

**HAP 780**

**Data Mining in Healthcare**

**Final Project**

**TITLE: USING DATA MINING  
TECHNIQUES TO ESTIMATE  
OBESITY LEVELS**

**BY**

**HEMA LATHA. K**

**G01384405**

## **TABLE OF CONTENTS:**

**Title**

**Abstract**

**I. Introduction**

**II. Literature review**

**III. Data Mining**

**IV. DATA**

**a. Data source**

**b. Variables**

**V. Methodology: Data Preprocessing**

**a. Importing data into SQL Server and updating data types**

**b. Data cleaning**

**c. Binarization of data**

**d. Calculate BMI**

**e. Categorization of BMI**

**f. Joining the tables**

**g. Splitting the data for data analysis.**

**VI. Data analysis using WEKA.**

**VII. Results**

**VIII. Discussion**

**IX. Conclusion**

**X. Limitations**

**XI. References**

## **ABSTRACT**

### **BACKGROUND:**

Obesity is a medical condition that is characterized by excess body fat accumulation, which can lead to negative health consequences such as heart disease, diabetes, and certain types of cancer. Obesity is typically measured using the body mass index (BMI), which is calculated as an individual's weight in kilograms divided by their height in meters squared. According to the World Health Organization (WHO), individuals with a BMI of 25 or higher are considered overweight, and those with a BMI of 30 or higher are considered obese. However, it's important to note that BMI is not a perfect measure of obesity, as it does not account for factors such as muscle mass, bone density, and overall body composition. Obesity is currently a significant global issue, particularly in the United States, where approximately 41.9% of the adult population is obese. According to the 19th Annual Report from Truth of America's Health, over 19 states in the US have an adult obesity rate of 35% or higher. Various factors contribute to obesity, and these can vary from person to person. Some key factors include poor diet, inadequate sleep, sedentary lifestyles, and genetic predisposition to higher body weight.

### **OBJECTIVE:**

The main aim of this study is to detect obesity levels in association with a family history of overweight and obesity risk.

### **DATASET:**

This study uses the obesity dataset from the UC Irvine Machine Learning Repository, which included obesity information about individuals from Mexico, Peru, and Colombia. The dataset consisted of 16 attributes and 2,111 instances, and it was used to estimate obesity levels based on eating habits and physical condition.

### **METHODS:**

Cleaning of the data by removing decimals, and nulls. Then attribute coding of the data and calculated the BMI and their obesity levels were determined based on their BMI. The data set was split into training and testing with 10-fold cross-validation after features were selected using the wrapper method in WEKA software. The data was loaded into the Weka 3.8.6 software, and classification models were constructed to classify the data into underweight, normal weight, and obese categories.

### **RESULTS:**

The study revealed some associations between a family history of being overweight, Smoking, Frequency of Food intake in-between meals, consumption of high-calorie foods, regular alcohol consumption, and obesity. Different classification models like naïve bays, decision trees, random forest, and logistic regression were evaluated. In this Random Forest and Random tree models showed the best results in accurately classifying the data into underweight, normal weight, and obese categories.

### **CONCLUSION:**

This study helps to determine which model works better in determining obesity levels. Addressing obesity is crucial for reducing the prevalence of related comorbidities such as cardiovascular conditions and diabetes. By gaining insights into the factors contributing to obesity, we can work towards developing effective strategies and policies to combat this global health issue and improve overall public health outcomes.

**KEYWORDS:** Data mining, Obesity, Family history, correlation, Logistic regression, random forest method, naïve Bayes classification, Decision tree, FP-growth, prediction.

## **I. INTRODUCTION:**

Obesity is a medical condition that is characterized by excess body fat accumulation, which can lead to negative health consequences such as heart disease, diabetes, and certain types of cancer. Obesity is typically measured using the body mass index (BMI), which is calculated as an individual's weight in kilograms divided by their height in meters squared.

According to the World Health Organization (WHO), individuals with a BMI of 25 or higher are considered overweight, and those with a BMI of 30 or higher are considered obese. However, it's important to note that BMI is not a perfect measure of obesity, as it does not account for factors such as muscle mass, bone density, and overall body composition. The prevalence of obesity has been increasing globally in recent decades, with nearly 40% of adults worldwide being overweight and over 13% being obese, according to the WHO. Obesity is influenced by a complex interplay of genetic, environmental, and lifestyle factors, including diet, physical activity, sleep, stress, and socioeconomic status.

The negative health consequences of obesity are well-documented, including an increased risk of heart disease, stroke, type 2 diabetes, certain types of cancer, and other chronic health conditions. Managing and preventing obesity requires a multi-faceted approach that includes lifestyle modifications (e.g., healthy eating and physical activity), behavioral interventions, and in some cases, medical treatment.

Data mining is a process of discovering patterns, relationships, and insights from large datasets. When it comes to detecting obesity levels, data mining techniques can be applied to analyze various factors and variables associated with obesity to identify patterns and make predictions. The background of using data mining for detecting obesity levels involves collecting relevant data from individuals, such as their demographic information, lifestyle habits, dietary intake, physical activity levels, medical history, and body measurements (e.g., weight, height, body mass index). This data can be obtained through surveys, medical records, wearable devices, or other sources. Once the data is collected, data mining techniques can be applied to uncover meaningful patterns and correlations. By leveraging these data mining techniques, researchers and healthcare professionals can gain insights into the factors associated with obesity, develop predictive models, and design targeted interventions and strategies to prevent and manage obesity at both the individual and population levels.

Identifying risk factors that contribute to the onset of obesity is crucial for understanding and addressing this global health concern. Such a risk factor is a positive family history of obesity, although the correlation between the two has been insufficiently established in many studies (Nielsen et al., 2015). However, a study conducted in Delhi focused on 444 females aged 18 to 22 years and revealed intriguing findings. The participants who had a positive family history of obesity were notably more obese compared to those without such a family history (Mangala et al., 2019). This study provides valuable evidence suggesting a strong association between family history and obesity, emphasizing the need for further research in this area. By better understanding the influence of family history on obesity, healthcare professionals can develop targeted interventions to prevent and manage this prevalent health issue.

The main purpose of this study is to detect obesity levels in all age groups from 14 to 61 using data mining techniques. 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.

## **II. LITERATURE REVIEW:**

In a paper titled "Obesity level estimation software based on decision trees" by E. De-La-Hoz-Correa, F. E. Mendoza-Palechor, A. DeLa-Hoz-Manotas, R. C. Morales-Ortega, and S. H. B. Adriana, published in the Journal of Computer Science, volume 15, issue 1, pages 67-77 in 2019, the researchers addressed the issue of obesity and conducted an analysis of the disease to develop web tools. They applied the SEMMA data mining methodology, which involves selecting, exploring, modeling, and assessing the dataset. Three techniques, namely Bayesian networks, Logistic Regression, and Decision trees, were chosen for the analysis. Among these techniques, decision trees yielded the most favorable results based on various evaluation metrics, including precision, true positive (TP) rate, false positive (FP) rate, and recall. Using the WEKA data mining tool, the decision trees technique achieved an impressive precision rate of 97.4%.

In another paper "R. C. Cervantes and U. M. Palacio, "Estimation of obesity levels based on computational intelligence," Informatics Med. Unlocked, vol. 21, no. November 2020, the authors focused on the growing prevalence of obesity in children, teenagers, and adults. They proposed a computational intelligence-based approach that utilized data mining techniques, specifically Decision Trees, Support Vector Machines (SVM), and K-Means clustering. The study collected data from male and female students aged 18-25 from Colombia, Peru, and Mexico. A comparative analysis was conducted to enhance the proposed tool, and the best approach was combined with the clustering technique based on the classification results.

In a paper titled "Hossain et al. proposed a method called PRMT (Predicting Risk Factor of Obesity among Middle-Aged People Using Data Mining Techniques)". The study, published in Procedia Computer Science in 2018, employed data mining techniques to forecast the risk factors of obesity in middle-aged individuals. the authors employed data mining techniques to forecast the risk factors of obesity in middle-aged individuals in Bangladesh. They proposed the risk mining technique (PRMT) to predict obesity class-based risk factors. The evaluation and data analysis was conducted using different machine learning algorithms in WEKA software. The Naïve Bayes technique demonstrated the best results through 10-fold cross-validation.

In another paper, Singh and Tawfik developed a machine-learning approach for predicting weight gain risks in young adults. The study was presented at the 2019 10th International Conference on Dependable Systems and Services Technology. The authors recognized the importance of identifying individuals at risk of obesity and predicting BMI rates accurately and quickly, particularly for young adults. They used machine-learning techniques and experimented with multivariate regression algorithms and multi-layer perceptron feed-forward artificial neural networks (MLPFFANN). The results showed that MLPFFANN outperformed regression algorithms, achieving 90% prediction accuracy.

### III. DATA MINING:

Data mining is a process of discovering patterns, relationships, and insights from large datasets. When it comes to detecting obesity levels, data mining techniques can be applied to analyze various factors and variables associated with obesity to identify patterns and make predictions. The background of using data mining for detecting obesity levels involves collecting relevant data from individuals, such as their demographic information, lifestyle habits, dietary intake, physical activity levels, medical history, and body measurements (e.g., weight, height, body mass index). This data can be obtained through surveys, medical records, wearable devices, or other sources.

Once the data is collected, data mining techniques can be applied to uncover meaningful patterns and correlations. Here are some common approaches used in data mining for detecting obesity levels:

**Classification:** Classification algorithms can be used to build predictive models that classify individuals into different obesity categories (e.g., normal weight, overweight, obese). These algorithms use features such as age, gender, dietary habits, physical activity levels, and other variables to predict the likelihood of a person falling into a particular obesity category.

**Association Rule Mining:** Association rule mining helps identify associations or relationships between different variables. For example, it can identify patterns such as "people who consume high amounts of sugary drinks are more likely to be obese." These associations can provide valuable insights into the factors contributing to obesity.

**Clustering:** Clustering algorithms group individuals based on their similarities, allowing for the identification of distinct subgroups within a population. By clustering individuals based on variables such as lifestyle habits, dietary patterns, or genetic factors, researchers can gain a better understanding of the different factors contributing to obesity and identify specific risk profiles.

**Feature Selection:** Feature selection techniques help identify the most relevant variables or features that contribute significantly to predicting obesity levels. By reducing the dimensionality of the dataset, researchers can focus on the most informative variables and improve the accuracy and efficiency of obesity prediction models.

**Data Visualization:** Data mining results can be visualized to provide intuitive representations of the patterns and relationships discovered. Visualizations such as scatter plots, heatmaps, and decision trees can help researchers and healthcare professionals understand complex data and communicate findings effectively.

By leveraging these data mining techniques, researchers and healthcare professionals can gain insights into the factors associated with obesity, develop predictive models, and design targeted interventions and strategies to prevent and manage obesity at both the individual and population levels.

#### **IV. DATA :**

##### **a. Data Source:**

The data set that we are using is from UCI Machine learning repositories. This data is for the estimation of obesity levels in people from the countries of Mexico, Peru, and Colombia, with ages between 14 and 61 and diverse eating habits and physical conditions as mentioned, data were collected using a web platform with a survey where anonymous users answered each question, then the information was processed obtaining 17 attributes and 2111 records. The attributes related to eating habits are Frequent consumption of high-caloric food (FAVC), Frequency of consumption of vegetables (FCVC), Number of main meals (NCP), Consumption of food between meals (CAEC), Consumption of water daily (CH20), and Consumption of alcohol (CALC). The attributes related to the physical condition are Calories consumption monitoring (SCC), Physical activity frequency (FAF), Time using technology devices (TUE), and Transportation used (MTRANS), other variables obtained were Gender, Age, Height, and Weight. Finally, all data were labeled and the class variable NObesity was created with the values of Insufficient Weight, Normal Weight, Overweight Level I, Overweight Level II, Obesity Type I, Obesity Type II, and Obesity Type III, based on Equation and information from WHO and Mexican Normativity. The data contains numerical data and continuous data, so it can be used for analysis based on algorithms of classification, prediction, and association.

##### **b. Variables:**

In this study, participants ranging from 14 to 61 years of age were included, with an average age of 24.31 years. The wide age range covered in the available dataset allowed for the examination of various age groups, including adolescents. To analyze the data effectively, several independent variables were chosen, namely age, gender, height, weight, family history with overweight, high-calorie intake, and alcohol intake. These variables were selected based on their potential influence on obesity. Additionally, two dependent variables were created, namely BMI (Body Mass Index) and Obesity level, to assess the impact of the independent variables on weight-related measures. By considering these variables, the study aimed to explore the complex relationship between various factors and obesity, providing valuable insights for better understanding and addressing this health issue.

#### **V. METHODOLOGY:**

##### **DATA PREPROCESSING:**

Data preprocessing is a vital step in machine learning, as it involves preparing raw data to make it suitable for analysis and modeling. In many cases, the initial data may not be in a clean or formatted state, requiring preprocessing to handle inconsistencies, missing values, and other issues.

In the pre-processing stage of the data, the following steps were undertaken using SQL Server Management Studio (SSMS):

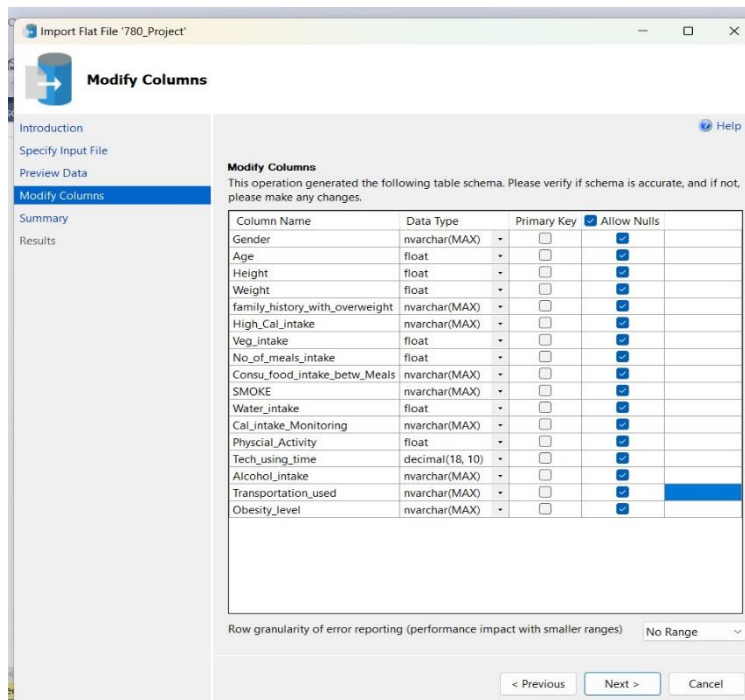
- Importing data into My SQL Server and updating their data types.

- Data cleaning - Removing the Nulls if present and Altering the table by removing Decimals.
- Binarization of data
- Calculating BMI and dropping unwanted columns
- Categorizing the BMI attribute
- Joining the tables-Training data

**a. Importing data into My SQL Server and updating their data types:**

This is the main crucial step in Data preprocessing. The dataset was initially imported with Null values, as the file could not be accepted without them. However, during the import process, these Nulls were substituted with random values based on the corresponding attribute categories. For example, if the attribute "Water\_intake" had categories 1, 2, and 3, the Null values were replaced with random values chosen from within this categorical range.

The data types within the dataset underwent alterations during the import process. This modification was essential due to certain constraints imposed by the file acceptance criteria. This data transformation was necessary to ensure that the dataset could be processed and utilized effectively for analysis in accordance with the research objectives.



**Import Flat File '780\_Project'**

**Modify Columns**

Introduction  
Specify Input File  
Preview Data  
**Modify Columns**  
Summary  
Results

**Modify Columns**  
This operation generated the following table schema. Please verify if schema is accurate, and if not, please make any changes.

Column Name	Data Type	Primary Key	Allow Nulls
Gender	nvarchar(MAX)	<input type="checkbox"/>	<input checked="" type="checkbox"/>
Age	float	<input type="checkbox"/>	<input checked="" type="checkbox"/>
Height	float	<input type="checkbox"/>	<input checked="" type="checkbox"/>
Weight	float	<input type="checkbox"/>	<input checked="" type="checkbox"/>
family_history_with_overweight	nvarchar(MAX)	<input type="checkbox"/>	<input checked="" type="checkbox"/>
High_Cal_intake	nvarchar(MAX)	<input type="checkbox"/>	<input checked="" type="checkbox"/>
Veg_intake	float	<input type="checkbox"/>	<input checked="" type="checkbox"/>
No_of_meals_intake	float	<input type="checkbox"/>	<input checked="" type="checkbox"/>
Consu_food_intake_betw_Meals	nvarchar(MAX)	<input type="checkbox"/>	<input checked="" type="checkbox"/>
SMOKE	nvarchar(MAX)	<input type="checkbox"/>	<input checked="" type="checkbox"/>
Water_intake	float	<input type="checkbox"/>	<input checked="" type="checkbox"/>
Cal_intake_Monitoring	nvarchar(MAX)	<input type="checkbox"/>	<input checked="" type="checkbox"/>
Physical_Activity	float	<input type="checkbox"/>	<input checked="" type="checkbox"/>
Tech_using_time	decimal(18, 10)	<input type="checkbox"/>	<input checked="" type="checkbox"/>
Alcohol_Intake	nvarchar(MAX)	<input type="checkbox"/>	<input checked="" type="checkbox"/>
Transportation_used	nvarchar(MAX)	<input type="checkbox"/>	<input checked="" type="checkbox"/>
Obesity_level	nvarchar(MAX)	<input type="checkbox"/>	<input checked="" type="checkbox"/>

Row granularity of error reporting (performance impact with smaller ranges) No Range

< Previous Next > Cancel



Gender	Age	Height	Weight	family_history...	high_cal_intake	veg_intake	num_meals	food_bte_meals	smoke	water_intake	cal_monitor	phy_activity	tech_time	alcohol_intake	transport	obese_level
Female	21	1.62000...	64	yes	no	2	3	Sometimes	no	2	no	0	1	no	Public_Transportation	Normal_Weight
Female	21	1.51999...	56	yes	no	3	3	Sometimes	yes	3	yes	3	0	Sometimes	Public_Transportation	Normal_Weight
Male	23	1.79999...	77	yes	no	2	3	Sometimes	no	2	no	2	1	Frequently	Public_Transportation	Normal_Weight
Male	27	1.79999...	87	no	no	3	3	Sometimes	no	2	no	2	0	Frequently	Walking	Overweight_Level_I
Male	22	1.77999...	89.80...	no	no	2	1	Sometimes	no	2	no	0	0	Sometimes	Public_Transportation	Overweight_Level_II
Male	29	1.62000...	53	no	yes	2	3	Sometimes	no	2	no	0	0	Sometimes	Automobile	Normal_Weight
Female	23	1.5	55	yes	yes	3	3	Sometimes	no	2	no	1	0	Sometimes	Motorbike	Normal_Weight
Male	22	1.63999...	53	no	no	2	3	Sometimes	no	2	no	3	0	Sometimes	Public_Transportation	Normal_Weight
Male	24	1.77999...	64	yes	yes	3	3	Sometimes	no	2	no	1	1	Frequently	Public_Transportation	Normal_Weight
Male	22	1.72000...	68	yes	yes	2	3	Sometimes	no	2	no	1	1	no	Public_Transportation	Normal_Weight
Male	26	1.85000...	105	yes	yes	3	3	Frequently	no	3	no	2	2	Sometimes	Public_Transportation	Obesity_Type_I
Female	21	1.72000...	80	yes	yes	2	3	Frequently	no	2	yes	2	1	Sometimes	Public_Transportation	Overweight_Level_II
Male	22	1.64999...	56	no	no	3	3	Sometimes	no	3	no	2	0	Sometimes	Public_Transportation	Normal_Weight
Male	41	1.79999...	99	no	yes	2	3	Sometimes	no	2	no	2	1	Frequently	Automobile	Obesity_Type_I
Male	23	1.76999...	60	yes	yes	3	1	Sometimes	no	1	no	1	1	Sometimes	Public_Transportation	Normal_Weight
Female	22	1.70000...	66	yes	no	3	3	Always	no	2	yes	2	1	Sometimes	Public_Transportation	Normal_Weight
Male	37	1.67000...	173	yes	yes	3	1	Sometimes	no	1	no	1	0	Sometimes	Public_Transportation	Overweight_Level_II

## b. DATA CLEANING:

After the dataset import, several categorical attributes were assigned decimal values due to the automated process during the "Import Flat File" function. To ensure that these attributes maintained their categorical nature and were represented as whole numbers, we applied the Round () function in combination with the ALTER function. This process involved rounding values to the nearest whole number for attributes such as Age, veg\_intake, no.of\_meals, water\_intake, phy\_activity, tech\_time, Height, and weight.

Following the rounding process, we conducted a thorough check for the presence of any Null values within the dataset. Any instances of Null values were subsequently removed, ensuring a more robust and complete dataset for analysis purposes.

```

----Data cleaning
----Removing the nulls
DELETE FROM dbo.obesity_data
WHERE Age IS NULL
OR gender IS NULL
OR Veg_intake IS NULL
OR No_of_meals_intake IS NULL
OR Water_intake IS NULL
OR Physcial_Activity is null
OR Height is null
OR weight is null --- (0 rows affected)

---Removing decimals from columns
update dbo.Obesity_data SET Age = Round(Age,0) ----(2111 rows affected)
update dbo.Obesity_data SET Veg_intake = Round(Veg_intake,0) ----(2111 rows affected)
update dbo.Obesity_data SET No_of_meals_intake = Round(No_of_meals_intake,0) ----(2111 rows affected)
update dbo.Obesity_data SET Water_intake = round(Water_intake,0) ----(2111 rows affected)
update dbo.Obesity_data SET Physcial_Activity = round (Physcial_Activity,0) ----(2111 rows affected)
update dbo.Obesity_data SET Tech_using_time= round (Tech_using_time, 0) ----(2111 rows affected)
ALTER TABLE dbo.obesity_data ALTER COLUMN Tech_using_time decimal(9,0)
ALTER TABLE dbo.obesity_data ALTER column Height decimal(9,2)
ALTER TABLE dbo.obesity_data ALTER COLUMN Weight decimal (9,0)

```

```
Select Age,Height,Weight,Veg_intake,No_of_meals_intake,Water_intake,
Physcial_Activity,Tech_using_time from dbo.Obesity_data
```

	Height	Weight
1	1.62000000476837	64
2	1.51999998092651	56
3	1.79999995231628	77
4	1.79999995231628	87
5	1.77999997138977	89.8000030517578
6	1.62000000476837	53
7	1.5	55
8	1.63999998569489	53
9	1.77999997138977	64
10	1.72000002861023	68

	Height	Weight
1	1.62	64
2	1.52	56
3	1.8	77
4	1.8	87
5	1.78	89.8
6	1.62	53
7	1.5	55
8	1.64	53
9	1.78	64

	Age	Height	Weight	Veg_intake	No_of_meals_intake	Water_intake	Physcial_Activity	Tech_using_time
1	21	1.62	64.0	2	3	2	0	1
2	21	1.52	56.0	3	3	3	3	0
3	23	1.80	77.0	2	3	2	2	1
4	27	1.80	87.0	3	3	2	2	0
5	22	1.78	89.8	2	1	2	0	0
6	29	1.62	53.0	2	3	2	0	0
7	23	1.50	55.0	3	3	2	1	0
8	22	1.64	53.0	2	3	2	3	0
9	24	1.78	64.0	3	3	2	1	1
10	22	1.72	68.0	2	3	2	1	1
11	26	1.85	105.0	3	3	3	2	2
12	21	1.72	80.0	2	3	2	2	1
13	22	1.65	56.0	3	3	3	2	0
14	41	1.80	99.0	2	3	2	2	1
15	23	1.77	60.0	3	1	1	1	1
16	22	1.70	66.0	3	3	2	2	1
17	27	1.02	102.0	2	1	1	1	0

### c. BINARIZATION OF DATA:

To facilitate the modeling process in Weka and conduct frequent itemset mining for association analysis, we transformed the categorical attributes into a binary representation. This involved binarizing the attributes into '0' and '1' categories, based on specific criteria. By binarizing the categorical attributes in this manner, we were able to transform the dataset into a format suitable for conducting association mining using frequent itemset algorithms in Weka. The table below provides further details on how the attributes were categorized:

Attributes	Categories	Processed changes
<b>Gender</b>	female	None
	male	None
<b>Age</b>	numeric	None
<b>Height</b>	numeric	None
<b>Weight</b>	numeric	None
<b>Family_history_with_overweight</b>	no	0
	yes	1
<b>High_cal_intake</b>	no	0
	yes	1
<b>Veg_intake</b>	never	0
	sometimes	1
	always	

<b>No.of_meals_intake</b>	1	0
	2	
	3	1
	4	
<b>Conc_food_intake_btw_meals</b>	no	0
	sometimes	1
	frequently	
	always	
<b>Smoke</b>	no	0
	yes	1
<b>Water_intake</b>	1- less than a liter	0
	2 - 2 liters	1
	3- more than 2liter	
<b>Cal_monitor</b>	no	0
	yes	1
<b>Phy_activity</b>	0 - none	0
	1 - 1 to 2 days	
	2- 2 to 4 days	1
	3 - 4-5 day	
<b>Alcohol_intake</b>	no	0
	sometimes	1
	frequently	
	always	
<b>BMI</b>	numeric	Calculated using height and weight
<b>Obesity_levels</b>	Underweight	0 (Only for Association)

	Normal	
	Overweight	1 (Only for Association)

Results Messages

	Gender	Age	Height	Weight	family_history_with_overweight	High_Cal_intake	Veg_intake	No_of_meals_intake	Consu_food_intake_betw_Meals	SMOKE	Water_intake	Cal_intake_Monitoring	Physical_Activity	Alcohol_I
1	Female	16	1.65	86	1	1	1	1	1	0	0	0	1	0
2	Female	17	1.55	55	0	1	1	0	1	0	1	1	0	1
3	Female	17	1.60	65	1	1	1	1	1	0	1	1	0	1
4	Female	17	1.63	65	0	1	1	0	1	0	1	0	0	0
5	Female	17	1.63	86	1	1	1	1	1	0	0	0	1	0
6	Female	17	1.65	67	1	1	1	0	1	0	1	0	0	0
7	Female	18	1.45	53	0	1	1	1	1	0	1	1	0	1
8	Female	18	1.55	56	0	1	1	1	1	0	0	0	0	0
9	Female	18	1.58	48	0	1	1	1	1	0	1	0	0	0
10	Female	18	1.60	56	1	1	1	0	1	0	1	1	0	1
11	Female	18	1.62	68	0	0	1	0	1	0	0	0	0	0
12	Female	18	1.63	63	1	1	0	1	1	0	1	1	1	1
13	Female	18	1.64	56	1	1	1	1	1	0	0	0	0	0
14	Female	18	1.81	153	1	1	1	1	1	0	1	1	1	1
15	Female	19	1.51	59	1	1	1	1	1	0	0	1	0	1
16	Female	19	1.53	42	0	1	1	0	1	0	0	1	0	1

Query executed successfully. HEMA (15.0 RTM) | HEMA\hema1 (55) | 780\_Project | 00:00

#### d. Calculate BMI :

Body Mass Index (BMI) is a metric used to estimate an individual's body fat based on their weight and height. It is calculated by dividing the weight in kilograms by the square of the height in meters. Here's the formula:

$$\text{BMI} = \text{weight (kg)} / (\text{height (m)})^2$$

-----Calculate BMI

```
ALTER TABLE dbo.obesity_data ADD BMI AS round((weight/(Height*Height)),2) persisted;
Select * from dbo.Obesity_data
```

Results Messages

	Height	Weight	BMI
1	1.62	84	24
2	1.52	56	24
3	1.80	77	24
4	1.80	87	27
5	1.78	90	28
6	1.62	53	20
7	1.50	55	24
8	1.54	53	20
9	1.78	54	20
10	1.72	68	23
11	1.85	105	31
12	1.72	80	27
13	1.65	56	21
14	1.80	99	31
15	1.77	60	19
16	1.70	66	23
17	1.63	42	16

Query executed successfully. HEMA (15.0 RTM) | HEMA\hema1 (53) | 780\_Project | 00:00:00 | 2,111 rows

#### e. Categorization of BMI Attribute as underweight, normal, overweight:

To categorize the BMI attribute into underweight, normal, and overweight ranges, standard BMI classification thresholds can be used. The following ranges are commonly employed:

- Underweight: BMI less than 18.5
- Normal weight: BMI between 18.5 and 24.9
- Overweight/obese: BMI equal to or greater than 25

By comparing the calculated BMI value with these thresholds, the BMI attribute can be categorized accordingly. The classes created and stored in a #BMI\_table.

```
-----Categorization of BMI Attribute as underweight, normal, overweight
---adding obesity levels:
Select height, weight, BMI,
CASE
When BMI <=18 then 'underweight'
When BMI between 18 and 25 Then 'Normal'
When BMI >25 then 'Overweight'
END as obesity_level into #BMI_table From dbo.Obesity_data ---2111 rows affected

Select * from #BMI_table
```

Results		Messages		
	height	weight	BMI	obesity_level
1	1.62	64	24....	Normal
2	1.52	56	24....	Normal
3	1.80	77	24....	Normal
4	1.80	87	27....	Overweight
5	1.78	90	28....	Overweight
6	1.62	53	20....	Normal
7	1.50	55	24....	Normal
8	1.64	53	20....	Normal
9	1.78	64	20....	Normal
10	1.72	68	23....	Normal
11	1.85	105	31....	Overweight
12	1.72	80	27....	Overweight
13	1.65	56	21....	Normal
14	1.80	99	31....	Overweight
15	1.77	60	19....	Normal
16	1.70	66	23....	Normal
17	1.82	102	27....	Overweight

✓ Query executed successfully.

### Dropping unwanted Tables from the data:

As part of our analysis and model-building process, we made the decision to drop three attributes: tech\_time, transport, and obese\_level. The rationale behind dropping tech\_time and transport attributes was that they were not relevant to our research focus or objectives.

Additionally, I chose to remove the obese\_level attribute since we had created a new attribute called obesity\_level, which was derived based on the BMI values. The obesity\_level attribute was deemed more suitable for our analysis as it directly captured the information regarding the individuals' obesity status, derived from their BMI measurements. By dropping these attributes, we streamlined the dataset to include only the most relevant and useful features for our research purposes, enabling more focused and accurate analysis.

```
-----drop unwanted columns from the table:
```

```

Alter table dbo.obesity_data drop column Tech_using_time
Alter table dbo.obesity_data drop column obesity_level
Alter table dbo.obesity_data drop column transportation_used

```

#### f. Join the tables :

The classes of obesity\_levels were stored in a temporary table. So, Joined the temporary table with the obesity table based on a common key or identifier. This allowed us to combine the information from both tables into a single joined table. After the join operation, we obtained the final joined table, which included the attributes from the obesity table as well as the corresponding obesity\_level classes.

```

----Join the tables-Training data
----Joining #BMI_table and Obesity_data
SELECT a.*, b.obesity_level
INTO dbo.final_table1
FROM dbo.Obesity_data a
JOIN #BMI_table b ON b.BMI = a.BMI
GROUP BY a.Gender, a.Age, a.Height, a.Weight, a.family_history_with_overweight,
a.High_Cal_intake, a.Veg_intake, a.No_of_meals_intake,
a.Consu_food_intake_betw_Meals, a.SMOKE, a.Alcohol_intake,
a.Water_intake, a.Cal_intake_Monitoring, a.Physical_Activity, a.BMI,
b.obesity_level
ORDER BY a.Gender, a.Age, a.Height, a.Weight, a.family_history_with_overweight,
a.High_Cal_intake, a.Veg_intake, a.No_of_meals_intake,
a.Consu_food_intake_betw_Meals, a.SMOKE, a.Alcohol_intake,
a.Water_intake, a.Cal_intake_Monitoring, a.Physical_Activity, a.BMI,
b.obesity_level; ----1766 rows affected

Select * from dbo.final_table1

```

	Gender	Age	Height	Weight	family_history_with_overweight	High_Cal_intake	Veg_intake	No_of_meals_intake	Consu_food_intake_betw_Meals	SMOKE	Water_intake	Cal_i...	Physical_Activity	Alcohol...	BMI	obesity_level
1	Female	16	1.65	86	1	1	1	1	1	0	0	0	1	0	32...	Overweight
2	Female	17	1.55	55	0	1	1	0	1	0	1	1	0	1	23...	Normal
3	Female	17	1.60	65	1	1	1	1	1	0	1	1	0	1	25...	Normal
4	Female	17	1.63	65	0	1	1	0	1	0	1	0	0	0	24...	Normal
5	Female	17	1.63	86	1	1	1	1	1	0	0	0	1	0	32...	Overweight
6	Female	17	1.65	67	1	1	1	0	1	0	1	0	0	0	25...	Normal
7	Female	18	1.45	53	0	1	1	1	1	0	1	1	0	1	25...	Normal
8	Female	18	1.55	56	0	1	1	1	1	0	0	0	0	0	23...	Normal
9	Female	18	1.58	48	0	1	1	1	1	0	1	0	0	0	19...	Normal
10	Female	18	1.60	56	1	1	1	0	1	0	1	1	0	1	22...	Normal
11	Female	18	1.62	68	0	0	1	0	1	0	0	0	0	0	26...	Overweight
12	Female	18	1.63	63	1	1	0	1	1	0	1	1	1	1	24...	Normal
13	Female	18	1.64	56	1	1	1	1	1	0	0	0	0	0	21...	Normal
14	Female	18	1.81	153	1	1	1	1	1	0	1	1	1	1	47...	Overweight
15	Female	19	1.51	59	1	1	1	1	1	0	0	1	0	1	26...	Overweight
16	Female	19	1.53	42	0	1	1	0	1	0	0	1	0	1	18...	underweight
17	Female	19	1.63	49	0	1	1	0	1	0	1	1	0	1	19...	underweight

Query executed successfully.

HEMA (15.0 RTM) | HEMA\hema1 (55) | 780\_Project | 00:00:00 | 1,766 rows

#### g. Splitting the data for Data Analysis:

To conduct data analysis, the final dataset has 1766 records & they were split into training and testing sets using the 80/20 rule. The dataset was first randomly shuffled to ensure an unbiased distribution of instances. Subsequently, 80% of the shuffled data was assigned for training purposes, while the remaining 20% was reserved for evaluating the performance of the predictive models. This split enabled the construction of various predictive models using Weka's algorithms on the training data, allowing for learning and

pattern discovery. The testing data, which was unseen during model development, was then used to assess the models' generalization capabilities and determine their effectiveness in predicting outcomes. This rigorous evaluation process ensured the reliability and applicability of the models for real-world scenarios.

```
-----Code for splitting data
select distinct Age
into #OBA from dbo.final_table1---40 rows affected

select top 80 percent *
into #train_obA
from #OBA---32 rows affected

select *
into [trainingOB]
from dbo.final_table1
where Age in (select * from #train_obA) ---1512 rows affected

select *
into [testingOB2]
from dbo.final_table1
where Age not in (select * from #train_obA)---254 rows
```

## VI. DATA ANALYSIS USING WEKA:

In the data analysis and model-building phase using Weka, the processed dataset was utilized. The dataset was divided into training and testing data using the Percentage Split method, and 4 predictive models were built based on this approach. The models were then evaluated using the testing data to assess their performance and predictive capabilities.

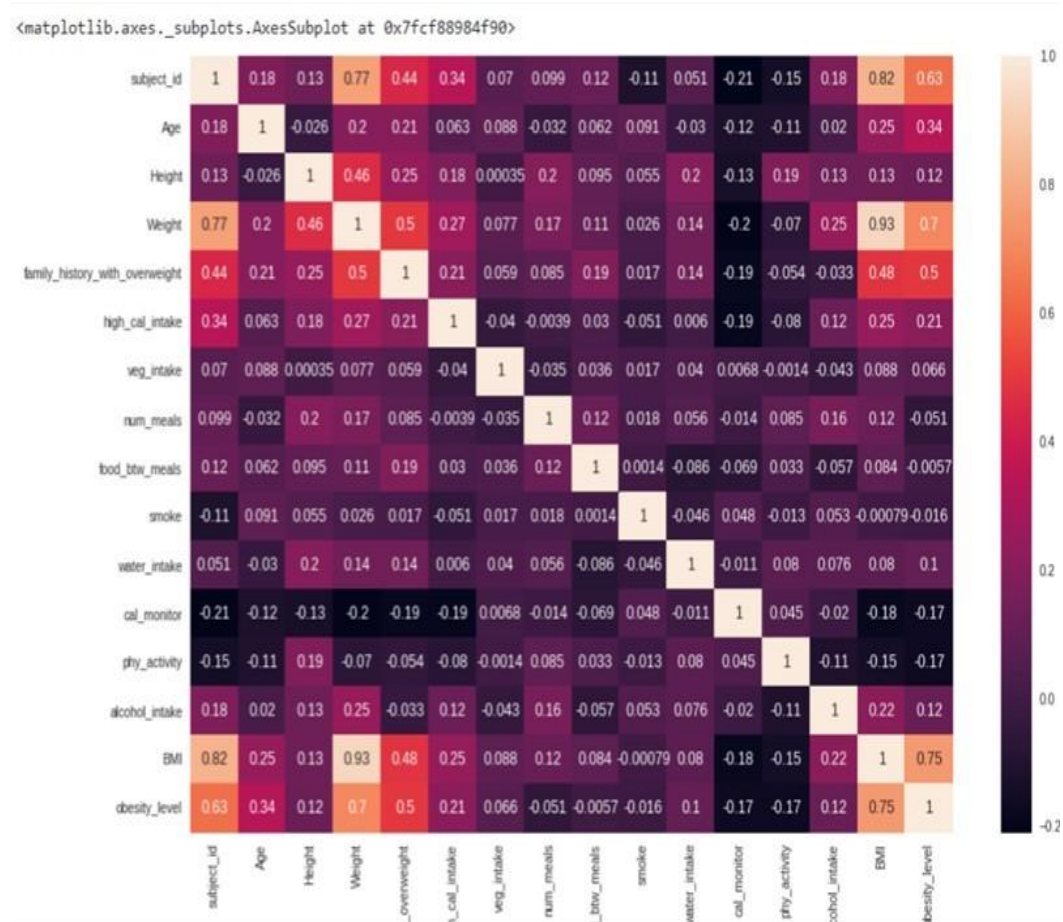
Additionally, Association Rule Mining (ARM) was conducted on the dataset to explore patterns and relationships among the attributes. For this purpose, the FP-Growth method was employed, as it is a widely used algorithm for the efficient mining of frequent itemsets and association rules.

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)



During the association rule mining analysis using the FP-Growth algorithm, the first rule discovered indicates a significant relationship between factors such as a family history with overweight, high obesity level, and high alcohol intake with a high-calorie intake.



The rule suggests that when these factors are present together, there is a 95% confidence level that they contribute to a high-calorie intake.

To further explore the correlations between attributes, a heat map was generated. The heat map revealed a positive correlation of 0.5 between the attributes family\_history\_with\_overweight and obesity\_level. This indicates that individuals with a family history of overweight tend to have a higher likelihood of being classified as obese. Additionally, a correlation matrix was visualized. This matrix provided an overview of the correlations between all the attributes. Notably, the heatmap displayed a positive correlation of 0.7 between family history with overweight and obesity levels. This strong positive correlation suggests that a family history of overweight is closely associated with a higher likelihood of being classified as obese.

Four predictive models, including Random tree, Logistic Regression, Naïve Bayes, and Random Forest, were constructed using the training data, which constituted 80% of the dataset. These models were then evaluated on the remaining 20% of the data using metrics such as ROC area, precision, and recall values. The ROC area assessed the discrimination ability of the models, while precision measured their ability to minimize false positives and recall evaluated their ability to minimize false negatives. Comparing the results of these metrics allowed for a comprehensive analysis of the models' accuracy and performance, enabling conclusions to be drawn regarding their effectiveness in predicting outcomes.

## VII. RESULTS:

This analysis reveals a correlation between family predisposition to obesity and the current state of obesity in adolescent and adult age groups. In the below Table 1 presents the results of the Percentage Split Method, indicating the accuracy, precision, recall, and ROC area for each model on the training set. Random Forest & Random tree achieves the highest accuracy of 100% and a ROC area of 1, demonstrating superior performance compared to other models.

TABLE 1 TRAINING Set				
Model	Accuracy	Precision	Recall	ROC Area
Naïve Bayes	99.5	0.995	0.995	1
Logistic	99.3	0.993	0.993	0.994
Random tree	100	1	1	1
Random forest	100	1	1	1

In the below Table 2 displays the results of the testing set evaluation. Naïve Bayes achieves an accuracy of 88.67%, while Logistic Regression and Random Forest have accuracies of 87.19% and 89.2% respectively. Random Tree performs the best on the testing set, achieving an accuracy

of 91.6%. These results highlight the models' performance on unseen data, with Random Tree outperforming the other models in terms of accuracy and ROC area.

Table 2 Testing Set				
Model	Accuracy	Precision	Recall	ROC Area
Naïve Bayes	88.67	0.872	0.887	0.855
Logistic	87.19	0.862	0.872	0.762
Random forest	89.2	0.892	0.892	0.878
Random tree	91.6	0.923	0.916	0.975

## VIII. DISCUSSION:

This study highlights the significant role of family predisposition to obesity in determining an individual's health status throughout their life. Often, the importance of family history is overlooked when discussing the etiology and risk factors of obesity. However, my findings indicate that considering family history can aid in identifying individuals who may be at a higher risk of obesity and allow for early interventions and lifestyle modifications.

I observed a strong correlation between a family history of obesity and high-calorie food intake in the dataset. This suggests that family history is closely linked to the food habits of individuals, which in turn impact their lifestyle and obesity levels. The combination of family history and high-calorie intake emerged as a major contributing factor to obesity.

To further explore this relationship, future studies could incorporate the attribute of physical activity. By examining the impact of physical activity alongside family history and high-calorie intake, we can gain a deeper understanding of how lifestyle modifications, such as engaging in regular exercise, can potentially mitigate the influence of family history on obesity levels.

Recognizing the significance of family history in determining health outcomes allows for early interventions and empowers individuals to make informed choices about their lifestyles. Further investigation into the interplay between family history, dietary habits, physical activity, and obesity can contribute to developing effective preventive strategies and interventions for managing obesity and improving overall health.

## **IX. CONCLUSION:**

In conclusion, this study highlights the significance of family history in determining the risk of obesity in individuals. While family history is not the sole determinant of obesity, it does increase the chances of an individual becoming obese. I found that lifestyle choices, including alcohol consumption and high-calorie food intake, contribute to the onset of obesity.

However, my findings suggest that individuals who are aware of their family history can make informed choices about their eating habits and adopt a healthy lifestyle, thereby reducing their risk of obesity. By making appropriate dietary choices and maintaining an active lifestyle, individuals can mitigate the impact of family history and decrease their susceptibility to obesity.

This study emphasizes the importance of a comprehensive approach to obesity prevention that takes into account both genetic factors and lifestyle choices. By promoting awareness and providing individuals with the knowledge and tools to make healthy choices, I can empower them to proactively manage their weight and reduce the risk of obesity. Ultimately, this research underscores the significance of personalized interventions and lifestyle modifications in preventing obesity and promoting overall health.

## **X. LIMITATIONS:**

This study has several limitations that should be considered. Firstly, the training dataset used in this study consisted of a limited number of rows (~2111), which may result in underfitting of the predictive models and affect their robustness. A larger dataset would provide more diverse and representative samples, allowing for better model training and generalizability.

Secondly, the level of awareness among the public regarding the risk factors and long-term consequences of obesity may vary. This could lead to a bias in the dataset, potentially resulting in an overrepresentation of individuals with higher obesity rates. Consequently, the findings may not fully reflect the broader population's obesity patterns.

Additionally, it is worth noting that while logistic regression and random forest models achieved high accuracy, the optimization of accuracy for other models may have been limited due to resource constraints. Further exploration and optimization of alternative models could provide additional insights and potentially improve the overall performance of the predictive models.

## **XI. REFERENCES:**

1. H. B. Hubert, M. Feinleib, P. M. McNamara, and W. P. Castelli, "Obesity as an independent risk factor for cardiovascular disease: A 26-year follow-up of participants in the Framingham Heart Study," *Circulation*, vol. 67, no. 5, pp. 968–977, 1983, doi: 10.1161/01.CIR.67.5.968.
2. A. Must, J. Spadano, E. H. Coakley, A. E. Field, G. Colditz, and W. H. Dietz, "The disease burden associated with overweight and obesity," *J. Am. Med. Assoc.*, vol. 282, no. 16, pp. 1523–1529, 1999, doi: 10.1001/jama.282.16.1523.
3. B. Guy-Grand, "Beyond body mass index," *Cah. Nutr. Diet.*, vol. 49, no. 3, pp. 93–94, 2014, doi: 10.1016/j.cnd.2014.05.002.
4. E. Alyahyan and D. Dusteaor, "Decision trees for very early prediction of student's achievement," 2020 2nd Int. Conf. Comput. Inf. Sci. ICCIS 2020, 2020, doi: 10.1109/ICCIS49240.2020.9257646.
5. N. Lavrač, "Selected techniques for data mining in medicine," *Artif. Intell. Med.*, vol. 16, no. 1, pp. 3–23, 1999, doi: 10.1016/S0933-3657(98)00062-1.
6. M. H. J. and P. Jian and Kamber, "Data Mining Techniques, Third Edition," p. 847, 2011.
7. M. Khajehei and F. Etemady, "Data mining and medical research studies," *Proc. - 2nd Int. Conf. Comput. Intell. Model. Simulation, CIMSIm 2010*, no. September 2010, pp. 119–122, 2010, doi: 10.1109/CIMSIm.2010.24.
8. R. C. Cervantes and U. M. Palacio, "Estimation of obesity levels based on computational intelligence," *Informatics Med. Unlocked*, vol. 21, no. November 2020, doi: 10.1016/j.imu.2020.100472.
9. Singh and H. Tawfik, "A Machine Learning Approach for Predicting Weight Gain Risks in Young Adults," *Conf. Proc. 2019 10th Int. Conf. Dependable Syst. Serv. Technol. DESSERT 2019*, pp. 231–234, 2019.
10. R. Hossain, S. M. H. Mahmud, M. A. Hossin, S. R. Haider Noori, and H. Jahan, "PRMT: Predicting Risk Factor of Obesity among Middle-Aged People Using Data Mining Techniques," *Procedia Comput. Sci.*, vol. 132, pp. 1068–1076, 2018, doi: 10.1016/j.procs.2018.05.022.
11. Z. Zheng and K. Ruggiero, "Using machine learning to predict obesity in high school students," *Proc. – 2017 IEEE Int. Conf. Bioinforma. Biomed. BIBM 2017*, vol. 2017-Janua, pp. 2132–2138, 2017, doi: 10.1109/BIBM.2017.8217988.
12. M. K. Uçar, Z. Uçar, F. Köksal, and N. Daldal, "Estimation of body fat percentage using hybrid machine learning algorithms," *Meas. J. Int. Meas. Confed.*, vol. 167, 2021, doi: 10.1016/j.measurement.2020.108173.
13. N. Daud, N. L. Mohd Noor, S. A. Aljunid, N. Noordin, and N. I. M. Fahmi Teng, "Predictive Analytics: The Application of J48 Algorithm on Grocery Data to Predict Obesity," 2018 IEEE Conf. Big Data Anal. ICBDA 2018, pp. 1–6, 2019, doi: 10.1109/ICBDAA.2018.8629623.