Work Log –Spring 2025 Semester:

Week 1 (Jan 13- Jan 19)

- **What pieces you need to address this semester? Do not be ambitious, be deep in everything you do.**
  Modeling
  Math part
  Machine Learning

Week 3 (Jan 7 – Feb 2)

- Research Question - *What are the most significant lifestyle factors contributing to different levels of obesity, and can we build a predictive model to classify individuals into obesity categories based on these factors?*
- *DS overview*
- *EDA Findings*
- *Modeling Results*
- *Insights*

Train the model to predict NObeyesdad based on selected features.

Use Random Forest to identify significant factors

Potential hypothesis:

Higher physical activity frequency (FAF) correlates with lower obesity levels.

1. Family history of being overweight significantly increases the likelihood of higher obesity levels.

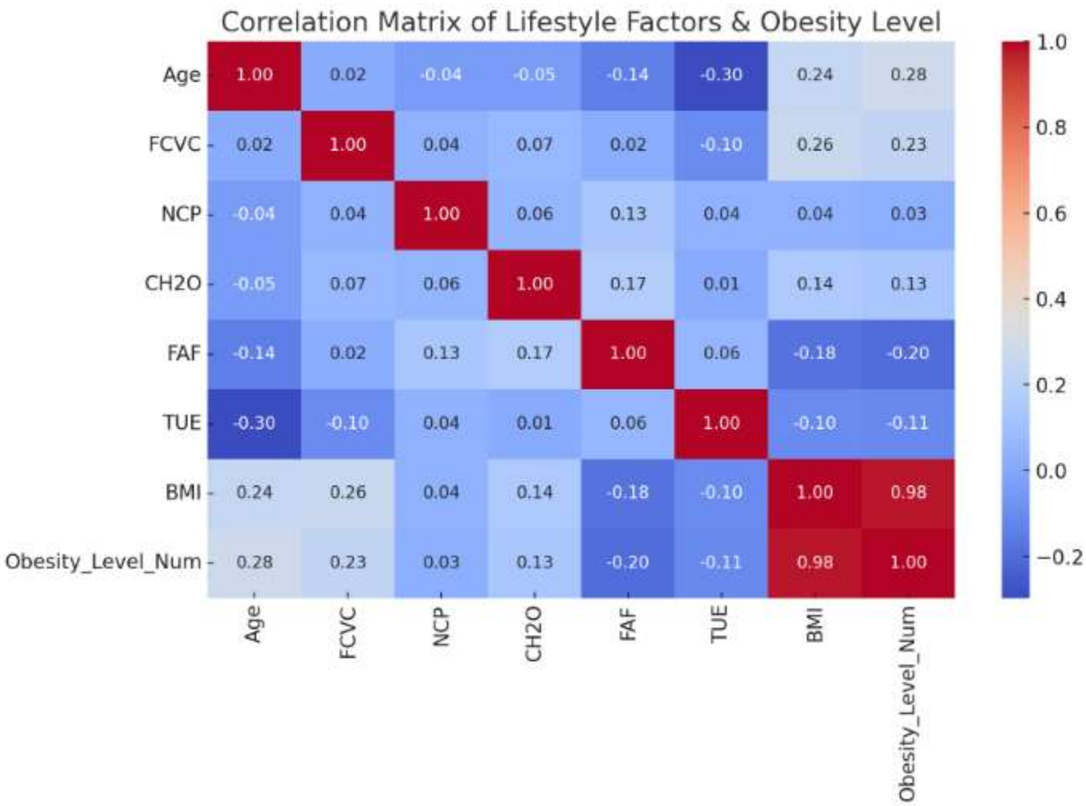
Week 4,5,6 (Feb 3 – Feb 24)

- **Dataset description**

Variable Name	Role	Type	Demographic	Description	Units	Missing Values
Gender	Feature	Categorical	Gender			no
Age	Feature	Continuous	Age			no
Height	Feature	Continuous				no
Weight	Feature	Continuous				no
family_history_with_overweight	Feature	Binary		Has a family member suffered or suffers from overweight?		no
FAVC	Feature	Binary		Do you eat high caloric food frequently?		no
FCVC	Feature	Integer		Do you usually eat vegetables in your meals?		no
NCP	Feature	Continuous		How many main meals do you have daily?		no
CAEC	Feature	Categorical		Do you eat any food between meals?		no
SMOKE	Feature	Binary		Do you smoke?		no

-

- Having columns with height and weight I was able to find BMI = weight(kg)/height(m^2)
- First thing I did - Correlation Analysis (Finding Relationships)
    - **Pearson correlation** for numerical variables (e.g., Age, Exercise, Water Intake).
    - **Chi-Square** for categorical variables (e.g., Mode of Transport, Smoking).
    - **Created a heatmap** to visualize which factors have the highest correlations with obesity.

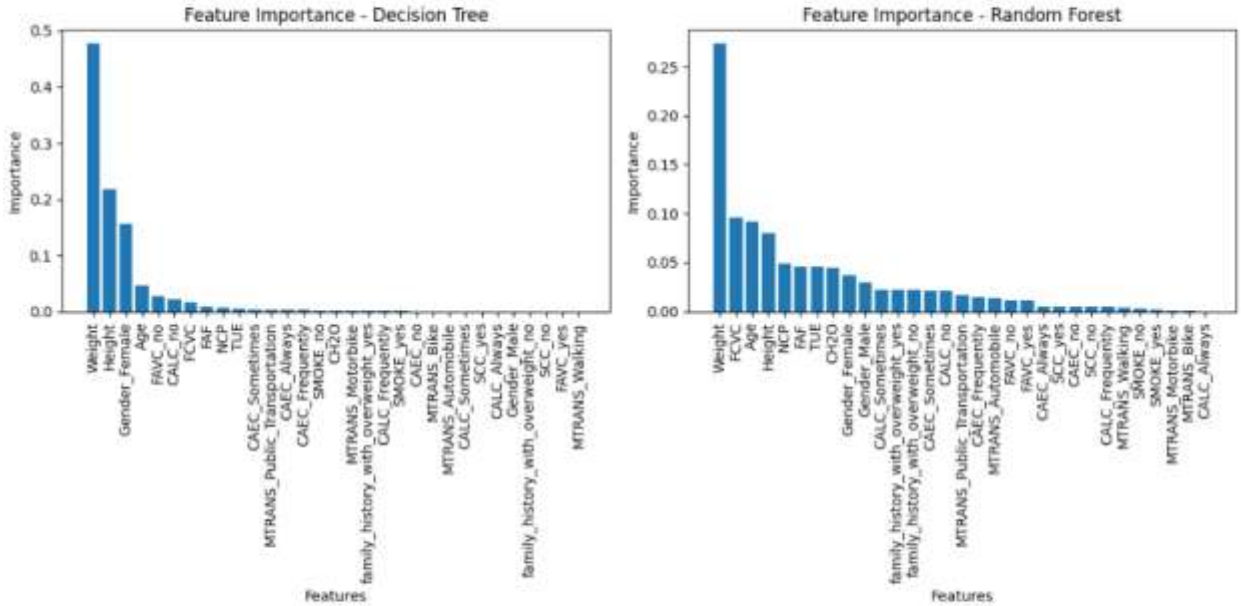


Summary:

- **0.98 (BMI)**: Very strong positive correlation, meaning BMI is almost perfectly related to obesity level.
- **0.28 (Age)**: Weak positive correlation, suggesting a slight trend where older individuals tend to have higher obesity levels.
- **0.23 (Vegetable Consumption - FCVC)**: Very weak positive correlation, but still a slight trend that people with higher obesity levels may eat more vegetables.
- **0.13 (Water Intake - CH2O):** Very weak positive correlation, indicating minimal impact.
- **-0.20 (Physical Activity - FAF):** Weak negative correlation, meaning that more physical activity is slightly related to lower obesity levels.
- **-0.11 (Time Using Technology - TUE):** Very weak negative correlation, indicating a very small association between screen time and lower obesity levels.

The second step is to Run Feature Importance Analysis to find the most influential factors

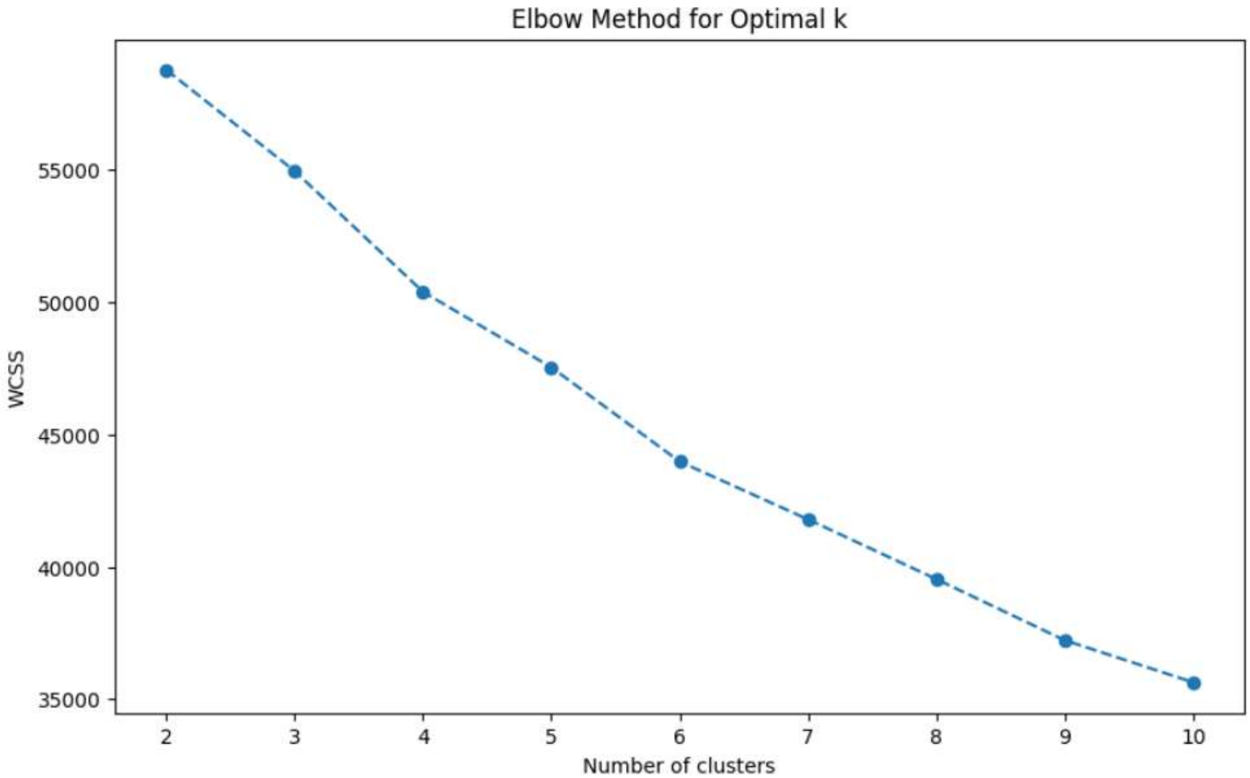
- Using Random Forest and Decision Tree
- Results:



Key observations:

Key Aspect	Decision Tree Insights	Random Forest Insights	Potential Actions
Most Important Feature	Weight is the most important feature.	Weight is the most important feature.	Focus on weight management interventions.
Physical Characteristics	Emphasizes Weight, Height, and Gender_Female.	Highlights Weight and Height, but less emphasis on Gender_Female.	Tailor interventions to gender and age.
Lifestyle Factors	CAEC (Consumption of Food Between Meals) is influential.	FCVC (Frequency of Consumption of Vegetables) is highly important.	Promote vegetable consumption; emphasize balanced diet and physical activity.
Age	Age is considered important.	Age is considered important.	Tailor strategies based on age, considering metabolism and lifestyle changes.
Family History	Low importance assigned to family_history_with_overweight.	Low importance assigned to family_history_with_overweight.	Investigate family history further; consider data quality issues.
Physical Activity and Screen Time	FAF (Physical Activity Frequency) and TUE (Time Using Electronic Devices) have negligible importance.	FAF and TUE are moderately important.	Encourage more physical activity and reduced screen time.

1. **Clustering**



Clustering Output Analysis:

- **Silhouette Score:** The Silhouette Score of 0.159 indicates poor cluster separation and significant overlap, suggesting that the number of clusters and features may not be optimal. (-1 to 1, should be closer to 1)
- **Cluster Breakdown:**
    - **Cluster 0:** Slightly older individuals with higher weight and moderate physical activity. Uses public transport more.
    - **Cluster 1:** Younger, heavier individuals with the highest obesity levels. Primarily use public transport.
    - **Cluster 2:** Older individuals, with moderate activity levels and higher alcohol consumption. Primarily use cars.
    - **Cluster 3:** Young females with lower weight, lower obesity levels, and moderate activity levels. Rely more on public transport.

Elbow Method Analysis:

- **Elbow Plot:** The "elbow" suggests that 4 to 6 clusters may be more appropriate, as the rate of decrease in WCSS diminishes after this point.

Recommendations:

1. **Re-evaluate Number of Clusters:** Try 4, 5, or 6 clusters.
2. **Feature Selection:** Focus on the most important features like weight, age, FCVC, and FAF to create more meaningful clusters.

Next Steps:

- Experiment with different clustering parameters and algorithms.
- Analyze the clusters in-depth for patterns contributing to obesity levels.

2. **Predictive Modeling**

Obesity Level (NObeyesdad):

- **This is the most direct and obvious prediction target.** As you have already started to do, you can build classification models to predict the obesity category (e.g., "Normal_Weight," "Overweight_Level_I," "Obesity_Type_II," etc.) based on the other features in the dataset.
- **The feature importance plots give you insights into which features will be most useful for this prediction.** Weight, height, age, gender, and vegetable consumption (FCVC) are likely to be strong predictors.

2. BMI Category:

- Since you've already calculated BMI, you could create your own BMI categories (e.g., "Underweight," "Normal Weight," "Overweight," "Obese") and use the other features to predict these categories.
- **This might be a simpler and more interpretable prediction task than directly predicting the NObeyesdad categories.**

Obesity Prediction Model – Summary

Model Overview

- Developed a **Random Forest Classifier** to predict obesity levels based on lifestyle factors.
- Achieved **99.29% accuracy**, with strong precision and recall across all obesity categories.
- Utilized **feature engineering** (computed BMI), **data preprocessing** (label encoding, scaling), and **model evaluation** (confusion matrix, feature importance).

Key Findings

- **Feature Importance:**
  - **Weight, BMI, and Physical Activity (FAF)** were the most influential predictors.
  - **Family history of obesity and food consumption (FCVC)** also had moderate importance.
- **Confusion Matrix Analysis:**
  - **Minimal misclassifications**, confirming the model's effectiveness.
  - The highest error occurred between neighboring obesity levels, but overall classification was highly accurate.

Mathematical Foundation

- **Random Forest** builds multiple decision trees and aggregates results using majority voting.
- **Feature importance** was computed based on how much each feature reduces impurity in splits.
- **Confusion matrix** helped evaluate classification performance by comparing actual vs. predicted labels.

Next Steps

- **Hyperparameter tuning** to further refine accuracy.
- **Deploying the model** via a user-friendly interface for real-time predictions.
- **Further exploration** of clustering to uncover obesity-related patterns.

User friendly interface to predict obesity levels. (Possibly – MICS 2025)

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