MULTIMEDIA UNIVERSITY

TDS2101 Fundamental Data Science

Trimester 2, 2022/2023

Individual Report :

Exploring & Developing a Model for Obesity Estimation Based on Eating Habits and Physical Condition

Name	Student ID	Email
Lee Tian Xin	1211301744	1211301744@student.mmu.edu.my

Table of Contents

1. Introduction

Obesity is a major health problem that is becoming more common worldwide. Therefore, estimating obesity levels is crucial for identifying individuals who require intervention and monitoring.

The aim of this project is to develop a predictive model that can accurately estimate obesity levels based on the provided dataset. Besides that, identify the key factors that contribute to obesity and provide insights into the associated risk factors.

This project utilises a dataset that includes data from individuals in Mexico, Peru, and Colombia, based on their eating habits and physical condition. After cleaning, processing the dataset, and exploring each variable, I will employ a supervised classification technique.

The proposed methodology involves feature selection and model construction using Random Forest, decision tree, KNN, SVM, and Logistic Regression. These models will be compared to determine the most suitable approach for estimating the obesity level. After identifying the most suitable model, I will proceed with its evaluation using cross validation to determine its performance in estimating obesity levels. Additionally, I will analyse the model's key predictors to gain insights into the factors that significantly contribute to obesity.

Therefore, this project will show the most suitable methodology for estimating obesity levels and identifying the key factors contributing to this health issue.

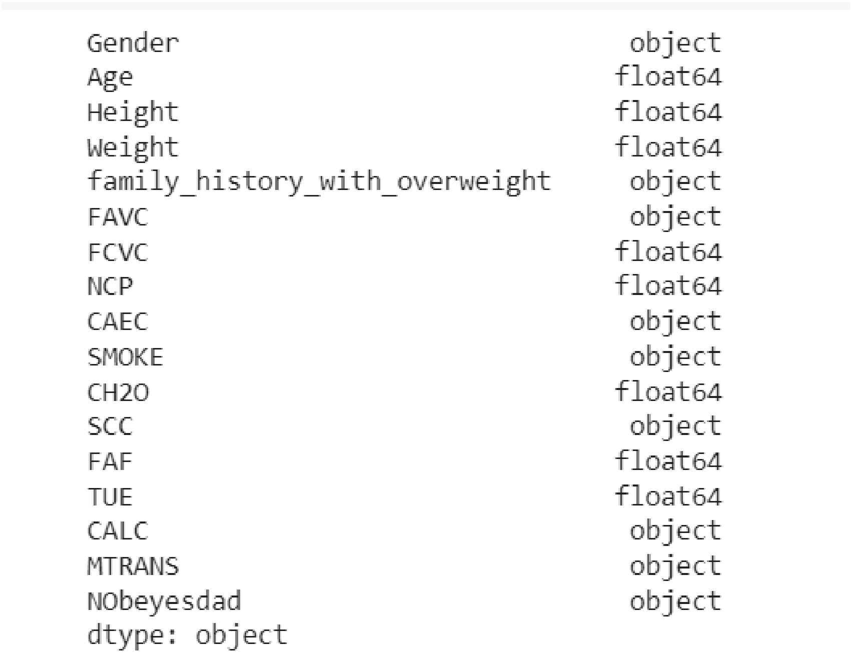
2. Project Objective

I. Develop a predictive model that can accurately estimate obesity levels based on the provided dataset.

II. To identify the key predictors of obesity from the dataset and investigate the relationships between these factors.

3. Data Collection

The data was collected from UCI, and it included 17 variables with 2111 columns. As the info of the variables were showing below:



```
Gender                              object
Age                                float64
Height                             float64
Weight                             float64
family_history_with_overweight      object
FAVC                                object
FCVC                               float64
NCP                                float64
CAEC                                object
SMOKE                               object
CH2O                               float64
SCC                                 object
FAF                                float64
TUE                                float64
CALC                                object
MTRANS                              object
NObeyesdad                          object
dtype: object
```

Figure 3.1: Data Type of variables

Below are the dataset variable description:

- Gender - The gender of the individual (Female / Male)
- Age - The age of the individual (Numeric)
- Height - The height of the individual (Numeric)
- Weight - The weight of the individual (Numeric)
- Family_history_with_overweight - Indicates whether the individual has a family history of overweight or obesity. (Yes / No)
- FAVC - Frequent consumption of high caloric food(Yes / No)
- FCVC - Frequency of consumption of vegetables (1, 2, 3)
- NCP - Number of main meals consumed per day (1, 2, 3 or 4)
- CAEC - Eat any food between meals (No, Sometimes, Frequently or Always)
- SMOKE - Smoking habit (Yes/ No).
- CH20 - Daily water consumption (1, 2 or 3)
- SCC - Calorie monitoring (Yes/ No)
- FAF - Frequency of physical activity performed (0, 1, 2 or 3)
- TUE - Time using technology devices per day (0, 1 or 2)
- CALC - Consumption of Alcohol (No, Sometimes, Frequently or Always)

- MTRANS - Mode of transportation

(Automobile, motorbike, bike, public transportation or walking)

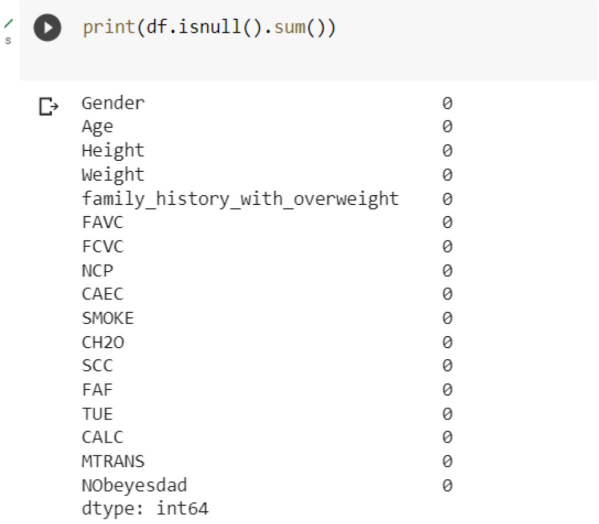
- Nobeyesdad - Degree of obesity

(Insufficient_Weight, Normal_Weight, Overweight_Level_I, Overweight_Level_II, Obesity_Type_I, Obesity_Type_II, Obesity_Type_III)

4. Data Processing and Cleaning

4.1. Data Cleaning

- All of the data are unique and contains null values



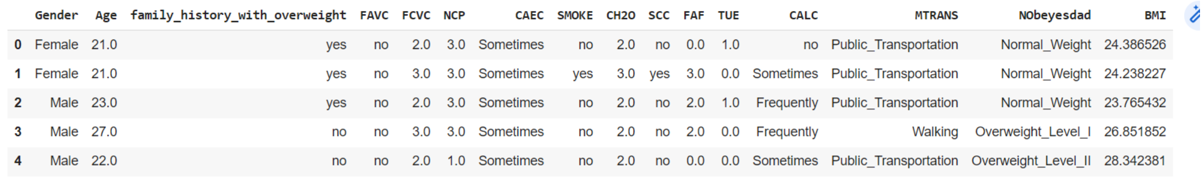
```
print(df.isnull().sum())
```

```
Gender                            0
Age                               0
Height                            0
Weight                            0
family_history_with_overweight    0
FAVC                              0
FCVC                              0
NCP                               0
CAEC                              0
SMOKE                             0
CH2O                              0
SCC                               0
FAF                               0
TUE                               0
CALC                              0
MTRANS                            0
NObeyesdad                        0
dtype: int64
```

Figure 4.1: Data null checking

4.2. Data Transformation

- Combining Height and Weight into BMI
- Drop height and weight



	Gender	Age	family_history_with_overweight	FAVC	FCVC	NCP	CAEC	SMOKE	CH2O	SCC	FAF	TUE	CALC	MTRANS	NObeyesdad	BMI
0	Female	21.0	yes	no	2.0	3.0	Sometimes	no	2.0	no	0.0	1.0	no	Public_Transportation	Normal_Weight	24.386526
1	Female	21.0	yes	no	3.0	3.0	Sometimes	yes	3.0	yes	3.0	0.0	Sometimes	Public_Transportation	Normal_Weight	24.238227
2	Male	23.0	yes	no	2.0	3.0	Sometimes	no	2.0	no	2.0	1.0	Frequently	Public_Transportation	Normal_Weight	23.765432
3	Male	27.0	no	no	3.0	3.0	Sometimes	no	2.0	no	2.0	0.0	Frequently	Walking	Overweight_Level_I	26.851852
4	Male	22.0	no	no	2.0	1.0	Sometimes	no	2.0	no	0.0	0.0	Sometimes	Public_Transportation	Overweight_Level_II	28.342381

Figure 4.2: Data after transformation

5. Exploratory Data Analysis

In this section, the analysis is divided into two parts: univariate analysis and bivariate analysis.

5.1. Univariate Analysis

The univariate analysis further consists of two parts: categorical data and numerical data.

The first graph displays a 3x3 plot of categorical data, consisting of nine bar graphs, as shown below:

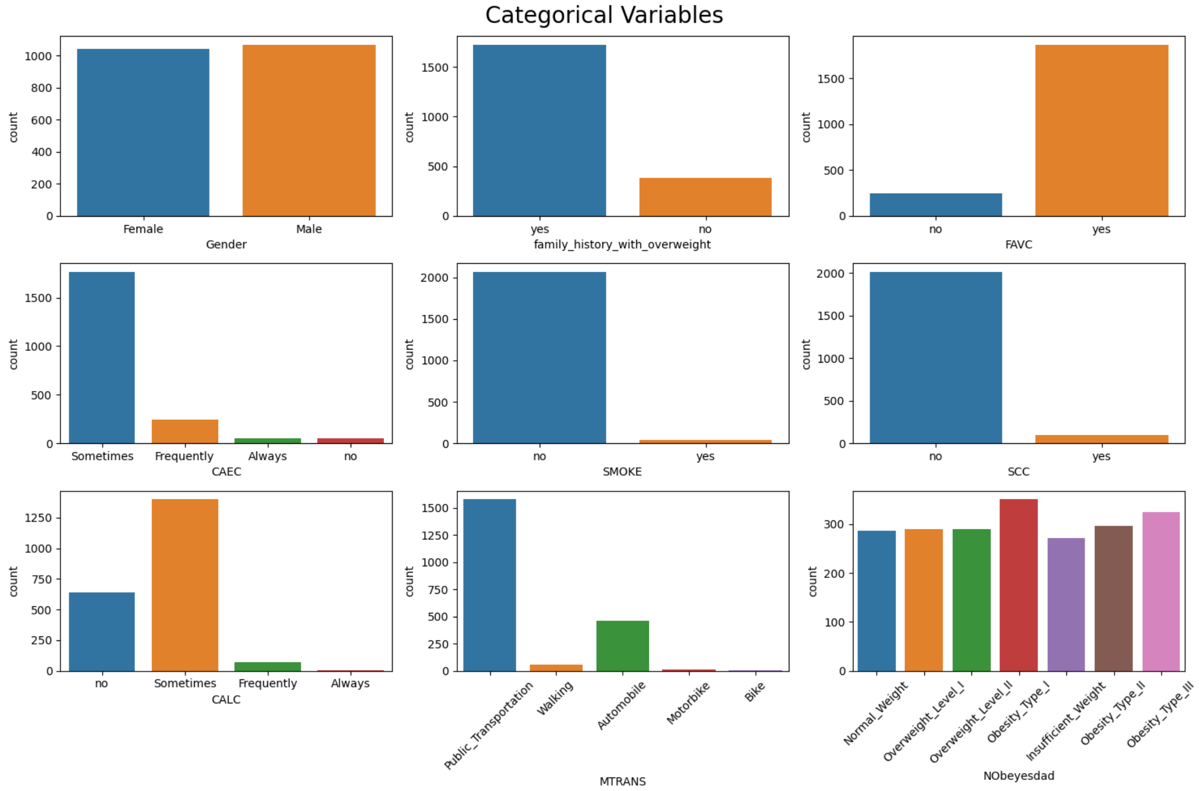


Figure 5.1.1: Bar chart of Categorical variables

Here is the description for each bar plot (Figure 5.1.1), following from left to right and row by row:

Gender:

The bar chart shows an almost equal distribution between female and male individuals.

Family History with Overweight:

Most of the people in the dataset indicate having a family history of being overweight.

FAVC (Frequent Consumption of High Caloric Food):

Most of the people in the dataset consume high caloric food.

CAEC (Eat any food between meals)

Among the individuals in the dataset, the majority (around 90%) reported eating any food between meals sometimes, while only a small percentage indicated frequent, constant, or no eating any food between meals.

SMOKE (Smoking Habit):

Nearly 99% of people in the dataset don’t have a smoking habit.

SCC (Calorie Monitoring):

Nearly 99% of people in the dataset don’t have calorie monitoring.

CALC (Consumption of Alcohol):

Among the individuals in the dataset, the majority (around 60%) reported engaging in CALC, 30% of people not CALC, and followed frequently and always.

MTRANS (Mode of Transportation):

Most people (around 70%) used to use public transportation, followed by automobiles, walking, motorbikes and bikes.

NObeyesdad (Obesity Level):

People are evenly distributed across the different obesity levels, only obesity type I slightly outnumber the other categorical.

The second graph displays a 2x4 plot of numerical data, consisting of seven histogram, as shown below:

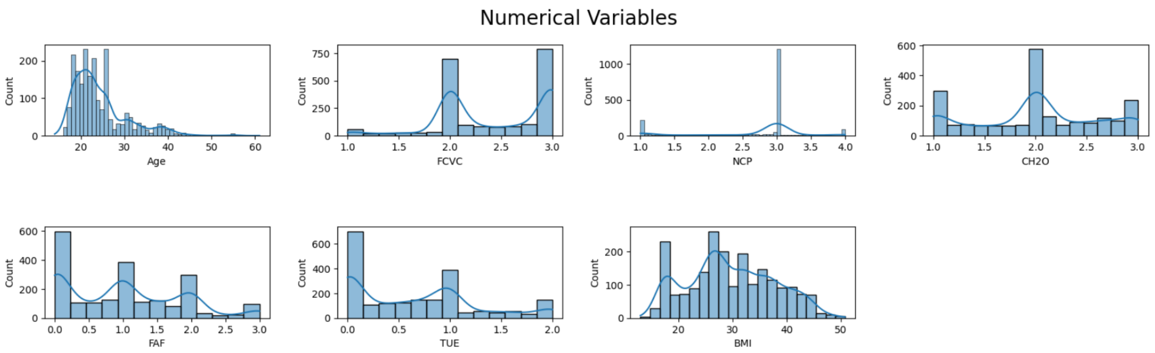


Figure 5.1.2: Histogram of Numerical variables

Here is the description for each histogram (Figure 5.1.2), following from left to right and row by row:

- **Age - The age of the individual (Numeric)**

The graph shows that the people in the data set are around 20- 30 years old, it is a right skewed graph.

- **FCVC - Frequency of consumption of vegetables (1, 2, 3)**

It can also be viewed as ordinal data, as we can see frequency of consumption of vegetables is the highest , followed by 2 and 1.

- **NCP - Number of main meals consumed per day (1, 2, 3 or 4)**

Most people consumed 3 meals per day as seen in histogram and followed by 1, 4, 2.

- **CH20 - Daily water consumption (1, 2 or 3)**

Most people consumed water at 2L as seen in histogram, followed by 1 L and 3L.

- **FAF - Frequency of physical activity performed (0, 1, 2 or 3)**

We can assume 3 as always, 2 as frequently, 1 as sometimes, 0 as seldom. As the graph shows, most people don't do any physical activity followed by 2 , 1 and 0.

- **TUE - Time using technology devices per day (0, 1 or 2)**

Most people using technology devices only 0 , followed by 1 and 2, have some people in intervals 0-1 or 1-2.

- **BMI**

Most people BMI is around 18 - 30 (kg/m^2)

For checking outliers, I have implemented box plot as shown below:

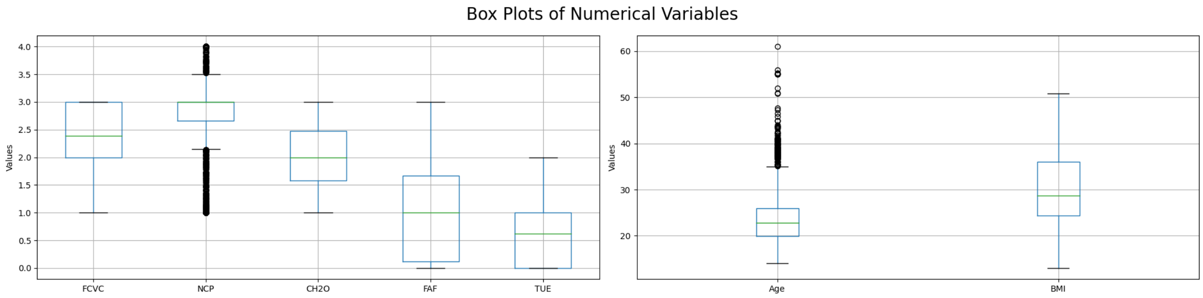


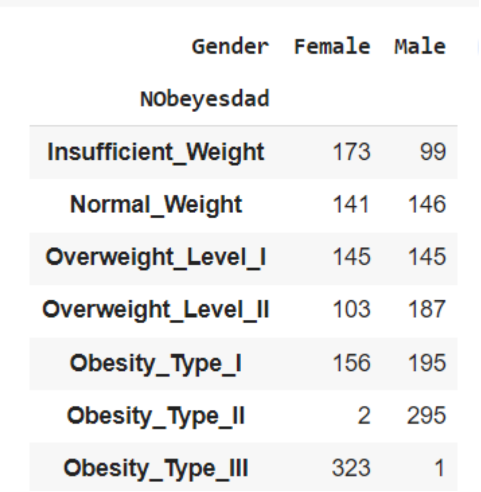
Figure 5.1.3 : Box Plot of Numerical Variables

- For the Age box plot, it is mostly located around 24, and it has many upper outliers.
- For the FCVC box plot, it is mostly located around 2.3, and it has no outliers.
- For the NCP box plot, it is mostly located around 2.7, with a significant number of outliers, located both above and below this range.
- For the CH2O box plot, it is mostly located around 2, and there are no outliers.
- For the FAF box plot, it is mostly located around 1, and there are no outliers.
- For the TUE box plot, it is mostly located around 0.7, and there are no outliers.
- For the BMI box plot, it is mostly located around 29, and there are no outliers.

5.2. Bivariate Analysis

For Bivariate analysis, I have explored the relationship between categorical variables with obesity level. I have used crosstab and barchart.

Gender vs Obesity Level:



Gender	Female	Male
NObeyesdad		
Insufficient_Weight	173	99
Normal_Weight	141	146
Overweight_Level_I	145	145
Overweight_Level_II	103	187
Obesity_Type_I	156	195
Obesity_Type_II	2	295
Obesity_Type_III	323	1

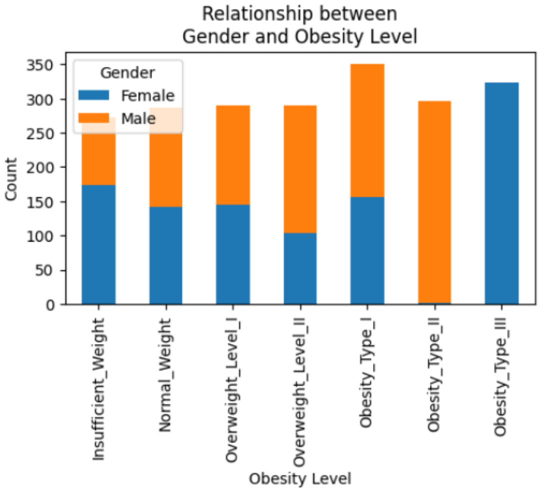
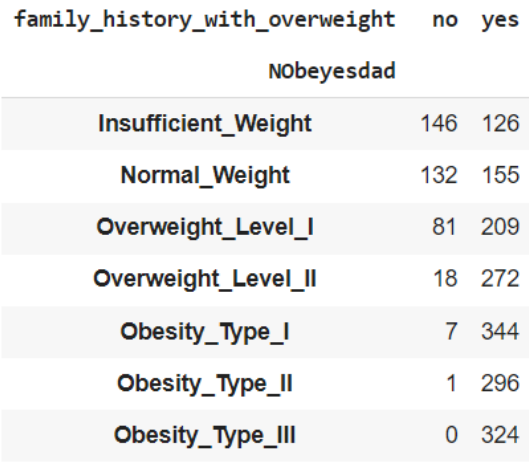


Figure 5.2.1: Crosstab and Bar chart of relationship between gender and obesity

Bar chart shows that females tend to have insufficient weight.

Family History with Overweight vs Obesity Level:



family_history_with_overweight	no	yes
NObeyesdad		
Insufficient_Weight	146	126
Normal_Weight	132	155
Overweight_Level_I	81	209
Overweight_Level_II	18	272
Obesity_Type_I	7	344
Obesity_Type_II	1	296
Obesity_Type_III	0	324

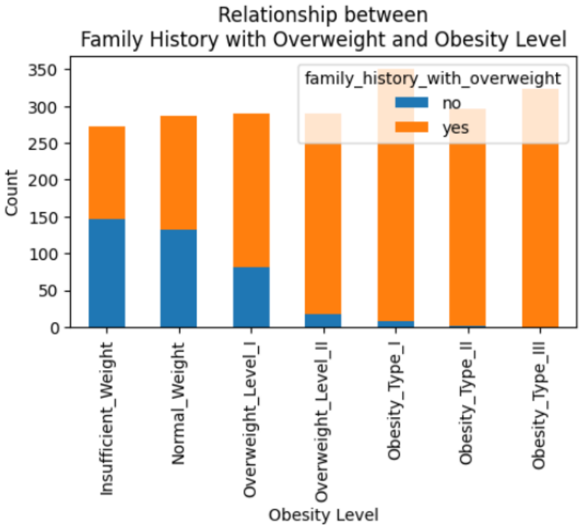
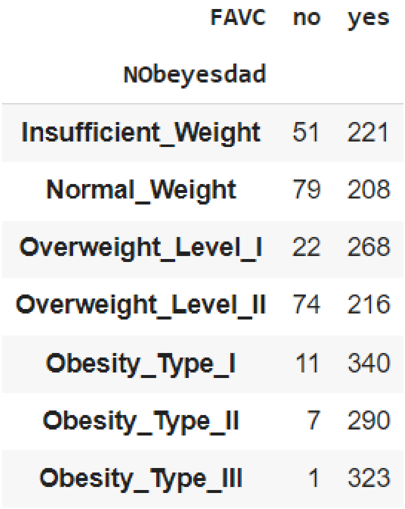


Figure 5.2.2: Crosstab and Bar chart of relationship between family history and obesity level

Bar chart shows that individuals with a family history of overweight are more likely to be in the higher categories of obesity levels.

FAVC (Frequent Consumption of High Caloric Food) vs Obesity Level:



FAVC	no	yes
NObeyesdad		
Insufficient_Weight	51	221
Normal_Weight	79	208
Overweight_Level_I	22	268
Overweight_Level_II	74	216
Obesity_Type_I	11	340
Obesity_Type_II	7	290
Obesity_Type_III	1	323

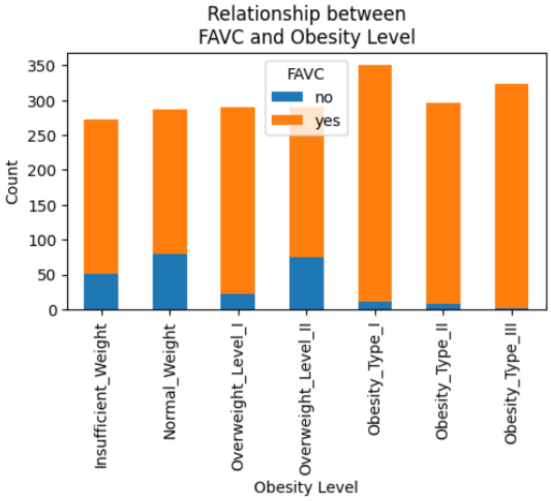
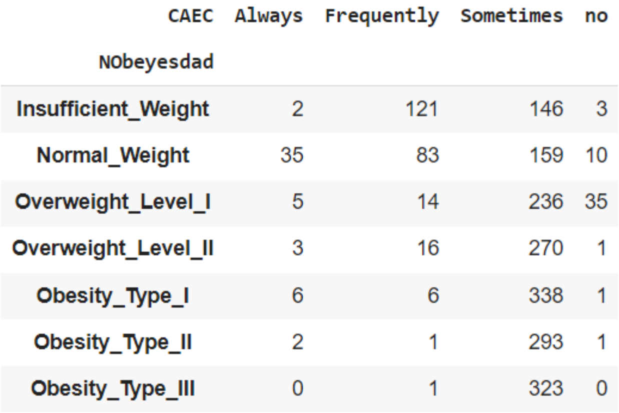


Figure 5.2.3: Crosstab and Bar chart of relationship between FAVC and obesity level

Bar chart shows that individuals with frequent consumption of high caloric food tend to have potential to get obese.

CAEC (Eat any food between meals) vs Obesity Level:



CAEC NObeyesdad	Always	Frequently	Sometimes	no
Insufficient_Weight	2	121	146	3
Normal_Weight	35	83	159	10
Overweight_Level_I	5	14	236	35
Overweight_Level_II	3	16	270	1
Obesity_Type_I	6	6	338	1
Obesity_Type_II	2	1	293	1
Obesity_Type_III	0	1	323	0

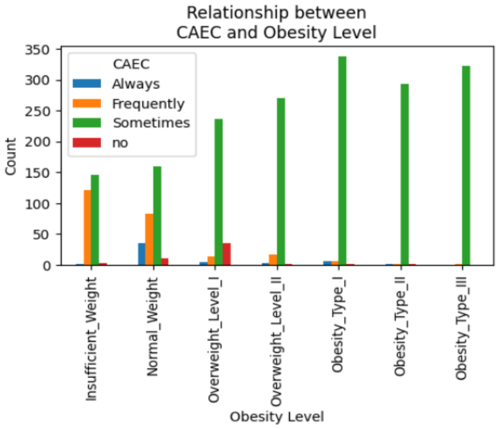
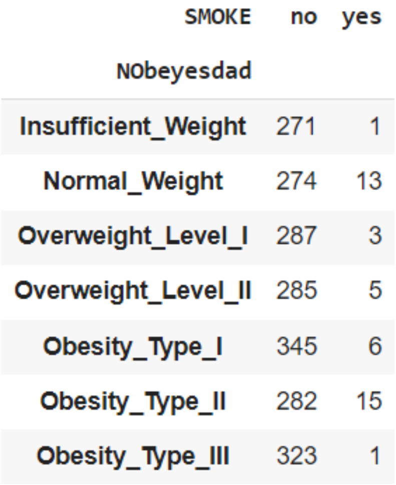


Figure 5.2.4: Crosstab and Bar chart of relationship between CAEC and obesity level

The bar chart highlights a notable trend: individuals who sometimes eat between meals have an increasing trend across weight categories. In contrast, those who frequently indulge in snacks between meals show a decreasing trend across weight categories.

SMOKE (Smoking Habit) vs Obesity Level :



SMOKE NObeyesdad	no	yes
Insufficient_Weight	271	1
Normal_Weight	274	13
Overweight_Level_I	287	3
Overweight_Level_II	285	5
Obesity_Type_I	345	6
Obesity_Type_II	282	15
Obesity_Type_III	323	1

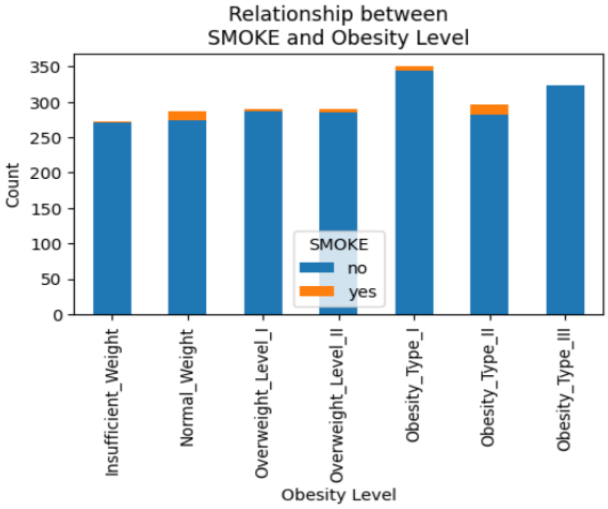
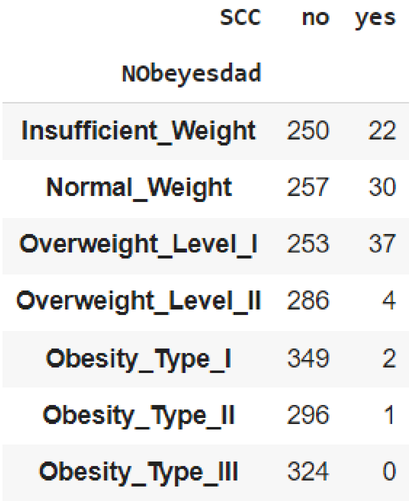


Figure 5.2.5: Crosstab and Bar chart of relationship between Smoke and obesity

The bar chart illustrates the distribution of obesity levels among individuals categorised as 'smokers' and 'non-smokers.' It shows that there is a relatively balanced distribution of obesity levels between the two groups, with no significant difference observed in the prevalence of obesity based on smoking status.

SCC (Calorie Monitoring) vs Obesity Level:



SCC NObeyesdad	no	yes
Insufficient_Weight	250	22
Normal_Weight	257	30
Overweight_Level_I	253	37
Overweight_Level_II	286	4
Obesity_Type_I	349	2
Obesity_Type_II	296	1
Obesity_Type_III	324	0

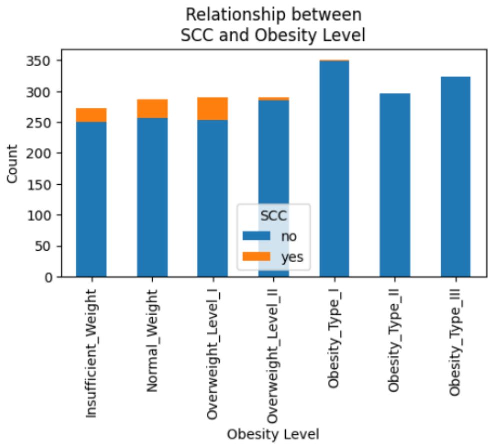
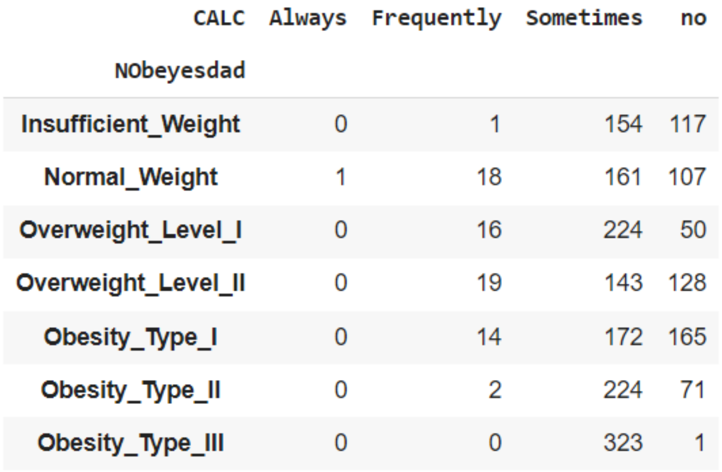


Figure 5.2.6: Crosstab and Bar chart of relationship between Scc and obesity

Bar chart shows that the person who has calorie monitoring tends to have normal weight.

CALC (Consumption of Alcohol) vs Obesity Level:



CALC NObeyesdad	Always	Frequently	Sometimes	no
Insufficient_Weight	0	1	154	117
Normal_Weight	1	18	161	107
Overweight_Level_I	0	16	224	50
Overweight_Level_II	0	19	143	128
Obesity_Type_I	0	14	172	165
Obesity_Type_II	0	2	224	71
Obesity_Type_III	0	0	323	1

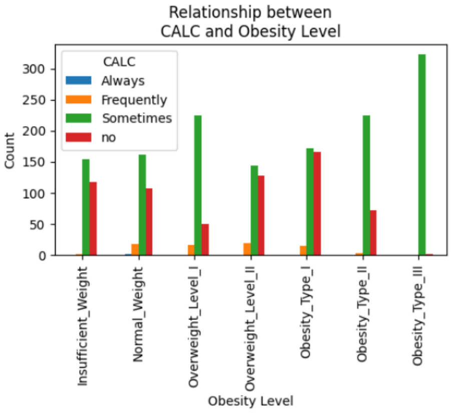
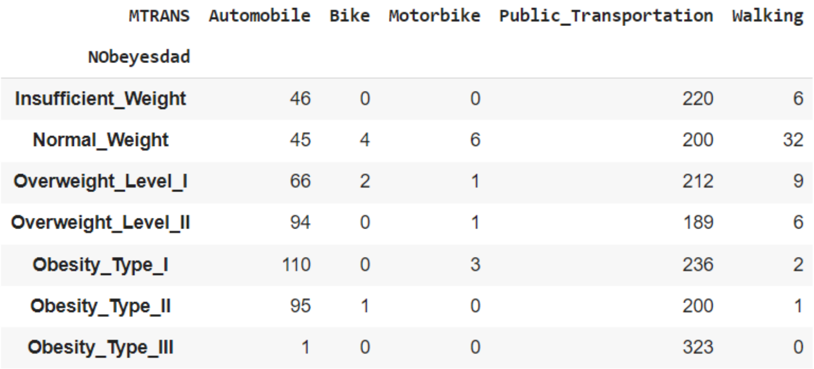


Figure 5.2.7: Crosstab and Bar chart of relationship between CALC and obesity

Bar chart shows that the person who consumption of alcohol sometimes have more potential to have obesity compared to person who no consumption of alcohol

MTRANS (Mode of Transportation) vs Obesity level :



MTRANS	Automobile	Bike	Motorbike	Public_Transportation	Walking
NObeyesdad					
Insufficient_Weight	46	0	0	220	6
Normal_Weight	45	4	6	200	32
Overweight_Level_I	66	2	1	212	9
Overweight_Level_II	94	0	1	189	6
Obesity_Type_I	110	0	3	236	2
Obesity_Type_II	95	1	0	200	1
Obesity_Type_III	1	0	0	323	0

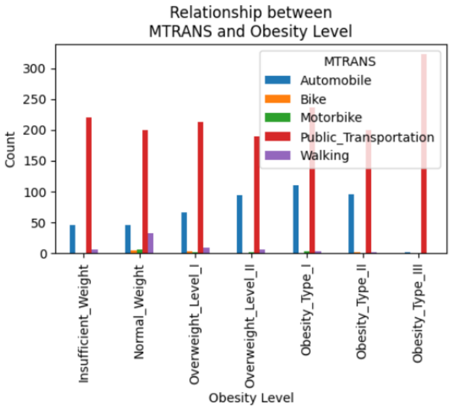


Figure 5.2.8: Crosstab and Bar chart of relationship between MTRANS and obesity level

The bar chart shows varying obesity levels based on transportation modes. Public transportation users have the highest prevalence of obesity type III, while walking, motorbike, and bike users tend to have lower obesity rates, indicating a normal weight tendency. Automobile users show a higher likelihood of obesity type I. Transportation choices may influence obesity levels, with active modes associated with healthier weights and public transportation linked to higher obesity rates.

To explore the relationship between numerical variables, we have implemented heatmap.

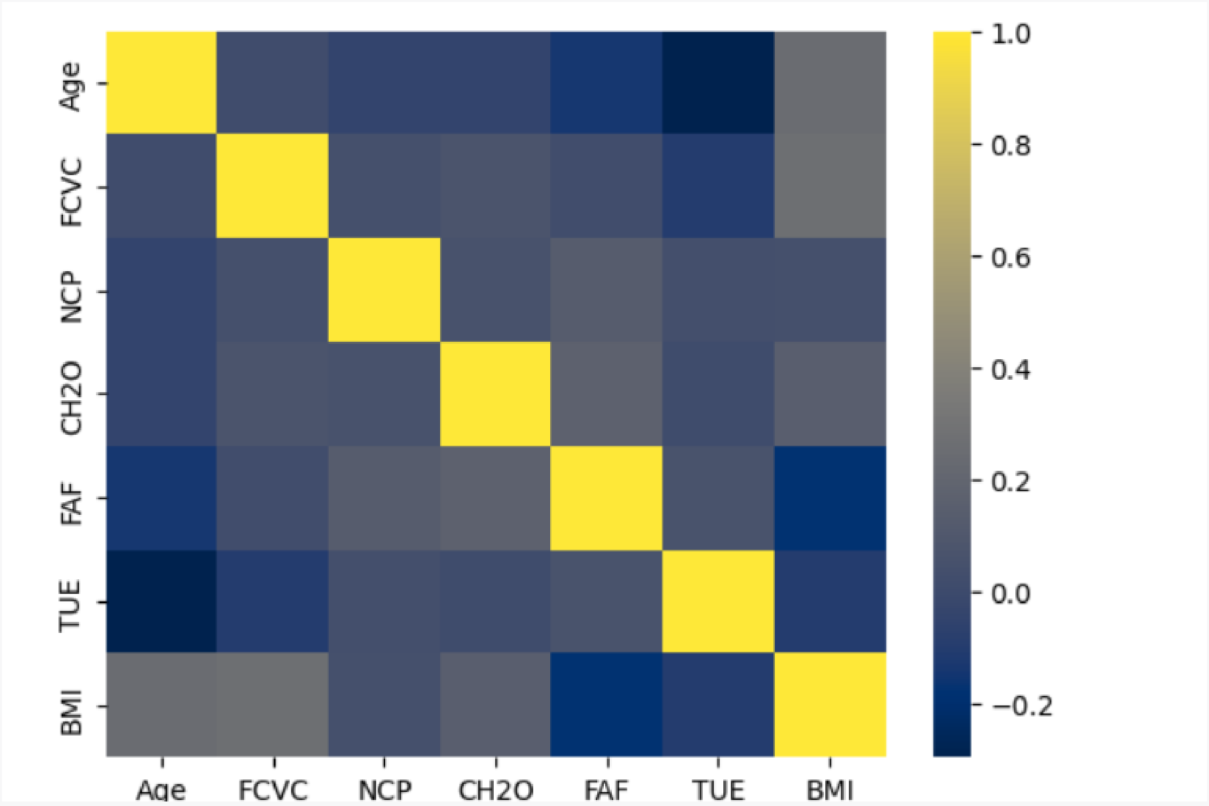
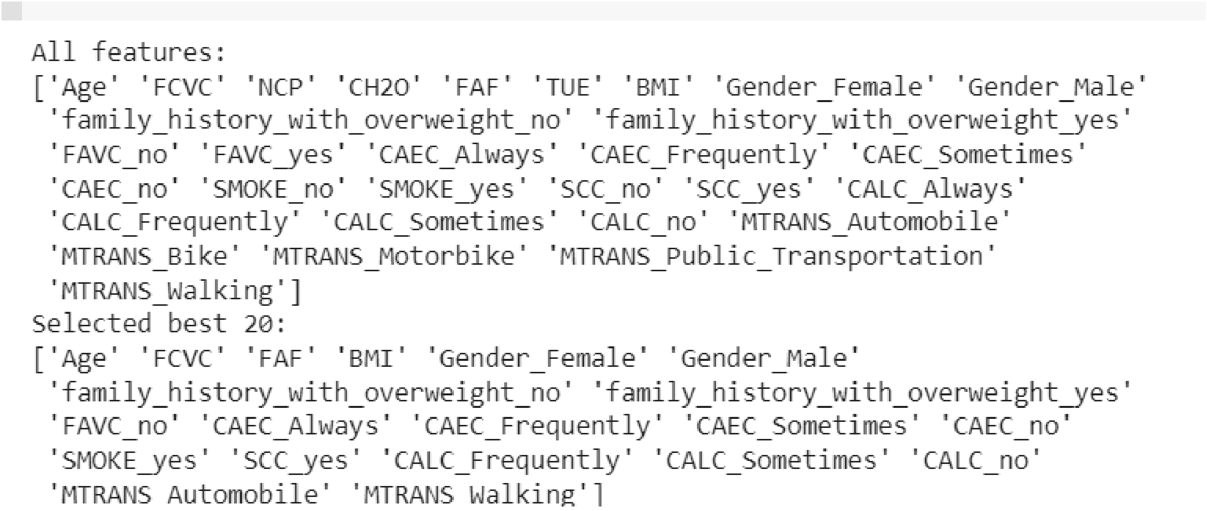


Figure 5.2.8: Heatmap for numerical variables

As we observe from the heatmap, all numericals variables have weak relationships between each other. To dive into more details, we can observe that FCVC, Age, and CH2O exhibit a moderate association with BMI.

6. Feature Selection

After building the data, I implemented feature selection using chi2 scoring and selected the top 20 best features from the training data.



```
All features:
['Age' 'FCVC' 'NCP' 'CH2O' 'FAF' 'TUE' 'BMI' 'Gender_Female' 'Gender_Male'
 'family_history_with_overweight_no' 'family_history_with_overweight_yes'
 'FAVC_no' 'FAVC_yes' 'CAEC_Always' 'CAEC_Frequently' 'CAEC_Sometimes'
 'CAEC_no' 'SMOKE_no' 'SMOKE_yes' 'SCC_no' 'SCC_yes' 'CALC_Always'
 'CALC_Frequently' 'CALC_Sometimes' 'CALC_no' 'MTRANS_Automobile'
 'MTRANS_Bike' 'MTRANS_Motorbike' 'MTRANS_Public_Transportation'
 'MTRANS_Walking']
Selected best 20:
['Age' 'FCVC' 'FAF' 'BMI' 'Gender_Female' 'Gender_Male'
 'family_history_with_overweight_no' 'family_history_with_overweight_yes'
 'FAVC_no' 'CAEC_Always' 'CAEC_Frequently' 'CAEC_Sometimes' 'CAEC_no'
 'SMOKE_yes' 'SCC_yes' 'CALC_Frequently' 'CALC_Sometimes' 'CALC_no'
 'MTRANS Automobile' 'MTRANS Walking']
```

Figure 6.1 : Feature selection result

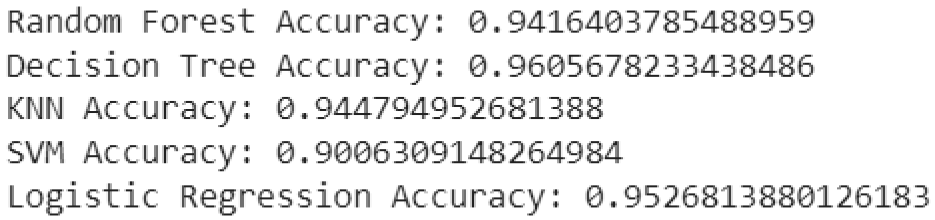
From the figure 6.1, we can see that it has selected 20 best features from the 30 features

7. Model Construction and Comparison

To predict the obesity level, I have implemented those model that from classification to finding the best model, The project has using

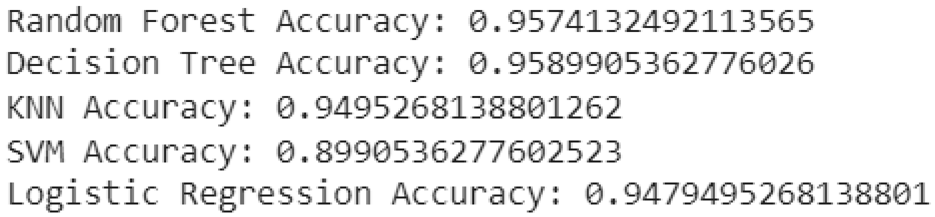
- Random Forest
- Decision Trees
- Support Vector Machines (SVM)
- K Nearest Neighbors
- Logistic Regression

After evaluating the performance for each model, the results are shown below:



```
Random Forest Accuracy: 0.9416403785488959
Decision Tree Accuracy: 0.9605678233438486
KNN Accuracy: 0.944794952681388
SVM Accuracy: 0.9006309148264984
Logistic Regression Accuracy: 0.9526813880126183
```

Figure 7.1 : Result for the model that no using feature selection



```
Random Forest Accuracy: 0.9574132492113565
Decision Tree Accuracy: 0.9589905362776026
KNN Accuracy: 0.9495268138801262
SVM Accuracy: 0.8990536277602523
Logistic Regression Accuracy: 0.9479495268138801
```

Figure 7.2 : Result for the model that using feature selection

From the result Figure 7.1 and Figure 7.2, we can see that Feature selection had a minimal impact on the accuracy of the random forest and decision tree models. The highest accuracy was achieved by the decision tree model without using feature selection. Therefore, it can be concluded that feature selection did not significantly improve the performance of the models. The decision tree model emerged as the best performer in terms of accuracy.

I have generated a confusion matrix for the decision tree model, which provides an overview of the model's performance, as shown below:

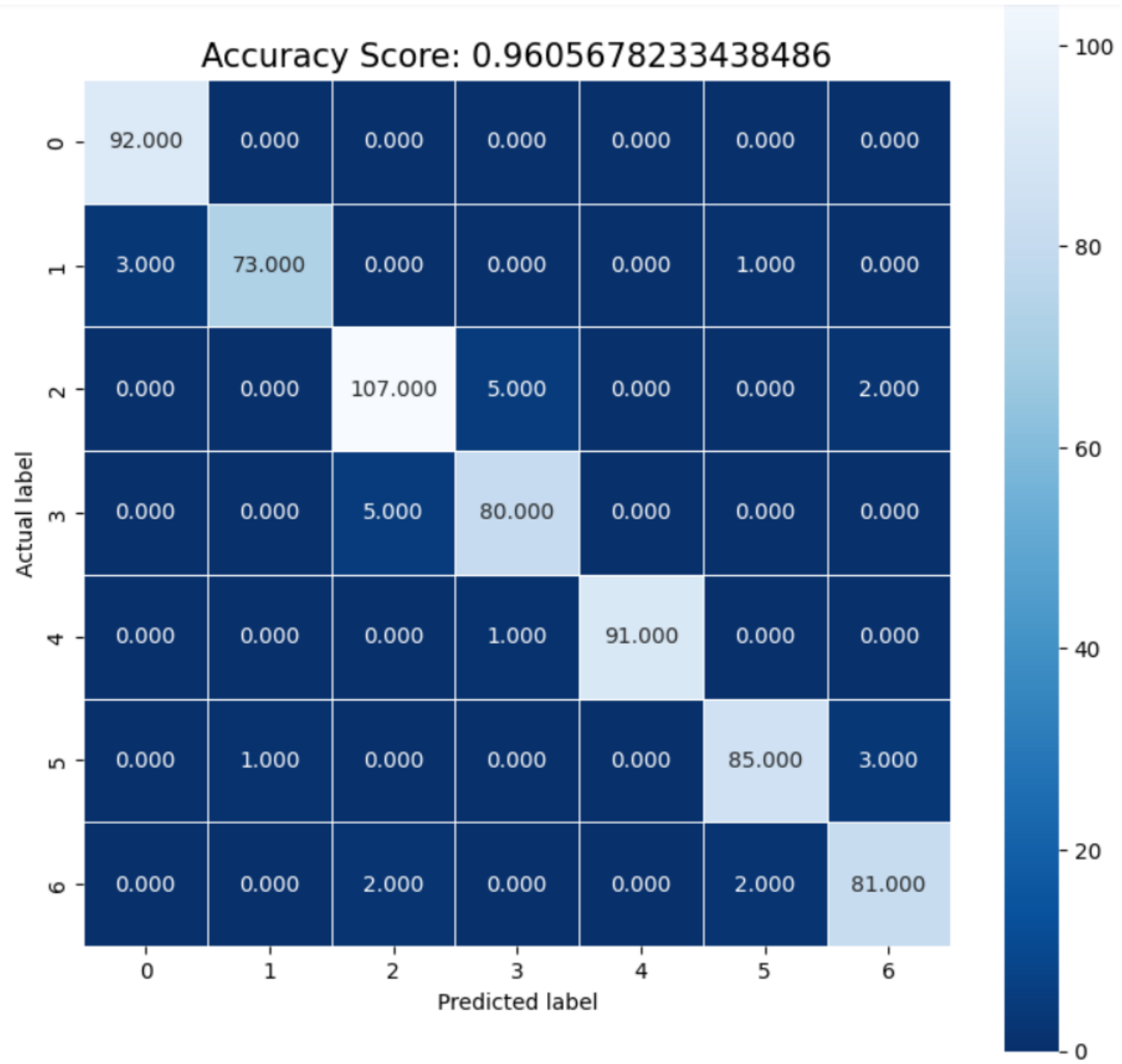


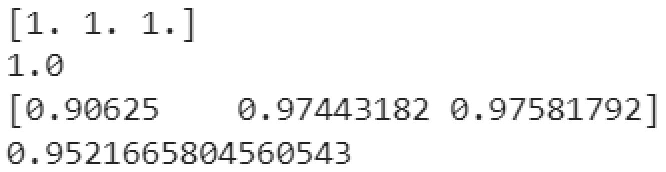
Figure 7.3 : Confusion matrix for decision tree

To further evaluate the decision tree model, I used cross-validation. The model was trained on the numerical features and the target variable, with a cross-validation of 3 folds. The results of the cross-validation are as follows:

Training scores: [train_score_1, train_score_2, train_score_3]

Mean training score: mean_train_score
Test scores: [test_score_1, test_score_2, test_score_3]

Mean test score: mean_test_score



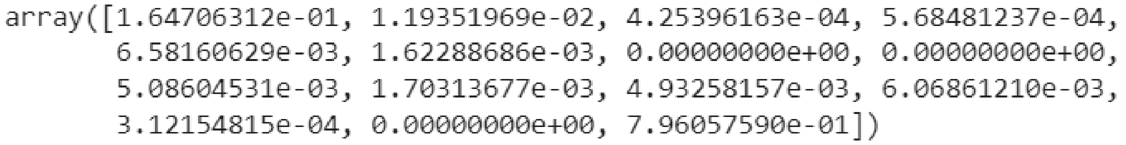
```
[1. 1. 1.]
1.0
[0.90625    0.97443182 0.97581792]
0.9521665804560543
```

Figure 7.4 : Test score for decision tree model

The decision tree model was trained using cross-validation with 3 folds. From Figure 7.4, we can see that the training scores for each fold were [1.0, 1.0, 1.0], resulting in a mean training score of 1.0, indicating a perfect fit to the training data.

For the test scores, the model achieved scores of [0.90625 0.97443182 0.97581792] for each fold, with a mean test score of 0.9521665804560543. These scores indicate that the model performs well on unseen data, with an average accuracy of approximately 95.2%.

Among the features, the decision tree model identified the following variables as having potential impact on the obesity level, ranked by their feature importances:



```
array([1.64706312e-01, 1.19351969e-02, 4.25396163e-04, 5.68481237e-04,
       6.58160629e-03, 1.62288686e-03, 0.00000000e+00, 0.00000000e+00,
       5.08604531e-03, 1.70313677e-03, 4.93258157e-03, 6.06861210e-03,
       3.12154815e-04, 0.00000000e+00, 7.96057590e-01])
```

Figure 7.4 : Feature importance for decision tree model

From figure 7.4, we can see that the feature ranking show as below

BMI (0.796)
Gender (0.165)
Age (0.012)
CH20 (0.006)
FAVC (0.006)
TUE (0.006)
SCC (0.005)
NCP (0.005)
CALC (0.004)
CAEC (0.002)
FAF (0.002)
FCVC (0.001)
MTRANS (0.0)
Family_history_with_overweight (0.0)
SMOK (0.0)

These feature importances indicate the relative importance of each variable in determining the obesity level according to the decision tree model.

8. Conclusion and Future Enhancements

In conclusion, our exploratory analysis has revealed several important insights regarding the factors contributing to obesity. It is evident that eating habits play a significant role in obesity, with individuals who consume high-caloric foods and alcohol showing a higher tendency to become obese. On the other hand, individuals who engage in regular physical activity, such as walking, are more likely to maintain a normal weight. Additionally, having a family history of being overweight emphasises the need for heightened health awareness.

During our exploration of the data, I observed that certain variables exhibit a moderate to high association with BMI. Notably, factors such as FCVC (Frequency of consumption of vegetables), age, and CH2O (Daily water consumption) demonstrated notable importance. Furthermore, I discovered that individuals who actively monitor their calorie intake tend to maintain a healthier weight.

In our feature selection process, I identified the most influential variables for predicting obesity. However, the results indicated that feature selection did not significantly enhance the model's performance in this particular context. Despite this, I proceeded to compare different models and found that the decision tree model proved to be suitable and accurate for estimating obesity levels. To validate its effectiveness, I employed cross-validation, which yielded an impressive accuracy rate of 95%. This underscores the appropriateness of the decision tree model for this specific problem.

Furthermore, I derived a ranking of the variables that have the most potential impact on obesity. The order of importance begins with BMI, followed by gender, age, and CH2O, among others. These insights can guide future research and interventions aimed at addressing and preventing obesity.

To further improve this project, I recommend implementing clustering techniques. By identifying patterns and relationships within the data, clustering can provide more accurate recommendations for targeted interventions and health management strategies.

Overall, this project sheds light on the complex factors associated with obesity and highlights the efficacy of the decision tree model in estimating obesity levels.

9. Deployment

To access the project data file, please follow the steps below:

1. Click on the link provided to download the file: https://archive.ics.uci.edu/dataset/544/estimation+of+obesity+levels+based+on+eating+habits+and+physical+condition

2. Once the download is complete, extract the contents of the ZIP file to your desired location on your computer.

3. Open Google Colab by clicking on the provided link:

https://colab.research.google.com/drive/1G5-zkYH3RxpWF6DDs6rRCgELmbb-KQ0A?usp=sharing

4. In Google Colab, navigate to the file explorer on the left-hand side and click on the "Upload" button.

5. Select the extracted dataset files “ObesityDataSet_raw_and_data_sinthetic.csv” from your computer and upload them to Google Colab, make sure the filename is same as “ObesityDataSet_raw_and_data_sinthetic.csv”

6. Once you have uploaded the dataset to Google Colab, you can use the uploaded files in your code and perform the necessary analysis.