Student ID: 21156028

Assignment 1

Data Exploration and Classification

Semester 1 2024

Student Name: Min Thiha Ko Ko
Student ID: 21156028
Paper Name: Foundations of Data Science

PAPER CODE: COMP615

Due Date: Sunday 14 April 2024 (midnight)

TOTAL MARKS: 100

INSTRUCTIONS:

1. **The following actions** may be deemed to constitute a breach of the General Academic **Regulations Part 7: Academic Discipline,**
   - Communicating with or collaborating with another person regarding the Assignment
   - Copying from any other student work for your Assignment
   - Copying from any third-party websites unless it is an open book Assignment
   - Uses any other unfair means
2. **Please email DCT.EXAM@AUT.AC.NZ if you have any technical issues with your Assessment/Assignment/Test submission on Canvas immediately**
3. **Attach your code for all the datasets in the appendix section**.

Table of Contents

Figures List

Tables List

Estimating Obesity Levels and Identifying Contributing Factors of Obesity via Classification Model

1 Introduction

Statement of Problem

Obesity is the global health concern, which could lead to increased risk of type 2 diabetes, heart disease and various chronic disease. New study in 2022 stated that more than 1 billion people in the world are now living with obesity. Main goal of report is to classifiy the obesity levels of certain population. Accurately estimating obesity levels for the population is crucial for developing effective prevention and cure. It is also important to identify the main factors causing obesity to implement prevention measures for future generation.

How do we estimate the obesity levels?

This problem of estimating obesity levels and identifying main factors will be addressed and investigated in this report through utilization of data mining and machine learning technique.

Specifically, this report will focus on a comprehensive dataset known as “Estimation of Obesity Levels” from UCI Machine Learning Repository. This dataset include data for the estimation of obesity levels in individuals from the countries of Mexico, Peru and Colombia, based on their eating habits and physical condition.

The aim of this work is to develop a machine learning model that can accurately classifiy and predict an individual’s obesity level based on their dietary habits and physical activities. By implementing a reliable machine learning model, any dataset can be utilized to analyse the relationship between obesity level and its contributing factors, and effectively predicting the root cause of obesity.

The research questions that need to be answer are:

- What are the key contributing features (predictiors) of obesity levels among individuals in Mexico, Peru and Columbia?
- Is it sufficient to develop an accurate model using the provided features in this dataset?

To address therse questions, this study will employ various machine learning techniques, with a particular focus on decision trees. Decision trees are chosen for their ability to provide interpretable results, handle both numerical and categorical data and identify important features in the prediction process.

Assumptions:

- The dataset provides accurate and reliable information about the individuals' attributes.
- The features included in the dataset are relevant for predicting obesity levels.
- The dataset represents data from diverse sample from Mexico, Peru and Columbia to ensure the effectiveness of classification model.
- The decision tree model can effectively capture the complex relationships between the various features and obesity levels.

2 Data Exploration

The dataset consists of several attributes that are relevant to the analysis and modelling task. In this section, these features will be explored and discussed to determine which are the most relevant for the classification model. The aim of this exploration is to learn about the characteristics of dataset, number of features and to check the cleanliness of the data. This process will help achieve a meaningful dataset which will improve the machine learning model in next stages.

The **Estimation of Obesity Levels** dataset consists of 17 attributes and 2,111 instances (rows). 16 of the attributes are independent features and the target to predict is the obesity level. Features include both numerical and categorical data types (Table 1).

Numerical Features

- Age (Years)
- Height (Meters)
- Weight (Kilograms)
- FCVC (Frequency of consumption of vegetables)
- NCP (Number of main meal)
- CH20 (Consumption of water daily)
- FAF (Physical activity frequency)
- TUE (Time using technology devices)

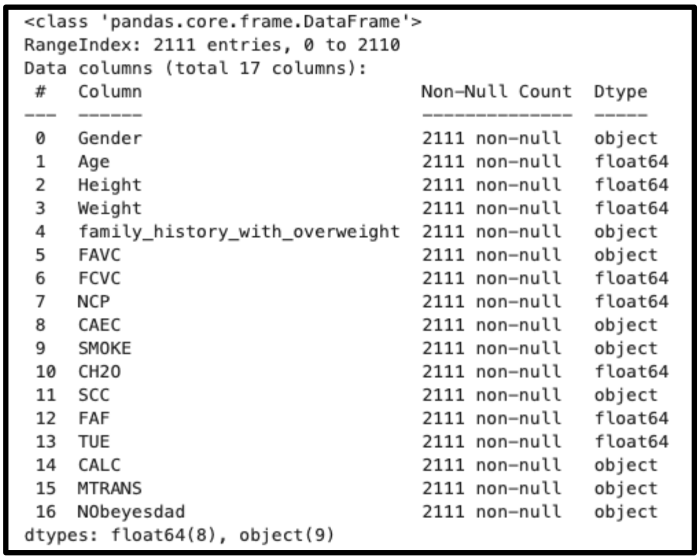
All the numerical data are continuous, stored as float datatype in the dataset.

Categorical Features

- Gender (Male/Female)
- family_history_with_overweight (Binary)
- FAVC (Frequent consumption of high caloric food) (Binary)
- CAEC (Consumption of food between meals)
- SMOKE (Smoking) (Binary)
- SCC (Calories consumption monitoring) (Binary)
- CALC (Consumption of alcohol)
- MTRANS (Transportation used)

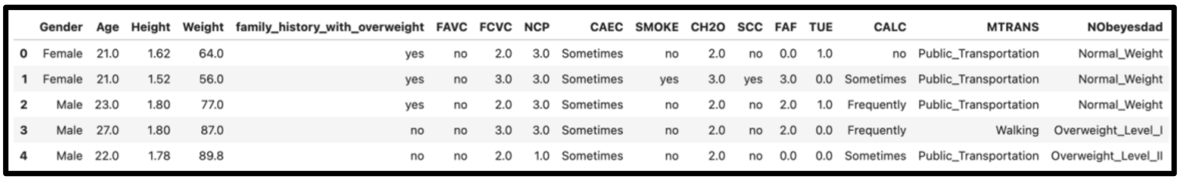
Target Variable (Class) – NObeyesdad (Obesity Levels)

Most of the categorical data are binary, meaning they have two distinct categories (example: Yes/No) whiles some categorical features such as CAEC have multiple categories.



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2111 entries, 0 to 2110
Data columns (total 17 columns):
 #   Column                          Non-Null Count  Dtype
---  ------                          --------------  -----
 0   Gender                          2111 non-null   object
 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   object
 5   FAVC                            2111 non-null   object
 6   FCVC                            2111 non-null   float64
 7   NCP                             2111 non-null   float64
 8   CAEC                            2111 non-null   object
 9   SMOKE                           2111 non-null   object
 10  CH2O                            2111 non-null   float64
 11  SCC                             2111 non-null   object
 12  FAF                             2111 non-null   float64
 13  TUE                             2111 non-null   float64
 14  CALC                            2111 non-null   object
 15  MTRANS                          2111 non-null   object
 16  NObeyesdad                      2111 non-null   object
dtypes: float64(8), object(9)
```

Table 1 - Dataset Infromation

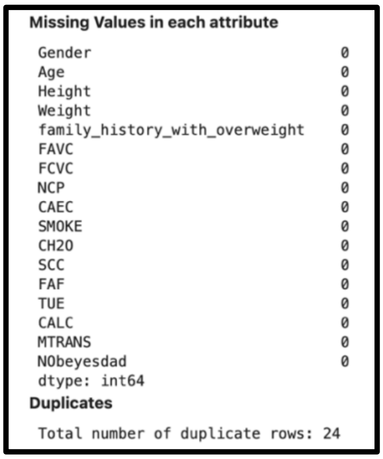


	Gender	Age	Height	Weight	family_history_with_overweight	FAVC	FCVC	NCP	CAEC	SMOKE	CH2O	SCC	FAF	TUE	CALC	MTRANS	NObeyesdad
0	Female	21.0	1.62	64.0	yes	no	2.0	3.0	Sometimes	no	2.0	no	0.0	1.0	no	Public_Transportation	Normal_Weight
1	Female	21.0	1.52	56.0	yes	no	3.0	3.0	Sometimes	yes	3.0	yes	3.0	0.0	Sometimes	Public_Transportation	Normal_Weight
2	Male	23.0	1.80	77.0	yes	no	2.0	3.0	Sometimes	no	2.0	no	2.0	1.0	Frequently	Public_Transportation	Normal_Weight
3	Male	27.0	1.80	87.0	no	no	3.0	3.0	Sometimes	no	2.0	no	2.0	0.0	Frequently	Walking	Overweight_Level_I
4	Male	22.0	1.78	89.8	no	no	2.0	1.0	Sometimes	no	2.0	no	0.0	0.0	Sometimes	Public_Transportation	Overweight_Level_II

Table 2 - Head of Dataset

Data Cleaning and Visualization

- Dataset has no missing values in any of the columns.
- However, it has 24 duplicates rows. In this type of datasets, the presence of duplicate instances could indicate that same individual was measured multiple times, which would introduce bias and affect the model's performance in next step. Therefore, duplicates rows were dropped.



Missing Values in each attribute

```
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
```

Duplicates

```
Total number of duplicate rows: 24
```

Table 3 - Missing Value and Duplicates Counts

- Figure 1 and Figure 2 show the distribution of numerical data to check if there is any outliers in each feature.
- Boxplot of age describes that there might be potential outliers lying outside the range. However, this dataset is recorded based on various age groups to get obesity levels across different demographics. Thus, these outliers in age may reflect the natural diversity of the population rather than data error. Removing them may affects the outcome of the model.
- Similarly, NCP feature seems to have potential outliers. However, NCP (Number of Main Meals) could vary among individuals based on dietary habits, cultural practices and lifestyle choices. While most of the population in the dataset have 3 meals a day, other might have less or more than 3. Therefore, keeping the diverse range of data is essential for ensuring the model's robustness.

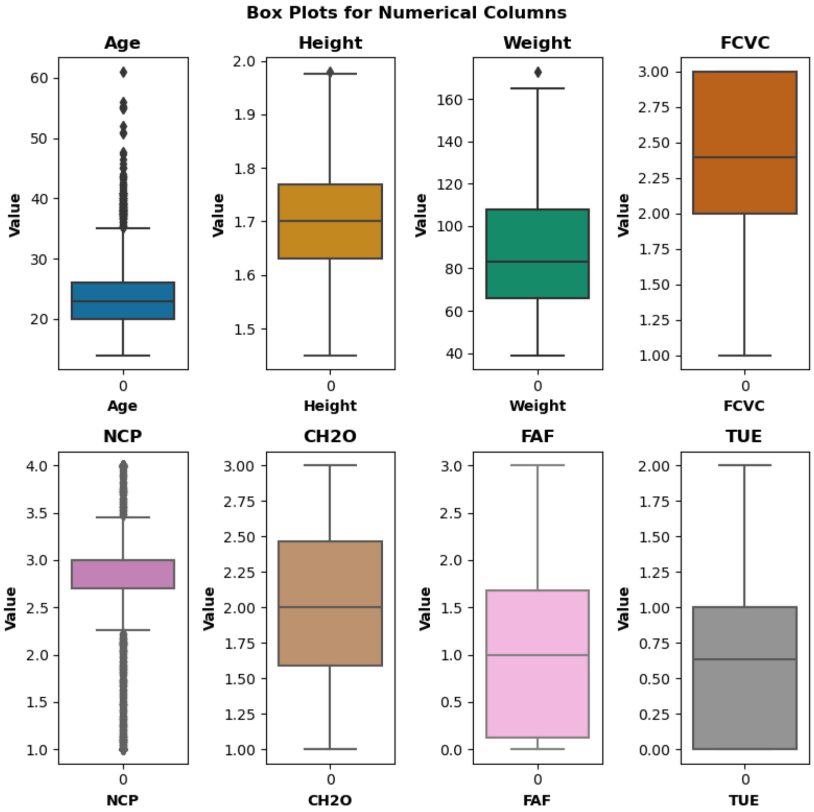


Figure 1 - Box Plots of Numerical Data

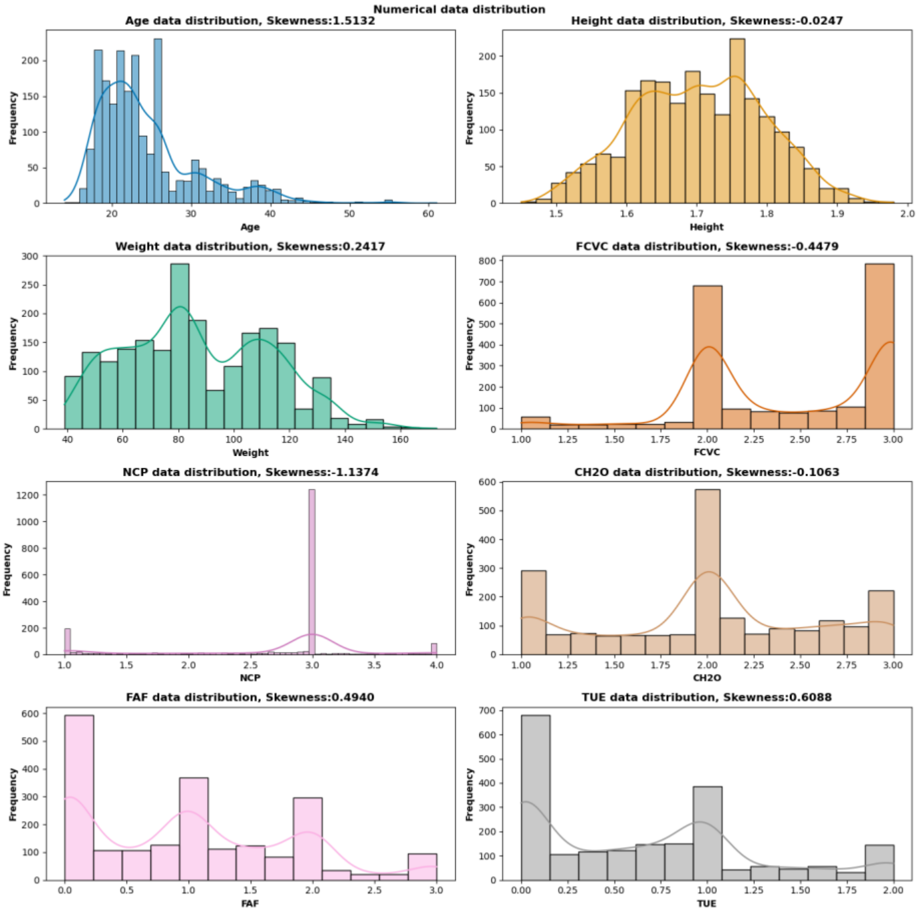


Figure 2 - Distribution of Numerical Data

Upon examination of distribution and skewness of each numerical data (Figure 2), it is evident that age data exhibit a positive skewness. This indicates that there is a large proportion of young indiviuduals in the sample.

Relationship Between Some Notable Features

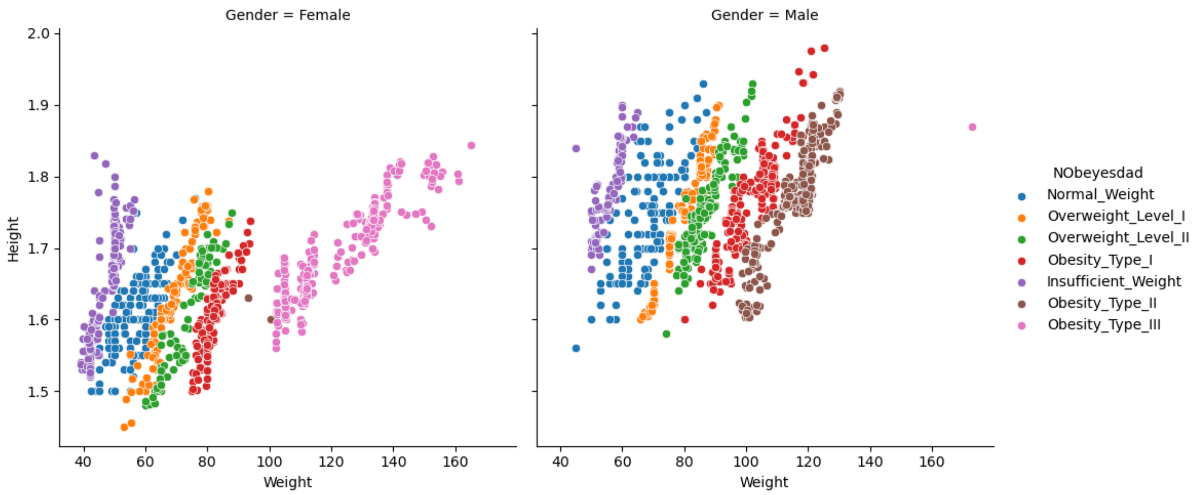
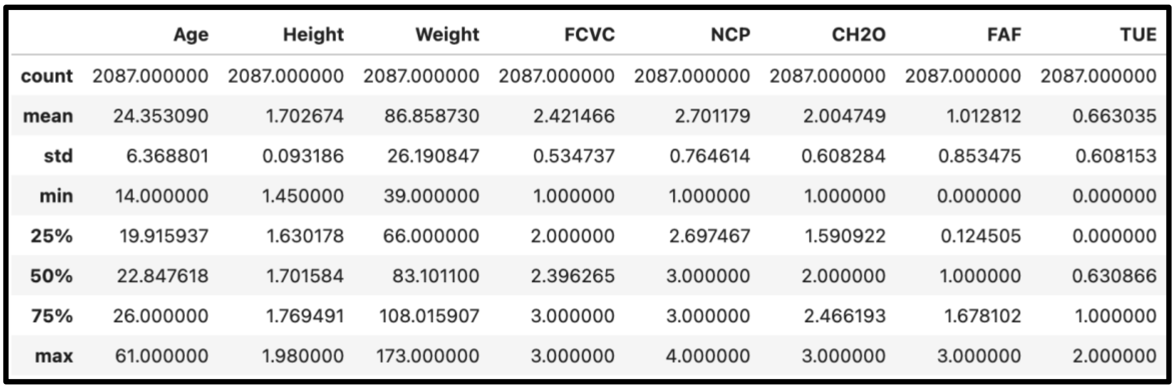


Figure 3 - Relationship Between Weight and Height by Gender, Colored by Obesity Level

The above groups scatterplot indicates the relationship between weight and height, categorised by Male and Female. It is evident that there is a positive correlation between height and weight, meaning that as height increases, weight tends to increase as well. This relationship is true for both male and female. However, men generally have higher heights compared to women according to the plot. It is also notable that there is a significant amount of individuals with Obesity Type 3 among female compared to male.

Summary Statistics of Numerical Features

The following table (Table 4) shows the summary statistics of numerical data. Afer dropping duplicates rows, all the columns haave 2087 instances. Average age of the sample in the dataset is 24.35 years. Average height is around 1.7 m and average weight is 86.85 kg. Using this summary statistic, we can see the general information about this dataset.



	Age	Height	Weight	FCVC	NCP	CH2O	FAF	TUE
count	2087.000000	2087.000000	2087.000000	2087.000000	2087.000000	2087.000000	2087.000000	2087.000000
mean	24.353090	1.702674	86.858730	2.421466	2.701179	2.004749	1.012812	0.663035
std	6.368801	0.093186	26.190847	0.534737	0.764614	0.608284	0.853475	0.608153
min	14.000000	1.450000	39.000000	1.000000	1.000000	1.000000	0.000000	0.000000
25%	19.915937	1.630178	66.000000	2.000000	2.697467	1.590922	0.124505	0.000000
50%	22.847618	1.701584	83.101100	2.396265	3.000000	2.000000	1.000000	0.630866
75%	26.000000	1.769491	108.015907	3.000000	3.000000	2.466193	1.678102	1.000000
max	61.000000	1.980000	173.000000	3.000000	4.000000	3.000000	3.000000	2.000000

Table 4 - Summary Statistics of Numerical Data

Samples in Each Category of Target Variable

The target variable in the dataset is **NObeyesdad**, which has 7 categories. The following Figure 4 shows distribution of samples across each category. It can be seen that data spread relatively uniform across different categories, suggesting that there is no class imbalance problem. This means that dataset is well-suited for classification model. Obesity_Type_1 category has the highest instances with 351 while only 267 individuals were classified in insufficient weight category.

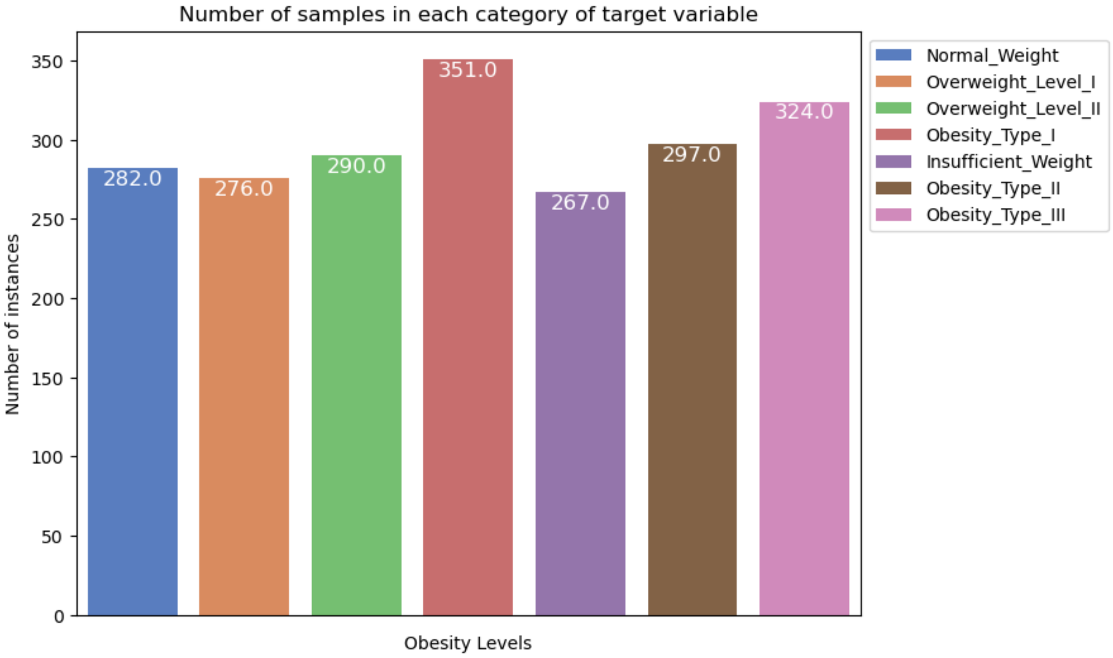


Figure 4 – Counts of Samples in Each Target Category

Distribution of Categorical Data Across Obesity Levels

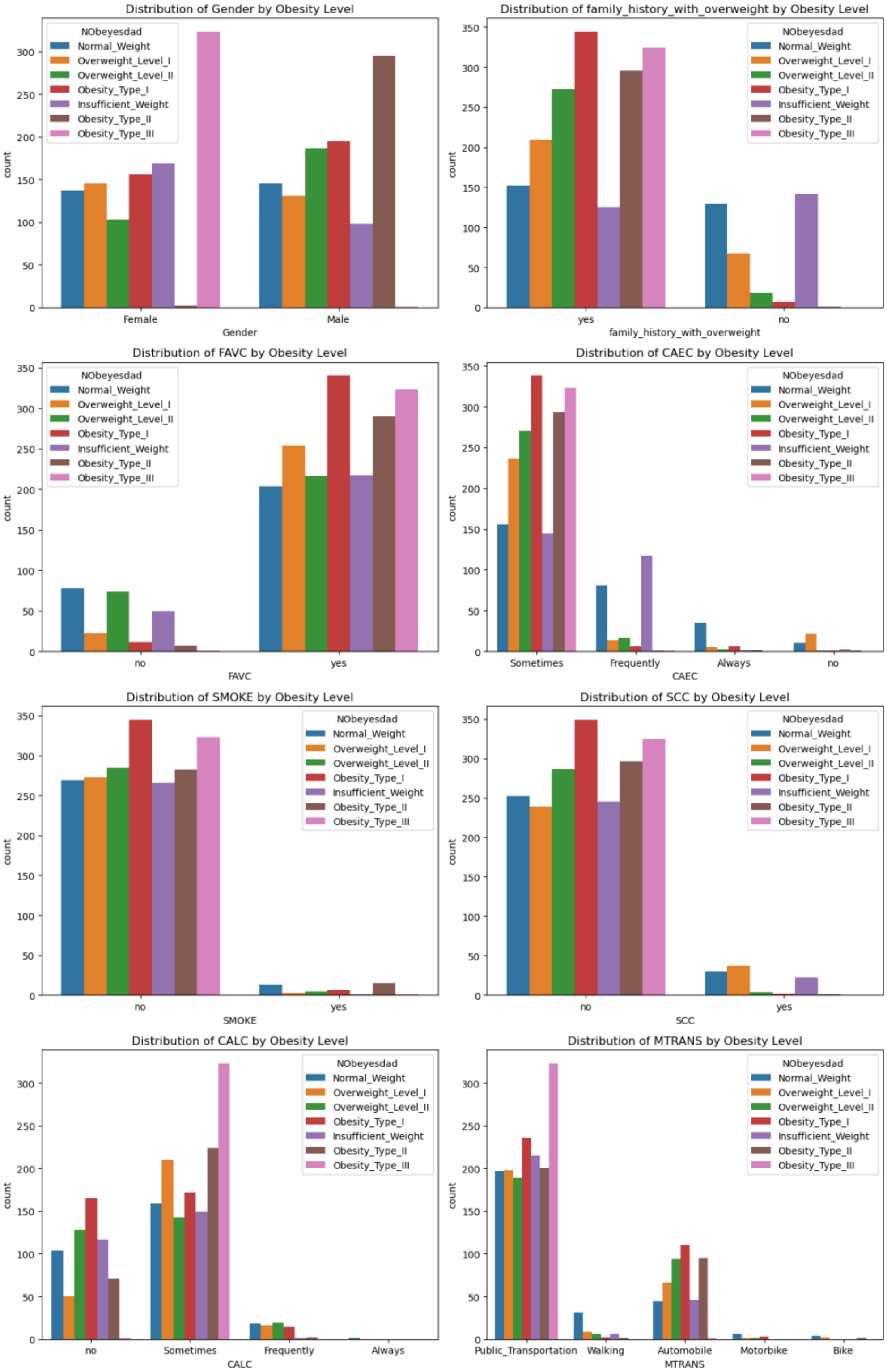


Figure 5 - Distribution of Categorical Data

The above figure show the distribution of various categorical features across different obesity levels in the dataset. Each plot represents the relationship between each feature and the obesity level. One notable observation in the plot is that SMOKE and SCC (Calories consumption monitoring) features has most of the data in NO section, meaning that the individuals in the dataset do not smoke or usually monitor their daily calories intake.

Another interesting point is that the relationship between the FAVC (Frequency of consuming high caloric food) and obesity level. We observe that individuals that classified as Obesity Type 1, Obesity Type 2 and Obesity Type 3, all tend to have answered "Yes" in consuming high-caloric food. This explained that people who often eats high caloric food might have a higher likelihood of being classified into higher obesity levels.

3 Classification Models

In this section, the dataset will be used to create classication models. Based on evalution results of each model, we will determine the most sutiable model for classification of obesity levels.

Report on Data Preprocessing

Before creating a classification model, it is important to decide how to deal with the missing values, duplicates and outliers. As seen in section 2, there is no missing values in this dataset, however there are some duplicates. All duplicated rows were removed. For outliers, there are some notable outliers values in Age and NCP. As discussed in section 2, they are not necessarily outliers. Those extreme values are the representation of diversity within dataset. Thus, all those extreme values (outliers) were kept.

Before creating a model, the dataset is typically divided into training and testing sets to evaluate model performance. Train set is used to train the classification model and test set is used to evaluate how the classification model performs. Train/Test split is usally done by random splitting around 70% of data into train set and the remaining 30% into test set. Since, it is a random process, it is important to note that if the train, test data is changed by running the code again, the outcome of the model will be changed.

Base Model

The base tree model is constructed by using Decision Tree Classifier. A baseline model refers to one where no parameters are explicitly controlled or modified, and only default values are used. By using the default setting and any parameters are not defined, the model will continue to build until all the leaf nodes are pure, meaning they contain only samples of a single class.

The criterion used in this model is Gini Impurity. A lower Gini means the node is more pure, so that all leaves have gini value of zero, meaning that the node is purest without impurity.

Samples indicate number of samples that reach to this node during training.

Since, this dataset has 16 features and 2087 instances, the baseline decision tree is relatively large with 215 nodes. It is not ideal to interpret the tree with large numbers of nodes. The common way to interpret and evaluate the model is using confusion matrix and model accuracy.

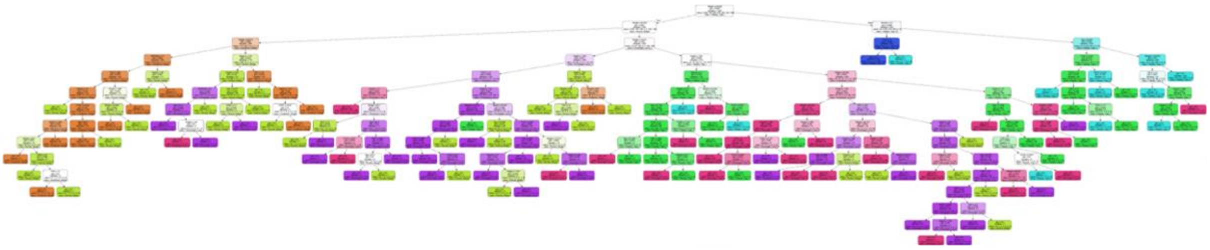


Figure 6 - Base Decision Tree Model

Confusion Matrix and Classification Report of Base Model

This following confusion matrix represents the performance of a baseline classification model across different clsses. The rows represent the actual categories from the dataset.The columns represent the predicted classes or categories assigned by the model. The diagonal elements (from top-left to bottom-right) represent the number of instances where the predicted class matches the actual class. These values indicate correct predictions.

Other cells that are not diagonals represent instances where the predicted class does not match the actual class. These values indicate misclassifications.

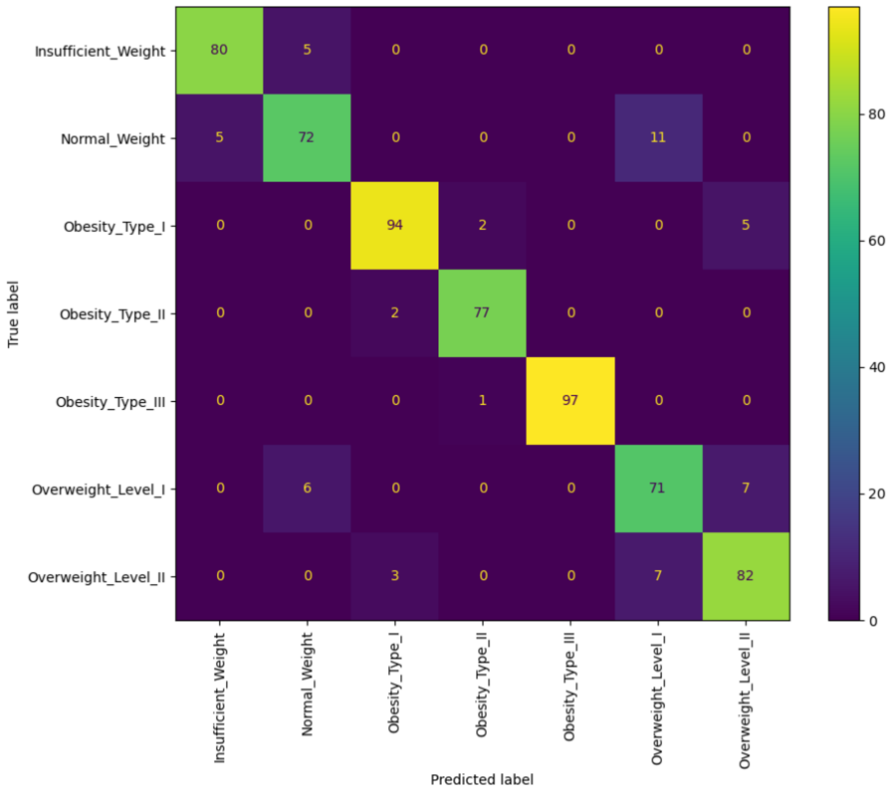


Figure 7 - Confusion Matrix of Base Model

The above confusion matrix can be interpreted as:

1. Row 1 (Insufficient Weigtht)

   80 instances were correctly classified as insufficient weight. 5 were wrongly classified as normal weight.

2. Row 2 (Normal Weight)

   72 instances were correctly classified as normal weight. 5 were wrongly classified as insufficient weight while 11 were put in overweight level 1.

3. Row 3 (Obesity Type 1)

   94 were correctly classified as obesity type 1. 2 were misclassified as obesity type 2 while 5 were wrongly put in overweight level 3.

4. Row 4 (Obesity Type 2)

   77 were correctly classified while 2 were misclassified.

5. Row 5 (Obesity Type 3)

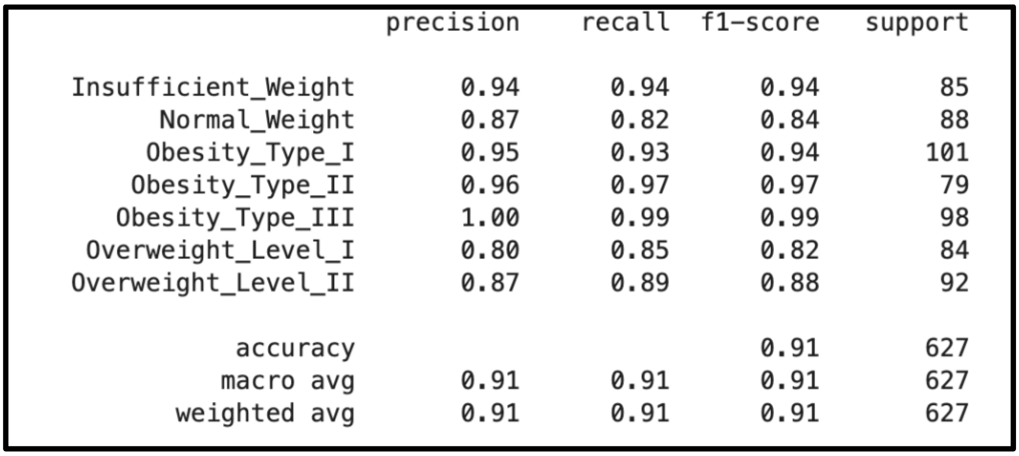
   Only 1 were misclassified while 97 were in correct category.

6. Row 6 (Overweight Level 1)

   71 were correctly classified as Overweight Level 1 while 6 were put in normal weight and 7 were misclassified as overweight level 2.

7. Row 7 (Overweight Level 2)

   82 were in the right category while 3 were misclassified as overweight type 1 and 7 were wrongly put in overweight levl 1.



```
                     precision    recall  f1-score   support

 Insufficient_Weight       0.94      0.94      0.94        85
       Normal_Weight       0.87      0.82      0.84        88
      Obesity_Type_I       0.95      0.93      0.94       101
     Obesity_Type_II       0.96      0.97      0.97        79
    Obesity_Type_III       1.00      0.99      0.99        98
  Overweight_Level_I       0.80      0.85      0.82        84
 Overweight_Level_II       0.87      0.89      0.88        92

            accuracy                           0.91       627
           macro avg       0.91      0.91      0.91       627
        weighted avg       0.91      0.91      0.91       627
```

Table 5 - Classification Report of Base Model

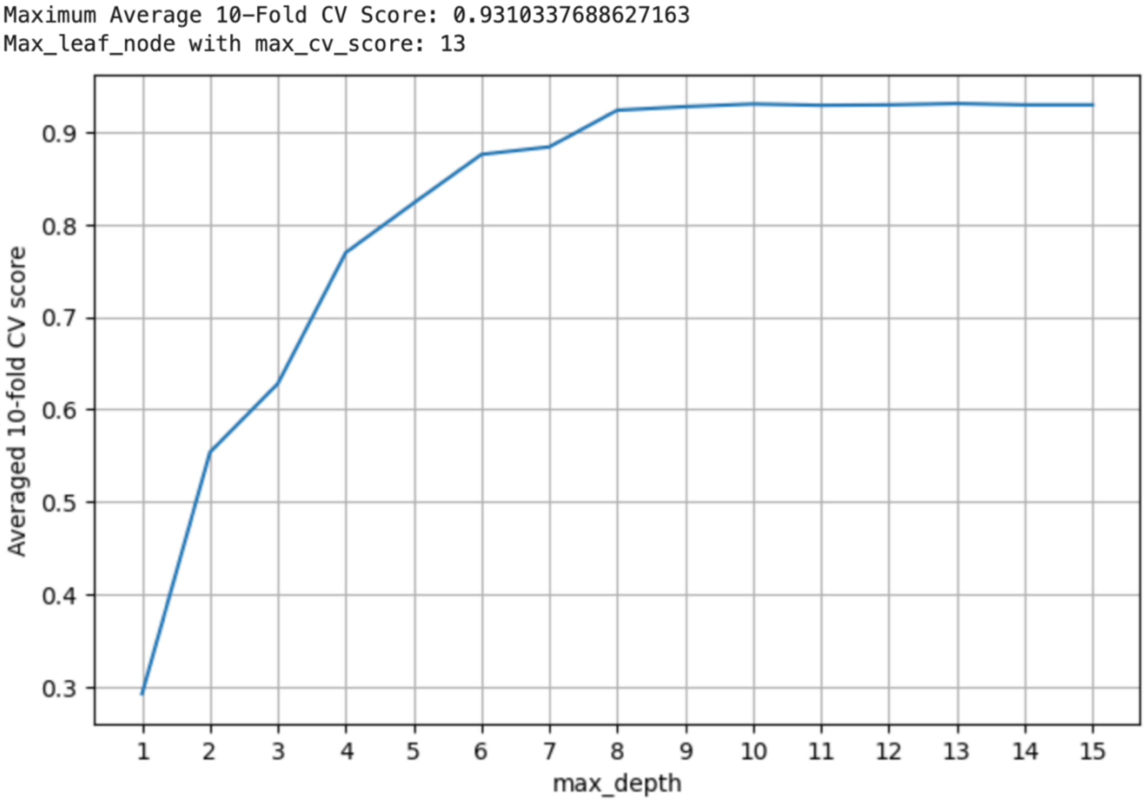
This is the classification report for the base model. The accurary of the model is 91 percent. Accuracy is an overall measure of correct prediction, regardless of the class (positive or negative). High recall implies that very few positives are misclassified as negatives. High precision implies very few negatives are misclassified as positives. Most of the classes have high precision and recall score. Only overweight level 1 has a bit low precision compared to other classes meaning that few of the instances from other classes are missclassifed as overweight level 1. Normal weight has 0.82 recall value which is a bit lower than other classes, meaning that a few of the normal weight instances are misclassified as other category.

Tuning Maximum Depth Parameter

The maximum depth parameter in the Decision Tree Classifier controls the maximum depth of the tree, which directly impacts the model's complexity and ability to capture different patterns in the data. A deeper tree can potentially learn more complex relationships between features but also risks overfitting, where the model memorizes noise in the training data rather than learning meaningful patterns.

The maximum depth will be tuned by using Average 10-fold cross validating scores. By running through a certain range, we find the max average cross validating scores and pick the max depth value which give that max score.

The following graph is the plot between averaged 10-fold cv score and max_depth. It can be seen that the maximum depth 13 has the higest scores. Therefore, it is chosen as a parameter.



```
Maximum Average 10-Fold CV Score: 0.9310337688627163
Max_leaf_node with max_cv_score: 13
```

Figure 8 - Max_Depth vs Average 10 Fold CV Score

Maximum Depth Tuned Model

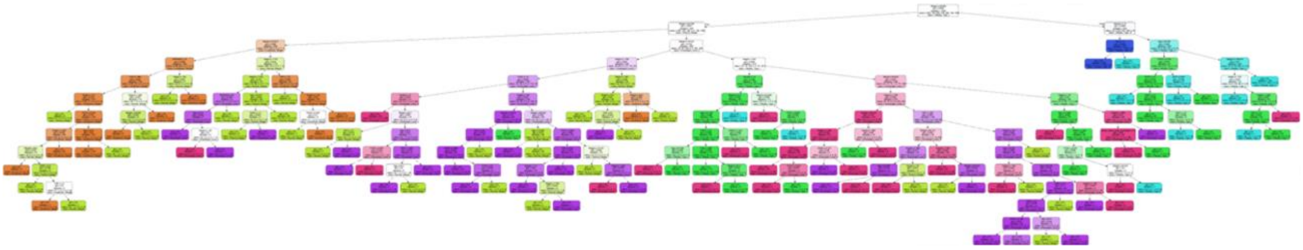
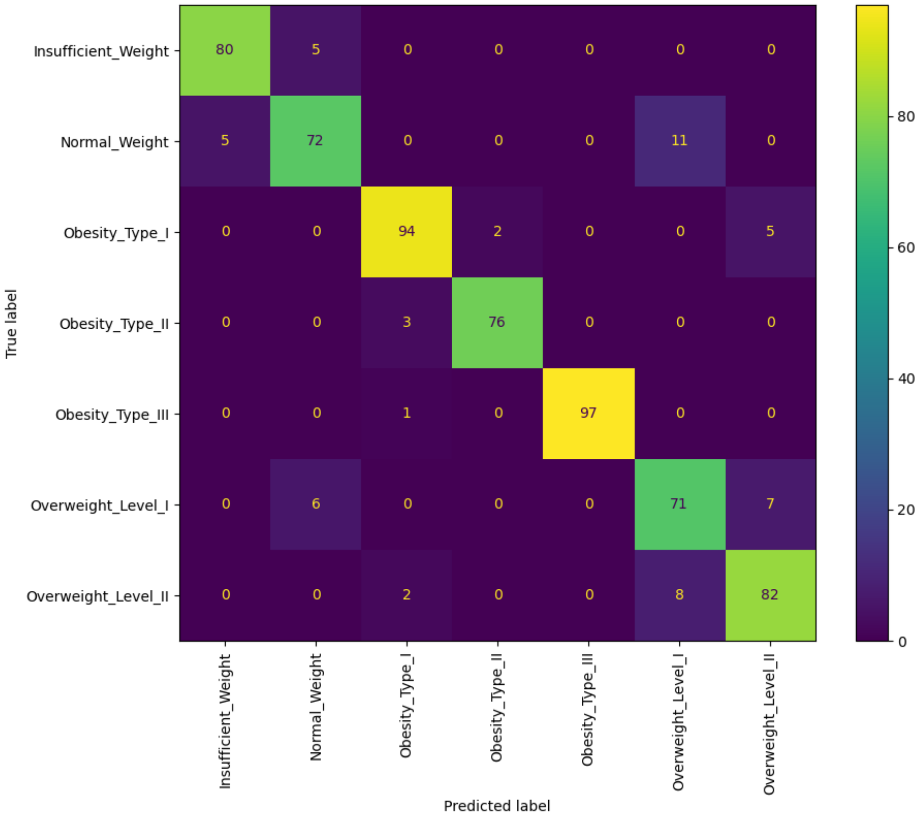


Figure 9 - Max_Depth Tuned Tree Model

After tuning the maximum depth, the tree model is built using the tuned value (13). The new model has 213 nodes which is only 2 nodes less than the base model. This model can be evaluated by using confusion matrix and classification report.

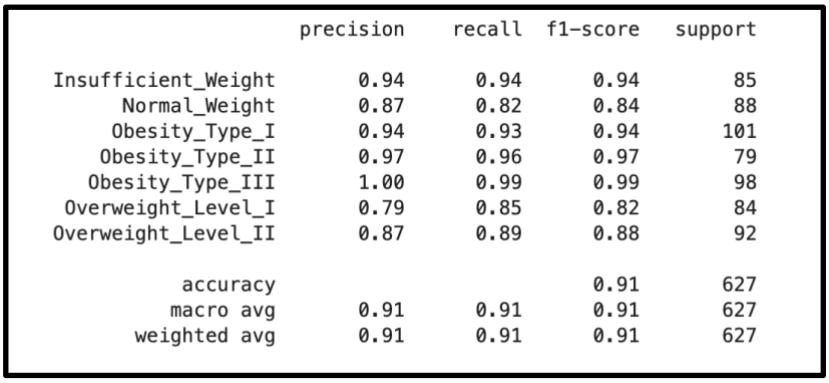
Confusion Matrix and Classification Report of Depth Tuned Model



True label \ Predicted label	Insufficient_Weight	Normal_Weight	Obesity_Type_I	Obesity_Type_II	Obesity_Type_III	Overweight_Level_I	Overweight_Level_II
Insufficient_Weight	80	5	0	0	0	0	0
Normal_Weight	5	72	0	0	0	11	0
Obesity_Type_I	0	0	94	2	0	0	5
Obesity_Type_II	0	0	3	76	0	0	0
Obesity_Type_III	0	0	1	0	97	0	0
Overweight_Level_I	0	6	0	0	0	71	7
Overweight_Level_II	0	0	2	0	0	8	82

Figure 10 - Confusion Matrix for Depth Tuned Model

The confusion matrix of tuned model is the mostly the same as basemodel since only a few nodes were reduced. Only a minimal changes were occurred in this matrix.

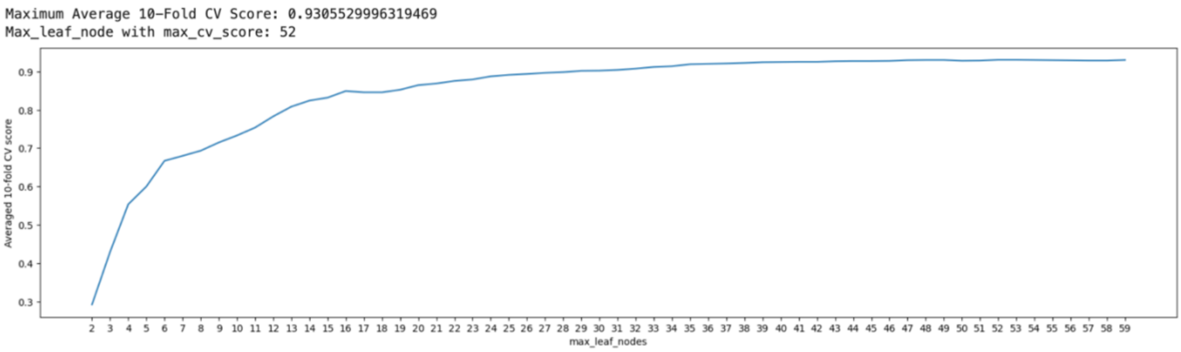


```
                     precision    recall  f1-score   support

Insufficient_Weight       0.94      0.94      0.94        85
      Normal_Weight       0.87      0.82      0.84        88
     Obesity_Type_I       0.94      0.93      0.94       101
    Obesity_Type_II       0.97      0.96      0.97        79
   Obesity_Type_III       1.00      0.99      0.99        98
 Overweight_Level_I       0.79      0.85      0.82        84
Overweight_Level_II       0.87      0.89      0.88        92

           accuracy                           0.91       627
          macro avg       0.91      0.91      0.91       627
       weighted avg       0.91      0.91      0.91       627
```

Table 6 - Classification Report for Depth Tuned Model

The result of the precision matrix is similar to that of the base model. One notal change is precision score of overweight level 1 seems to reduce for a small amount.

Tuning Maximum Leaf Node Parameter



```
Maximum Average 10-Fold CV Score: 0.9305529996319469
Max_leaf_node with max_cv_score: 52
```

Figure 11 - Max_Leaf_Nodes vs Average 10 Fold CV Score

The maximum leaf node parameter is also tuned by using the same method as maximum depth. Average 10 fold cross validation score were conducted using different max_leaf_node values while max depth parameter is set according to the previous tuned value (13). Once the highest score is obtained, we choose the maximum leaf node parameter that corresponds to that highest score. This max_leaf_nodes parameter is also important for the model since it controls the size and complexity of the tree. The tuned value we obtained is 52, which is relatively large but optimal for balancing model complexity and performance.

Final Model

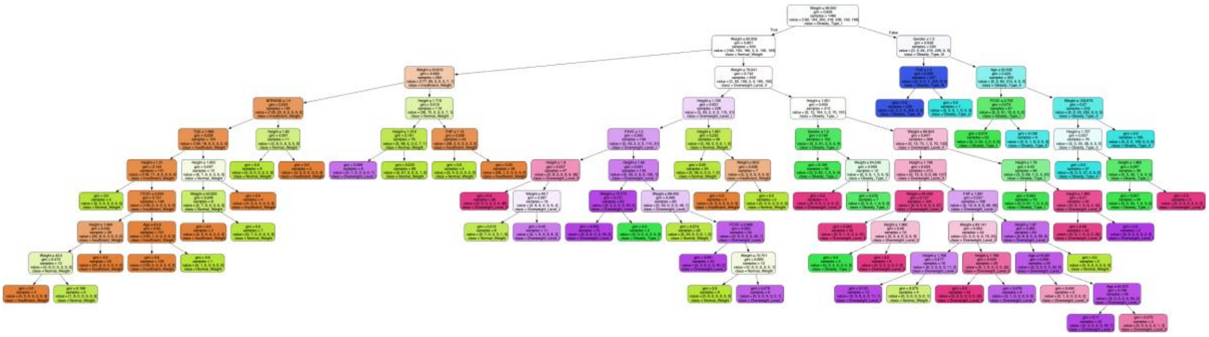
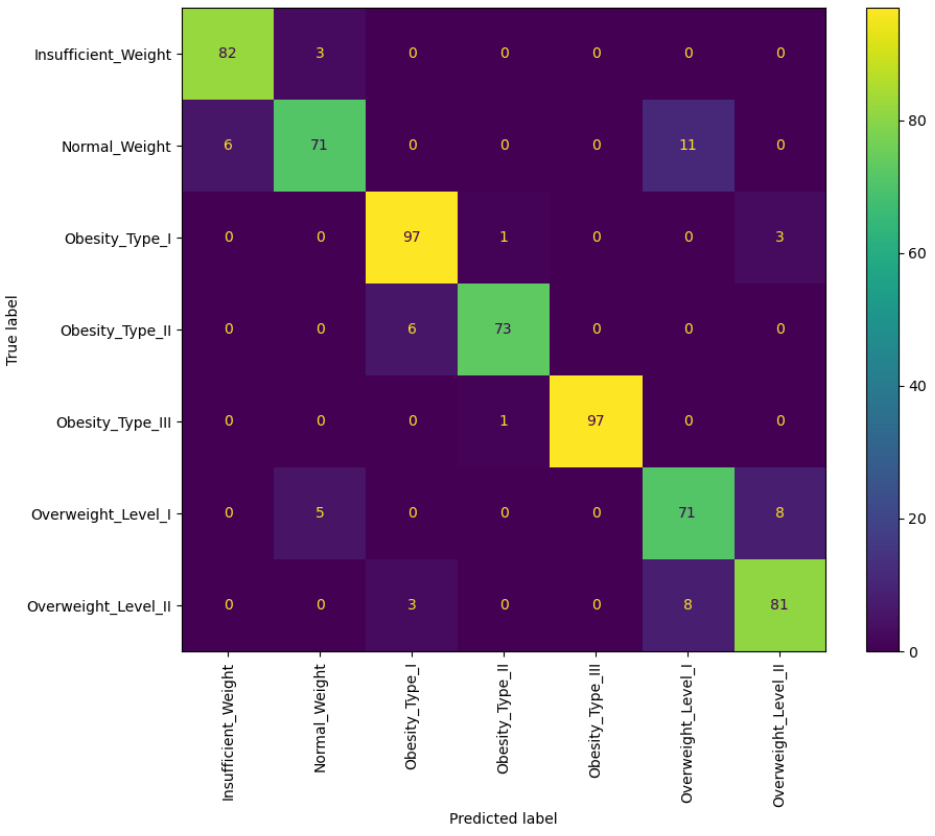


Figure 12 - Final Decision Tree Model

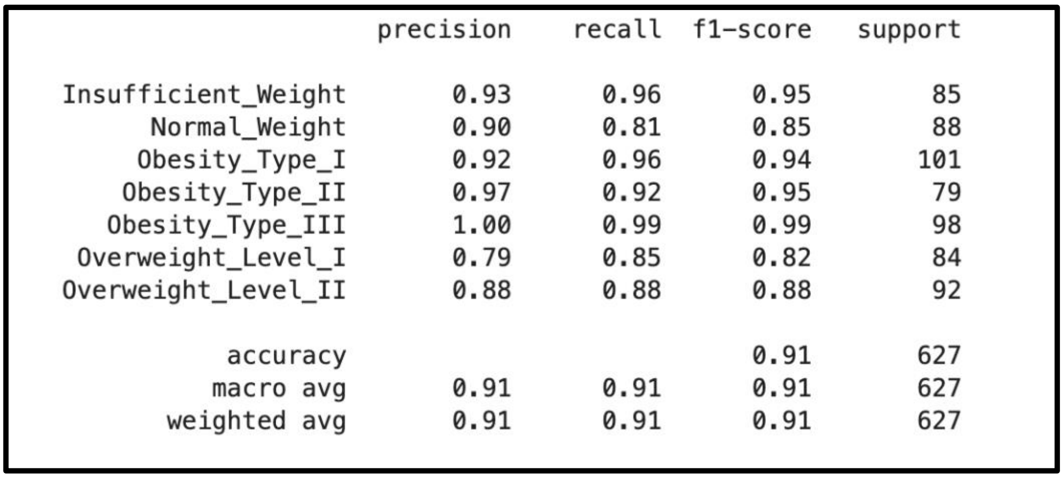
The final model is built by tuning max_depth 13 and max_leaf_nodes 52. The model has 103 nodes which is more than half of the previous model, but it is relative large since the dataset has 16 predictors.

Confusion Matrix and Classification Report for Final Model



	Insufficient_Weight	Normal_Weight	Obesity_Type_I	Obesity_Type_II	Obesity_Type_III	Overweight_Level_I	Overweight_Level_II
Insufficient_Weight	82	3	0	0	0	0	0
Normal_Weight	6	71	0	0	0	11	0
Obesity_Type_I	0	0	97	1	0	0	3
Obesity_Type_II	0	0	6	73	0	0	0
Obesity_Type_III	0	0	0	1	97	0	0
Overweight_Level_I	0	5	0	0	0	71	8
Overweight_Level_II	0	0	3	0	0	8	81

Figure 13 - Confusion Matrix for Final Model



```
                     precision    recall  f1-score   support

Insufficient_Weight       0.93      0.96      0.95        85
      Normal_Weight       0.90      0.81      0.85        88
     Obesity_Type_I       0.92      0.96      0.94       101
    Obesity_Type_II       0.97      0.92      0.95        79
   Obesity_Type_III       1.00      0.99      0.99        98
 Overweight_Level_I       0.79      0.85      0.82        84
Overweight_Level_II       0.88      0.88      0.88        92

           accuracy                           0.91       627
          macro avg       0.91      0.91      0.91       627
       weighted avg       0.91      0.91      0.91       627
```

Table 7 - Classification Report for Final Model

The final model's confusion matrix and classification report do not vary much from the previous models. The accuracy of the model remain the same with 0.91 %, which is relatively accuate for a classification mode.

Role of Max Depth and Max Leaf Node Parameters

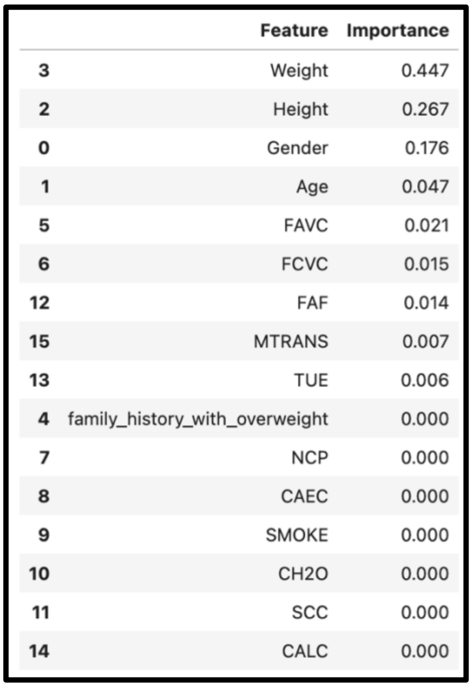
The max_depth parameter controls the maximum depth of the decision tree, limiting the number of splits and preventing the tree from growing too deep. A deeper tree can lead to overfitting, capturing noise and irrelevant patterns in the training data.

The max_leaf_nodes parameter limits the maximum number of leaf nodes in the tree, effectively controlling the tree's overall size and complexity. A larger number of leaf nodes can potentially capture more intricate patterns but may also lead to overfitting.

While the optimal values obtained for this dataset (max_depth=13 and max_leaf_nodes=52) improved the model's size and did not affect much of the model's accuracy, these values are specific to the characteristics and complexity of this particular dataset. Using the same values for other datasets may not necessarily improve the accuracy, as each dataset has its own unique patterns, features, noises and complexities.

The appropriate values for these parameters should be determined through cross-validation or a separate validation set for each new dataset, as they can vary significantly based on the data's characteristics and the problem being addressed.

Feature Importance



	Feature	Importance
3	Weight	0.447
2	Height	0.267
0	Gender	0.176
1	Age	0.047
5	FAVC	0.021
6	FCVC	0.015
12	FAF	0.014
15	MTRANS	0.007
13	TUE	0.006
4	family_history_with_overweight	0.000
7	NCP	0.000
8	CAEC	0.000
9	SMOKE	0.000
10	CH2O	0.000
11	SCC	0.000
14	CALC	0.000

Table 8 - Feature Importance

Based on the provided feature importance scores, we can interpret the importance of each feature in the decision tree model as follows:

- Weight (Importance: 0.447):
  Weight is the most important feature for predicting obesity levels according to the final decision tree model. This means that an individual's weight plays a crucial role in determining their obesity level.
- Height (Importance: 0.267):
  Height is the second most important feature in the model since height and weight are key factors in calculating body mass index (BMI), which is widely used to access obesity levels.
- Gender (Importance: 0.176):
  Gender is the third most important feature, suggesting that there is the differences in obesity levels between males and females in the dataset.
- Age (Importance: 0.047):
  Age has a relatively low importance score, indicating that according to this model and dataset, age may not be a strong predictor of obesity levels.
- FAVC (Frequent consumption of high caloric food) (Importance: 0.021):
  The feature FAVC, which represents the frequent consumption of high-caloric food, has a low importance score but still has a few impact on classification. This suggests that dietary habits may play a role in predicting obesity levels, although its importance is relatively low compared to factors like weight and height.
- FCVC (Frequency of consumption of vegetables) (Importance: 0.015): Similar to FAVC, FCVC, which represents the frequency of consumption of vegatables, has a low importance score, indicating that it may not be a strong predictor but still has small impact on the obesity level.
- FAF (Physical activity frequency) (Importance: 0.014): The feature FAF, which represents the frequency of physical activity, has similar importance scores as FCVC. This means it has some influence but minimal to the obesity level.
- Other features have importance scores of near zero or zero, indicating that they did not contribute to predictions of obesity levels.

4 Conclusion

The results of three models are as followed:

Model 1: Decision Tree with default parameters
Accuracy: 0.91
Precision: 0.91(macro average)
Recall: 0.91(macro average)
F1-score: 0.91(macro average)

Model 2: Decision Tree with max_depth=13 (optimized)
Accuracy: 0.91
Precision: 0.91 (macro average)
Recall: 0.91(macro average)
F1-score: 0.91 (macro average)

Model 3: Decision Tree with max_leaf_nodes=52 (optimized)
Accuracy: 0.91
Precision: 0.91(macro average)
Recall: 0.91 (macro average)
F1-score: 0.91(macro average)

It seems that all models have equal accuaracy scores. Despite achieving equal accuracy scores, the three models exhibit variations in their parameters.

Model 1, with default parameters, demonstrates that the decision tree classifier can achieve a commendable level of accuracy without any parameter tuning. However, this model's simplicity may limit its ability to capture complex patterns in the data. Another drawback is that base model is relatively large since all its leaf nodes are pure. Therefore, it is not ideal for interpreting.

Model 2, optimized with a max_depth of 13, emphasizes depth control to balance complexity and interpretability. By limiting the tree's depth, this model aims to prevent overfitting while still allowing for the capture of meaningful patterns. The resulting performance metrics indicate successful optimization without sacrificing accuracy.

Model 3, optimized with a max_leaf_nodes value of 52, focuses on controlling the tree's size to prevent overfitting. By setting a maximum number of leaf nodes, this model ensures that the decision tree remains manageable in size. The equal performance metrics compared to Model 1 and model 2 suggest that both depth and leaf node control can lead to similar levels of accuracy with reduced and more controlled and manageable tree size.

Overall, the equal accuracy scores among the three models show the effectiveness of decision tree classifiers in capturing patterns within the dataset. However, parameter configuration depends on the specific dataset, goal and problem. Three models’ scores might be completely different if the same procedure were performed in different dataset.

In conclusion, the decision on which model to use depends on the desired balance between accuracy, interpretability and efficiency.