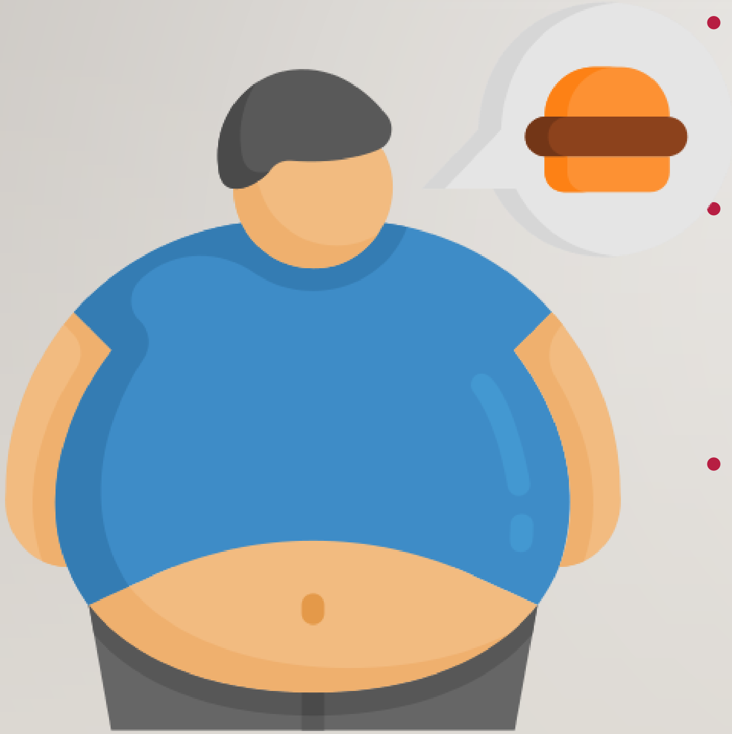
PREDICTING OBESITY LEVELS: DATA ANALYTICS PROJECT PRESENTATION

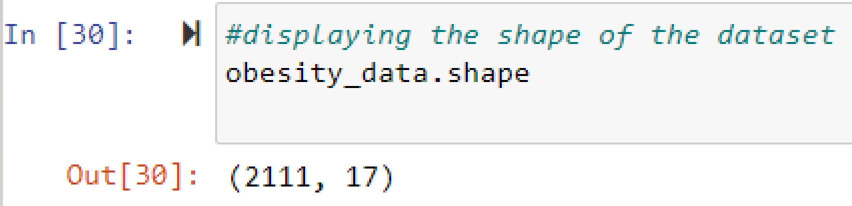
BY GETRUDE SHABIHA

INTRODUCTION



- Business Problem
  - Predict obesity levels based on eating habits and physical condition.
- Project Objectives:
  - Develop a reliable predictive model for classifying obesity levels accurately.
- Importance of Predicting Heart Disease Risk:
  - Obesity is a significant public health concern.
  - Predictive model can aid in prevention and management strategies.

DATASET OVERVIEW



```
In [30]:   #displaying the shape of the dataset
           obesity_data.shape

Out[30]:   (2111, 17)
```

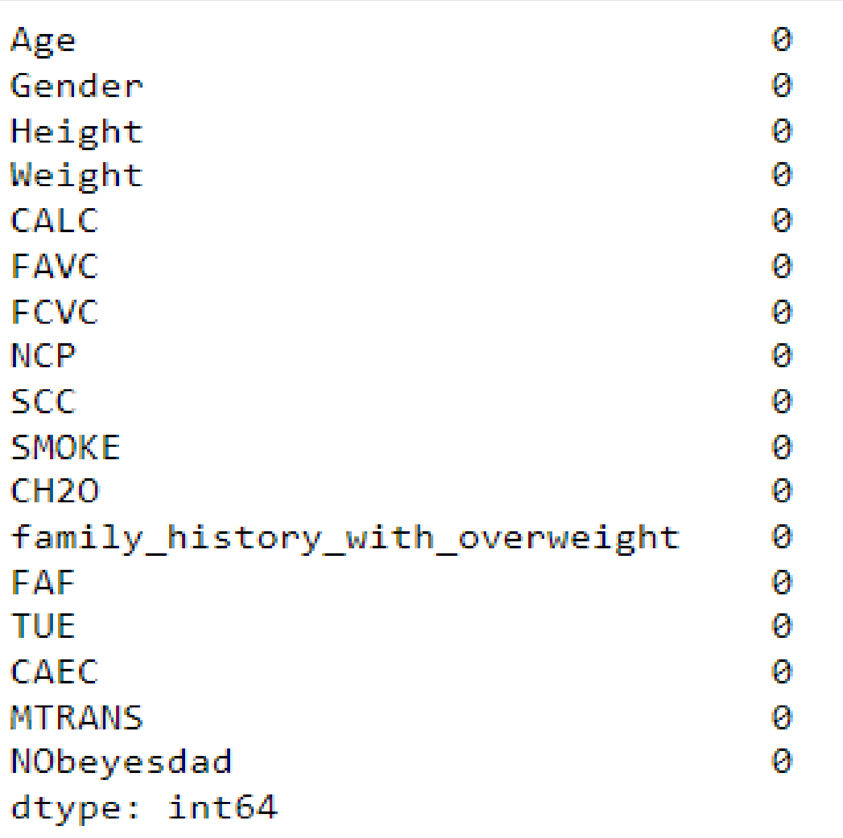
Brief Description of Dataset:

- The obesity dataset contains dataset from Mexico, Peru, and Colombia.
- Combination of synthetic and user-collected data.
- The dataset has a total of 2111 rows and 17 columns of data

Significance of Dataset:

- Foundation for predicting the factors that contribute to obesity.
- The dataset has multiple data points of people who experience obesity problems.

DATASET OVERVIEW



```
Age                               0
Gender                            0
Height                            0
Weight                            0
CALC                              0
FAVC                              0
FCVC                              0
NCP                               0
SCC                               0
SMOKE                             0
CH2O                              0
family_history_with_overweight    0
FAF                               0
TUE                               0
CAEC                              0
MTRANS                            0
NObeyesdad                        0
dtype: int64
```

- The initial data cleaning phase revealed that no data rows had any null values.

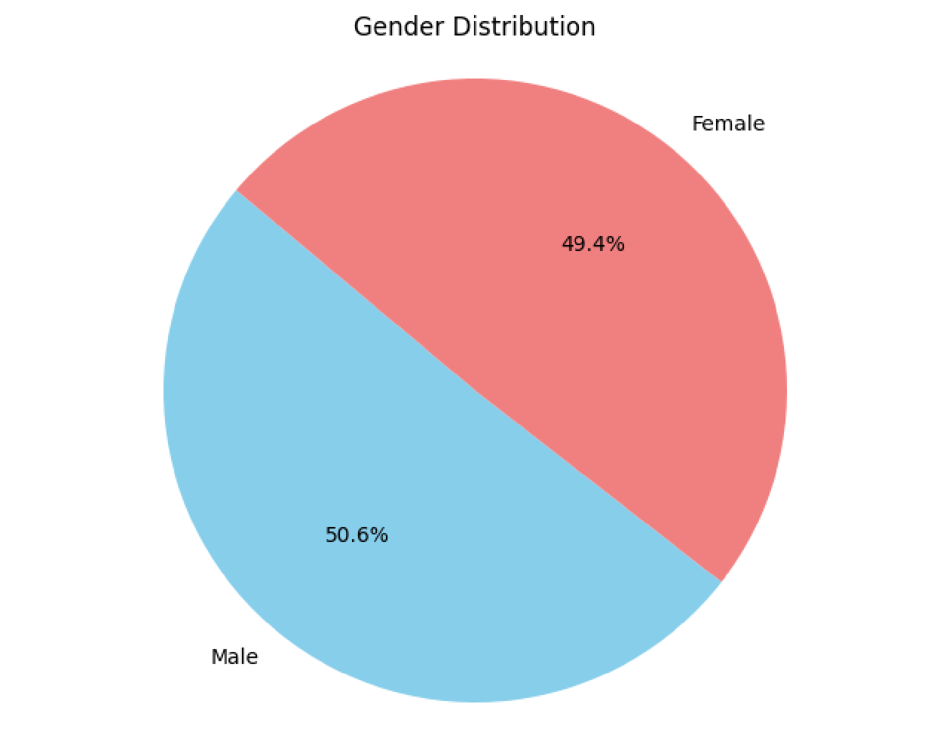
DATASET VARIABLE DESCRIPTION

Variable Name	Description	Data type
Gender	This is the gender of the participant	Categorical
Age	This is the age of the participant	Continuous
Height	This is the height of the participant	Continuous
Weight	This is the weight of the participant	Continuous
family_history_with_overweight	this is value shows any presence of obesity in immediate family members	Feature, Binary
FAVC	Do you eat high caloric food frequently?	Feature, Binary
FCVC	Do you usually eat vegetables in your meals?	Feature, Integer
NCP	How many main meals do you have daily?	Feature, Continuous
CAEC	Do you eat any food between meals?	Feature, Categorical

DATASET VARIABLE DESCRIPTION

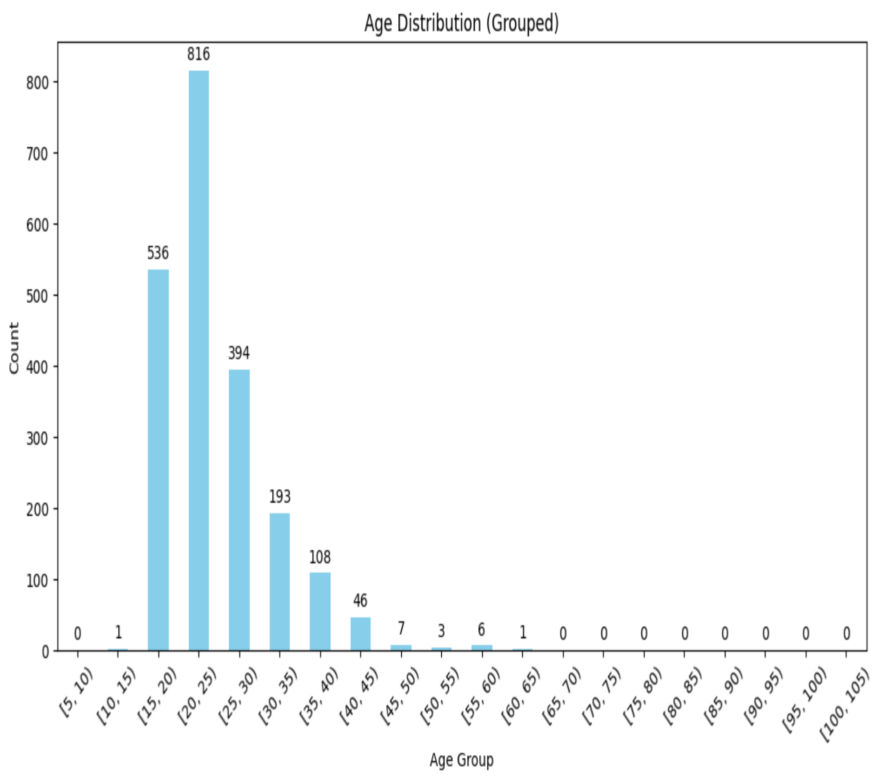
Variable Name	Description	Data Type
SMOKE	Do you smoke?	Feature, Smoke
CH2O	How much water do you drink daily?	Feature, Continuous
SCC	Do you monitor the calories you eat daily?	Feature, Binary
FAF	How often do you have physical activity?	Feature, Continuous
TUE	How much time do you use technological devices such as cell phone, videogames, television, computer and others?	Feature, Integer
CALC	How often do you drink alcohol?	Feature, Categorical
MTRANS	Which transportation do you usually use?	Feature, Categorical
NObeyesdad	Obesity level	Target, Categorical

EXPLORATORY DATA ANALYSIS (EDA)



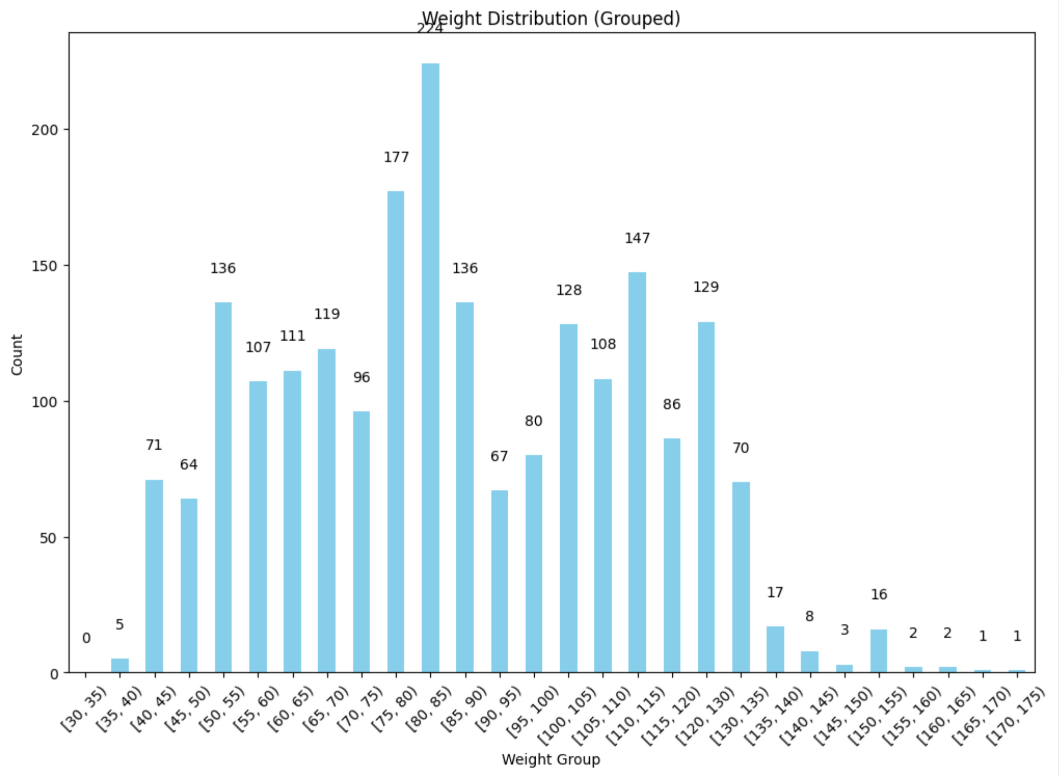
- This exploration shows the gender distribution in the dataset
- The dataset shows a close to equal distribution with 49.4% female participants vs 50.6% male participants

EXPLORATORY DATA ANALYSIS (EDA)



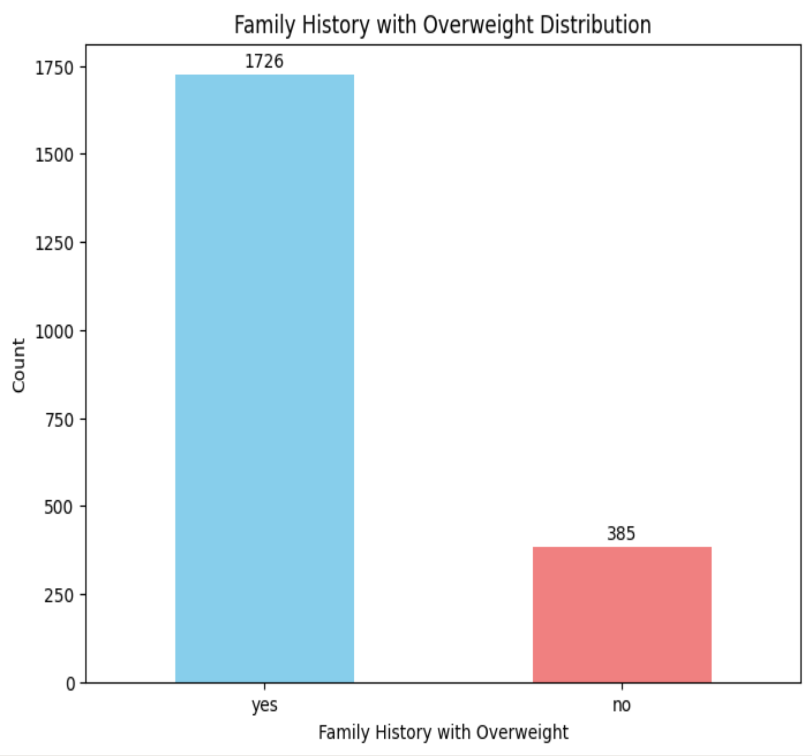
- This exploration shows the age distribution of the participants
- Majority of the participants are in the age bracket of 20 to 25
- This is closely followed by 15 to 20 and 25 to 30 consecutively

EXPLORATORY DATA ANALYSIS (EDA)



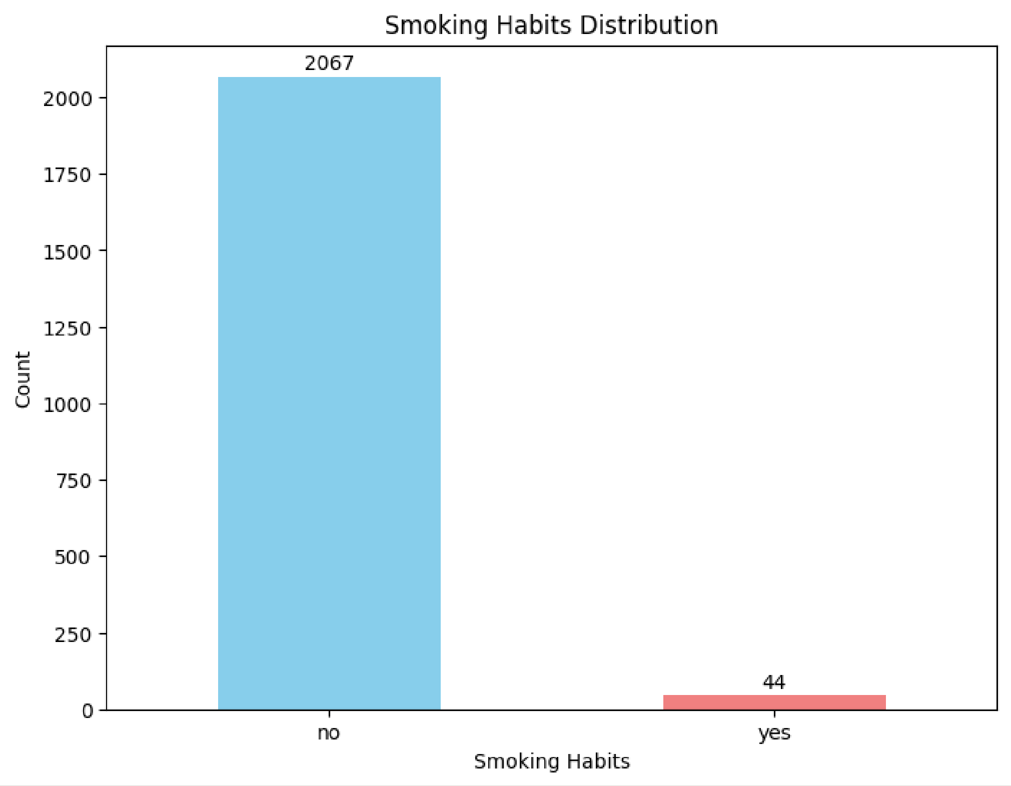
- This exploration shows the weight distribution of the participants in the dataset
- Majority of the members had a weight range of 85 to 90
- Majority of the participants have a weight range between 50 to 200

EXPLORATORY DATA ANALYSIS (EDA)



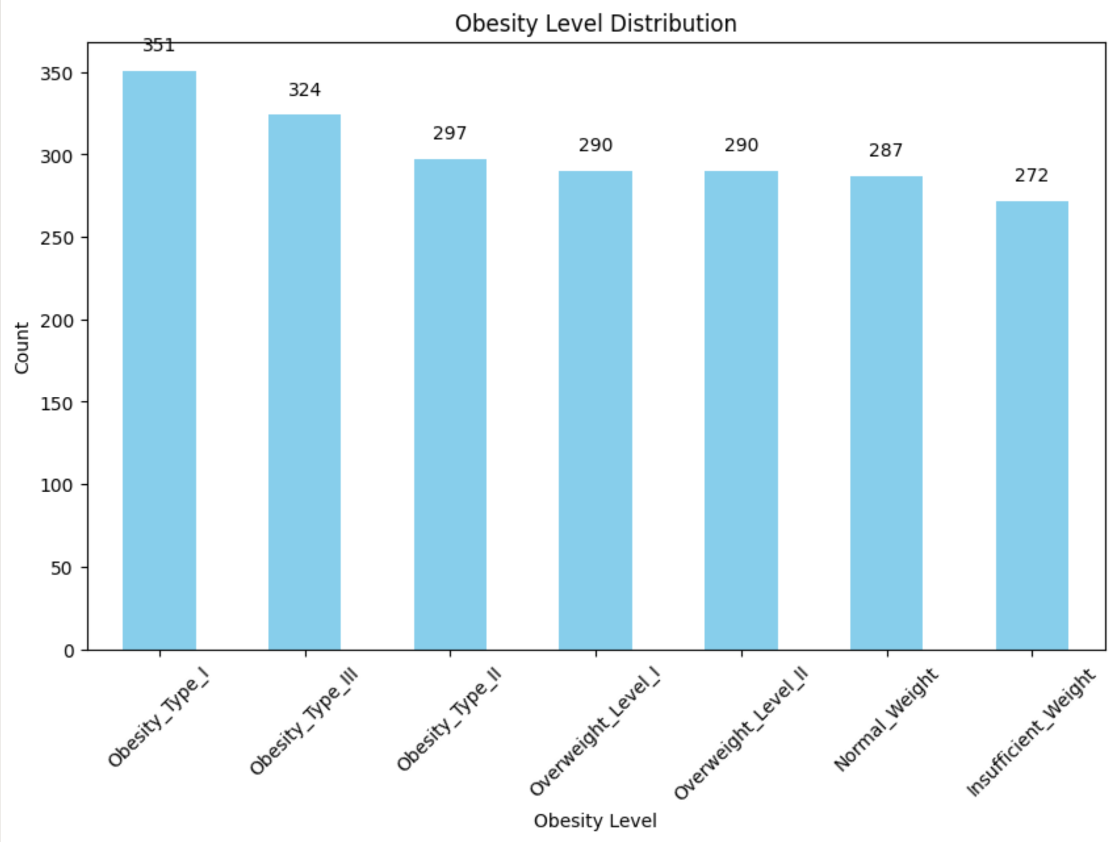
- This exploration shows the distribution of participants with family members who have had obesity problems in the past.
- 1726 members which represents 81.7% reportedly have a family history of obesity.
- This shows a high possibility of obesity in individuals who have family members with the condition

EXPLORATORY DATA ANALYSIS (EDA)



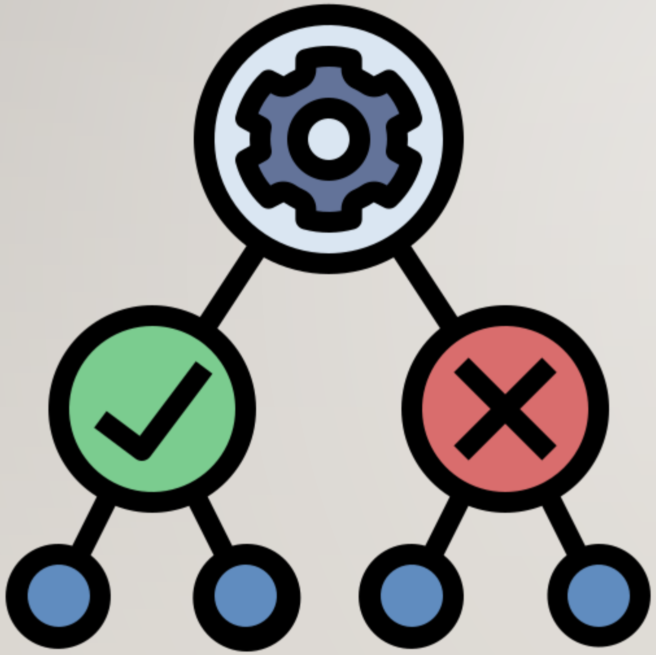
- This exploration shows participants who are actively engaged in smoking habits.
- 2067 members responded as to not engaging in any smoking habits. This represented 97.9% of the total.

EXPLORATORY DATA ANALYSIS (EDA)



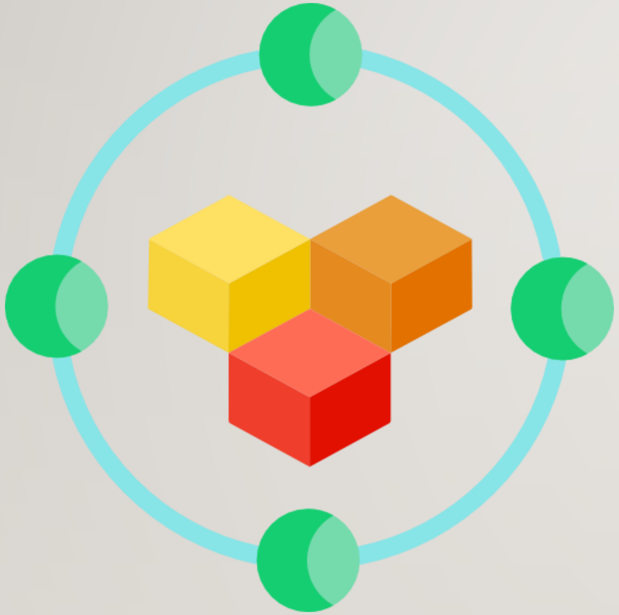
- This exploration shows the distribution of weight problems.
- The prevalent form of obesity is obesity type I followed by obesity type III and finally obesity type II
- This is the target variable

MODEL SELECTION



- The following machine learning algorithms were chosen for the data analysis.
  - Logistic regression.
  - Decision tree.
  - Random forest.
- Reason behind the model selections
  - Considerations for interpretability, performance, and suitability.

MODEL TRAINING AND VALIDATION



- The training and validation process for each model includes splitting the dataset into training and test samples.
- Each model is trained and validated individually.
- The performance and accuracy of each model is evaluated based on its accuracy score, confusion matrix and roc graph.

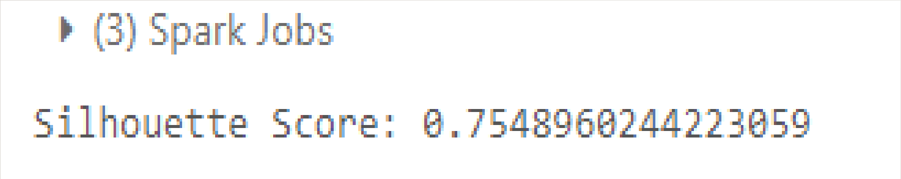
MODEL RESULTS

- Below are the performance results of the models.

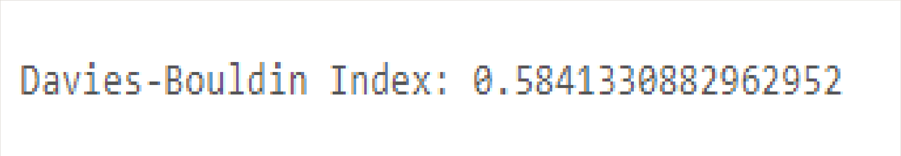
Model	Accuracy	Recall	F1-Score	ROC_AUC
Logistic Regression	0.85	0.96	0.92	0.88
Decision Tree	0.94	0.98	0.97	0.84
Random Forest	0.95	0.99	0.97	0.99

ADDITIONAL MODELS: CLUSTERING MODELS RESULTS

- A clustering algorithm was used to train and fit the dataset with a K-means of 2 and a seed of 42
- Two commonly used evaluation metrics are the Silhouette Score and the Davies-Bouldin Index.



```
Silhouette Score: 0.7548960244223059
```



```
Davies-Bouldin Index: 0.5841330882962952
```

Silhouette Score: 0.75

- A Silhouette Score of 0.75 indicates strong clustering structure, with well-separated and distinct clusters.

Davies-Bouldin Index: 0.58

- The Davies-Bouldin Index of 0.58 confirms good cluster quality, with clear separation between clusters.

CONCLUSION

- Machine learning algorithms can be used to predict the possibility of obesity based on risk factors.
- Some lifestyle factors have a greater contribution to obesity than others.