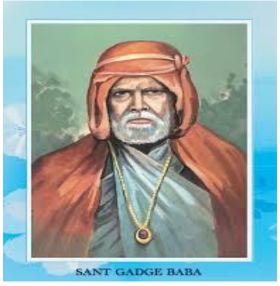
“EDUCATION THROUGH SELF-HELP IS OUR MOTTO”

RAYAT SHIKSHAN SANSTHA’S

SADGURU GADAGE MAHARAJ COLLEGE, KARAD



(An Autonomous College)

Project report on,

“Statistical Analysis of Factors Affecting on Obesity”

Submitted by,

Mr. Jadhav Rahul Arun (M. Sc. II)

Roll No.- 14

2021-2022

CERTIFICATE

This is to certify that the Project report entitled "Statistical analysis of factors affecting on Obesity" being submitted by **Mr. Jadhav Rahul Arun** as a partial fulfillment for the M.Sc. - II in Statistics of Sadguru Gadage Maharaj College, Karad is a record of bonafide work carried out by him under my supervision and guidance. To the best of our knowledge and belief, the matter presented in this project report is original and has not been submitted elsewhere for any other purpose.

Place: Karad

Date:

Dr.Mrs.Chavan R.V.	**Examiner**	**Smt Mahajan S. V.**
(Project Guide)		**Head Dept. of Statistics**

Acknowledgement

I am privileged to express my sincere thanks with great respect and gratitude to **Smt. Mahajan S.V.** (Head, Department of Statistics), **Dr.Mrs.Chavan R.V.** for their aspiring guidance. They all helped with kind of co-operation and constant encouragement. I am grateful to thank them for providing me with all necessary facilities.

Also, I would like to thank all the non-teaching staff of the department for their help and co-operation. I thank all my friends and the teaching staff for their co-operation and help which I received from them during the work throughout.

I am indebted to my parents for their encouragement and patience throughout my study and for the trust they have on me.

Yours Sincerely,
Mr. Jadhav Rahul Arun
M.Sc - II
Department Of Statistics.

CONTENTS

Introduction :

Obesity is a complex disease involving an excessive amount of body fat. Obesity isn't just a cosmetic concern. It's a medical problem that increases the risk of other diseases and health problems, such as heart disease, diabetes, high blood pressure and certain cancers.

There are many reasons why some people have difficulty losing weight. Usually, obesity results from inherited, physiological and environmental factors, combined with diet, physical activity and exercise choices.

The good news is that even modest weight loss can improve or prevent the health problems associated with obesity. A healthier diet, increased physical activity and behavior changes can help you lose weight. Prescription medications and weight-loss procedures are additional options for treating obesity.

Body mass index (BMI) is often used to diagnose obesity. To calculate BMI, multiply weight in pounds by 703, divide by height in inches and then divide again by height in inches. Or divide weight in kilograms by height in meters squared.

If BMI is below 18.5 then the weight status is underweight , if it is 18.5 to 24.9 then the individual's weight status is normal , for 25-29.9 it is Overweight and if BMI is 30 and more then it is obesity which is not good. Asians with BMI of 23 or higher may have an increased risk of health problems.

For most people, BMI provides a reasonable estimate of body fat. However, BMI doesn't directly measure body fat, so some people, such as muscular athletes, may have a BMI in the obesity category even though they don't have excess body fat.

Many doctors also measure a person's waist circumference to help guide treatment decisions. Weight-related health problems are more common in men with a waist circumference over 40 inches (102 centimeters) and in women with a waist measurement over 35 inches (89 centimeters).

Causes : Although there are genetic, behavioral, metabolic and hormonal influences on body weight, obesity occurs when you take in more calories than you burn through normal daily activities and exercise. Your body stores these excess calories as fat.

In the United States, most people's diets are too high in calories — often from fast food and high-calorie beverages. People with obesity might eat more calories before feeling full, feel hungry sooner, or eat more due to stress or anxiety.

Many people who live in Western countries now have jobs that are much less physically demanding, so they don't tend to burn as many calories at work. Even daily activities use fewer calories, courtesy of conveniences such as remote controls, escalators, online shopping and drive-through banks.

Objectives

1) To study the impact of all factors or variable mentioned in the dataset on the target variable i.e. Obesity.
2) To check the relation between Genetics and Obesity by using EDA.
3) To build the model for predicting the Obesity based on habits.
4) To find relationship between Gender and Obesity.

About Data

This project data 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 condition as mentioned, data was 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.

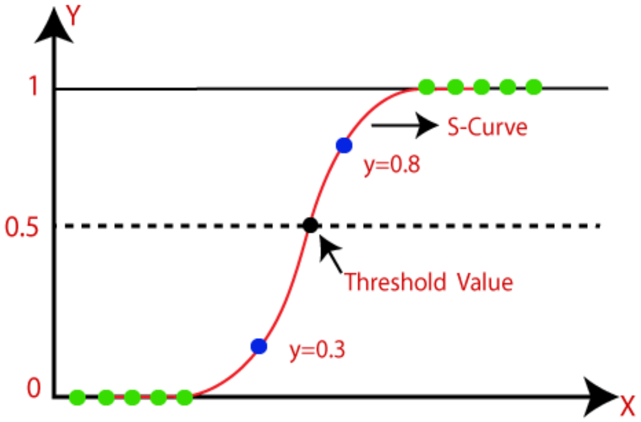
The attributes related with eating habits are: Frequent consumption of high caloric food (FAVC), Frequency of consumption of vegetables (FCVC), Number of main meals (NCP), Consumption of food between meals (CAEC), Consumption of water daily (CH20), and Consumption of alcohol (CALC). The attributes related with the physical condition are: Calories consumption monitoring (SCC), Physical activity frequency (FAF), Time using technology devices (TUE), Transportation used (MTRANS), other variables obtained were: Gender, Age, Height and Weight. Finally, all data was labelled 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 below equation,

$$\text{Body Mass Index} = \frac{\text{Weight (kg)}}{\text{Height}^2 \text{ (m)}}$$

Methodology

Logistic Regression in Machine Learning

- Logistic regression is one of the most popular Machine Learning algorithms, which comes under the Supervised Learning technique. It is used for predicting the categorical dependent variable using a given set of independent variables.
- Logistic regression predicts the output of a categorical dependent variable. Therefore the outcome must be a categorical or discrete value. It can be either Yes or No, 0 or 1, true or False, etc. but instead of giving the exact value as 0 and 1, **it gives the probabilistic values which lie between 0 and 1**.
- Logistic Regression is much similar to the Linear Regression except that how they are used. Linear Regression is used for solving Regression problems, whereas **Logistic regression is used for solving the classification problems**.
- In Logistic regression, instead of fitting a regression line, we fit an "S" shaped logistic function, which predicts two maximum values (0 or 1).
- The curve from the logistic function indicates the likelihood of something such as whether the cells are cancerous or not, a mouse is obese or not based on its weight, etc.
- Logistic Regression is a significant machine learning algorithm because it has the ability to provide probabilities and classify new data using continuous and discrete datasets.
- Logistic Regression can be used to classify the observations using different types of data and can easily determine the most effective variables used for the classification. The below image is showing the logistic function:



Logistic Function (Sigmoid Function):

- The sigmoid function is a mathematical function used to map the predicted values to probabilities.
- It maps any real value into another value within a range of 0 and 1.
- The value of the logistic regression must be between 0 and 1, which cannot go beyond this limit, so it forms a curve like the "S" form. The S-form curve is called the Sigmoid function or the logistic function.
- In logistic regression, we use the concept of the threshold value, which defines the probability of either 0 or 1. Such as values above the threshold value tends to 1, and a value below the threshold values tends to 0.

Assumptions for Logistic Regression:

- The dependent variable must be categorical in nature.
- The independent variable should not have multi-collinearity.

Logistic Regression Equation:

The Logistic regression equation can be obtained from the Linear Regression equation. The mathematical steps to get Logistic Regression equations are given below:

- We know the equation of the straight line can be written as:

$$\mathrm{y}= b_0 + b_1x_1 + b_2x_2 + b_3x_3 + \cdots + b_nx_n$$

- In Logistic Regression y can be between 0 and 1 only, so for this let's divide the above equation by (1-y):

$$\frac{\mathbf{y}}{\mathbf{1-y}}\ \textbf{; 0 for y= 0, and infinity for y=1}$$

- But we need range between -[infinity] to +[infinity], then take logarithm of the equation it will become:

$$\boldsymbol{log\left[\frac{y}{1-y}\right] = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + \cdots + b_nx_n}$$

The above equation is the final equation for Logistic Regression.

Variable Description

1) Gender: Male= 1 , Female=0
2) Age: Numerical Value
3) Height: Numeric value in meters
4) Weight: Numerical Valuekilograms
5) FHWO : family history with overweight i.e if the disease has came genetically
6) FAVC : Frequent consumption of high caloric food – whose responses are binary
7) FCVC : Frequency of consumption of vegetables.
8) NCP : Number of main meals – 1,2 or 3 meals a day.
9) CAEC : Consumption of food between meals – like as Sometimes , Always , Frequently and No which can be encoded as 1, 2, 3, 4.
10) Smoke – which is binary variable.
11) CH20 : Consumption of water daily - less than a liter =1 , Between 1 and 2 liter =2 , More than 2Liter =3.

    The attributes related with the physical condition.
12) SCC : Calories consumption monitoring .
13) CALC : Consumption of alcohol - like as Sometimes , Always , Frequently and No which are encoded as 1, 2, 3, 4.
14) MTRANS : Transportation used– Automobile , Motorbike , Bike , Public Transportation , Walking which are encoded.
15) FAF : Physical activity frequency- do not have , 1 or 2 days , 2 or 4 days , 4 or 5 days
16) TUE : Time using technology devices -technological devices such as cell phone, videogames, television, computer and others-- 0-2 hours , 3-5 hours , More than 5 hours.

Descriptive Statistics

- ❖ **Dimension of dataset :**

There is 2111 rows and 17 columns in this dataset , that means overall size of the dataset is 2111

- ❖ **To Check missing values and treat using suitable technique**

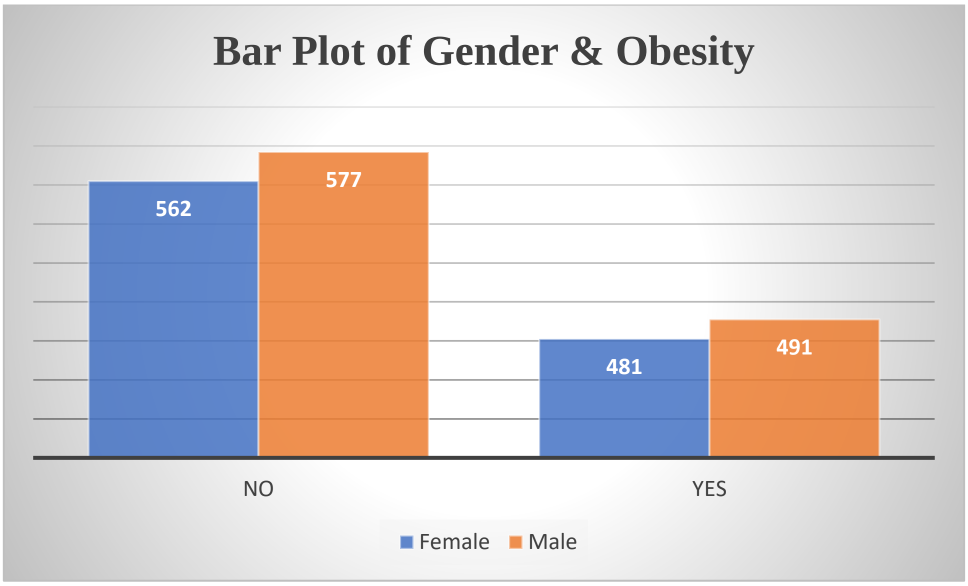
```
#   Column                                           Non-Null    Count    Dtype
---  ------                                  -------------  -----
0   Gender                                          2111 non-null   int64
1   Age                                             2111 non-null   float64
2   Height                                          2111 non-null   float64
3   Weight                                          2111 non-null   float64
4   family_history_with_overweight                  2111 non-null   int32
5   Frequent consumption of high caloric food       2111 non-null   int32
6   Frequent consumption of high caloric food       2111 non-null   float64
7   Number of main meals                            2111 non-null   float64
8   Consumption of food between meals               2111 non-null   int32
9   SMOKE                                           2111 non-null   int32
10  Consumption of water daily                      2111 non-null   float64
11  Calories consumption monitoring                 2111 non-null   int32
12  Physical activity frequency                     2111 non-null   float64
13  Time using technology devices                   2111 non-null   float64
14  Consumption of alcohol                          2111 non-null   int32
15  Transportation used                             2111 non-null   int32
16  obesity                                         2111 non-null   int64
```

We see that there is no any missing or null value in our dataset .

Exploratory Data Analysis

	Obesity	
Gender	No	Yes
Female	562	481
Male	577	491

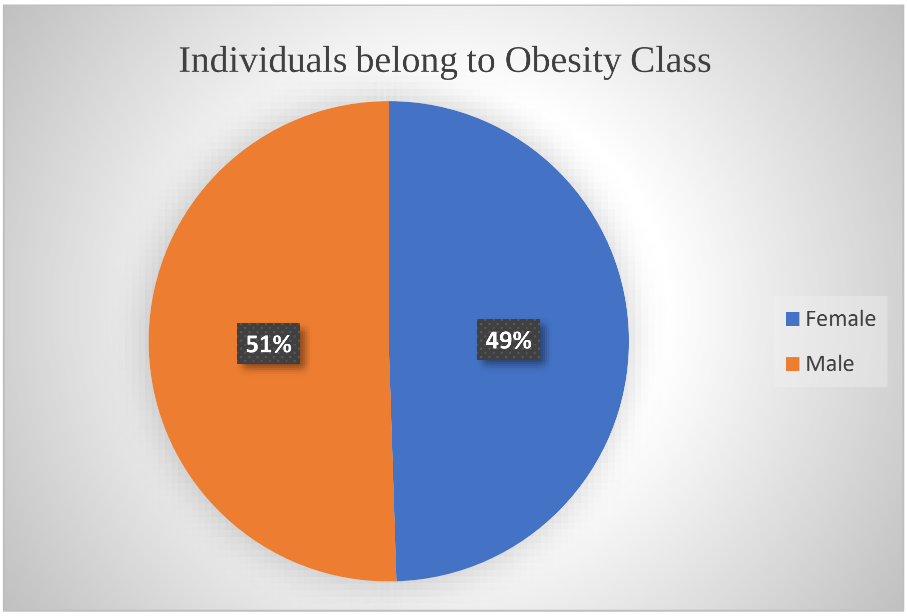
1) Bar plot :



From above Bar Plot of Gender & Obesity we clearly see that number of Male and Female individuals in the data are almost nearly equal i.e count of Male is 1068 (51%) and that of Female is 1043 (49%) . From this we conclude that data is not biased.

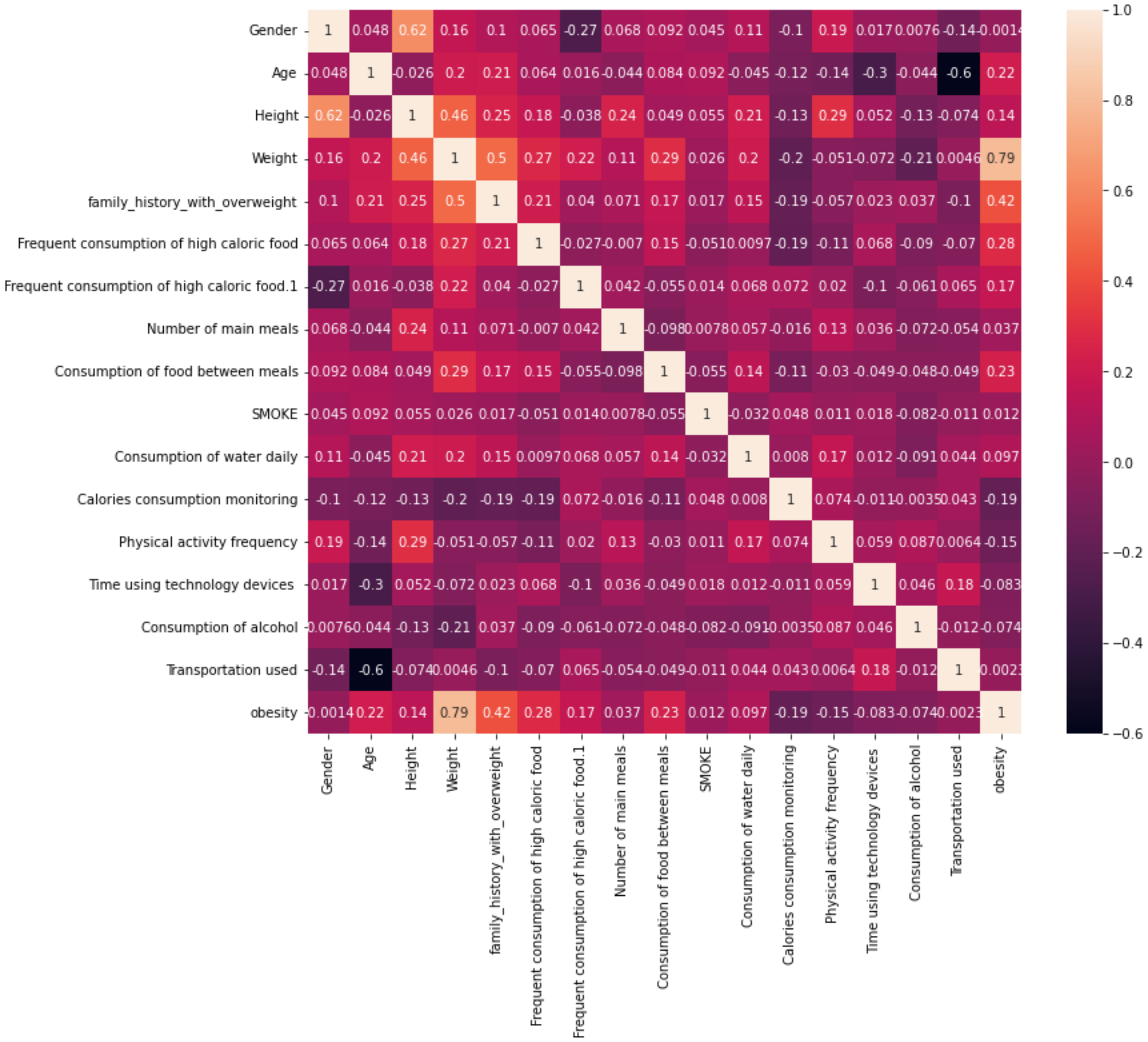
From the plot of Obesity we see that there are 972 (46%) individuals belongs to obesity class and remaining 1139 (54%) does not belongs to obesity class.

2) Pie Chart :



From above Pie chart of Individuals belong to Obesity Class we clearly see that 51% of Male belongs to obesity whereas 49% of Female belongs to obesity i.e there is slight difference. From this pie chart we can see that percentage of Male Individuals belong to Obesity Class is more than that of Female .

3) Analysis of Correlation between factors affecting on dataset by using Heatmap



Interpretation:

From above heatmap we observed that there is positive correlation between ‘family history with overweight’ and Obesity Type i.e Obesity and also there is positive correlation between “Weight” and “Obesity Type” which is obvious that is as Weight increases Obesity also increases.

There is negative correlation between “Calories consumption monitoring” and “Obesity Type” i.e as we monitor Calories consumption there is high probability that Obesity will decreases.

Normality Test

Shapiro – Wilk test for the assessment of normality:

Shapiro-Wilk Test:

The Shapiro-wilk test, proposed by Shapiro in 1965, is considered the most reliable test for non-normality for small to medium sized samples by many authors.

The test statistic is defined as:

$$W = \frac{\left(\sum_{i=1}^{N} a_i x_{(i)}\right)^2}{\sum_{i=1}^{N}(x_i - \bar{x})^2},$$

The test for normality has been found to be the most powerful test in most situations. It is the ratio of two estimates of the variance of a normal distribution based on a random sample of n observations. The numerator is proportional to the square of the best linear estimator of the standard deviation. The test statistic W may be written as the square of the Pearson correlation coefficient between the ordered observations and a set of weights which are used to calculate the numerator. W is roughly measure of the straightness of the normal quantile – quantile plot. Hence, the closer is to one, the more normal the sample is. The probability values for W are valid. Shapiro-Wilk test is fairly powerful test. Not good with samples or discrete data. Shapiro – wilk test good power with symmetrical, short and long tails, Good with symmetry.

Hypothesis :-

H_0: The data is normal.

H_1: The data is not normal.

Result :

ShapiroResult

statistic=0.38 , pvalue=0.0

Interpretation:

Since P-Value < Level of significance(0.05)

Therefore we reject null hypothesis.

We conclude that our data is not Normal.

Hypothesis Testing

Chi-Square Test :

Hypothesis :

H_0:Obesity and Gender are independent

H_1:Obesity and Gender are not independent

Calculation :

$$\chi^2_{cal} = \sum_{i=1}^{m} \sum_{j=1}^{n} \frac{(Oij-Eij)^2}{Eij} \sim \chi^2_{tab(n-1)} d.f.$$

$$\chi^2_{tab} = \chi^2_{(m-1)(n-1)}\text{d.f.}$$

2×2 contingency table:

Gender obesity	No	Yes	Total
Male	a = 562	b = 481	a+b = 1043
Female	c = 577	d = 491	c+d = 1068
Total	a+c = 1139	b+d = 972	N = 2111

Pearson's Chi-squared test with Yates' continuity correction

data: y

X-squared = 0.00049823, df = 1, p-value = 0.9822

Result:

Here P Value is greater than α (=0.05)

Conclusion:

We accept H_0 at 5% level of significance i.e Obesity and Gender are independent.

Data Mining Classifier

In this section we have started Modelling – Predicting the attrition dataset by using python software. we have split the dataset into training and testing data so that, the model is trained on training data and predicts the result on test data. Here, the target variable is ‘Obesity’. Different classifiers are used below and their performance is measured.

Logistic Regression:-

Logistic regression is a supervised learning classification algorithm used to predict the probability of a target variable. The nature of target or dependent variable is dichotomous, which means there would be only two possible classes. In simple words, the dependent variable is binary in nature having data coded as either 1 (stands for success/yes) or 0 (stands for failure/no). Mathematically, a logistic regression model predicts P(Y=1) as a function of X. It is one of the simplest ML algorithms that can be used for various classification problems such as spam detection, Diabetes prediction, cancer detection, heart failure prediction etc.

Confusion Matrix in Machine Learning:

Confusion matrix is a matrix used to determine the performance of the classification models for a given set of test data. It can only be determined if the true values for test data are known. The matrix itself can be easily understood, but the related terminologies may be confusing. Since it shows the errors in the model performance in the form of a matrix, hence also known as an error matrix. It looks like the below table:

n = Total Predictions	Predicted : No	Predicted: Yes
Actual: No	True Negative	False Positive
Actual: Yes	False Negative	True Positive

The above table has the following cases:

o True Negative: Model has given prediction No, and the real or actual value was also No.

o True Positive: The model has predicted yes, and the actual value was also true.

o False Negative: The model has predicted no, but the actual value was Yes, it is also called as Type-II error.

o False Positive: The model has predicted Yes, but the actual value was No. It is also called a Type-I error.

ACCURACY: Accuracy is used to find the correct values; it is the sum of all true values divided by total values.

ROC-AUC :

ROC AUC stands for Receiver Operating Characteristic - Area Under Curve. It is a technique to compare classifier performance. In this technique, we measure the area under the curve (AUC). A perfect classifier will have a ROC AUC equal to 1, whereas a purely random classifier will have a ROC AUC equal to 0.5

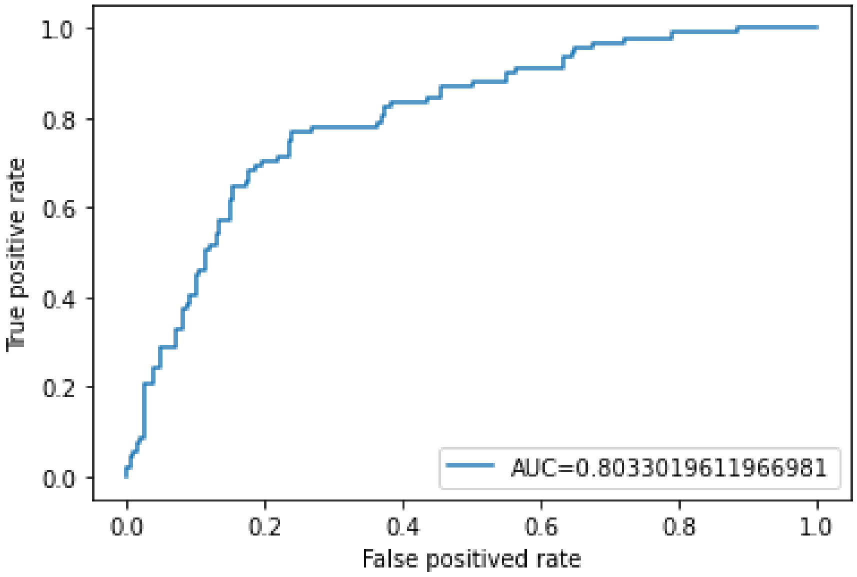
So, ROC AUC is the percentage of the ROC plot that is underneath the curve.

Logistic Regression:

Confusion Matrix :

n = 423	Predicted : No	Predicted: Yes
Actual: No	216	18
Actual : Yes	31	158

Receiver Operating Characteristic Curve :



Accuracy= 0.88

Interpretation:

From the confusion matrix, the classifier has made a total 423 predictions. Out of 423 predictions, 374 are true predictions and 49 are incorrect predictions according to test data. The model has predicted 18 which belong to obesity class but the actual value was not from obesity class, it is also Type-II error. The model has predicted 31 which does not belongs to obesity class and actually they belongs, which is Type-I error. For this model the error type is Type-II error. The accuracy of this model is 88% with 12% misclassification.

From the ROC-AUC graph we can see that Area Under Curve is 0.8 which is in between the range 0-1. And hence we can conclude that our classifier performance is good.

Hence the model represents good discrimination.

Conclusion

1) We conclude that the most affecting factor on obesity is family history with obesity i.e. obesity disease is transferred by genetically. If someone had in your family such disease then there are more chances of obesity in next generation.
2) From chi-square test we conclude that there is no relation between Gender and Obesity i.e Gender and Obesity are independent.
3) By using Logistic Regression which is best fitted to our data , we can predict the probability that the individual is going to belong to Obesity Class (i.e. our target variable) or not.
4) From heatmap we conclude that as we monitor Calories consumption there is high probability that Obesity will decreases.

References

i) Functional Programming in Python – David Mertz

ii) Logistic Regression Models – Joseph M. Hilbe

iii) www.google.com

iv) www.researchgate.com

Appendix

```
#chi-square test :
x=c(562,481,577,491);x
n=2;n
m=2;m
y=matrix(x,nrow=n,ncol=m);y
chisq.test(y)
```

```
#python
import pandas as pd
df = pd.read_csv("C:/Users/Desktop/EOBOH/data/Scrap Data coded n.csv")
df
df.isnull().sum()
print(df.info())
```

Label Encoder :

```
# import sklearn module to encode labled data
from sklearn.preprocessing import LabelEncoder
le = LabelEncoder()
df['Consumption of alcohol']=le.fit_transform(df['Consumption of alcohol'])
df['Gender']=le.fit_transform(df['Gender'])
df['Consumption of food between meals']=le.fit_transform(df['Consumption of food between meals'])
df['family_history_with_overweight']=le.fit_transform(df['family_history_with_overweight'])
df['Frequent consumption of high caloric food']=le.fit_transform(df['Frequent consumption of high caloric food'])
```

```
df['SMOKE']=le.fit
_transform(df['SMOKE'])
df['Calories consumption monitoring']=le.fit_transform(df['Calories consumption monitoring'])
df['Transportation used']=le.fit_transform(df['Transportation used'])
df
df.describe()
```

Feature Score :

```
dataframe = df
dataframe
array = dataframe.values
x = array[:,0:8]
y = array[:,8]
# Import the necessary libraries first
from sklearn.feature_selection import SelectKBest
from sklearn.feature_selection import chi2
bestfeature=SelectKBest(score_func=chi2,k=11)
fit=bestfeature.fit(x,y)
fit
dataScore=pd.DataFrame(fit.scores_)
datacolumns=pd.DataFrame(x.columns)
featureScores=pd.concat([datacolumns,dataScore],axis=1)
featureScores.columns=['factor','Score']
featureScores
```

```
pd.DataFrame(featureScores.nlargest(8,'Score'))
```

Heatmap :

```
import seaborn as sns

import matplotlib as plt

import matplotlib.pyplot as plt

plt.figure(figsize=(12,10))

sns.heatmap(df.corr(),annot=True)
```

Logistic Regression :

```
x=df[['Weight','Age','Calories consumption monitoring','family_history_with_overweight','Physical activity frequency','Frequent consumption of high caloric food','Consumption of food between meals']]
y=df['obesity']
from sklearn.model_selection import train_test_split
from sklearn.metrics import classification_report
from sklearn.metrics import accuracy_score
from sklearn.metrics import confusion_matrix
xtrain,xtest,ytrain,ytest=train_test_split(x,y,test_size=0.2,random_state=0)
xtrain
from sklearn.linear_model import LogisticRegression
model=LogisticRegression()
model.fit(xtrain,ytrain)


model.predict(xtest)

model.predict(ytest)

xpred = model.predict(xtest)

ypred = model.predict(ytest)
```

```
ypred
model.score(xtest, ytest)
```

Roc curve :

```
y_pred_proba=model.predict_proba(xtest)[::,1]
fpr,tpr,=metrics.roc_curve(ytest,ypred_proba)
auc=metrics.roc_auc_score(ytest,ypred_proba)
auc
TP=cm[0][0]
TN=cm[1][1]
FN=cm[1][0]
FP=cm[0][1]
print('Testing accuracy:',(TP+TN)/(TP+TN+FN+FP))
accuracy_score(ytest,ypred)
print(classification_report(ytest,ypred))
log_roc_auc=roc_auc_score(ytest,model.predict(xtest))
fpr,tpr,threshold=log_roc_auc(ytest,ypred[:,1])
plt.figure(figsize=(6,6))
plt.plot(fpr,tpr,color='red',label="Logit Model 1 (area=%0.2f)"%log_roc_auc)
plt.plot([0,1][0,1],color='darkblue',linestyle='--')
plt.xlim([-0.05,1.05])
plt.yim([-0.05,1.05])
plt.xlabel("False Positive Rate")
plt.yabel("True Positive Rate")
plt.title("Reciever Opertaing Characteristics")

plt.legend(loc="lower right")

plt.savefig("log_ROC")

plt.show()
```