I. Introduction

In this report, we delve into a detailed analysis of the "Obesity or CVD Risk" dataset, sourced from Kaggle. This dataset presents a unique opportunity to explore the intricate relationship between various health metrics and the risk of obesity and cardiovascular diseases (CVD). Our primary objective is to uncover patterns, correlations, and insights that could contribute to a better understanding of these health conditions and potentially aid in their prediction

II. Description of the Dataset

- **Nature of the Dataset**:
  The obesity dataset is a rich collection of data aimed at estimating obesity levels in individuals. It encompasses a wide demographic from the countries of Mexico, Peru, and Colombia, covering ages between 14 and 61. This dataset is notable for its comprehensive coverage of 17 distinct attributes across 2,111 records. These attributes are meticulously gathered to reflect various aspects of eating habits and physical conditions that are crucial in understanding obesity levels.
  Attributes:
  →Related with eating habits: Frequent consumption of high caloric food (FAVC), Frequency of consumption of vegetables (FCVC), Number of main meals (NCP), Consumption of food between meals (CAEC), Consumption of water daily (CH20), and Consumption of alcohol (CALC)
  →Related with physical condition: Calories consumption monitoring (SCC), Physical activity frequency (FAF), Time using technology devices (TUE), Transportation used (MTRANS)
  →Variables obtained : Gender, Age, Height and Weight.
  →Others: family_history_with_overweight, Smoker or not(SMOKE)
  Outcome: Obesity level deducted(NObesity), values classifying into
  •Underweight Less than 18.5
  •Normal 18.5 to 24.9
  •Overweight 25.0 to 29.9
  •Obesity I 30.0 to 34.9
  •Obesity II 35.0 to 39.9
  •Obesity III Higher than 40

- **Interest Factor**:
  The dataset stands out for its comprehensive and detailed approach to understanding obesity, a critical global health issue. It offers a rich blend of demographic, dietary, and physical activity data across a diverse population from Mexico, Peru, and

Colombia. This makes it an invaluable resource for predictive healthcare modeling, public health policy development, and academic research. Its potential to provide insights into the varying factors contributing to obesity and cardiovascular diseases across different cultures and age groups is particularly intriguing, highlighting its significance in both the healthcare sector and public health initiatives.

III. Acquisition of the Dataset

- **Method of Acquisition**: Acquired via file download on Kaggle Dataset https://www.kaggle.com/datasets/aravindpcoder/obesity-or-cvd-risk-classifyregressorcluster

- **FAIRness Evaluation**:
  -The dataset is well-annotated with comprehensive metadata
  - The attributes are clearly defined
  -License: The dataset is licensed under CC BY-SA 4.0, which is clearly stated. This license is permissive and encourages sharing and adapting the dataset, provided appropriate credit is given and any changes are distributed under the same license.

IV. Data Preprocessing

- **Dataset Head overview:**

	Gender	Age	Height	Weight	family_history_with_overweight	FAVC	FCVC	NCP	CAEC	SMOKE	CH2O	SCC	FAF	TUE	CALC	MTRANS	NObeyesdad
0	Female	21.0	1.62	64.0	yes	no	2.0	3.0	Sometimes	no	2.0	no	0.0	1.0	no	Public_Transportation	Normal_Weight
1	Female	21.0	1.52	56.0	yes	no	3.0	3.0	Sometimes	yes	3.0	yes	3.0	0.0	Sometimes	Public_Transportation	Normal_Weight
2	Male	23.0	1.80	77.0	yes	no	2.0	3.0	Sometimes	no	2.0	no	2.0	1.0	Frequently	Public_Transportation	Normal_Weight
3	Male	27.0	1.80	87.0	no	no	3.0	3.0	Sometimes	no	2.0	no	2.0	0.0	Frequently	Walking	Overweight_Level_I
4	Male	22.0	1.78	89.8	no	no	2.0	1.0	Sometimes	no	2.0	no	0.0	0.0	Sometimes	Public_Transportation	Overweight_Level_II
5	Male	29.0	1.62	53.0	no	yes	2.0	3.0	Sometimes	no	2.0	no	0.0	0.0	Sometimes	Automobile	Normal_Weight
6	Female	23.0	1.50	55.0	yes	yes	3.0	3.0	Sometimes	no	2.0	no	1.0	0.0	Sometimes	Motorbike	Normal_Weight
7	Male	22.0	1.64	53.0	no	no	2.0	3.0	Sometimes	no	2.0	no	3.0	0.0	Sometimes	Public_Transportation	Normal_Weight
8	Male	24.0	1.78	64.0	yes	yes	3.0	3.0	Sometimes	no	2.0	no	1.0	1.0	Frequently	Public_Transportation	Normal_Weight
9	Male	22.0	1.72	68.0	yes	yes	2.0	3.0	Sometimes	no	2.0	no	1.0	1.0	no	Public_Transportation	Normal_Weight

- **Handling Missing Data**: There is no missing values for all columns

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

- **Handling Categorical Data:**

Column: Gender
Female => 0
Male => 1

Column: family_history_with_overweight
no => 0
yes => 1

Column: FAVC
no => 0
yes => 1

Column: CAEC
Always => 0
Frequently => 1
Sometimes => 2
no => 3

Column: SMOKE
no => 0
yes => 1

Column: SCC
no => 0
yes => 1

Column: CALC
Always => 0
Frequently => 1
Sometimes => 2
no => 3

Column: MTRANS
Automobile => 0
Bike => 1
Motorbike => 2
Public_Transportation => 3
Walking => 4

Column: NObeyesdad
Insufficient_Weight => 0
Normal_Weight => 1
Obesity_Type_I => 2
Obesity_Type_II => 3
Obesity_Type_III => 4
Overweight_Level_I => 5
Overweight_Level_II => 6

V. Summary Statistics and Potential Misinterpretations

- **Basic summary statistics for columns containing numeric data type**:

	Age	Height	Weight	FCVC	NCP	CH2O	FAF	TUE
count	2111.000000	2111.000000	2111.000000	2111.000000	2111.000000	2111.000000	2111.000000	2111.000000
mean	24.312600	1.701677	86.586058	2.419043	2.685628	2.008011	1.010298	0.657866
std	6.345968	0.093305	26.191172	0.533927	0.778039	0.612953	0.850592	0.608927
min	14.000000	1.450000	39.000000	1.000000	1.000000	1.000000	0.000000	0.000000
25%	19.947192	1.630000	65.473343	2.000000	2.658738	1.584812	0.124505	0.000000
50%	22.777890	1.700499	83.000000	2.385502	3.000000	2.000000	1.000000	0.625350
75%	26.000000	1.768464	107.430682	3.000000	3.000000	2.477420	1.666678	1.000000
max	61.000000	1.980000	173.000000	3.000000	4.000000	3.000000	3.000000	2.000000

According to this sample of data:

Ranges:

-Age: From 14 To 61 years,

-Height: From 1.45 To 1.98,

-Weight: From 39 To 173,

-FCVC: From 1 To 3.00,

-NCP: From 1 To 4.00,

-CH2O: From 1 To 3.00,

-FAF: From 0.00 To 3.00

-TUE: From 0.00 To 2.00

- **Check the Skewness and the Kurtosis of the numerical data:**

	variable	skewness	kurtosis
0	Age	1.53	2.83
1	Height	-0.01	-0.56
2	Weight	0.26	-0.70
3	FCVC	-0.43	-0.64
4	NCP	-1.11	0.39
5	CH2O	-0.10	-0.88
6	FAF	0.50	-0.62
7	TUE	0.62	-0.55

Skewness

-Between -0.5 and 0.5, the data are fairly symmetrical.

-Between -1 and -0.5 or between 0.5 and 1, the data are moderately skewed.

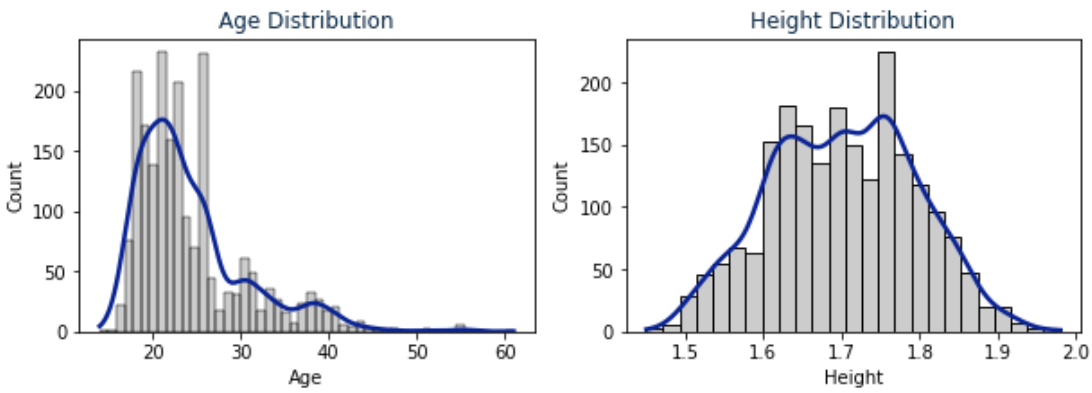
-Less than -1 or greater than 1, the data are highly skewed.

Kurtosis

-The general guideline is that if the kurtosis is greater than +2, the distribution is too peaked.

-Likewise, a kurtosis of less than −2 indicates a distribution that is too flat.

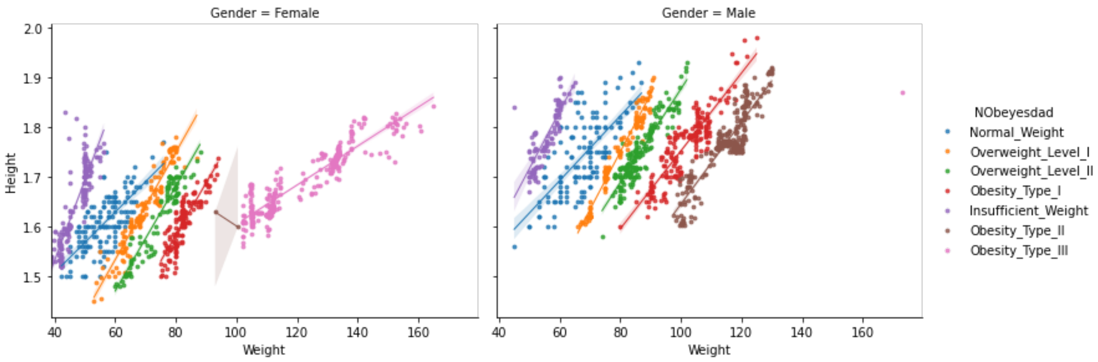
So we visualize the two variables "Age" and "Height", to see the shape of skewness = 1.53 and -0.01 and kurtosis = 2.83 and -0.56



We surprisingly find that the "Age" variable is significantly skewed and has outliers (Most ages are between 19 and 25), whereas the "Height" variable is distributed normally.

- **Very interesting figure**

The figure consists of two separate plots for 'Female' and 'Male', allowing for a gender-specific analysis of the relationship between 'Weight' and 'Height'.



We can see that **Obesity_Type_III,** the majority of the points go to women, with one point referring to a single man; **Obesity_Type_II** the majority of the points go to men, with two points going to women.

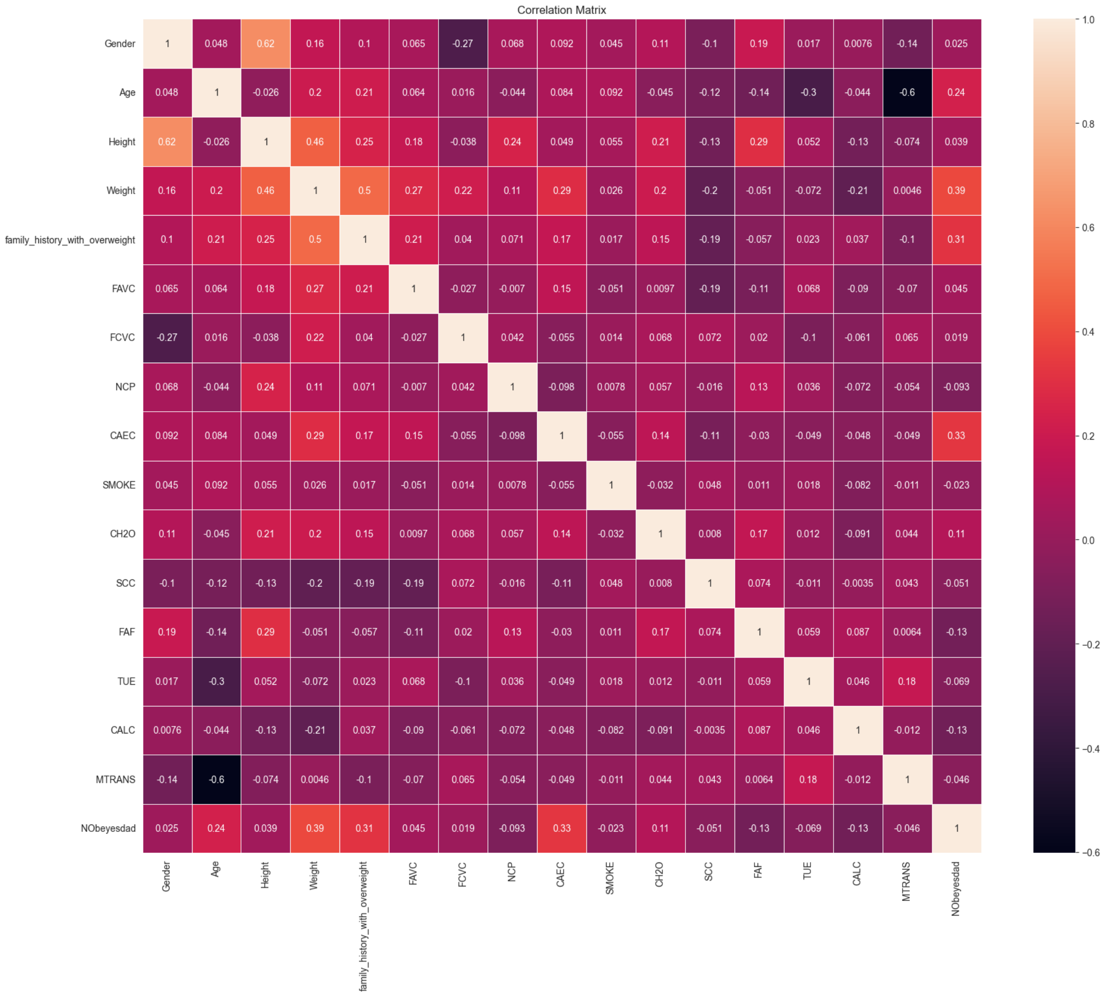
However, let's affirm this figure with numbers:

	Gender	NObeyesdad	count
0	Female	Obesity_Type_III	323
1	Male	Obesity_Type_II	295
2	Male	Obesity_Type_I	195
3	Male	Overweight_Level_II	187
4	Female	Insufficient_Weight	173
5	Female	Obesity_Type_I	156
6	Male	Normal_Weight	146
7	Female	Overweight_Level_I	145
8	Male	Overweight_Level_I	145
9	Female	Normal_Weight	141
10	Female	Overweight_Level_II	103
11	Male	Insufficient_Weight	99
12	Female	Obesity_Type_II	2
13	Male	Obesity_Type_III	1

- **Correlation heatmap**

We can visualize the strength and direction of the linear relationships between various variables.

- Age and Transportation used (MTRANS) seem to have a negative correlation, suggesting that as age increases, people are more likely to choose automobile or bike.
- 'Family_history_with_overweight' has a moderately positive correlation with 'Weight', which might indicate a genetic or lifestyle influence on weight



VI. Data Analysis

- **Chosen Analytical Methods**: KNN
- **Rationale for Analysis**: KNN is chosen for its simplicity and effectiveness in classification tasks. We choose 80% of data as training set, 20% as test set, K = 5
- **Results**:

  The confusion matrix shows how the predicted categories compare with the actual labels, with the diagonal representing correct predictions. The classification report provides precision, recall, and F1-score for each class.

```
[[51  3  0  0  0  2  0]
 [17 28  6  0  0  2  9]
 [ 0  0 72  2  0  1  3]
 [ 0  0  1 57  0  0  0]
 [ 0  0  0  0 63  0  0]
 [ 3  9  4  0  0 38  2]
 [ 1  2  3  3  1  2 38]]
              precision    recall  f1-score   support

           0       0.71      0.91      0.80        56
           1       0.67      0.45      0.54        62
           2       0.84      0.92      0.88        78
           3       0.92      0.98      0.95        58
           4       0.98      1.00      0.99        63
           5       0.84      0.68      0.75        56
           6       0.73      0.76      0.75        50

    accuracy                           0.82       423
   macro avg       0.81      0.82      0.81       423
weighted avg       0.82      0.82      0.81       423
```

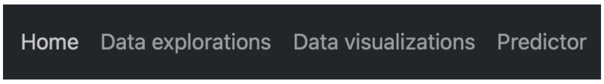
- **Surprises and Insights**:
  - The confusion matrix shows some misclassifications, particularly for the class labeled '1'.
  - The results show that some categories, such as those labeled '3' and '4', have very high precision and recall, meaning the model performs exceptionally well in classifying these categories. In contrast, category '1' shows notably lower precision and recall, indicating the model struggles with this particular class.
  - The overall accuracy of the model is 0.82
- **Validation of Analyses**: The analysis is validated by using a standard train-test split to evaluate model performance on unseen data, ensuring the evaluation metrics are indicative of the model's ability to generalize. The accuracy score presented shows the overall accuracy across all classes.

VII. Server API and Web Front-End

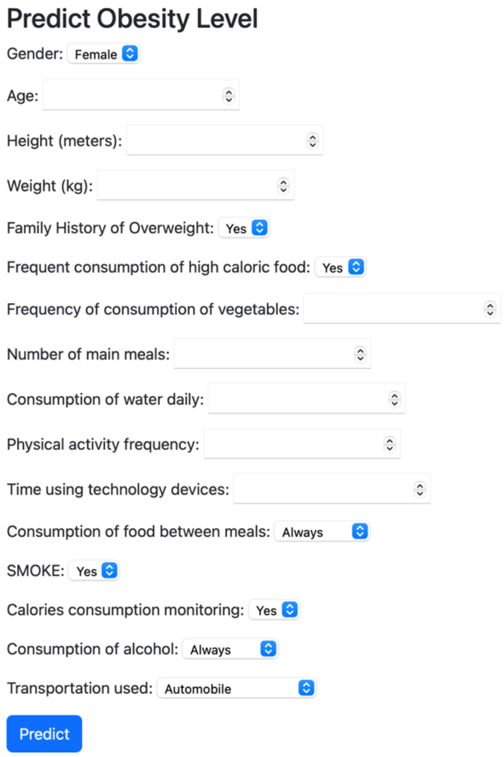
- **Server API Description**:
  The server API serves as the communication layer between the web front-end and the data processing backend, typically designed using Flask for Python. In this project, the server API's architecture would consist of endpoints (e.g.` **calc_corr.py**`, ` **knn_predict.py**`) that correspond to different analytical functions, such as data exploration, visualization, and prediction. The server API handles HTTP requests, executes the relevant Python code, and sends the results back in the response.

- **Web Front-End Overview**:
  The web front-end is designed to provide an interactive interface for users to engage with the data and analysis tools. This part of the application is built using HTML, CSS,

and JavaScript, with HTML templates like `**predict.html**` offering forms where users can input their data. Here's a breakdown of the web front-end's features:



- **Home Page**: Introduction of the project
- **Data Explorations**: Allows users to view initial data explorations, such as summary statistics and distributions.
- **Data Visualizations**: Users can navigate to view various interactive data visualizations, which include correlation matrices and distribution charts.
- **Predictor**: This page appears to provide a form where users can input various features to predict obesity levels using the implemented model.



In summary, the project employs a Flask server to create an API that processes data analysis and prediction requests, and a web front-end that provides a user-friendly interface for interacting with the data, visualizing results, and using the prediction tools.

VIII. Challenges and Surprising Findings

- **Unexpected Difficulties**:
  During the course of the project, various challenges have emerged, one of which could be related to the HTML format and the implementation of the model results' transmission to the prediction results web interface.

- **Surprising Results**:

  An unexpected result from the analysis was that category '1', which represents the 'Normal_Weight' group, showed a notably very low precision. This is surprising because one might expect the 'Normal_Weight' category to be the most balanced and well-represented in the dataset, leading to more accurate predictions. This finding may suggest that the features that distinguish 'Normal_Weight' from other categories are not as clear-cut or that the data may not be as well-separated as for other categories, which might be easier for the model to predict due to more distinct characteristics.