RSE4207

AI & Machine Learning

Milestone 2

Group 4

Name:

Derek Lim Jun Hao

Muhammad Daniel Bin Rosli

Muhammad Syamim Bin Shamsulbani

Calvin Khor Zhen Loong

[Introduction 1](#_Toc1406005858)

[Data Pre-Processing 2](#_Toc830176444)

[Extreme/Outlier/”Noisy” Data 4](#_Toc1811605150)

[Categorical Encoding 11](#_Toc922307318)

[Algorithm & Software 12](#_Toc1546727325)

[Performance 13](#_Toc1423213125)

[Logistic Regression 14](#_Toc1862675988)

[Confusion Matrix 14](#_Toc1154336581)

[Precision Score 14](#_Toc1061725647)

[Recall Score 14](#_Toc352002289)

[Accuracy 15](#_Toc626053621)

[Error Matrix 15](#_Toc647831869)

[AUC (Receiver Operating Curve) 15](#_Toc914417842)

[MCC 15](#_Toc447274722)

[MAE 15](#_Toc1363632120)

[MSE 15](#_Toc1661203112)

[RMSE 15](#_Toc861637655)

[Random Forest 16](#_Toc1146909706)

[Confusion Matrix 17](#_Toc1311953307)

[Precision Score 17](#_Toc2017698215)

[Recall Score 17](#_Toc40453830)

[Accuracy 18](#_Toc1322588069)

[Error Matrix 18](#_Toc1893589461)

[AUC (Receiver Operating Curve) 18](#_Toc519024218)

[MCC 18](#_Toc1994495594)

[MAE 18](#_Toc37310283)

[MSE 19](#_Toc446452412)

[RMSE 19](#_Toc632308510)

[K-Nearest Neighbor 19](#_Toc1234859982)

[Confusion Matrix 20](#_Toc1894712341)

[Precision Score 21](#_Toc1516032497)

[Recall Score 21](#_Toc876745428)

[Accuracy 21](#_Toc1554966017)

[Error Matrix 21](#_Toc1664552655)

[AUC (Receiver Operating Curve) 21](#_Toc1338473781)

[MCC 21](#_Toc1080962465)

[MAE 22](#_Toc2044555391)

[MSE 22](#_Toc784836434)

[RMSE 22](#_Toc760025868)

[Reflection on System Performance 22](#_Toc1941521509)

[Conclusion 22](#_Toc1736408830)

[References 22](#_Toc231180469)

# Introduction

The advent of advanced data analytics and machine learning techniques has opened new avenues for healthcare customization. Leveraging on these technologies, our team seeks to harness the power of predictive modelling to categorize patients into distinct obesity levels, enabling healthcare providers to tailor interventions that align with the unique characteristics of each individual. By combining clinical data, lifestyle factors, and other relevant variables, our program aims to offer a comprehensive and precise assessment of a patient's obesity status. The report discusses how data was processed, rationale for algorithm selection and performance of chosen AI and ML model.

# Data Pre-Processing

1. We analyse the dataset to determine the relevance of how the data might affect the results of the patients with high possibility of developing obesity or are already developing obesity.

The “Patient ID” column has been classified as data to better identify each entry, hence these data will not be included in the training model.

In the pre-processing phase, we designed a program to clean the dataset and change all alphabetical values to numerical values so it can be better analysed by our ML model later on.

Age data were rounded off to whole numbers.

*Rationale: Age are generally measured in whole numbers and there is no reason to measure*   *them in decimals*

Gender was changed to binary also where Male: 1, Female: 0

*Rationale: Gender in most cases is a binary choice (With exceptions of LGBT which are not*   *considered in this dataset)*

Height and Weights data were rounded off to 1 decimal place

*Rationale: Height and Weight data are measured to 1 DP in most cases, need no more than 1DP*   *to determine height and weight of a person*

fam\_hist\_over-wt/SMOKE/ SCC and FAVC data were changed to binary where yes: 1, no: 0

*Rationale: Yes/No is a binary decision which can be translated to 1s and 0s*

FCVP / CH2O / FAF / TUE and NCP data were rounded off to whole numbers as

*Rationale: They are just values that if higher, means more frequent. Decimals would be too small*   *changes so it's better to just measure it to the nearest whole number.*

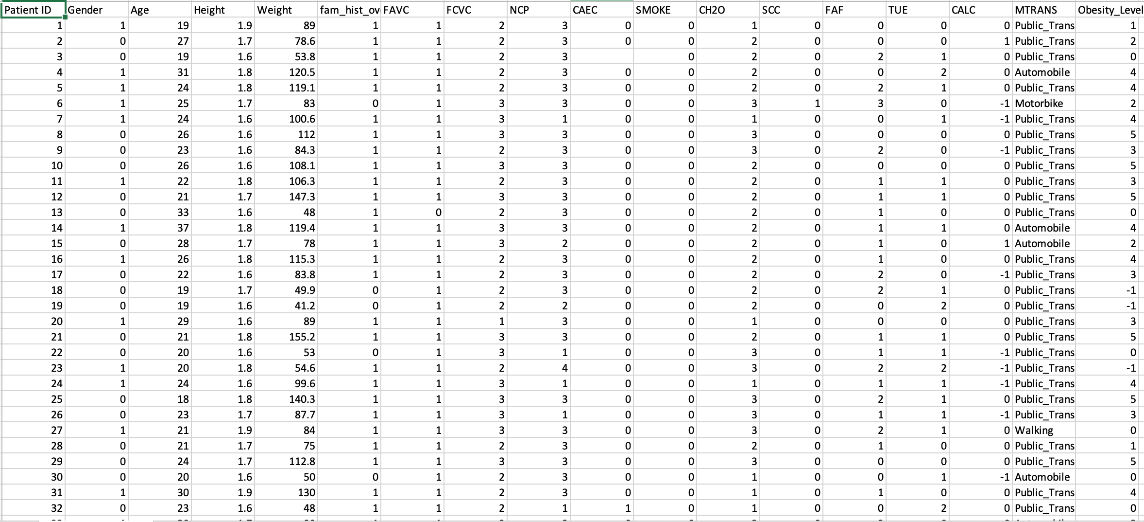
CAEC/CALC data were changed to numerical representation where no: -1, Sometimes: 0, Frequently: 1

Rationale: For analysis purposes

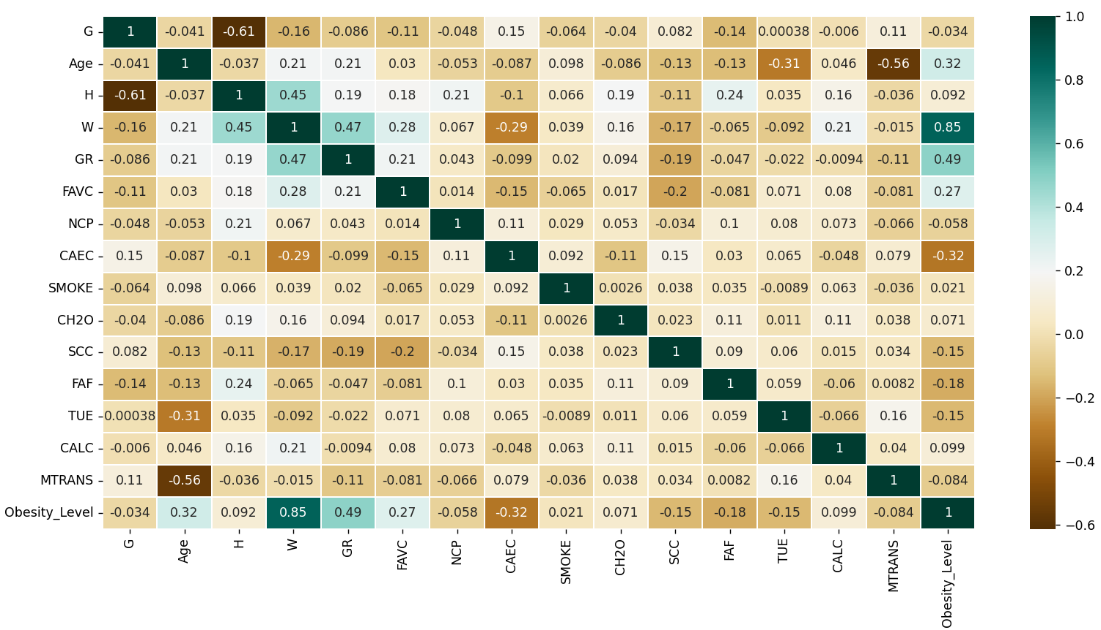
Obesity\_Level data were changed to numerical representation where

Obesity\_Type\_I : 3, Obesity\_Type\_II : 4, Obesity\_Type\_III : 5, Overweight\_Level\_II : 2, Overweight\_Level\_I : 1, Normal\_Weight : 0, Insufficient\_Weight : -1

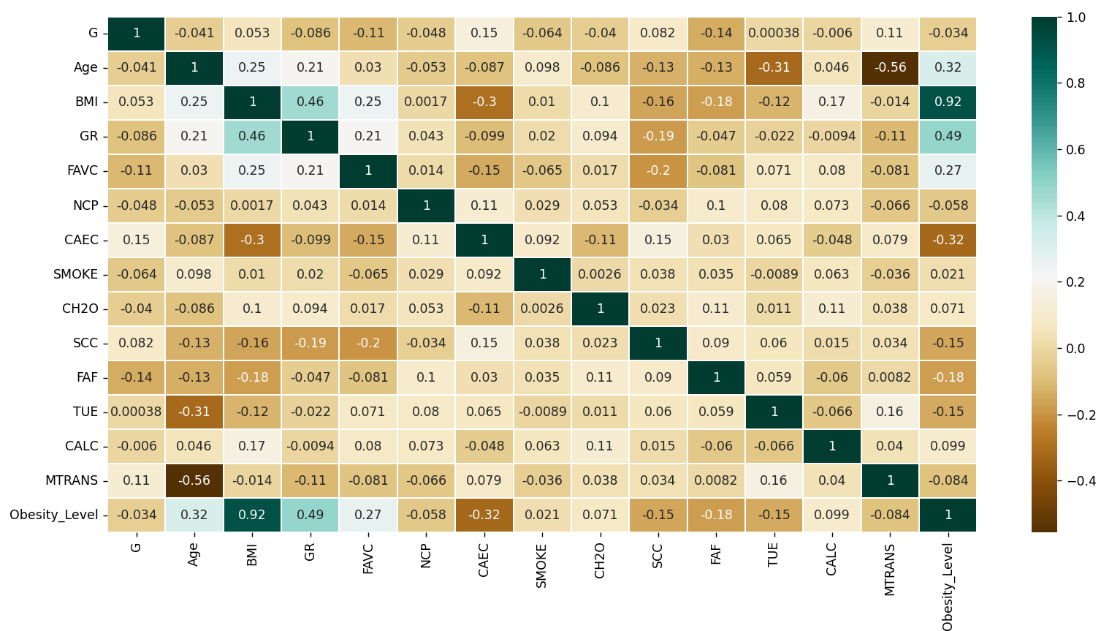
These modified fields will be saved to a new CSV file which will be the input to MS2\_abnormalies.py for detecting outliers and extreme abnormal datas.



2. We then proceed to analyse the correlation between each data, to determine if any pairs of columns have significant relationship to each other.



Ignoring the “Obesity\_Level”, there are a few pairs of columns that have some significant correlation in the correlation matrix, with the highest being “Gender” and “H” at –0.61, which indicates that male might be taller than female on average. This could be due to the classification of male as value “0” and female as value “1”, which the matrix interprets that the gender (Gender) with “0” values tend to have higher values for their height (H), explaining the inverse correlation. Another interesting note is that one of the higher correlations, between “H” and “W” at 0.45, has a known formula which is known as Body Mass Index (BMI). After combining "H” and “W” into “BMI”, we evaluate the correlation again and the correlation index with “Obesity\_Level” did increase as compared to either “H” or “W”, from 0.85 to 0.92.



3. Then, the following processes are performed to clean the dataset that will be used to train our model.

## Extreme/Outlier/”Noisy” Data

Initially during our data preprocessing, we encountered some of the data to be abnormal (e.g., Age with decimal value). If this data were not detected, it might skew the evaluation of the columns that such data exists in there.

To ensure that the data is valid before checking the extreme/outlier/noisy data, we designed a program where you can select which column the user would like to detect the abnormal data.

Summary of Patient likely to be obese

With the visualized data, we can infer the effect of how each data columns may affect the result of patient being obese.

Obese vs Gender:

A graph of blue and orange bars

Description automatically generated

Gender of the patient does not affect the result of them being obese by a significant amount as both genders has similar probability.

Obese vs Age:

A graph of blue rectangular bars

Description automatically generated

Patients within the age group 25 to 52 have a higher chance of being obese compared to those younger or older.

Obesity vs Height:

A graph of blue rectangular bars

Description automatically generated

Taller patient has a slightly higher probability of being obese.

Weight vs Obese:

A graph with blue bars

Description automatically generated

Patient above 85.74kg has a significant increased probability of being obese.

Obese vs Overweight Family History:

A graph with blue and orange squares

Description automatically generated

Patient with family that has history of being overweight significantly increase their probability of being obese.

Obese vs Frequently Consumes High Calorie Food:

A graph of a bar

Description automatically generated

Patient that frequently consumes high calorie food has higher probability of being obese.

Obese vs Vegetables Consumption Frequency:

A graph with numbers and a number of bars

Description automatically generated with medium confidence

Patient with frequency index of more than 1.4 for consuming vegetables has higher probability of being obese.

Obese vs Main Meals Consumption:

A graph of blue rectangular bars

Description automatically generated

Consuming more main meals does not affect the probability of patient becoming obese significantly.

Obese vs Consumption of Additional Food Between Meals:

A graph of a bar chart

Description automatically generated

Patient who doesn’t consume additional food between meals or consumes additional food between meals sometimes has higher probability of being obese.

Obese vs Smoke:

A graph of smoke and smoke

Description automatically generated

Smoking does not seem to affect the probability of patient being obese by a significant amount as patient who do not smoke has a slightly lower probability of being obese than those who do.

Obese vs Daily Water Consumption Frequency:

A graph of blue bars

Description automatically generated

Daily water consumption frequency does not seem to affect the probability of patient being obese by a significant amount, as there is no observable trend in the probability.

Obese vs Self-monitor:

A graph of a bar

Description automatically generated

Patient who do not keep track of their calorie intake has higher probability of being obese.

Obese vs Physical Activity Frequency:

A graph of blue rectangular bars

Description automatically generated

Patient with less than 1.2 index of physical activity frequency have higher probability of being obese.

Obese vs Time Spent Using Technological Devices:

A graph of blue rectangular bars

Description automatically generated

Time spent using technological devices does not have a clear effect on patient being obese, as there is no observable pattern in the probability of patient being obese.

Obese vs Alcohol Consumption Frequency:

A bar graph with text and a bar chart

Description automatically generated with medium confidence

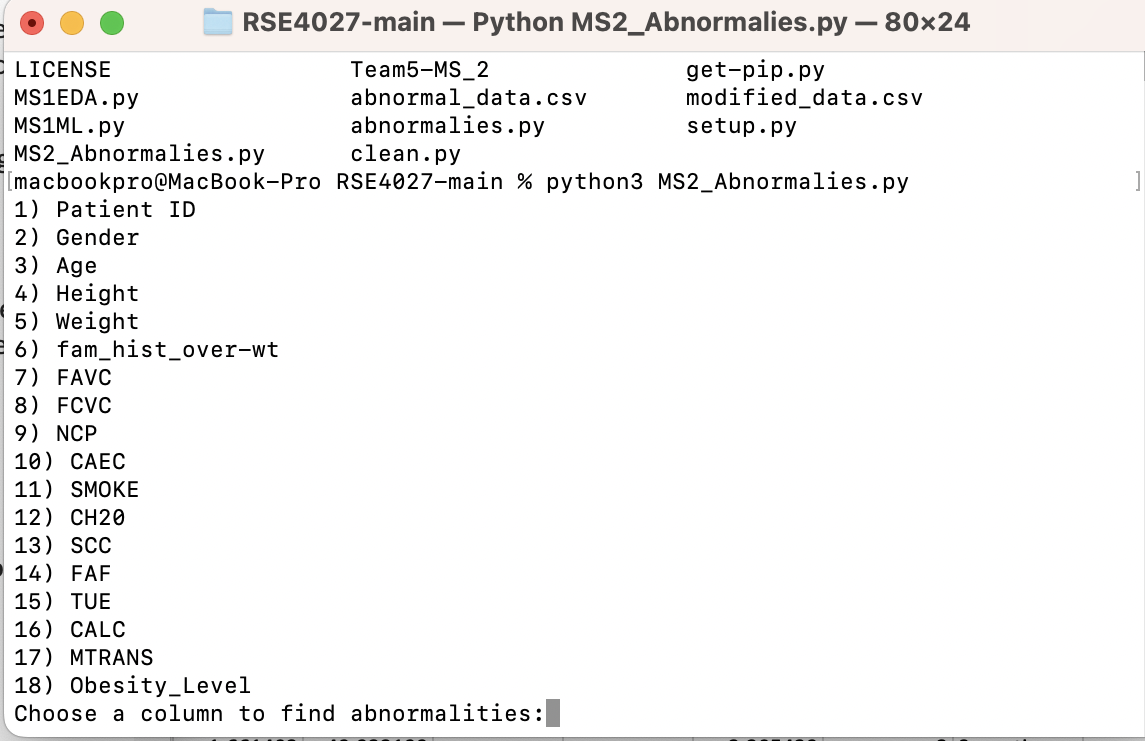
Alcohol consumption frequency does not have a have a clear effect on patient being obese, as there is no observable pattern in the probability of patient being obese.

Obese vs Mode of Transportation:

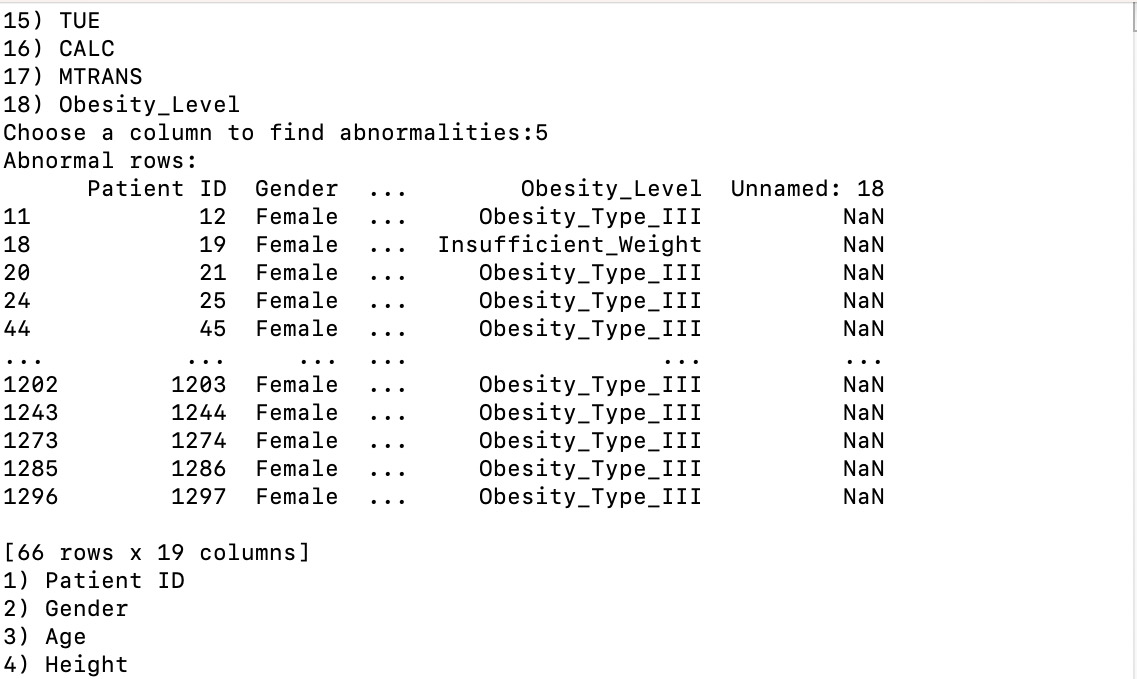
A graph of a patient mode of travelling

Description automatically generated

There appears to be some sort of pattern in which the less effort required for the patient’s mode of transport, the higher the probability of them being obese.



Selecting the number corresponding to the data column will display the abnormal/outlier data based on that column, for example; selecting '5' will display all abnormal data under 'Weight' which can be used later to train a set of data which doesn't include these outlier data.



To determine these outlier data, we incorporate the Isolation Forest Model for anomaly detection. We chose to use this model as it employs binary tress to detect anomalies, resulting in linear time complexity and low memory usage that is good for processing large datasets which is suitable for our dataset with over 500 data.

We can save the data that are not valid into a separate csv file for future pattern finding



## Categorical Encoding

Since the model uses algorithms that functions directly with numerical value, we have decided to convert the values in (“SMOKE” and “Gender” etc) column into numerical format. The values in these columns mentioned are categorical data, by assigning a unique integer value to each possible category, it gives the category itself some form of meaning, which is also known as label encoding. This process allows the model to train with categorical data that are “interpreted” beforehand, hence the numerical values tied to the data have better meaning.

# Algorithm & Software

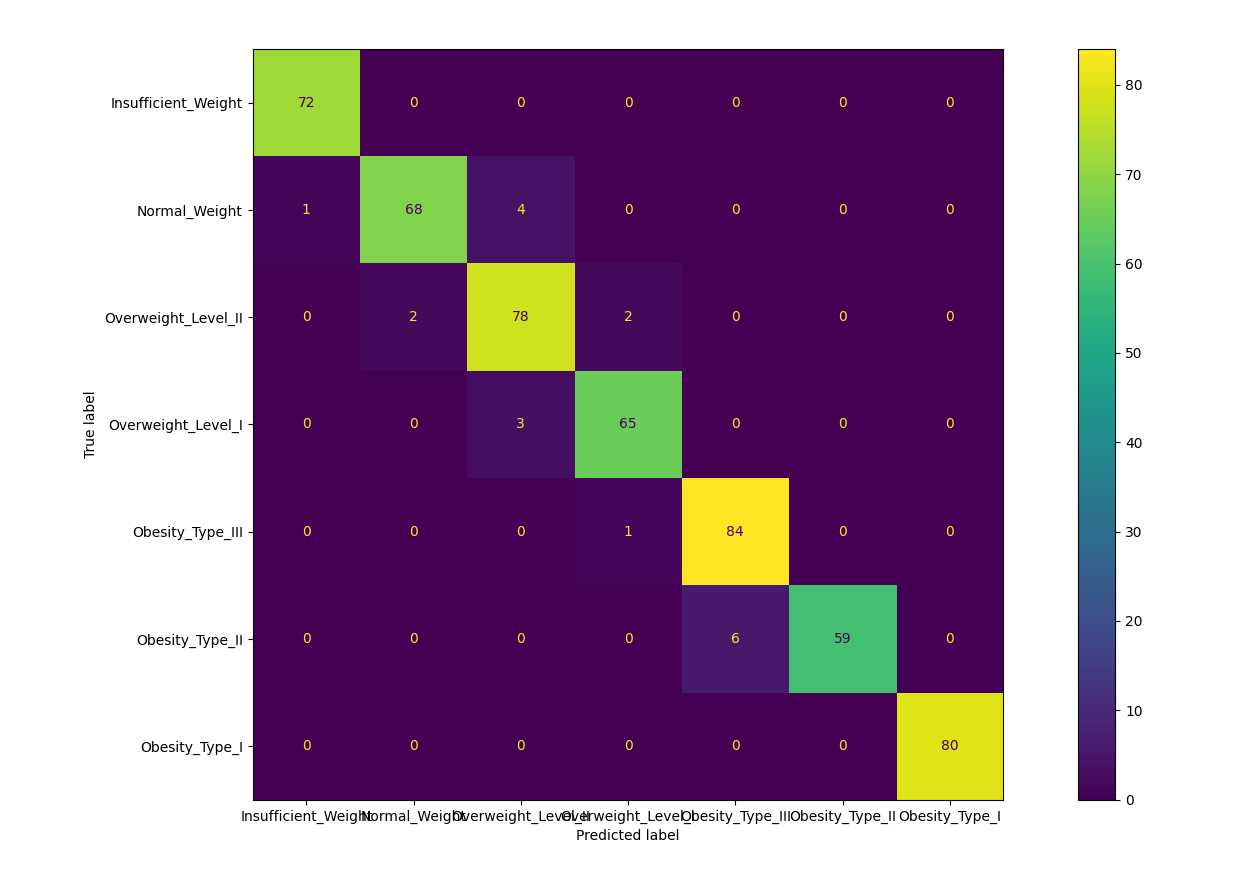
Our team explored multiple AI algorithms such as Logistic Regression, Random Forest, and K-Nearest Neighbour. In the Performance section, we discuss the usage of all three prediction models to show why we decided on using them. The models are trained using the given dataset which is split into training and testing sets to evaluate their performance.

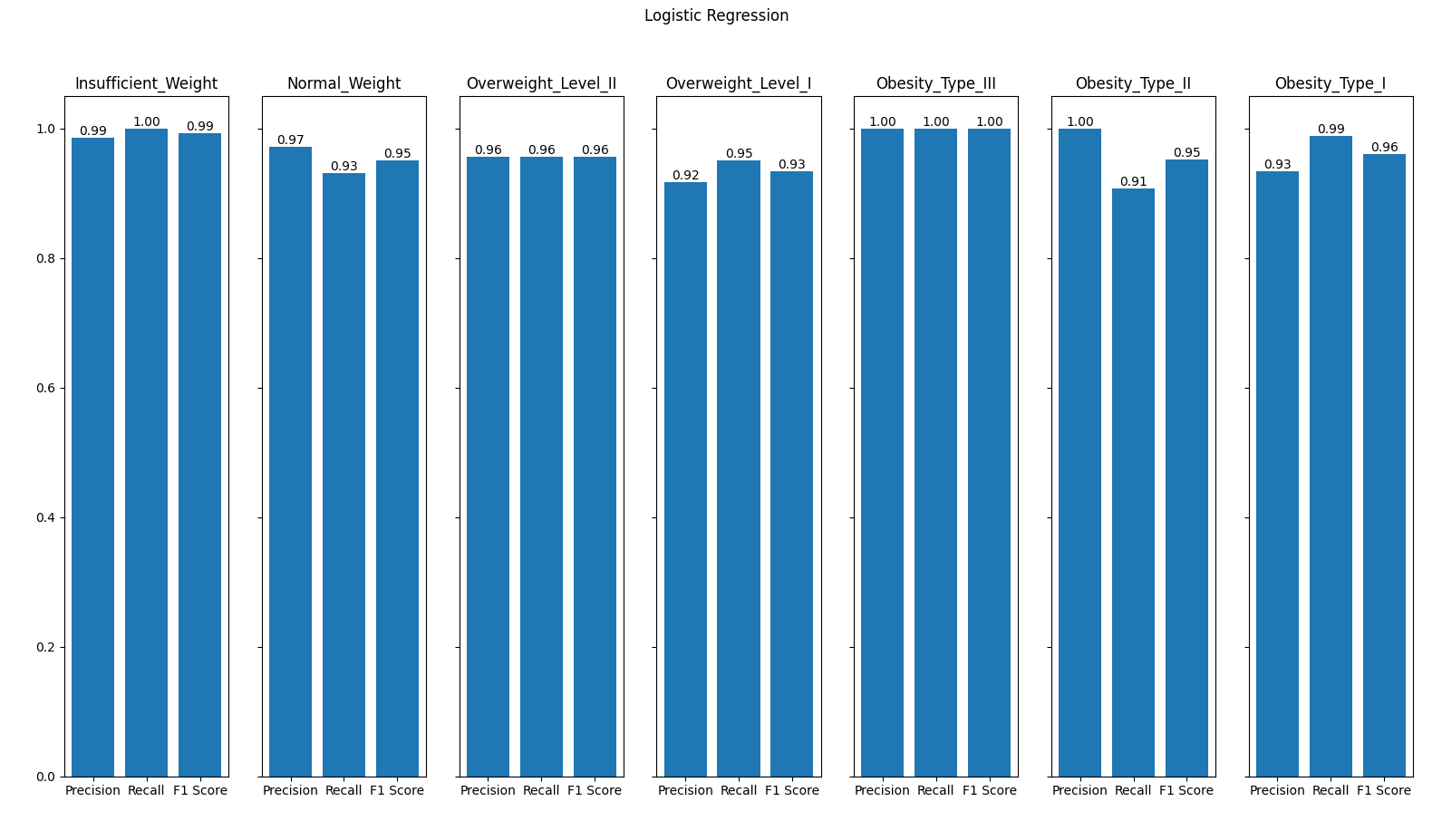
Since the criterion of our system is to determine the patients who are at risk of obesity, this model fits the requirement as we have multiple variables that we can analyse to determine the probability of a patient’s risk classification.

Utilizing our chosen ML model, we can see that there are some relations between certain factors. These data and factors are crucial for decision making. In this case, our program aids the client in correctly identifying which patient to administer a miracle drug for their clinical trials. Individuals who are most likely to benefit are those that have high probability of developing obesity or are on the way to becoming obese.

# Performance

## Logistic Regression





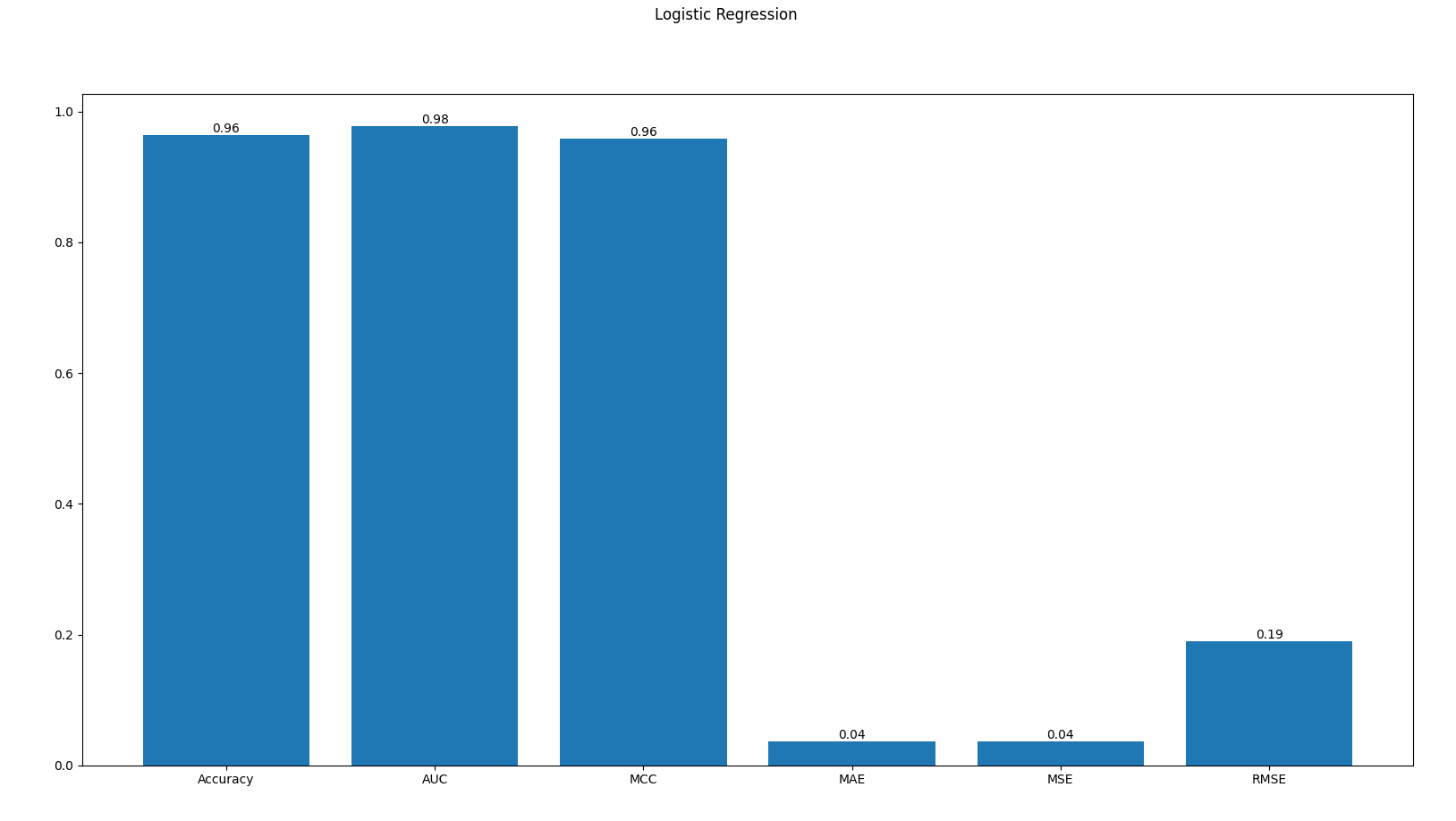
### Confusion Matrix

### Precision Score

By taking the ration of True Positives vs all predicted Positives per class, we can derive the precision on a class-by-class basis. From this, we observe that the performance of the Logistic Regression model is very well with the lowest precision coming in at 92% for classifying Overweight Level 1 patients. However, considering that our goal is to find patients that are at risk and/or going to be obese, this includes those in the Overweight Level 1 group. It would be better if we could increase the precision for specifically the Overweight groups.

### Recall Score

By taking the ratio of True Positives vs all actual Positives per class, we can derive the recall on a class-by-class basis. From this we can observe that the performance of Logistic Regression is very well with the lowest score coming in at 91% for Obesity Type 2 patients. This is acceptable as the miss classification of Type 2 Obesity patients in this context will not result in the rise of opportunity costs as they are not the target of the medicine.



### Accuracy

The average accuracy of the Logistic Regression model across all classes is 96% which results in a rather accurate model for overall classification of patient’s obesity levels.

### Error Matrix

#### AUC (Receiver Operating Curve)

Since this is a multi-class classification problem, the area under the curve has been computed via a ‘one vs one’ criteria. This means that the values are computed in respect to each class individually rather than being measured against all predictions. The areas of each class are then averaged out to give us an AUC score of 98%.

#### MCC

By taking the Matthews Correlation Coefficient of the model’s predictions, we get a score of 96%. This indicates a close to perfect prediction model. This indicates that the model is not overfitted and has enough generalizations to be used as a predictor.

#### MAE

Taking the mean absolute errors, we get a value of 0.04 or 4%. From this, we can infer that the logistic regression model has an average error rate of 4%. From this we can observe that the total amount of errors is very small.

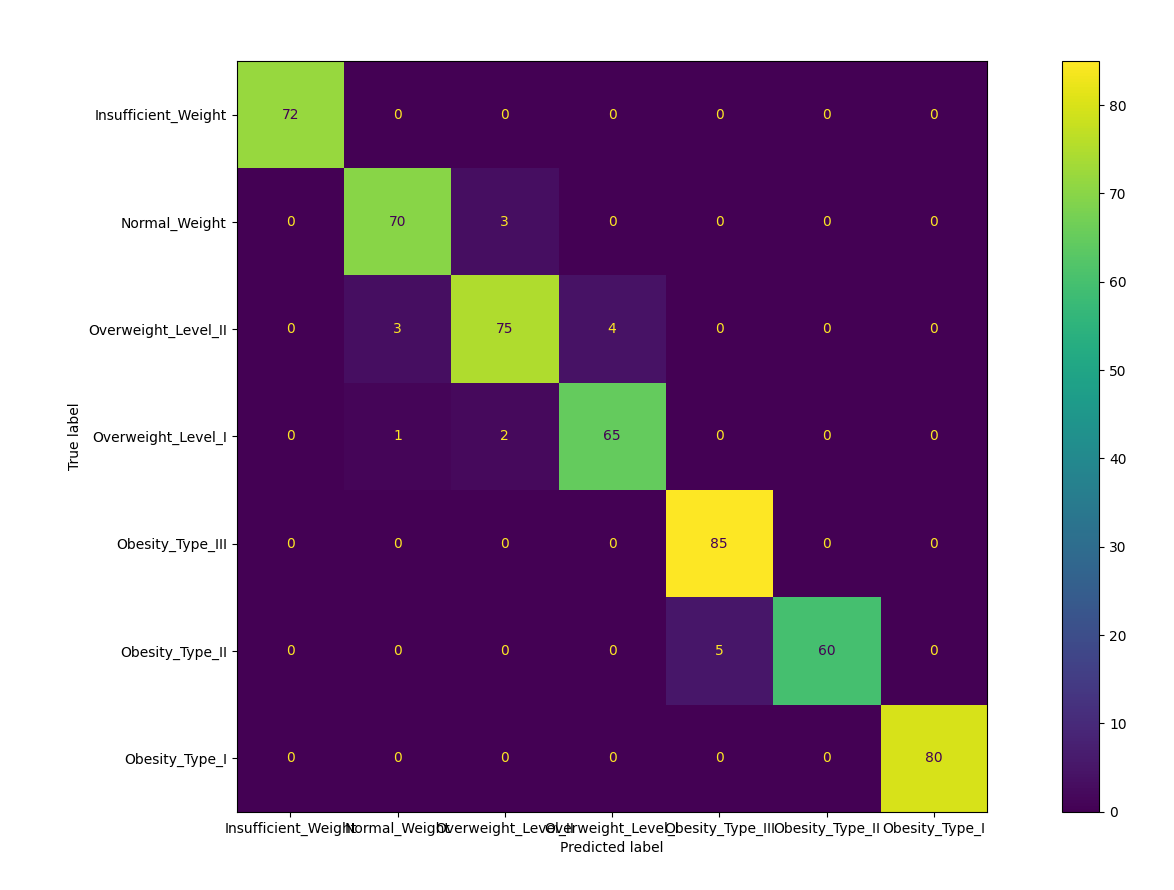
#### MSE

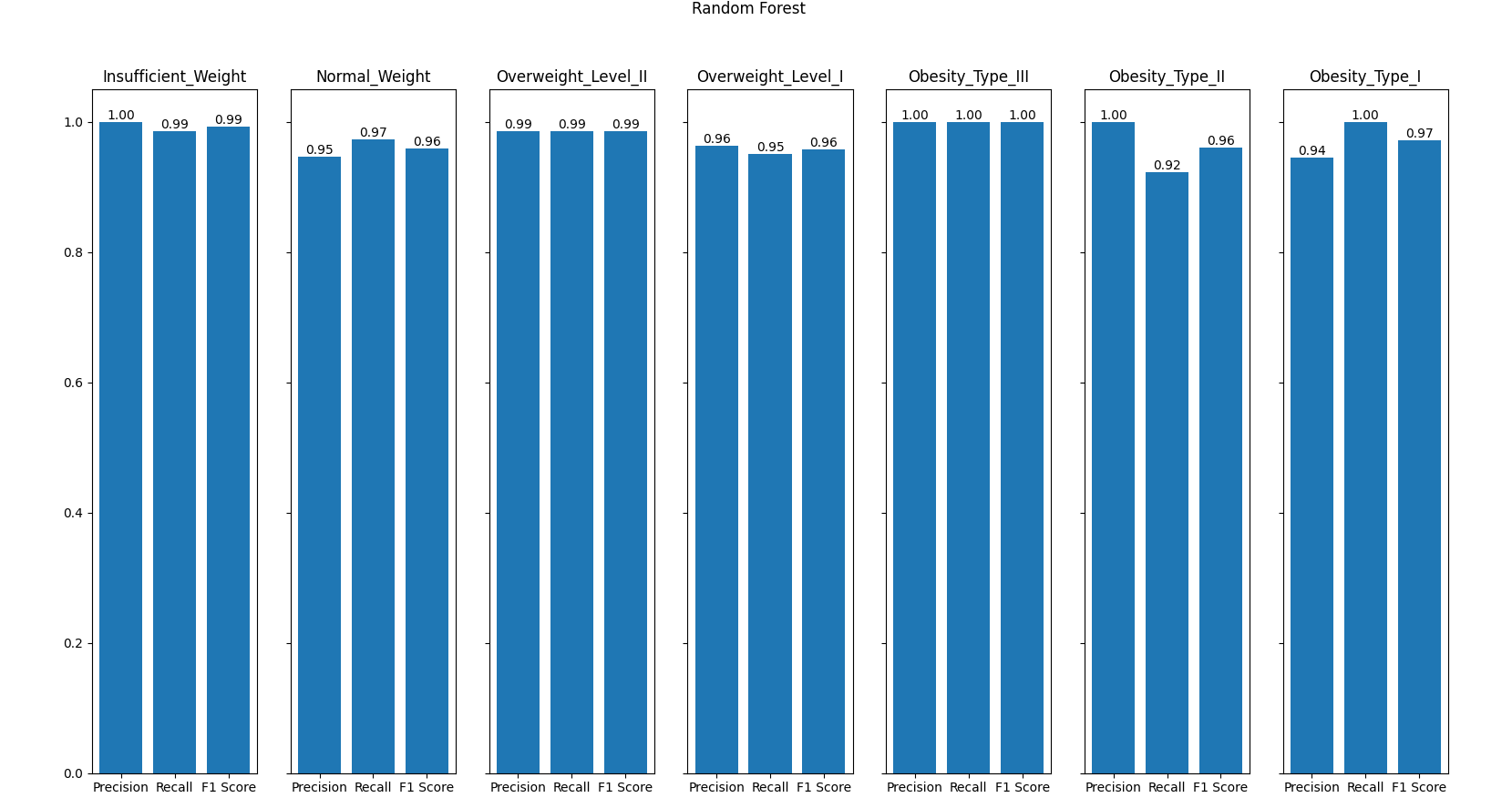
Taking the mean square of the error results in no change in the error score, this indicates that there is no outlier information in the training dataset.

#### RMSE

Taking the root mean square of the error results in an increase in the error score, however the error is small to begin with which results in a rather insignificant metric.

## Random Forest





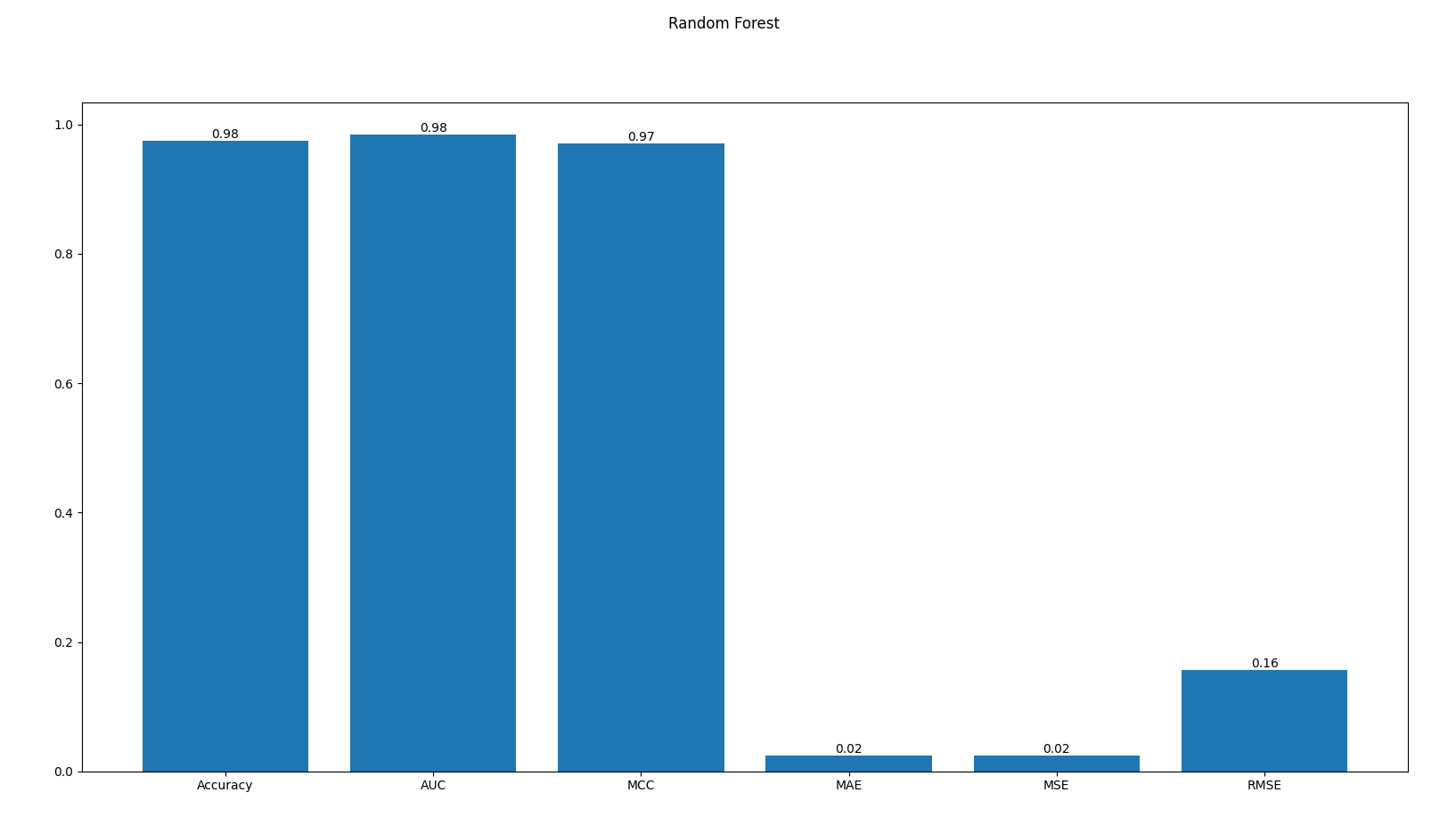
### Confusion Matrix

### Precision Score

By taking the ration of True Positives vs all predicted Positives per class, we can derive the precision on a class-by-class basis. From this, we observe that the performance of the Random Forest model is very well with the lowest precision coming in at 94% for classifying Obesity Type 1 patients. However, since the main goal of the model is to find those at risk or going to be obese, which are those in the Overweight category with precisions of 96% and 99%, this shows that the Random Forest model has outperformed the Logistic Regression model in precision.

### Recall Score

By taking the ratio of True Positives vs all actual Positives per class, we can derive the recall on a class-by-class basis. From this we can observe that the performance of Random Forest is very well with the lowest score coming in at 92% for Obesity Type 2 patients. This is acceptable as the miss classification of Type 2 Obesity patients in this context will not result in the rise of opportunity costs as they are not the target of the medicine.



### Accuracy

The average accuracy of the Logistic Regression model across all classes is 98% which results in a rather accurate model for overall classification of patient’s obesity levels.

### Error Matrix

Since this is a multi-class classification problem, the area under the curve has been computed via a ‘one vs one’ criteria. This means that the values are computed in respect to each class individually rather than being measured against all predictions. The areas of each class are then averaged out to give us an AUC score of 98%.

#### AUC (Receiver Operating Curve)

Since this is a multi-class classification problem, the area under the curve has been computed via a ‘one vs one’ criteria. This means that the values are computed in respect to each class individually rather than being measured against all predictions. The areas of each class are then averaged out to give us an AUC score of 98%.

#### MCC

By taking the Matthews Correlation Coefficient of the model’s predictions, we get a score of 97%. This indicates a close to perfect prediction model. This indicates that the model is not overfitted and has enough generalizations to be used as a predictor.

#### MAE

Taking the mean absolute errors, we get a value of 0.02 or 2%. From this, we can infer that the Random Forest model has an average error rate of 2%. From this we can observe that the total amount of errors is very small, and it has 50% less errors that the Logistic Regression model.

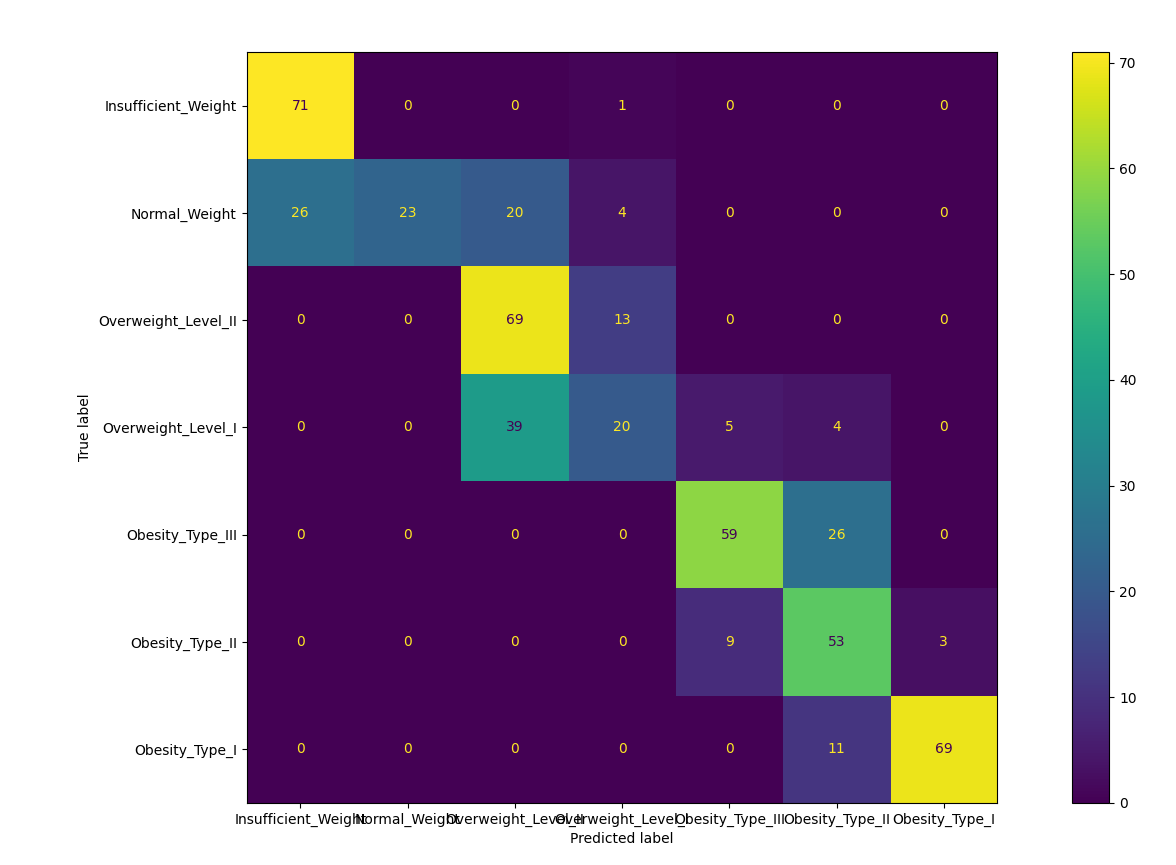
#### MSE

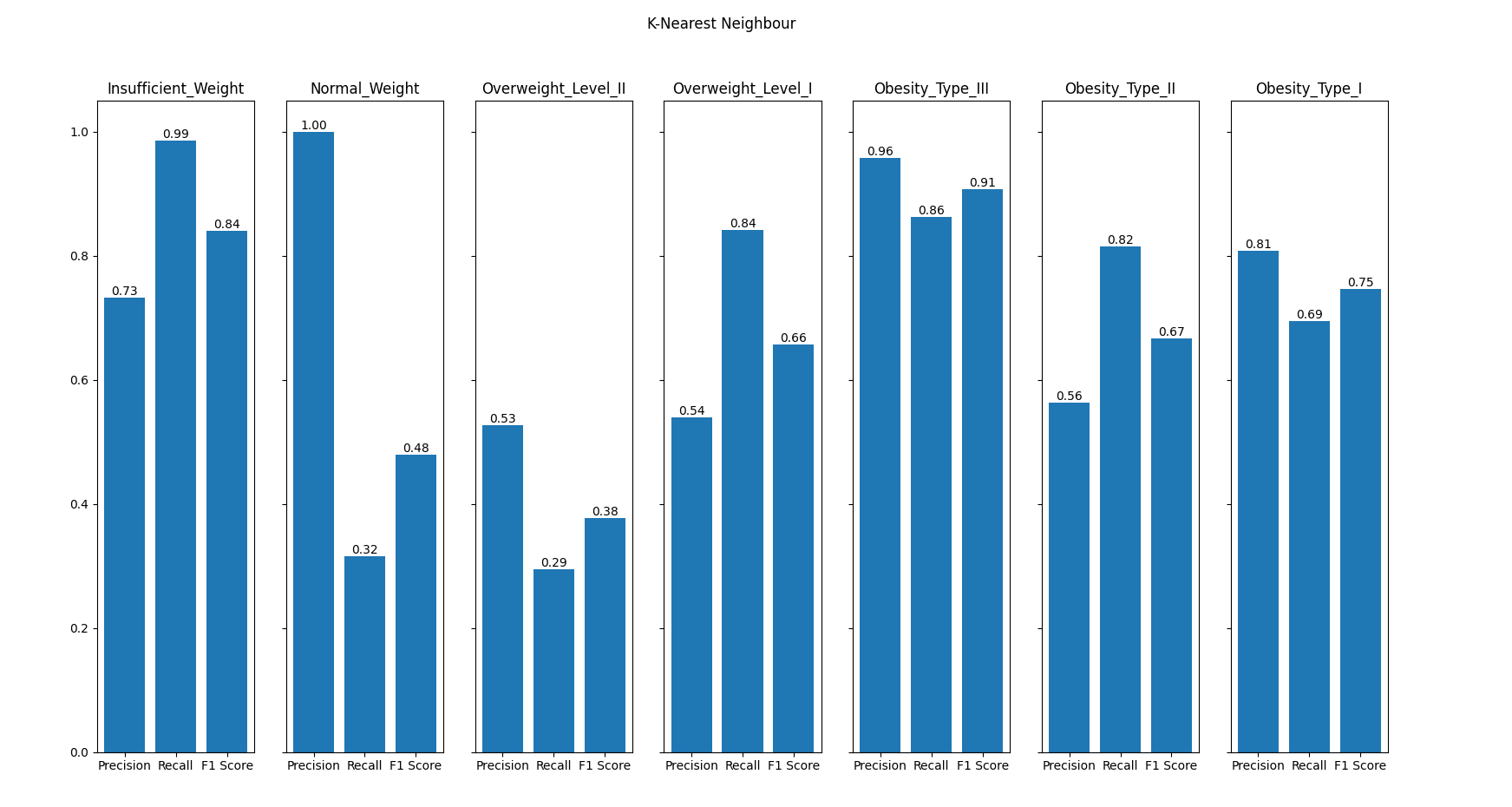
Taking the mean square of the error results in no change in the error score, this indicates that there is no outlier information in the training dataset.

#### RMSE

Taking the root mean square of the error results in an increase in the error score, however the error is small to begin with which results in a rather insignificant metric.

## K-Nearest Neighbor





### Confusion Matrix

By calculating the respective True Positive and False Positive of each class, the precision for each class can be calculated to measure the accuracy of the positive predictions. The recall can also be calculated using the False Negative values of each class as well.

The confusion matrix shows the distribution of True Positive, False Positive, False Negative, and True Negative values based on the KNN model predictions. Each value is used to calculate the precision, recall, and F1 score of the model on each class and the values are represented in the following bar graph.

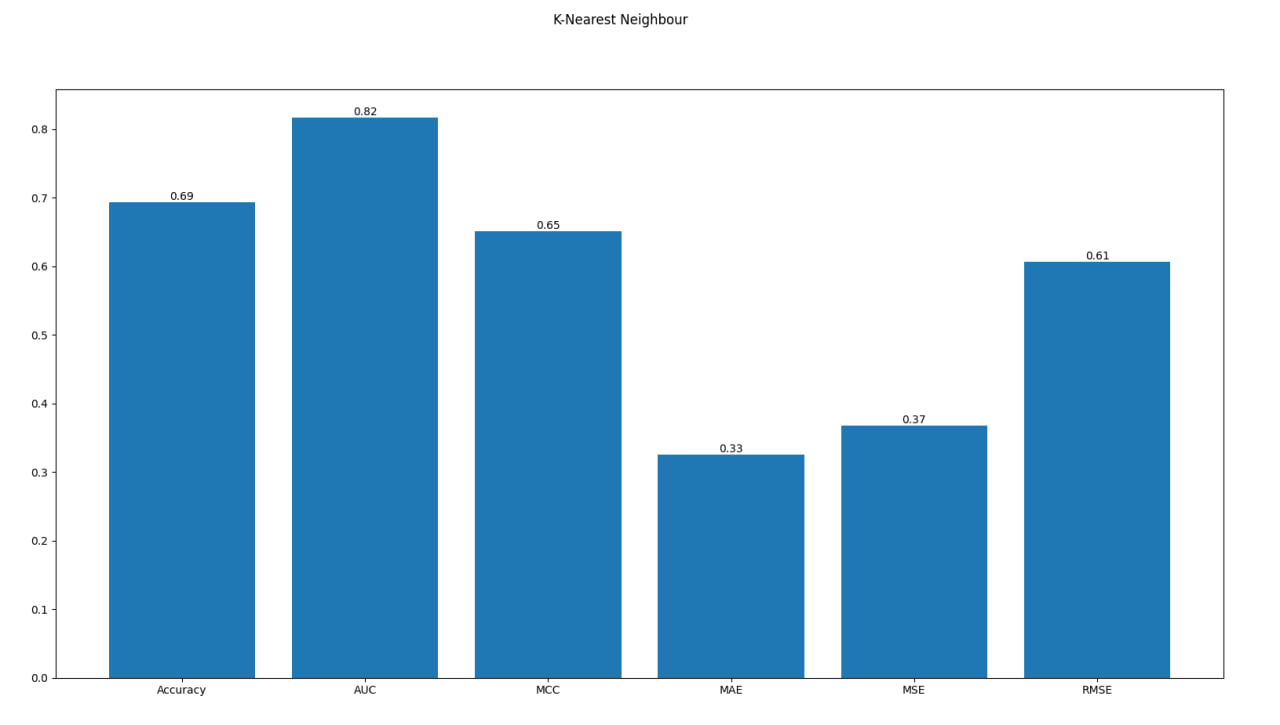
The scores represented in the bar graph above are the respective precision, recall, and F1 Score of the model in predicting the various classes. Overall, the higher the value, meaning a higher bar, corresponds to a better model.

### Precision Score

While KNN has a relatively good performance in predicting the classes, Insufficient Weight and Obesity Type III, it still struggles with most of the classes. Its lowest precision score, 54%, is in predicting for Overweight Level I. As the goal of the model is to find patients that are at risk of and/or going to be obese, the model struggling with the overweight class shows that it is unable to identify those who are likely to be at risk.

### Recall Score

The recall score on average is performing worse than the precision meaning that the model is unable to correctly identify many classes. It is again struggling with those in the overweight classes which is the focus of the prediction model.



### Accuracy

The accuracy of the KNN model is 69% which is not entirely a bad performance in predicting the correct class values.

### Error Matrix

### *AUC (Receiver Operating Curve)*

The Area Under Curve is calculated using a ‘one vs one’ criteria for analysis on the multiclass problem. Its value is calculated for each pair of classes and then averaged out instead of comparing each class against the rest. The AUC value of the KNN model is 82%.

### *MCC*

Matthew Correlation Coefficient is a metric that evaluates the performance of the prediction model and considers both the correct and incorrect predictions across all classes. The MCC value for the KNN model is 65% which is decent but is just barely getting its predictions correct.

#### MAE

Taking the mean absolute errors, we get a value of 0.33 or 33%. From this, we can infer that the KNN model has an average error rate of 33%.

#### MSE

The mean square error value is at 37%. While not entirely bad, its prediction values deviate much more than the other prediction models used.

#### RMSE

By taking the square root of the MSE value we get the RMSE value which is at 67%. The value is now more easily interpreted to show that the model has a poor performance.

# Reflection on System Performance

From error metrics and confusion matrix using the Logistic Regression method, with data from the Precision and MAE, the client can trust the program 83% of the time given an error of 17% as calculated above in our program. Therefore, the program is relatively trustworthy for the client to use.

# Conclusion

As compared to Milestone 1, this task provided us with much more columns of data, which is why Exploratory Data Analysis step is even more vital to assist in making the judgement of dropping certain columns to avoid the curse of dimensionality. Visualizing the data makes it easier to infer how does each column can contribute to the result of prediction and help understand the presence of correlation in certain data, which was Height and Weight for this task. Additionally, we were able to gather a general summary of the characteristics of patient who are obese that we can then apply into our Machine Learning models.

# References