Estimation of obesity levels based on eating habits and physical condition

INDEX

Problem statement

The increasing prevalence of obesity has become a significant public health concern globally. Obesity is linked to various health conditions such as diabetes, cardiovascular diseases, and certain cancers, making it crucial to understand and address its underlying causes. This project aims to estimate obesity levels based on individuals' eating habits and physical conditions, providing insights into patterns and risk factors associated with obesity and a more comprehensive and individualized approach to assessing and addressing obesity.

.

Problem Definition

The project not only aims to predict obesity levels based on obvious factors like height and weight but also other important factors which will be introduced in further sections of the project since using only height, and weight alone to determine obesity levels can lead to ethical concerns, including stigmatization, misclassification and a narrow focus on health.

Our problem is a classification problem where we classify data points into different levels of obesity. Also since majority of the data (about 77%) is synthetic and balanced which does not reflect the data we see in real life, we aim to obtain a imbalanced dataset from the dataset we already have to emulate a real life dataset and apply the machine learning models we build to see their difference in performance between balanced and imbalanced scenarios. We also aim to build a neural network, and a transformer-based model to answer the same problem.

Data Source and reference:

• Dataset

UCI Machine Learning Repository

https://archive.ics.uci.edu/dataset/544/estimation+of+obesity+levels+based+on+eating+habits+and+physical+condition

• Publication Reference

Dataset for estimation of obesity levels based on eating habits and physical condition in individuals from Colombia, Peru and Mexico - ScienceDirect

Data Description:

The data is classified using the values of Insufficient Weight, Normal Weight, Overweight Level I, Overweight Level II, Obesity Type I, Obesity Type II, and Obesity Type III. The data comprises 2111 records and 17 attributes (The attributes related with eating habits are: 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). The attributes related with the physical condition are - Calories consumption monitoring (SCC), Physical activity frequency (FAF), Time using technology devices (TUE), Transportation used (MTRANS), other variables obtained were - Gender, Age, Height and Weight. Finally, all data was labeled and the class variable NObesity was created with the values of: Insufficient Weight, Normal Weight, Overweight Level I, Overweight Level II, Obesity Type I, Obesity Type II, and Obesity Type III. The two other factors are family history with obesity and Smoking habits of a person (SMOKE)). The records are labeled with the class variable NObesity (Obesity Level). 23% of the data was gathered directly from consumers via a web platform, while the remaining 77% of the data was artificially created using the Weka tool and the SMOTE filter. For reference the different obesity levels can be related to BMI index as follows:

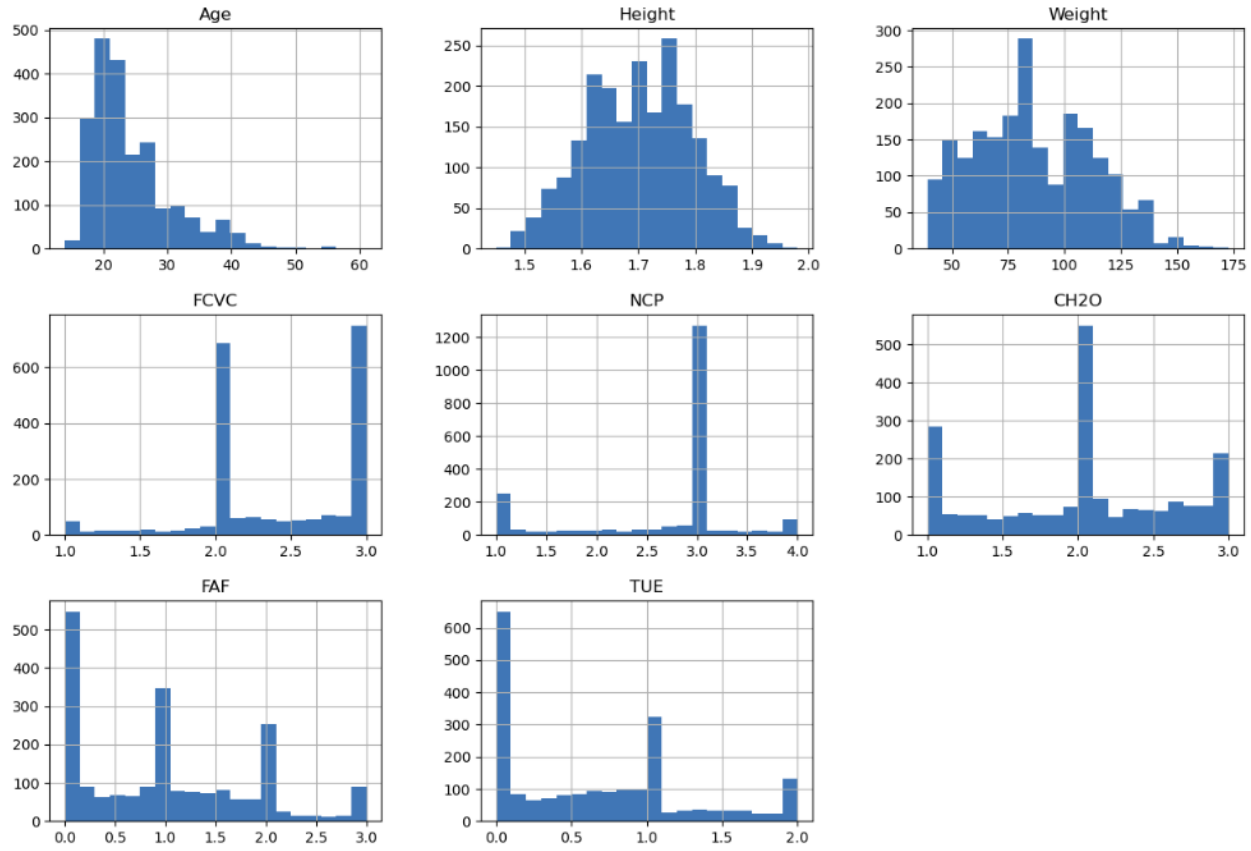
Insufficient Weight	<18.5
Normal Weight	18.5 – 24.9
Overweight Level I	25.0 – 27.4
Overweight Level II	27.5 – 29.9
Obesity Type I	30.0 – 34.9
Obesity Type II	35.0 – 39.9
Obesity Type III	>40.0

Data Exploration:

To better comprehend our data and provide the groundwork for our predictive modeling, we began using EDA and produced some visualizations, as follows.

Histogram of numerical features

By constructing histograms of numerical/already encoded features we derived the following information.



1. **Age**: Shows the distribution of ages in the dataset. The histogram indicates that most individuals are in their twenties, with fewer people as the age increases.

2. **Height**: Displays the distribution of heights. It appears most heights are centered around 1.7 to 1.8 meters, with fewer individuals being taller or shorter.

3. **Weight**: Represents the weight distribution. The data seems to be slightly right skewed, indicating a higher number of individuals with weights around 75 to 100 kilograms, with some heavier weights as well.

4. **FCVC (Frequent Consumption of Vegetables)**: Shows how often individuals consume

vegetables. The categories are likely coded as 1 (seldom), 2 (frequently), and 3 (always). The majority consume vegetables frequently.

5. NCP **(Number of Main Meals)**: Represents the number of main meals consumed daily by individuals. The categories might include options like 1, 2, 3, and more than 3 meals per day. Most individuals consume around 3 meals a day.

6. CH2O **(Consumption of Water Daily)**: Illustrates the distribution of daily water consumption levels. Most individuals consume around 2 to 3 liters of water per day.

7. FAF **(Physical Activity Frequency)**: Shows how frequently individuals engage in physical activity. The histogram suggests that many individuals do not engage in frequent physical activity, with a spread across other frequencies.

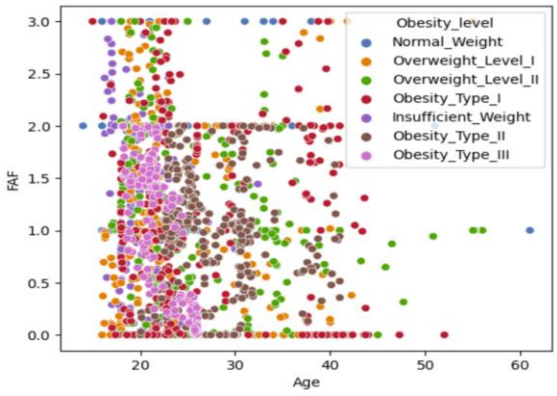
8. TUE **(Time Using Electronic Devices)**: Depicts the time spent using electronic devices. The histogram shows a high number of individuals spending less than 0.5 units of time (hours) using devices, with fewer individuals spending more time.

Pairplot and Scatterplots

We visualized pair plots to identify any interesting relationships between features. Since it did not show anything noteworthy, we focused on two relationships we thought were interesting. i.e.

Age vs. Physical Activity Frequency (FAF):

- The distinct colors represent various obesity levels from normal weight to obesity type III. The plot shows that individuals across all age groups fall into different obesity categories, with no clear trend indicating that age or physical activity frequency alone dictates obesity level.
- Individuals of the same age group show diverse physical activity levels, indicating variability in lifestyle choices among people of similar ages.

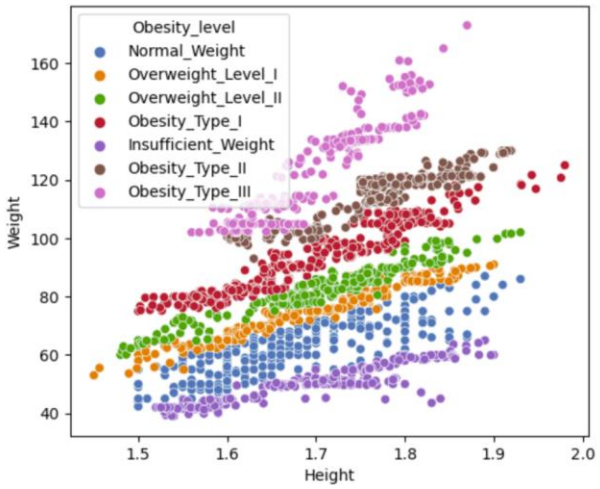


Height vs. Weight

Each color representing different obesity levels shows that:

- Higher weights and moderate heights often correlate with higher obesity levels.
- Normal weights are distributed uniformly across different heights.
- Obesity Type III tends to cluster at higher weights across a variety of heights.
- Insufficient weight is found primarily at lower weight ranges, regardless of height.

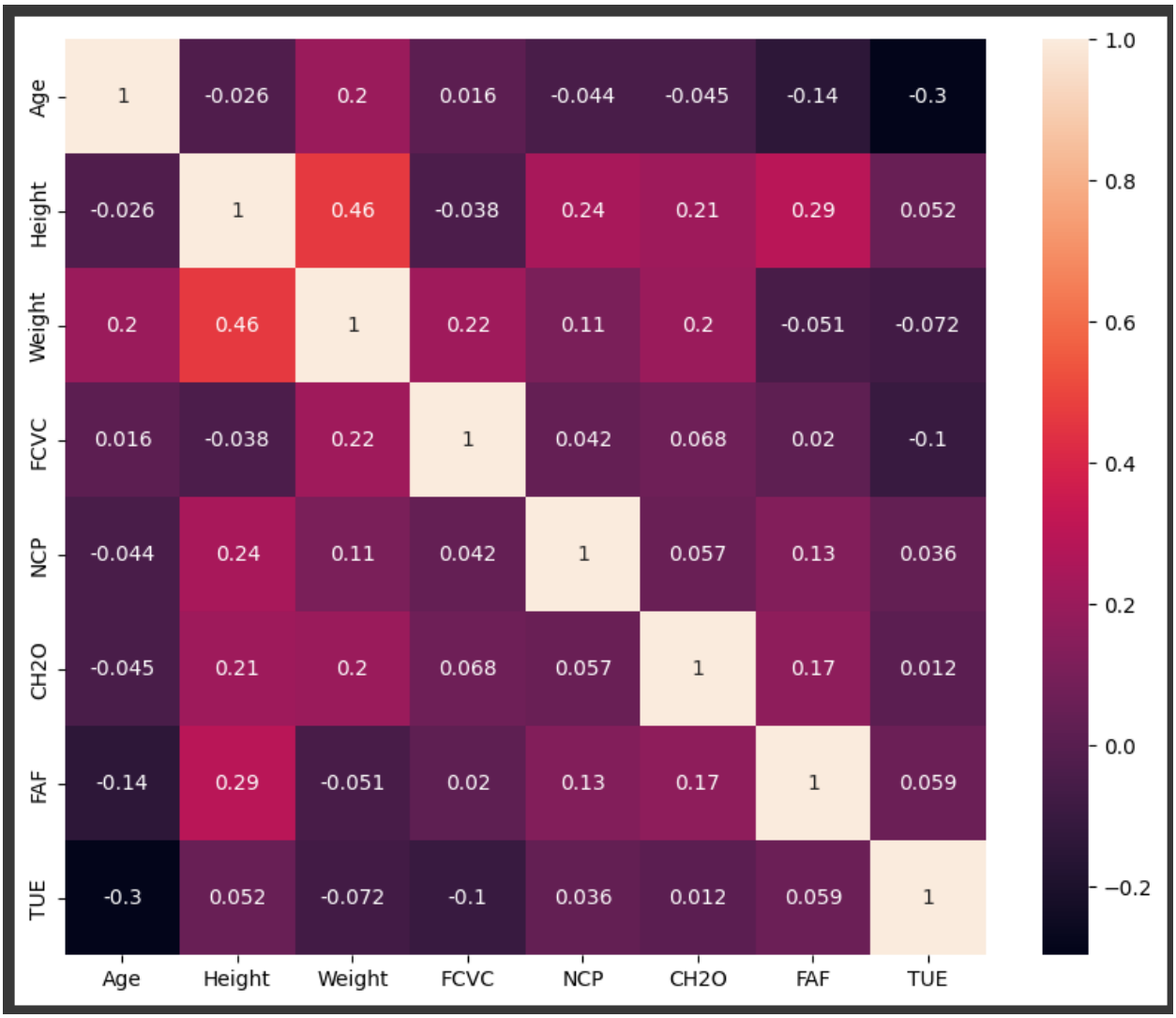
The plot underscores the importance of considering both height and weight when assessing obesity, as the same weight can represent different health statuses in individuals of different heights.



(P.S. Not displaying the pair plot as it takes up too much space)

Correlation Heatmap of numerical/already encoded features

This heatmap illustrates the strength and direction of the linear relationships between different pairs of features.



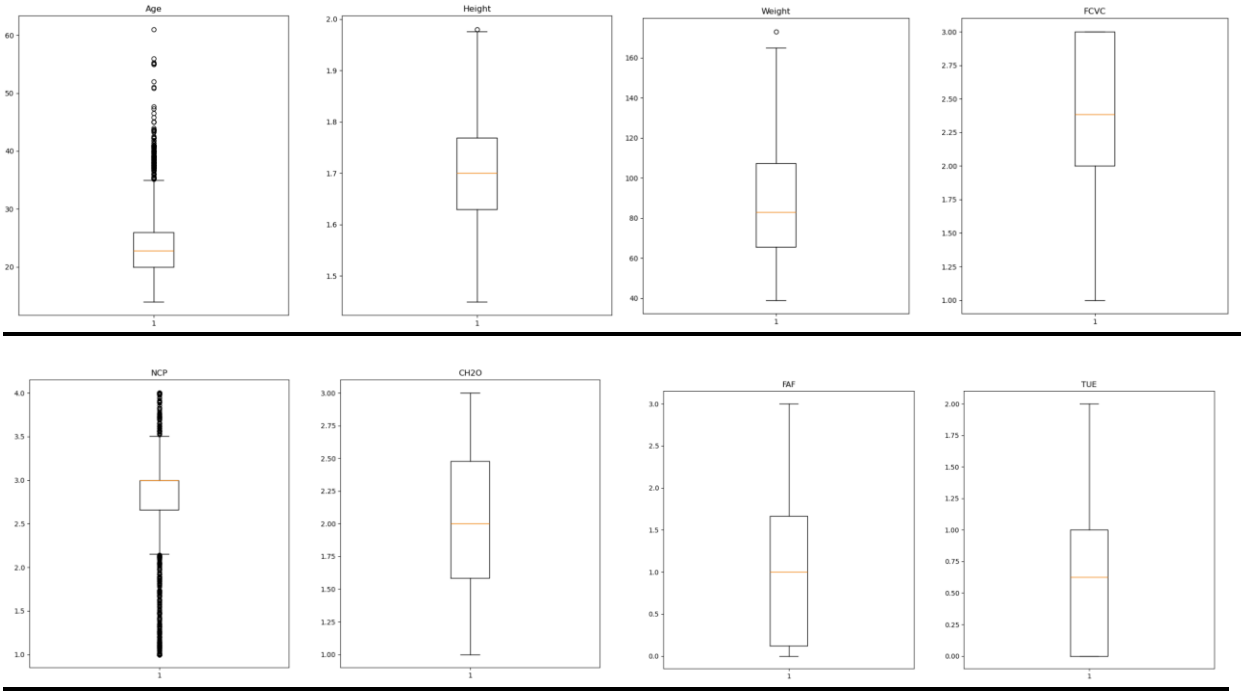
1. **Height and Weight (0.463136)**: This moderate positive correlation indicates that typically, as height increases, weight also increases. This is a common physiological pattern.

2. **Age and TUE (-0.296931)**: This moderate negative correlation suggests that older individuals tend to spend less time using electronic devices compared to younger individuals.

3. **FAF and CH2O (0.167236)**: A weak positive correlation indicating that individuals who engage more frequently in physical activities tend to consume more water. This aligns with general health advice about hydration and physical activity.

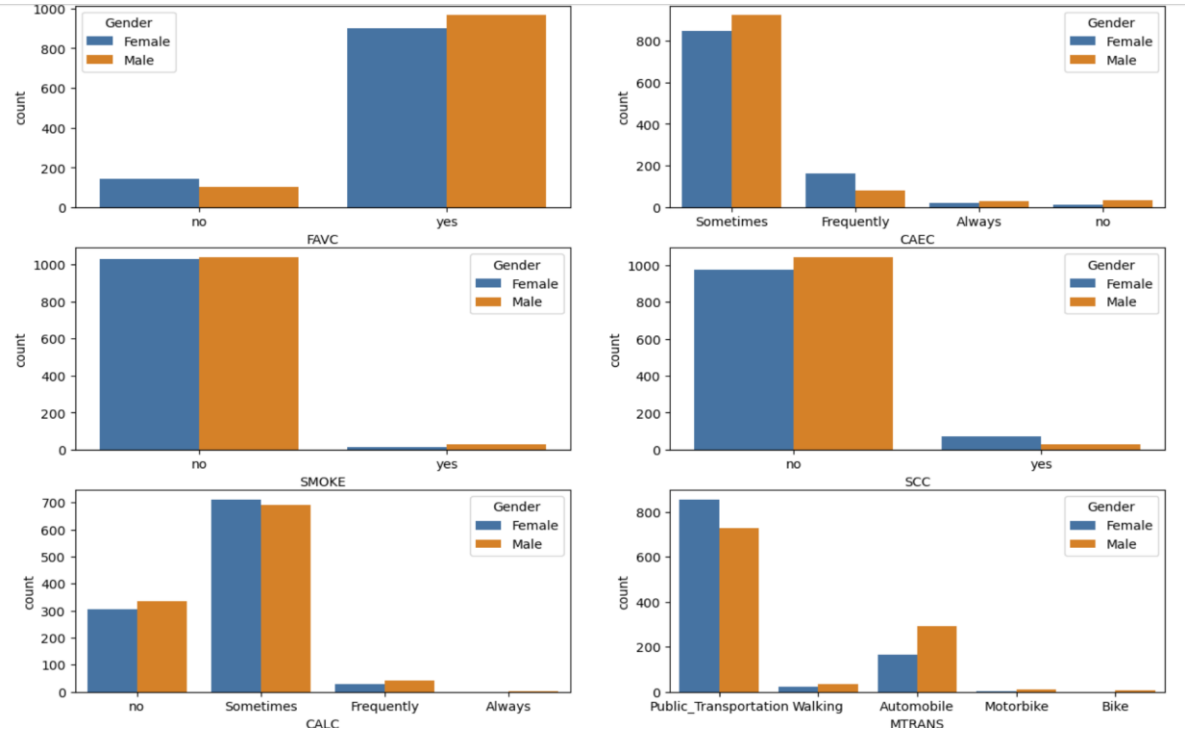
4. **Weight and FAF (-0.05)**: A weak negative correlation, which might imply that individuals with higher weight engage less in physical activities, although this relationship is weak.

Boxplots



1. **Age**: Outliers consist of individuals over 35, suggesting the presence of older adults in a younger dataset.
2. **Height**: A few outliers are taller than 1.9 meters, indicating some exceptionally tall individuals.
3. **Weight**: Outliers on both ends, especially above 160 kilograms, point to a small group with significantly higher weights.
4. **FCVC (Frequency of Consumption of Vegetables)**: There are no significant outliers, showing a general consistency in vegetable consumption habits.
5. **NCP (Number of Main Meals)**: Outliers consuming more than 3 meals suggest some individuals may have higher than usual meal frequencies.
6. **CH2O (Daily Water Consumption)**: Minimal outliers, with some consuming exceptionally high or low amounts of water daily.
7. **FAF (Physical Activity Frequency)**: Outliers at both extremes highlight a variation from individuals with no physical activity to those engaging very frequently.
8. **TUE (Time Using Electronic Devices)**: Outliers show some individuals with unusually high electronic device usage, potentially indicating heavy reliance on technology.

Bar Charts



1. FAVC (Frequent Consumption of High-Calorie Food)

- Both genders have a high count of individuals who frequently consume high-calorie foods, with males slightly outnumbering females.

2. SMOKE (Smoking Behavior)

- A large majority of both genders do not smoke, with a very small proportion of both males and females who do smoke.

3. CAEC (Consumption of Food Between Meals)

- Most individuals sometimes eat between meals, with males slightly more likely than females to eat frequently between meals. Very few always eat between meals.

4. SCC (Calories Consumption Monitoring)

- Most individuals do not monitor their calorie intake, with similar proportions between males and females.

5. CALC (Alcohol Consumption)

- Most individuals either do not consume alcohol or only sometimes consume it. A smaller number of individuals consume alcohol frequently, with very few always consuming it. Males are slightly more likely than females to consume alcohol frequently.

6. MTRANS (Modes of Transportation)

- The most common mode of transportation for both genders is walking, followed by public transportation and automobiles. Few individuals use motorbikes or bikes, with males slightly more likely to use these modes.

P.S. We have also visualized data in violin and distribution plots but chose not to include them in the report as they did not offer any new/different insights.

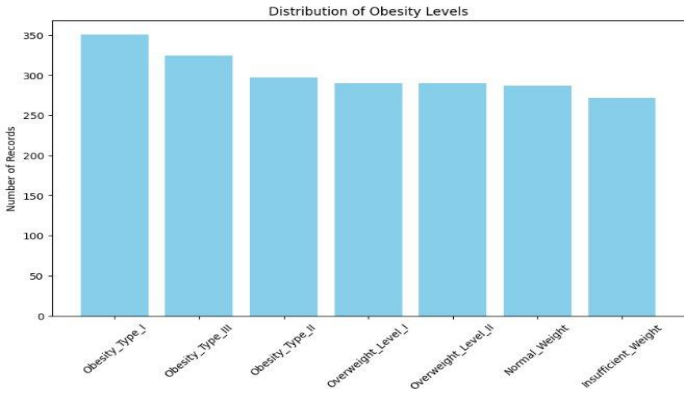
Data Mining tasks

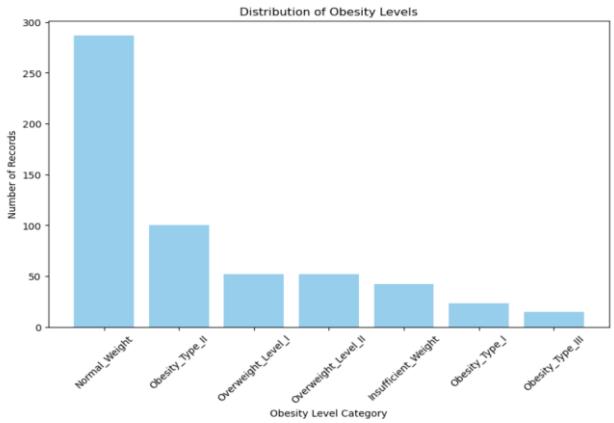
In our project we investigated several data mining tasks to prepare the dataset for predictive use:

1. **Importing and Reading Data**: The dataset had to be imported into our environment, and its contents had to be read. We were able to learn the format and data structure in a basic way as a result.
2. **Missing values**: We checked if our dataset contained any missing values, but there were not any, hence we did not have to go for any sort of imputation methods
3. **Duplicate values**: We chose to retain the duplicate rows as they are important to balance the dataset which was done using SMOTE filter during synthetic data generation
4. **Scaling data:** We used the StandardScaler from sklearn.preprocessing to standardize the numerical columns in the dataset. This process adjusts each numerical feature to have a mean of zero and a standard deviation of one. The transformation involves subtracting the mean and dividing it by the standard deviation for each column. This standardization is crucial for many machine learning algorithms as it ensures that all features contribute equally to the analysis, preventing bias toward features with larger magnitudes.
5. **Encoding Categorical columns:** We used the pd.factorize() function to convert categorical variables in the dataset into numerical codes. This is essential for preparing the dataset for machine learning algorithms, which typically require numerical input. Each categorical column such as 'Gender', 'family_history_with_overweight', and others are encoded into integers, replacing the original string categories. This transformation facilitates the computational handling of the data in statistical and machine learning models

6. **Feature selection:** We used a decision tree classifier to assess and rank the importance of features in a dataset. Decision trees are particularly useful for feature importance analysis because they measure how effectively each feature reduces uncertainty (or impurity) when creating splits in the data. This is quantified using metrics like the Gini impurity or information gain. Combined with logical interpretation and the results of Feature importance determined by Decision tree classifier we chose to exclude 'family_history_with_overweight' and SCC as features in our ML models

7. **Generating imbalanced data**: Since we wanted to build models which can be applied to real-world datasets which tend to be more imbalanced, we decided to derive a dataset from our original balanced dataset which has majority of "Normal Weight" values in our class variable. We made use of "resample" method to derive an imbalanced dataset from an originally balanced one based on the frequency distribution specified by us for each value of class variable. The class variable distributions in both unbalanced and balanced datasets are shown below respectively





Data Mining Models/Methods:

After Splitting the balanced and unbalanced datasets into train and test we used 3 machine learning models as baseline to decide which ML model to focus on for improvement

Naïve Bayes:

Naive Bayes is a suitable algorithm for estimating obesity levels based on its simplicity and efficiency, particularly when handling large datasets with categorical features. It operates quickly and provides probabilistic outputs, which help in assessing the certainty of predictions, a valuable trait in medical applications where understanding risk levels is crucial.

One of its strengths is its ability to handle categorical data effectively, such as dietary and lifestyle habits, which is a major part of the dataset. The algorithm assumes that all features are independent given the class label, which simplifies computations and can yield good performance even when some features are irrelevant. This makes Naive Bayes robust in scenarios where feature interactions are complex or unknown.

While the independence assumption may not fully capture the interactions between variables, Naive Bayes can still serve as a powerful baseline model. It provides a reference point for evaluating the performance of more sophisticated models, helping to establish a foundation for further improvements in predicting obesity levels.

Training Accuracy on balanced dataset: 0.68
Test Accuracy on balanced dataset: 0.65
Training Accuracy on unbalanced dataset: 0.50
Test Accuracy on unbalanced dataset: 0.29

Logistic Regression:

Logistic regression is a strong choice for estimating obesity levels because it can handle both binary and multiclass classification problems, making it ideal for categorizing individuals into various obesity categories. It is particularly valued for its interpretability, as the model

provides coefficients that reveal the direction and strength of the relationship between each predictor and the outcome. This feature is crucial in medical and public health fields, where understanding how lifestyle factors influence obesity is essential for designing effective interventions.

The model produces probabilistic outputs, which are useful for assessing the likelihood of an individual falling into a particular obesity category. This probabilistic approach facilitates informed decision-making and risk assessment, allowing healthcare providers to tailor interventions based on predicted probabilities.

Logistic regression efficiently manages datasets with both categorical and continuous variables, which are present in our dataset.

Training Accuracy on balanced dataset: 0.90
Test Accuracy on balanced dataset: 0.89
Training Accuracy on unbalanced dataset: 0.79
Test Accuracy on unbalanced dataset: 0.77

Support Vector Classification

Support Vector Classification (SVC) is a powerful method for estimating obesity levels due to its ability to handle complex, non-linear relationships in data.

SVC can model non-linear relationships using kernel functions such as the radial basis function (RBF). This is particularly useful in obesity prediction, where relationships between lifestyle factors and obesity levels are often complex and non-linear

By maximizing the margin between different classes, SVC tends to generalize well to new, unseen data, reducing the risk of overfitting, which is crucial for reliable predictions in medical applications.

SVC supports various kernels (linear, polynomial, RBF, etc.), allowing the model to adapt to the specific nature and complexity of the dataset. This flexibility is essential for capturing

interactions between eating habits, physical conditions, and obesity levels.

SVC works well with datasets that have many features, making it suitable for obesity studies that consider many factors like dietary habits, exercise levels, and demographic information. Hence in case we want to extend the number of features or add more medically relevant biochemical data to our dataset, SVC will be able to handle them

SVC can handle both binary and multiclass classification problems, making it appropriate for categorizing individuals into different obesity levels.

Training Accuracy on balanced dataset: 0.95
Test Accuracy on balanced dataset: 0.93

Training Accuracy on unbalanced dataset: 0.86
Test Accuracy on unbalanced dataset: 0.82

Since SVC had the best performance out of all 3 baseline models, we decided to use it as the main model for our problem
First, we applied the SVC model trained on the balanced dataset to the unbalanced dataset to see if our assumption that applying the model trained on the balanced data gives a better performance when applied to the imbalanced data (emulating the original data i.e. obtained from various sources in the real world) and upon doing that we got the following results.
Here the unbalanced data acts as a validation set since it is not exposed to the model in balanced data.

Using whole imbalanced dataset as validation set: 0.91

Then we decided to tune the SVC model using GridSearchCV, since GridSearchCV is a powerful tool for hyperparameter tuning of SVC in the context of obesity level estimation. By systematically searching through a predefined parameter space and using cross-validation to assess performance, GridSearchCV ensures that the SVC model is fine-tuned for optimal

performance. This results in a more accurate and generalizable model, capable of effectively capturing the complex relationships between eating habits, physical conditions, and obesity levels.

Best Parameters: {'C': 100, 'degree': 2, 'gamma': 1, 'kernel': 'linear'}
Best Estimator: SVC (C=100, degree=2, gamma=1, kernel='linear')
Training accuracy on tuned balanced model: 0.98
For whole imbalanced dataset as validation set: 0.98

As we can see, the best estimator is Linear Kernel in this case, which means there is no transformation applied on the existing dataset and it is a linear hyperplane. We also see that the gamma value is 1, indicating that the model gives more emphasis to correct classification. It might lead to a bit of overfitting.

But our focus when it comes to metrics was to make sure the obese classes are not misclassified as not obese hence, we focus on Recall as a primary metric of evaluation. All the Recall information for SVM in case of balanced, imbalanced and hyperparameter tuning and Model trained on balanced dataset applied on imbalanced dataset are included in the performance evaluation.
One thing to note is that there is a gradual increase in performance from models on imbalanced dataset to model trained on balanced and applied on imbalanced to model just trained and applied on balanced and finally to model trained on balanced and hyperparameter tuned against imbalanced data using GridSearchCV.

Neural Network

We implemented a simple Sequential Neural Network model to see how the model performs on this dataset as our dataset also contains multiple features which might have non-linearity between them.
For implementation of Neural Network, the target variable is encoded to show numerical value using one hot encoding. In terms of architecture, the model uses 2 hidden layers with 128 and 64 neurons respectively with ReLu as activation function. As it is multiclass classification

problem, SoftMax is used as the activation function at the output layer and the loss function used here is Categorical Cross Entropy.
For this model, we have considered the train split of the balanced dataset as training data, test split of the balanced dataset as the validation set and the entire unbalanced dataset as the test dataset.
Validation accuracy on balanced model: 0.96
For whole imbalanced dataset as test set: 0.98
As we can see, the model is generalizing too well on the unseen data as the test accuracy is higher than the validation set. We can conclude that in this case, considering a Neural network model might not be good idea as simpler models like SVM are performing better.

Transformer

We also tried a simple transformer model on our dataset. Transformer based models generally work well with sequential data for Natural Language Processing. But as proof of concept for our project, we tested the Transformer model on the tabular data.
Like Neural network, we had to encode the target variable using one hot encoding technique and transform it into PyTorch tensors to fit it in the Transformer model.
On the model architecture, it is a simple transformer model with the encoder part only using 20 epochs and Cross Entropy as the Loss function. Below is the accuracy of the transformer model:

Train accuracy on balanced model: 0.98
For whole imbalanced dataset as test set: 0.92

In this case, we can see the model is a little overfitting on the train dataset as the training accuracy is higher than the test accuracy. We can see that using a Transformer model might not be very effective as other simpler models are performing better and can save the computation cost for this dataset.

Performance Evaluation:

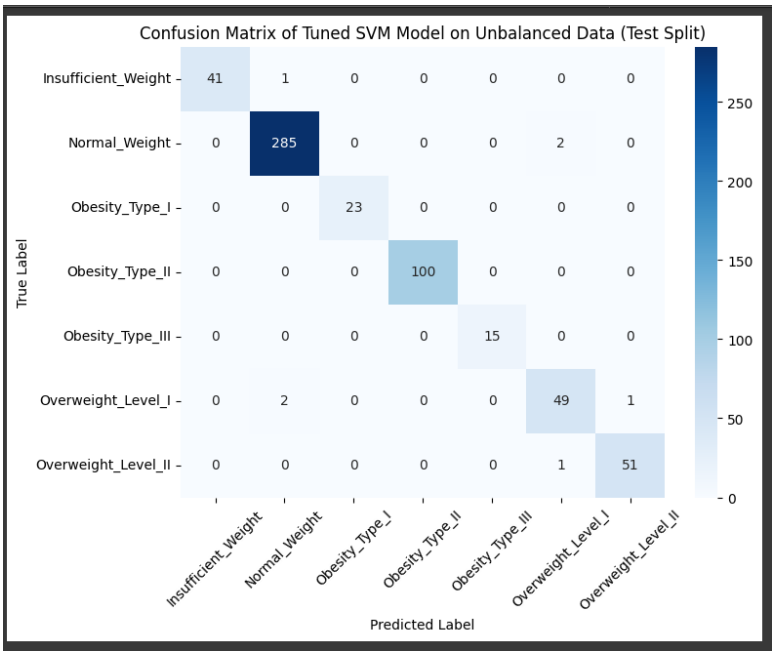
Initially we have the accuracy metrics of all models which we referred to see which model we wanted to focus more on as follows

Model	Train Accuracy	Test Accuracy
Naïve Bayes on balanced Dataset	~68%	~65%
Naïve Bayes on unbalanced dataset	~50%	~29%
Logistic Regression on balanced dataset	~90%	~89%
Logistic Regression on unbalanced dataset	~79%	~77%
SVM on balanced dataset	~90%	~95%
SVM on unbalanced dataset	~80%	~85%
Tuned SVM on unbalanced dataset	~97%	~98%
Neural Network	~98%	~98%
Transformer Model	~98%	~92%

Based on how both train and test data splits performed in our baseline models (i.e. Naïve Bayes, Logistic Regression and SVM) we decided to focus on SVM as our main model to experiment with (Please note that Neural Network and Transformer models were done after the Baseline models)

Below is the confusion matrix containing all possible values in the target class

Which shows what is the misclassification count for each class. As we can from the matrix, the diagonals are fully populated, that means, that all the target class is being predicted correctly for most of the times. Also, all other values other than the diagonal is zero indicating minimal misclassification.



In SVM even-though we focus mostly on recall, we made binary classification reports for all 7 possible values in the Target class which contains information about Accuracy, Precision, Macro avg, Weighted avg, Recall, F-1 score and Support (Number of records in the class for that particular category) for reference along with the confusion matrix for all above. This was done when SVM(SVC) was applied to balanced dataset, imbalanced dataset, balanced model on imbalanced dataset, and hyperparameter tuned model (all of which can be seen in the jupyter notebook/html file of the same submitted as a supplement) but since displaying all of them here would make the report too verbose, we are doing the same only for the most important implementation i.e. when the model developed on balanced dataset was applied to the imbalanced dataset which acts as a validation set, which can be seen as below.

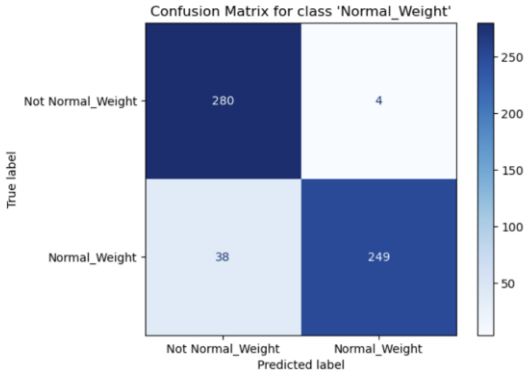
Insufficient Weight:

	precision	recall	f1-score	support
Not Insufficient_Weight	1.00	0.97	0.98	529
Insufficient_Weight	0.72	0.98	0.83	42
accuracy			0.97	571
macro avg	0.86	0.97	0.91	571
weighted avg	0.98	0.97	0.97	571

Confusion Matrix for class 'Insufficient_Weight':
[[513 16]
[1 41]]

Normal Weight:

	precision	recall	f1-score	support
Not Normal_Weight	0.88	0.99	0.93	284
Normal_Weight	0.98	0.87	0.92	287
accuracy			0.93	571
macro avg	0.93	0.93	0.93	571
weighted avg	0.93	0.93	0.93	571



Confusion Matrix for class 'Normal_Weight':
[[280 4]
[38 249]]

Obesity_Type_1:

	precision	recall	f1-score	support
Not Obesity_Type_I	1.00	1.00	1.00	548
Obesity_Type_I	1.00	0.96	0.98	23
accuracy			1.00	571
macro avg	1.00	0.98	0.99	571
weighted avg	1.00	1.00	1.00	571

Confusion Matrix for class 'Obesity_Type_I':
[[548 0]
[1 22]]

Obesity_Type_2:

	precision	recall	f1-score	support
Not Obesity_Type_II	1.00	1.00	1.00	471
Obesity_Type_II	1.00	1.00	1.00	100
accuracy			1.00	571
macro avg	1.00	1.00	1.00	571
weighted avg	1.00	1.00	1.00	571

Confusion Matrix for class 'Obesity_Type_II':
[[471 0]
[0 100]]

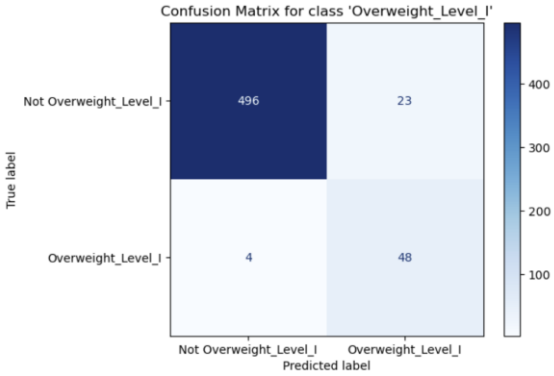
Obesity_Type_3:

	precision	recall	f1-score	support
Not Obesity_Type_III	1.00	1.00	1.00	556
Obesity_Type_III	1.00	1.00	1.00	15
accuracy			1.00	571
macro avg	1.00	1.00	1.00	571
weighted avg	1.00	1.00	1.00	571

Confusion Matrix for class 'Obesity_Type_III':
[[556 0]
[0 15]]

Overweight_Level_1:

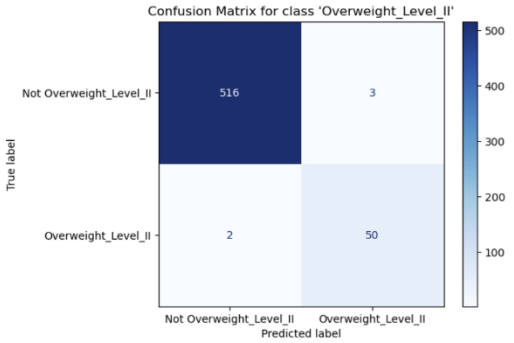
	precision	recall	f1-score	support
Not Overweight_Level_I	0.99	0.96	0.97	519
Overweight_Level_I	0.68	0.92	0.78	52
accuracy			0.95	571
macro avg	0.83	0.94	0.88	571
weighted avg	0.96	0.95	0.96	571



Confusion Matrix for class 'Overweight_Level_I':
[[496 23]
[4 48]]

Overweight_Level_2:

	precision	recall	f1-score	support
Not Overweight_Level_II	1.00	0.99	1.00	519
Overweight_Level_II	0.94	0.96	0.95	52
accuracy			0.99	571
macro avg	0.97	0.98	0.97	571
weighted avg	0.99	0.99	0.99	571



Confusion Matrix for class 'Overweight_Level_II':
[[516 3]
[2 50]]

A more comprehensive collection of recall values which we have chosen as our primary metric for evaluation to avoid misclassification of Obese as Non obese is given in the following table to see a gradual increase in performance from models on imbalanced dataset to model trained on balanced and applied on imbalanced to model just trained and applied on balanced and finally to model trained on balanced and hyperparameter tuned against imbalanced data using GridSearchCV

SVM	Insufficient_weight	Normal_Weight	Obesity_type_1	Obesity_type_2	Obesity_type_3	Overweight_level_1	Overweight_level_2
Trained and tested on Unbalanced	0.33	1	0.43	0.94	1	0.22	0.75
Trained on Balanced Tested on Unbalanced	0.98	0.87	0.96	1	1	0.92	0.96
Trained and tested on Balanced	0.97	0.81	0.97	0.98	1	0.92	0.84
(Hyper tuned)Trained on balanced Tested on unbalanced	0.98	1	1	1	1	0.94	0.98

Project Results:

In the project we aimed to classify data points from an individual based on their physical attributes and lifestyle choices to provide a more comprehensive approach on classifying obesity levels. The predictive capability of Naïve Bayes and Logistic regression as baseline models was sub-par (even with cross validation). SVC, as a model was good at classification on which we focused more by applying it individually on both imbalanced data(which we generated by sampling the balanced data) and balanced data and further applied the model trained on balanced dataset on imbalanced dataset (which emulates real-life dataset) to improve the performance of the model which we further improved by performing hyperparameter tuning on the same.

We also experimented with achieving the same goal using Neural Network and a Simple Transformer Model but realized that we are achieving almost the same results using SVC with less computational complexity. Also, we determined that SVC is even more computationally complex than Naïve Bayes and Logistic regression, the trade off in classificational performance validates using SVC.

Conclusion

- In conclusion, this project highlights the critical role of dietary habits and physical conditions in understanding and addressing obesity, emphasizing the need for personalized and comprehensive strategies to combat this global health issue.
- The ML model (SVC) created here can be improved and extended in the medical domain to help analyze and predict obesity levels, considering a multitude of factors, thereby supporting the development of targeted interventions and informed public health policies aimed at reducing obesity-related health complications.