University of Science and Technology of Hanoi

Analysis of Spatial and Temporal Data Final Report

Obesity Classification

Group 5
BI12-389 Nguyen Son
BI12-447 An Minh Tri
BI12-375 Nguyen Cong Quoc
BI12-314 Truong Tuan Nghia
BI12-357 Nguyen Duc Phuong
BI12-390 Vu Hung Son

Lecturer: *Nguyen Xuan Thanh (Obese)*

Academic Year 3 - ICT Department
February 2024

I. Introduction

Staying healthy is so important these days. With fast food everywhere and many of us living more sedentary lifestyles, maintaining a healthy weight can be a real challenge. The stats are pretty alarming - almost 1 in 3 adults are considered overweight, over 2 in 5 are obese, and about 1 in 11 have severe obesity. Those numbers really show the scale of the obesity problem, especially since being overweight or obese puts you at serious risk for medical issues down the line.

In this report, we looked at two key aspects of the obesity issue. First, we examined obesity rate trends from 1975 to 2016 for men and women around the world, with a specific focus on Vietnam. By studying how rates have changed over time in different locations, we hoped to gain more insight into what drives obesity and how it impacts various populations.

Secondly, we used obesity data to build a computer model that tries to predict whether teachers are obese or not. We wanted to see if any patterns emerged in the data that could help us better understand obesity among teachers. Our ultimate goal was that by shedding light on these patterns, we might discover better ways to support people in maintaining a healthy lifestyle and avoiding obesity-related health problems.

II. Obesity by Country Dataset

1. Data Preprocessing

After loading the data, we noticed that it included some rows that didn't contain useful information, such as “Prevalence of obesity among adults, BMI ≥ 30 (age-standardized estimate) (%)” and “18+ years.” We skipped these rows in the dataset. Following this cleanup, the dataset was organized with two header rows indicating the **Year** (spanning from 1975 to 2016), **Gender** (categorized as Male, Female, and Both sexes), and a header column **Country** listing the names of 195 countries from around the world in alphabetical order.

Within the dataset, each cell presented data in the format x [y-z], where x represents the percentage of obese individuals of a particular gender in a country, and y and z indicate the lower and upper bounds, respectively, of that country's data. We focused on extracting the valuable numeric information x, y, and z from each cell.

During our data cleaning process, we ensured that there were no missing values. However, we found that a few countries, namely Monaco, San Marino, South Africa, and Sudan, lacked data, so we skipped these rows as well.

2. Filter the Dataset

After completing the final cleanup, our new dataset now includes data from 191 countries that we can analyze. We've organized this data based on Gender, Continent, and specifically for Vietnam.

The dataset identifies two genders, Male and Female, and also includes data that combines both, labeled "Both Sexes." As a result, we've divided the original dataset into three smaller subsets: one for Male, one for Female, and one for Both Sexes. This division allows us to more easily examine the percentage of obesity among different genders in various countries from 1975 to 2016.

Regarding continent-based organization, it's generally recognized that there are six continents. However, since Antarctica doesn't have any countries relevant to our project, we've categorized the dataset into five regions instead: Asia, America, Europe, Africa, and Oceania. This breakdown helps us analyze the data with a focus on specific geographical areas.

Lastly, since we all are located in Vietnam, we decided it would be beneficial to have a separate dataset specifically for Vietnam. This allows us to give special attention to data relevant to our immediate context.

3. Data Analysis

Across the globe, obesity rates have been steadily rising for both males and females, especially females over the decades. This increase has been a significant public health concern, with implications for various chronic diseases and healthcare systems.

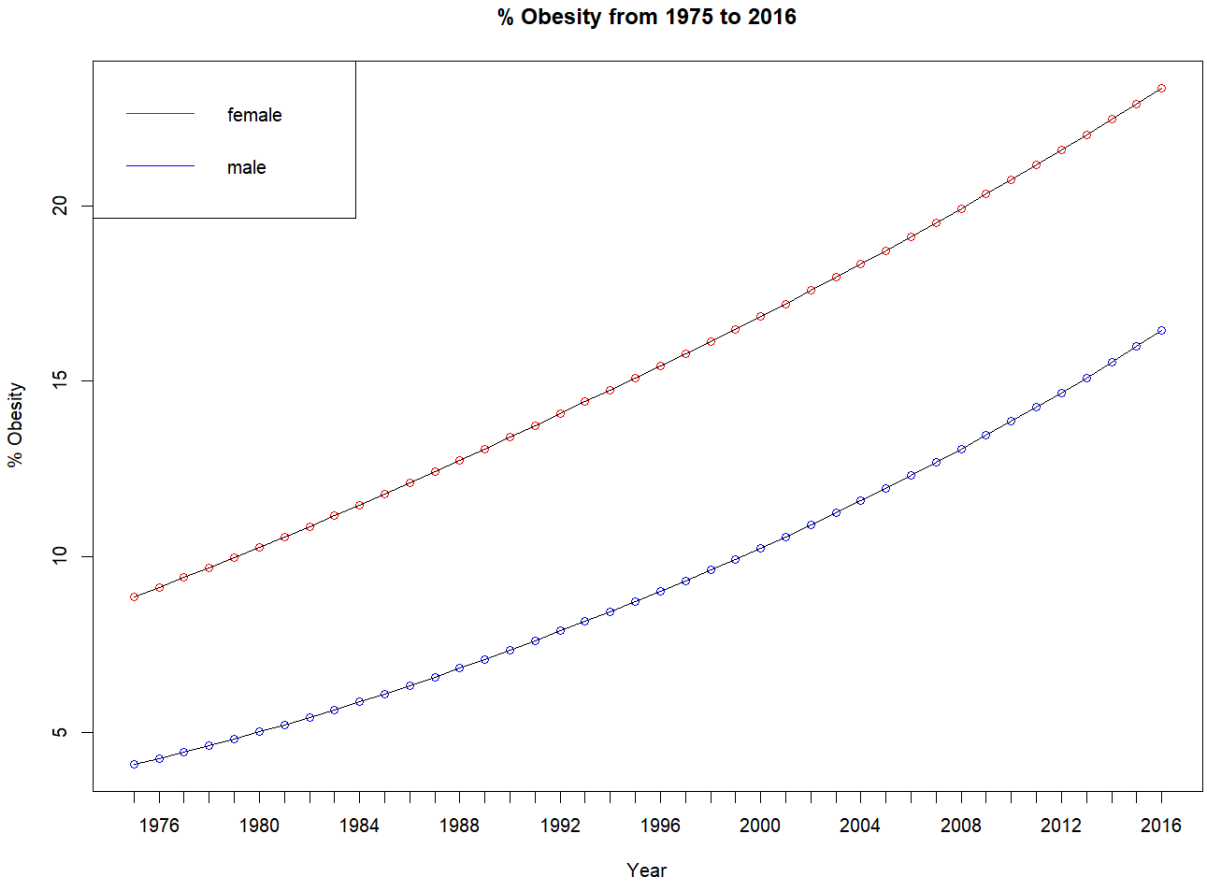


Figure 1. Obesity Trend of Females and Males from 1975 to 2016 Worldwide

Both genders trend the same as females and males since it is basically the average of both. Here we can see the upward trend of obesity worldwide more clearly which is a very persistent problem.

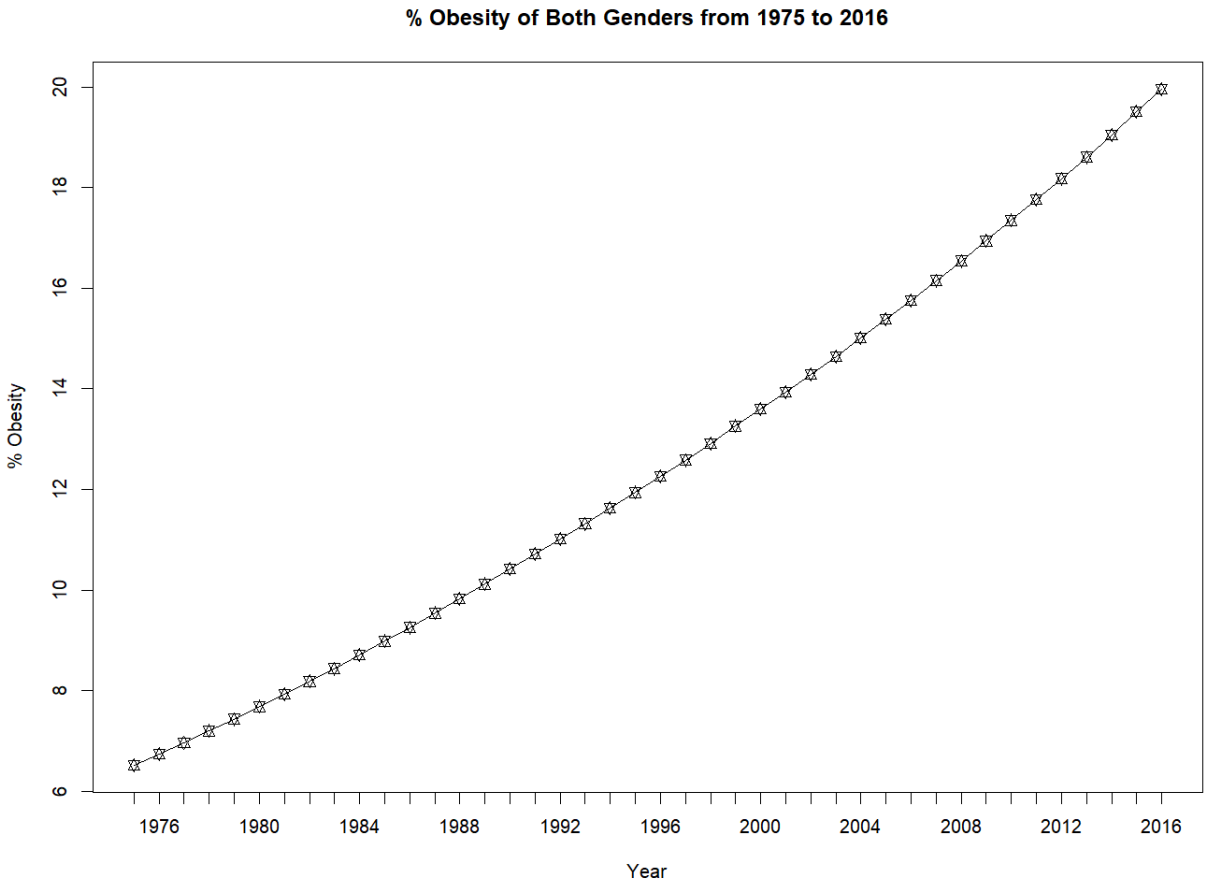
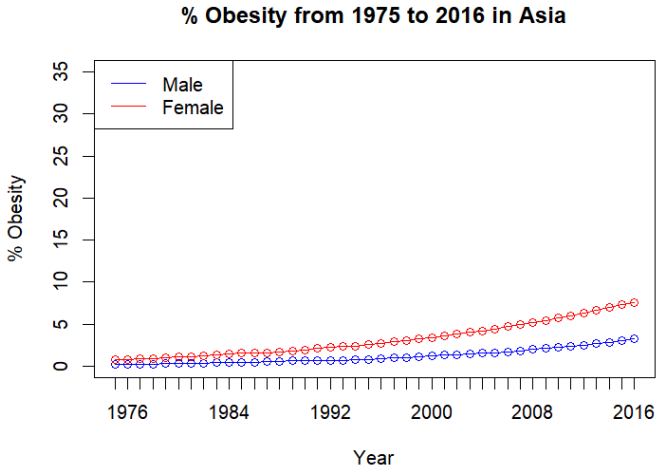
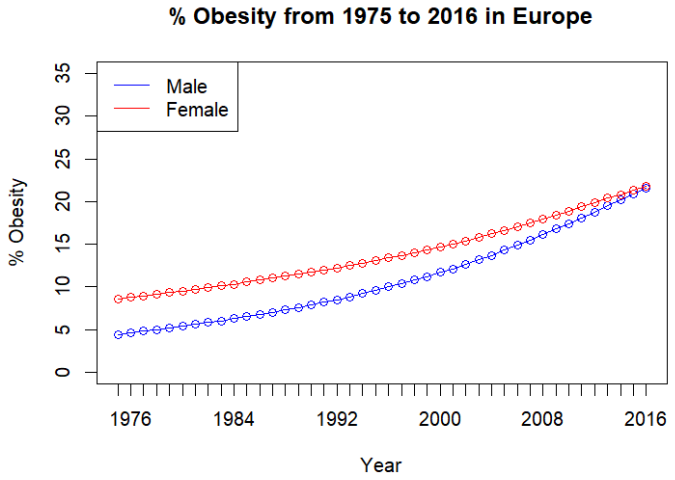


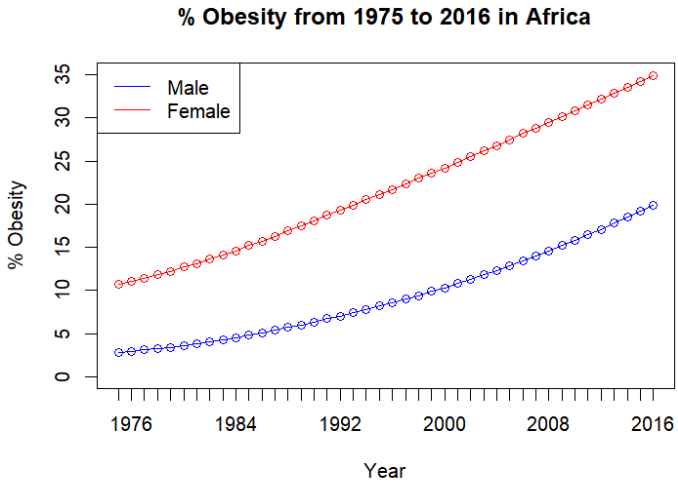
Figure 2. Obesity Trends of Both Sexes from 1975 to 2016 Worldwide

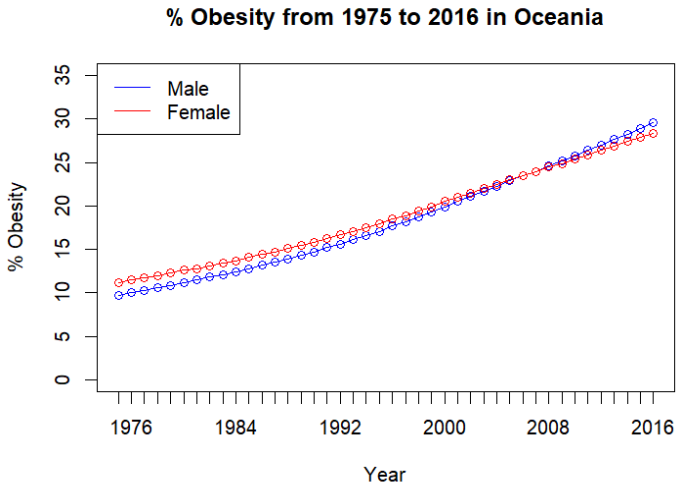
Obesity Classification

Asia: The percentage of obese people started the same for both males and females but from time to time, the obesity trend for females started to increase more drastically compared to males.
Europe: From the graph below, the percentage of obesity for females was higher than for males by around 5%, but they started to converge as time went by.
Africa: The percentage of obesity for females was higher than for males at the start and increased more rapidly as well, further extending the gap between the genders.
Oceania: The obesity trend lines of both genders are almost aligned with each other with the males surpassing the females at around the year 2007-2008.
South/North America: Both males and females have seen a significant increase in obesity rates over the past few decades.









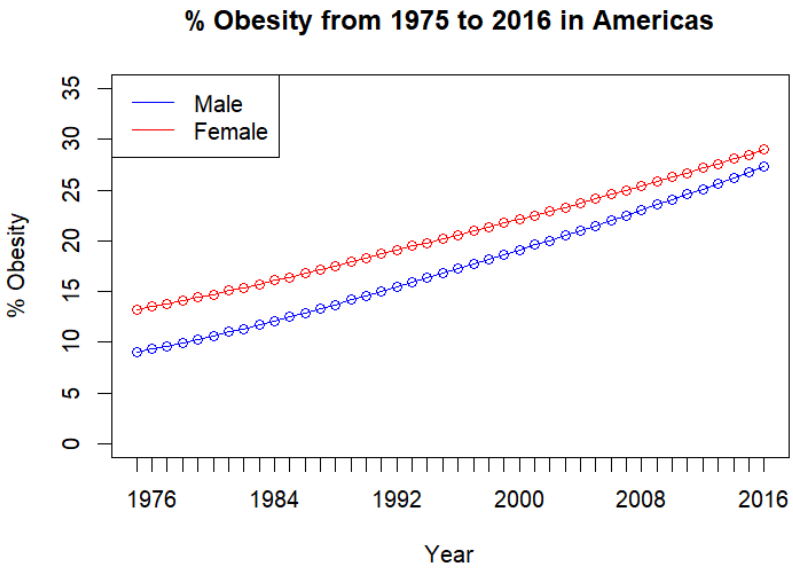


Figure 3-7. Obesity Trend of Females and Males from 1975 to 2016 Divided by Continents

The obesity trend for both females and males in Vietnam from 1975 to 2016 shows a notable increase over the decades. This rise aligns with global trends but the increase is rather small compared to other countries and the starting point of obesity was almost at 0% for both genders

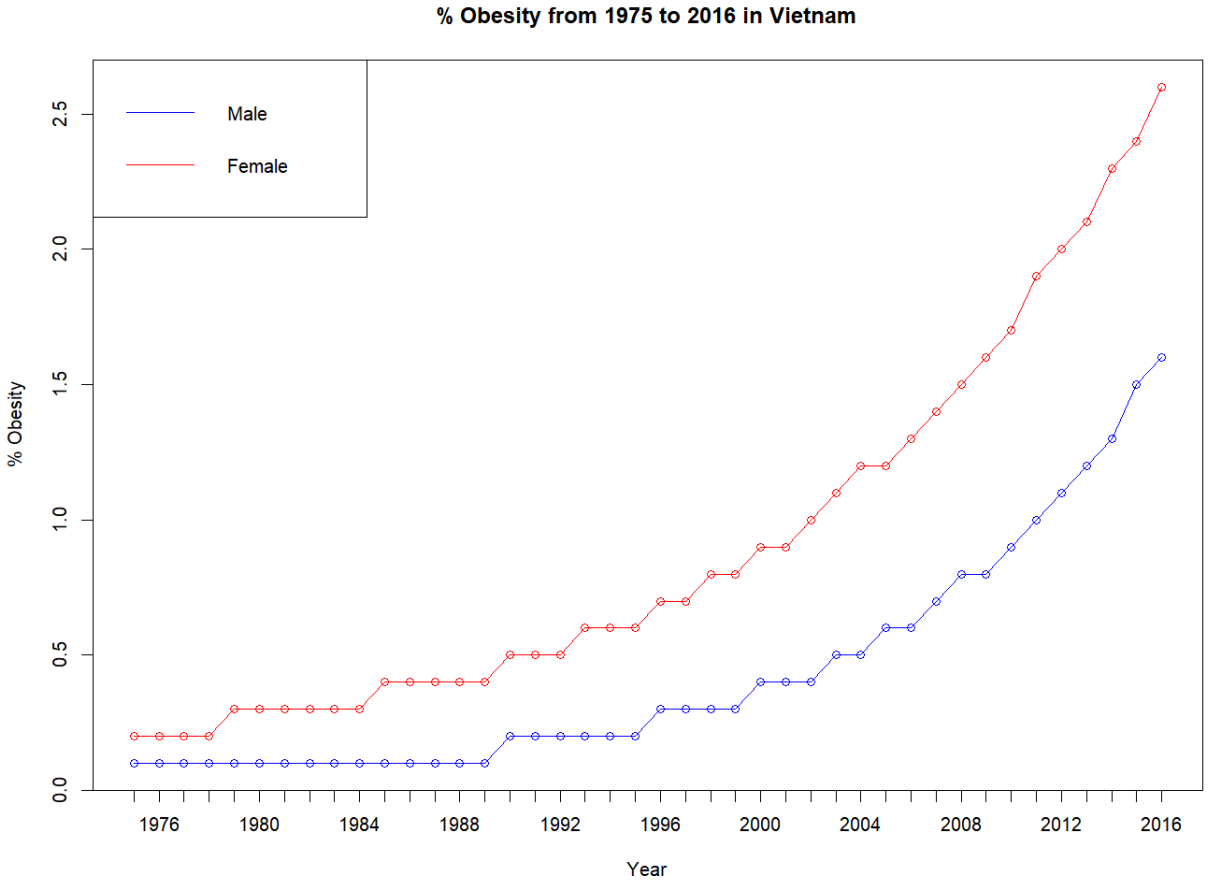


Figure 8. Obesity Trend of Females and Males from 1975 to 2016 in Vietnam

III. Obesity Classification Dataset

1. Data Observations

The dataset encompasses a wide demographic in terms of age, gender, and obesity categories, making it suitable for analyzing factors influencing obesity. This consists of 1,000 entries and 7 columns, with no missing values in any of the columns:

- **Age:** Integer values represent the age of the individuals from 18 to 79 years.
- **Gender:** Categorical data with two categories - 'Male' and 'Female'.
- **Height:** Float values representing the height of individuals in centimeters.
- **Weight:** Float values representing the weight of individuals in kilograms.
- **BMI:** Float values representing the Body Mass Index of individuals.

- **PhysicalActivityLevel:** Integer values ranging from 1 to 4, indicating the level of physical activity.
- **ObesityCategory:** Categorical data with categories indicating the obesity classification, such as 'Normal weight', 'Obese', and 'Overweight'.

Statistical Summary of Numerical Columns
Age: Ranges from 18 to 79 years with a mean of approximately 49.86 years.
Height: Varies between 136.12 cm and 201.42 cm, with an average height of 170.05 cm.
Weight: Ranges from 26.07 kg to 118.91 kg, with a mean weight of 71.21 kg.
BMI (Body Mass Index): Ranges from 8.47 to 50.79, with a mean BMI of 24.89, indicating a mix of underweight to obese categories within the dataset.
Physical Activity Level: Levels range from 1 to 4, with a mean value of 2.53, suggesting a varied distribution of physical activity levels among individuals.

Gender Distribution
Male: 523 (52.3%)
Female: 477 (47.7%)

Obesity Category Distribution
Normal weight: 371 individuals
Overweight: 295 individuals
Obese: 191 individuals
Underweight: 143 individuals

2. Data Analysis

Figure 9 overlaid with a density plot provides insight into the age distribution of the study population. The histogram's bin width is set to 5 years, allowing for a detailed view of age distribution within the population. The majority of the participants are distributed around the middle age ranges, with notable peaks suggesting higher frequencies of participants in these age categories. The density plot, with its smooth curve, indicates the distribution pattern and shows that there are two age groups particularly prevalent in the dataset. It is important to note that there is a tailing off in the number of participants in the older age categories.

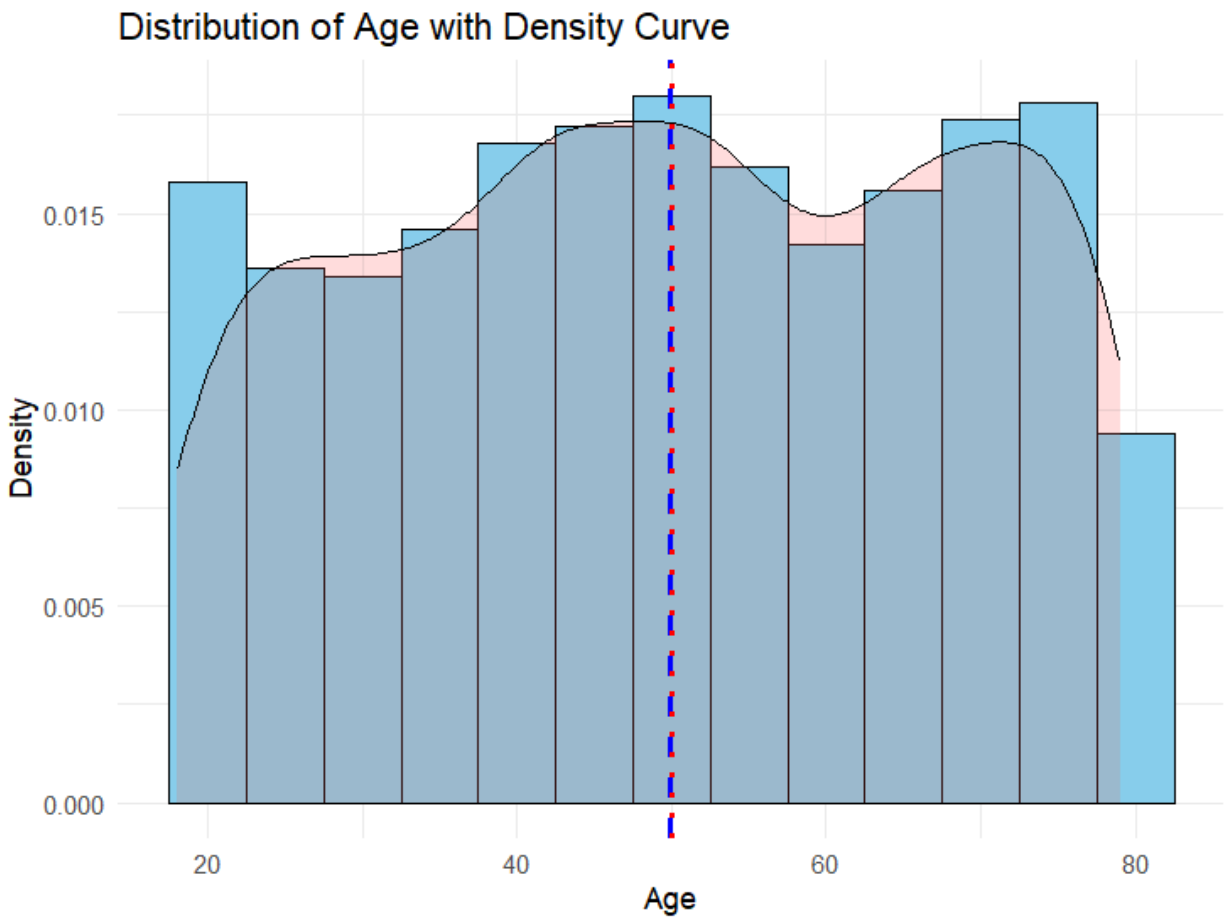


Figure 9. Distribution of Age with Density Curve

Figure 10 illustrates the distribution of BMI across different age ranges, separated by gender. The data is categorized into seven age ranges, allowing for a comparison of BMI distribution between males and females across different stages of life. For most age ranges, the median BMI appears to be consistent between genders, with a slight increase in median BMI in older age ranges. The distribution also shows a wide range of BMI values within each category, indicated by the spread of the box and whiskers, as well as outliers that may represent individuals with particularly high or low BMI values.

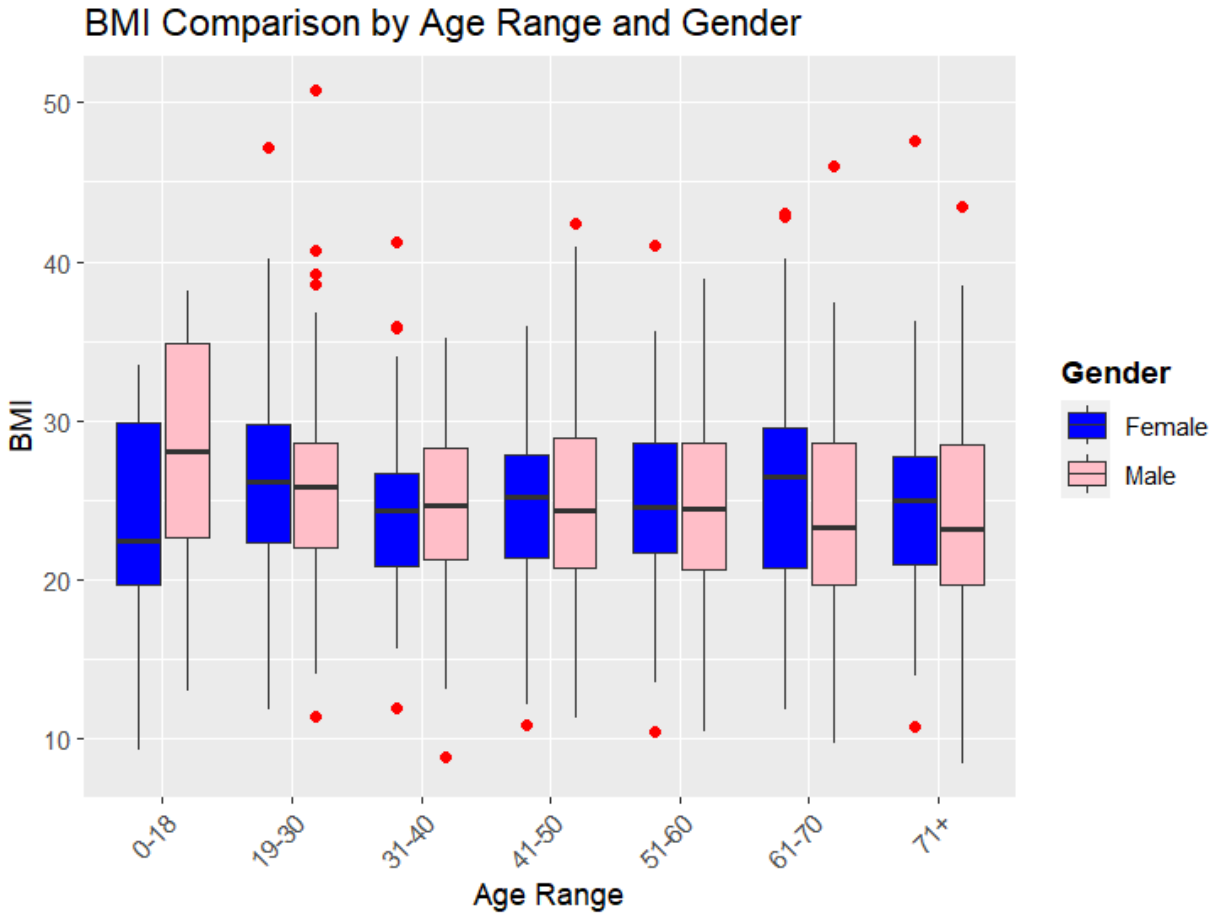


Figure 10. BMI Comparison by Age Range and Gender

Figure 11 categorizes individuals by their self-reported physical activity level, ranging from 1 (least active) to 4 (most active), and compares these groups in terms of BMI. The plot reveals that there is a wide spread of BMI values within each activity level category, with no clear trend indicating a lower median BMI for higher activity levels. However, it is noteworthy that the highest level of physical activity (4) has a slightly lower median BMI than the other categories. This plot suggests that while there is variation in BMI across physical activity levels, the relationship is not straightforward and warrants further investigation to understand other contributing factors.

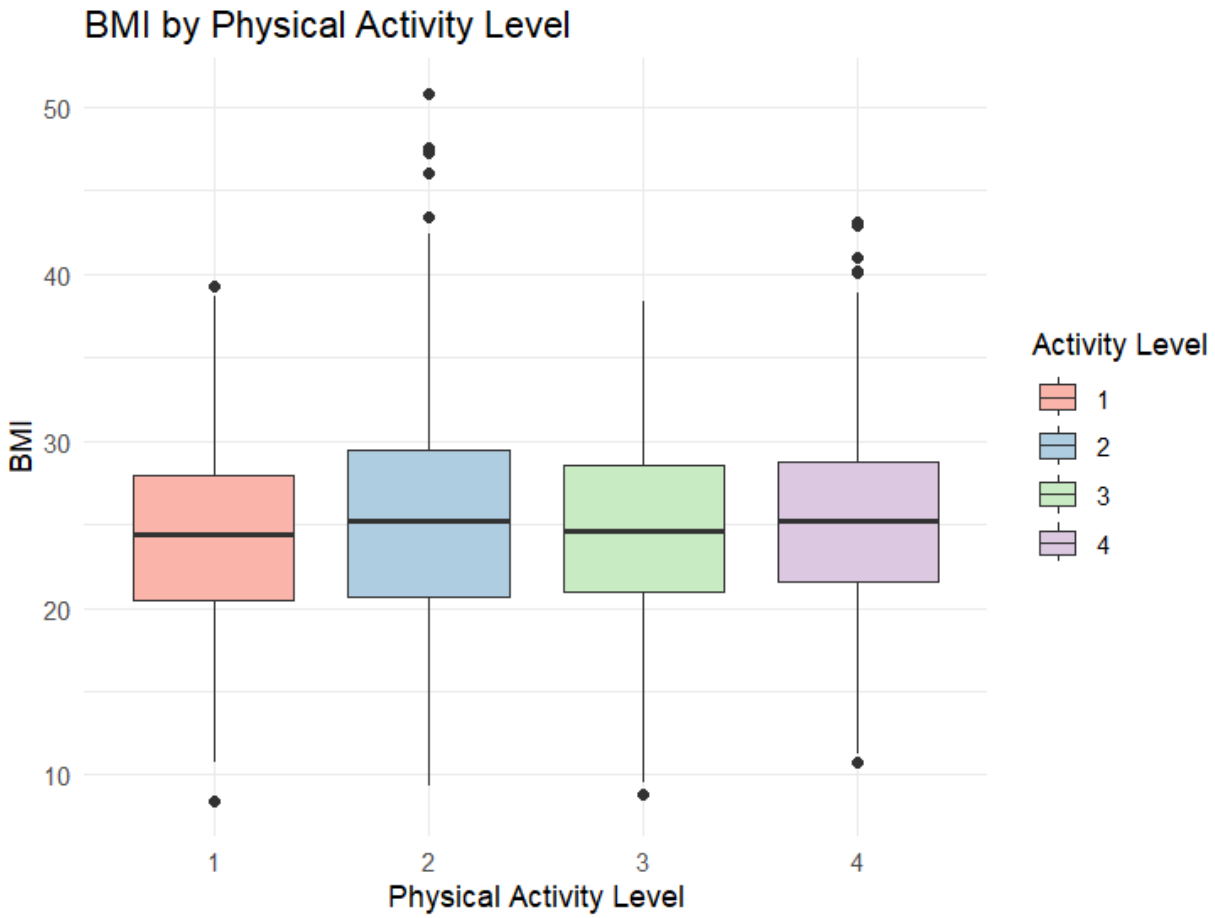


Figure 11. BMI by Physical Activity Level

The histogram below (Figure 12) illustrates that while there is some overlap between the categories, especially between 'Normal weight' and 'Overweight'—each category has a distinct peak that corresponds to the typical BMI ranges associated with those labels. The graph provides a clear visualization of how BMI is distributed within this dataset and underscores the differences between each obesity classification.

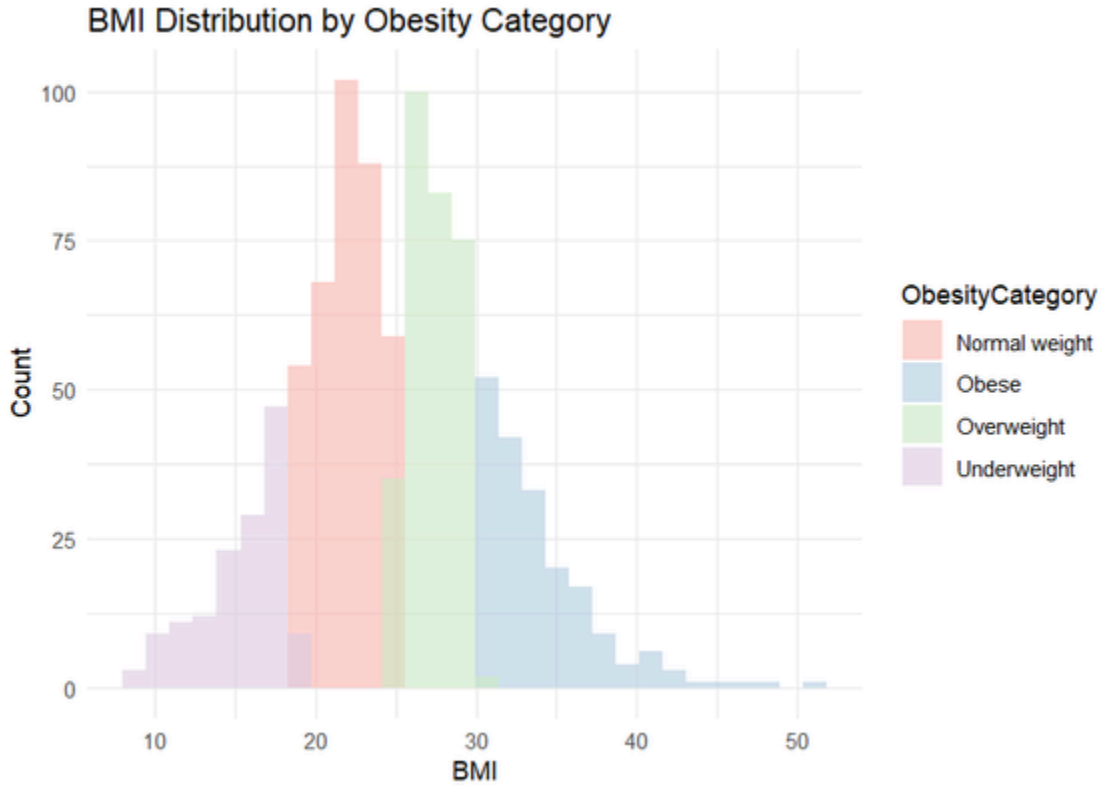


Figure 12. BMI Distribution by Obesity Category

IV. Classification Model

1. Random Forest

Random Forest is like a team of decision-makers (trees) where each makes a decision based on the information given (like age, weight, and height). They all vote, and the most common vote is the final decision. This method works well for guessing categories, like figuring out if someone is underweight, normal weight, overweight, or obese based on their characteristics.

2. Random Forest in Obesity Classification

In this assignment on obesity classification, we used Random Forest to understand how different factors, such as a person's age, gender, height, weight, how active they are, and body mass index (BMI), can tell us about their obesity level.

First, we made sure our data was ready to be used by turning categories like gender into numbers so the Random Forest could understand it. We also split our data into two parts: one part to teach the model (training data) and another to check how well it learned (test data).

We then trained our Random Forest model with the training data. We set some limits (like the number of trees and how deep they can go) to make sure it learns well without taking too much time or getting confused by too many details.

After training, we checked how good the model was by using the test data. We looked at its accuracy (how often it was right) and used a confusion matrix, which helps us see where it's doing well and where it could improve.

We can also see how the model had become after the training through this graph below.

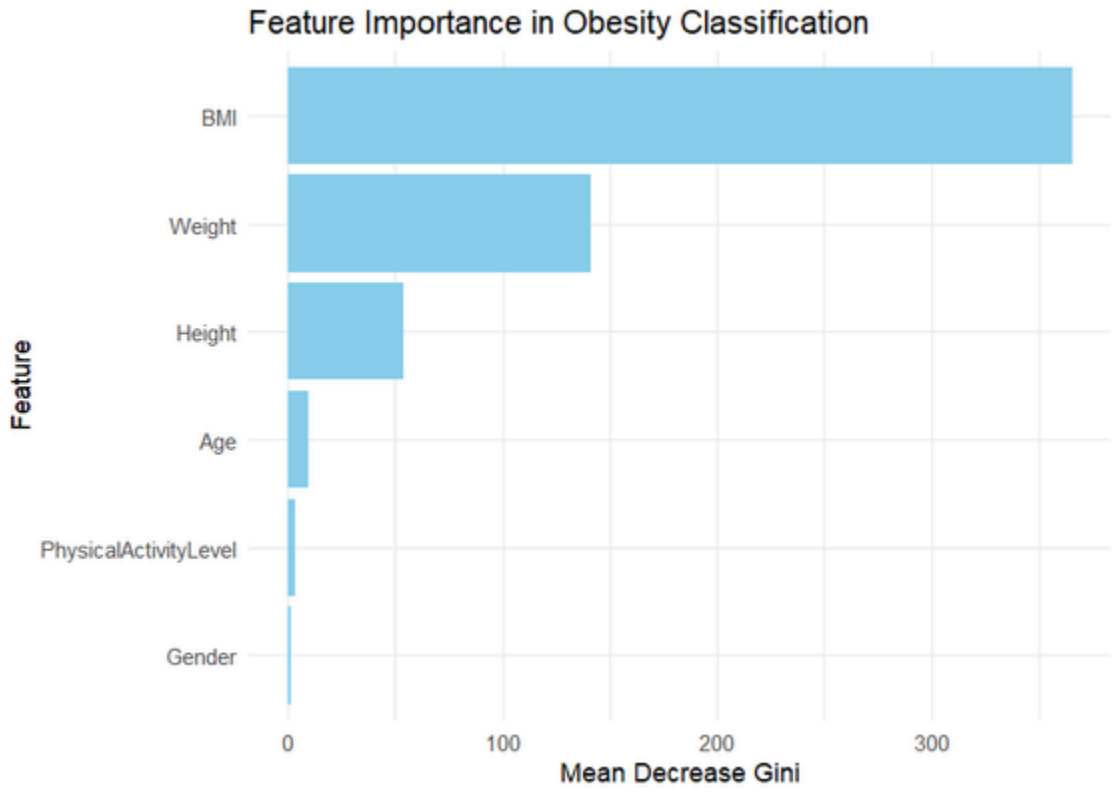


Figure 13. How are Features Weighted after Training

The 'BMI' feature has the highest mean decrease in Gini, which indicates that it is the most significant predictor of obesity in the model, followed by Weight, Height, Age, Physical Activity Level then Gender respectively.

Finally, we used our trained model to predict the obesity levels of new individuals based on their information. This shows how our model can be applied in real life to help identify people's health risks based on their physical stats and activity levels.

Using Random Forest for this task is great because it's accurate and can handle many different information. It's a powerful way to predict obesity levels, which can help doctors and health workers better understand and help their patients. By learning from a wide range of data, our model can contribute to healthier lifestyles and better health planning.

V. Evaluation

After the training part, we can now evaluate the model.

The model has achieved an accuracy of 100%, which is quite remarkable. This means that all of the predictions made by the model align with the actual data.

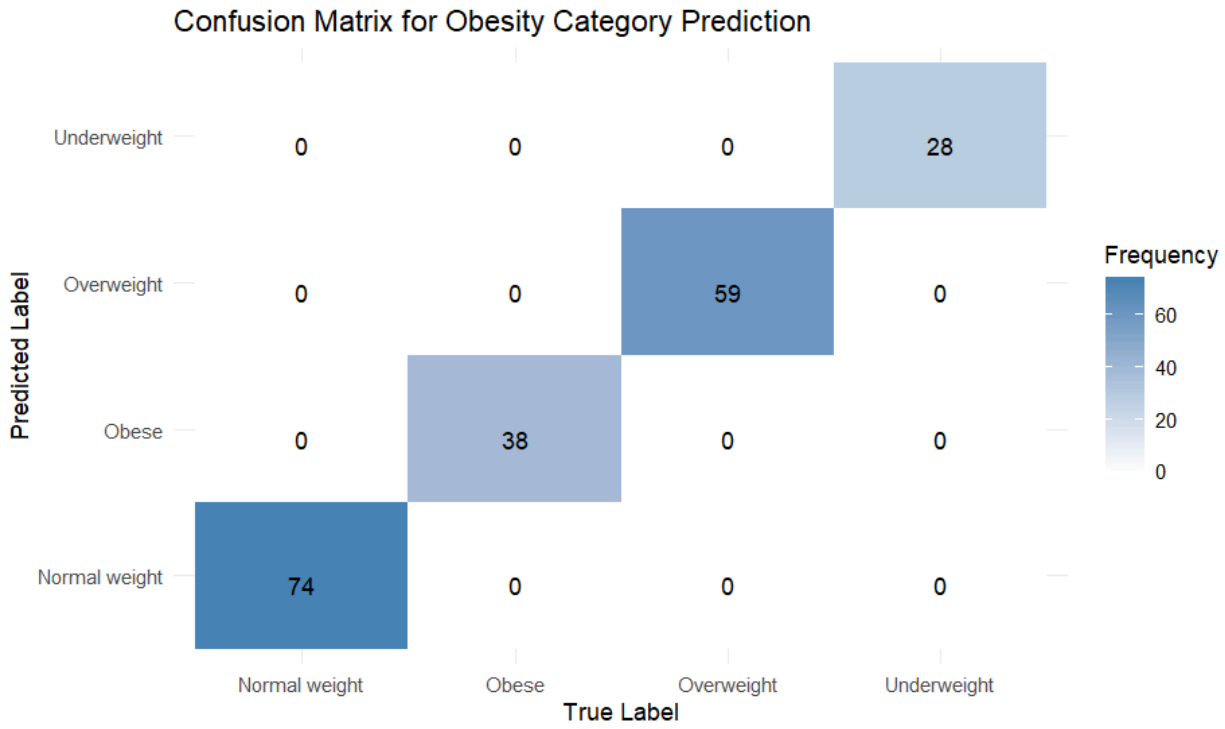


Figure 14. Confusion Matrix for Obesity Category Prediction

It perfectly identified all individuals in the first and second groups without any errors.

Diving deeper, our model not only boasts high sensitivity and specificity across the board, but it's also consistently excellent in its predictive values. It correctly identifies the presence of a condition (sensitivity) and the absence of a condition (specificity) with high reliability.

The statistical tests underline the model's capability with a p-value less than 2.2e-16, indicating that the model's accuracy is significantly better than a random guess.

The Kappa score of 0.986 reinforces the model's accuracy, signifying a high level of agreement between the predictions and the actual values.

Obesity Classification

We also tested the model on a custom dataset given by our lecturer Mr.Nguyen Xuan Thanh, more specifically:

- Mr.Thanh: Height = 179 cm; Weight = 106 Kg;
- Mr.Tung (SA): Height = 175 cm; Weight = 98 Kg;
- Mr.Quan (SA): Height = 180 cm; Weight = 90 Kg.

Here are our results on the dataset given above:

Table: Predicted Obesity Categories for Individuals

Age	Gender	Height	Weight	PhysicalActivityLevel	BMI	PredictedCategory
30	Male	179	106	1	33.08261	Obese
30	Male	175	98	1	32.00000	Obese
30	Male	180	90	1	27.77778	Overweight

> |

From the results above, we classified:

- Mr.Thanh in the Obese class
- Mr.Tung (SA) in the Obese class
- Mr.Quan (SA) in the Overweight class

VI. Conclusion

The upward trend in obesity rates from 1975 to 2016 paints a concerning picture of global health, highlighting the significant challenge that fast food and sedentary lifestyles pose to maintaining a healthy weight. In Vietnam, similar to global trends, obesity rates have shown an alarming increase, signaling a need for targeted health interventions and policy actions.

Through our analysis, we have identified key trends and patterns that contribute to obesity rates across different demographics. By comparing the trends between men and women globally, and within Vietnam, we have gathered evidence that will help in crafting tailored strategies to combat obesity.

Our computational model, focused on predicting obesity among teachers, represents a microcosm of the larger issue. The patterns we have uncovered not only enhance our understanding of obesity within this specific professional group but also serve as a beacon for developing preventive measures. It's clear from our findings that obesity does not occur in isolation; it is a complex interplay of lifestyle choices and environmental factors.

As we conclude this report, we hope that the insights gained will be a stepping stone towards more effective interventions. The statistical evidence underscores the urgency for educational campaigns and systemic changes to encourage healthier living. By leveraging the power of predictive modeling, stakeholders can identify at-risk groups and provide them with the necessary support to lead healthier lives.