Deriving Insights from an Obesity Dataset

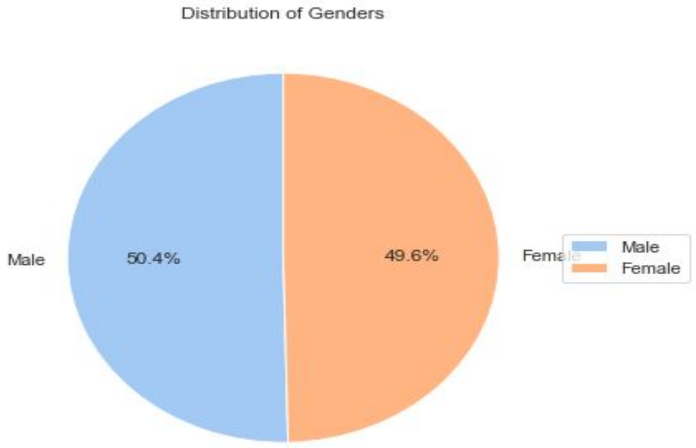
Group 14
Alokita Karmokar
Anbo Feng
Benjamin Nguyen
Christopher Lin
Sampada Shelke

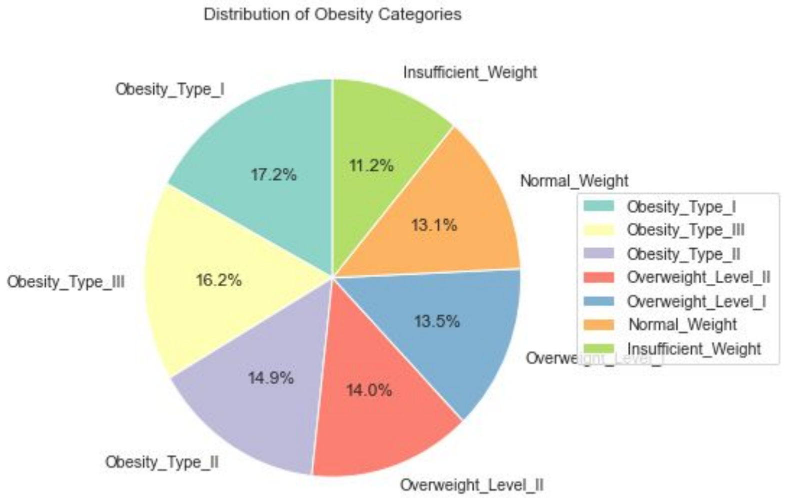
Background of data

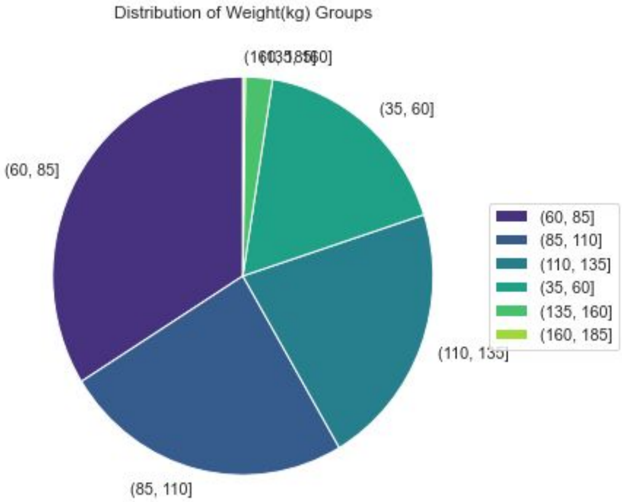
- **Gender:** Categorical - Female or Male
- **Age:** Numeric - Age of the individual
- **Height:** Numeric - Height in meters
- **Weight:** Numeric - Weight in kilograms
- **Family History:** Binary - Family history of obesity (Yes or No)
- **FCHCF:** Binary - Frequent consumption of high caloric food (Yes or No)
- **FCV:** Categorical - Frequency of vegetable consumption (1 = Never, 2 = Sometimes, 3 = Always)
- **NMM:** Categorical - Number of main meals (1, 2, 3, or 4)
- **CFBM:** Categorical - Consumption of food between meals (1 = No, 2 = Sometimes, 3 = Frequently, 4 = Always)
- **Smoke:** Binary - Smoking habit (Yes or No)
- **CW:** Categorical - Consumption of water (1 = Less than a liter, 2 = 1-2 liters, 3 = More than 2 liters)
- **CCM:** Binary - Calorie consumption monitoring (Yes or No)
- **PAF:** Categorical - Physical activity frequency per week (0 = None, 1 = 1 to 2 days, 2 = 2 to 4 days, 3 = 4 to 5 days)
- **TUT:** Categorical - Time using technology devices a day (0 = 0-2 hours, 1 = 3-5 hours, 2 = More than 5 hours)
- **CA:** Categorical - Consumption of alcohol (1 = Never, 2 = Sometimes, 3 = Frequently, 4 = Always)
- **Transportation:** Categorical - Mode of transportation (Automobile, Motorbike, Bike, Public Transportation, Walking)
- **Obesity:** Categorical - Weight classification (Insufficient weight, Normal weight, Level I overweight, Level II overweight, Type I obesity, Type II obesity, Type III obesity)

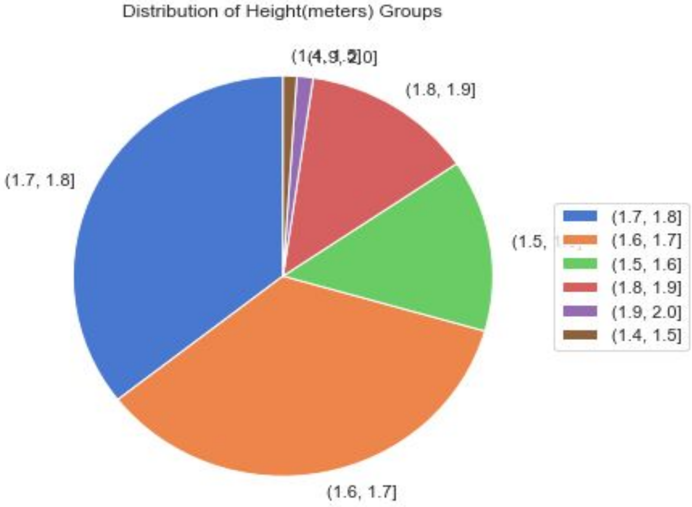
- Dataset hail from Mexico, Peru, and Colombia
- Dataset sourced from UC Irvine’s Machine Learning Repository, encompassing information on 2111 individuals aged 14 to 61.

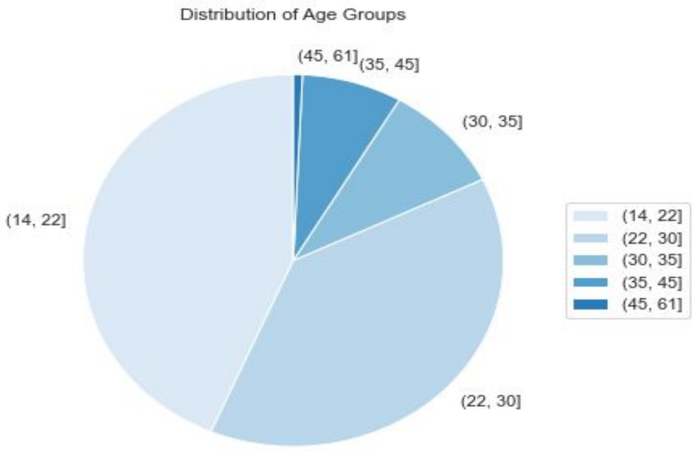
Distribution of Respondents





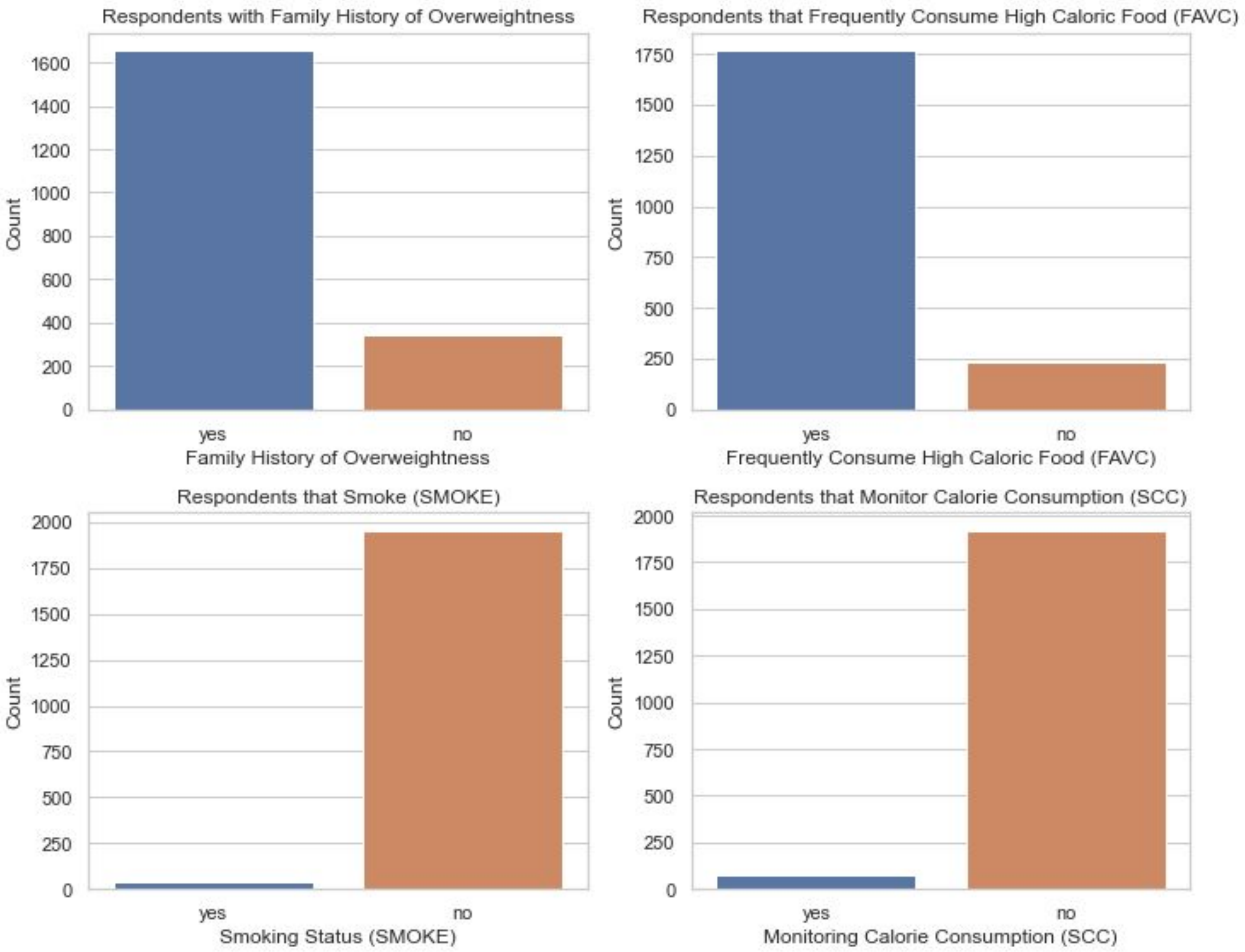






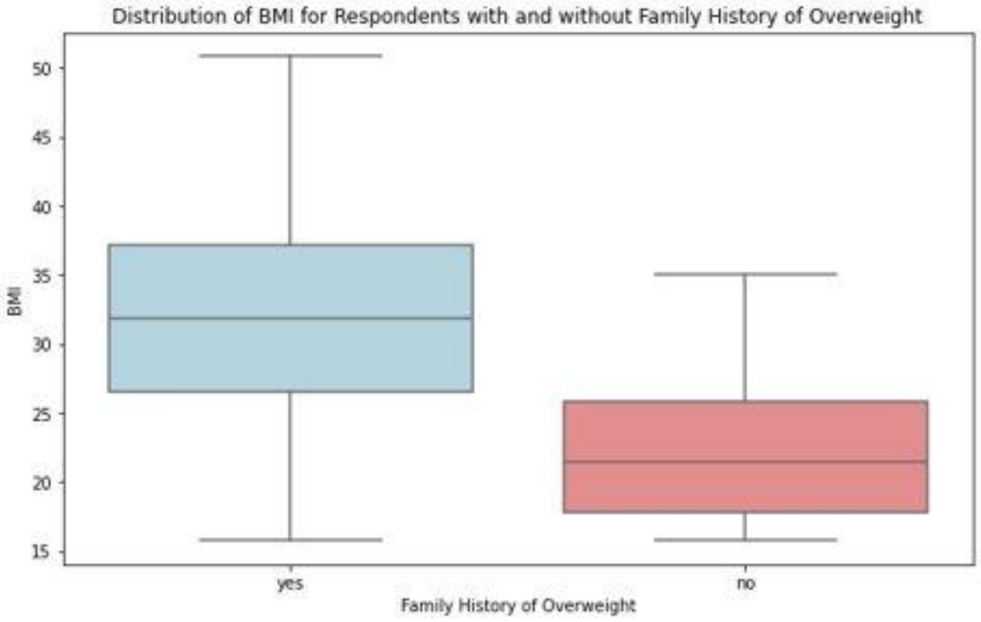
Which data should we analyze?

- **Not much data on respondents who smoke or monitor their calorie so we will not analyze those.**

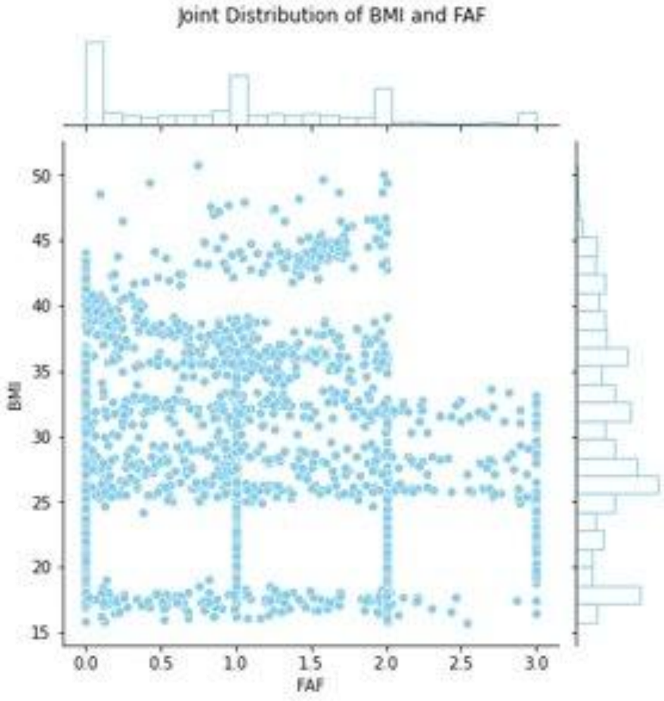


How does the prevalence of obesity differ between those with and without a family history of obesity?

- Pearson correlation coefficient is a correlation coefficient that measures linear correlation between two sets of data.
- T-test is a type of statistical analysis used to compare the averages of two groups and determine whether the differences between them are more likely to arise from random chance.
- Result: people who have family history tend to have a higher chance of obesity.

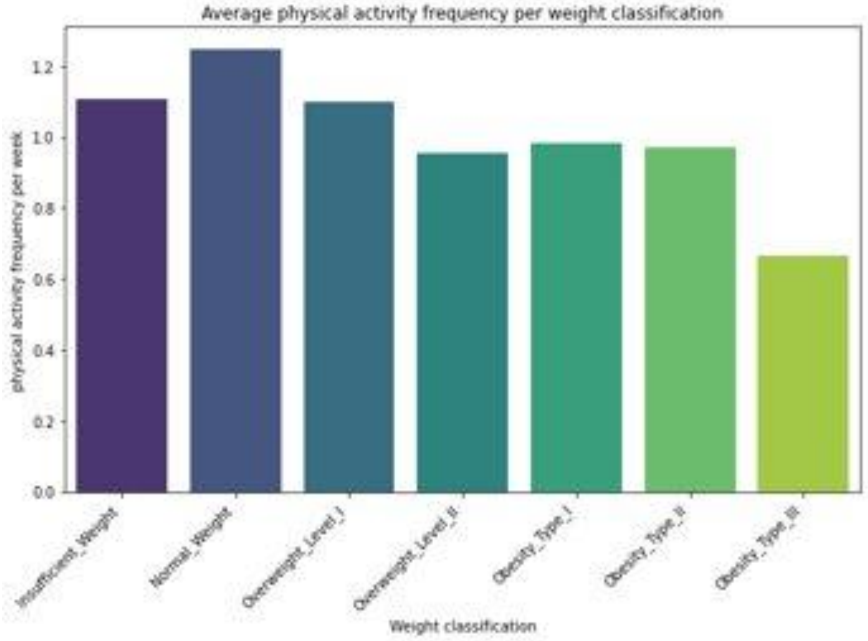



```
Average BMI with family history: 31.89
Average BMI without family history: 21.80

T-test results:
T-statistic: 34.6979
P-value: 0.0000
The difference in average BMI is statistically significant.

Correlation results:
Pearson correlation coefficient: 0.4820
P-value: 0.0000
The correlation between BMI and family history is statistically significant.
```

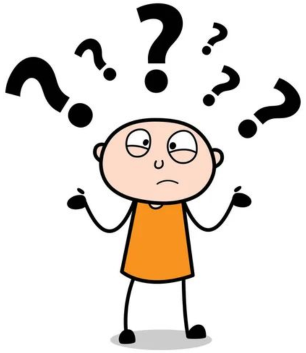


Is There is correlation Between Physical Activity Frequency and BMI

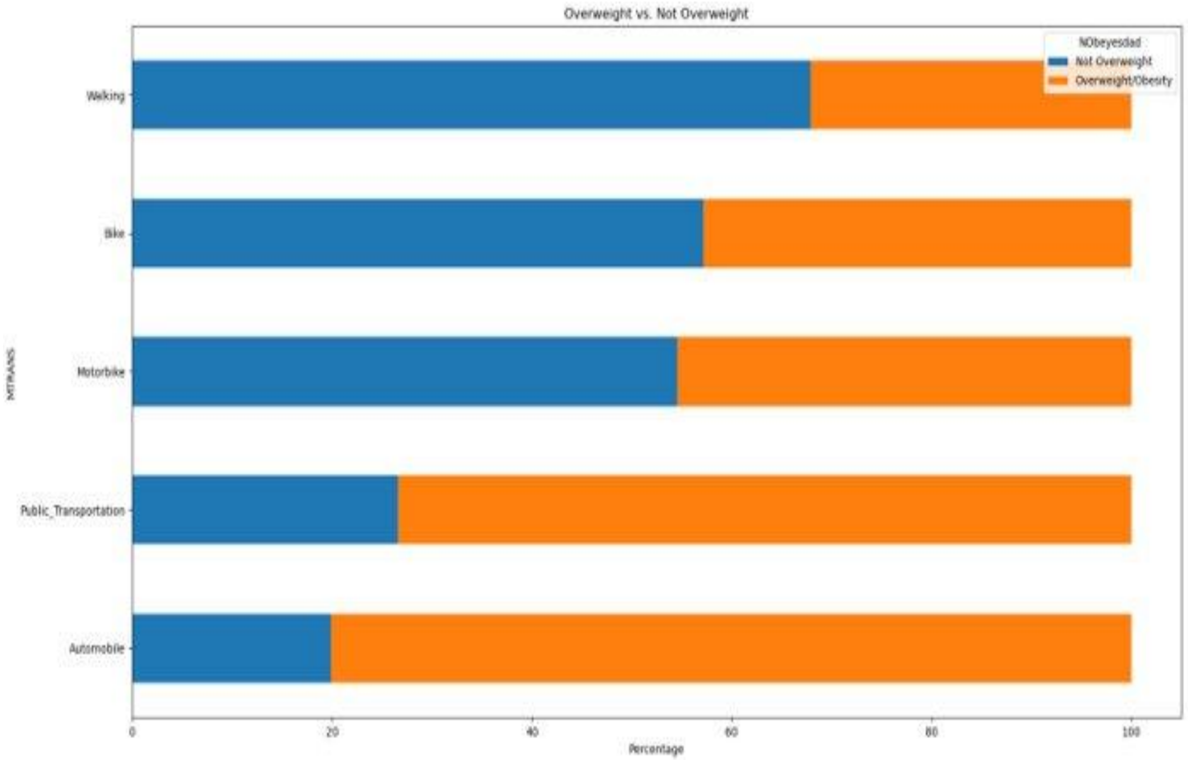
- No correlation?




```
Average FAF per NObeyesdad:
           NObeyesdad       FAF
0  Insufficient_Weight  1.107083
1        Normal_Weight  1.248092
2       Obesity_Type_I  0.982697
3      Obesity_Type_II  0.971857
4     Obesity_Type_III  0.664817
5   Overweight_Level_I  1.098116
6  Overweight_Level_II  0.955990

Correlation between FAF and BMI Result:
Pearson correlation coefficient: -0.1481
Low(close to NONE) degree of correlation
```



How is Physical Activity Define?



Conclusion:

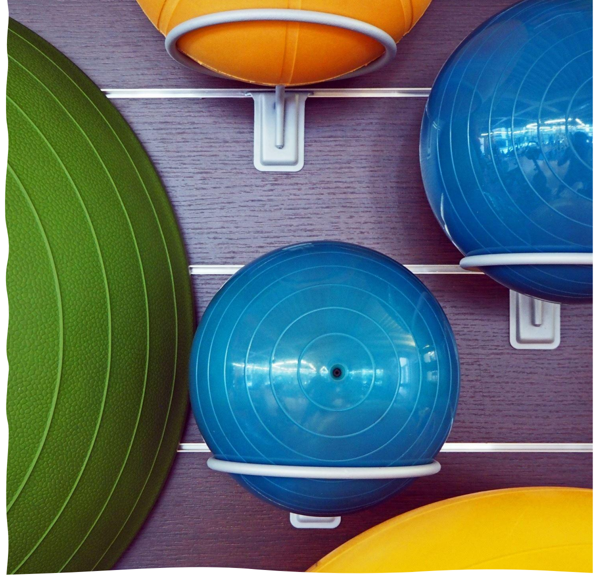
- Analyzing data from indirect questions without due diligence can steer us toward misleading conclusions.
- While genetic factors significantly impact BMI, acknowledging the role of individual lifestyle choices is essential for effective health management.

Classifier

LightGBM: A Highly Efficient Gradient Boosting Decision Tree

- LightGBM is a gradient boosting framework that uses tree based learning algorithms.
- It is designed to be distributed and efficient with the following advantages:
- Faster training speed and higher efficiency.
- Lower memory usage.
- Better accuracy.
- Support of parallel and GPU learning.
- Capable of handling large-scale data.
- LightGBM is a gradient boosting framework that uses tree based learning algorithm.

Other Relevant Features to Detect Obesity?



- **Feature Engineering**
- We decided to drop 3 features:
- -Weight
- - Strongly correlates with one’s obesity level
- - Drop it to prevent multicollinearity and reduce the risk of ov
- -Smoke, Calorie Monitoring
- - Highly imbalanced features introduce bias in the model, whi generalization performance and leads to false negatives and
- - Machine learning models often require a sufficient number category to generalize well, otherwise the model may struggl patterns from the minority.

Methodology

Weight,
Smoke,
Calorie Monitoring

```
Yes/no enode columns: no-0 yes-1
family_history_with_overweight, FAVC, TUE

Gender encode columns: female-0 male-1
Gender

Frequency encode columns: Always-0, Frequently-1, No -2, Sometimes -3
CAEC, CALC

Transport encode columns: auto-0, bike-1, motor-2, public-3, walk-4
MTRANS

Obesity encode columns: insufficient-0, normal-1, obes_I-2, obes_II-3, obes_III-4, over_I-5, over_II-6
NObeyesdad
```

Faster training speed

Lower Memory Usage

Decision Tree on Steroids

Normal: 70: 30 split of data

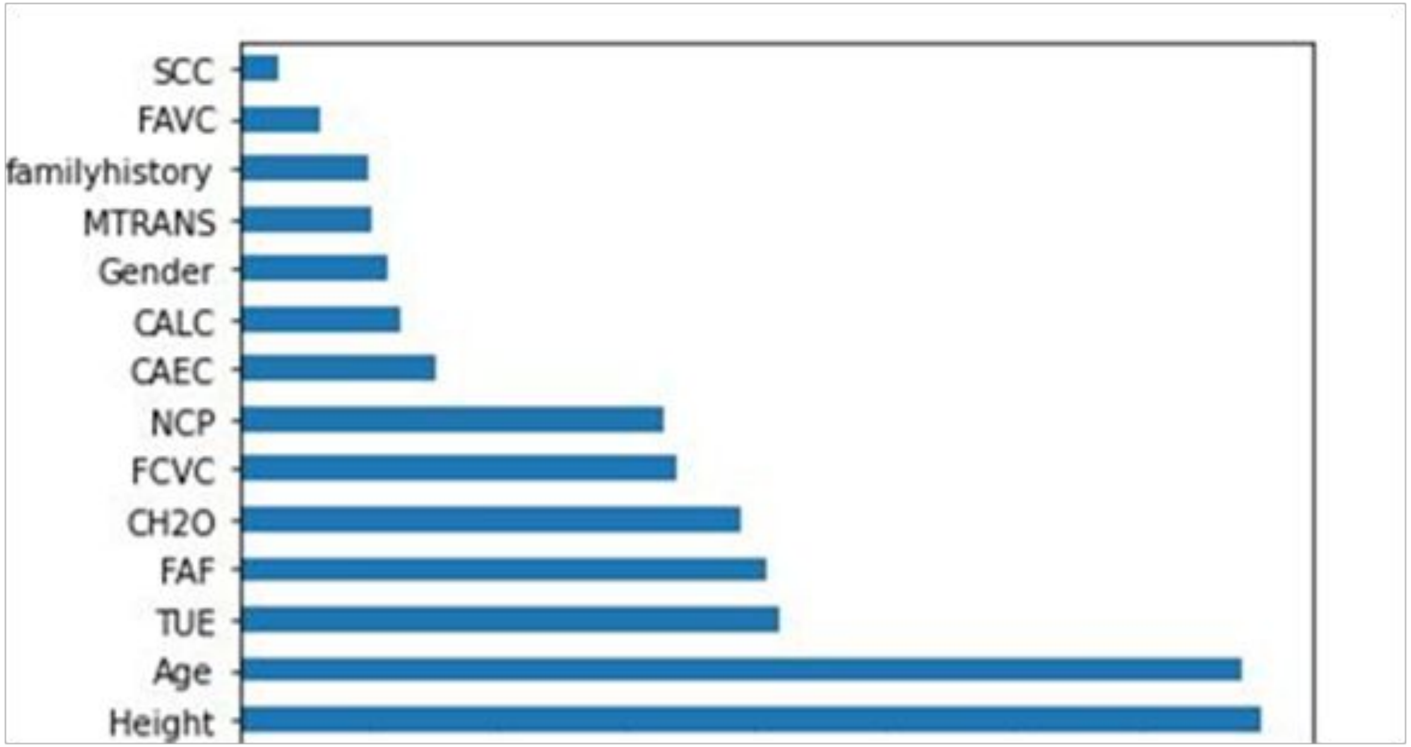
Results

- Confusion Matrix details high degree of accuracy
- Model Accuracy - 0.873

Predicted

Insufficient_Weight	[[79	6	0	0	0	1	0]	Actual
Normal_Weight	[3	76	5	0	0	3	6]	
Obesity_Type_I	[0	9	87	3	0	0	3]	
Obesity_Type_II	[0	3	1	84	0	0	0]	
Obesity_Type_III	[0	1	0	0	97	0	0]	
Overweight_Level_I	[1	11	3	1	0	70	2]	
Overweight_Level_II	[1	7	3	3	0	4	61]]	

Histogram shows the most informative features



- **Time spent on Tech Devices**
- **Physical Activity Frequency**
- **Daily Consumption of Water**
- **Consumption of Veggies**
- **Number of Main Meals**

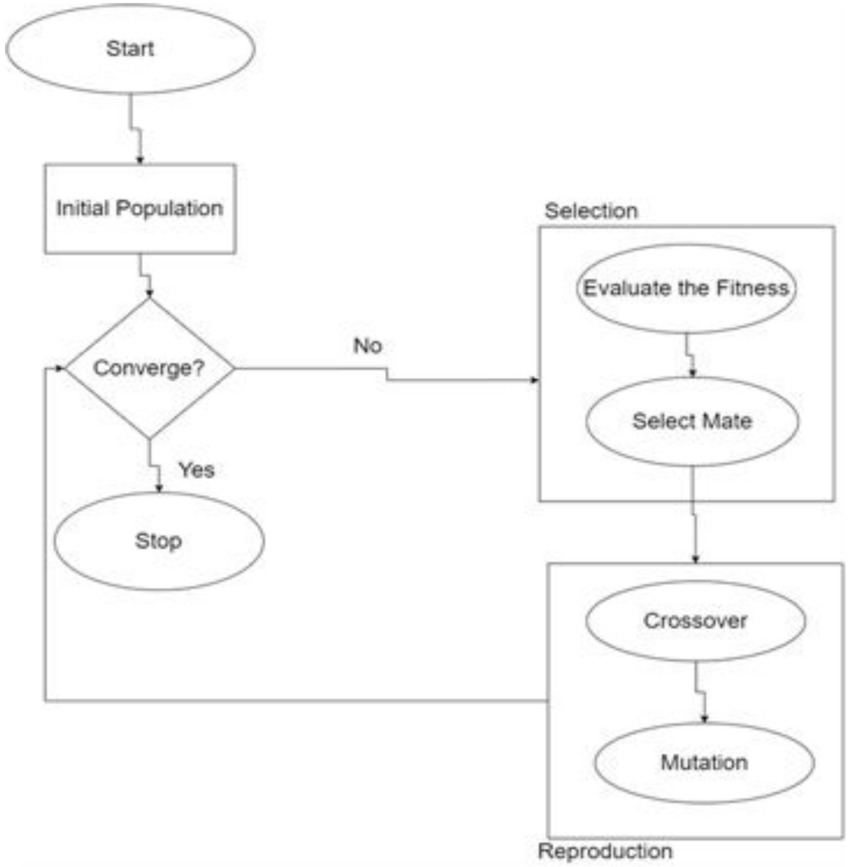
Evaluation

	precision	recall	f1-score
Insufficient_Weight	0.94	0.92	0.93
Normal_Weight	0.67	0.82	0.74
Obesity_Type_I	0.88	0.85	0.87
Obesity_Type_II	0.92	0.95	0.94
Obesity_Type_III	1.00	0.99	0.99
Overweight_Level_I	0.90	0.80	0.84
Overweight_Level_II	0.85	0.77	0.81
accuracy			0.87
macro avg	0.88	0.87	0.87
weighted avg	0.88	0.87	0.88

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{Recall} = \frac{TP}{TP + FN}$$

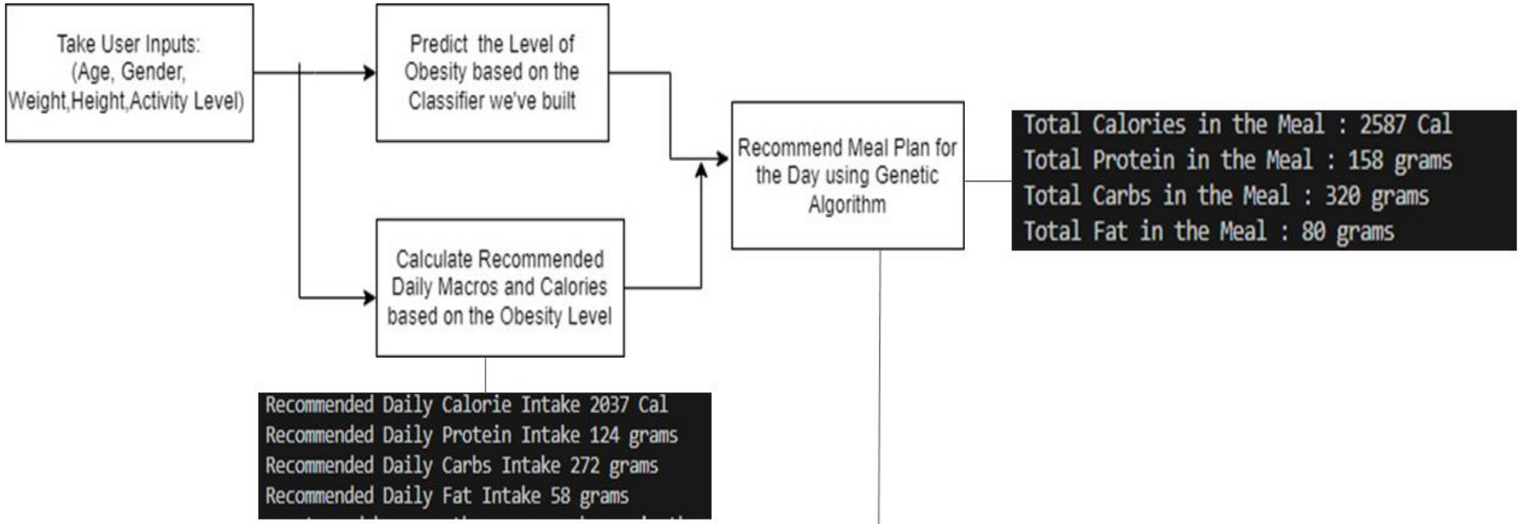
$$\text{F1 Score} = \frac{TP}{TP + \frac{1}{2}(FP + FN)}$$



Can we recommend meals based on the classification of Obesity Level ?

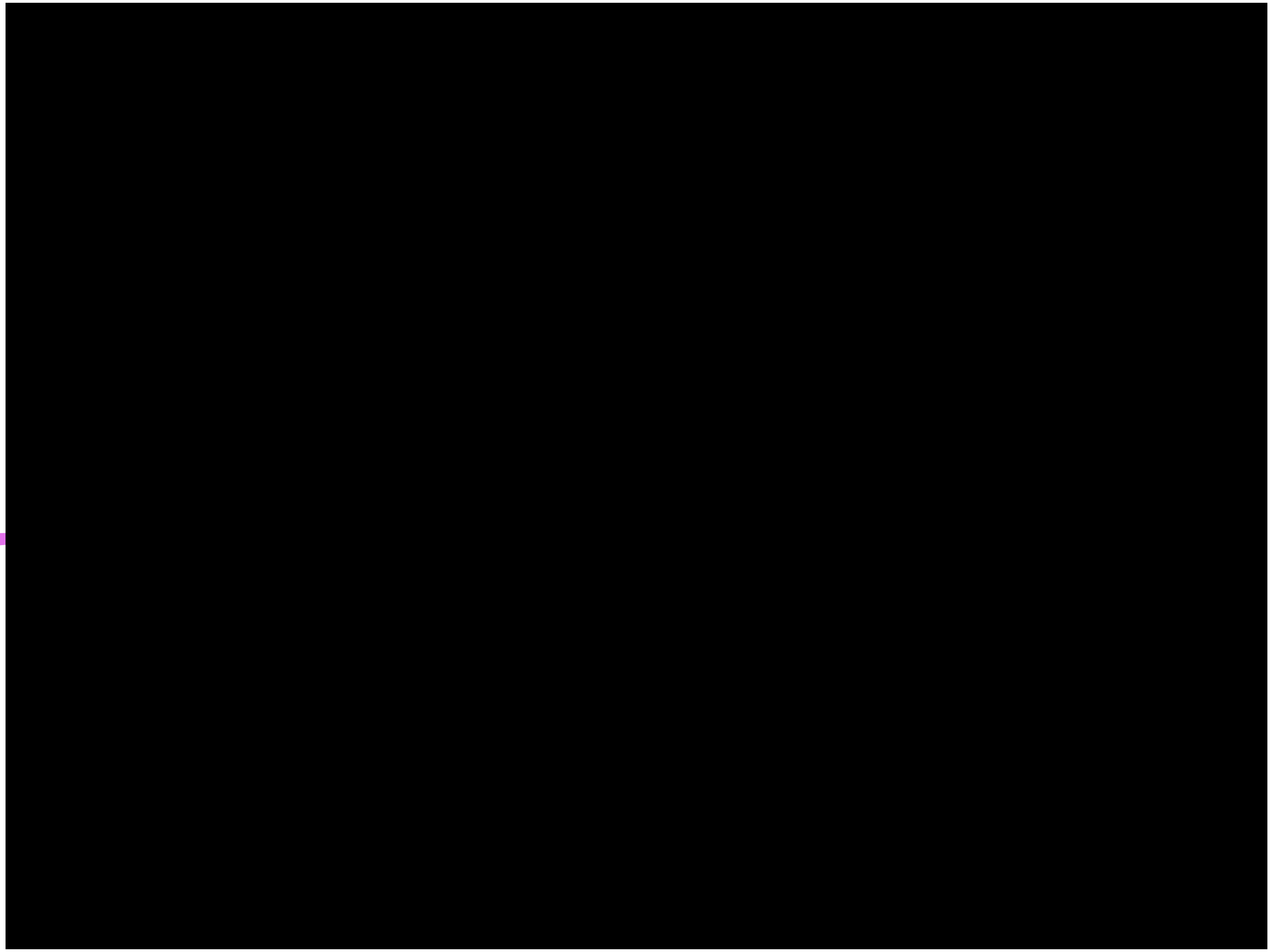
- **Obesity Level-> Know Your Goals-> Calculate Recommended Macros**
- **Aim: Equate Recommended Macros and Meal Macros**
- **Method: Genetic Algorithm**
- **Population: Recipe Dataset (Contains Recipe Names, Meal Types, Macros, Recipe links)**
- **Fitness Function: Parametrized by Macros(Carbs,Protein,Fat), Calories**
- **Convergence: Ends iteration and gives the output once the recommended Macros and Calories based on the Classification of Obesity Level and Meal Macros in the Recipe Dataset are close enough and do not alter.**

Methodology for the Meal Recommender



	Meal	Calories	Protein	Fat	Carbs	RecipeName	Recipe Links
0	Breakfast	330	5	9	59	No Sugar Added Apple Pie Overnight Oats	https://www.budgetbytes.com/2016/06/no-sugar-a...
1	Breakfast	233	19	14	7	Spinach, Mushroom, and Feta Crustless Quiche	http://www.budgetbytes.com/2011/11/spinach-mus...
2	Entrée	296	22	8	36	Chipotle Chicken Taco Salad	http://www.myrecipes.com/recipe/chipotle-chick...
3	Entrée	331	18	7	48	Lentil Dal with Hearty Greens	http://www.thekitchn.com/recipe-lentil-dal-wit...
4	Protein	293	29	6	33	Crispy Baked Chicken Breasts	https://aseasyasapplepie.com/crispy-baked-chic...
5	Protein	124	14	5	4	Skillet Steak with Onions and Mushrooms	http://www.skinnytaste.com/quick-skillet-steak...
6	Side	130	1	9	11	Creamy Cucumber and Tomato Salad	https://natashaskitchen.com/2010/08/16/creamy-...
7	Side	278	5	16	36	Grilled Corn and Poblano Salad	https://www.bonappetit.com/recipe/grilled-corn...
8	Soup	233	20	1	39	Chicken Spinach Soup	http://joyinmykitchen.blogspot.com/2010/02/chi...
9	Soup	339	25	5	47	Sausage and Kale Cassoulet	http://www.budgetbytes.com/2012/12/sausage-kal...

Demo Video for the Streamlit Application



Thank You!

Any Questions?