HAP 780

Data Mining in Healthcare

Classifying Obesity Levels based on Body Mass Index and Finding Correlations between Family History and Obesity

Preeti Ankush Bhople

Sindhu Mallala

Shubham Khaladkar

Table of Contents

[ABSTRACT 5](#_Toc121565706)

[Keywords: 5](#_Toc121565707)

[INTRODUCTION 6](#_Toc121565708)

[METHOD 6](#_Toc121565709)

[2.1 Dataset 6](#_Toc121565710)

[2.2 Variables 6](#_Toc121565711)

[2.3 Data Preprocessing 8](#_Toc121565712)

[2.4 Data Analysis 14](#_Toc121565713)

[RESULTS 17](#_Toc121565714)

[DISSCUSSION: 18](#_Toc121565715)

[CONCLUSION: 18](#_Toc121565716)

[LIMITATIONS 18](#_Toc121565717)

[BIBLIOGRAPHY 19](#_Toc121565718)

# ABSTRACT

Obesity is one of the leading issues faced by the world currently now. In the United States of America around 41.9% adult population is obese. Over 19 states in America have a 35% of adult obesity rate according to Truth of America’s Health 19th Annual Report. Many factors contribute to obesity, and it varies from person to person. Few factors to mention would be poor diet, lack of sleep, inactive lifestyle, and body weight genes. Obesity also causes comorbidities such as cardiovascular conditions and diabetes. In this study we tried to study how family history with overweight becomes an increased risk factor in obesity. This study uses the obesity dataset provided by the UC Irvine Machine Learning Repository. This dataset includes data for the estimation of obesity levels in individuals from the countries of Mexico, Peru, and Colombia, based on their eating habits and physical condition. It consists of 16 attributes with 2111 instances. These subjects' Body Mass Index (BMI) was calculated and further their obesity levels were determined based on BMI. The data was then loaded to the Weka 3.8.6 software, and classification model were built to classify the data as underweight, normal, and obese. The study also found some association between family history with overweight, consumption of high calorie food and regular consumption of alcohol with obesity. As per the results of the evaluation of different models built during this study Random Forrest and Logistic were found to best results for classifying the data into underweight, normal, and obese.

Keywords:

Obesity, Family History, Correlation, Classification, Prediction

# INTRODUCTION

Prevalence of Obesity has increased worldwide. According to WHO, by 2025, about 167 million people (adults and children) will suffer from obesity-related diseases. There is a need to identify risk factors that influence the onset of obesity in people. Family History is one of the risk factors that plays a role in predicting obesity but very few studies establish a correlation between both. (Nielsen et.al 2015). A study conducted in Delhi with 444 females age group 18-22 years showed that participants with a positive family history of obesity were comparably more obese than the other participants (Mangala et. Al 2019). However, most of the studies have focused on either childhood and adolescent or adult obesity but there are few studies that have included all the age groups.

# METHOD

## 2.1 Dataset

The data set that we are using is from [UCI Machine learning repositories.](https://archive.ics.uci.edu/ml/datasets/Estimation+of+obesity+levels+based+on+eating+habits+and+physical+condition+) The data set consists of 16 attributes with 2111 instances. The individuals belonging to the data set come from Mexico, Peru, and Colombia. The data set also includes many other interesting attributes like Family history, Calorie intake, Alcohol consumption along with demographics relevant to the study like individuals age and gender.

## 2.2 Variables

The individuals in this study belong to age 14- 61 years with an average age being 24.31 years. The available data set aided us in studying different age groups as it includes adolescents too. After a thorough examination of data, we choose independent variables such as age, gender, height, weight, family\_history\_with\_overweight, high\_cal\_intake, alcohol\_intake. Further we created dependent variables BMI (Body Mass Index) and Obesity level with the help of these independent variables.  

|  |  |  |
| --- | --- | --- |
| Attribute | Categories | Processed change |
| Gender | female | None |
| male | None |
| Age | numeric | None |
| Height | numeric | None |
| Weight | numeric | None |
| family\_history\_with\_overweight | no | 0 |
| yes | 1 |
| high\_cal\_intake | no | 0 |
| yes | 1 |
| veg\_intake | never | 0 |
| sometimes | 1 |
| always |
| num\_meals | 1 | 0 |
| 2 |
| 3 | 1 |
| 4 |
| food\_btw\_meals | no | 0 |
| sometimes | 1 |
| frequently |
| always |
| smoke | no | 0 |
| yes | 1 |
| water\_intake | 1- less than a liter | 0 |
| 2 - 2 liters | 1 |
| 3- more than 2liter |
| cal\_monitor | no | 0 |
| yes | 1 |
| phy\_activity | 0 - none | 0 |
| 1 - 1 to 2 days |
| 2- 2 to 4 days | 1 |
| 3 - 4-5 day |
| alcohol\_intake | no | 0 |
| sometimes | 1 |
| frequently |
| always |
| BMI | numeric | Calculated using height and weight |
| obesity\_levels | Underweight | 0 (Only for Association) |
| Normal |
| Overweight | 1 (Only for Association) |

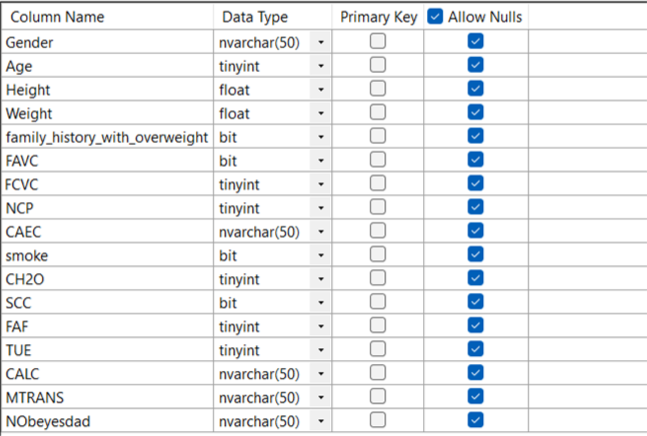
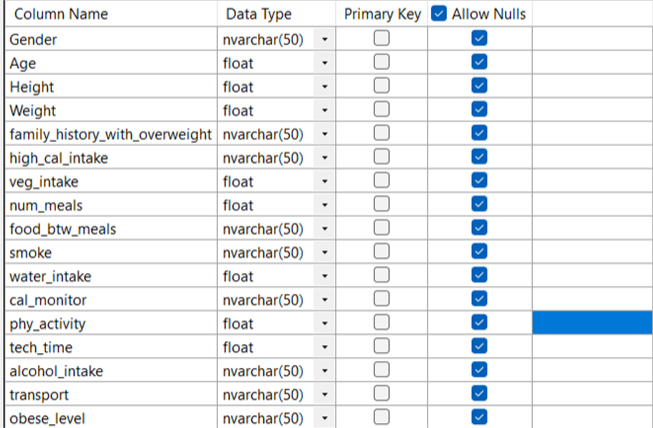
## 2.3 Data Preprocessing

The data is pre-processed in the SQL Server Management Studio (SSMS) tool. After the partial raw data (csv file) is imported in the software, the data is pre-processed using six different steps:

* Updating attributes names and their data types.
* Round-off values
* Converting categorical variables to binary form
* Evaluating BMI and drop trivial attributes
* Categorizing BMI attribute
* Join the tables – Training data

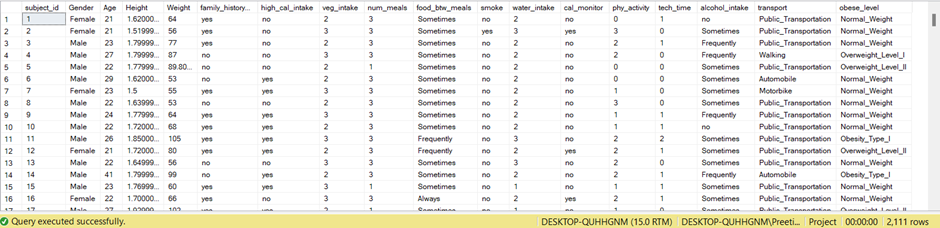
**STEP 1: Updating attributes names and their data types**

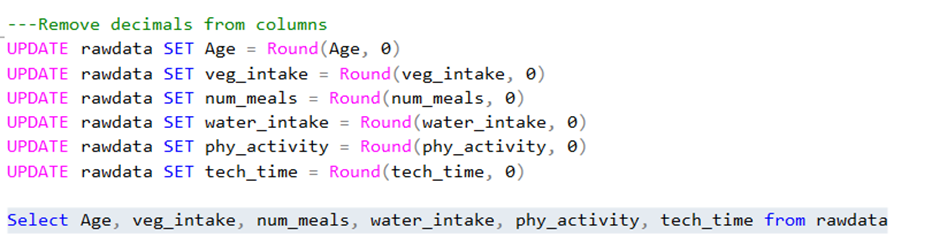
We renamed the column names to avoid ambiguity with the column names provided in the original dataset.  The dataset was imported with Nulls as the file would not be accepted without Nulls. However, when the file was imported the Nulls were replaced by random values as per the attributes category. For instance, if veg\_intake had categories such as 1, 2 and 3 the Nulls were replaced with random values between these categorical numbers.

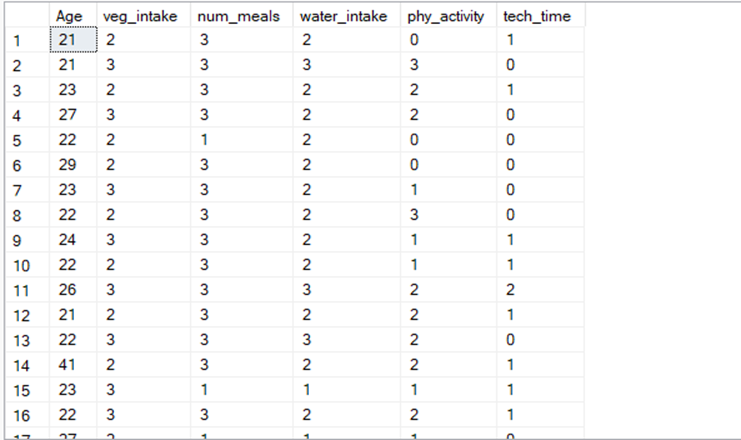
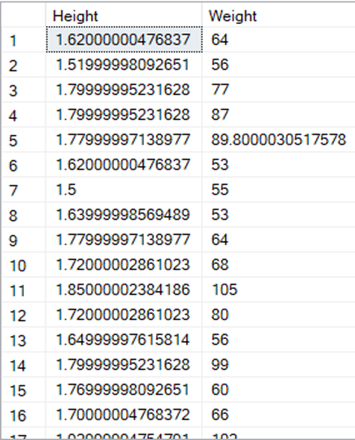
 

**STEP 2: Round-off values**

After importing dataset, most of the categorical attributes had decimal values as a result if automation during the “Import Falt File” function. To make the attributes categorical and whole we rounded the values to the nearest whole number using Round() function. This was done for Age, veg\_intake, num\_meals, water\_intake, phy\_activity, and tech\_time. However, Height and weight were rounded to the nearest two decimal places.



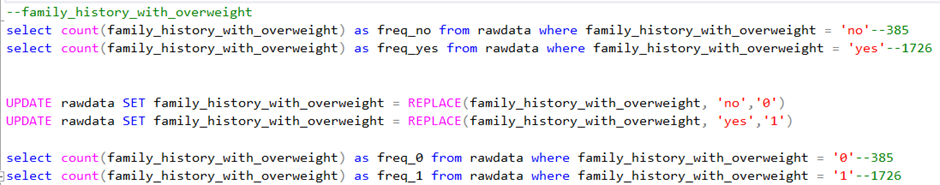


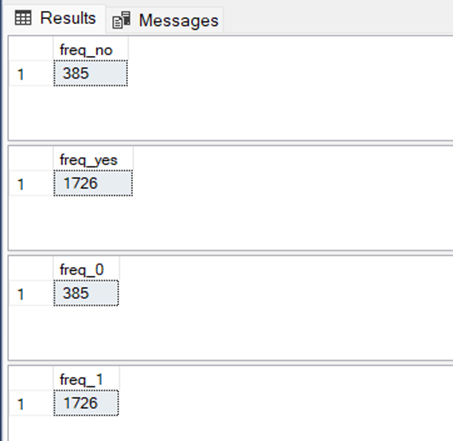
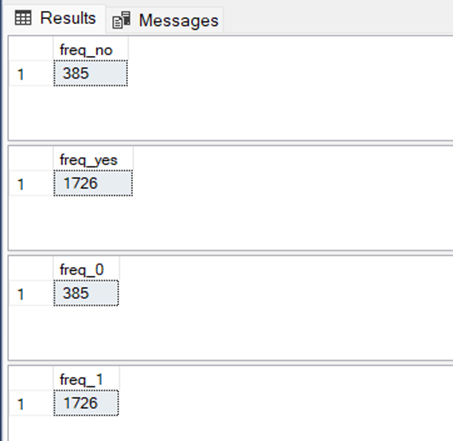
            

**STEP 3: Converting categorical variables into binary**

-----Add respective Tables of yes and no -----

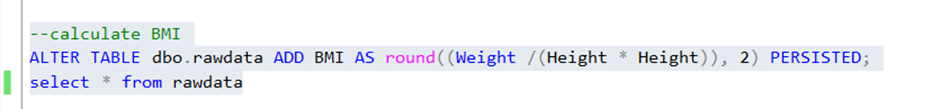
To build models in Weka and to find frequent itemset (association mining) we binarized all the categorical attributes. The attributes were binarized into ‘0’ and ‘1’. Further information is provided in the table below about how they were put into two categories.

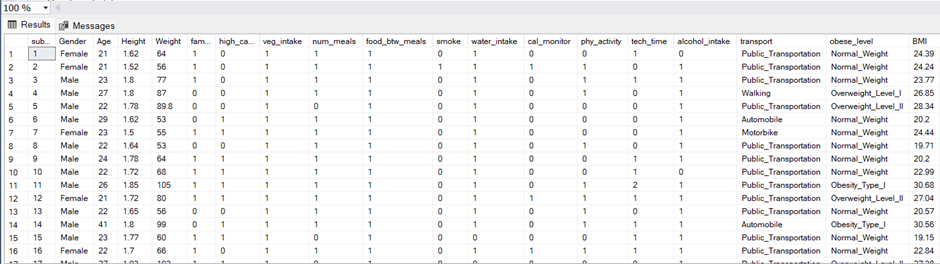


**STEP 4: Evaluating BMI and drop trivial attributes**

Body Mass Index (BMI) is a measure of overweight and obesity. BMI is calculated using the formula given below.



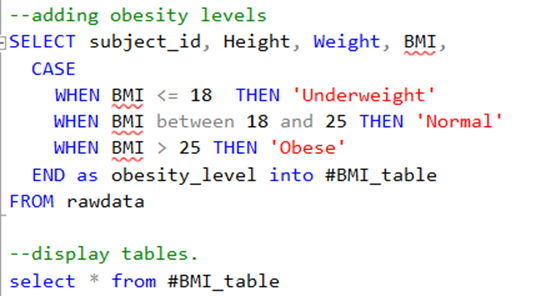
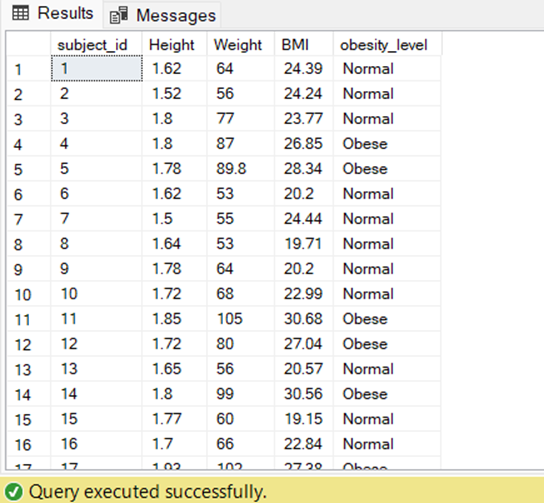


**STEP 5: Categorizing BMI attribute**

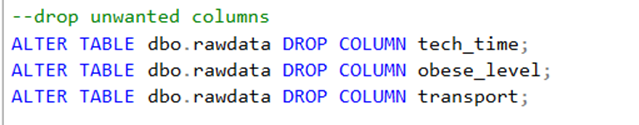
The dataset was categorizedinto three classes, Underweight, Normal and Obese. The determining factor for these were as follows:

|  |  |
| --- | --- |
| BMI | Class |
| Underweight | less than 18 |
| Normal | between 18 and 24 |
| Overweight | between 25 and 30 |
| Obese | above 30 |

The classes created were stored on a temporary table using the below query.

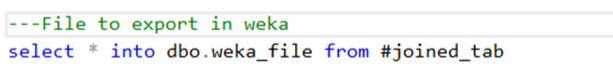
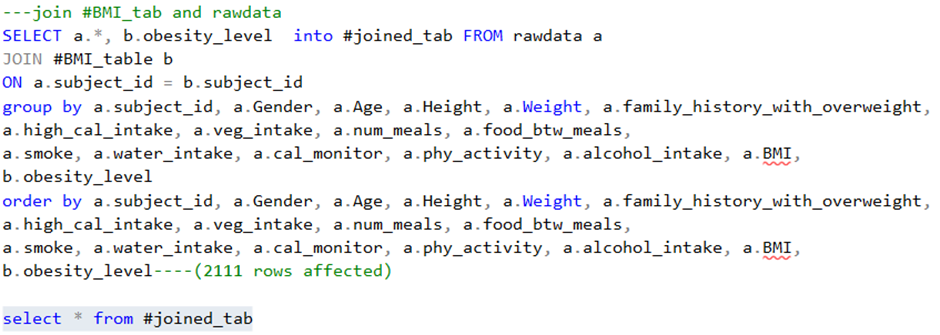
  

Further we dropped three attributes tech\_time and transport as we were not focused on it. Attribute obese\_level was dropped as we had created a new attribute obesity\_level, which was dependent on BMI.



**STEP 6: Join the tables – Training data**

As the classes of obesity\_levels were stored in a temporary table, we joined it with the raw data temple. After joining the tables, it was exported to perform analysis in Weka.



## 2.4 Data Analysis

Data Analysis and Model building was done using Weka. Using the processed dataset, 6 predictive models were built based on Percentage Split method. These 6 models were further tested on testing data.

Association Rule Mining (ARM) was also performed on the dataset. The method used here was FP-Growth method, as research question required to study the patterns among the attributes. Following are the results we found:

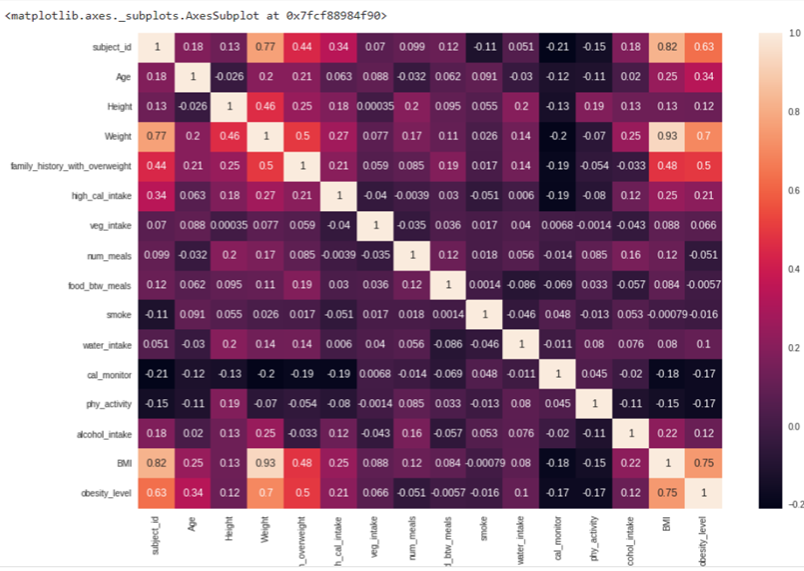
FP-Growth found 11 rules (displaying top 10)

1. **[family\_history\_with\_overweight=1, obesity\_level=1, alcohol\_intake=1]: 1054 ==> [high\_cal\_intake=1]: 1002   <conf:(0.95)> lift:(1.08) lev:(0.03) conv:(2.31)**
2. **[obesity\_level=1, alcohol\_intake=1]: 1125 ==> [high\_cal\_intake=1]: 1064   <conf:(0.95)> lift:(1.07) lev:(0.03) conv:(2.11)**
3. [high\_cal\_intake=1, obesity\_level=1, alcohol\_intake=1]: 1064 ==> [family\_history\_with\_overweight=1]: 1002   <conf:(0.94)> lift:(1.15) lev:(0.06) conv:(3.08)
4. [high\_cal\_intake=1, obesity\_level=1]: 1424 ==> [family\_history\_with\_overweight=1]: 1341   <conf:(0.94)> lift:(1.15) lev:(0.08) conv:(3.09)
5. [obesity\_level=1, alcohol\_intake=1]: 1125 ==> [family\_history\_with\_overweight=1]: 1054   <conf:(0.94)> lift:(1.15) lev:(0.06) conv:(2.85)
6. [family\_history\_with\_overweight=1, alcohol\_intake=1]: 1191 ==> [high\_cal\_intake=1]: 1114   <conf:(0.94)> lift:(1.06) lev:(0.03) conv:(1.77)
7. [obesity\_level=1]: 1539 ==> [family\_history\_with\_overweight=1]: 1439   <conf:(0.94)> lift:(1.14) lev:(0.09) conv:(2.78)
8. [family\_history\_with\_overweight=1, obesity\_level=1]: 1439 ==> [high\_cal\_intake=1]: 1341   <conf:(0.93)> lift:(1.05) lev:(0.03) conv:(1.69)
9. [obesity\_level=1]: 1539 ==> [high\_cal\_intake=1]: 1424   <conf:(0.93)> lift:(1.05) lev:(0.03) conv:(1.54)
10. [family\_history\_with\_overweight=1]: 1726 ==> [high\_cal\_intake=1]: 1580   <conf:(0.92)> lift:(1.04) lev:(0.03) conv:(1.36)

Using one of the association rule mining algorithms, the first rule of FP-Growth indicates that the factors such as family history with overweight, high obesity level, and high alcohol intake can have high calories intake with a confidence level of 95%.

To understand the correlation between attributes a heat map was produced. One can see that the family\_history\_with\_overweight had positive correlation of 0.5 with obesity\_level.

Moreover, to understand the correlations between all the attributes, correlation matrix is visualized in python seaborn using heatmap. One can see that the family history with overweight and obesity level has a positive correlation of 0.7



RESULTS

The analysis indicate that a correlation exists between the family predisposition of obesity of an individual with current state of obesity in adolescent and adult age groups.  The results of the split method can be seen in the table 1 below.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Training Set- Percentage Split Method | | | | |
| Model | **Accuracy** | **Precision** | **Recall** | **ROC Area** |
| Naïve Bayes | 94.7 | 0.949 | 0.947 | 0.997 |
| Logistic | 94.41 | 0.947 | 0.944 | 0.982 |
| SMO | 92.05 | 0.923 | 0.921 | 0.97 |
| Ibk | 89.7 | 0.896 | 0.897 | 0.923 |
| Random forest | 99.11 | 0.991 | 0.991 | 1 |
| Random tree | 97.64 | 0.976 | 0.978 | 0.988 |

Table1:  Accuracy, Precision, Recall, and ROC found in Percentage Split Method

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Testing Set | | | | |
| Model | **Accuracy** | **Precision** | **Recall** | **ROC Area** |
| Naïve Bayes | 65 | 0.64 | 0.65 | 0.825 |
| Logistic | 55.3 | 0.578 | 0.553 | 0.703 |
| SMO | 72.64 | 0.72 | 0.726 | 0.806 |
| Ibk | 80.29 | 0.797 | 0.803 | 0.909 |
| Random forest | 76.76 | 0.778 | 0.768 | 0.923 |
| Random tree | 71.47 | 0.713 | 0.715 | 0.762 |

Table2:  Accuracy, Precision, Recall, and ROC found in Testing Dataset

Table 1 shows the result of Percentage- spilt classification model. ROC for Random Forest is 1 followed by Logistic which is 0.982. Of the two Random Forest has shown better performance on training data. Also, random forest has the highest accuracy of 99.1% compared to other models. Table 2 shows results of evaluation of classification models built. For As one can notice that Logistic gave a good performance on training set, however it performed lower on test data.

As we look at evaluation results, one can see that Random Forest performed the best out of other models for evaluated.

DISSCUSSION: 

Often family predisposition of Obesity is given less importance when talking about an etiology and risk factors of obesity. Our study shows that family history plays a significant enough role in determining the level of health in an individual in their life. Knowing this factor aids people perform interventions at an early stage of their life and modify their lifestyle. We also noticed that family history of obesity combined with high calorie food intake is a major cause for obesity in the given data set. This indicates that the family history is also related to the food habits of people which affect the lifestyle of individuals and in turn affect their obesity levels. Further study can be done by adding physical activity as an attribute to verify that fact that despite of family history and high calorie intake with the help of physical activity one can modify their health status.

# CONCLUSION:

During the analysis of the data, we observed in this study that if an individual has a family history of overweight and/or obesity, he/she may be at increased risk for obesity. The study does not suggest that family history is the sole reason for obesity in a person. However, it increases their chances of being obese. We have also observed that the other lifestyle choices such as alcohol consumption and high calorie food consumption become a contributing factor in onset of obesity. If an individual is aware of his/her family history by making appropriate choices about their eating habits and maintaining a healthy lifestyle they can avoid obesity.

# LIMITATIONS

The training data set consist of ~2000 rows which is quite limited to train the model and can create a case of underfitting the model. The model ultimately cannot be robust.

The awareness of this study, from risk factors to future problems of obesity is less known to the public and hence more obesity.

 In a nutshell, other than logistic regression and random forests, the accuracy could not be optimized and lead to use of fewer resources.

# BIBLIOGRAPHY

1. World Obesity Atlas 2022. (n.d.). World Obesity Federation. <https://www.worldobesity.org/resources/resource-library/world-obesity-atlas-2022>
2. State of Obesity 2022: Better Policies for a Healthier America. (n.d.). Tfah. <https://www.tfah.org/report-details/state-of-obesity-2022/#:~:text=Nationally%2C%2041.9%20percent%20of%20adults>
3. ‌Boston, 677 H. A., & Ma 02115 +1495‑1000. (2012, October 21). Obesity Causes. Obesity Prevention Source. <https://www.hsph.harvard.edu/obesity-prevention-source/obesity-causes/#:~:text=Many%20factors%20influence%20body%20weight>
4. ‌National Heart, Lung, and Blood Institute. (2019). Assessing Your Weight and Health Risk. Nih.gov. <https://www.nhlbi.nih.gov/health/educational/lose_wt/risk.htm>
5. Mangla, A. G., Dhamija, N., Gupta, U., & Dhall, M. (2019). Familial Background as a Hidden Cause for Obesity among College Going Girls. Journal of Biosciences and Medicines, 07(04), 1–13. <https://doi.org/10.4236/jbm.2019.74001>
6. ‌Traversy, G., & Chaput, J.-P. (2015). Alcohol Consumption and Obesity: An Update. Current Obesity Reports, 4(1), 122–130. <https://doi.org/10.1007/s13679-014-0129-4>
7. ‌Corica, D., Aversa, T., Valenzise, M., Messina, M. F., Alibrandi, A., De Luca, F., & Wasniewska, M. (2018). Does Family History of Obesity, Cardiovascular, and Metabolic Diseases Influence Onset and Severity of Childhood Obesity? Frontiers in Endocrinology, 9. <https://doi.org/10.3389/fendo.2018.00187>
8. ‌Bahreynian, M., Qorbani, M., Khaniabadi, B. M., Motlagh, M. E., Safari, O., Asayesh, H., & Kelishadi, R. (2017). Association between Obesity and Parental Weight Status in Children and Adolescents. Journal of Clinical Research in Pediatric Endocrinology, 9(2), 111–117. <https://doi.org/10.4274/jcrpe.3790>
9. CDC. (2011). Body Mass Index: Considerations for Practitioners What is BMI?

https://www.cdc.gov/obesity/downloads/bmiforpactitioners.pdf