**Course Work Project Description and Rubric**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Semester** | **202410** | | **Division** | | CIS |
| **Assessment title in Syllabus** | **Project** | | **Program** | | **IT and IS** |
| **1** |  | |  | |  |
| **Course Code** | **CIS 2423** | | | | |
| **Course Title** | **Programming for Data Analytics** | | | | |
| **CLOs** | **All CLOs** | | **Accreditation Body** | | **CAA & CIPS** |
| **Course Instructor** |  | | **CRN** | |  |
| **Assessment Weight** | **40%** | | **Submission Date** | | **Week 14** |
| **For Group Work submissions an additional individual assessment will be conducted.**  **Grades for the students in one group will vary based on the individual performance in the additional assessment.** | | | | | |
|  | | | | | |
| **Student Declaration**:  **Academic Integrity Statement**  In accordance with the HCT Academic Integrity Policy  • Students are required to refrain from all forms of academic integrity breaches as defined and explained by HCT.  • A student found guilty of having committed acts of academic integrity breach(es) will be subject to the relevant sanctions as outlined by HCT.  إفادة النزاهة الأكاديمية  **وفقًا لسياسة كليات التقنية العليا للنزاهة الأكاديمية**  **• على الطلبة الإلتزام بلوائح وقواعد النزاهة الأكاديمية، كما هو مبيّن وموضح في السياسات والإجراءات الخاصة بكليات التقنية العليا.**  **• في حالة ارتكاب الطالب أي شكل من أشكال الإخلال بالنزاهة الأكاديمية، سيتعرض الى العقوبات الموضحة في السياسات ذات الصلة.**  This assignment is entirely my own work except where I have duly acknowledged other sources in the text and listed those sources at the end of the assignment.  I have not previously submitted this work to the HCT, or any other entity. I understand that I may be orally examined on my submission.  **Student (s) Signature: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**  **\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_** | | | | | |
|  | | | | | |
| **Student Name(s):** | **Maryam Jameel Almahri** |  | |  | |
| **Student HCT ID(s):** | H00468784 | H00 | | H00 | |

**For Examiner’s Use Only**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Group (50%)** | | | | | **Individual (50%)** |  |  |
| **C**LO | **1** | **2** | **3** | **4** | **Report Formatting** | **Oral Defense** | **Total** | **%** |
| **Marks Allocated** | 10 | 10 | 42 | 26 | 12 | **50** | **100** | **4**0 |
| **Marks Obtained** |  |  |  |  |  |  |  |  |

* 1. **Purpose of Data Analysis for Gym Members Exercise Tracking Dataset**

The goal of analyzing the "gym\_members\_exercise\_tracking" dataset is to gain insights into how different factors affect the exercise habits and overall health of gym members. Here’s a breakdown of what we aim to discover:

1. **Age and Gender**: We want to see if age and gender influence how people work out and their fitness results.
2. **Body Measurements**: By looking at members' weight, height, and Body Mass Index (BMI), we can understand how these measurements relate to their fitness levels and exercise performance.
3. **Heart Rate**: We will analyze the relationship between maximum heart rate, average heart rate, and resting heart rate with how intense and long their workouts are.
4. **Workout Details**: We’ll explore the different types of workouts members do, how long they exercise, and how many calories they burn during their sessions.
5. **Fitness Levels**: We will compare members' experience levels with how often they work out and their performance metrics to see if experience affects results.
6. **Nutrition and Hydration**: Finally, we want to look at how water intake and fat percentage relate to members' exercise outcomes.

**2.1 Type of Programming Used: Python**

**Why WE Use Python for Data Analysis:**

1. **Useful Libraries**:
   * **Pandas**: Makes it easy to work with data.
   * **NumPy**: Helps with numbers and calculations.
   * **Matplotlib** and **Seaborn**: Great for making charts and graphs.
2. **Easy to Learn**: Python is simple to read and understand, which is perfect for beginners.
3. **Lots of Help Available**: There are many online resources, tutorials, and communities where you can find support.
4. **Works with Different Data**: Python can handle various data types, like CSV and Excel files and html web.
5. **Can Do Statistics**: You can perform important calculations to understand the data better.
6. **Machine Learning**: Python has tools to help you make predictions based on the data.
7. **Shareable Code**: You can easily share your Python code so others can use it too or improve it .

**3.1 Type of Machine Learning Algorithm: Regression**

Predicting Values: We will use regression to predict numerical outcomes. For example, we can predict:

Calories Burned based on factors like age, weight, height, and workout duration.

BMI using just weight and height.

Finding Relationships: This algorithm helps us understand how different factors, such as age or workout type, influence results like calories burned. This insight is valuable for analyzing fitness data.

Making Informed Decisions: By applying regression, trainers and gym managers can improve their decisions about workout plans and nutrition recommendations for gym members.

Regression is a useful tool for predicting outcomes and understanding how various factors are related to the gym members' dataset. This will help us gain insights into fitness and health

**4.1 Independent and Dependent Variables**

**Independent Variables:**

These are the factors we think might influence the results we’re looking at. In our dataset, the independent variables include:

1. **Age**: We believe that age can impact how members exercise and their fitness levels.
2. **Gender**: Male and female members may have different workout habits and preferences.
3. **Weight (kg)**: A member’s weight could affect how many calories they burn during workouts.
4. **Height (m)**: Height might also play a role in fitness results.
5. **Session\_Duration (hours)**: Longer workouts usually mean more calories burned.
6. **Workout\_Type**: Different types of workouts can lead to different results in calorie burning.
7. **Fat\_Percentage**: This can affect overall health and fitness levels.
8. **Workout\_Frequency (days/week)**: Working out more often can improve fitness outcomes.
9. **Experience\_Level**: The amount of time someone has been exercising can impact their results.
10. **BMI**: Body Mass Index is another important measure we want to analyze because it relates to weight and height.
11. **Resting\_BPM**

**Dependent Variables:**

These are the outcomes we want to focus on and measure. In our dataset, the dependent variables include:

1. **Calories\_Burned**: This is the primary outcome we want to predict based on the independent variables.
2. **Max\_BPM**: The maximum heart rate can indicate workout intensity and effort.
3. **Avg\_BPM**: The average heart rate during a workout reflects how hard someone is working.
4. **Water\_Intake (liters)**: Staying hydrated might influence workout performance.

**Justification:**

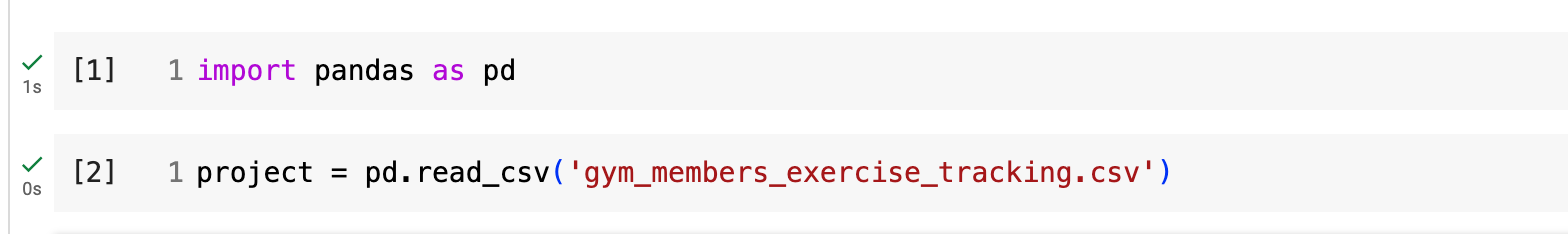
We chose these independent variables because we believe they can significantly influence how many calories a person burns and their BMI. The dependent variables (Calories Burned and BMI) are what we’re interested in understanding better through our analysis.

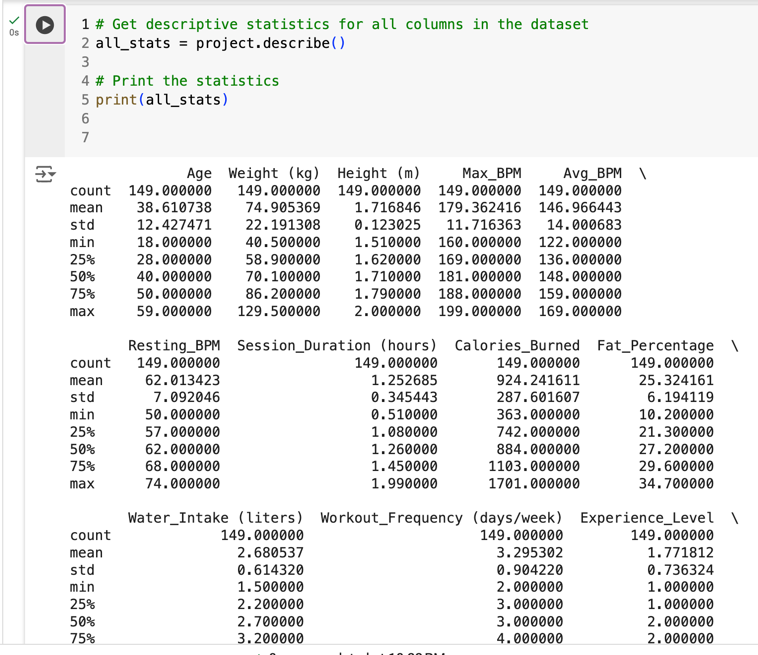
**2.5Why Perform Descriptive Analysis**

doing descriptive analysis helps us understand the data and find useful information to improve fitness programs.I want to do descriptive analysis on the gym members' dataset for a few reasons:

1. **Get to Know the Data**: It helps us understand the basic details, like the average age and weight of the members.
2. **Find Patterns**: We can see trends, like whether older members work out more or if certain exercises burn more calories.
3. **Spot Unusual Values**: This analysis helps us find any strange numbers, like someone with a very high or low BMI, which we might need to check further.
4. **Support Decision-Making**: The insights we gather can help gym managers and trainers improve workout programs and better meet member needs.
5. **Prepare for Further Analysis**: Understanding the basics helps us choose the right methods for deeper analysis later on.

**2.6 Descriptive Statistics Summary for Gym Members' Dataset**

****

****

**A white background with black text

Description automatically generated**

When you run project.describe(), it will provide the following information for each numeric column in your dataset:

* count: The number of non-missing entries in each column.
* mean: The average value of each column.
* std: The standard deviation, showing how much the values vary from the mean.
* min: The minimum value in each column.
* 25%: The 25th percentile (first quartile) of each column.
* 50%: The 50th percentile (median) of each column.
* 75%: The 75th percentile (third quartile) of each column.
* max: The maximum value in each column**.**

**2.7 Random Sampling and Descriptive Statistics Calculation for Dependent Variables**

**A screenshot of a computer

Description automatically generated**

This code takes a random sample of 150 entries from the dataset and analyzes key details—like the average and range—for four main variables: Calories\_Burned, Max\_BPM, Avg\_BPM, and Water\_Intake. It then prints these details, giving us a quick overview of the sample’s characteristics for each variable.

**2.8**

**2.9**

**2.10.A Scatter Plot of Calories Burned vs. Workout Frequen**

**A close-up of a white background

Description automatically generated**

**A graph of red dots

Description automatically generated**

This scatter plot shows the relationship between Workout Frequency (days per week) and Calories Burned. Each red dot represents an individual data point, helping to visualize any correlation between exercise frequency and calorie expenditure.

**2.10.B Box Plot of Age**

**A screen shot of a graph

Description automatically generated**

This box plot displays the distribution of Age in the dataset, showing the median, quartiles, and any outliers. It provides insights into the spread and central tendency of ages.

**2.10.C Histogram of Experience Level**

**A computer code with text

Description automatically generated**

**A graph of a person with a blue line

Description automatically generated with medium confidence**

This histogram shows the distribution of Experience Level in the dataset, with the frequency of each level represented by the bars. The plot includes a Kernel Density Estimate (KDE) curve to visualize the overall distribution.

**2.10.D Heat Map of Correlation Matrix**

**A screenshot of a computer code

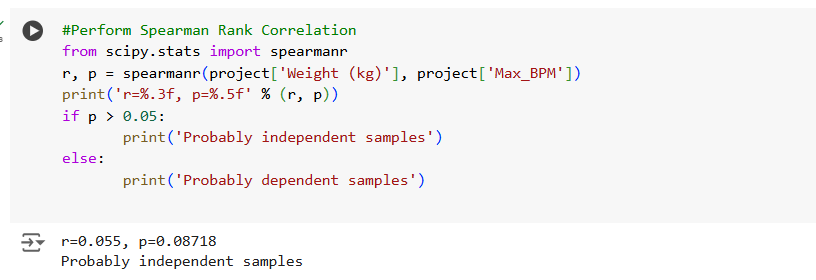
Description automatically generated**

**A screenshot of a graph

Description automatically generated**

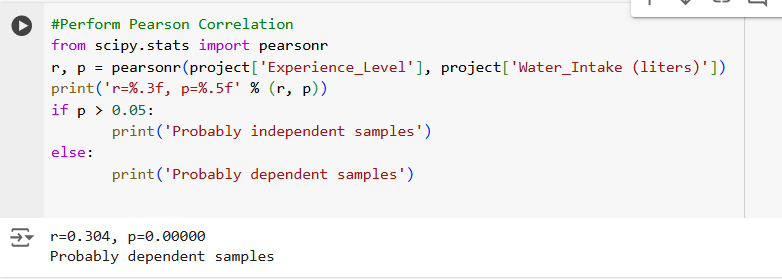
This heat map visualizes the correlation between numeric variables in the dataset. The color intensity indicates the strength of the correlation, with values annotated for clarity. It helps identify potential relationships between variables

**2.11 A Spearman Rank Correlation between Weight and max BPM**

****

This Spearman rank correlation test evaluates the monotonic relationship between Weight (kg) and MAX BPM in the dataset. The test calculates the correlation coefficient (r) and the p-value. If the p-value is greater than 0.05, so its probably independent samples.

**2.11 B person correlation between experience level and water intake**



This pearson correlation test evaluates the monotonic relationship between experience level and water intake in the dataset. The Pearson correlation test calculates the correlation coefficient (r) and the p-valu**e** Based on the results (r = 0.304, p = 0.00000), we can conclude that :the p-value is smaller than 0.05 so they are probably dependent.

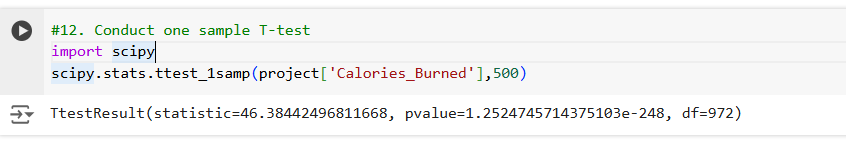
**2.11 C Chi-Square Test for Gender and Workout Type Association**

A screenshot of a computer program

Description automatically generated

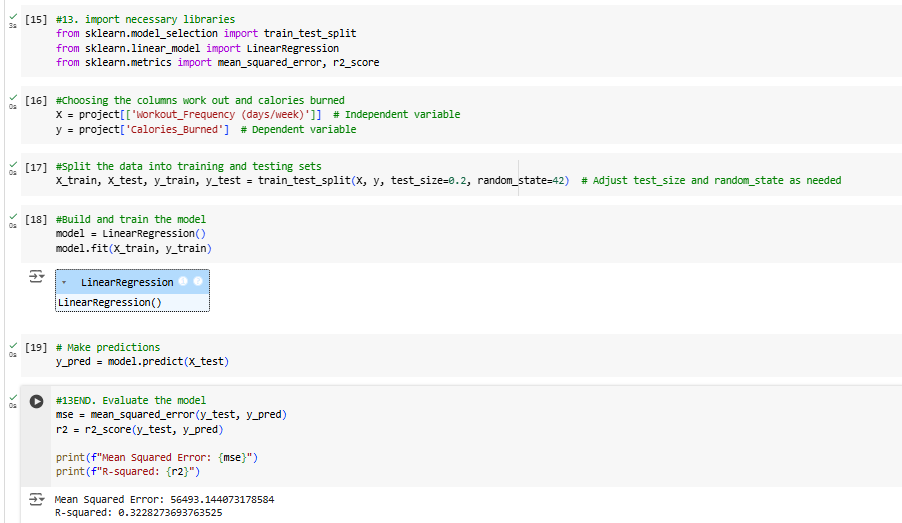
This chi-square test checks if there is an association between Gender and Workout Type. A p-value less than 0.05 indicates a dependency, while a p-value greater than 0.05 suggests independence.

**2.12 One-Sample T-Test for Calories Burned**

****

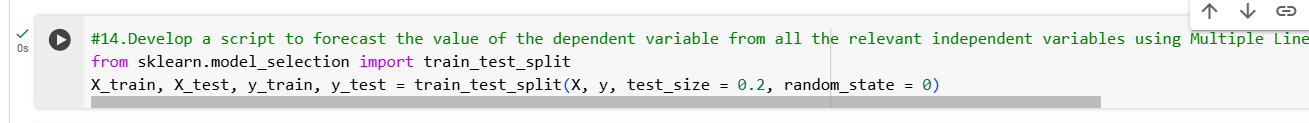
We Select the dependent variable to analyzes. Let's say it's calories burned . and Let's assume the population mean for calories burned is 500. So n this case, the p-value (1.2524745714375103e-248) is extremely small and much less than the typical significance level of 0.05. This means there is very strong evidence against the null hypothesis. Therefore, we **reject the null hypothesis**.

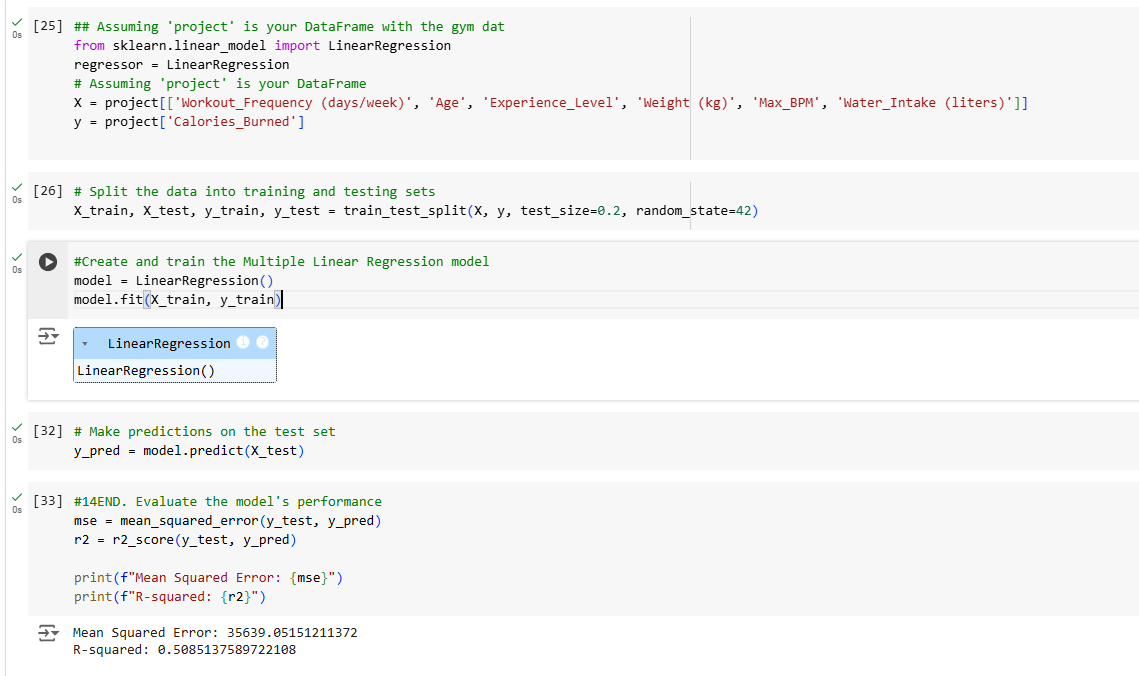
**2.13build, train, develop and evaluate using simple regression for chosen dataset**



Here we import libraries for data manipulation, model selection, model building, and model evaluation, then we identify the dependent and independent variables and create X and y accordingly, also we split the data into training and testing sets to evaluate the model's performance on unseen data. we create a Linear Regression object and train it using the training data ,then  We use the trained model to make predictions on the testing data. And finally, We evaluate the model's performance using metrics like Mean Squared Error and R-squared.

**2.14develop a script forecast the value of the dependent variable from all the relevant independent variables using multiple linear regression**

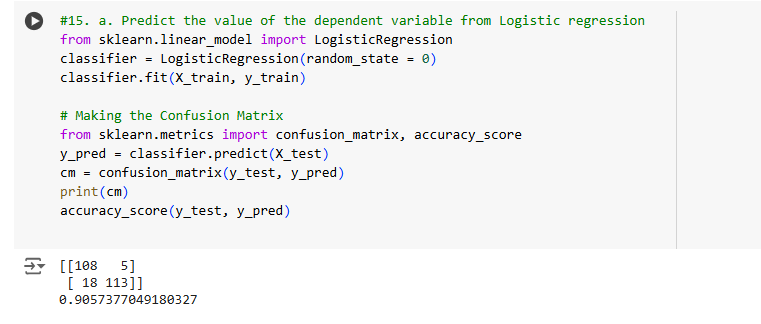




For the multiple linear regression , fist we import the dataset Prepare the data ,then Splits the data into training and testing sets using train split test to evaluate the model's performance on unseen data, also we Build and train the mode by Creates a Linear Regression model using Linear Regression() and Trains the model on the training data using model.fit(). Then Make prediction . lastly Evaluate the model by Calculates the Mean Squared Error (MSE) and R-squared (R2) to measure the model's accuracy. Prints the MSE and R2 values.

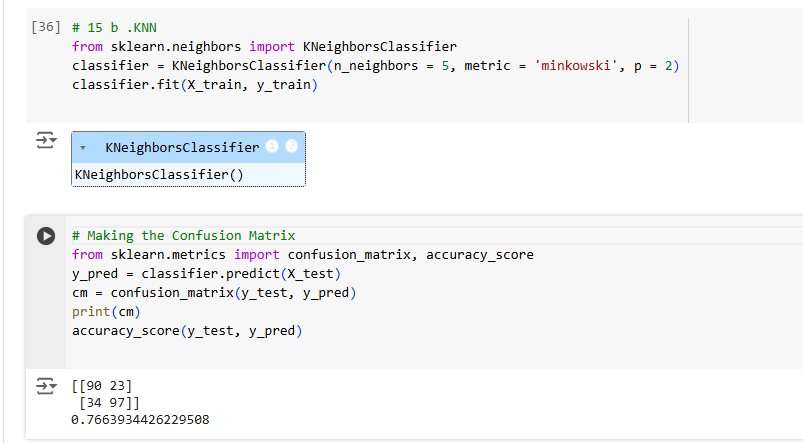
1. **15.predict the value of the dependent variable fron the different classifier such as logistic regression, KKN, naïve-bayes and decision tree**

**A** . **logistic regression**



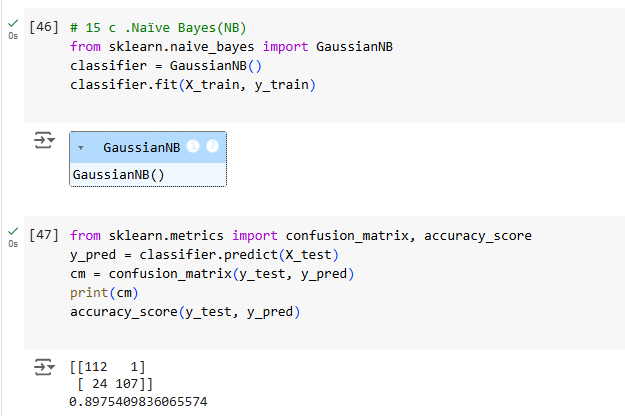
Therefore, for the given dataset, 23 errors (5+18) are found and accuracy of the model is 90.57%

* **b. KKN K-Nearest Neighbours-**



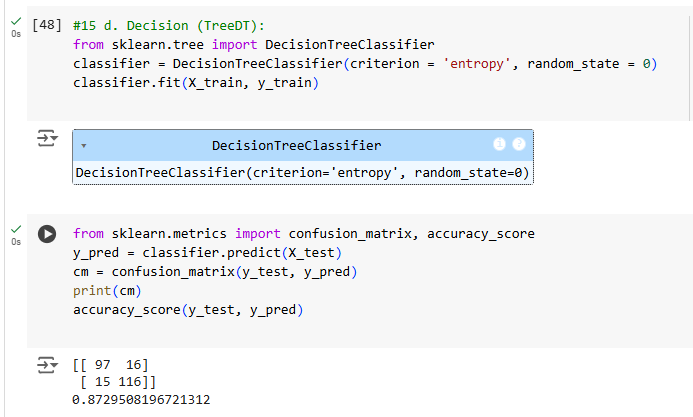
Therefore, for the given dataset, 57errors (23+34) are found and accuracy of the model is 76.63%

**C.** **Naïve Bayes**



Therefore, for the given dataset, 25 errors (1+24) are found and accuracy of the model is 89.75

**D.** **Decision Tree**



Therefore, for the given dataset, 31 errors (16+15) are found and accuracy of the model is 87.29%

**16 .evaluate the performace of each model using confusion matrix and accuracy and identify the best fit classifier for the chosen dataset** .

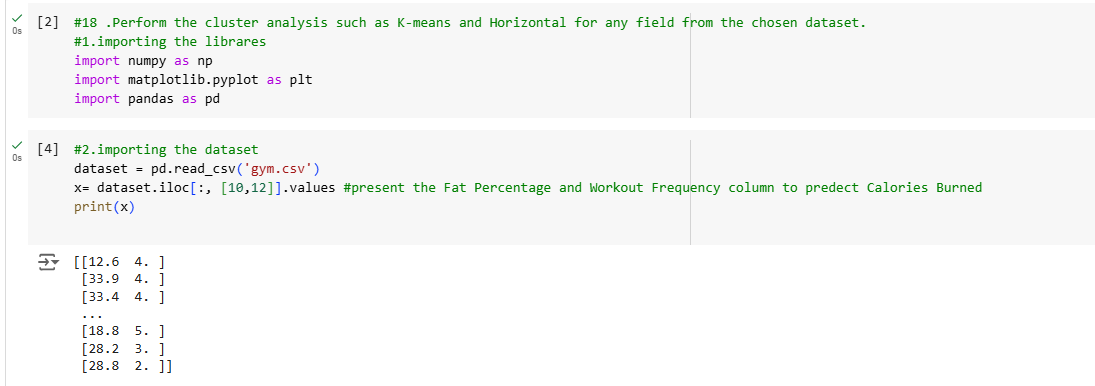
Below is the result of performance metrics(confusion matrix) for classification models of the given data set:

|  |  |  |
| --- | --- | --- |
| **Model** | **Errors** | **Accuracy** |
| LR | 23 | 90.57% |
| KNN | 57 | 76.63% |
| NB | 25 | 89.75 |
| DT | 31 | 87.29% |

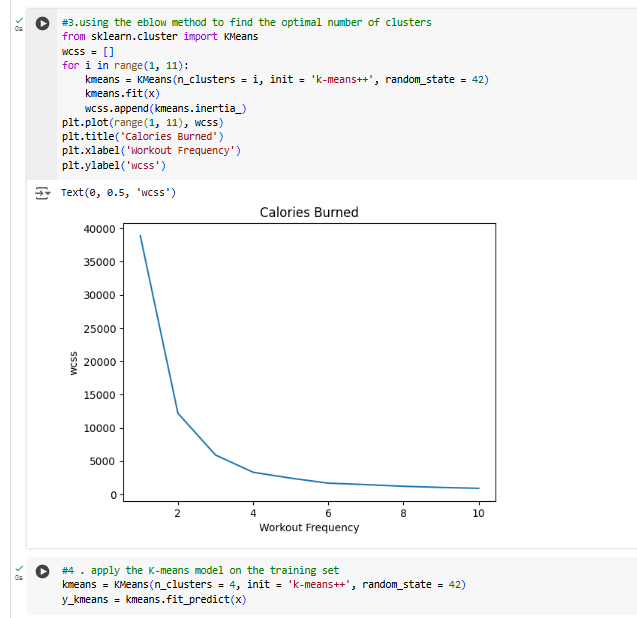
**17. predict the dependent variable by using best-fit classifier**.

Thus, By comparing the accuracy and confusion matrices for all of the models we can identify the best fit classifier for your dataset is logistic regression because its have a minimum number of errors .

**18.preform the cluster analysis such as K-means and horizontal for any field from the chosen dataset .**

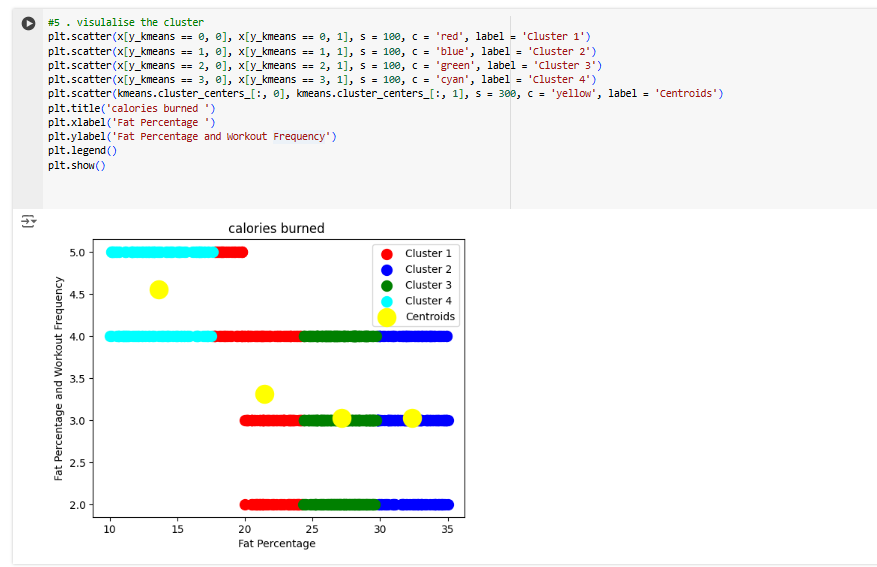


First we We must visualize the cluster at the conclusion of the analysis, which necessitates two columns (X-axis, Y-axis). Therefore, we select the fat percentage and workout frequency



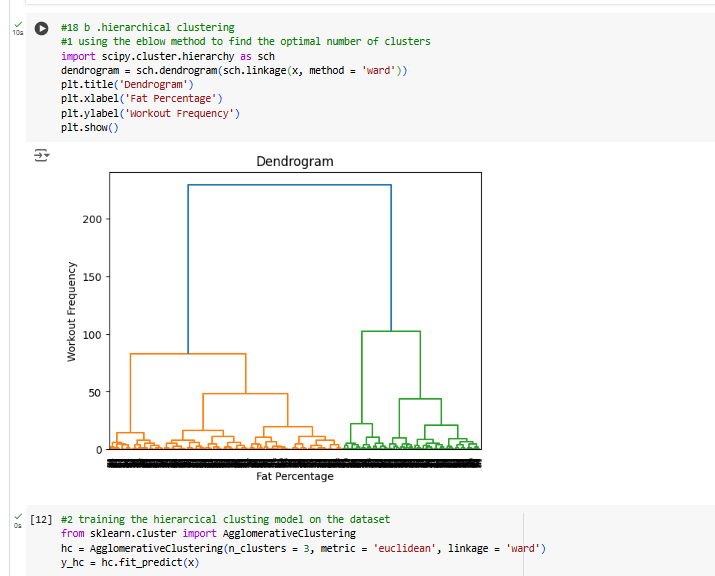
We Find the number of cluster by analyzing the graph where wcss value reduces drastically.

Here, when number of cluster=4, wcss reduces drastically

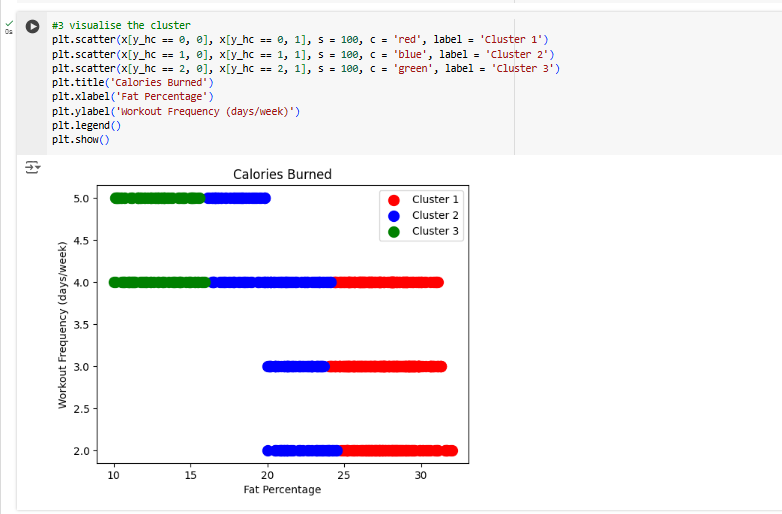


Finaly this is the result of k-mean

**18 b .horizontal clustering**

****

Using the dendrogram to find the optimal number of clusters and Training the Hierarchical Clustering model on the dataset

****

**19.explain the strategy for improving the system after viewing the cluster diagram**

The strategy is to visually represent the clusters formed by hierarchical clustering. It achieves this by plotting the data points on a scatter plot, using different colors for different clusters, and providing labels for clarity. This allows you to see how data points are grouped together based on the hierarchical clustering algorithm and to understand the relationships between clusters in the data

**20. create a new repo for project in git hub**

