# Machine Learning-Based Classification of Obesity Levels from Lifestyle Indicators in Adult

VENKATA SAI MANIKANTA MALIREDDY

VIVEK ATCHUTANNA

MAHENDER TANNIRU

**COURSE:** AIT-736 DL-1

**GEORGE MASON UNIVERSITY**

**EMANUELA MARASCO**

## ➢ Introduction:

Globally, the increasing incidence of obesity has emerged as a serious public health issue that affects people's physical health and places a heavy strain on healthcare systems. A complex disorder, obesity is impacted by lifestyle choices, environmental factors, and genetics. Our project's goal is to investigate how lifestyle factors, particularly alcohol use, screen time, physical activity, and smoking behaviors, can predict obesity and associated health outcomes.

This project aims to apply machine learning techniques to investigate the predictive power of various lifestyle factors in determining obesity and related metrics such as Body Mass Index (BMI) and body weight. Specifically, we address the following research questions:

We will do regression and classification studies using a structured dataset that includes comprehensive data on people's physical characteristics, habits, and demographics. This data-driven method makes it possible to identify important obesity predictors, providing information that could guide individualized health interventions and preventative measures. The research also functions as a real-world application of machine learning in behavioral epidemiology and public health.

## ➢ Abstract:

One of the biggest public health issues facing people worldwide is obesity, which is a major factor in the rise in chronic illnesses like diabetes, heart disease, and some types of cancer. The purpose of this study is to investigate the relationship between a person's risk of obesity and a variety of lifestyle characteristics, including the amount of time spent on screens each day, the frequency of physical activity, and the consumption of vegetables. Three research questions that direct our investigation and offer a framework for comprehending the multifaceted nature of obesity have been developed in order to obtain deeper insights.

In order to guarantee its appropriateness for machine learning applications, the dataset utilized in this study was meticulously selected and assembled from a variety of sources. Handling missing values, eliminating duplicate entries, label encoding of categorical variables, and, if necessary, normalizing numerical characteristics were all preprocessing stages.

We used a variety of statistical and machine learning methods, such as decision trees, logistic regression, KNN, support vector machines (SVM), and linear regression, to answer the research objectives. These models were selected to evaluate the linear and non-linear associations between obesity levels and lifestyle choices. Model performance was compared using evaluation criteria like accuracy, precision, recall, and F1-score.

Our results demonstrate that there is no single cause of obesity but rather a confluence of environmental and behavioral factors, underscoring the necessity of individualized and focused preventative measures. This work advances our understanding of obesity and aids in the creation of more potent public health treatments by utilizing data-driven insights.

## ➢ **Dataset Description:**

By collecting a wide range of real-world behavioral, lifestyle, and demographic characteristics that contribute to obesity, this dataset tackles a crucial global health issue. Knowing the underlying causes is crucial because obesity is becoming a major risk factor for chronic diseases including diabetes, cardiovascular problems, and several types of cancer, and its prevalence are rising globally. This dataset makes it possible to gain data-driven insights into the ways that daily routines like as eating, exercising, using screens, and drinking water affect body weight and health outcomes. Researchers and medical practitioners can provide more individualized health recommendations, focused interventions, and improved prevention measures by examining these characteristics. The collection also aids in the creation of predictive technologies that can detect at-risk persons early on, which will ultimately enhance public health and enable better decision-making.

The information used in this study came from a number of trustworthy sources, such as publicly accessible health records, studies, and lifestyle surveys about human behavior and obesity. Selecting characteristics that are directly related to predicting obesity, such as age, food habits, physical activity, water consumption, screen time, and genetic variables, was a priority during the data gathering phase.

2087 entries and 16 attributes make up the dataset used in this project, which captures a variety of behavioral, lifestyle, and health aspects important for predicting obesity levels. It contains both qualitative and numerical data, with a person represented by each row. A person's age, gender, alcohol use, frequency of high-calorie food intake, frequency of vegetable intake, average number of meals per day, and whether or not they track their caloric intake are important characteristics. Additional characteristics include types of food consumed between meals, physical activity levels, mobile screen time, smoking habits, daily water consumption, family history of obesity, and mode of transportation. Obesity_level, the target variable, divides people into groups like normal weight, overweight, and obese.Each person's Body Mass Index (BMI) is also included in the dataset. Through the analysis of behavioral and physiological aspects, this extensive dataset offers a well-rounded foundation for training machine learning models to predict obesity.

## ➢ Variable description:

| | |
|---|---|
| % - CALC: Categorical | captures the frequency of alcohol consumption. |
| % - FAVC: Binary | indicates the consumption of high caloric food frequently. |
| % - FCVC: Numeric | frequency of vegetable consumption. |
| % - NCP: Numeric | average number of main meals. |
| % - SCC: Binary | indicates if the individual consults a calorie consumption monitoring. |
| % - SMOKE: Binary | represents smoking habits. |
| % - CH2O: Numeric | daily water consumption in liters. |
| % - family_history_with_overweight: Binary | indicates a family history of overweight. |
| % - FAF: Numeric | frequency of physical activity per week. |
| % - TUE: Numeric | time using technology devices in hours. |
| % - CAEC: Categorical | consumption of food between meals. |
| % - MTRANS: Categorical | usual mode of transportation. |
| % - NObeyesdad: Categorical | denotes the obesity level of the individual. |

## ➢ **Data types:**

| Column names | Measurements | NOIR |
|---|---|---|
| Age | Numeric | Ratio |
| Bmi_index | Numeric | Ratio |
| Gender | Categorical | Nominal |
| Alcohol_consumption | Categorical | Nominal |
| High_calories_foodconsumption_frequency | Categorical | Nominal |
| Vegetable_consumption_frequency | Numeric | Ordinal |
| Avg_no_of_meals | Numeric | Ordinal |
| Calorie_consumption_monitoring | Categorical | Nominal |
| Smoking_habits | Categorical | Nominal |
| Water_consumption | Numeric | Ratio |
| Family_history | Categorical | Nominal |
| Physical_activity | Numeric | Ratio |
| Mobile_screentime | Numeric | Ratio |
| Food_consumed_between_meals | Categorical | Nominal |
| Transpotation | Categorical | Nominal |
| obesity_level | Categorical | Nominal |

## ➢ **Dataset preprocessing steps:**

For any machine learning effort to produce accurate and trustworthy results, data cleansing is an essential first step. Missing values, duplicates, or inconsistent datasets can seriously skew analysis and impair model performance. We preprocessed the dataset in a number of ways to guarantee consistency and enhance data quality.First of all, it was discovered that a number of column names were either formatted with unclear abbreviations or were excessively long. In order to improve readability and streamline model development, all column names were changed to be more logical and understandable. Inspection also revealed that decimal values, which are invalid for representing age, were included in the Age field. By casting the column to the proper data type, these values were transformed into integers.

We computed the Body Mass Index (BMI), a crucial metric for forecasting obesity levels, using the dataset's original Height and Weight columns. To prevent repetition, the original Height and Weight columns were removed after calculating the Bmi_index. Additionally, one column had one missing value; this was fixed by using the mean of the corresponding column to impute the missing value, maintaining the balance of the data's statistical distribution.Additionally, we used Python's df.drop_duplicates() function to find and eliminate 24 duplicate entries. To avoid model bias and guarantee that every record represented a distinct instance, duplicates had to be eliminated. To avoid mistakes during feature selection and model training, a number of column names also had abnormalities, such as leading or trailing whitespaces (e.g., "Gender"), which were fixed using string manipulation techniques.

To guarantee a repeatable and automated preprocessing workflow, Python was used to carry out these data cleaning procedures: resolving missing values, eliminating duplicates, standardizing column names, fixing invalid entries, and managing outliers. The dataset was subsequently converted into a dependable, consistent, and clean format that was ideal for efficient machine learning modeling.

## ➢ Research Questions:

### 1.Can smoking and alcohol habits predict whether someone likely to be obese?

Using the information, we developed three main research questions in order to provide a thorough investigation of obesity and the lifestyle factors that contribute to it. The first study topic looks into how a person's drinking and smoking habits affect their risk of obesity. We concentrate on the variables that reflect alcohol use and smoking frequency for this research. In order to ascertain if substance-related behaviours can serve as major predictors of obesity, we investigate the relationship between these behavioural characteristics and obesity levels using statistical methods and visualisation tools

### 2.To what extent does screen time or physical activity affect a person's weight?.

The effect of screen time and exercise on an individual's weight is the subject of the second question. We make use of the features in the dataset that describe the frequency of physical exercise and daily mobile screen usage. To comprehend their individual effects on BMI and weight categories, these variables are examined and visualised. Our goal is to ascertain whether excessive screen time, which is a sign of sedentary behaviour, has a negative correlation with physical fitness and raises obesity rates.

### 3.Can we predict a person’s obesity level from their lifestyle factors?

Finally, determining how a person's overall lifestyle decisions affect their obesity status is the third and more general goal of this research. The dataset contains a variety of lifestyle characteristics that are used as input variables for predictive modelling, such as food patterns, activity levels, and behavioural patterns. We aim to determine which lifestyle factors are most closely linked to obesity by using machine learning approaches. By using a comprehensive approach, we can derive valuable insights that can be applied to the creation of individualised weight-management plans and focused public health interventions.

## ➢ Model Evaluation:

### 1.Can smoking and alcohol habits predict whether someone likely to be obese?

Two categorical behavioural features—alcohol use and smoking habits—with values like Yes, No, and Sometimes are included in the dataset. Normal Weight, Obesity Level I, and Obesity Level II are among the categories for the target variable, obesity_level. This study uses two machine learning models—Random Forest Classifier and Linear Regression—to investigate the connection between these behaviours and obesity levels. Even while these lifestyle characteristics have some predictive power, they are not very good indicators on their own and have a low categorisation accuracy.

First, we used Python to create linear regression, utilising Scikit-learn for modelling and Pandas for data preprocessing. The model's performance was evaluated using Mean Squared Error (MSE) and R2 Score after the data was divided into training and testing sets (80:20). To enhance classification, we then used a Random Forest Classifier with 100 decision trees. A Confusion Matrix, Classification Report, and Accuracy Score were among the evaluation measures used to gauge the model's capacity to forecast obesity levels in relation to alcohol and smoking behaviours.

### 2.To what extent does screen time or physical activity affect a person's weight?

This analysis makes use of Python-based data exploration and visualisation techniques to investigate the degree to which a person's weight is impacted by their mobile screen time and physical activity. After loading and previewing the dataset, correlation analysis is made possible by mapping the category variable obesity_level to a numerical scale. Scatter plots are created using Seaborn and Matplotlib to show the association between BMI and physical activity as well as between BMI and mobile screen time. Obesity levels are underlined for categorical clarity. These graphics provide preliminary information about whether a higher BMI is associated with sedentary behaviour and low levels of physical activity.

The strength of the associations between BMI, screen time, physical activity, and obesity levels is then measured by computing a correlation matrix. Regression plots, also known as lmplots, are created to further depict these trends by superimposing linear regression lines on scatter plots to draw attention to directional trends and possible predictive patterns. This method makes it possible to comprehend how these two lifestyle factors may affect a person's weight status and risk of obesity both statistically and visually.

## 3.Can we predict a person’s obesity level from their lifestyle factors?

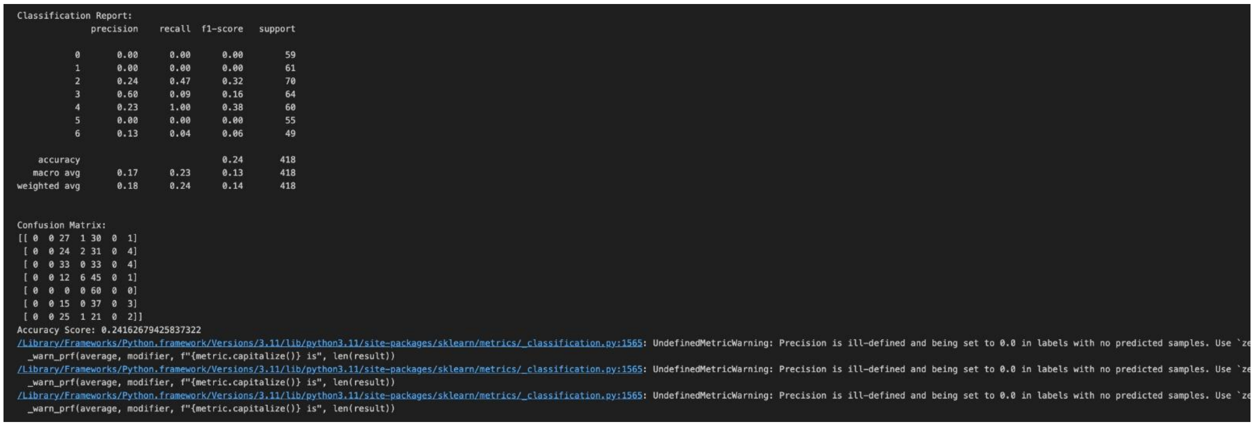
The Support Vector Machine (SVM) is a flexible technique that may be applied to both regression and classification problems. In regression (also known as Support Vector Regression, or SVR), it fits a function within a certain margin to reduce prediction error and preserve model simplicity, but in classification, it finds the best hyperplane to divide classes.

In this project, the BMI index was predicted using SVR with an RBF (Radial Basis Function) kernel based on specific features. Label-encoding was used for categorical variables (apart from the target obesity_level), and mean imputation was used to address missing data. After standardising the characteristics to enhance SVR performance, an 80:20 train-test split was conducted. The R2 score, which gauges how well the model predicts BMI variance, was used to assess model performance following training. To illustrate prediction accuracy, a scatterplot of real versus predicted BMI values was also produced.Assuming a linear connection between input data and output, the supervised learning process known as linear regression is used to predict a continuous target variable. In order to improve generalisation and decrease overfitting, we additionally employed Ridge Regression, a regularised version of Linear Regression that adds an L2 penalty.Both models were used in this study to predict body mass index (BMI) based on a variety of factors. With the exception of the target obesity_level, categorical variables were label-encoded, and the mean was used to impute missing values. The data was divided into training and test sets in an 80:20 ratio after the feature set was standardised using StandardScaler. The R2 score and Root Mean Squared Error (RMSE) were used to train and assess both Ridge Regression (with alpha=1.0) and Linear Regression. Additionally, a scatter plot was produced to show the actual and predicted BMI values for Ridge Regression, offering information on the correctness of the model.

## ➢ **Evaluation Results:**

**Research question 1:**

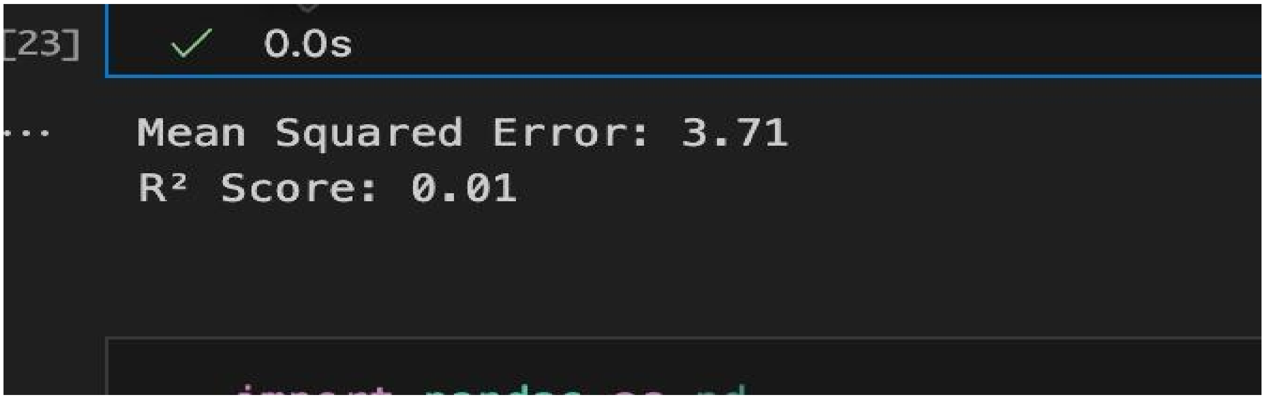
Both classification and regression models were used to evaluate the predictive potential of drinking and smoking habits on BMI and obesity levels in order to address this topic. The findings, however, suggest that these two characteristics by themselves have little predictive power.Based on the output of the Random Forest Classifier, the model's accuracy score was a poor 24.16%. All obesity categories had poor accuracy, recall, and F1-scores in the classification report; certain classes (such as classes 0, 1, 2) have zero precision and recall, meaning the model was unable to predict them at all. This is further supported by the confusion matrix, which shows notable misclassifications in practically every category, particularly for higher obesity levels. This implies that alcohol and smoking behaviours by themselves are insufficient to correctly categorise people into particular obesity groups.



```
Classification Report:
              precision    recall  f1-score   support

           0       0.00      0.00      0.00        59
           1       0.00      0.00      0.00        61
           2       0.24      0.47      0.32        70
           3       0.60      0.09      0.16        64
           4       0.23      1.00      0.38        60
           5       0.00      0.00      0.00        55
           6       0.13      0.04      0.06        49

    accuracy                           0.24       418
   macro avg       0.17      0.23      0.13       418
weighted avg       0.18      0.24      0.14       418


Confusion Matrix:
[[ 0  0 27  1 30  0  1]
 [ 0  0 24  2 31  0  4]
 [ 0  0 33  0 33  0  4]
 [ 0  0 12  6 45  0  1]
 [ 0  0  0  0 60  0  0]
 [ 0  0 15  0 37  0  3]
 [ 0  0 25  1 21  0  2]]
Accuracy Score: 0.24162679425837322
/Library/Frameworks/Python.framework/Versions/3.11/lib/python3.11/site-packages/sklearn/metrics/_classification.py:1565: UndefinedMetricWarning: Precision is ill-defined and being set to 0.0 in labels with no predicted samples. Use `ze
  _warn_prf(average, modifier, f"{metric.capitalize()} is", len(result))
/Library/Frameworks/Python.framework/Versions/3.11/lib/python3.11/site-packages/sklearn/metrics/_classification.py:1565: UndefinedMetricWarning: Precision is ill-defined and being set to 0.0 in labels with no predicted samples. Use `ze
  _warn_prf(average, modifier, f"{metric.capitalize()} is", len(result))
/Library/Frameworks/Python.framework/Versions/3.11/lib/python3.11/site-packages/sklearn/metrics/_classification.py:1565: UndefinedMetricWarning: Precision is ill-defined and being set to 0.0 in labels with no predicted samples. Use `ze
  _warn_prf(average, modifier, f"{metric.capitalize()} is", len(result))
```

The outcomes were likewise poor in the Linear Regression model, which sought to forecast BMI as a continuous variable. Only 1% of the variance in BMI could be explained by the model, as indicated by the Mean Squared Error (MSE) of 3.71 and the R2 score of just 0.01. This supports the idea that a person's weight or level of obesity cannot be accurately predicted by their drinking and smoking habits alone.

All things considered, these behavioural characteristics are not very good predictors of obesity on their own, even though they might help when paired with other factors (including nutrition, exercise, genetics, and health issues). To enhance prediction performance, a wider range of physiological and lifestyle factors should be included in future modelling initiatives.



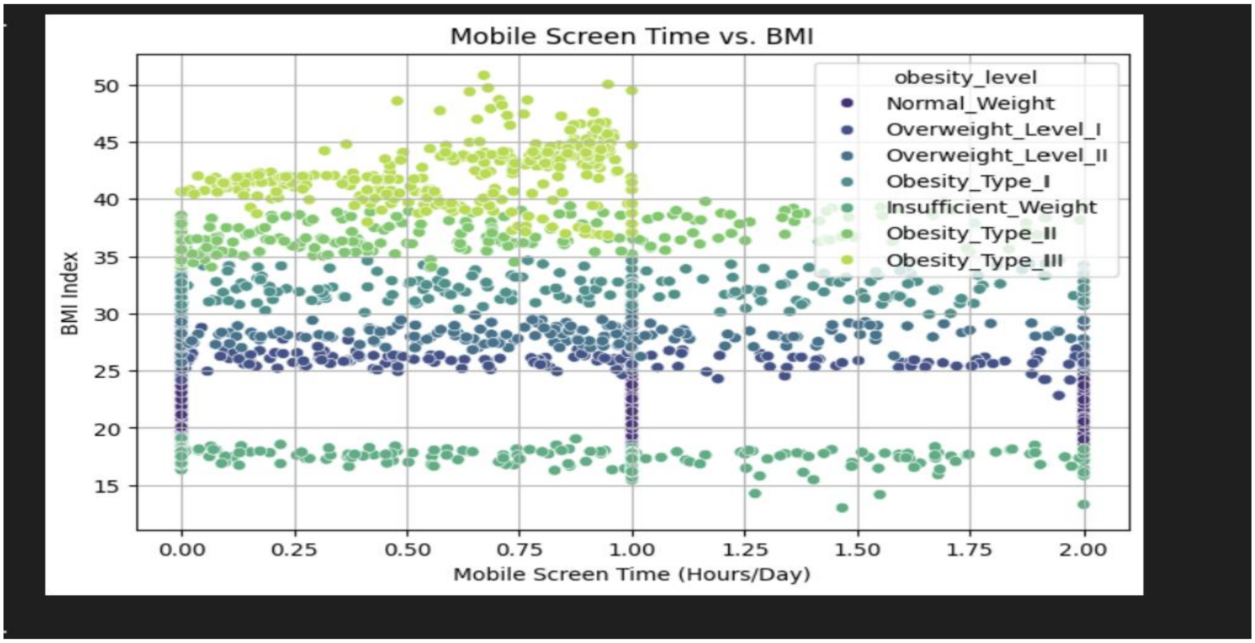
```
[23]   ✓  0.0s

...    Mean Squared Error: 3.71
       R² Score: 0.01

       import pandas as pd
```

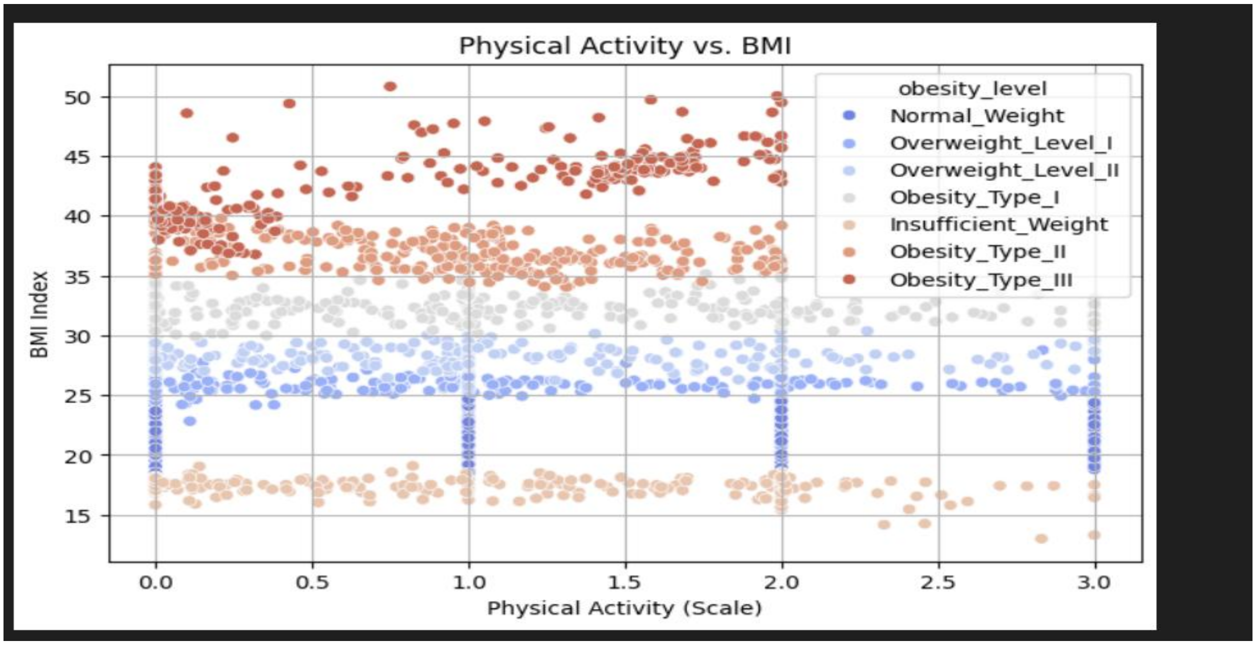
**<u>Research question 2:</u>**

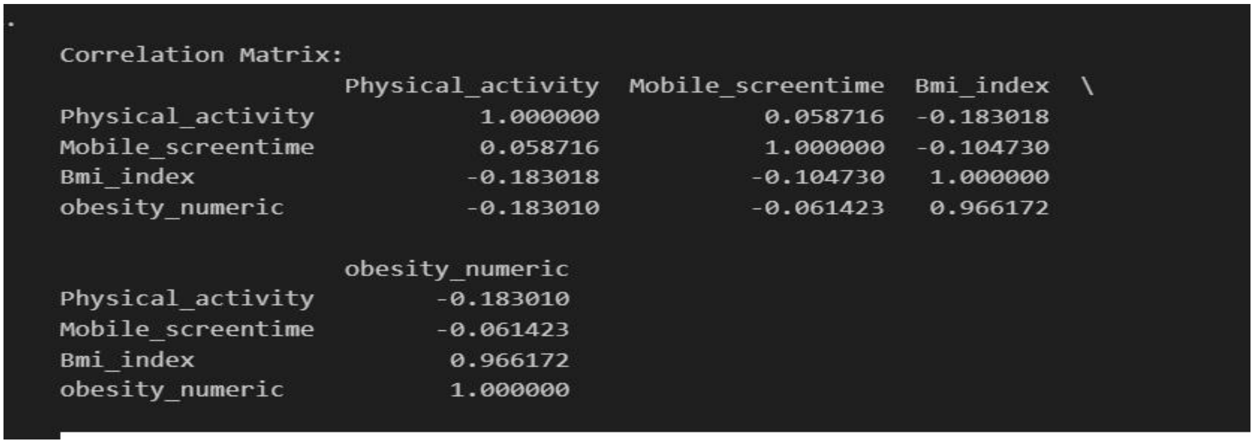
Using scatter plots, regression lines, and a correlation matrix, we conducted exploratory data analysis to assess how screen time and physical activity affected an individual's weight (as determined by BMI).

According to the scatter plots and regression visualisations, people who are more physically active typically have lower BMI values, whereas people who are less or not physically active are more frequently linked to higher BMI and severe obesity categories (such as Obesity Type II and III). The plot for mobile screen time, on the other hand, indicates a marginal rise in BMI with increased screen usage; however, this association is weaker and more diffuse than for physical activity.
These findings are corroborated by the correlation matrix. The BMI Index and physical activity have a somewhat negative association (-0.18), meaning that as physical activity rises, BMI tends to fall. In contrast, there is a slight negative correlation (-0.10) between mobile screen time and BMI, indicating a restricted connection.

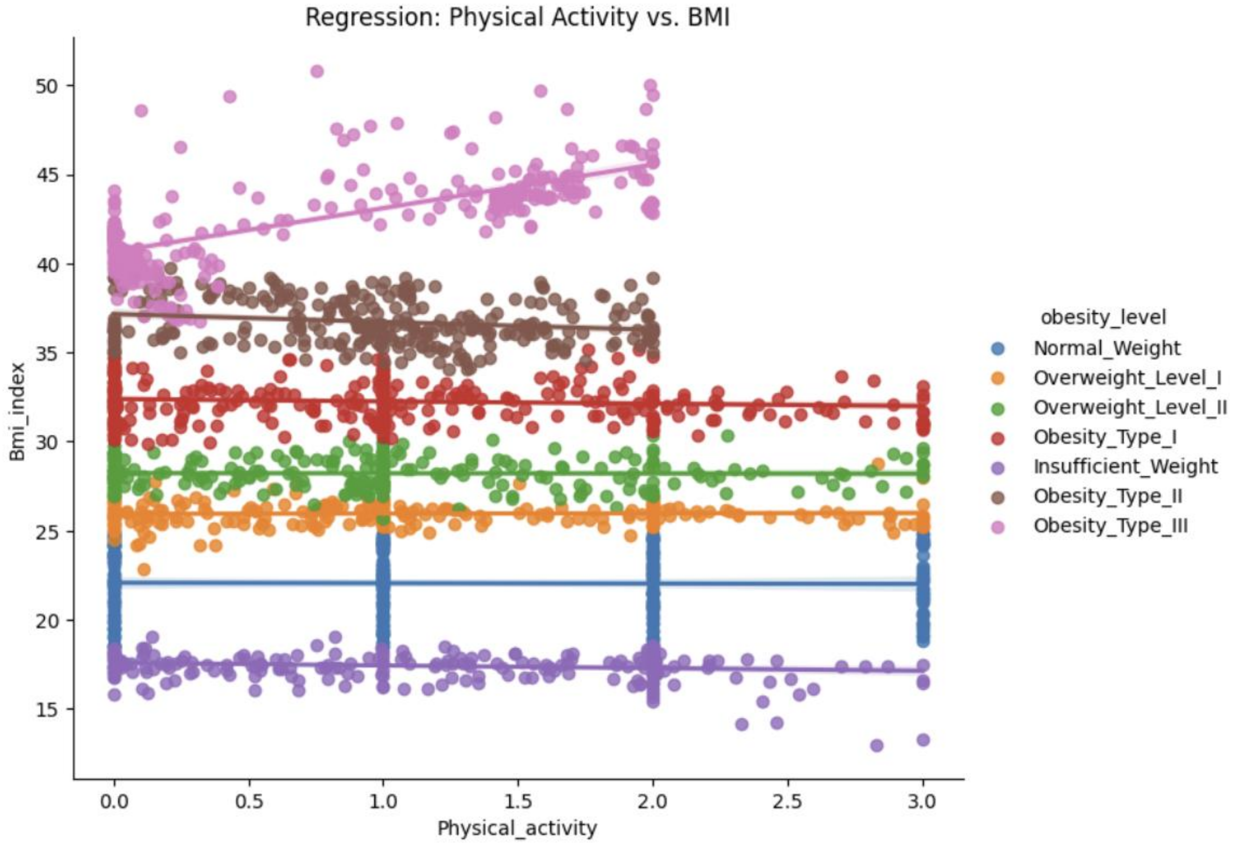
Overall, the data demonstrates that, in comparison to screen time, physical activity is more strongly linked to reduced BMI levels. Although BMI is influenced by both factors, physical activity has a more significant and persistent effect on weight, making it a more trustworthy lifestyle indicator for controlling or predicting obesity.

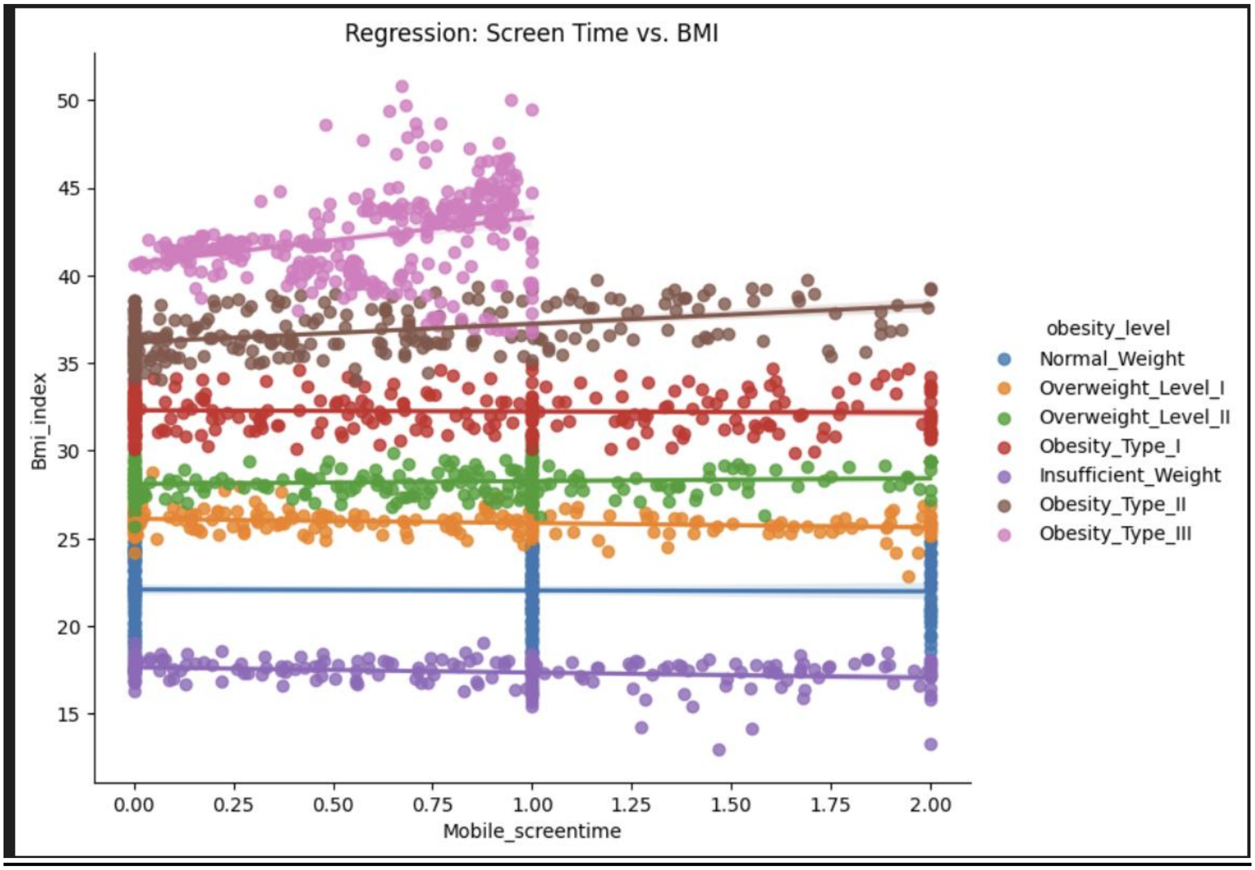

Mobile Screen Time vs. BMI
BMI Index
50
45
40
35
30
25
20
15
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
2.00
Mobile Screen Time (Hours/Day)
obesity_level
Normal_Weight
Overweight_Level_I
Overweight_Level_II
Obesity_Type_I
Insufficient_Weight
Obesity_Type_II
Obesity_Type_III


Physical Activity vs. BMI
BMI Index
50
45
40
35
30
25
20
15
0.0
0.5
1.0
1.5
2.0
2.5
3.0
Physical Activity (Scale)
obesity_level
Normal_Weight
Overweight_Level_I
Overweight_Level_II
Obesity_Type_I
Insufficient_Weight
Obesity_Type_II
Obesity_Type_III



```
.
  Correlation Matrix:
                     Physical_activity  Mobile_screentime  Bmi_index  \
  Physical_activity           1.000000           0.058716  -0.183018
  Mobile_screentime           0.058716           1.000000  -0.104730
  Bmi_index                  -0.183018          -0.104730   1.000000
  obesity_numeric            -0.183010          -0.061423   0.966172

                     obesity_numeric
  Physical_activity        -0.183010
  Mobile_screentime        -0.061423
  Bmi_index                 0.966172
  obesity_numeric           1.000000
```



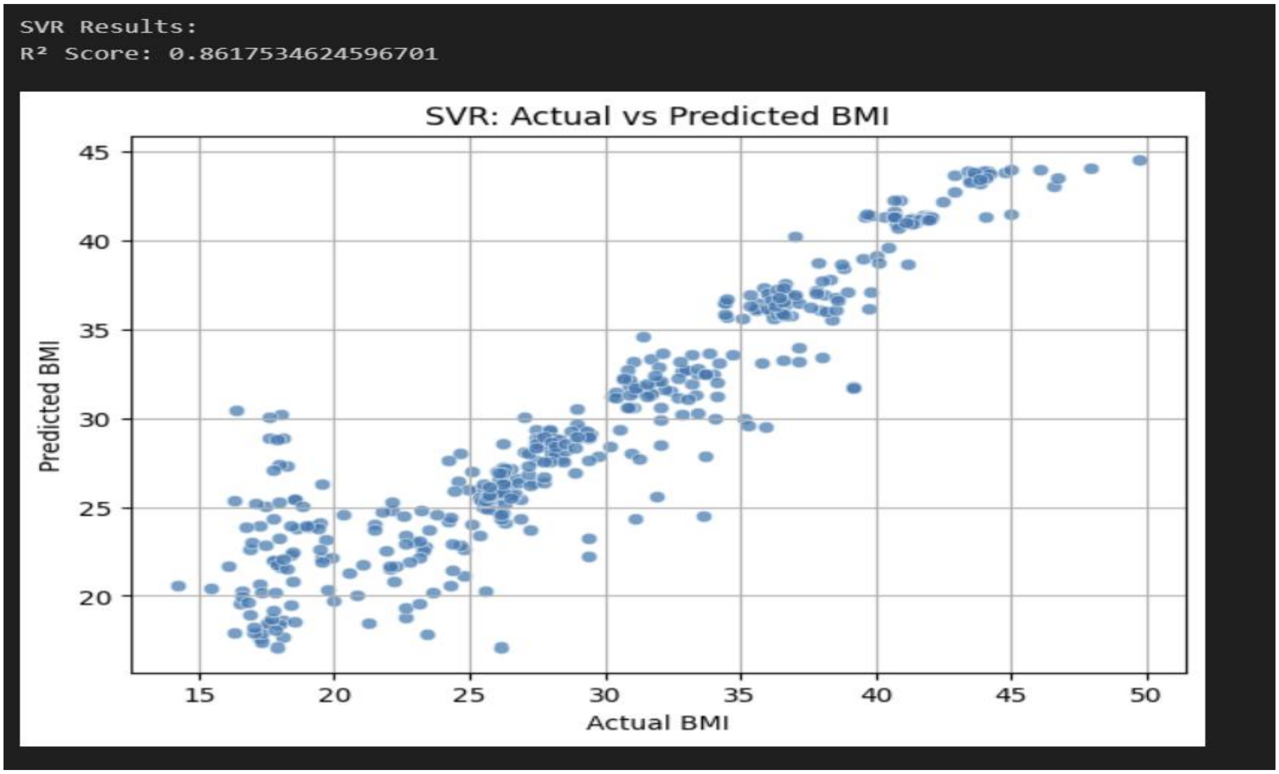


## Research question 3:

We employed a Support Vector Regression (SVR) model to predict the BMI index, a continuous measure of obesity, in order to determine whether a person's degree of obesity could be predicted from their lifestyle choices. Numerous lifestyle characteristics, including screen time, food habits, and physical activity, were included in the dataset. To guarantee consistency and the best possible model performance, all categorical variables (apart from the goal obesity_level) were label-encoded, missing values were imputed using the mean, and features were standardised using StandardScaler.

The scaled data was used to train the SVR model after the dataset was divided into training and testing sets (80:20). With a high R2 value of 0.86, the model was able to account for 86% of the variation in BMI that might be attributed to lifestyle factors. The model's predictions closely match the actual values, as evidenced by the tight linear alignment in the scatter plot comparing the actual and anticipated BMI values.

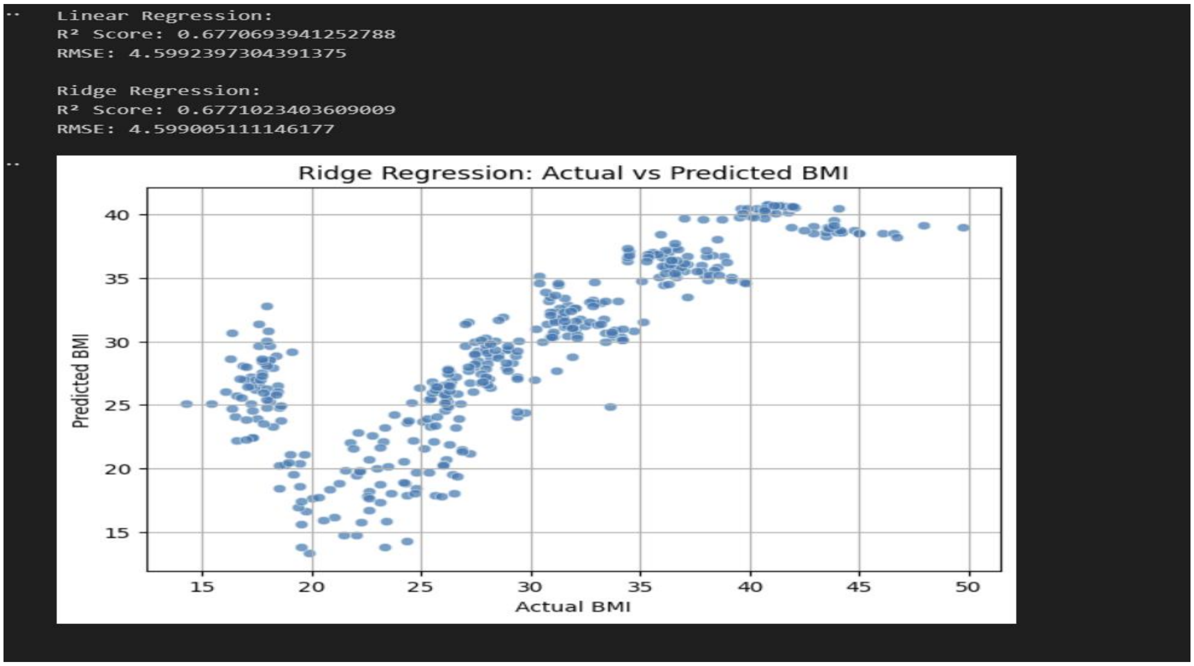
These findings show that lifestyle factors have a high predictive value for determining an individual's BMI and, consequently, their likelihood of falling into an obesity category. This demonstrates that, particularly when processed through strong regression techniques like SVR, obesity levels may be accurately modelled utilising factors like diet, activity levels, and behavioural habits.



```
SVR Results:
R² Score: 0.8617534624596701
```

We estimated BMI using both Linear Regression and Ridge Regression models based on parameters including food habits, physical activity, screen time, and other lifestyle-related inputs in order to ascertain whether lifestyle factors may predict an individual's level of obesity. Both models were trained and assessed using R2 Score and RMSE (Root Mean Squared Error) as performance measures following the encoding of categorical variables and the standardisation of the dataset.

The input lifestyle characteristics may account for almost 68% of the variance in BMI, according to the Linear Regression model's R2 value of 0.6771. A moderate average error in predicting BMI was indicated by the RMSE of 4.59. With an R2 score of 0.6771 and RMSE of 4.60, the Ridge Regression model, which uses L2 regularisation to avoid overfitting, performed almost as well. A distinct positive linear trend can be seen in the scatter plot of actual versus anticipated BMI values, indicating that the models successfully identified important correlations in the data.These findings imply that lifestyle characteristics can, in fact, be used to predict a person's BMI and, consequently, their degree of obesity. Even if they are not flawless, the performance measures show a high capacity for prediction, particularly when regularisation is used to enhance model generalisation.



```
Linear Regression:
R² Score: 0.6770693941252788
RMSE: 4.5992397304391375

Ridge Regression:
R² Score: 0.6771023403609009
RMSE: 4.599005111146177
```

## ➢ Analysis and conclusion from the results:

This study investigated whether a person's BMI or degree of obesity may be accurately predicted by a number of lifestyle characteristics, including smoking, drinking, exercising, and screen time. To obtain insights from various analytical viewpoints, we used a variety of machine learning models, such as Support Vector Regression (SVR), Random Forest Classifier, Ridge Regression, and Linear Regression.

Smoking and drinking are not enough to correctly categorise obesity levels, according to the results of the Random Forest classification models, which exhibited low accuracy (24%) and poor precision/recall across all obesity categories. In a similar vein, Linear Regression produced an extremely low R2 score (0.01) when employed just with these two characteristics, indicating very little predictive value.
Models that employed a wider range of lifestyle characteristics, however, fared noticeably better. With an R2 score of 0.86, the SVR model showed excellent predictive ability when calculating BMI from lifestyle factors. With R2 scores of about 0.677, both Ridge and Linear Regression demonstrated respectable accuracy, demonstrating that adding a variety of lifestyle factors enhances model performance.

Additionally, screen time only exhibited a mild link with BMI, but physical activity showed a moderately unfavourable correlation, according to exploratory data analysis. This suggests that screen time has a less significant and stable effect on an individual's weight than physical activity.

## ➢ **Referances:**

### 1. OpenML Dataset:

OpenML. (n.d.). Blood donation dataset (id: 45969). OpenML. Retrieved April 11, 2025, from https://www.openml.org/search?type=data&status=active&id=45969&sort=runs

### 2. Tableau on Data Cleaning:

Tableau. (n.d.). What is data cleaning? Tableau Software. Retrieved April 11, 2025, from https://www.tableau.com/learn/articles/what-is-data-cleaning#:~:text=Data%20cleaning%20is%20the%20process,to%20be%20duplicated%20or%20mislabeled.

### 3. W3Schools on Linear Regression:

W3Schools. (n.d.). Python machine learning – Linear regression. W3Schools. Retrieved April 11, 2025, from https://www.w3schools.com/python/python_ml_linear_regression.asp

### 4. IBM on Support Vector Machines:

IBM. (n.d.). What are SVMs? IBM. Retrieved April 11, 2025, from https://www.ibm.com/think/topics/support-vector-machine#:~:text=How%20SVMs%20work-,What%20are%20SVMs%3F,in%20an%20N%2Ddimensional%20space.

### 5. Scikit-learn Random Forest Documentation:

Scikit-learn. (n.d.). sklearn.ensemble.RandomForestClassifier. Scikit-learn. Retrieved April 11, 2025, from https://scikitlearn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html