Estimation of Obesity Levels Machine Learning Approach Based On Eating Habits and Physical Conditions

Naume Kizza Nabanjala
Department of Computer Science, COCIS
Makerere University
2024/HD05/21931U, 2400721931 Kampala, Uganda
kizza.naumenabanjala@students.mak.ac.ug

Jasper Okabo
Department of Computer Science, CoCIS.
Makerere University.
2024/HD05/21943U, 2400721943 Kampala, Uganda
jasper.okabo@students.mak.ac.ug

***Abstract*—Obesity by definition, is the excessive accumulation of adipose tissue or build-up of too much fat in the body that leads to various health risks. This study aimed to predict obesity levels using a classification-based machine-learning approach focusing on factors such as physical activity and nutritional habits.**

Methods: The research employed an observational design where data was collected via a web-based survey from a public dataset that included attributes such as gender, age, height, weight, family history of being overweight, dietary patterns, physical activity frequency, and more. Emphasis was placed on eating habits and physical activity levels. Data preprocessing involved filling gaps, removing duplicates, handling missing data, assessing unique data, and encoding categorical values (e.g., gender, transportation models etc). The data was then divided into training and testing data to ensure effective model evaluation. Logistic regression (LR) and random forest (RF) were applied to classify obesity levels. The performance of these different models was measured using metrics such as accuracy, recall, precision, F1-score, and precision-recall curve. Feature selection further improved the performance of LR and RF models. The RF model showed the best performance across most metrics with a 97% accuracy, followed by LR.

This study contributes to the understanding of obesity classification using machine learning by incorporating both physical activity and nutritional habits. The RF model proved the most robust and reliable for predicting obesity levels. The findings underline the importance of addressing both physical activity and diet in efforts to combat the growing obesity epidemic while showcasing the potential of machine-learning techniques to analyze health-related data effectively.

***Index Terms*—Obesity, Classification, Machine Learning**

I. Introduction

Obesity is one of the most prevalent health issues in the world today, affecting individuals of all ages and socioeconomic backgrounds. The World Health Organization (WHO) reports that obesity has tripled globally since 1975, with over 890 million adults classified as obese in 2022 alone. The fundamental cause of obesity or overweight is an energy imbalance between calories consumed and calories expended.

Obesity means having too much body fat, hence raising the chances of getting serious health problems like heart disease, diabetes, high blood pressure, and some types of cancer.

If there was a way to predict the occurrence of obesity based on the factors that influence or cause obesity, people would then know what to do to prevent or control it and with this solution, better health would be achieved in the world. Accuracy in predicting this still has a bit of a challenge for both health institutes and individuals. To address this challenge, this research aims to develop a prediction model for obesity categories based on physical activity and eating habits data using Logistic regression and Random Forest. By utilizing a public dataset for obesity prediction, various attributes including demographics, living habits, and individual health indicators are collected, with the aim of supporting predictions of obesity prevalence rates using 6 input variables (predictors) and 1 output variable (prediction).

II. Background and motivation

A. Obesity

Obesity is a global health issue with rising prevalence, marked by excessive fat accumulation that poses health risks. It is typically measured by Body Mass Index (BMI), where a BMI of 30 or higher indicates obesity, whereas when the BMI is between 25 and 30, the person is considered overweight [2] Contributing factors include poor eating habits, lack of physical activity, and genetics, which elevate the risk of chronic diseases like type 2 diabetes, heart disease, stroke, and cancer.

Physical activity has an important role in managing body weight and preventing obesity. Various studies have shown that increasing physical activity can help with weight loss and improve overall health. However, challenges in measuring and predicting an individual's physical activity levels often become obstacles in efforts to prevent and manage obesity (Wulandari et al., 2024).

In recent years, developments in technology and data science have opened new opportunities to analyze and predict obesity categories based on physical activity (Rahmawati et al., 2024). Machine learning algorithms, which are part of artificial intelligence, have shown great potential in processing large amounts of data and discovering patterns that cannot be identified by conventional methods. By using this algorithm, we can develop a more accurate and effective prediction model to categorize obesity levels based on individual physical activity (Wildan et al., 2024).

B. Classification

Classification is a task where the model predicts the category or class of given data points based on the input features. For example in our dataset context, it can determine whether Males being the highest number of obese individuals is ”True” or ”Not True”. Classification in machine learning also involves training a model to assign input data to specific categories or classes based on its features. It

C. Machine Learning

Machine learning, a branch of computer science, focuses on enabling computers to handle complex tasks without being directly programmed for every step. Instead, computers learn and improve at tasks over time through experience. This field typically deals with large, high-dimensional datasets and prioritizes accurate predictions over traditional hypothesis-driven analysis.

It usually works with large, complex datasets and focuses on making accurate predictions rather than testing specific ideas. As Big Data grows for example the ever-increasing huge amounts of information created by corporate financial companies, machine learning methods work well, even when older statistical methods face challenges.

D. Problem Statement

The rising prevalence of obesity worldwide poses significant health risks, including diabetes, cardiovascular disease, and reduced quality of life. Early detection of individuals at high risk of obesity is crucial for implementing preventative measures and promoting healthier lifestyles. However, accurately predicting obesity risk based on lifestyle and physical condition data remains challenging, particularly in diverse populations with varied dietary and exercise habits.

This project aims to develop a predictive model using the UCI dataset on obesity levels to assess individuals’ obesity risk based on eating habits, physical activity, and demographic information. By improving obesity risk prediction, this project seeks to support targeted interventions and enhance public health outcomes.

III. Literature Review

Over the recent past years, researchers have carried out a few predicting models for numerous obesity datasets and their results have been promising especially yielding some great outcomes in hospitals. A few of them are:

A. Literature Review

Dugan et al. (**Dugan et al., 2015**) did an excellent job of predicting obesity in children after age two. They used six models to test for their study. Their models are Random Tree, Random Forest, ID3, J48, Naïve Bayes, and Bayes Net trained on CHICA which are clinical decision support system. They got the best performance from the model ID3, which was highly accurate at 85% and sensitive at nearly 90%

Research by **Admojo & Rismayanti (2024)** utilized machine learning, specifically the Decision Tree algorithm, to analyze obesity determinants in Mexico, Peru, and Colombia. The study, involving 2,111 individuals and 5-fold cross-validation, achieved an impressive accuracy of **95%**, highlighting the importance of eating habits and physical activity as predictors of obesity. However, it acknowledged limitations due to self-reported data and called for a more diverse dataset. Their comparative studies showed that the Decision Tree algorithm outperformed K-Nearest Neighbor and Naïve Bayes with an accuracy of **93.6%**, identifying key risk factors such as body weight, gender, age, family history of obesity, and high-calorie food consumption. The model’s interpretability adds to its value.

Predicting Obesity in the General Population: A Machine Learning Approach by **Wang et al. (2020)**: The authors used a variety of machine learning models, including Logistic Regression, Decision Trees, and Random Forests, to predict obesity in a population sample. They used data from a health survey that included demographic information, physical activity levels, and dietary habits. Their key finding was that Random Forest outperformed other models, achieving an accuracy of 82%. Physical activity and diet were the strongest predictors of obesity.

Jindal et al. (**Jindal et al., 2018**) worked on predicting obesity using ensemble machine-learning approaches. Their predicted value of obesity was 89.68% accurate.

Additionally, Nasser & Abu-Naser (2023) found that the Random Forest model achieved an accuracy of **97.23%** in predicting obesity and cardiovascular disease risks, demonstrating its effectiveness with complex health data.

B. Research Gaps in the Literature

Limitations of Self-Reported Data: This type of data can be prone to inaccuracies because individuals may not always remember details correctly, exaggerate or downplay certain behaviors, or may have biases in reporting their habits. As a result, relying solely on self-reported data can lead to potential errors in the analysis and conclusions drawn from the study.

Limited exploration of hybrid models: Many researchers investigate individual algorithms but little research is carried out on hybrid models that can combine the strengths of different vast techniques for enhanced performance.

Call for a More Diverse Dataset: The dataset might not encompass a wide range of variables or demographic groups. A diverse dataset would include a broader spectrum of participants with varying characteristics (e.g., age, gender, socioeconomic status, geographic location) and additional relevant factors (e.g., genetic predispositions, psychological aspects). This diversity is important because it can help ensure that the findings are more generalizable and applicable to different populations, improving the robustness of the conclusions regarding obesity predictors.

C. Contributions carried out

This project aims to develop a predictive model using the UCI dataset on obesity levels to assess individuals' obesity risk based on eating habits, physical activity, and demographic information. By improving obesity risk prediction, this project seeks to support targeted interventions and enhance public health outcomes.

With the use of the ICU Estimating Obesity Level dataset, it helps solve the issue of using a more diverse dataset as this dataset includes more characteristics like gender, age, psychological aspects like smoking, and utilization of internet.

IV. Methodology

The focus for this research is to be able to predict health occurances that are associated with obesity using the factors provided by the dataset. This helps the health centers to be able to know the outcomes that come with the factors in the dataset and plan for them accordingly while letting the people know of the preventive measures to take so that they can avoid getting obese.

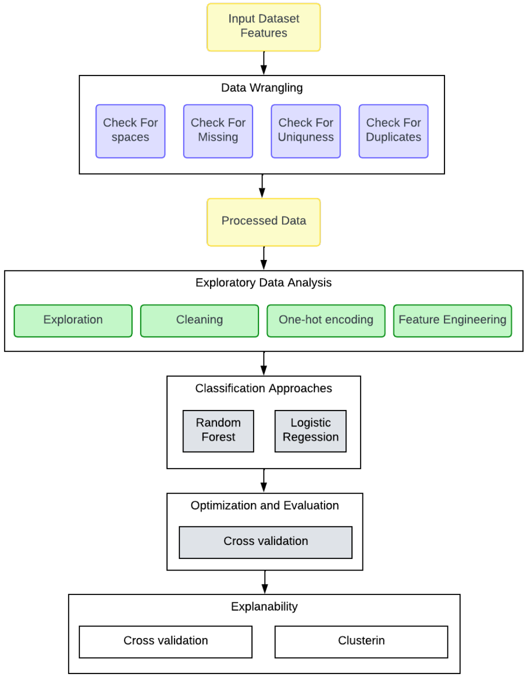


Fig. 1. Methodology followed

A. Dataset Description

The dataset used for the study is from a UCI repository. It consists of data collected from individuals in Mexico, Peru, and Colombia, focusing on their eating habits and physical condition to estimate obesity levels. The dataset contains 2,111 records with 17 attributes, including the target variable which is the ”NObeyesdad” (Obesity Level), which categorizes the data into seven classes: Insufficient Weight, Normal Weight, Overweight Level I, Overweight Level II, Obesity Type I, Obesity Type II, and Obesity Type III.

How the data was collected: It was collected through a survey, in which 16 questions related to the interviewees' dietary habits and physical condition were applied. These question were then considered as the basis factors that affect the obesity level. We then use them to get some performance metrics to determine the leading factor. These questions/factors include:

Gender: categorical variable that shows the biological sex of the individual (male or female).

Age: numerical variable that shows the individual's age in years.

Height: numerical variable that shows the individuals' height in meters.

Weight: numerical variable that shows the individuals' weight in kilograms.

Family history of overweight: categorical variable that shows if the individual has a family member who is overweight or obese (yes or no).

Frequently consumed high-calorie food (FAVC): categorical variable that shows if the individual often eats high-calorie food (yes or no).

Frequency of consumption of vegetables (FCVC): ordinal variable that shows how often the individual eats vegetables (1 = never, 2 = sometimes, 3 = always).

Number of main meals (NCP): ordinal variable that shows how many main meals the individual has daily (1 = between 1 and 2, 2 = three, 3 = more than three, 4 = no answer).

Consumption of food between meals (CAEC): ordinal variable that shows how often the individual eats food between meals (1 = no, 2 = sometimes, 3 = frequently, 4 = always).

SMOKE: categorical variable that shows whether the individual smokes or not (yes or no).

Consumption of water daily (CH2O): ordinal variable that shows how much water the individual drinks daily (1 = less than a liter, 2 = between 1 and 2 L, 3 = more than 2 L).

Monitor calorie intake (SCC): categorical variable that shows if the individual keeps track of their caloric intake (yes or no).

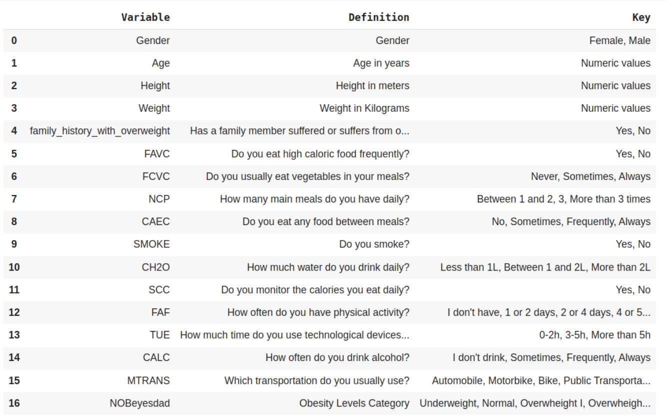
Frequency of physical activity (FAF): ordinal variable that shows how often the individual does physical activity (1 = never, 2 = once or twice a week, 3 = two or three times a week, 4 = four or five times a week).

Time using electronic devices (TUE): ordinal variable that shows how long the individual uses electronic devices (0 = none, 1 = less than an hour, 2 = between one and three hours, 3 = more than three hours).

Consumption of alcohol (CALC): ordinal variable that shows how often the individual drinks alcohol (1 = no, 2 = sometimes, 3 = frequently, 4 = always).

Type of transportation used (MTRANS): categorical variable that shows what kind of transportation the individual uses (automobile, motorbike, bike, public transportation, walking).

Level of obesity according to body mass index (NObesity): ordinal variable that shows the obesity level of the individual according to their BMI (insufficient weight normal weight, overweight level I, overweight level II, obesity type I, obesity type II, obesity type III).



	Variable	Definition	Key
0	Gender	Gender	Female, Male
1	Age	Age in years	Numeric values
2	Height	Height in meters	Numeric values
3	Weight	Weight in Kilograms	Numeric values
4	family_history_with_overweight	Has a family member suffered or suffers from o...	Yes, No
5	FAVC	Do you eat high caloric food frequently?	Yes, No
6	FCVC	Do you usually eat vegetables in your meals?	Never, Sometimes, Always
7	NCP	How many main meals do you have daily?	Between 1 and 2, 3, More than 3 times
8	CAEC	Do you eat any food between meals?	No, Sometimes, Frequently, Always
9	SMOKE	Do you smoke?	Yes, No
10	CH2O	How much water do you drink daily?	Less than 1L, Between 1 and 2L, More than 2L
11	SCC	Do you monitor the calories you eat daily?	Yes, No
12	FAF	How often do you have physical activity?	I don't have, 1 or 2 days, 2 or 4 days, 4 or 5...
13	TUE	How much time do you use technological devices...	0-2h, 3-5h, More than 5h
14	CALC	How often do you drink alcohol?	I don't drink, Sometimes, Frequently, Always
15	MTRANS	Which transportation do you usually use?	Automobile, Motorbike, Bike, Public Transporta...
16	NOBeyesdad	Obesity Levels Category	Underweight, Normal, Overwheight I, Overwheigh...

Fig. 2. Dataset attributes

B. Data Preparation and Exploratory Data Analysis

The dataset contains 2,111 records with 17 attributes, including the target variable which is the ”NObeyesdad” (Obesity Level), which categorizes the data into seven classes: Insufficient Weight, Normal Weight, Overweight Level I, Overweight Level II, Obesity Type I, Obesity Type II, and Obesity Type III. the aim for this is to

- Explore how different variables, such as physical activity, and calorie intake, are correlated with obesity levels to better understand risk patterns.
- Identify the Key Factors Contributing to Obesity i.e Identify the lifestyle, physical, and health factors (e.g., age, diet, physical activity) that have the most significant impact on obesity risk.
- Understand the distribution of key variables like age, weight and physical activity levels to see how they relate to obesity across different populations.
- Identify any outliers or data anomalies that could skew the analysis or provide insights into unusual cases of obesity risk.

The Exploratory Data Analysis (EDA) followed three main steps:

- Establishing questions or inquiries to examine the assumptions about the dataset and what is needed from it.
- Data Wrangling: For the data wrangling, three steps were followed:
  - Check for white spaces.
  - check for unique columns
  - check for duplicates
  - Data Cleaning: This process involved identifying and rectifying inaccuracies, missing values, and inconsistencies in the data. Three main steps were followed:
    - ∗ **Identify the Problem:** It was important to carefully examine the data to uncover issues such as missing values, outliers, and incorrect data types, among others.
    - ∗ **Develop Solutions in Code:** Code was written to resolve the identified issues. This included tasks like adding an ID column, to create uniqueness.
    - ∗ **Validate the Fixes:** After running the code, it was crucial to verify that the problems were resolved. This step involved reviewing a sample of the data or conducting statistical checks to ensure the data was clean.
- Data Visualization for the comparisons in the attributes.

Finally, the insights from the dataset were communicated and moved on to the next step of model development.

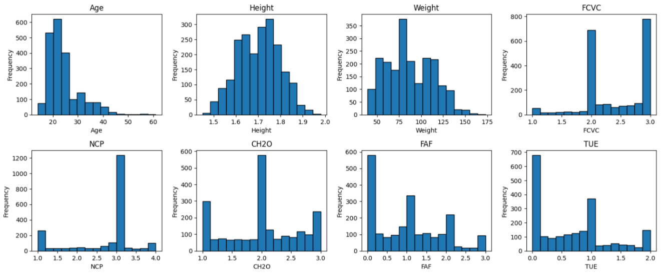
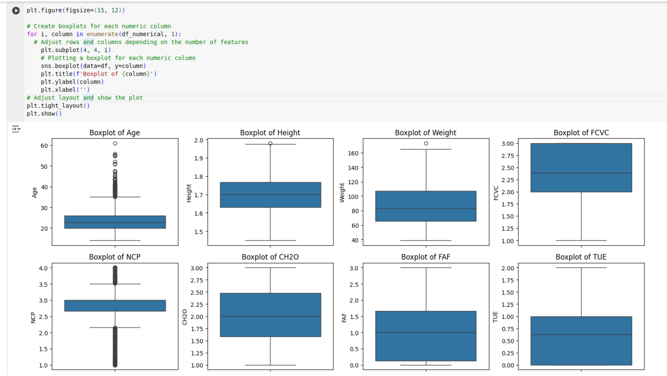


Fig. 3. Plotting graphs to relate to target value



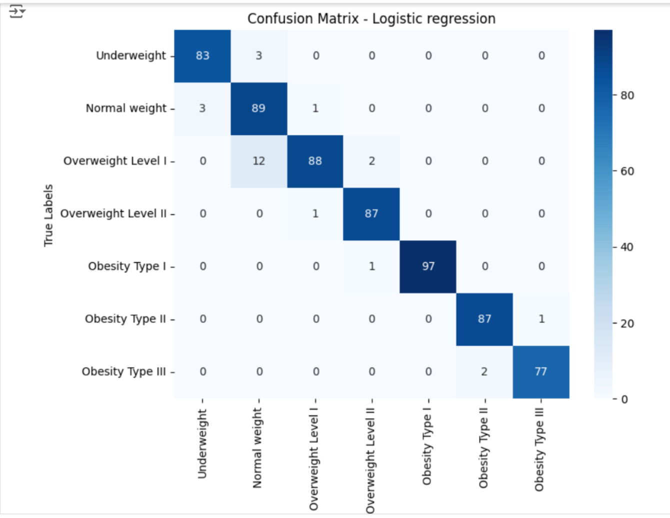
```
plt.figure(figsize=(15, 12))

# Create boxplots for each numeric column
for i, column in enumerate(df_numerical, 1):
  # Adjust rows and columns depending on the number of features
    plt.subplot(4, 4, i)
    # Plotting a boxplot for each numeric column
    sns.boxplot(data=df, y=column)
    plt.title(f'Boxplot of {column}')
    plt.ylabel(column)
    plt.xlabel('')
# Adjust layout and show the plot
plt.tight_layout()
plt.show()
```

Fig. 4. Box plot graphs to determine outliers

C. ML model selection and optimization

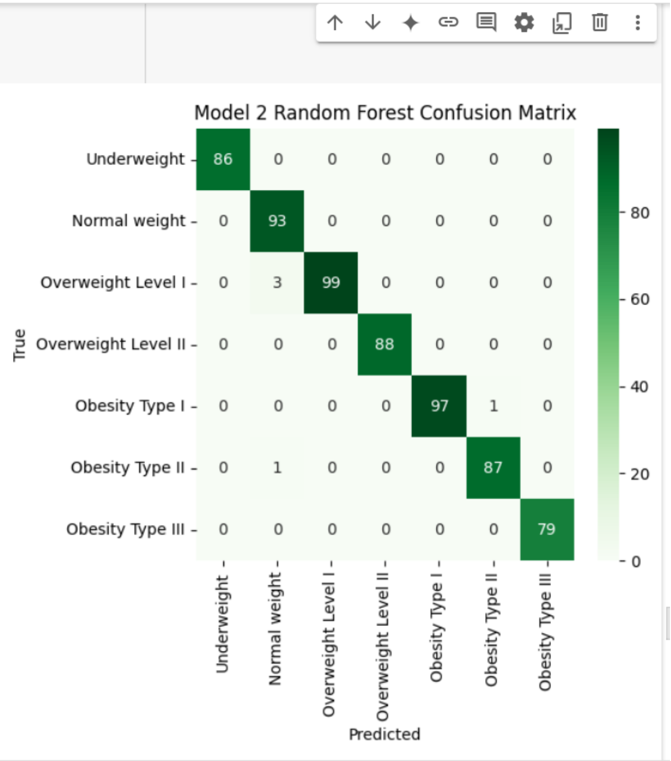
Logistic Regression



	Underweight	Normal weight	Overweight Level I	Overweight Level II	Obesity Type I	Obesity Type II	Obesity Type III
Underweight	83	3	0	0	0	0	0
Normal weight	3	89	1	0	0	0	0
Overweight Level I	0	12	88	2	0	0	0
Overweight Level II	0	0	1	87	0	0	0
Obesity Type I	0	0	0	1	97	0	0
Obesity Type II	0	0	0	0	0	87	1
Obesity Type III	0	0	0	0	0	2	77

Fig. 5. Confusion matrix for logistic regression

Random Forest



	Underweight	Normal weight	Overweight Level I	Overweight Level II	Obesity Type I	Obesity Type II	Obesity Type III
Underweight	86	0	0	0	0	0	0
Normal weight	0	93	0	0	0	0	0
Overweight Level I	0	3	99	0	0	0	0
Overweight Level II	0	0	0	88	0	0	0
Obesity Type I	0	0	0	0	97	1	0
Obesity Type II	0	1	0	0	0	87	0
Obesity Type III	0	0	0	0	0	0	79

Fig. 6. Confusion matrix for random forest

D. ML model selection Accountability

AI accountability refers to ensuring that an AI system operates correctly and responsibly throughout its entire lifecycle. An important part of this is being able to understand what the AI is doing and why and how it makes certain decisions. Within the AI field, this concept is widely recognized as model explainability.

In this project, SHAP (SHapley Additive exPlanations) was chosen as the explainable AI (XAI) technique. SHAP is a popular method that helps explain how each feature influences a model's predictions. It calculates the contribution of each feature to the final result using Shapley values from cooperative game theory.

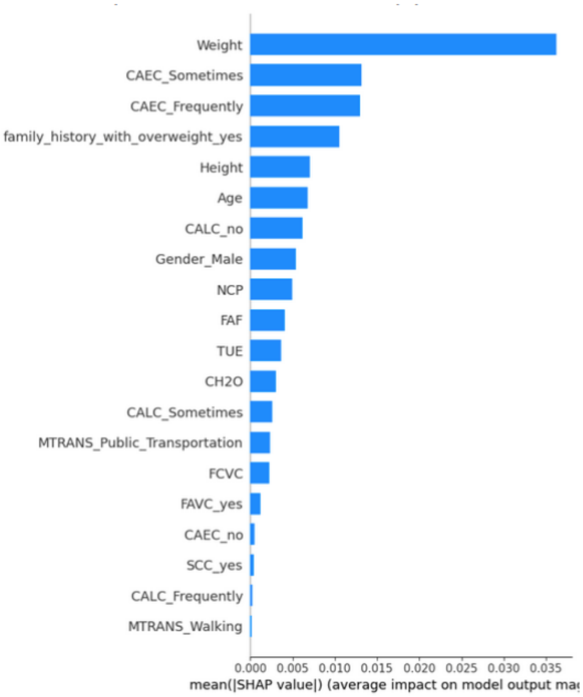


Fig. 7. Explainability using SHAP

Key Observations

Weight is the most important feature in predicting obesity levels, which makes sense as weight is a direct indicator of obesity.

CAEC (Consumption of High-Calorie Food) ”Sometimes” and ”Frequently” categories have high importance, which indicates that eating habits significantly influence obesity levels.

Family History with Overweight is another major factor in predicting obesity, as shown by the high contribution of this feature.

Height and Age: also play important roles. Height may relate to BMI calculation, and age can influence metabolism and lifestyle habits.

CALC (Consumption of Alcohol): The ”no” category has notable importance, suggesting that alcohol consumption patterns can correlate with obesity levels.

Other Features:

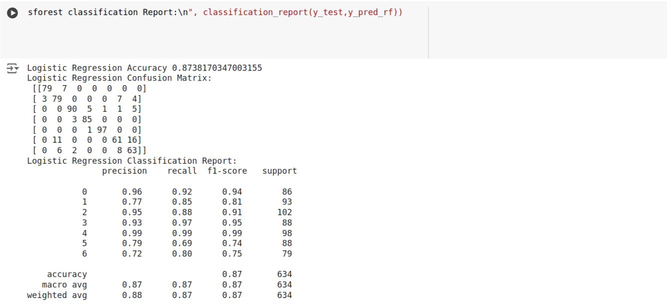
Features like **Gender (Male)**, **NCP (Number of Main Meals per Day)**, **FAF (Frequency of Physical Activity)**, and **TUE (Time Using Technology)** have moderate importance.

Lifestyle-related factors (like physical activity and screen time) are less influential than direct indicators like weight or eating habits.

Least Important Features:

Features such as **MTRANS (Means of Transportation - Walking)** and **CALC (Frequently)** have minimal impact on predictions.

E. Results and discussion



```
sforest classification Report:\n", classification_report(y_test,y_pred_rf))
```

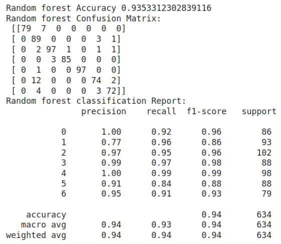
```
Logistic Regression Accuracy 0.8738170347003155
Logistic Regression Confusion Matrix:
 [[79  7  0  0  0  0  0]
 [ 3 79  0  0  0  7  4]
 [ 0  0 90  5  1  1  5]
 [ 0  0  3 85  0  0  0]
 [ 0  0  0  1 97  0  0]
 [ 0 11  0  0  0 61 16]
 [ 0  6  2  0  0  8 63]]
Logistic Regression Classification Report:
               precision    recall  f1-score   support

           0       0.96      0.92      0.94        86
           1       0.77      0.85      0.81        93
           2       0.95      0.88      0.91       102
           3       0.93      0.97      0.95        88
           4       0.99      0.99      0.99        98
           5       0.79      0.69      0.74        88
           6       0.72      0.80      0.75        79

    accuracy                           0.87       634
   macro avg       0.87      0.87      0.87       634
weighted avg       0.88      0.87      0.87       634
```

Fig. 8. Results from Logistic regression

As seen in the above representation Fig. 7, we managed to achieve the following results from precision, recall and f1-score under logistic regression.



```
Random forest Accuracy 0.9353312302839116
Random forest Confusion Matrix:
 [[79  7  0  0  0  0  0]
 [ 0 89  0  0  0  3  1]
 [ 0  2 97  1  0  1  1]
 [ 0  0  3 85  0  0  0]
 [ 0  1  0  0 97  0  0]
 [ 0 12  0  0  0 74  2]
 [ 0  4  0  0  0  3 72]]
Random forest classification Report:
               precision    recall  f1-score   support

           0       1.00      0.92      0.96        86
           1       0.77      0.96      0.86        93
           2       0.97      0.95      0.96       102
           3       0.99      0.97      0.98        88
           4       1.00      0.99      0.99        98
           5       0.91      0.84      0.88        88
           6       0.95      0.91      0.93        79

    accuracy                           0.94       634
   macro avg       0.94      0.93      0.94       634
weighted avg       0.94      0.94      0.94       634
```

Fig. 9. Results for random forest

As seen in the Fig 8 above, we managed to achieve the following results from precision, recall, and f1-score under Random Forest

Below is a table for the results. As seen in the representation below, we managed to achieve the following results from precision, recall and f1-score under **Random Forest**. The Random Forest model was able to better predict obesity levels based on the input features. However, further tuning or additional features could potentially improve the model performance. The model's application could extend to other related use cases in public health for obesity prediction. The Random Forest model also achieved a higher F1 score across most obesity categories, demonstrating its robustness in classification tasks.

Model	**Precision**	**Recall**	**F1-Score**
Logistic Regression	0.87	0.87	0.87
Random Forest	0.94	0.97	0.97

Conclusions and Future work

The Random Forest model was able to better predict obesity levels based on the input features. However, further tuning or additional features could potentially improve the model performance. The model's application could extend to other related use cases in public health for obesity prediction. The Random Forest model also achieved a higher F1 score across most obesity categories, demonstrating its robustness in classification tasks. The model successfully predicted obesity levels with an accuracy of 94% which was for the random forest.

Key lifestyle factors like fast food consumption, physical activity, and transportation methods significantly impact obesity.

Our Future Work is to Incorporate additional external data (e.g., socioeconomic status), Improve model performance through hyperparameter tuning and feature selection. We also plan to explore the use of more advanced algorithms like Neural Networks.

Dataset and Python source code

Final Python source code: Colab Notebook

Dataset used: Dataset

Term paper powerpoint slides: Powerpoint slides

References

[1] World Health Organization (WHO), Obesity and overweight, 2021. URL: https://www.who.int/news-room/fact-sheets/detail/obesity-and-overweight.

[2] J. H. Rosen, "Is Obesity A Disease or A Behavior Abnormality? Did the AMA Get It Right?". Missouri medicine, 111(2014): 104–108.

[3] M. Blüher, "Obesity: global epidemiology and pathogenesis". Nature Reviews Endocrinology, 15(2019): 288–298. doi: 10.1038/s41574-019-0176-8..

[4] World Health Organization (WHO), Body mass index–BMI, 2020. URL: https://www.euro.who.int/en/health-topics/disease-prevention/nutrition/a-healthy-lifestyle/bodymass-index-bmi.

[5] T. Bhurosy, R. Jeewon, "Overweight and obesity epidemic in developing countries: a problem with diet, physical activity, or socioeconomic status?". The Scientific World Journal, (2014). doi:10.1155/2014/964236.