BITI 2513: Introduction to Data Science

Exploring the Factors Contributing to Obesity using Data Analysis Approach

NAMES	MATRIC NUMBER
1. LEE YUN KANG	B032010390
2. TAN WEI HAN	B032010026
3. KISHWANTH A/L HARI KRISHNAN	B032010185
4. GIRIDHEV A/L MAHESAN	B032010451

Introduction

Obesity is a growing public health concern worldwide, and its prevalence has been on the rise in recent decades. Obesity is caused by a variety of factors, many of which are still poorly understood, in addition to genetics and lifestyle choices. It is essential to do accurate and data-driven research on the factors that contribute to obesity in order to address this problem and create effective therapies.

In this study, we will analyse all of the factors that lead to obesity using an artificial intelligence (AI) data analysis technique. We aim to find patterns and relationships that can shed light on the underlying causes of obesity by analysing huge datasets that include data on lifestyle and health-related variables. We will also look at how these factors change among various populations.

This study can help design specific approaches that address the particular causes of obesity in various populations, which has significant implications for public health policy and intervention efforts. Finally, by developing more efficient and long-lasting solutions to this important public health issue, we can better comprehend the complicated and complex nature of obesity.

Objective

1. Identifying the lifestyle and health-related factors that are associated with obesity, using large datasets that contain information on these factors.
2. Analyse the patterns and interactions between these factors and determine which combinations of factors are most strongly associated with obesity.
3. Develop predictive models that can accurately identify individuals at high risk of developing obesity based on their lifestyle and health-related factors.

Aims of Artificial Intelligence

The use of Artificial Intelligence (AI) can significantly help in exploring the factors contributing to obesity using a data analysis approach. There are two aims of AI in this study. First is to identify the relevant factors. AI can be used to automatically identify variables that are most strongly associated with obesity, including lifestyle and health-related factors. This can facilitate data analysis and guarantee that every relevant factor is considered and taken into account.

Next is for the prediction. AI can be used to develop predictive models that can accurately identify individuals at high risk of developing obesity based on their lifestyle and health-related factors. This can help to target interventions and prevention efforts to those who are most in need.

In this project, we used the logistic regression algorithm where it is a statistical method used in data science for predicting binary outcomes (for example: outcomes that take only one of two values, such as yes or no, true or false, etc.). It is a type of generalized linear model that uses a logistic function to model the relationship between the independent variables and the binary response variable. In logistic regression, the goal is to estimate the probability of a binary outcome based on one or more predictor variables. The logistic function is used to transform the linear combination of the predictor variables into a probability value between 0 and 1. The logistic regression model estimates the parameters that define the relationship between the predictor variables and the binary response variable. These coefficients are used to calculate the logic of the probability of the binary outcome. The log-odds are then transformed back into a probability value using the logistic function.

How to Deploy

1. Cleaning of data includes dropping of unnecessary column data which are unused variables. For example: The code uses the drop method to remove columns from the two separate pandas DataFrames, df1 and df2. Columns that are not required for the analysis or modelling that will be done on the DataFrames are removed using the drop method. Only the pertinent columns are kept when these columns are dropped, which also removes the appropriate data from the DataFrame.
2. Check the null data from the dataset. Because they can produce errors or biased findings when employing specific machine learning algorithms or statistical techniques, null values are typically deleted from datasets. For example in the code we used, it is ensured that the 'bmi' column does not include any missing or null values by eliminating rows with missing 'bmi' values, which can help avoid mistakes or inaccurate results in subsequent studies.
3. Besides that, the percentage of obesity in each key factor categories are calculated.
4. Bar graphs and histogram are created for data analysis using google collab code.
5. Replace the data with null data.
6. Convert it into numeric data.
7. Prediction is done by using Logistic Regression.
8. Using the train_test_split method with a test size of 0.3, the data is divided into training and testing sets, with 30% of the data being utilised for testing and the remaining 70% for training.
9. Finally for the training model, the accuracy obtained is 0.6973, whereas the accuracy for the testing model is 0.6935.

Dataset

The datasets were taken from the Kaggle website. Two different datasets were taken from the website, and we combined it into one dataset which has similar variables. We took 2 different datasets that has similar variables to ensure that the AI predictive model has sufficient amount of data to be trained and tested. The unused variables were dropped during the data cleaning process.

Electronic Health Records (EHRs) are the primary source of data for the Diabetes Prediction dataset. EHRs are digital versions of patient health records that contain information about their medical history, diagnosis, treatment, and outcomes. The data in EHRs is collected and stored by healthcare providers, such as hospitals and clinics, as part of their routine clinical practice. To create the Diabetes Prediction dataset, EHRs were collected from multiple healthcare providers and aggregated into a single dataset. The data was then cleaned and pre-processed to ensure consistency and remove any irrelevant or incomplete information. Finally, EHRs are widely used in clinical practice, making the Diabetes Prediction dataset relevant to real-world healthcare settings.

Before Cleaning and Merging Both Dataset

First Dataset

The Diabetes prediction dataset is a collection of medical and demographic data from patients, along with their diabetes status (positive or negative). In this dataset there were 9 variables which are the gender, age, hypertension, heart disease, smoking history, BMI, HbA1c level, blood glucose level, and diabetes. The gender refers to the biological sex of the individual, which can have an impact on their susceptibility to diabetes. As for the age, it is an important factor as diabetes is more commonly diagnosed in older adults. Age ranges from 0-80 in our dataset. Hypertension is a medical condition in which the blood pressure in the arteries is persistently elevated. Heart disease is another medical condition that is associated with an increased risk of developing diabetes. Both hypertension and heart disease have values of 0 and 1. The value 0 indicates no hypertension or heart disease, whereas for the value 1 it indicates the individual has hypertension or heart disease.

Next variable is the smoking history where it is also considered a risk factor for diabetes and can exacerbate the complications associated with diabetes. In our dataset we have 5 categories (for example: not current, former, No Info, current, never and ever). BMI (Body

Mass Index) is a measure of body fat based on weight and height. Higher BMI values are linked to a higher risk of diabetes. The range of BMI in the dataset is from 10.16 to 71.55. BMI less than 18.5 is underweight, 18.5-24.9 is normal, 25-29.9 is overweight, and 30 or more is obese. HbA1c (Hemoglobin A1c) level is a measure of a person's average blood sugar level over the past 2-3 months. Higher levels indicate a greater risk of developing diabetes. Mostly more than 6.5% of HbA1c Level indicates diabetes. Blood glucose level refers to the amount of glucose in the bloodstream at a given time. High blood glucose levels are a key indicator of diabetes. Diabetes is the target variable being predicted, with values of 1 indicating the presence of diabetes and 0 indicating the absence of diabetes.

	gender	age	hypertension	heart_disease	smoking_history	bmi	HbA1c_level	blood_glucose_level	diabetes
0	Female	80.0	0	1	never	25.19	6.6	140	0
1	Female	54.0	0	0	No Info	27.32	6.6	80	0
2	Male	28.0	0	0	never	27.32	5.7	158	0
3	Female	36.0	0	0	current	23.45	5.0	155	0
4	Male	76.0	1	1	current	20.14	4.8	155	0
...	...	...	...	...	...	...	...	...	...
95	Male	19.0	0	0	never	27.32	6.1	80	0
96	Female	67.0	0	0	never	27.32	6.2	159	1
97	Male	11.0	0	0	No Info	27.32	6.1	90	0
98	Female	30.0	0	0	No Info	50.13	6.0	100	0
99	Male	29.0	0	0	current	27.32	4.8	158	0

100 rows × 9 columns

Second Dataset

The Stroke prediction dataset is a collection of medical and demographic data from patients, along with their stroke status (positive or negative). In this dataset there were 12 variables which are the id, sex, age, hypertension, heart disease, ever married, work type, residence type, average glucose level, BMI, smoking status, and stroke.

	id	gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status	stroke
0	9046	Male	67.0	0	1	Yes	Private	Urban	228.69	36.6	formerly smoked	1
1	51676	Female	61.0	0	0	Yes	Self-employed	Rural	202.21	NaN	never smoked	1
2	31112	Male	80.0	0	1	Yes	Private	Rural	105.92	32.5	never smoked	1
3	60182	Female	49.0	0	0	Yes	Private	Urban	171.23	34.4	smokes	1
4	1665	Female	79.0	1	0	Yes	Self-employed	Rural	174.12	24.0	never smoked	1
...	...	...	...	...	...	...	...	...	...	...	...	...
95	2458	Female	78.0	0	0	Yes	Private	Rural	235.63	32.3	never smoked	1
96	35512	Female	70.0	0	0	Yes	Self-employed	Rural	76.34	24.4	formerly smoked	1
97	56841	Male	58.0	0	1	Yes	Private	Rural	240.59	31.4	smokes	1
98	8154	Male	57.0	1	0	Yes	Govt_job	Urban	78.92	27.7	formerly smoked	1
99	4639	Female	69.0	0	0	Yes	Govt_job	Urban	82.81	28.0	never smoked	1

After Cleaning and Merging Both Dataset

In the first dataset, the columns HbA1c Level and diabetes are dropped. As for the second dataset, the columns ID, ever married, work type, residence type and stroke are dropped. These variables or factors are not needed for our study. Hence, both datasets have the same variables which can be combined and used. The figure below shows the variables from the combined dataset.

	gender	age	hypertension	heart_disease	smoking_status	bmi	blood_glucose_level
0	Female	80.0	0	1	never	25.19	140.0
1	Female	54.0	0	0	No Info	27.32	80.0
2	Male	28.0	0	0	never	27.32	158.0
3	Female	36.0	0	0	current	23.45	155.0
4	Male	76.0	1	1	current	20.14	155.0
...	...	...	...	...	...	...	...
95	Male	19.0	0	0	never	27.32	80.0
96	Female	67.0	0	0	never	27.32	159.0
97	Male	11.0	0	0	No Info	27.32	90.0
98	Female	30.0	0	0	No Info	50.13	100.0
99	Male	29.0	0	0	current	27.32	158.0

Tools and Programming Language

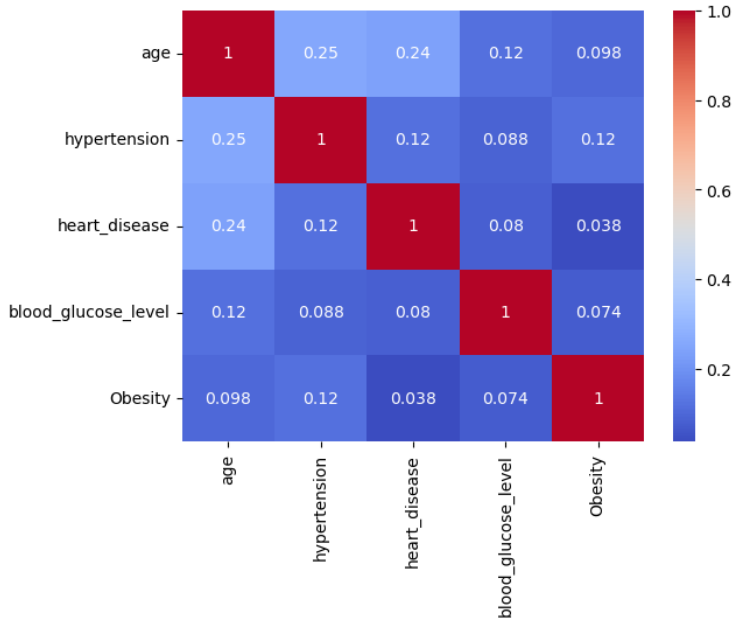
The tools we used for our data science project is Google Colab which is a powerful and convenient platform for data science projects, particularly when working on large datasets that require powerful computing resources.

Python is a popular programming language used extensively in the field of data science. Python is popular among data scientists and analysts because it offers an extensive collection of libraries and tools for data research. It is a useful tool for working with data because of its simplicity, adaptability, and potent powers.

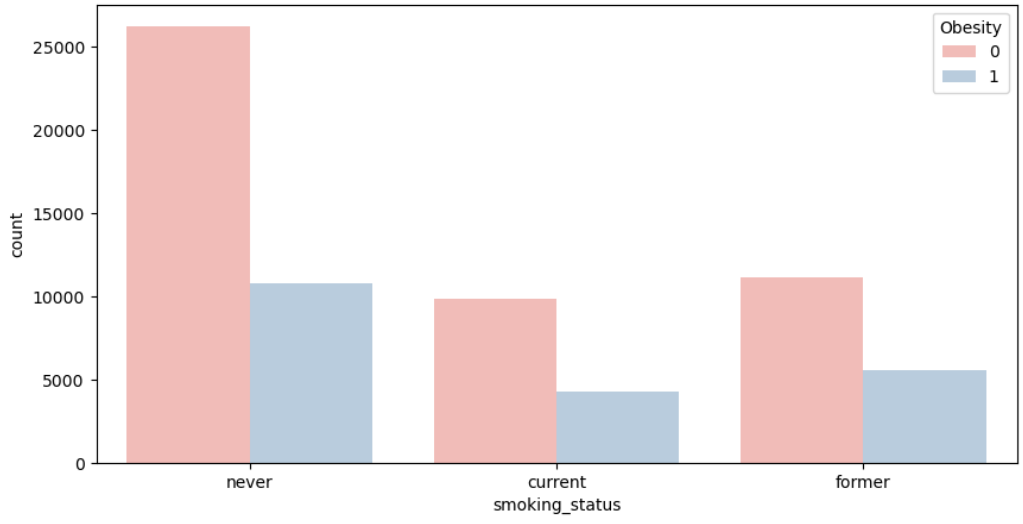
Exploratory Data Analysis

Exploratory data analysis (EDA) is a crucial step in understanding the data and identifying patterns and relationships. EDA can involve tasks such as descriptive statistics, data visualization, and correlation analysis.

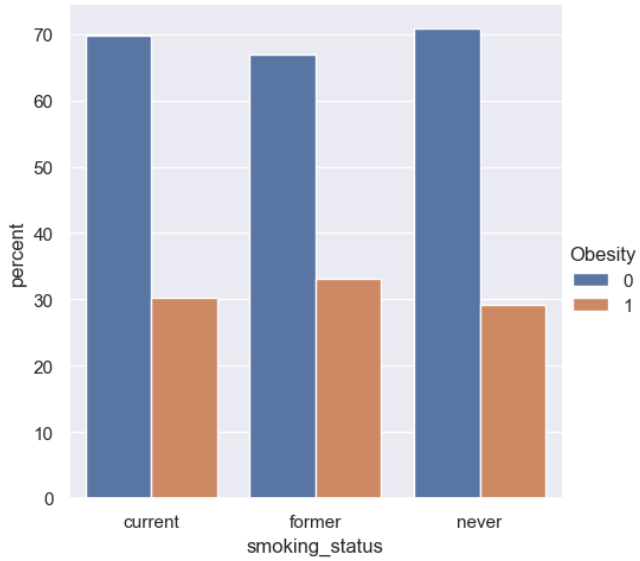
The figure below shows the heatmap of the variables used in the dataset.



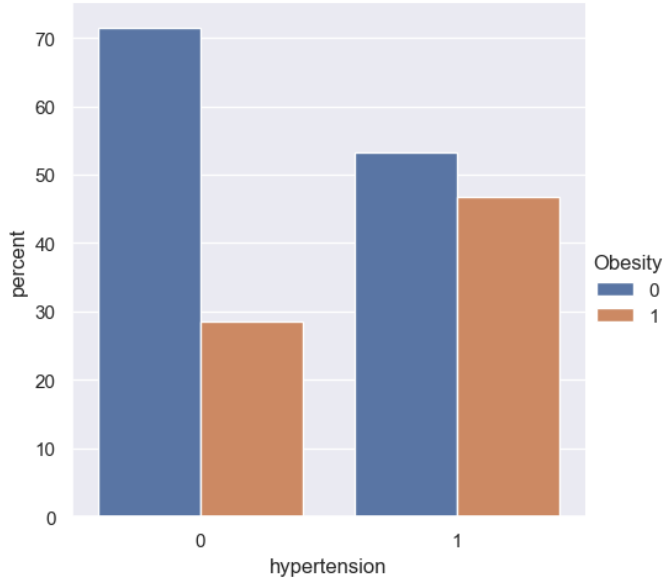
The figure below shows relationship on the number of individuals and how the smoking status variable can affect to obesity. According to our data, 10.8k people in the demographic being investigated have never smoked, 5.5k are past smokers, and 4.2k are current smokers. Further investigation suggests that people who have never smoked have a higher prevalence of obesity than those who have smoked in the past.



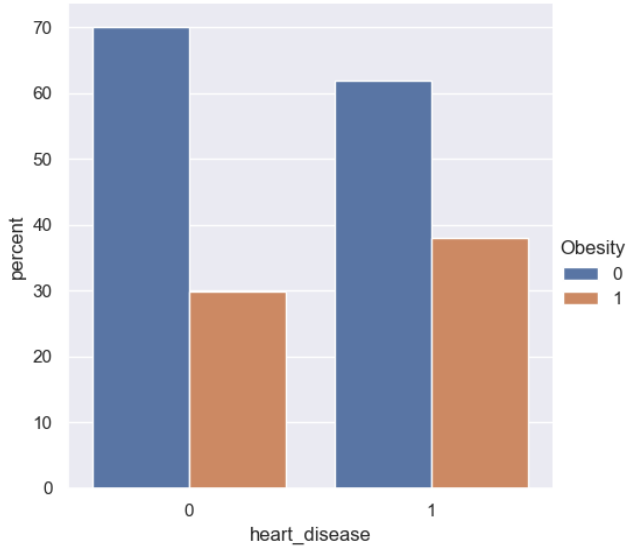
The figure below shows the relationship on the percentage and how the smoking status variable can affect to obesity. Based on our data analysis, it is possible to conclude that around 70% of current smokers are not obese, whereas the remaining 30% are. On the other hand, around 68% of ex-smokers are obese. Finally, nearly 71% of people who have never smoked are not obese. These findings shed light on the association between smoking status and obesity in our study population.



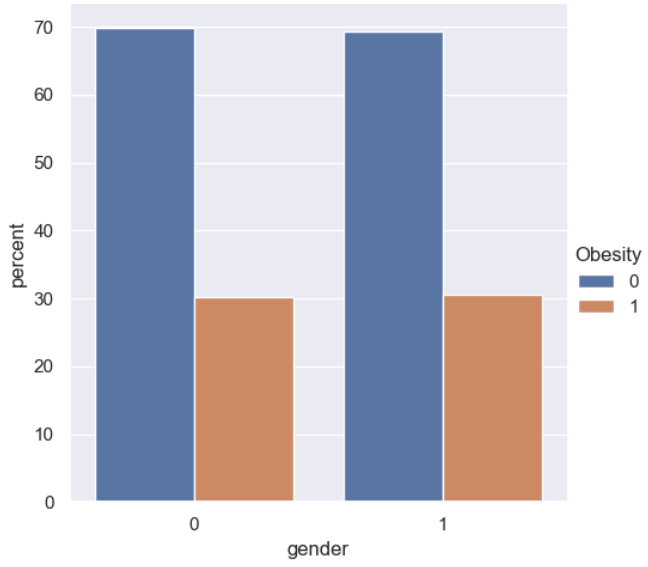
The figure below shows the relationship on the percentage and how the hypertension factor can affect to obesity. We analysed the data and discovered that nearly 70% of the people who do not have obesity or heart disease. On the other hand, approximately 62% of people with obesity and heart disease. This suggests that there is a moderate link between obesity and heart disease, and that being obese may raise the chance of acquiring heart disease.



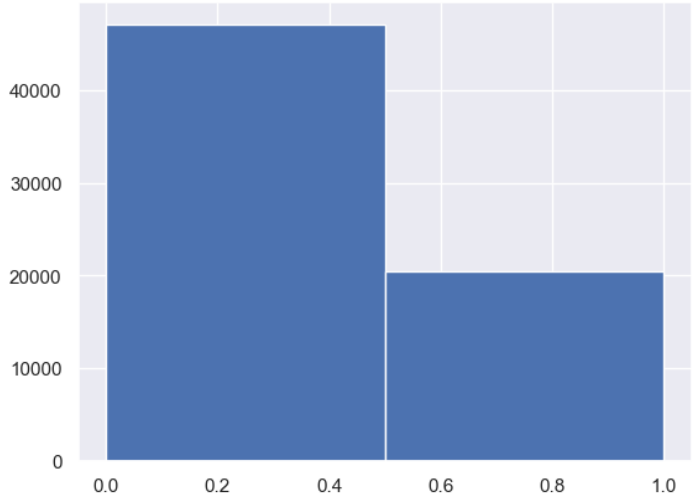
The figure below shows the relationship on the percentage and how the heart disease factor can affect to obesity. We analysed the data and discovered that nearly 70% of the people who do not have obesity or heart disease. On the other hand, approximately 62% of people with obesity and heart disease. This suggests that there is a moderate link between obesity and heart disease, and that being obese may raise the chance of acquiring heart disease.



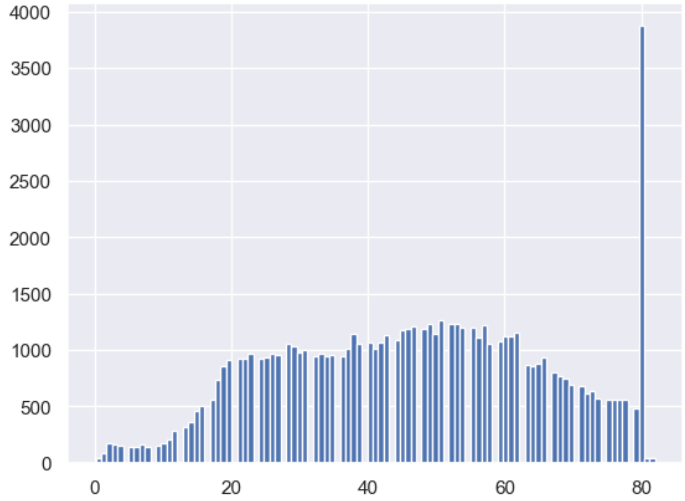
The figure below shows the relationship on the percentage and how the gender factor can affect to obesity. According to our data research, the percentage of ladies who are not fat is over 70%, whereas the comparable percentage for males is over 30%. This disparity could be attributed to a variety of causes such as lifestyle differences, genetics, or socio-cultural norms. More research is needed to identify and treat gender-based health disparities.



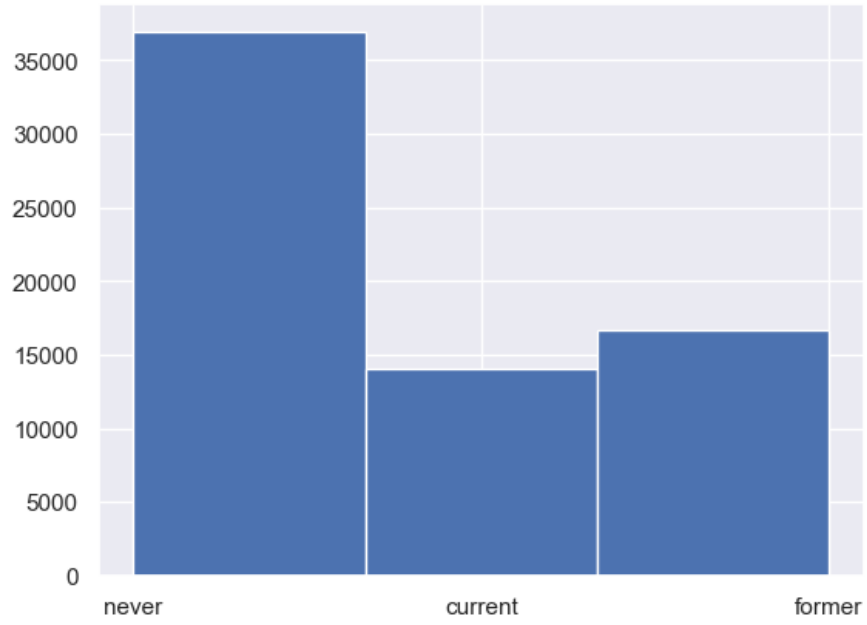
The figure below shows the number of individuals who faces obesity where the left-hand side (value 0) indicates no obesity whereas the right-hand side (value 1) indicates the individual has obesity. It's important to note that the dataset is unbalanced with around 70% of the population being obese and only 30% being non-obese. This could potentially lead to biased or inaccurate results and should be taken into consideration when analyzing the data. Specifically, there are around 47,000 individuals who are obese and 20,000 individuals who are not obese in the dataset.



The figure below shows the number of individuals who faced obesity according to their age range. According to the dataset analysis, the most populous age group is around 80 years old, with approximately 3800 persons. The number of people declines dramatically beyond this age group. There are less than 200 persons between the ages of 0 and 10, while the number of people between the ages of 60 and 80 ranges from 1000 to 500. The remaining age categories have a population of roughly 1000 people on average.



The figure below shows the number of individuals who faced obesity according to their smoking status. There are approximately 35,000 people in this dataset who have never smoked, while 14,500 people currently smoke. On the other hand, the number of ex-smokers ranges between 15,000 and 17,000 people.



Results

Training Results

```
=========LOGISTIC REGRESSION============
TRAINIG RESULTS:
===============================
CONFUSION MATRIX:
[[32578   446]
 [13879   414]]
ACCURACY SCORE:
0.6973
CLASSIFICATION REPORT:
                    0             1  accuracy     macro avg  weighted avg
precision    0.701251      0.481395  0.697255      0.591323      0.634839
recall       0.986495      0.028965  0.697255      0.507730      0.697255
f1-score     0.819768      0.054643  0.697255      0.437205      0.588648
support    33024.000000  14293.000000  0.697255  47317.000000  47317.000000
```

Testing Results

```
TESTING RESULTS:
===============================
CONFUSION MATRIX:
[[13892   188]
 [ 6028   172]]
ACCURACY SCORE:
0.6935
CLASSIFICATION REPORT:
                    0             1  accuracy     macro avg  weighted avg
precision    0.697390      0.477778  0.693491      0.587584      0.630250
recall       0.986648      0.027742  0.693491      0.507195      0.693491
f1-score     0.817176      0.052439  0.693491      0.434808      0.583381
support    14080.000000  6200.000000  0.693491  20280.000000  20280.000000
```

The prediction of the impacts that causes obesity.

	feature	coefficient
2	hypertension	0.182931
1	age	0.140062
5	blood_glucose_level	0.121749
3	heart_disease	0.025383
4	smoking_status	0.019679
0	gender	-0.000630

Based on the coefficients obtained, hypertension is most probably the main cause of obesity with the highest coefficient value of 0.1829. This followed by age, blood glucose level, heart disease, smoking status and gender which might be the least chance to be a factor of obesity with the lowest coefficient value -0.00063. Besides that the accuracy that we got is not that high with just a rate of 0.6935 because the dataset we got is imbalance.

References

Kaggle website, Mohammad Mustafa. *NA.* Diabetes prediction dataset. Retrieved from https://www.kaggle.com/datasets/iammustafatz/diabetes-prediction-dataset

Kaggle website, Prosper Chuks. *NA*. Diabetes, Hypertension and Stroke prediction. Retrieved from https://www.kaggle.com/datasets/prosperchuks/health-dataset?select=stroke_data.csv