**MSBD566 – Predictive Modeling and Analysis**

Name: Sravan Phani Kumar Kopparthi

Date:10/22/2025

Assignment: Midterm Project

## Project Description

Obesity is a significant, prevalent, and costly chronic illness. It is characterized by an excess accumulation of body fat that impairs health, with a Body Mass Index (BMI) of 30 or higher.[2] The Region of the Americas exhibits the highest prevalence among all regions of the World Health Organization, with 62.5% of adults classified as overweight or obese (64.1% of men and 60.9% of women). The prevalence of obesity is estimated at 28% within the adult population (26% of men and 31% of women).[3] The condition is categorized into various classes based on BMI ranges: Class I (BMI 30–34.9), Class II (BMI 35–39.9), and Class III (Severe/Morbid, BMI 40 or greater). Obesity elevates the risk of developing serious diseases such as type 2 diabetes, cardiovascular diseases, specific cancers, and joint disorders. Additionally, it impacts quality of life and mental health. The aim is to predict the risk of obesity prior to the onset of severe conditions.

The data set provides estimates of obesity levels among individuals from Mexico, Peru, and Colombia, based on their eating habits and physical condition. It includes 17 attributes and 2111 records, with each record labeled by the class variable NObesity (Obesity Level). This allows data to be classified, such as Insufficient weight, Normal Weight, Overweight Level I, Overweight Level II, Obesity Type I, Obesity Type II, and Obesity Type III. About 77% of the data was synthetically generated using the Weka tool and the SMOTE filter, while the remaining 23% was collected directly from users via a web platform.  Using this dataset, we would like to build a classification model.

## Data Description:

Data for estimating obesity levels in individuals from Mexico, Peru, and Colombia, aged 14 to 61, with varied eating habits and physical conditions as noted by [1], was gathered through a web-based survey where anonymized users answered each question. The data was processed to extract 17 attributes and 2111 records following a balancing procedure. The attributes related to eating habits are: Frequent consumption of high-calorie 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 to the physical condition are: Calorie 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 were labeled, and the class variable NObeyesdad was created with the following values: Insufficient Weight, Normal Weight, Overweight Level I, Overweight Level II, Obesity Type I, Obesity Type II, and Obesity Type III, based on information from the WHO and Mexican Normativity. The data includes numerical and categorical data, making it suitable for analysis using classification, prediction, segmentation, and association algorithms. We want to explore the classification models with this data initially.

| **Variable Name** | **Role** | **Type** | **Demographic** | **Description** |
| --- | --- | --- | --- | --- |
| Gender | Feature | Categorical | Gender |  |
| Age | Feature | Continuous | Age |  |
| Height | Feature | Continuous |  |  |
| Weight | Feature | Continuous |  |  |
| family\_history\_with\_overweight | Feature | Binary |  | Has a family member suffered or suffers from overweight? |
| FAVC | Feature | Binary |  | Do you eat high caloric food frequently? |
| FCVC | Feature | Integer |  | Do you usually eat vegetables in your meals? |
| NCP | Feature | Continuous |  | How many main meals do you have daily? |
| CAEC | Feature | Categorical |  | Do you eat any food between meals? |
| SMOKE | Feature | Binary |  | Do you smoke? |
| CH2O | Feature | Continuous |  | How much water do you drink daily? |
| SCC | Feature | Binary |  | Do you monitor the calories you eat daily? |
| FAF | Feature | Continuous |  | How often do you have physical activity? |
| TUE | Feature | Integer |  | How much time do you use technological devices such as cell phone, videogames, television, computer and others? |
| CALC | Feature | Categorical |  | How often do you drink alcohol? |
| MTRANS | Feature | Categorical |  | Which transportation do you usually use? |
| NObeyesdad | Target | Categorical |  | Obesity level |

## Method and Analysis:

We don’t have any null values in the Dataset. As the initial step, we conducted a review of the dataset. We analyzed the data using the head, info(), and describe() functions. The dataset description is in Figure 1.

1. Describe

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Age** | **Height** | **Weight** | **FCVC** | **NCP** | **CH2O** | **FAF** | **TUE** |
| **count** | 2111.000 | 2111.000 | 2111.000 | 2111.000 | 2111.000 | 2111.000 | 2111.000 | 2111.000 |
| **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 |

Fig (1)

EDA

Gender Distribution: Class Distribution

A screen shot of a graph

AI-generated content may be incorrect.A graph of different colored bars

AI-generated content may be incorrect.

Fig (2) Gender Distribution Fig (3) Class Distribution

Correlation Heatmap: Mean Age per Obesity Class

A screenshot of a graph

AI-generated content may be incorrect. A graph of different colored rectangular bars

AI-generated content may be incorrect.

Fig (4) Correlation Heatmap Fig (5) Mean Age per Obesity Class

Methods: We chose the random forest model for its balance of accuracy, interpretability, robustness, and versatility for our first model. We used all the variables in the dataset to build the model with 200 trees, fully expanded, and random\_state=42. We have 0.957 accuracy. Precision ranges from 0.84 to 1.00 across classes, indicating the model's accuracy in identifying positive examples within each class. Recall is very high (mostly above 0.90), meaning the model effectively captures most true cases without missing many (low false negatives). F1-scores (harmonic mean of precision and recall) are all close to or above 0.90, confirming balanced performance for each class even under multi-class settings.

**Classification Report:**

precision recall f1-score support

0 1.00 0.93 0.96 54

1 0.84 0.97 0.90 58

2 0.97 0.97 0.97 70

3 0.98 0.98 0.98 60

4 1.00 0.98 0.99 65

5 0.96 0.90 0.93 58

6 0.97 0.97 0.97 58

accuracy 0.96 423

macro avg 0.96 0.96 0.96 423

weighted avg 0.96 0.96 0.96 423

Confusion Matrix: Very few misclassifications between classes (off-diagonal values are low). Most confusion occurs between classes 1 and 5, but at low counts, highlighting strong discriminating power. High diagonal counts show strong agreement between predicted and true classes.

[[50, 4, 0, 0, 0, 0, 0]

[0, 56, 0, 0, 0, 2, 0]

[0, 0, 68, 0, 0, 0, 2]

[0, 0, 1, 59, 0, 0, 0]

[0, 0, 0, 1, 64, 0, 0]

[0, 6, 0, 0, 0, 52, 0]

[0, 1, 1, 0, 0, 0, 56]]

## Evaluation:

* Random Forest model is highly effective for this multi-class classification problem, demonstrating both high accuracy and balanced class-wise performance.
* The confusion matrix and classification report suggest minor room for improvement in a few classes, potentially addressable by hyperparameter tuning or adding more training data.
* This makes Random Forest a strong and interpretable choice for multi-class obesity classification problems.

We also want to compare the performance of the other models we've selected: decision trees, KNN, and XGBoost. We chose the Decision Tree because it is easy to interpret and visualize, handles all data types, and trains quickly. KNN was selected for its simplicity and intuitiveness, and it works for any number of classes. XGBoost was chosen for its extremely high accuracy, ability to handle missing data, robustness with tabular data, and fast training due to boosting optimizations.

The following are the accuracy, precision, recall, and F1-score for the models. Precision, recall, and F1-score are calculated using the weighted average method.

Models Accuracy Comparison: Model Precision Comparison

A graph showing different colored bars

AI-generated content may be incorrect.A graph showing different colored bars

AI-generated content may be incorrect.

Fig.(6) Models Accuracy Comparison Fig.(7) Model Precision Comparison

Models Recall Comparison: Model F1 score Comparison:

A graph showing a number of different colored bars

AI-generated content may be incorrect.A graph showing different colored bars

AI-generated content may be incorrect.

Fig.(8) Models Recall Comparison Fig.(9) Model F1 score Comparison

We achieve the best results with Random Forest and XGBoost. Less performance with the KNN model.

Interpretation:

* Random Forest and XGBoost have the highest accuracy and F1 score, meaning they are both highly reliable at correctly classifying all categories in your dataset.
* Decision Tree scores are slightly lower but still strong—this model is more interpretable but is more prone to overfitting and has lower generalization than ensemble methods.
* KNN performs well but is outperformed by ensemble and boosting methods, showing it may not capture complex boundaries as effectively for your data.

Metric for all models:

* Accuracy: Ratio of true predictions to all predictions. High values (close to 1.0) show the model is rarely wrong.​
* Precision: Fraction of relevant instances among retrieved instances. High precision means very few false positives.
* Recall: Fraction of relevant instances that have been retrieved. High recall means very few false negatives.
* F1 Score: Harmonic mean of precision and recall. High value shows a good balance between not missing true cases and not having many false alarms.

Takeaway:

* Random Forest and XGBoost are the top options for your dataset, offering a good balance of accuracy, sensitivity, and reliability.
* Decision Tree is suitable if you need a simple, interpretable model but accept some loss in predictive accuracy.
* KNN is the simplest, but it is less robust.

Reference:

[1] Fabio Mendoza Palechor, Alexis de la Hoz Manotas, “Dataset for estimation of obesity levels based on eating habits and physical condition in individuals from Colombia, Peru and Mexico,” Data in Brief, Volume 25, 2019, 104344, ISSN 2352-3409, <https://doi.org/10.1016/j.dib.2019.104344>.

[2] <https://www.cdc.gov/obesity/adult-obesity-facts/?CDC_AAref_Val=https%3A%2F%2Fwww.cdc.gov%2Fobesity%2Fdata%2Fadult.html&utm_campaign=CHD_g-plans-review#:~:text=Many%20U.S.%20adults%20have%20obesity,BMI%20of%2040.0%20or%20higher>.

[3] <https://www.paho.org/en/topics/obesity-prevention>