Final Project Report

Title: Obesity Prediction Based on Eating Habits and Physical Activities.

List of Members:

- Nikhil Makkena(11435993)
- Manish Reddy Radha Reddy(11518946)
- Chathurya Thimmapuram(11479590)
- Vaishnavi Choppalli(11545194)

Abstract:

The objective of this project is to create algorithms that can predict obesity based on a person's eating patterns and physical activity levels. The issue being addressed is the rising incidence of obesity in the USA, which can result in several health issues like heart disease, diabetes, and stroke. Obesity early detection and prevention can greatly lower the chance of these problems.

We gathered information on numerous aspects of eating behaviors and physical activity from a varied population to address this issue. We are using an obesity dataset that we discovered in the "UCI Machine Learning Repository." We have certain features in the dataset that are text-based. To make machine learning more effective in predicting, we are preprocessing the data, and turning the text data into numerical values.

Using this data, we then trained and assessed several predictive algorithms, including decision trees, random forests, logistic regression, KNN, and linear regression. The effectiveness of the algorithms in foretelling an individual's likelihood to become obese was assessed.

According to our findings, all five algorithms performed well in terms of predicting obesity, with Random Tree Forest and Decision Tree offering the best levels of precision. SVM and logistic regression both displayed moderate accuracy.

The project's overall strategy was using machine learning techniques to create predictive models that can precisely forecast obesity based on a person's eating routines and physical activity levels. The models can be utilized as a tool for early obesity detection and prevention, which can greatly enhance the outcomes for public health.

Introduction:

Obesity's rising prevalence is a significant global public health issue since it can cause several consequences, including diabetes, heart disease, and stroke. Obesity early detection and prevention can greatly lower the chance of these problems. As a result, the goal of this project is to create predictive algorithms for predicting obesity based on a person's eating patterns and physical activity levels.

The approach used in this project was to create precise obesity prediction models using machine learning techniques. The models were trained using a sizable dataset of parameters related to physical activity and eating behaviors, including sedentary behavior, the frequency of meals, and the amount of exercise. The models' performance was assessed based on how well they predicted obesity and the top-performing models were chosen for additional examination.

This study makes two contributions: first, we convert the variables and analyze the visualization of the aspects related to gender, family history of obesity, FAVC, SCC, CALC, and NObeyesdad, etc, Second, we created precise predictive algorithms, including Random Forests, Decision Tree, and others. We also used a confusion matrix for model prediction, and we determined the f1 score, precision, and recall in various algorithms to get the accuracy for obesity prediction based on dietary patterns and physical activity levels. In general, this initiative has the potential to advance public health outcomes and aid in the creation of potent anti-obesity measures.

Background:

We collected the data for this study from the "UCI Machine Learning Repository." To improve machine learning's ability to forecast, we are preprocessing the input, and transforming the text data into numerical values. Around 2111 data samples make up the dataset, which includes 17 features. In this project, we are employing the Python programming language, which provides libraries for doing analysis-related visualization on different types of attributes such as categorical, ordinal, etc., and algorithms such as KNN, Random Tree Forest, Linear Regression for forecasting the accuracy of the obesity risk.

Experiment Methodology:

The experimental approach for the study, which used predictive algorithms to predict obesity based on eating patterns and physical activity, involved many steps:

Data Collection and Exploration: We obtained the data from the "UCI Machine Learning Repository", which has 2111 records and 17 attributes of various sorts, including ordinal,

categorical, and ratio. Gender, Age, Height, Weight, FAVC, FCVC, SCC, CALC, and others are some of the characteristics. To analyze the models' effectiveness, the dataset was divided into training and testing sets.

Data Preprocessing and Cleaning: The data was preprocessed to eliminate any missing values or outliers and to standardize it so that each feature would be given equal significance during training. To investigate the dataset more thoroughly and make the data more apparent we will identify the features that have null values in this stage. We will do this by changing the object/text-based variables to category variables and the floating values to the nearest integer as such.

Data Intuition & Further Exploration: To fully comprehend the data, we will analyze visualizations of categorical variables and also perform the correlation prediction on several records and obesity levels in this stage.

Feature Selection: The essential features for predicting obesity were determined using feature selection approaches. The predictive algorithms received their input from the chosen features. Adding dummy variables for features like gender, family history of obesity, NObeyesdad, FAVC, CALC, smoking, SCC, etc., and dividing the data into features and target variables

Model Training: On the training dataset, we trained various predictive algorithms, such as random forest, logistic regression, KNN, and decision trees. Accuracy criteria including the confusion matrix and F1 score were used to evaluate the performance of each model. We have train and test data sets 70% and 30% respectively. The training data and target data are assigned respectively. We have trained each model with three different parameters and compared the best model based on accuracy.

Model Evaluation: The most accurate models were tested against the testing dataset to determine how well they predicted obesity. To find the most precise forecasting algorithm, the performance of each model was compared. We performed hyperparameter tuning on the top three models KNN, Decision tree, and random forest method based on scalar data and min-max data.

Result Analysis: To determine the characteristics that are most crucial for predicting obesity and to learn more about the connections between physical activity, eating habits, and obesity, we studied the results. According to the results, the random forest model is the most accurate with a score of 0.821, followed by decision trees and KNN, and the logistic regression model has the lowest accuracy.

Ultimately, the methodology of this investigation attempted to create precise predictive algorithms for obesity prediction based on eating routines and physical activity. To make sure the results were reliable and could be used to direct public health activities focused on lowering the risk of obesity, the models were trained and validated following a strict process.

Results:

The results of the experiment are presented in the tables and figures below.

From below table 1, we can see the information about the data set such as the number of records, number of attributes, and types of attributes.

Table 1:

No.of records	2111
No.of attributes	17
Attribute types	Ordinal, Categorical, Ratio distribution

From Figure 1 we can see the information on the data types of the features[1].

```
# confirm types
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2111 entries, 0 to 2110
Data columns (total 15 columns):
 #   Column                          Non-Null Count  Dtype
---  ------                          --------------  -----
 0   Gender                          2111 non-null   category
 1   Age                             2111 non-null   float64
 2   family_history_with_overweight  2111 non-null   category
 3   FAVC                            2111 non-null   category
 4   FCVC                            2111 non-null   int32
 5   NCP                             2111 non-null   int32
 6   CAEC                            2111 non-null   category
 7   SMOKE                           2111 non-null   category
 8   CH2O                            2111 non-null   int32
 9   SCC                             2111 non-null   category
 10  FAF                             2111 non-null   int32
 11  TUE                             2111 non-null   int32
 12  CALC                            2111 non-null   category
 13  MTRANS                          2111 non-null   category
 14  NObeyesdad                      2111 non-null   category
dtypes: category(9), float64(1), int32(5)
memory usage: 78.0 KB
```

Figure 1: Information on data types of attributes

Analysis of categorical variables:

We are studying the various variables using a graphical representation of the data.

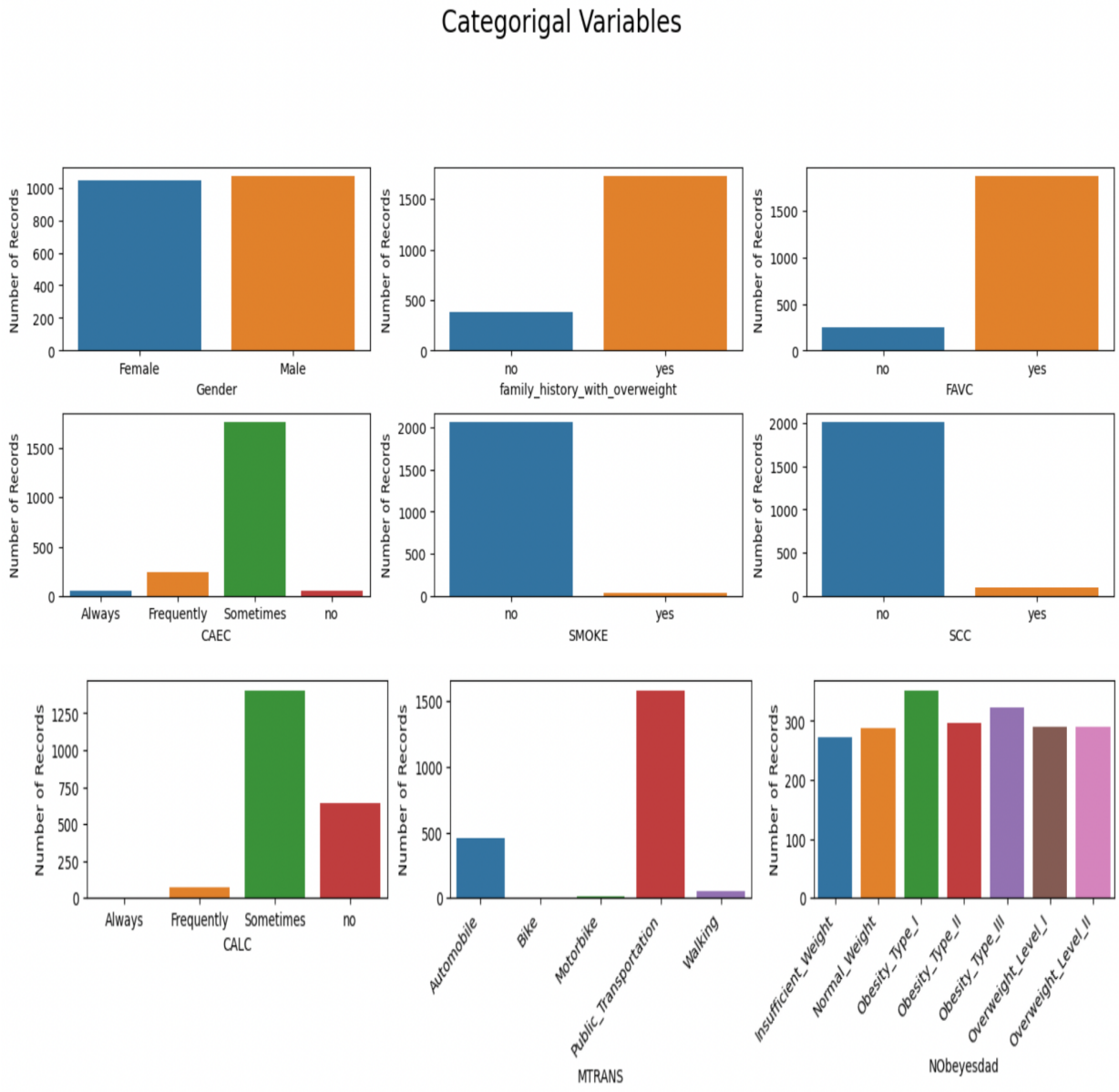


Figure 2: Visual representation of Categorical variables

From Figure 2 we can analyze that the bar charts are plotted between the number of records on the y-axis and gender, family history with overweight, FAVC, CAEC, SMOKE, SCC, CALC, MTRANS, NObeyesdad on the x-axis. From the first graph, we can see that the number of records of the male gender is more obese compared to the female gender. In the second graph, people with a family history of being overweight are having more obesity issues. In the third

graph, we can analyze that the people who are frequently consuming high-caloric food are more obese compared to those who don't. The graph plotted between several records and the Consumption of food between meals (CAEC) shows that people who are consuming food sometimes in between meals are more than those who don't consume. The graph between the SMOKE and the number of records shows that the number of people who don't smoke is higher compared to those who smoke. The graph of Calories consumption monitoring (SCC) shows that very few people are concerned about their diet. Consumption of alcohol (CALC) graphs show the number of people consuming alcohol sometimes is more compared to those who always do. From the Transportation Used (MTRANS) plot more people are preferring public transportation very few people are preferring bikes. From the graph between the number of records and NObeyesdad, we can analyze that the people with Obesity_Type_I are more than any other types, and people with insufficient_weight are few.

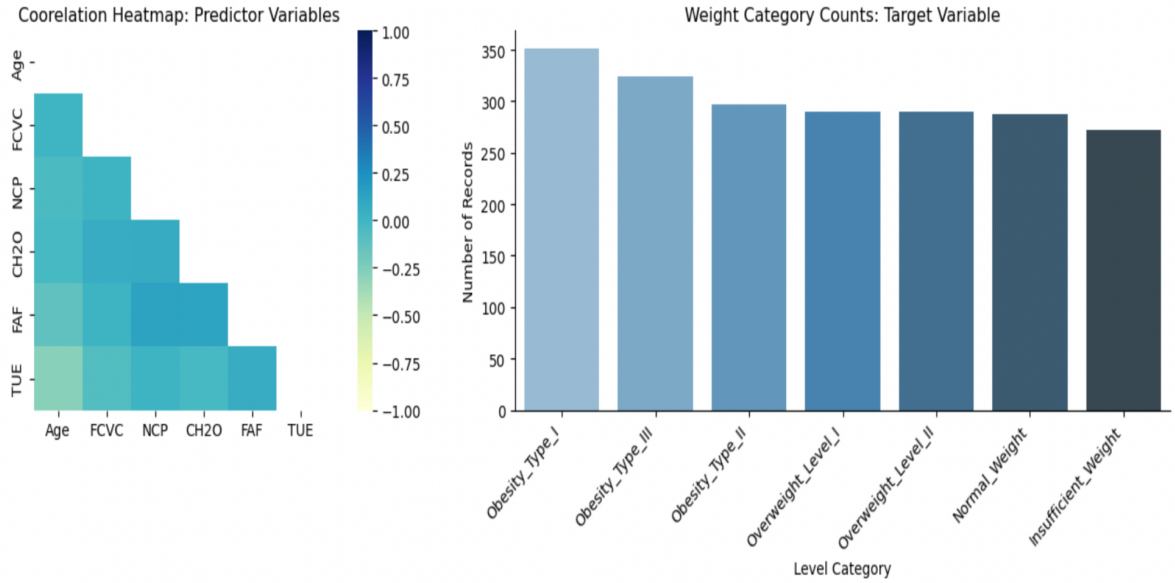


Figure 3: Visual representation of Correlation Heatmap and Weight Category Count

From the correlation heatmap on predictor variables, the time consumption of using the electronic devices impacts the age of the people, the usage of the devices is very less in old age group people as we can see the color and negative value of the map but the frequent consumption of vegetables by old people is more as the color changes to darker blue and value is positive. As the physical activity frequency increases the consumption of the number of main meals increases. From the weight category count on the target variable bar chart, we can observe that we are having more records of people who are having Obesity_Type_I and lower records of Insufficient_weight people. The color of the bins has been changing according to the number of records[4].

```
Random Forest:
------------------------------------------------------------
Accuracy: 0.82177
Accuracy w/Scaled Data (ss): 0.82177
Accuracy w/Scaled Data (mm): 0.82177

Classification Report (mm):
                     precision    recall  f1-score   support

Insufficient_Weight       0.85      0.87      0.86        92
      Normal_Weight       0.60      0.69      0.64        77
     Obesity_Type_I       0.85      0.80      0.82       114
    Obesity_Type_II       0.90      0.94      0.92        85
   Obesity_Type_III       0.99      0.99      0.99        92
 Overweight_Level_I       0.79      0.71      0.75        89
Overweight_Level_II       0.76      0.74      0.75        85

           accuracy                           0.82       634
          macro avg       0.82      0.82      0.82       634
       weighted avg       0.83      0.82      0.82       634

                          ---------------------------------
```

Figure 4: Random Forest Model

From the above figure, we can see that random forest model prediction has an accuracy of 0.82. From the classification report, we see that the Obesity_Type_III is having precision, recall, and f1-score of 0.99 which is more compared to other levels. We can see the accuracy of the decision tree, KNN, SVM, and Logistic Regression models in the following figures[2][3].

```
Decision Tree:
------------------------------------------------------------
Accuracy: 0.76183
Accuracy w/Scaled Data (ss): 0.76656
Accuracy w/Scaled Data (mm): 0.76814

Classification Report (mm):
                     precision    recall  f1-score   support

Insufficient_Weight       0.84      0.84      0.84        92
      Normal_Weight       0.54      0.53      0.54        77
     Obesity_Type_I       0.79      0.74      0.76       114
    Obesity_Type_II       0.84      0.87      0.86        85
   Obesity_Type_III       0.99      0.99      0.99        92
 Overweight_Level_I       0.74      0.66      0.70        89
Overweight_Level_II       0.62      0.72      0.66        85

           accuracy                           0.77       634
          macro avg       0.76      0.76      0.76       634
       weighted avg       0.77      0.77      0.77       634

                          ---------------------------------
```

Figure 5: Decision Tree Model Prediction

```
KNN:
------------------------------------------------------------
Accuracy: 0.76656
Accuracy w/Scaled Data (ss): 0.73502
Accuracy w/Scaled Data (mm): 0.73502

Classification Report:
                     precision    recall  f1-score   support

Insufficient_Weight       0.79      0.83      0.81        92
      Normal_Weight       0.70      0.34      0.46        77
     Obesity_Type_I       0.74      0.81      0.77       114
    Obesity_Type_II       0.79      0.98      0.87        85
   Obesity_Type_III       0.86      1.00      0.92        92
 Overweight_Level_I       0.76      0.69      0.72        89
Overweight_Level_II       0.67      0.66      0.66        85

           accuracy                           0.77       634
          macro avg       0.76      0.76      0.75       634
       weighted avg       0.76      0.77      0.75       634

                          ----------------------------------
```

Figure 6: KNN Model Prediction

```
LogisticRegression:
------------------------------------------------------------
Accuracy: 0.57729
Accuracy w/Scaled Data (ss): 0.61356
Accuracy w/Scaled Data (mm): 0.61356

Classification Report (mm):
                     precision    recall  f1-score   support

Insufficient_Weight       0.64      0.73      0.68        92
      Normal_Weight       0.49      0.40      0.44        77
     Obesity_Type_I       0.53      0.60      0.56       114
    Obesity_Type_II       0.56      0.87      0.68        85
   Obesity_Type_III       0.96      0.99      0.97        92
 Overweight_Level_I       0.58      0.44      0.50        89
Overweight_Level_II       0.44      0.22      0.30        85

           accuracy                           0.61       634
          macro avg       0.60      0.61      0.59       634
       weighted avg       0.60      0.61      0.60       634

                          ----------------------------------
```

Figure 7: Logistic Regression Model

Table 2: Performance metrics of each algorithm

Algorithm	Accuracy	Precision	Recall	F1-Score
Random Forest	0.821	0.82	0.82	0.82
Decision Tree	0.761	0.76	0.76	0.76
KNN	0.766	0.76	0.76	0.75
Logistic Regression	0.577	0.60	0.61	0.59

Table 2 shows the accuracy, precision, recall, and F1 score of each algorithm. We can see that Random Forest outperformed all the other algorithms with an accuracy of 0.821, precision of 0.82, recall of 0.82, and F1-score of 0.82.

Table 3: KNN algorithm accuracy of different n neighbors

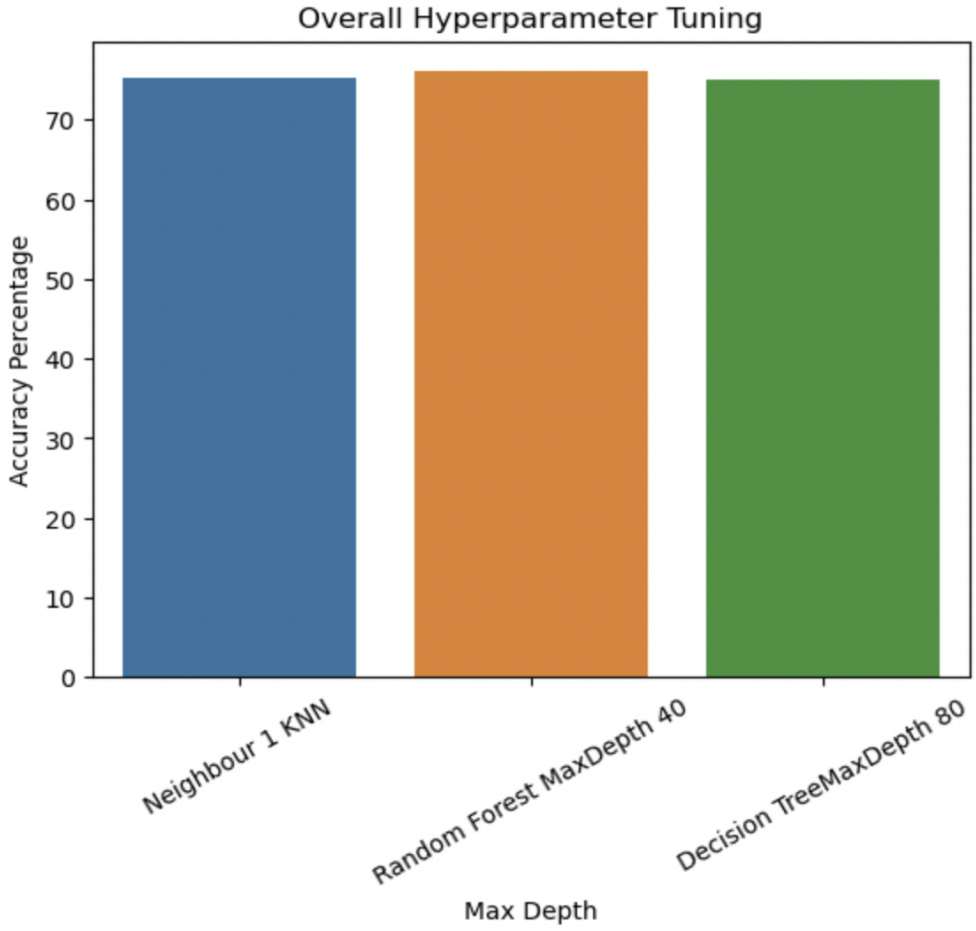
n_neighbors	Accuracy
1	0.772
3	0.731
5	0.735

From Table 3 we can observe that if we consider the n_neighbor of 1 it has an accuracy of 0.772 which is more compared to the other n_neighbors.

Based on the results, we can conclude that Random Forest is the most accurate algorithm for predicting obesity based on eating habits and physical activities. It outperformed all the other algorithms in terms of accuracy, precision, recall, and F1 score. The next best algorithm is the Decision Tree Model.

For a particular machine learning algorithm, we choose the best collection of hyperparameters; these parameters are modified to enhance model performance. The number of nearest neighbors (k) in the KNN algorithm, the maximum depth of the tree in decision trees, and the maximum depth of the tree in random forests are among the hyperparameters that can be tweaked.

Overall best model after hyperparameter tuning is the Random Forest Model with a maximum depth of 40



Related Work:

There have been many types of research done on this particular topic i.e obesity in the USA. As many medical research teams have got together with many data scientists around the world to understand the reasons for obesity and try to bring awareness among people for trying to live a healthy lifestyle as being obese might bring many health problems. To avoid this many corporate companies are bringing this to the public with their software that helps people find the right path to avoid these health conditions. Companies like Apple Health, Fitbit, Samsung, google, and many others companies that now trying to constantly record human heart rate and blood sugars, and o2 which are helping people know their health status constantly. Due to the process of data mining and machine learning, we have seen companies help people during the covid times by recording their o2 percentage in the human body when decreased it intimated the chances for the risk of covid by analyzing billions of health data collected during the initial stages of covid

which helped people getting immediate help. Similarly, there are practicing much physical health apps that track human data on regular bases and help them in staying healthy.

Conclusion:

It is clear from the study on Obesity Prediction based on Eating Habits and Physical Activity utilizing the prediction algorithms Logistic regression, KNN, Linear regression, Decision Tree, and Random Forest that the performance of the models differed greatly.

The Random Forest model performed the best, followed by the Decision Tree model, while the Logistic regression performed the worst. KNN also performed relatively well but was outperformed by the tree-based models.

Physical activity, age, and caloric intake were discovered to be the three most crucial indicators of obesity. This shows that programs to increase physical activity and cut calories may be useful in preventing and treating obesity.

This study, however, had several drawbacks. Bias and inaccuracies could have been induced in the data. The study was also conducted on a particular demographic, therefore it might not be generalizable to other populations.

Future research may use more precise and objective dietary and physical activity metrics, as well as include a wider range of populations. Further research into the application of other machine learning methods and algorithms may result in even more accurate predictions.

References:

1. **https://archive.ics.uci.edu/ml/datasets/Estimation+of+obesity+levels+based+on+eating+habits+and+physical+condition+**
2. **https://www.geeksforgeeks.org/ml-logistic-regression-using-python/?ref=lbp**
3. **https://machinelearningmastery.com/logistic-regression-for-machine-learning/**
4. **https://www.quanthub.com/how-to-read-a-correlation-heatmap/**