression line then the results that we obtain are far better than what we can achieve with the standard linear regression. Some of the uses of polynomial regression are as follows:

- The growth rate of tissues
- Progression of epidemic diseases
- Economic forecasting

The basic goal of regression analysis is to model the expected value of a dependent variable in terms of an independent variable.

Advantages:

- Polynomial regression can capture a wide range of relationships, from linear to highly nonlinear, making it suitable for diverse datasets.
- Unlike some complex machine learning models, polynomial regression equations are often interpretable, allowing practitioners to understand the impact of each predictor variable on the outcome.
- Polynomial regression can directly model the nonlinear relationships without data manipulation.

Disadvantages:

- As the degree of the polynomial increases, the model becomes increasingly complex and may overfit the training data.
- Higher-degree polynomials can significantly increase the computational complexity of the model, especially with large datasets.
- They are too sensitive to outliers.

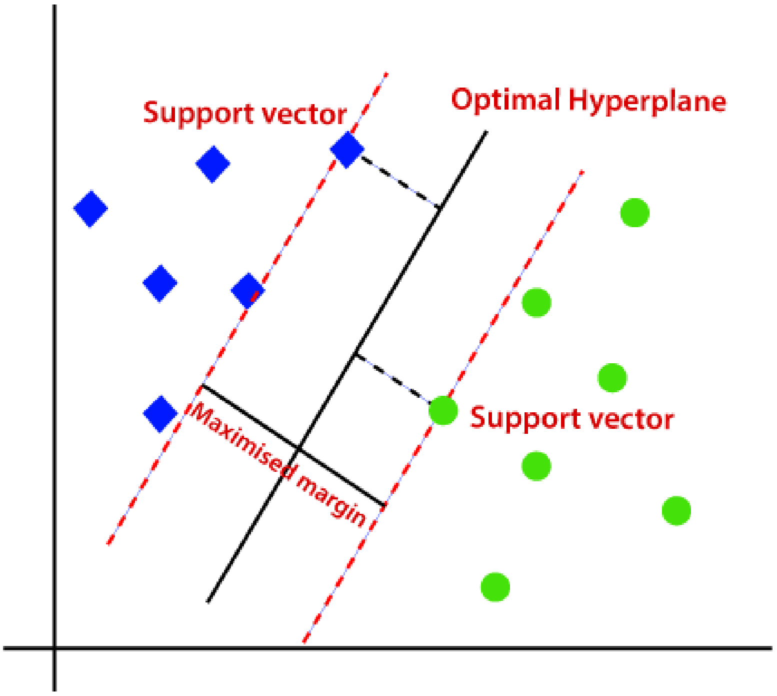
2.7.3 SUPPORT VECTOR REGRESSION

Support Vector Machines (SVMs) are supervised learning models used for classification and regression tasks. The main concept of Support Vector Machine is finding the hyperplane (a flat subspace of n-dimensional space that divides the space into two parts) with the maximum margin that separates classes in the feature space, thereby ensuring minimal classification. The key idea is to identify support vectors, which are data points closest to the hyperplane and crucial for defining its position. Some of the terminologies used in SVM are:

- Hyperplane: It is a separation line between two classes.
- Kernel: It is a function used to map a lower-dimensional data into higher dimensional data.

- Support Vectors: Support vectors are the datapoints that are nearest to the hyperplane.
- Boundary Line: They are two lines apart from hyperplane, which creates a margin for datapoints.

SVMs are characterized by their capacity for margin maximization and generalization. Unlike some other classifiers that may focus solely on reducing training error, SVMs prioritize maximizing the margin between classes, which helps to reduce overfitting and improve the model's ability to generalize the data.



Applications of SVR:

- Text classification and sentiment analysis
- Object recognition in image recognition
- Malware detection in cybersecurity
- Stock price prediction
- Patient outcome prediction

Advantages:

- SVMs can be used in high-dimensional feature spaces, making them suitable for tasks with a large number of features.
- SVMs have built-in mechanisms to combat overfitting.
- By relying only on a subset of training data (support vectors), SVMs maintain memory efficiency, making them suitable for applications with large datasets.

Disadvantages:

- Training SVMs, especially with nonlinear kernels, can be computationally intensive, which limits their suitability for large datasets.
- SVMs don't naturally give probability estimates. We often use technique cross-validation to estimate probabilities instead.
- The effectiveness of SVMs depends on the kernel function and its parameters, so choosing and tuning them carefully is crucial for achieving the best results.

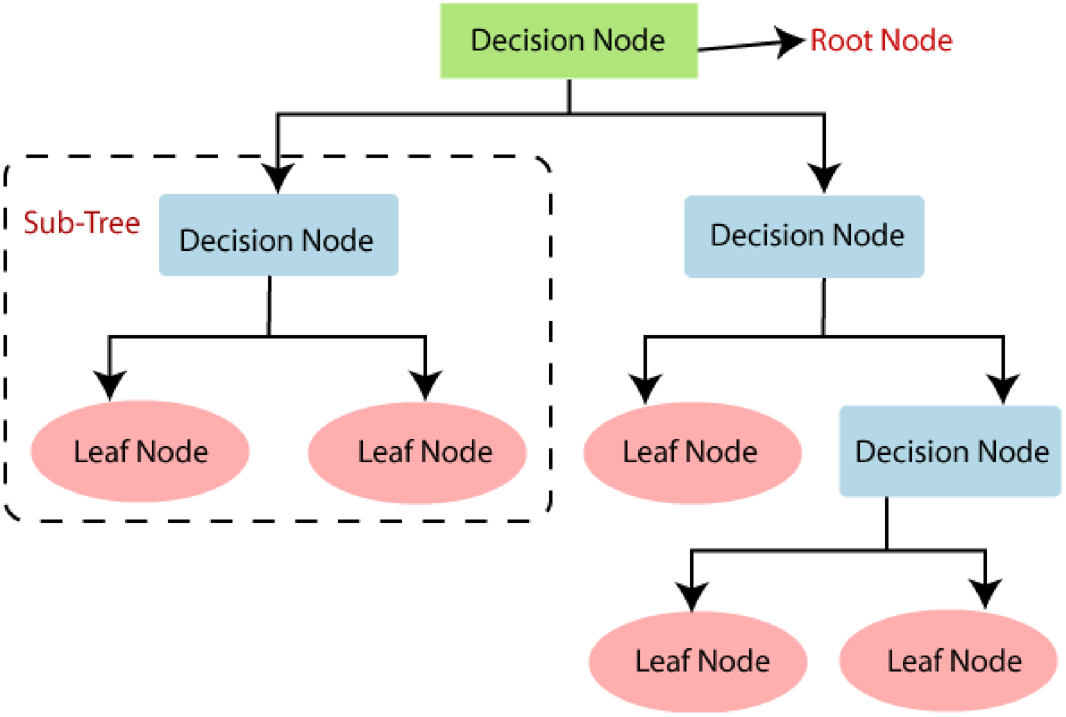
2.7.4 DECISION TREE

A decision tree is a supervised machine learning algorithm used for both classification and regression. It models decisions using a tree-like structure where each internal node represents a decision based on a feature, each branch represents the outcome of that decision, and each leaf node represents the final decision or outcome.

A decision tree or a classification tree is a tree in which the paths from root to leaf represent classification rules. At each node of a decision tree, different types of questions are asked based on the features or attributes of the data. The goal is to split the data into subsets that are more homogeneous with respect to the target variable. Decision trees are used in various applications such as medical diagnosis, anomaly detections, classification and regression tasks. It is one of the most widely used and practical methods for inductive inference.

There are mainly two types of decision tress.

1) Classification Trees
2) Regression Trees



Working of a Decision Tree Algorithm

The various steps involved in the working of a decision tree are:

1. Data Preparation: The process starts with collecting and preprocessing the dataset. This includes handling missing values, encoding categorical variables, and splitting the data into training and testing sets.
2. Splitting: It is the process of partitioning of data into subsets. The decision tree algorithm selects the best feature to split the data at each node. The decision tree is constructed by splitting the data at each node based on the selected feature. The dataset is partitioned into subsets, and the process continues until a stopping criterion is met.
3. Prediction: Once the decision tree is constructed, it can be used to make predictions on new instances.
4. Evaluation: Finally, the performance of the decision tree model is evaluated using metrics like mean squared error. Then we find the smallest tree that fits the data.

Advantages:

1) A decision tree is simple to understand as it follows the same process which a human follow while making any real-life decision.
2) Decision trees can handle both numerical and categorical data.
3) They make no assumptions about the distribution of the data. This makes them flexible and applicable to a wide range of problems.
4) They can handle missing values in the dataset by simply ignoring the missing values during the splitting process.

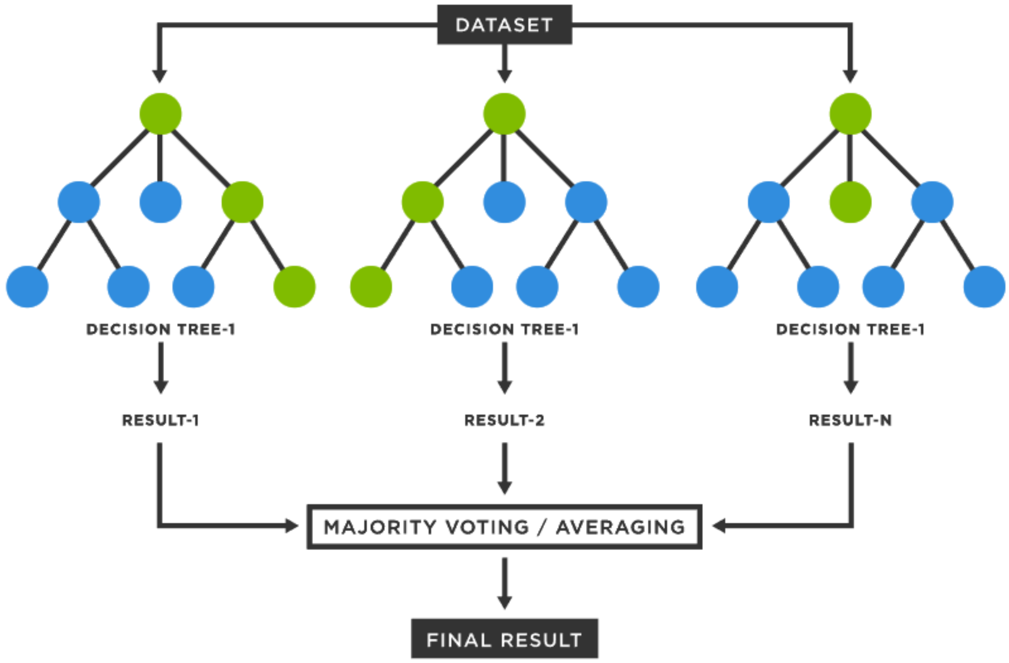
Disadvantages:

1) Decision trees may not capture complex relationships or interactions between features as effectively as other algorithms.
2) Only if the information is precise and accurate, the decision tree will deliver promising results. Even if there is a slight variation in the input data, it can cause large changes in the tree.
3) Decision trees are prone to overfitting, especially when the tree grows too deep. This can lead to poor performance of unseen data.

2.7.5 RANDOM FOREST

Random forest is an ensemble learning method where multiple decision trees are constructed and they are merged to get a more accurate prediction.

The main idea behind Random Forest is the combination of multiple decision trees, each trained on a random subset of the training data and using a random subset of features for splitting at each node. This randomness injected into the training process helps to reduce overfitting and increase the model's performance. Random forests are known for their robustness and high performance across various domains.



Working of a Random Forest Algorithm

Here is an outline of the random forest algorithm.

1. Creating Multiple Decision Trees: The algorithm begins by creating a collection of decision trees, each trained on a random subset of the training data.
2. Random Sampling: It then creates multiple datasets by randomly sampling with replacement from the original dataset. Each decision tree is trained on one of these datasets.
3. Splitting: At each node of each decision tree, a random subset of features is considered for splitting. This will help preventing overfitting.
4. Averaging: For regression tasks, the predictions of all the decision trees are averaged to produce the final prediction. It means that the final prediction is based on the collective predictions of multiple decision trees. This process is called ensemble learning.

Advantages:

1) It can run efficiently on large data bases.

2) It is less prone to overfitting compared to individual decision trees.
3) Random forest provides high accuracy in both classification and regression tasks.
4) Random forest algorithm can be applied to a wide range of machine learning tasks, including classification and regression.

Disadvantages:

1) Random forest algorithms, when used for regression, cannot predict beyond the range in the training data, and they may over-fit data sets that are particularly noisy.
2) The sizes of the models created by the random forest may be very large. It may take hundreds of megabytes of memory and may be slow to evaluate.
3) The algorithm is less interpretable compared to individual decision trees.

2.7.6 RIDGE REGRESSION

Ordinary least squares (OLS) regression provides unbiased estimates with minimum variance within the linear class. However, relaxing the unbiasedness condition allows for biased estimators with smaller variance. Ridge regression introduces bias by shrinking coefficients, reducing variance. This trade-off makes ridge regression advantageous, especially in scenarios of multicollinearity or high-dimensional data, where it outperforms OLS by giving importance to variance reduction over strict unbiasedness.

Ridge regression, also known as L2 regression is one of the most robust versions of linear regression in which a small amount of bias is introduced so that we can get better long-term predictions. The amount of bias added to the model is known as ridge regression penalty. We can compute this penalty term by multiplying the tuning parameter (λ) with the sum of squared coefficients. A linear or polynomial regression will fail if there is high collinearity between the independent variables. To solve this problem, we use ridge regression. It helps to solve the problems if we have more parameters than samples.

Ridge regression is widely used in various fields, including economics, finance, biology, and engineering, where predictive modelling with correlated predictors is common. It's a powerful tool for improving the stability and performance of regression models, particularly in situations where multicollinearity and overfitting are concerns.

Applications of Ridge Regression:

1) Ridge regression is utilized to improve accuracy in predicting outcomes, such as stock prices or sales forecasts.
2) In medical research, ridge regression is employed to predict health outcomes and personalize treatment plans.
3) Ridge regression assists in weather forecasting, climate change prediction, and environmental impact assessment.
4) Ridge regression is applied in image and signal processing to denoise images, compress data, and extract features, improving the accuracy of processing algorithms.

Advantages:

1) Ridge regression effectively deals with multicollinearity, stabilizing estimates by penalizing large coefficients.
2) It protects the model from overfitting.
3) Ridge regression can deal with situations where the number of predictors exceeds observations.
4) Model complexity is reduced.

Disadvantages:

1) Solving ridge regression can be computationally complex, especially for large datasets.
2) Unlike Lasso regression, ridge regression does not perform variable selection, making it less suitable for feature selection tasks.
3) Ridge regression is sensitive to outliers which can leading to biased results.

2.7.7 LASSO REGRESSION

Lasso regression is another regularization technique to reduce the complexity of the model. It is similar to ridge regression except that penalty term contains only the absolute weights instead of squared weights. Since it takes absolute values, it can shrink the slope to 0, whereas in ridge regression, it can be shrunk to near 0.

Lasso stands for Least Absolute Shrinkage and Selection Operator. It is also called L1 regularization. It is frequently used in machine learning to handle high dimensional data as it facilitates automatic feature selection with its application.

Working of Lasso Regression:

1) Objective Function: Lasso regression minimizes the residual sum of squares between observed and predicted values while adding a penalty term based on the absolute values of coefficients and a regularization parameter λ.
2) Feature Selection: Lasso regression automatically selects relevant features by shrinking coefficients of less important variables to zero, thereby excluding them from the model.
3) Optimization: Lasso regression optimizes the objective function using iterative algorithms such as coordinate descent or gradient descent to update coefficients until convergence is achieved.
4) Regularization Parameter: The optimal value of regularization parameter λ is obtained using cross-validation techniques like k-fold cross-validation.
5) Model Evaluation: After determining the optimal value of λ value, the final Lasso regression model is trained on the entire dataset, and its performance is evaluated on new data using appropriate evaluation metrics.

Advantages:

1) Lasso regression automatically selects relevant features by shrinking coefficients to zero, simplifying models and enhancing interpretability.
2) It mitigates overfitting by penalizing large coefficients, improving the model's generalization performance.

3) It is effective in dealing with by selecting one variable from a group of highly correlated variables and setting the coefficients of the others to zero.
4) Lasso regression is relatively easy to implement and interpret compared to more complex models.

Disadvantages:

1) Lasso regression may introduce bias while reducing variance, particularly in complex relationships between predictor and the response variable.
2) It may arbitrarily select one of the correlated predictors, leading to instability in feature selection.
3) Lasso regression can face computational challenges with large sample sizes or high-dimensional data, increasing computational time and resource requirements.

2.8 MODEL EVALUATION

Model evaluation is a critical aspect of machine learning that determines the performance and effectiveness of predictive models. This paper explores various techniques and best practices for evaluating machine learning models, covering key concepts such as

performance metrics and cross-validation. By understanding and implementing evaluation methods, practitioners can make informed decisions about model selection, optimization, and deployment.

Metrics are used to evaluate the developed models and determine their overall performance. These metrics include the mean square error (MSE), mean absolute error (MAE), root mean square error (RMSE), R-square (R^2) and cross-validation (CV) score.

Mean Square Error:

Mean Squared Error (MSE) is a commonly used metric for evaluating the performance of regression models. It measures the average squared difference between the actual values and the predicted values produced by the model. The lower the MSE score, the better the model's performance. The formula to calculate MSE is:

$$MSE = \frac{\sum(y_i - \hat{y}_i)^2}{n}$$

Where,

- yi = Observed Value
- ŷl = Predicted Value
- n = number of observations

Mean Absolute Error:

Mean Absolute Error (MAE) is another common metric used for evaluating the performance of regression models. Similar to MSE, MAE measures the average absolute difference between the actual values and the predicted values produced by the model. Mathematically, MAE is calculated as:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|\widehat{y_i} - y_i|$$

Where:

$\widehat{y_i}$ = Predicted value for the ith data point

y_i= Actual value for the ith data point

n = number of observations

Root Mean Square Error:

Root Mean Square Error (RMSE) is yet another commonly used metric for evaluating the performance of regression models. It calculates the square root of the average of the squared differences between the predicted values and the actual values. The equation to calculate RMSE is as follows:

$$RMSE = \sqrt{\sum_{i=1}^{n}\frac{(\hat{y}_i - y_i)^2}{n}}$$

Where,

- yi = Observed Value
- ŷI = Predicted Value
- n = number of observations

RMSE provides a measure of the average magnitude of the errors between the predicted values and the actual values, taking into account both the magnitude and direction of the errors.

R^2 Value:

The R^2 value, also known as the coefficient of determination, is a statistical measure used to assess the goodness of fit of a regression model. It represents the proportion of the variance in the dependent variable that is explained by the independent variables in the model. It provides an assessment of how well the regression model fits the observed data. The R^2 value ranges between 0 and 1, with a higher value representing a better fit. The formula to calculate R^2 is given by:

$$R^2 = 1 - \frac{SS_{RES}}{SS_{TOT}} = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \overline{y})^2}$$

where,

- SS_{RES} is the sum of squared residuals
- SS_{TOT} is the total sum of squares

Overall, RMSE is a useful metric for evaluating the performance of regression models, providing insight into the typical magnitude of errors made by the model in its predictions.

Cross Validation Score (CV score):

Cross-validation is a technique used to evaluate the performance of a predictive model by splitting the dataset into multiple subsets or folds, training the model on a subset of the data, and then evaluating it on the remaining data.

The CV score typically represents the average performance metric obtained across all folds of the cross-validation process. It provides an estimation of how well the model performs on unseen data by evaluating its performance on multiple folds of the dataset. A higher CV score generally indicates better model performance, but it is essential to consider other metrics and evaluate the model's performance comprehensively.

CHAPTER 3
RESULTS

This chapter contains the results obtained by the comparison of various regression methods such as linear regression, decision tree and random forest for the prediction of calorie expenditure. In addition to that we have also given the graphical representation of data on the basis of various factors.

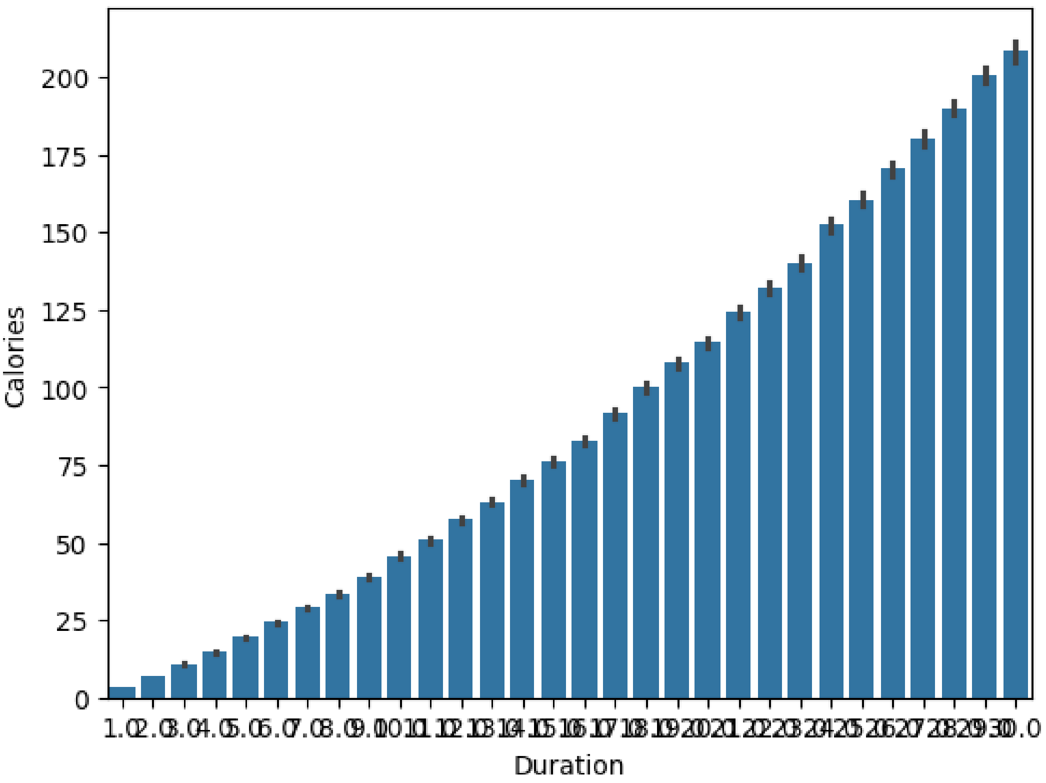
We collected the dataset from Kaggle.

By using libraries like pandas and numpy we clean and preprocess the data. For graphical visualization we use libraries like matplotlib, seaborn and demonstrate the visualizations using heatmaps, distplots, etc. Next, the data is split into training and testing data. Training data is used to train the different models and we test the models with the help of testing data and get the finest fulfil model out of all models.

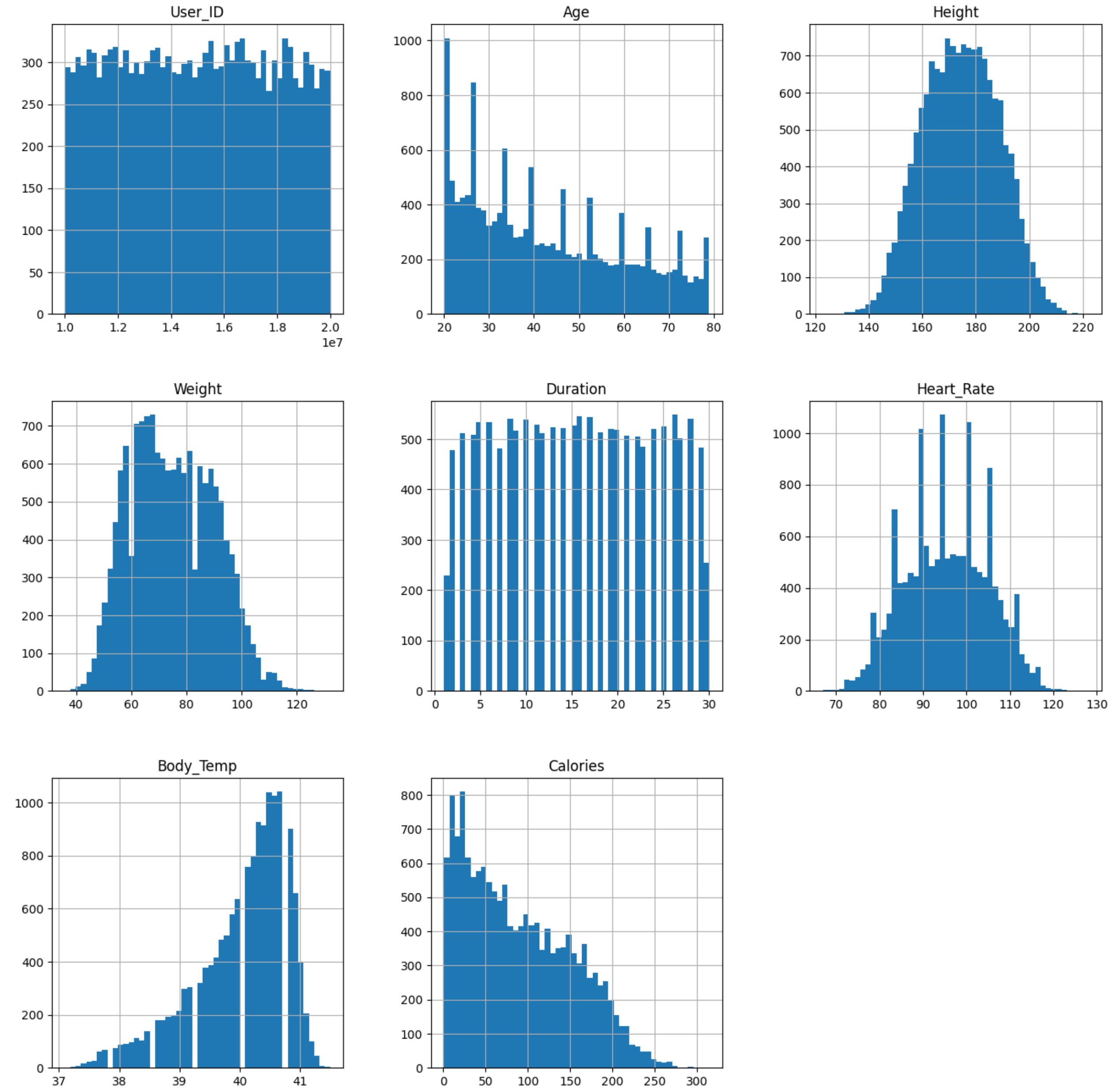
Figures plotted:

Barplot: Barplots are a type of graph that visually displays categorical data through rectangular bars, where the length or height of each bar corresponds to the frequency or value of the category it represents. Typically, the categories are displayed on the horizontal axis, while the frequency or values are represented on the vertical axis. Barplots are commonly used to compare the distribution or frequency of different categories, identify trends, and visualize relationships between categorical variables. They are commonly used in data analysis and visualization to illustrate relationships between categorical variables and to give insights about the distribution of data across different categories.

Barplot showing the comparison between ‘Calories’ and ‘Duration’ is given below.

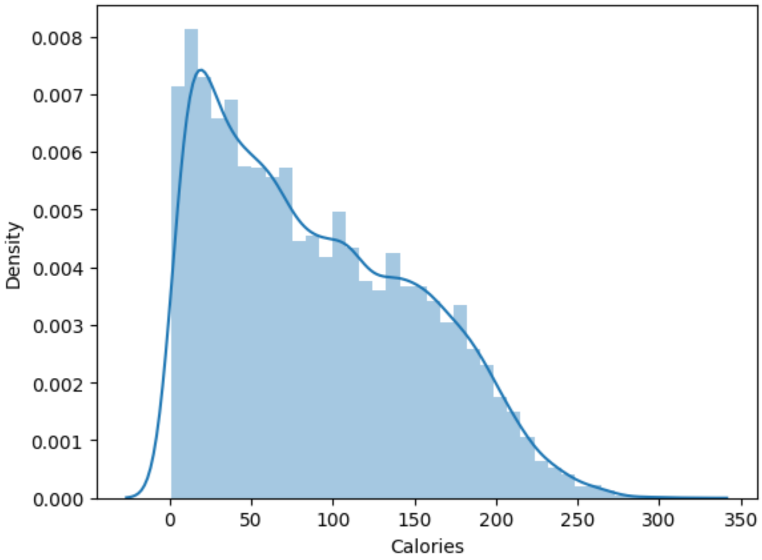


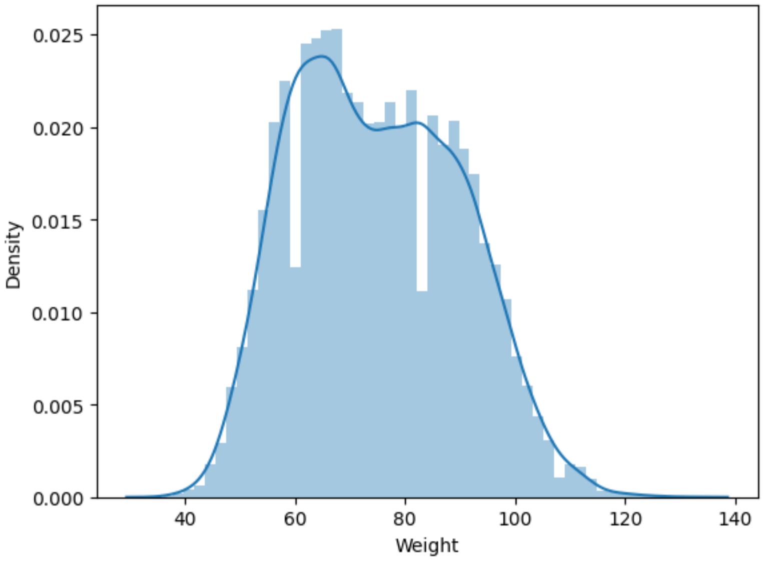
Histogram: A histogram is a graphical representation of the distribution of data. It consists of a series of adjacent rectangles, where the area of each rectangle corresponds to the frequency of the data within a specific range or "bin". The horizontal axis represents the range of values being measured, while the vertical axis represents the frequency of those values occurring. Histograms are commonly used to understand the shape, central tendency, and variability of a dataset, making them valuable tools for exploratory data analysis and data visualization. We can visualize our dataset using histograms:

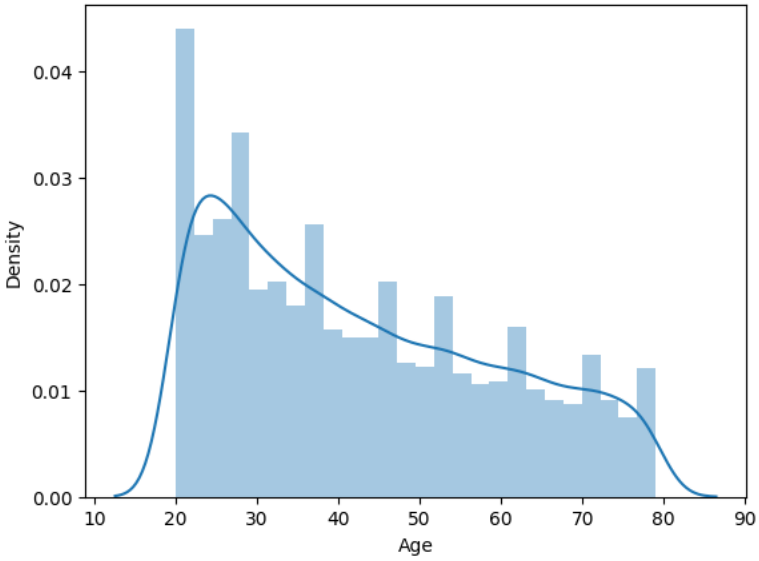


Density Plots: Density plots display the probability density function of the data, showing where values are concentrated and how they are spread out across different ranges. The density plot consists of a smoothed curve that estimates the probability density of the variable, often using techniques like kernel density estimation. The x-axis represents the values of the variable being analysed, while the y-axis represents the frequency or density of occurrence. Density plots provide insights into the shape, central tendency, and spread of the distribution. They are particularly useful for visualizing the distribution of data when histograms may be too rough or when comparing multiple distributions in the same plot.

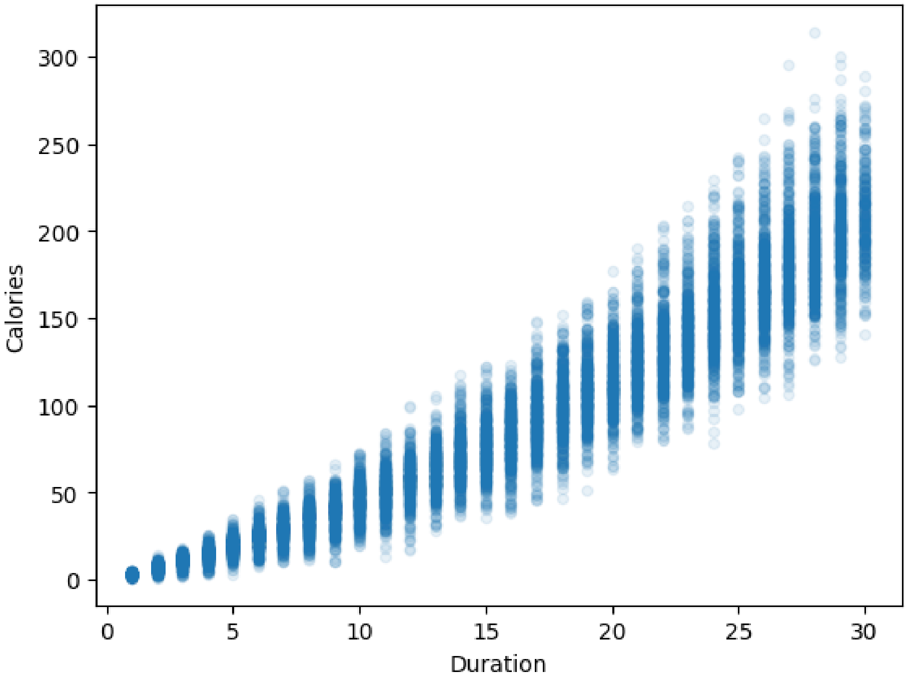
Density plot of some attributes:





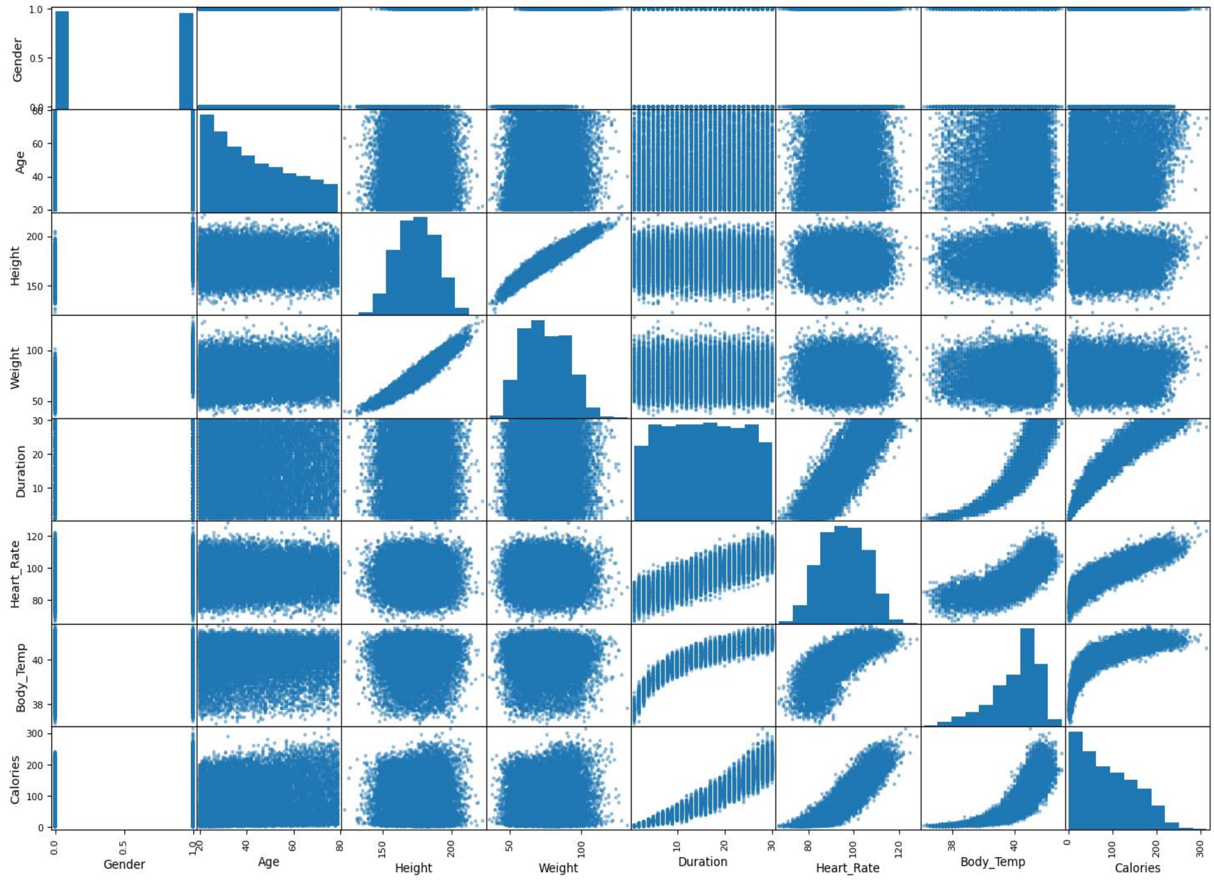


Scatter plot: Scatter plot displays the relationship between two variables by representing individual data points on a Cartesian coordinate system. Each data point is plotted as a dot with its position determined by the values of the two variables. The horizontal axis typically represents one variable (the independent variable), while the vertical axis represents the other variable (the dependent variable). Scatter plots are used to identify patterns, trends, and correlations between variables. They are particularly useful for exploring and visualizing bivariate data, aiding in understanding the nature of relationships.

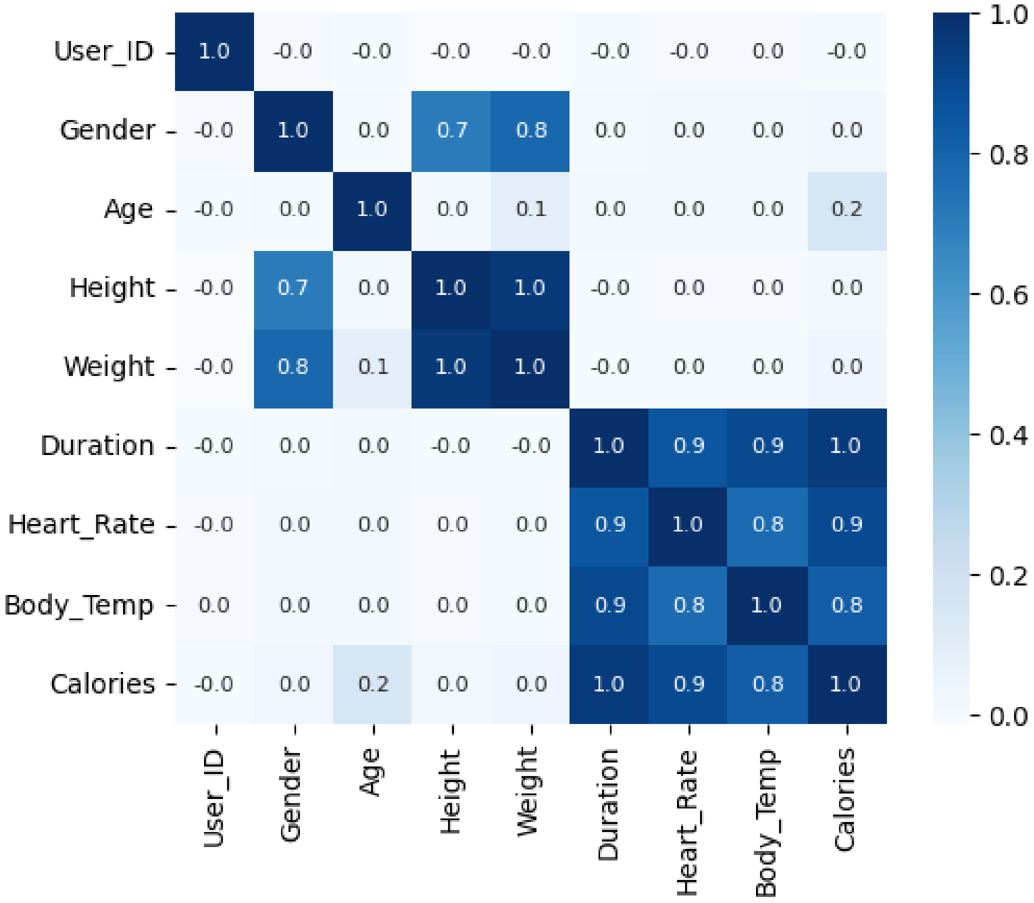


Scatter matrix is a grid of scatterplots where each variable in a dataset is plotted against every other variable. This graphical representation is particularly useful for exploring the relationships between multiple variables simultaneously.

The scatter matrix for various attributes in our data is given below.



Heatmap: A heatmap is a graphical representation of data where values in a matrix are represented as colors. Typically, the matrix is displayed as a grid, with rows and columns corresponding to different categories or variables. Each cell in the grid is colored according to the value it represents, with a color scale indicating the magnitude of the values. Heatmaps are particularly useful for visualizing large datasets and identifying patterns or trends within the data. They can reveal areas of high and low concentration, allowing for quick interpretation of complex relationships. Heatmaps are commonly used in fields such as data analysis, biology and finance visualize various types of data, including correlation matrices, gene expression levels, stock market returns, and geographical distributions.



The heatmap given shows the correlation among features. It indicates the interrelation among features of used data. The map shows that the individual’s caloric expenditure has a high positive correlation with the features ‘Heart_Rate’, ‘Duration’ and ‘Body_Temp’. It can also be seen that features like age, height, weight or even gender has little to no impact on the amount of calories burned.

3.1 MODEL SELECTION AND EVALUATION

In this project, various machine learning algorithms like linear regression, decision tree regression, random forest regression were used to predict caloric expenditure. The dataset has a total of 9 attributes, out of those only 8 attributes were considered for the prediction of caloric expenditure.

LINEAR REGRESSION

```
Mean Squared Error: 138.12408611460899
Mean Absolute Error: 8.479071745987955
Root Mean Squared Error: 11.752620393538157
R-squared value: 0.9655977245826504
CV scores: [0.96712832 0.96658977 0.96769213 0.96828562 0.96606908]
Mean CV score: 0.967152984018283
```

DECISION TREE

```
Mean Squared Error: 30.042
Mean Absolute Error: 3.4833333333333334
Root Mean Squared Error: 5.481058291972454
R-squared value: 0.9925175022897132
CV scores: [0.99324814 0.99244949 0.99304256 0.99263242 0.99295852]
Mean CV score: 0.9928662244988278
```

RANDOM FOREST

```
Mean Squared Error: 9.352970766666665
Mean Absolute Error: 1.809223333333333
Root Mean Squared Error: 3.0582627039982464
R-squared value: 0.9976704752564423
CV scores: [0.99804131 0.99798124 0.99798466 0.9976876  0.99800476]
Mean CV score: 0.9979399145854062
```

SUPPORT VECTOR REGRESSION

```
Mean Squared Error: 153.72801041174648

Mean Absolute Error: 8.372907296378877

Root Mean Squared Error: 12.398710030150172

R-squared value: 0.9617112879997057

CV scores: [0.96472358 0.96452881 0.96588848 0.96604203 0.96345961]

Mean CV score: 0.9649285030548607
```

RIDGE REGRESSION

```
Mean Squared Error: 138.12415168056407

Mean Absolute Error: 8.479061921349464

Root Mean Squared Error: 11.752623182956393

R-squared value: 0.9655977082522759

CV scores: [0.96712809 0.96658958 0.96769232 0.96828554 0.96606943]

Mean CV score: 0.9671529901042017
```

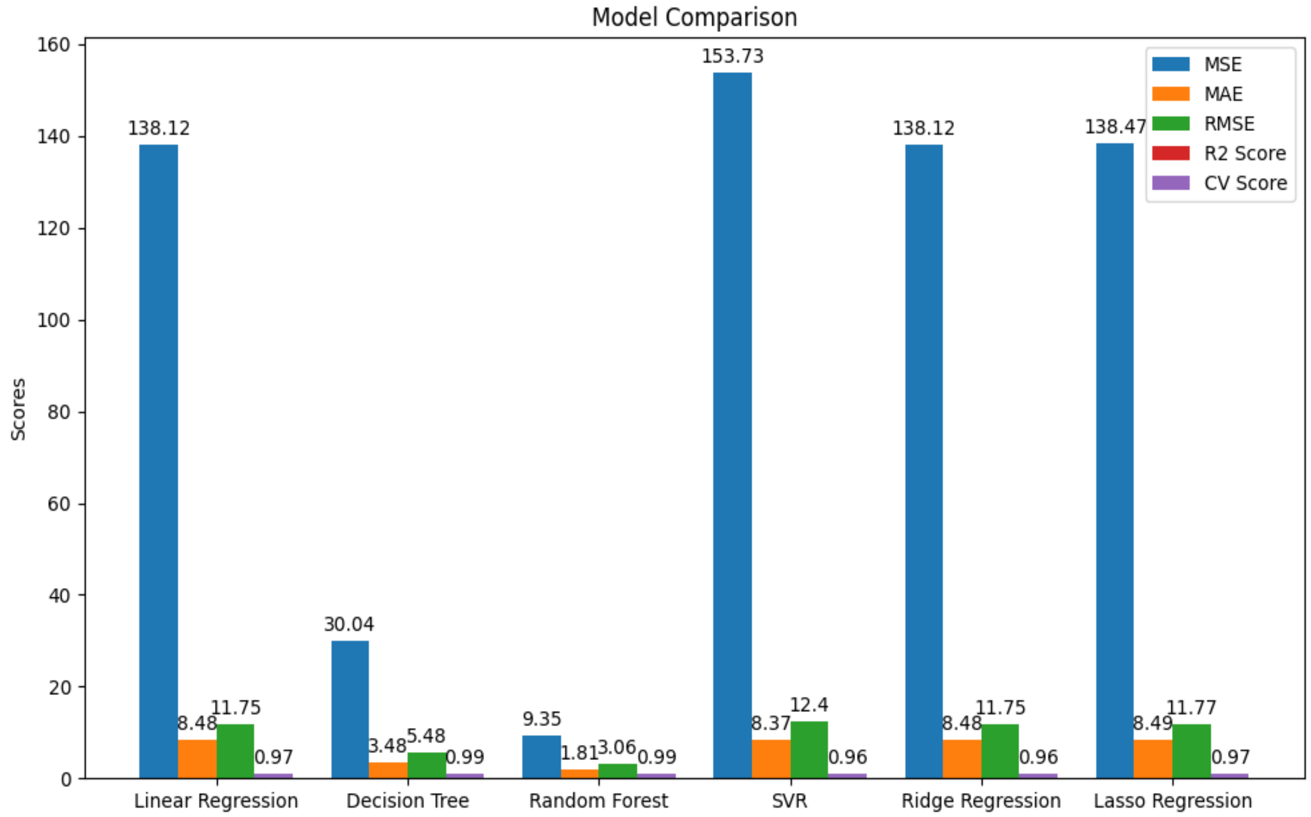
LASSO REGRESSION

```
Mean Squared Error: 138.472048981071

Mean Absolute Error: 8.485678477262354

Root Mean Squared Error: 11.767414711017496

R-squared value: 0.9655110582038617

CV scores: [0.96701829 0.96646919 0.96764276 0.96825862 0.9660856 ]

Mean CV score: 0.9670948936292485
```

Comparison of Different Regressor Algorithms:

Models	MSE	MAE	RMSE	R^2 Value	CV Score
Linear Regression	138.1241	8.4791	11.7526	0.9656	0.9672
Decision Tree	30.042	3.4833	5.4811	0.9925	0.99289
Random Forest	9.353	1.8092	3.0583	0.9977	0.9979
SVR	153.7280	8.3729	12.3987	0.9617	0.9649
Ridge Regression	138.1242	8.4791	11.7526	0.9656	0.9672
Lasso Regression	138.4721	8.4857	11.7674	0.9655	0.9671

Graphical Representation of Comparison of Different Regressor Models:



The table shows the various performances of different regressor models with average predicting accuracy 99%. Each model performed relatively well in terms of prediction accuracy. However, the Decision Tree and Random Forest models seem to perform exceptionally well with low errors and high R-squared values. The Linear Regression, SVR, Ridge Regression and Lasso Regression models show good results but they have high MSE indicating high errors.

CHAPTER 4
CONCLUSION

With the world becoming more data-oriented and as social media platforms and the internet continue to grow as the repository of these data, researchers should make use of these data to drive decision-making policies that can positively impact society.

This project aimed to recognize the number of calories our body burns, which depends on several factors such as age, gender, weight, height, body temperature, duration, and heart rate. We applied machine learning methods for caloric expenditure prediction based on the above factors. We applied data preprocessing techniques to improve the overall quality of the data. Later, we applied regression models to predict caloric expenditure.

Calories burnt can be predicted through different regression algorithms. Out of these regression algorithms, Random Forest regression gives the best accurate result. The MAE value of the Random Forest regressor is 1.8092 which is a good value. It means the errors are quite low. Therefore, Random Forest algorithm is the optimal algorithm for the calories burnt prediction so far. The Decision Tree model also gave an equally good result although it had a higher MAE value compared to Random Forest. We were also able to identify the factors affecting caloric expenditure. Variables such as heart rate, duration of exercise, and body temperature were found to be critical in predicting caloric expenditure. They were found to have a high positive correlation with the outcome. Whereas, factors such as gender, age, height and weight have little to no impact on the amount of calories burnt.

The findings of this project have significant practical implications. Accurate prediction of caloric expenditure can help individuals customize their exercise routines to achieve specific health and fitness goals, such as weight loss, cardiovascular health, or muscle building. Additionally, health professionals and fitness trainers can utilize these predictive models to develop more effective and personalized exercise programs for their clients, thereby improving overall health outcomes.

In summary, the application of machine learning to predict caloric expenditure based on exercise data holds great promise. Using various regression algorithms, we have identified the most effective approach for this task, thereby getting valuable insights into the complex relationship between exercise and caloric expenditure.

REFERENCES

[1] K. S. University, “Burning more calories is easier when working out with someone you perceive as better,” 26 November 2012. [Online]. Available: https://www.sciencedaily. com/releases/2012/11/121126130938.htm.

[2] Kalpesh, Jadhav, "Human Physical Activities Based Calorie Burn Calculator Using LSTM" Intelligent Cyber Physical Systems and Internet of Things: ICoICI 2022. Cham: SpringerInternational Publishing, 405-424, 2023.

[3] S. T and V. K, “PREDICTION OF USER’S CALORIE ROUTINE USING CONVOLUTIONAL NEURAL NETWORK,” International Journal of Engineering Applied Sciences and Technology, vol. 5, 189-195, 2020.

[4] Nipas, Marte,"Burned Calories Prediction using Supervised Machine Learning: Regression Algorithm." 2022 Second International Conference on Power, Control and Computing Technologies (ICPC2T). IEEE, 2022.

[5] G. vijayalakshmi and T. Sridurga, “COMPARING MACHINE LEARNING ALGORITHMS FOR PREDICTING CALORIES BURNED,” JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR), vol. 10, no. 3, 519-527, March 2023.

[6] S. P. Vinoy and B. Joseph, “Calorie Burn Prediction Analysis Using XGBoost Regressor and Linear Regression Algorithms,” in Proceedings of the National Conference on Emerging Computer Applications (NCECA), Kottayam, 2022.

[7] Tayade, Akshit Rajesh, and Hadi Safari Katesari, "A Statistical Analysis to Develop Machine Learning Models: Prediction of User Diet Type.”

[8] K. Westerterp, “Control of energy expenditure in humans,” European Journal of Clinical Nutrition, vol. 71, 340-344, 30 November 2016.

[9] Gour, Sanjay, "A Machine Learning Approach for Heart Attack Prediction." IntelligentSustainable Systems: Selected Papers of WorldS4 2021, vol. 1, 2022.

[10] Khan, Abdul Wahid, et al. "Factors Affecting Fitness Motivation: An Exploratory Mixed Method Study."IUP Journal of Marketing Management21.2 (2022).
https://www.medicalnewstoday.com/articles/319731a