Influential Factors Determining Obesity and Fitness

1st Gaurav Gupta
dept. of Computing
National College of Ireland
Dublin, Ireland
x22212311@student.ncirl.ie

2nd Vipin Sharma
dept. of Computing
National College of Ireland
Dublin, Ireland
x22207406@student.ncirl.ie

3rd Himani Sharma
dept. of Computing
National College of Ireland
Dublin, Ireland
x22224815@student.ncirl.ie

***Abstract*—This project design report states about the crucial factors affecting the obesity of an adult. There are multiple factors which we look deeply into such as family history of being overweight, smoking and alcoholic habits, eating habits, physical activities. This can be very useful for fitness trainers, the fitness industry for personalized training and recommending the products accordingly. We will use algorithms such as Random Forest and KNN(k Nearest Neighbour) for categorizing obesity level based on the data.**

***Index Terms*—Obesity, Health, Data Visualization**

I. Introduction and Dataset Description

A medical condition known as obesity is defined by a large amount of body fat to the point where it can be dangerous for a person's health. According to the World Health Organization (WHO) [1]"The definition of Overweight and obesity are the inappropriate or excessive accumulation of fat that may have negative health effects". Nowadays the obesity is the most important health problem around all over the world. Many other diseases are developing with obesity like heart disease, diabetes, and others which is not good for our good health. [2]A survey was taken in 2022 of a population from 34 countries and founded that the obesity problem was ranked fifth behind cancer, stress, mental health, covid 19 and heart attack. The main issue of childhood and teenage obesity is concerning because overweight children are more likely to be obese than adults who have disabilities or die young. The only way to avoid the obesity problem is by doing physical exercise, eating a healthy and balanced diet in your meals. The obesity depends on various different features which we describe in the dataset description section II. Many different techniques have been studied to identify the obesity trends and early detection can lead to quick obesity prevention. The author F. H. Yagin et al. [3]developed the Neural network-based classification model to predict obesity levels based on their physical activity levels and eating habits. The main goal of our study is to understand and use the obesity dataset that we downloaded from the UCI machine learning repository and to create prediction models by using a machine learning algorithm that predicts the obesity level of a person depending on the various parameters. By using this model the person takes immediate action to care of their health.

A. Dataset Description

Introduction This dataset was downloaded from the UCI machine-learning repository in the CSV format and the dataset contains the details for the estimation of obesity levels in individuals from the countries of Colombia, Mexico, and Peru with different age groups and also based on their physical conditions and eating habits. The dataset has 17 columns and 2112 rows which are shown in Figure 1. The columns have the data which is related to the eating habits and regular routine of a person in Colombia, Mexico, and Peru. The attributes of the dataset are define below:

	Gender	Age	Height	Weight	FHWOW	FAVC	FCVC	NCP	CAEC	SMOKE	CH2O	SCC	FAF	TUE	CALC	MTRANS	NObeyesdad
0	Female	21.0	1.62	64.0	yes	no	2.0	3.0	Sometimes	no	2.0	no	0.0	1.0	no	Public_Transportation	Normal_Weight
1	Female	21.0	1.52	56.0	yes	no	3.0	3.0	Sometimes	yes	3.0	yes	3.0	0.0	Sometimes	Public_Transportation	Normal_Weight
2	Male	23.0	1.80	77.0	yes	no	2.0	3.0	Sometimes	no	2.0	no	2.0	1.0	Frequently	Public_Transportation	Normal_Weight
3	Male	27.0	1.80	87.0	no	no	3.0	3.0	Sometimes	no	2.0	no	2.0	0.0	Frequently	Walking	Overweight_Level_I
4	Male	22.0	1.78	89.8	no	no	2.0	1.0	Sometimes	no	2.0	no	0.0	0.0	Sometimes	Public_Transportation	Overweight_Level_II
5	Male	29.0	1.62	53.0	no	yes	2.0	3.0	Sometimes	no	2.0	no	0.0	0.0	Sometimes	Automobile	Normal_Weight

Fig. 1. Dataset Description

1. Gender: It has two 2-categorical values, Male and Female.
2. Age: This column defines the age of the person in the dataset.
3. Height and Weight: This 2-column tells the height and weight of the person.
4. Family history with overweight: This is an important features that give the information that any family member suffered from being overweight previously.
5. FAVC: This Feature tells the consumption of high-caloric food frequently.
6. FCVC: This column tells whether user includes the vegetables in their meals.
7. NCP: This Feature gives information on How many meals users consume in a day.
8. CAEC: It tells the User to consume any food between the regular meals.
9. SMOKE: Any smoking habit.
10. CH2O: This column tells what is the daily water intake.
11. SCC: This attribute tells the user to monitor the calories daily when they eat any food.
12. FAF: The user performs any physical activities this will be defined in the column.
13. TUE: how much time a user spends in front of the mobile screen, television, computer, and others.
14. CALC: Frequency of alcohol consumption.
15. MTRANS: This column gives information on what mode of transportation is used by the user in the daily routine.
16. NObeyesdad: This column tells about the level of obesity.

II. Goal of the Project

The goal of this project is to get the insights of obesity for prediction of obese which can be used by personalized fitness trainers as well as in the fitness industry. This can be used to find the highly correlated features and most contributing factors for overweight. These insights can be used in the fitness industry in multiple ways such as analyzing nutritional guidance and can recommend products on the basis of number of meals taken, amount of calorie intake, how frequent the intake of vegetables. It can also help the fitness trainer to keep track of each individual by their physical activity, usage of technological devices, water intake and can give feedback based on this as well as make strategies for achieving goals. Furthermore, specific exercise plans for individual specific needs, according to their age, gender, medical situations. In Future, it can also be utilized as heart rate monitors, activity levels, sleep patterns and other important metrics.

This can also be used in the health domain by the calculation of BMI and also in the pre-determination of major diseases such as diabetes, cardiovascular diseases, joint problems. This is also an important factor in mental health issues.

III. Ethical Concerns

There are numerous ethical concerns related to obesity and its application in fitness trainers. However, the dataset is taken from the UCI Machine Learning Repository, an open-source community where data is available freely with the consent of individuals. Despite this, there should be measures in place to address the following ethical considerations:

1) **Confidentiality and Privacy**: The dataset contains personal details such as age, gender, and weight, which pose a potential risk to individuals. To address this, remove columns containing unnecessary personal information from the dataset[4].
2) **Consent of Individuals**: The dataset also includes other personal information such as family history of overweight and smoking habits. Ensure that consent is obtained from individuals to protect their mental and physical health.
3) **Gender Discrimination**: Avoid making assumptions or conclusions based on gender. Obesity is a sensitive topic, and stereotyping individuals based on gender can be harmful.
4) **Data Security and Accuracy**: Ensure proper data integrity and security checks to prevent breaches. The available data should be accurate and precise to maintain credibility.
5) **Transparency of Collected Data**: Provide transparency regarding how the data was collected. This allows for authentication and verification of the data's reliability. Assign accountability for the collected data.
6) **Transparency of Collected Data**: Provide transparency regarding how the data was collected. This allows for authentication and verification of the data's reliability. Assign accountability for the collected data.

Also, focus should be more on the positive outcome rather than focus on the negative consequences.

IV. The Business Value of the Project

There are multiple scenarios and industries which can use this type of data for business point of view -

- **Target Market:** This can be used in the healthcare and fitness industry by manufacturing and recommending products to the client[5]. This can also be used for product development of such products.
- **Personalized products for an individual:** These insights can be used by a fitness app which can analyze the data in real time and can suggest personalized advice to each individual. This can also be helpful for personalized meal plans, physical activities should be done, calorie intake as well as health lifestyle.
- **Health insurance companies:** Health care insurance can use these insights for potential risk of a person to be infected by any diseases. So, they can set a premium amount according to this for each individual. Individuals with smoking habits, drinking habits, and a higher obesity level have to pay a higher premium.
- **Prediction of health issues:** Hospitals can use these insights and analyze the data for potential diseases which a general population can get affected by it in future. So, with this they can take a potential step for betterment of them. This can also be used by creation of healthcare policies for a large general set of people.

Not only profit making organizations can use this but also researchers and academic students can use this for further development in the future and can help the society as whole.

V. Preliminary Visualisations

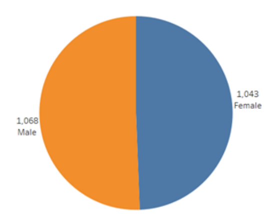

Fig. 2. Number of male and female

Fig. 2 shows the total number of males and females in the dataset. Using the pie chart visualization the yellow part shows the total number of males which is 1068 and the blue color shows the total number of females which is 1043.

The column chart in Fig. 3 shows the consumption habits of alcohol by males and females in our dataset. Both males and females show a very higher frequency **Sometimes** alcohol

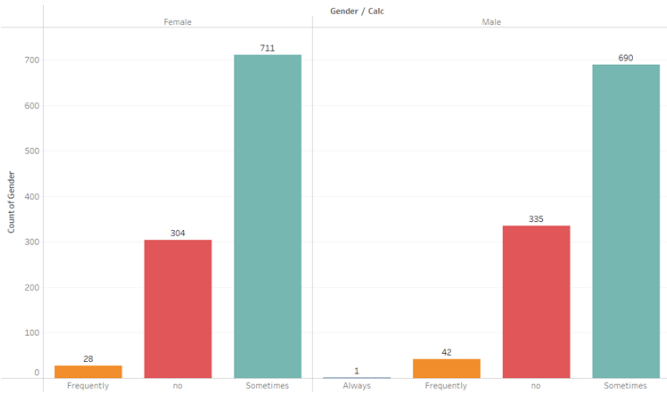

Fig. 3. How often males and females drink alcohol

drinking with 690 males and 711 females. Fewer females 304 responded **No** compared to males 335. There are a very small number of males 42 and females 28 drinking **frequently**.In the dataset, only one male reported **always** drinking and no females drinking **always**.

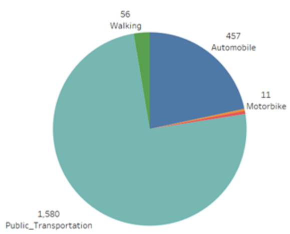

Fig. 4. Transportation used by public

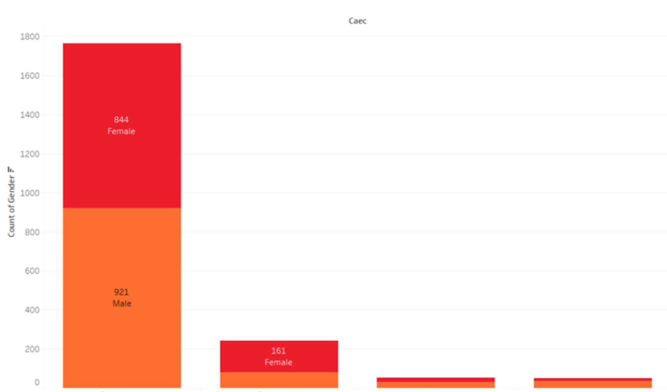

Fig. 5. Eating habits between meals

Fig. 4 indicates the different range of transportation choices by the people in the dataset. The pie chart shows that public transportation is the commonly chosen mode with 1580 males and females. 457 individuals rely on private vehicles, walking is favored by 56 people and Motorbikes are used by 11 people while 7 people use bicycles as a mode of transportation.

In the figure 5 shows the diverse eating habits between meals where 844 females and 921 males **sometimes** eat the food whereas 161 females and 81 males eat more **frequently** between meals. While 23 females and 30 males **always** consume food between meals. lastly, 15 females and 36 males do **not** eat between meals.

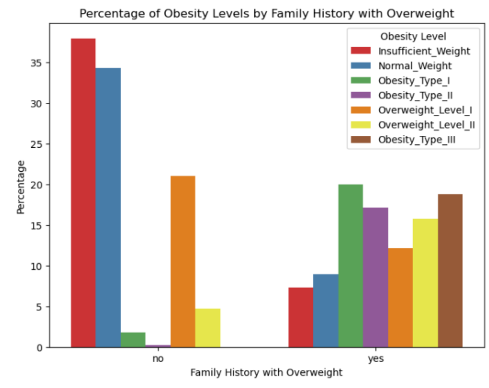

Fig. 6. Percentage of Obesity Levels by Family History with Overweight

In Fig 6, a stacked bar chart shows the percentage of obesity level by family history of being overweight. The graph illustrates that adults with family history of overweight tends to children with obesity is highly correlated specially for obesity type 1,2, and 3. This is called inter generational transmission.

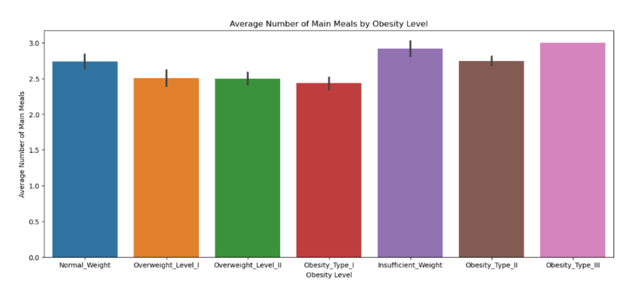

Fig. 7. Average Number of Main Meals by Obesity Level

Fig 7 shows a bar chart represents the average number of main meals by obesity level. It interprets that adults with obesity level 3 have intake of more than 3 meals. So, it can be concluded that there is a slight potential in the number of meals and obesity.

The bar chart in Fig 8 shows the average time spent using technological devices by obesity levels. The use of electronic devices also affects on our bodies and The figure clearly shows that there is not any relation between usage of multiple devices such as cell phone, video games, televisions, computer and others with obesity. And it is around 0.6 to 0.8 hours.

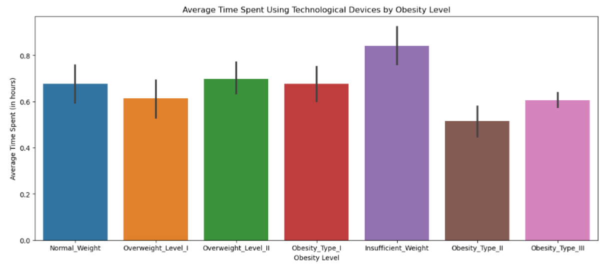

Fig. 8. Average Time Spent Using Technological Devices by Obesity Level

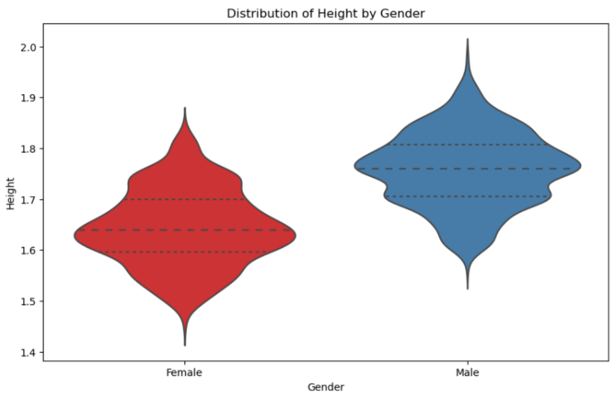

Fig. 9. Distribution of Height by Gender

Figure 9 shows a violin chart showing the distribution of heights among men and women. It is similar to a box plot where the central line describes the median and two dotted lines represent 25 percentile and 75 percentile of heights. The width of height represents a density of height. For female it is around 1.6m and for men it is 1.7m.

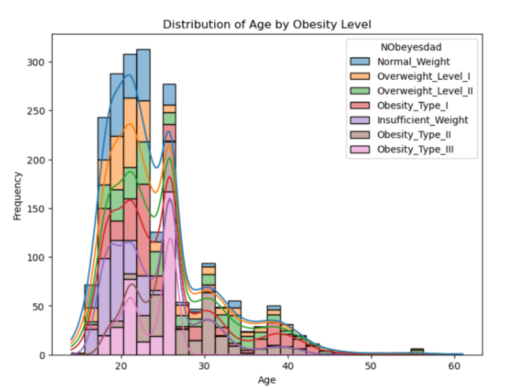

Fig. 10. Distribution of Age by Obesity Level

The Fig 10 illustrates a histogram distribution of Age by Obesity level. It illustrates that obesity type 3, is very common among 20's. While obesity level 2 is common among people in their early 30's.

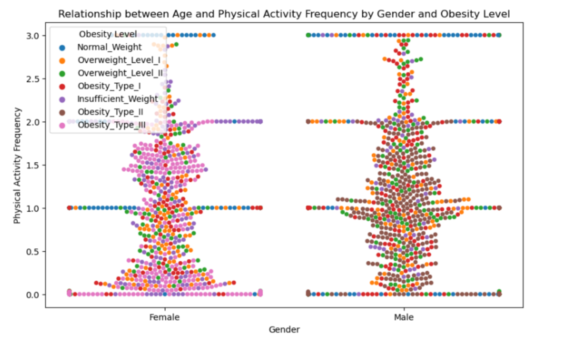

Fig. 11. Relationship between Age and Physical Activity Frequency by Gender and Obesity Level

The Fig 11 shows a swarmplot which describes a corelation of sex, how frequent of physical activity and obesity levels. Each color repreprents the obesity level corresponding to gender and age.

VI. List of Applicable Techniques

By using various data and business analytics skills, there are multiple machine learning algorithms such as Logistic Regression, KNN, and Random Forest[6] can be used for prediction, classification and recommendation.

Techniques to be considered for classifying obesity in terms of business aspect:

- Logistic Regression
- Random Forest
- K Nearest Neighbor

Tools used for implementation of the machine learning algorithms:

- Programming Language: Python
- Visualization: Power BI

References

[1] "Estimation of Obesity Levels Based on Decision Trees," IEEE Conference Publication — IEEE Xplore, May 01, 2021. https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=arnumber=9516516
[2] "Topic: Obesity worldwide," Statista, Dec. 18, 2023. https://www.statista.com/topics/9037/obesity-worldwide/topicOverview
[3] F. H. Yağın et al., "Estimation of Obesity Levels with a Trained Neural Network Approach optimized by the Bayesian Technique," Applied sciences, Mar. 18, 2023. https://doi.org/10.3390/app13063875
[4] Suzanna M. Azevedo & Lenny R. Vartania, "Ethical Issues for Public Health Approaches to obesity", Health Services and Programs, 26 June 2015
[5] Shari L. Barkin , William J. Heerman, Michael D. Warren , Christina Rennhoff, " Millennials and the World of Work: The Impact of Obesity on Health and Productivity", J Bus Psychol , 7 March 2010
[6] Shari L. Barkin , William J. Heerman , Michael D. Warren , Christina Rennhoff "Using random forest for reliable classification and cost-sensitive learning for medical diagnosis",BMC Bioinformatics , 30 Jan 2009