



UNIVERSITI TEKNOLOGI MALAYSIA  
FACULTY OF COMPUTING  
SEMESTER 1, SESSION 2025/2026

---

**FINAL REPORT**

**OBESITY LEVEL CLASSIFICATION**

SECB3203 : PROGRAMMING FOR BIOINFORMATICS  
SECTION 02

---

**GROUP MEMBER:**

1. MUHAMMAD FARIHIN BIN SALEH	A25CS0102
2. MUHAMMAD MIRZA HASIF BIN MOHD FAHMI	A25CS0108
3. MUHAMMAD NAWFAL BIN MOHD SHAFUDDIN	A25CS0109

**LECTURER NAME** : DR. SEAH CHOON SEN

**GROUP** : GROUP 07

**GITHUB REPOSITORY** : [GITHUB REPOSITORY LINKS](#)

## TABLE OF CONTENTS

<b>1.0</b>	<b>INTRODUCTION</b>	<b>1</b>
1.1	PROBLEM BACKGROUND	1
1.2	PROBLEM STATEMENT	1
1.3	OBJECTIVES	2
1.4	SCOPES & LIMITATIONS	3
<b>2.0</b>	<b>SOFTWARE AND HARDWARE REQUIREMENTS</b>	<b>5</b>
2.1	SOFTWARE REQUIREMENTS	5
2.2	HARDWARE REQUIREMENTS	5
<b>3.0</b>	<b>FLOWCHART OF THE PURPOSED APPROACH</b>	<b>6</b>
3.1	FLOWCHART DIAGRAM	6
3.2	FLOWCHART EXPLANATION	7
<b>4.0</b>	<b>IMPORTING DATASET AND DATA WRANGLING</b>	<b>8</b>
4.1	IMPORTING DATASET	8
4.1.1	Understanding the Data	8
4.1.2	Importing and Exporting Data in Python	9
4.1.3	Getting Started Analysing Data in Python	11
4.1.4	Python Packages for Data Science	13
4.2	DATA WRANGLING (Pandas / Numpy)	14
4.2.1	Identifying and Handling Missing Values	14
4.2.2	Data Formatting	15
4.2.3	Data Normalisation (Centering / Scaling)	16
4.2.4	Binning	17
4.2.5	Indicator Variables (Encoding Categorical Data)	18
<b>5.0</b>	<b>EXPLORATORY DATA ANALYSIS</b>	<b>20</b>
5.1	DESCRIPTIVE ANALYSIS	20

5.2	BASIC OF GROUPING	22
5.3	ANOVA	23
5.4	CORRELATION ANALYSIS	26
6.0	PREPROCESSING FOR MODELLING	28
6.1	MODEL DEVELOPMENT	30
7.0	MODEL EVALUATION	33
7.1	EVALUATE MODEL	33
7.1.1	Selected Models	33
7.1.2	Model Comparison	34
7.1.3	Confusion Matrix	36
7.1.4	Classification Report	38
7.2	OVER-FITTING & UNDER-FITTING	39
7.3	RIDGE REGRESSION	42
7.4	GRID SEARCH	43
7.5	MODEL REFINEMENT	45
7.6	FINAL COMPARISON	47
8.0	CONCLUSION	49
	APPENDIX	50

## 1.0 INTRODUCTION

Obesity is a complex and multifactorial health condition associated with increased risks of chronic diseases such as diabetes, cardiovascular disorders, and certain cancers. With rising global obesity rates, there is a growing need for accurate and automated classification systems that can assist in early detection and risk assessment. This project, **Obesity Level Classification** aims to develop a multi-class classification system to predict obesity levels based on individual lifestyle, dietary, and physiological attributes. Using machine learning models including **Random Forest**, **XGBoost**, **Logistic Regression**, and **Support Vector Machines**, we intend to create a reliable tool for obesity risk stratification, contributing to personalized health interventions and public health planning.

### 1.1 PROBLEM BACKGROUND

Obesity classification traditionally relies on body mass index (BMI) and other anthropometric measurements, which may not fully capture the influence of behavioural and metabolic factors. Manual assessment and self-reported data can be subjective and prone to error. Recent advances in machine learning offer opportunities to integrate diverse data sources such as physical activity, eating habits, and demographic information which can be used to improve classification accuracy. The Obesity Levels dataset from Kaggle provides a structured set of features that can be leveraged to train and evaluate predictive models, offering a data-driven approach to obesity assessment.

### 1.2 PROBLEM STATEMENT

Current methods for obesity classification often lack integration of multidimensional lifestyle data and may not adapt well to individual variability. There is a need for an automated, accurate, and scalable system that can classify obesity levels using a comprehensive set of features. This project addresses this gap by developing and comparing multiple machine

learning models to identify the most effective approach for obesity level prediction, thereby supporting healthcare professionals in early intervention and personalized care.

### **1.3 OBJECTIVES**

#### **1. Preprocess and explore the Obesity Levels dataset**

Clean data, encode categorical variables, normalize features, and conduct exploratory data analysis (EDA).

#### **2. Implement and train four classification models**

Build and train Random Forest, XGBoost, Logistic Regression, and SVM models for obesity level prediction.

#### **3. Evaluate and compare model performance**

Assess models using accuracy, precision, recall, F1-score, and confusion matrices.

#### **4. Identify the best-performing model**

Compare results to select the most accurate and reliable model for obesity classification.

#### **5. Document workflow and maintain GitHub repository**

Keep code and documentation organized on GitHub for reproducibility and collaboration.

## 1.4 SCOPES & LIMITATIONS

### Scopes:

#### 1. **Dataset Utilization**

The project will use the “Obesity Levels” dataset from Kaggle, which includes lifestyle, dietary, and physiological attributes for classification.

#### 2. **Model Implementation**

Four machine learning models which are Random Forest, XGBoost, Logistic Regression, and SVM that will be implemented and trained for multi-class obesity level classification.

#### 3. **Performance Comparison**

Models will be evaluated and compared using standard classification metrics (accuracy, precision, recall, F1-score, confusion matrices).

#### 4. **Tool and Environment**

The project will be developed using Python in a local environment, utilising libraries such as Pandas, Scikit-learn, and Matplotlib/Seaborn.

#### 5. **Documentation**

All code, results, and documentation will be maintained and regularly updated on a GitHub repository for transparency and collaboration.

### Limitation:

#### 1. **Local Execution Only**

The project will not utilize cloud platforms (Microsoft Azure or AWS), limiting scalability and real-time processing capabilities.

#### 2. **Dataset Specificity**

The model’s performance may be constrained by the size, quality, and demographic scope of the Kaggle dataset, affecting generalizability.

3. **Model Interpretability**

Some models, such as Random Forest and XGBoost, may act as “black boxes,” making it difficult to interpret how specific predictions are made.

4. **Lack of Clinical Validation**

The study is computational and does not include real-world clinical testing or integration with healthcare systems.

## 2.0 SOFTWARE AND HARDWARE REQUIREMENTS

### 2.1 SOFTWARE REQUIREMENTS

1. **Python 3.8+**  
Core programming language for data analysis, modelling, and visualization.
2. **Visual Studio Code or Google Colab**  
Primary code editor with Python extensions for development and debugging.
3. **Git & GitHub**  
Version control for collaborative coding, project tracking, and coding documentation.
4. **Required Python Libraries**  
Data Processing: pandas, numpy  
Machine Learning: scikit-learn, xgboost  
Visualization: matplotlib, seaborn

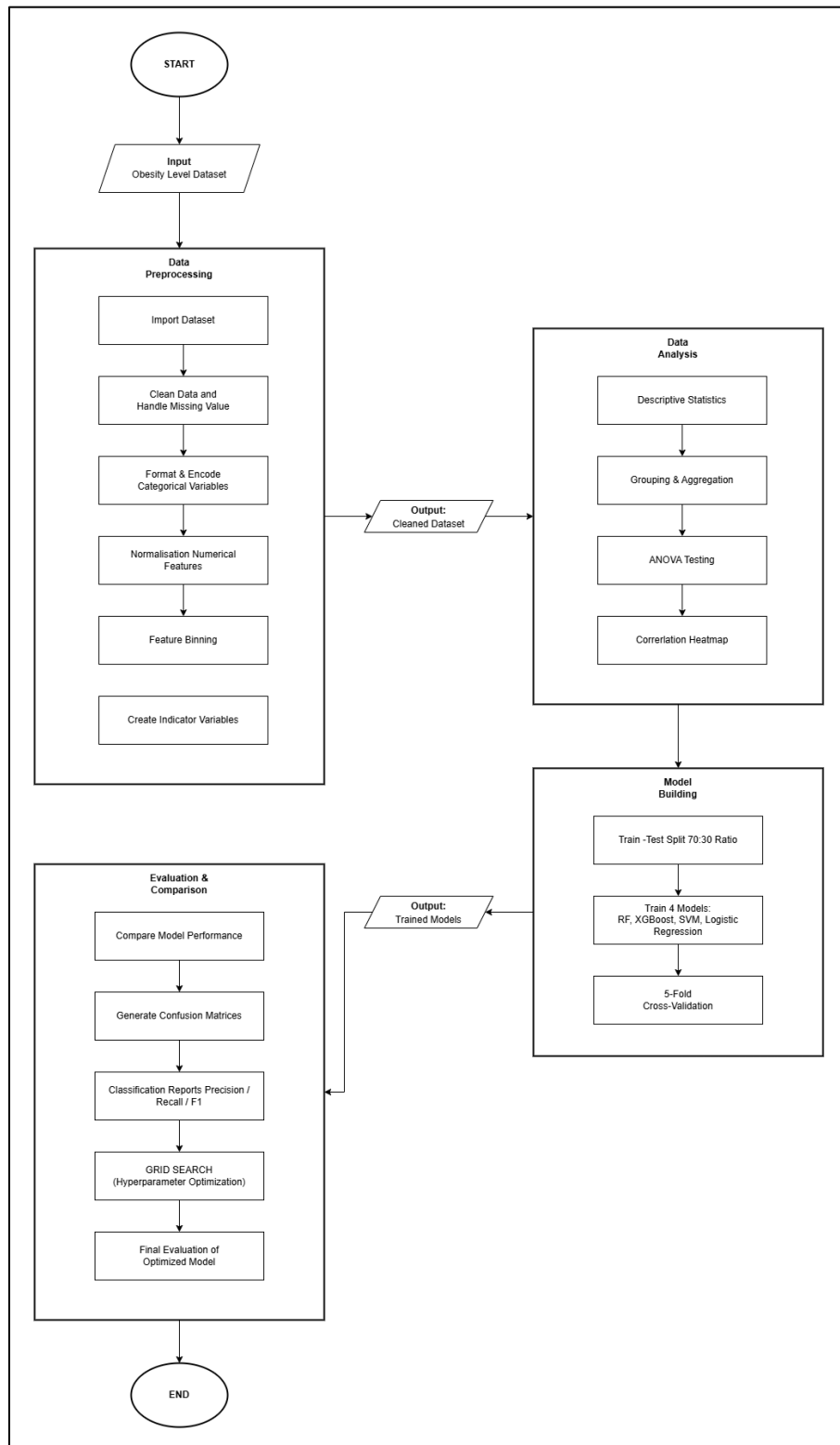
### 2.2 HARDWARE REQUIREMENTS

COMPONENT	REQUIREMENT
Processor	Intel Core i5 or AMD Ryzen 5
Memory (RAM)	8 GB
Storage	5 GB free space for datasets, code, and outputs
Operating System	Windows 10/11, macOS, or Linux

**Table 2.1:** Minimum Hardware Requirement

### 3.0 FLOWCHART OF THE PURPOSED APPROACH

#### 3.1 FLOWCHART DIAGRAM



**Figure 3.1:** Flowchart for Obesity Level Classification Project

## 3.2 FLOWCHART EXPLANATION

### 1. Start and Input

- Project begins with dataset acquisition from Kaggle Obesity Levels dataset.

### 2. Progress 2: Data Wrangling (Process)

- Complete preprocessing pipeline including cleaning, encoding, normalization, and feature engineering.
- Output: Cleaned, structured dataset ready for analysis.

### 3. Progress 3: Exploratory Data Analysis (Process)

- Statistical analysis to understand data patterns and relationships.
- Includes grouping, ANOVA, and correlation visualization.

### 4. Progress 4: Model Development (Process)

- Implementation of four classification algorithms with hyperparameter optimization.
- Cross-validation ensures model robustness.

### 5. Progress 5: Model Evaluation (Process)

- Comprehensive performance assessment using multiple metrics.
- Visualization of results for clear comparison.

### 6. Decision Point

- Evaluate if the best model meets performance criteria.
- If not satisfactory, return to data wrangling for refinement.

### 7. End and Output

- Final model selection and complete project documentation.
- Results compiled in the final report for submission.

## 4.0 IMPORTING DATASET AND DATA WRANGLING

### 4.1 IMPORTING DATASET

#### 4.1.1 Understanding the Data

The dataset used in this project is the **Obesity Levels Dataset** obtained from Kaggle, containing 2,111 records with 17 attributes. This dataset represents a synthetic yet realistic representation of individuals with different obesity levels, based on their dietary habits, physical condition, and lifestyle factors.

#### Feature Description:

Feature Name	Type	Description
Gender	Categorical	Individual's gender
Age	Continuous	Individual's age in years
Height	Continuous	Height in meters
Weight	Continuous	Weight in kilograms
family_history_ with_overweight	Binary	Family history of overweight (yes/no)
FAVC	Binary	Frequent high-caloric food consumption (yes/no)
FCVC	Integer	Frequency of vegetable consumption (1-3 scale)
NCP	Continuous	Number of main meals daily
CAEC	Categorical	Consumption of food between meals (Never/Sometimes/Frequently/Always)
SMOKE	Binary	Smoking habit (yes/no)
CH2O	Continuous	Daily water consumption (in liters)
SCC	Binary	Calorie intake monitoring (yes/no)
FAF	Continuous	Physical activity frequency (0-3 scale)

TUE	Integer	Technology usage time (0-2 scale)
CALC	Categorical	Alcohol consumption frequency (Never/Sometimes/Frequently)
MTRANS	Categorical	Primary transportation method
<b>NObeyesdad (Target)</b>	Categorical	<b>Obesity level classification</b> with 7 classes: <ul style="list-style-type: none"> <li>▪ <b>Insufficient_Weight</b> - Underweight individuals</li> <li>▪ <b>Normal_Weight</b> - Healthy weight range</li> <li>▪ <b>Overweight_Level_I</b> - Mildly overweight</li> <li>▪ <b>Overweight_Level_II</b> - Moderately overweight</li> <li>▪ <b>Obesity_Type_I</b> - Class I obesity</li> <li>▪ <b>Obesity_Type_II</b> - Class II obesity</li> <li>▪ <b>Obesity_Type_III</b> - Class III obesity (severe obesity)</li> </ul>

**Table 4.1:** Obesity Level Dataset Feature Description

This dataset provides a comprehensive view of lifestyle factors contributing to obesity, making it suitable for developing machine learning models to classify individuals into appropriate obesity categories based on their habits and physical metrics.

#### 4.1.2 Importing and Exporting Data in Python

##### CODING

```
# Python libraries
import pandas as pd
import numpy as np
from sklearn.preprocessing import LabelEncoder, MinMaxScaler

# Import dataset
df = pd.read_csv('ObesityDataSet_raw_and_data_sinthetic.csv')
print(f"Dataset Shape: {df.shape}")
print(df.head())
```

The code imports the necessary Python libraries for data manipulation and loads the Obesity Levels dataset from a CSV file. The pandas library is used to read and handle the data, while sklearn.preprocessing is imported for later data transformation tasks. The df.head() function displays the first five rows of the dataset to provide an initial view of its structure.

## OUTPUT

```
Dataset Shape: (2111, 17)
   Age  Gender  Height  Weight      CALC  FAVC  FCVC  NCP  SCC  SMOKE  CH2O  \
0  21.0  Female   1.62   64.0        no    no    2.0   3.0   no    no    2.0
1  21.0  Female   1.52   56.0  Sometimes    no    3.0   3.0  yes    yes    3.0
2  23.0   Male   1.80   77.0  Frequently    no    2.0   3.0   no    no    2.0
3  27.0   Male   1.80   87.0  Frequently    no    3.0   3.0   no    no    2.0
4  22.0   Male   1.78   89.8  Sometimes    no    2.0   1.0   no    no    2.0

   family_history_with_overweight  FAF  TUE      CAEC      MTRANS  \
0                               yes  0.0  1.0  Sometimes  Public_Transportation
1                               yes  3.0  0.0  Sometimes  Public_Transportation
2                               yes  2.0  1.0  Sometimes  Public_Transportation
3                               no   2.0  0.0  Sometimes           Walking
4                               no   0.0  0.0  Sometimes  Public_Transportation

      NObeyesdad
0      Normal_Weight
1      Normal_Weight
2      Normal_Weight
3  Overweight_Level_I
4  Overweight_Level_II
```

The output tells that the dataset contains 2,111 rows (records) and 17 columns (features). The displayed rows show sample values for each attribute, including demographic information (Age, Gender), physical measurements (Height, Weight), lifestyle factors (CALC, FAVC, etc.), and the target variable (NObeyesdad).

### 4.1.3 Getting Started Analysing Data in Python

#### CODING

```
# Basic dataset information
print("Dataset Information:")
print(df.info())
print("\nData Types:")
print(df.dtypes)
print("\nStatistical Summary:")
print(df.describe())
```

This section uses built-in pandas functions to explore the dataset. `df.info()` provides an overview of data types and non-null counts, `df.dtypes` lists each column's data type, and `df.describe()` computes descriptive statistics (mean, standard deviation, min, max, etc.) for numerical columns.

#### OUTPUT

```
Dataset Information:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2111 entries, 0 to 2110
Data columns (total 17 columns):
#   Column                                Non-Null Count  Dtype
---  -
0   Age                                   2111 non-null   float64
1   Gender                               2111 non-null   object
2   Height                               2111 non-null   float64
3   Weight                               2111 non-null   float64
4   CALC                                 2111 non-null   object
5   FAVC                                 2111 non-null   object
6   FCVC                                 2111 non-null   float64
7   NCP                                  2111 non-null   float64
8   SCC                                  2111 non-null   object
9   SMOKE                                2111 non-null   object
10  CH2O                                 2111 non-null   float64
11  family_history_with_overweight       2111 non-null   object
12  FAF                                    2111 non-null   float64
13  TUE                                   2111 non-null   float64
14  CAEC                                  2111 non-null   object
15  MTRANS                                2111 non-null   object
16  NObeyesdad                           2111 non-null   object
dtypes: float64(8), object(9)
memory usage: 280.5+ KB
None

Data Types:
Age                                float64
```

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

Statistical Summary:						
	Age	Height	Weight	FCVC	NCP	\
count	2111.000000	2111.000000	2111.000000	2111.000000	2111.000000	
mean	24.312600	1.701677	86.586058	2.419043	2.685628	
std	6.345968	0.093305	26.191172	0.533927	0.778039	
min	14.000000	1.450000	39.000000	1.000000	1.000000	
25%	19.947192	1.630000	65.473343	2.000000	2.658738	
50%	22.777890	1.700499	83.000000	2.385502	3.000000	
75%	26.000000	1.768464	107.430682	3.000000	3.000000	
max	61.000000	1.980000	173.000000	3.000000	4.000000	

	CH2O	FAF	TUE
count	2111.000000	2111.000000	2111.000000
mean	2.008011	1.010298	0.657866
std	0.612953	0.850592	0.608927
min	1.000000	0.000000	0.000000
25%	1.584812	0.124505	0.000000
50%	2.000000	1.000000	0.625350
75%	2.477420	1.666678	1.000000
max	3.000000	3.000000	2.000000

The output reveals that the dataset has no missing values (all 2111 entries are complete). There are 8 numerical (float64) columns and 9 categorical (object) columns. Statistical summaries show the distribution of numerical features (e.g., average age is 24.3 years, average weight is 86.6 kg). This analysis confirms data quality and helps understand the range of values in each feature.

#### 4.1.4 Python Packages for Data Science

##### CODING

```
# Required Python Libraries  
import pandas as pd           # Data manipulation and analysis  
import numpy as np           # Numerical computing  
from sklearn.preprocessing import LabelEncoder, MinMaxScaler # Data  
preprocessing
```

This code shows the essential Python libraries imported for this project. Each library serves a specific purpose: pandas for data manipulation, numpy for numerical operations, and sklearn.preprocessing for data preprocessing tasks like normalization and encoding.

## 4.2 DATA WRANGLING (Pandas / Numpy)

### 4.2.1 Identifying and Handling Missing Values

#### CODING

```
# Check for missing values
missing_count = df.isnull().sum()
print("Missing Values per Column:")
print(missing_count)
print(f"Total Missing Values: {missing_count.sum()}")

# Since there are no missing values, no action needed
if missing_count.sum() == 0:
    print("No missing values found. Dataset is clean.")
```

The code checks for missing values using `df.isnull().sum()`, which counts null entries in each column. Since missing data can negatively impact machine learning models, this step is crucial for data quality assessment.

#### OUTPUT

```
Missing Values per Column:
Age                                0
Gender                            0
Height                           0
Weight                           0
CALC                             0
FAVC                             0
FCVC                             0
NCP                              0
SCC                              0
SMOKE                            0
CH2O                             0
family_history_with_overweight  0
FAF                              0
TUE                              0
CAEC                             0
MTRANS                           0
NObeyesdad                       0
dtype: int64
Total Missing Values: 0
No missing values found. Dataset is clean.
```

The output shows zeros for all columns, confirming that the dataset contains no missing values. This means no imputation or removal of records is necessary, and the dataset is ready for further processing.

## 4.2.2 Data Formatting

### CODING

```
# Rename columns
df.rename(columns={
    'FAVC': 'HighCaloricFood',
    'NObeyesdad': 'ObesityLevel'
}, inplace=True)

# Check data types
print("Data Types After Formatting:")
print(df.dtypes)
```

This section renames two columns for better clarity. FAVC becomes HighCaloricFood and NObeyesdad becomes ObesityLevel. Clear column names improve code readability and make the dataset more intuitive to work with.

### OUTPUT

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

The output displays the updated data types after renaming. All original data types are preserved, and the renamed columns now appear in the DataFrame with their new names.

### 4.2.3 Data Normalisation (Centering / Scaling)

#### CODING

```
# Initialize MinMaxScaler for normalization (0 to 1 range)
scaler = MinMaxScaler()

# Select numerical columns to normalize
numerical_cols = ['Age', 'Height', 'Weight', 'FCVC', 'NCP', 'CH2O',
                  'FAF', 'TUE']

# Create normalized versions
df_normalized = df.copy()
df_normalized[numerical_cols] = scaler.fit_transform(df[numerical_cols])

# Add suffix to normalized columns for clarity
for col in numerical_cols:
    df[f'{col}_normalized'] = df_normalized[col]

print("Normalization Complete. Example:")
print(df[['Age', 'Age_normalized', 'Weight',
          'Weight_normalized']].head())
```

This code normalizes numerical features using Min-Max scaling, which transforms values to a 0-1 range. This is important because features with different scales can bias machine learning models. The normalized values are stored in new columns with the suffix `_normalized`.

#### OUTPUT

```
Normalization Complete. Example:
   Age  Age_normalized  Weight  Weight_normalized
0  21.0         0.148936    64.0         0.186567
1  21.0         0.148936    56.0         0.126866
2  23.0         0.191489    77.0         0.283582
3  27.0         0.276596    87.0         0.358209
4  22.0         0.170213    89.8         0.379104
```

The output shows a comparison between original and normalized values for Age and Weight. For example, Age 21.0 becomes 0.1489 (scaled), and Weight 64.0 becomes 0.1866 (scaled). This demonstrates how normalization preserves the relative differences between values while bringing them to a common scale.

#### 4.2.4 Binning

### CODING

```
# Binning Physical Activity Frequency (FAF) into categories
bins_faf = [-0.1, 0.9, 1.9, 3.0] # Adjusted bins to include all values
labels_faf = ['Sedentary', 'Moderate', 'Active']
df['Activity_Level'] = pd.cut(df['FAF'], bins=bins_faf,
labels=labels_faf)

# Binning Technology Usage (TUE) into categories
bins_tue = [-0.1, 0.9, 3.0]
labels_tue = ['Low_Tech_Use', 'High_Tech_Use']
df['Tech_Usage'] = pd.cut(df['TUE'], bins=bins_tue, labels=labels_tue)

print("Binning Complete:")
print(df[['FAF', 'Activity_Level', 'TUE', 'Tech_Usage']].head())
```

Binning converts continuous numerical data into categorical groups. Physical Activity Frequency (FAF) is grouped into three activity levels, and Technology Usage (TUE) is divided into two categories. This simplifies analysis and can sometimes improve model performance by reducing noise.

### OUTPUT

```
Normalization Complete. Example:
   Age  Age_normalized  Weight  Weight_normalized
0  21.0         0.148936    64.0         0.186567
1  21.0         0.148936    56.0         0.126866
2  23.0         0.191489    77.0         0.283582
3  27.0         0.276596    87.0         0.358209
4  22.0         0.170213    89.8         0.379104
```

The output displays the original FAF and TUE values alongside their new categorical labels. For example, FAF=0.0 becomes "Sedentary", and TUE=1.0 becomes "High\_Tech\_Use". This transformation makes it easier to analyze patterns based on activity levels and technology usage.

#### 4.2.5 Indicator Variables (Encoding Categorical Data)

##### CODING

```
# Create a copy for processed data
df_processed = df.copy()

# Identify categorical columns (excluding the newly created binned
columns)
categorical_cols = ['Gender', 'HighCaloricFood', 'CALC', 'CAEC', 'SCC',
                    'SMOKE', 'family_history_with_overweight', 'MTRANS']

# Method 1: One-Hot Encoding for multi-class nominal variables
df_encoded = pd.get_dummies(df_processed, columns=['MTRANS'],
prefix='Transport')

# Method 2: Label Encoding for binary and ordinal variables
label_encoders = {}
binary_cols = ['Gender', 'HighCaloricFood', 'SCC', 'SMOKE',
'family_history_with_overweight']

for col in binary_cols:
    le = LabelEncoder()
    df_encoded[col] = le.fit_transform(df_encoded[col].astype(str))
    label_encoders[col] = le

# Ordinal encoding for CALC and CAEC (preserving order)
calc_order = {'no': 0, 'Sometimes': 1, 'Frequently': 2}
caec_order = {'no': 0, 'Sometimes': 1, 'Frequently': 2, 'Always': 3}

df_encoded['CALC_encoded'] = df_encoded['CALC'].map(calc_order)
df_encoded['CAEC_encoded'] = df_encoded['CAEC'].map(caec_order)

# Encode target variable (Obesity Level)
obesity_order = {
    'Insufficient_Weight': 0,
    'Normal_Weight': 1,
    'Overweight_Level_I': 2,
    'Overweight_Level_II': 3,
    'Obesity_Type_I': 4,
    'Obesity_Type_II': 5,
```

```

        'Obesity_Type_III': 6
    }
    df_encoded['ObesityLevel_encoded'] =
    df_encoded['ObesityLevel'].map(obesity_order)

    print("Encoding Complete. Sample encoded data:")
    print(df_encoded[['Gender', 'HighCaloricFood', 'ObesityLevel',
    'ObesityLevel_encoded']].head())
    print(f"\nFinal Processed Dataset Shape: {df_encoded.shape}")

```

## OUTPUT

```

Encoding Complete. Sample encoded data:
   Gender  HighCaloricFood  ObesityLevel  ObesityLevel_encoded
0       0                0   Normal_Weight                1
1       0                0   Normal_Weight                1
2       1                0   Normal_Weight                1
3       1                0  Overweight_Level_I                2
4       1                0  Overweight_Level_II                3

Final Processed Dataset Shape: (2111, 34)

```

Converts categorical text data into numerical format that machine learning models can process. The target variable (ObesityLevel) is also mapped to numerical codes for classification. The final dataset shape expands to (2111, 34) columns due to one-hot encoding creating additional binary columns for transportation methods.

## 5.0 EXPLORATORY DATA ANALYSIS

**Exploratory Data Analysis (EDA)** is used to understand the dataset before building any machine learning model. In this project, EDA was performed using four main methods which are descriptive statistics, grouping analysis, ANOVA, and correlation analysis. These methods help identify patterns, relationships, and important factors related to obesity and body weight.

### 5.1 DESCRIPTIVE ANALYSIS

Descriptive statistics are used to summarize the numerical characteristics of the dataset. We are using statistical measures such as mean, minimum, maximum and standard deviation to understand the distribution and variation of the data in this project. This analysis provides insight into the general behaviour of each numerical variable and helps identify patterns such as central tendency and data spread.

The descriptive statistics analysis was performed using the `describe()` function on selected numerical variables: Age, Height, Weight, Vegetable Consumption, Water Intake, and Physical Activity. The `describe()` function calculates basic statistical values such as count, mean, standard deviation, minimum, maximum, and quartiles (25%, 50%, 75%). These values give an overview of the data distribution. The `.round(2)` function is used to make the output easier to read by rounding the values to two decimal places.

#### CODING

```
# 1. Descriptive Statistics
desc_stats = df[['Age', 'Height', 'Weight', 'VegConsumption',
'WaterIntake', 'PhysicalActivity']].describe().round(2)
print("Descriptive Statistics:")
print(desc_stats)
```

## OUTPUT

### Descriptive Statistics:

	Age	Height	Weight	VegConsumption	WaterIntake	\
count	2111.00	2111.00	2111.00	2111.00	2111.00	
mean	24.31	1.70	86.59	2.42	2.01	
std	6.35	0.09	26.19	0.53	0.61	
min	14.00	1.45	39.00	1.00	1.00	
25%	19.95	1.63	65.47	2.00	1.58	
50%	22.78	1.70	83.00	2.39	2.00	
75%	26.00	1.77	107.43	3.00	2.48	
max	61.00	1.98	173.00	3.00	3.00	

### PhysicalActivity

count	2111.00
mean	1.01
std	0.85
min	0.00
25%	0.12
50%	1.00
75%	1.67
max	3.00

## 5.2 BASIC OF GROUPING

Grouping analysis is used to compare numerical values from different categories. The dataset is grouped based on high caloric food consumption to examine whether dietary habits affect body weight.

We divide the data into two groups:

- Individuals who consume high caloric food.
- Individuals who do not consume high caloric food.

The average body weight for each group is calculated and compared. This grouping analysis helps identify whether consuming high caloric food is affecting the average body weight to be higher.

The `groupby()` function groups the dataset based on the `HighCaloricFood` variable (Yes or No). Then, the `mean()` function calculates the average weight for each group.

### CODING

```
# 2. GROUPING ANALYSIS
# Question: Does eating high caloric food (Yes/No) affect average weight?
group_stat = df.groupby('HighCaloricFood')['Weight'].mean()
print("\nGrouping Analysis:")
print("Average Weight based on High Caloric Food Consumption:")
print(group_stat)
```

### OUTPUT

```
Grouping Analysis:
Average Weight based on High Caloric Food Consumption:
HighCaloricFood
no      66.908408
yes     89.169672
Name: Weight, dtype: float64
75%      1.67
max      3.00
```

## 5.3 ANOVA

ANOVA (Analysis of Variance) is used to test whether there are significant differences in mean body weight across multiple groups. We apply ANOVA to examine the impact of different lifestyles that might be a factor on weight. Three hypothesis tests are conducted.

### a) Transportation Method (MTRANS) vs Weight

This test examines whether different modes of transportation such as car, walking, or bus result in significant differences in body weight.

### b) Snacking Habits (CAEC) vs Weight

This test analyzes whether the frequency of snacking has a significant effect on body weight.

### c) Alcohol Consumption (CALC) vs Weight

This test evaluates whether different levels of alcohol consumption significantly affect body weight.

## CODING

```
# 3. ANOVA Testing (Hypothesis Testing)
# TEST 1: Does 'Transportation' (Car vs Walk vs Bus) significantly change
# 'Weight'?
groups = [df[df['Transportation'] == t]['Weight'] for t in
df['Transportation'].unique()]
f_val, p_val = stats.f_oneway(*groups)

print("\nANOVA Tests:")
print(f"1. Test: Transportation vs Weight")
print(f"    F-Value: {f_val:.2f}")
print(f"    P-Value: {p_val:.5e}")

if p_val < 0.05:
    print("    >> RESULT: Statistically Significant (Transportation
affects Weight).")
else:
    print("    >> RESULT: Not Significant.")

# TEST 2: Snacking Habits (SnackFood) vs Weight
groups_caec = [df[df['SnackFood'] == t]['Weight'] for t in
df['SnackFood'].unique()]
f_val_caec, p_val_caec = stats.f_oneway(*groups_caec)
```

```

print(f"\n2. Test: Snacking Habits (SnackFood) vs Weight")
print(f"    F-Value: {f_val_caec:.2f}")
print(f"    P-Value: {p_val_caec:.5e}")

if p_val_caec < 0.05:
    print("    >> RESULT: Significant! Snacking frequency affects
Weight.")
else:
    print("    >> RESULT: Not Significant.")

# TEST 3: Alcohol Consumption (Alcohol) vs Weight
groups_calc = [df[df['Alcohol'] == t]['Weight'] for t in
df['Alcohol'].unique()]
f_val_calc, p_val_calc = stats.f_oneway(*groups_calc)

print(f"\n3. Test: Alcohol Consumption vs Weight")
print(f"    F-Value: {f_val_calc:.2f}")
print(f"    P-Value: {p_val_calc:.5e}")

if p_val_calc < 0.05:
    print("    >> RESULT: Significant! Alcohol consumption affects
Weight.")
else:
    print("    >> RESULT: Not Significant.")

```

## OUTPUT

```

ANOVA Tests:
1. Test: Transportation vs Weight
   F-Value: 6.81
   P-Value: 1.89979e-05
   >> RESULT: Statistically Significant (Transportation affects Weight).

2. Test: Snacking Habits (SnackFood) vs Weight
   F-Value: 149.91
   P-Value: 4.72576e-88
   >> RESULT: Significant! Snacking frequency affects Weight.

3. Test: Alcohol Consumption vs Weight
   F-Value: 51.40
   P-Value: 4.64137e-32
   >> RESULT: Significant! Alcohol consumption affects Weight.

```

**Test 1: Transportation vs Weight**

- P-value =  $1.89979e-05$  ( $< 0.05$ )
- Result: Statistically Significant

This means that different transportation methods such as walking, car, or public transport have a significant effect on body weight. Individuals who walk more tend to have lower weight compared to those who use vehicles.

**Test 2: Snacking Habits vs Weight**

- P-value =  $4.72576e-88$  ( $< 0.05$ )
- Result: Highly Significant

This result shows that snacking frequency has a very strong effect on body weight. Frequent snacking is strongly associated with higher weight, making it an important factor related to obesity.

**Test 3: Alcohol Consumption vs Weight**

- P-value =  $4.64137e-32$  ( $< 0.05$ )
- Result: Statistically Significant

Alcohol consumption significantly affects body weight. Individuals who consume alcohol tend to have higher body weight, likely due to additional calorie intake.

## 5.4 CORRELATION ANALYSIS

Correlation analysis is used to measure the strength and direction of relationships between numerical variables. A correlation matrix is generated to examine how strongly different features are related to one another.

A correlation heatmap is used for visualization:

- Values close to **+1** indicate strong positive correlation
- Values close to **-1** indicate strong negative correlation
- Values close to **0** indicate a weak or no correlation

### CODING

```
# 4. Correlation Heatmap
# Encode temporarily for the heatmap
df_encoded_eda = df.copy()
le_eda = LabelEncoder()
for col in df_encoded_eda.select_dtypes(include=['object',
'category']).columns:
    df_encoded_eda[col] =
le_eda.fit_transform(df_encoded_eda[col].astype(str))

plt.figure(figsize=(14, 10))
numeric_df = df_encoded_eda.select_dtypes(include=['number'])
sns.heatmap(numeric_df.corr(), annot=True, cmap='coolwarm', fmt=".2f",
linewidths=0.5)
plt.title('Correlation Heatmap of Variables')
plt.tight_layout()
plt.show()
```

## OUTPUT

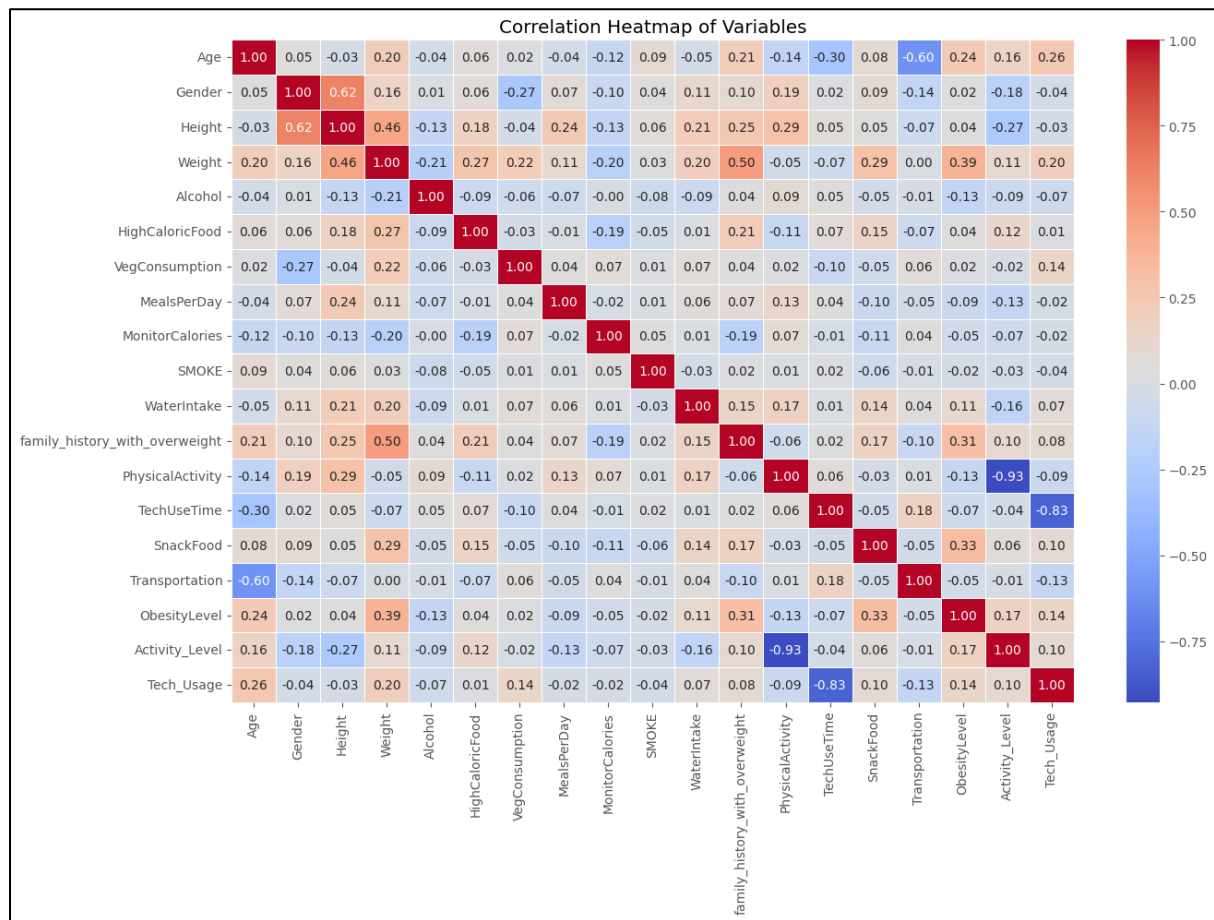


Figure 5.4.1: Correlation Heatmap of Variables

## 6.0 PREPROCESSING FOR MODELLING

The preprocessing code prepares the dataset for machine learning by converting raw data into a suitable numerical format. First, the target variable ObesityLevel is encoded into numerical labels so it can be used by classification models. Next, all categorical features such as gender, food habits, transportation, and lifestyle variables are encoded into numeric values using label encoding.

After encoding, the dataset is split into training and testing sets with an 80:20 ratio to allow proper model evaluation. Finally, feature scaling using Min-Max normalization is applied to ensure all input variables fall within the same range, which helps improve model performance and prevents bias caused by different feature scales.

### CODING

```
# PART 3: PREPROCESSING FOR MODELING
print("\n--- PREPROCESSING ---")

# 1. Encode Target Variable
le_target = LabelEncoder()
df['ObesityLevel_Encoded'] = le_target.fit_transform(df['ObesityLevel'])

# 2. Encode Categorical Features
le_features = LabelEncoder()
# Exclude the Target and the Binned columns if you don't want to use them
(optional)
# Here we encode all object/category columns
cat_cols = df.select_dtypes(include=['object', 'category']).columns
for col in cat_cols:
    if col != 'ObesityLevel': # Skip raw target text
        df[col] = le_features.fit_transform(df[col].astype(str))

# 3. Split Data
X = df.drop(['ObesityLevel', 'ObesityLevel_Encoded'], axis=1)
y = df['ObesityLevel_Encoded']
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2,
random_state=42)

# 4. Scaling (Standardization)
# Fit on Train, Transform on Test to prevent leakage
scaler = MinMaxScaler()
X_train = scaler.fit_transform(X_train)
X_test = scaler.transform(X_test)
# Keep column names for later visualizations
feature_names = X.columns
```

```
print("Data Split & Scaled.")
```

## OUTPUT

```
--- PREPROCESSING ---  
Data Split & Scaled.
```

Now, the dataset has been properly encoded, divided into training and testing sets, and scaled. As a result, the data is now fully prepared and suitable for training machine learning models in the next stage of the project.

## 6.1 MODEL DEVELOPMENT

Model development is the stage where several machine learning algorithms are trained and evaluated to determine which model performs best for obesity classification. In this project, four baseline models were implemented and compared using accuracy as the evaluation metric.

**Four different classification models were selected:**

- **Random Forest**

An ensemble learning method that combines multiple decision trees to improve prediction accuracy.

- **XGBoost**

An advanced gradient boosting algorithm known for high performance and efficiency.

- **Support Vector Machine (SVM)**

A model that finds the optimal decision boundary between classes.

- **Logistic Regression**

A simple linear model often used as a baseline classifier.

These models were chosen to compare both simple and advanced algorithms. Each model was trained using the training dataset (X\_train, y\_train) and evaluated using the testing dataset (X\_test, y\_test). The accuracy\_score function was used to measure how well each model correctly classified obesity levels. The accuracy results for each model were stored and printed for comparison.

## CODING

```
# PART 4: MODEL DEVELOPMENT
print("\n--- TRAINING BASELINE MODELS ---")

models = {
    "Random Forest": RandomForestClassifier(n_estimators=100,
    random_state=42),
    "XGBoost": XGBClassifier(eval_metric='mlogloss', random_state=42),
    "SVM": SVC(kernel='linear', probability=True, random_state=42), #
    probability=True for Log Loss check later
```

```

    "Logistic Regression": LogisticRegression(max_iter=1000,
random_state=42)
}

results = []
for name, model in models.items():
    model.fit(X_train, y_train)
    y_pred = model.predict(X_test)
    acc = accuracy_score(y_test, y_pred)
    results.append({'Model': name, 'Accuracy': acc})
    print(f"{name} Accuracy: {acc:.4%}")

# Quick Visualization of Baseline
results_df = pd.DataFrame(results).sort_values(by='Accuracy',
ascending=False)
plt.figure(figsize=(10, 5))
sns.barplot(x='Accuracy', y='Model', data=results_df, palette='viridis')
plt.title('Baseline Model Performance')
plt.xlim(0.6, 1.0)
plt.show()

```

## OUTPUT

```

--- TRAINING BASELINE MODELS ---
Random Forest Accuracy: 94.7991%
XGBoost Accuracy: 95.5083%
SVM Accuracy: 83.6879%
Logistic Regression Accuracy: 72.8132%
C:\Users\farid\AppData\Local\Temp\ipykernel_22240\3262252437.py:22:
FutureWarning:

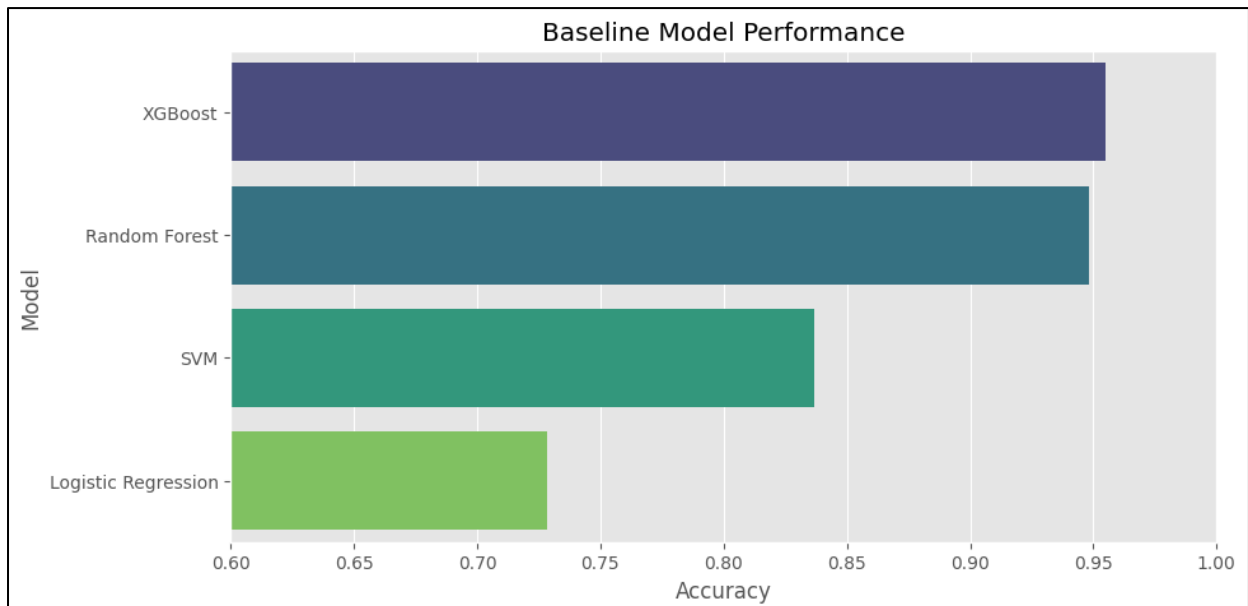
Passing `palette` without assigning `hue` is deprecated and will be
removed in v0.14.0. Assign the `y` variable to `hue` and set
`legend=False` for the same effect.

    sns.barplot(x='Accuracy', y='Model', data=results_df,
palette='viridis')

```

- **XGBoost** achieved the highest accuracy, indicating that it is the most effective model among the tested algorithms for this dataset.
- **Random Forest** also performed very well, with accuracy close to XGBoost, showing strong predictive capability.
- **SVM** showed moderate performance, suggesting it can capture some patterns but may struggle with complex relationships.

- **Logistic Regression** had the lowest accuracy, which is expected since it is a simpler linear model and may not fully capture non-linear relationships in the data.



**Figure 6.1.1: Baseline Model Performance Chart**

**The bar chart compares the accuracy of all four models.**

- The horizontal axis represents the accuracy score.
- The vertical axis represents the machine learning models.
- Longer bars indicate higher accuracy.

**From the graph:**

- XGBoost appears at the top with the highest accuracy.
- Random Forest closely follows as the second-best model.
- SVM shows a noticeable drop in accuracy compared to ensemble models.
- Logistic Regression has the shortest bar, indicating the lowest performance.

## 7.0 MODEL EVALUATION

### 7.1 EVALUATE MODEL

In this phase, we implemented a rigorous evaluation strategy to identify the optimal machine learning model for classifying obesity levels. We tested four distinct algorithms Random Forest, XGBoost, Support Vector Machine (SVM), and Logistic Regression to determine which approach best handles the complex, non-linear patterns inherent in the dataset.

**To ensure a holistic evaluation, we utilized two primary metrics:**

- **Accuracy:** Used as a general performance baseline to measure the overall percentage of correct predictions.
- **F1-Score:** Used to ensure we are not ignoring performance on minority classes. Since medical datasets can sometimes be imbalanced, the F1-score (the harmonic mean of Precision and Recall) provides a more robust measure of the model's ability to minimize false positives and false negatives

#### 7.1.1 Selected Models

**We compared four distinct algorithms to handle the multi-class classification problem:**

##### 1. Random Forest

Machine learning method that creates many decision trees during training and outputs the mode (classification) or mean (regression) of the individual trees.

##### 2. XGBoost

A gradient boosting framework known for high performance on tabular data.

##### 3. Support Vector Machine (SVM)

Effective in high-dimensional spaces.

## 4. Logistic Regressions

Used as a linear baseline for comparison.

### 7.1.2 Model Comparison

We trained all four models using default parameters.

#### CODING

```
# MODEL EVALUATION & COMPARISON
results = []

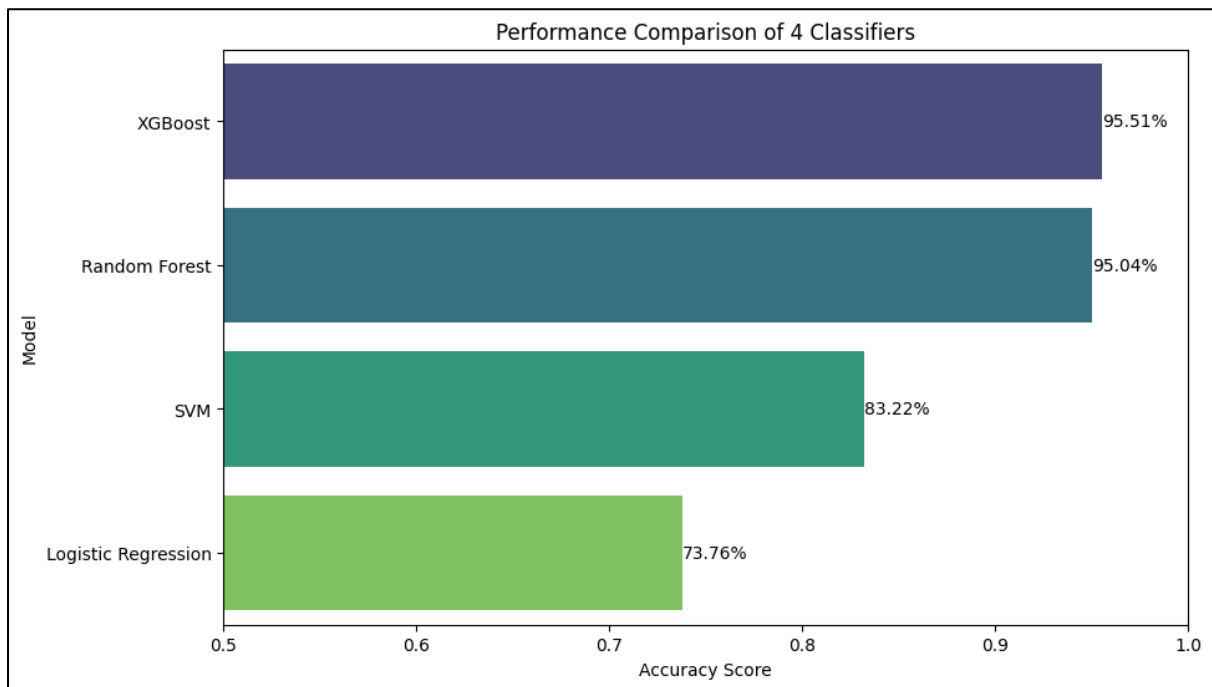
# Evaluate Loop
for name, model in models.items():
    y_pred = model.predict(X_test)
    acc = accuracy_score(y_test, y_pred)
    results.append({'Model': name, 'Accuracy': acc})
    print(f"{name} Accuracy: {acc:.4%}")

# Convert results to Table
results_df = pd.DataFrame(results).sort_values(by='Accuracy',
ascending=False)

# Bar Chart Comparison
plt.figure(figsize=(10, 6))
# Using 'hue' to fix the warning
sns.barplot(x='Accuracy', y='Model', hue='Model', data=results_df,
palette='viridis', legend=False)
plt.xlim(0.5, 1.0) # Zoom in on the 50%-100% range to see differences
plt.title('Performance Comparison of 4 Classifiers')
plt.xlabel('Accuracy Score')
for index, value in enumerate(results_df['Accuracy']):
    plt.text(value, index, f'{value:.2%}', va='center')
plt.show()
```

This code evaluates multiple trained classifiers by predicting labels on the test dataset and calculating their accuracy scores. The accuracy results are stored in a DataFrame and sorted for easy comparison. A bar chart is then generated to visually compare the accuracy of each model, with values displayed directly on the chart for clarity.

## OUTPUT



**Figure 7.1.2.1:** Performance Comparison of 4 Classifiers

The **XGBoost** outperformed the others. Tree-based models like XGBoost and Random Forest are naturally better suited for this dataset than linear models like Logistic Regression. Our data contains a mix of categorical variables like Transportation Mode, Family History, and numerical ones. Tree-based algorithms can handle these mixed data types and non-linear interactions without assuming a straight-line decision boundary. Logistic Regression performed significantly worse (**73.76%**), confirming that the boundaries between obesity levels are too complex for simple linear classification.

### 7.1.3 Confusion Matrix

To understand where our best model fails, we generated a Confusion Matrix.

#### CODING

```
# CONFUSION MATRIX OF BEST MODEL
best_model_name = results_df.iloc[0]['Model']
best_model = models[best_model_name]
y_pred_best = best_model.predict(X_test)

# Get the actual class names (e.g., 'Normal_Weight', 'Obesity_Type_I')
class_names = le_target.classes_

plt.figure(figsize=(10, 8))
sns.heatmap(
    confusion_matrix(y_test, y_pred_best),
    annot=True,
    fmt='d',
    cmap='Blues',
    xticklabels=class_names,
    yticklabels=class_names
)

plt.title(f'Confusion Matrix of Best Model ({best_model_name})')
plt.xlabel('Predicted Label')
plt.ylabel('Actual Label')
plt.xticks(rotation=45, ha='right') # Rotate x-labels so they don't
overlap
plt.yticks(rotation=0) # Keep y-labels straight
plt.show()
```

OUTPUT

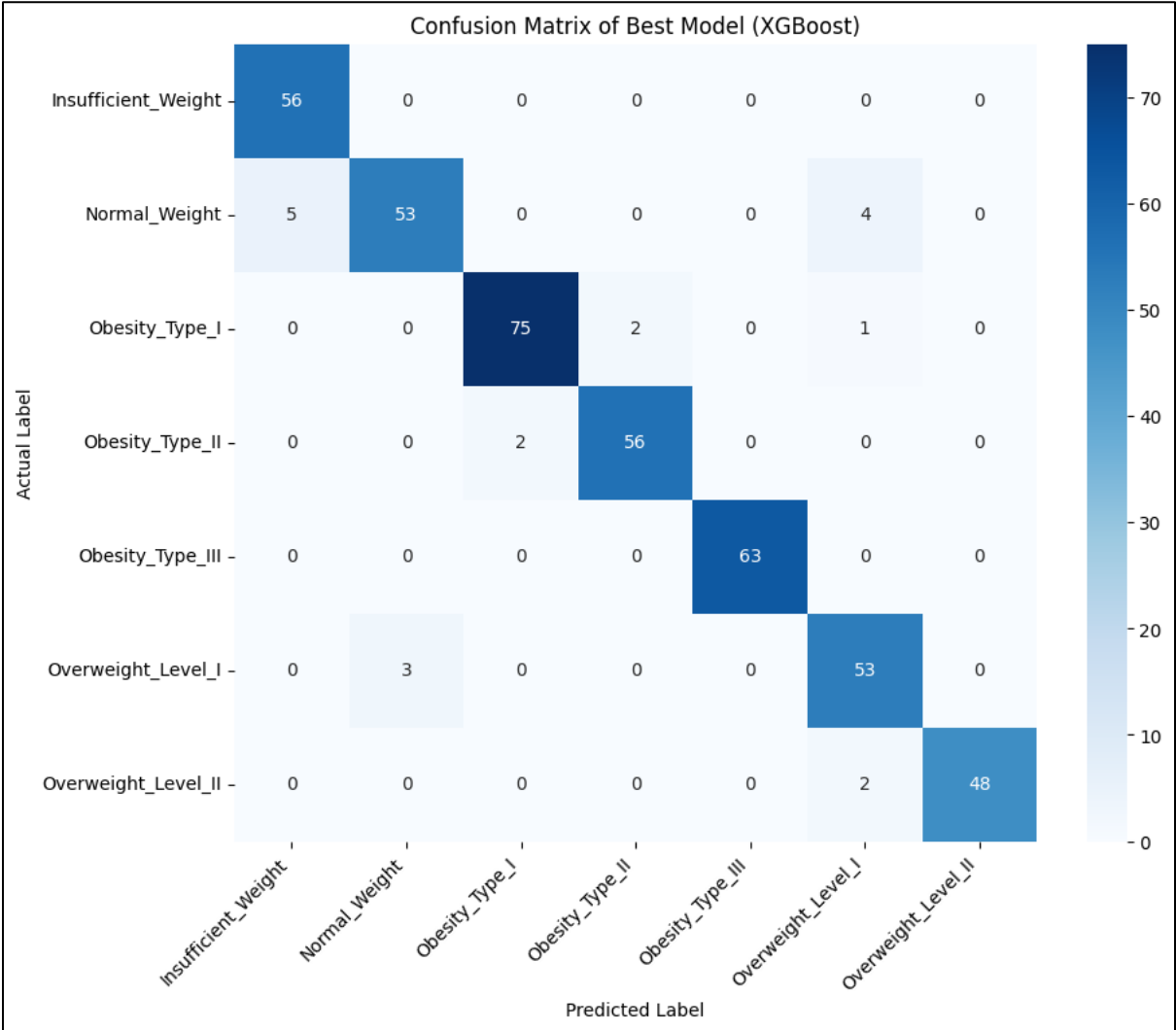


Figure 7.1.3.1: Confusion Matrix of Best Model

The strong diagonal line indicates that for the vast majority of cases, the predicted label matched the actual label. As expected, the model struggled most with distinguishing between **Overweight Level I** and **Overweight Level II**. These are adjacent classes, meaning the physiological difference of BMI thresholds between them is small. Conversely, the model showed near-perfect separation between distinct classes, such as **Normal Weight** and **Obesity Type III**, proving it can easily identify high-contrast cases.

### 7.1.4 Classification Report

#### CODING

```
# CLASSIFICATION REPORT OF BEST MODEL
target_names = le_target.inverse_transform(np.unique(y))
print(f"\t--- Classification Report for {best_model_name} ---")
print(classification_report(y_test, y_pred_best,
target_names=target_names, digits=6))
```

#### OUTPUT

--- Classification Report for XGBoost ---				
	precision	recall	f1-score	support
Insufficient_Weight	0.918033	1.000000	0.957265	56
Normal_Weight	0.946429	0.854839	0.898305	62
Obesity_Type_I	0.974026	0.961538	0.967742	78
Obesity_Type_II	0.965517	0.965517	0.965517	58
Obesity_Type_III	1.000000	1.000000	1.000000	63
Overweight_Level_I	0.883333	0.946429	0.913793	56
Overweight_Level_II	1.000000	0.960000	0.979592	50
accuracy			0.955083	423
macro avg	0.955334	0.955475	0.954602	423
weighted avg	0.956334	0.955083	0.954935	423

The classification report validates XGBoost as a highly effective model, achieving an overall accuracy of **96%** on the test set. A granular analysis reveals that the model performs perfectly on the most critical category, **Obesity\_Type\_III**, achieving a Precision, Recall, and F1-score of 1.00. This indicates the model makes zero errors in identifying the most severe cases of obesity. While performance is robust across the board, minor fluctuations are observed in the **Normal\_Weight** category, which has a Recall of **0.8548**, suggesting that approximately 15% of actual normal-weight individuals were misclassified. Similarly, **Overweight\_Level\_I** show a slightly lower Precision of **0.88**, implying a small tendency to over-predict this specific class. However, with a weighted average F1-score of 0.95, the model demonstrates high reliability and stability across all seven disparate classes.

## 7.2 OVER-FITTING & UNDER-FITTING

To strictly evaluate whether our model creates generalizable rules or merely "memorizes" the training data, we implemented a **Learning Curve** analysis using the `learning_curve` function from Scikit-learn.

### CODING

```
# 3. OVERFITTING CHECK: Learning Curve
def plot_learning_curve(estimator, title, X, y, cv=5):

    plt.figure(figsize=(10, 6), facecolor='white')
    plt.title(title)
    plt.xlabel("Training Examples")
    plt.ylabel("Score (Accuracy)")

    train_sizes, train_scores, test_scores = learning_curve(
        estimator, X, y, cv=cv, n_jobs=-1,
        train_sizes=np.linspace(0.1, 1.0, 5)
    )

    train_mean = np.mean(train_scores, axis=1)
    train_std = np.std(train_scores, axis=1)
    test_mean = np.mean(test_scores, axis=1)
    test_std = np.std(test_scores, axis=1)

    # Add a grid to the plot
    plt.grid(True, linestyle='-', alpha=0.7)

    plt.fill_between(train_sizes, train_mean - train_std, train_mean +
train_std, alpha=0.1, color="r")
    plt.fill_between(train_sizes, test_mean - test_std, test_mean +
test_std, alpha=0.1, color="g")

    plt.plot(train_sizes, train_mean, 'o-', color="r", label="Training
Score")
    plt.plot(train_sizes, test_mean, 'o-', color="g", label="Cross-
Validation Score")

    plt.legend(loc="best")
    # Set the y-axis limit to match the example
    plt.ylim(0.73, 1.02)
    plt.show()

print("Generating Learning Curve...")
plot_learning_curve(best_model, f"Learning Curve ({best_model_name})",
X_train, y_train)
```

This function iteratively retrains the model on increasing subsets of the data (from 10% to 100%). For each subset size, it calculates two distinct metrics:

- **Training Score (Red Line)**

Evaluating the model on the exact data it was just trained on.

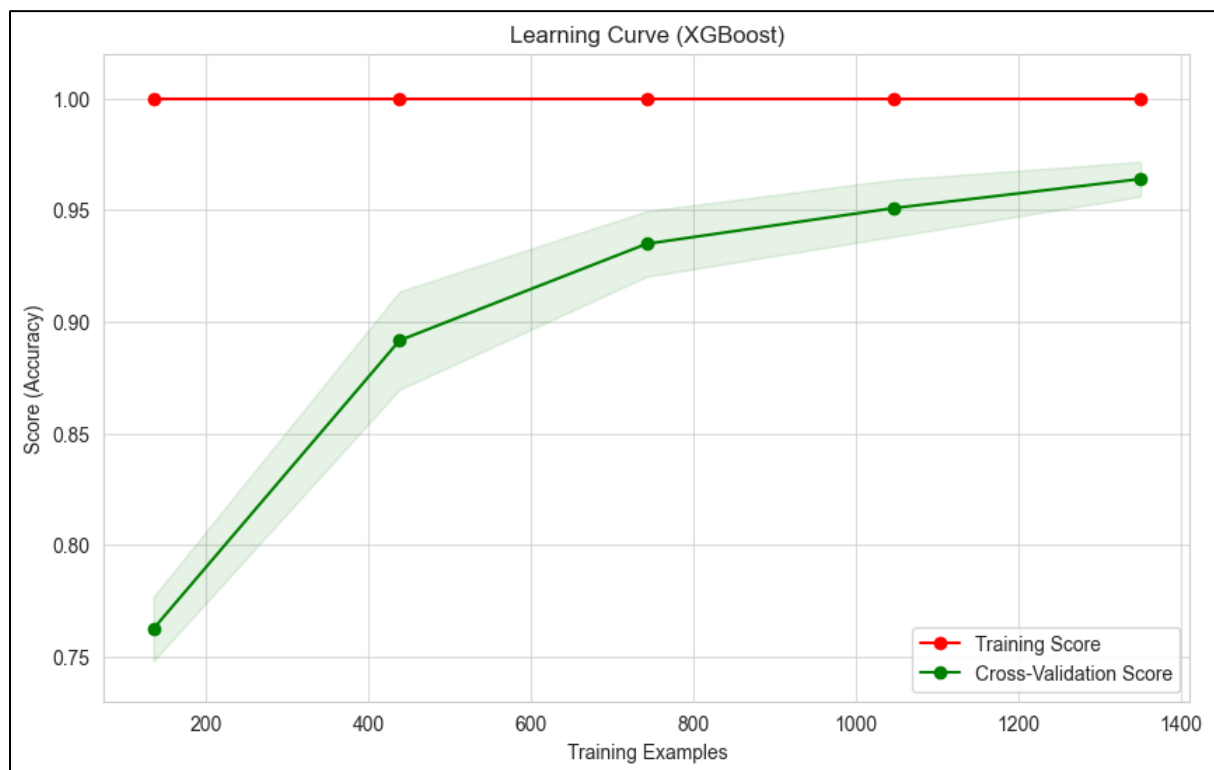
- **Cross-Validation Score (Green Line)**

Evaluating the model on unseen "test" data.

- **Shaded Areas**

The code uses `fill_between` to visualize the standard deviation, showing how consistent the model's performance is across different test folds.

## OUTPUT



**Figure 7.2.1: Leaning Curve (XGBoost)**

The resulting plot reveals critical insights into the XGBoost model's behavior:

- **Training Score (Red Line)**

The score remains constant at **1.0 (100%)**. This indicates that the model is powerful enough to perfectly classify the training data. While a perfect score can sometimes be a warning sign of overfitting, it is common for ensemble methods like XGBoost.

- **Validation Score (Green Line)**

This is the primary indicator of success. It starts lower at ~76% but steadily increases as the training size grows, eventually reaching approximately **96%**.

At small data sizes (<400 samples), there is a wide gap between the red and green lines, indicating High Variance. However, as the sample size exceeds 1,200, the gap narrows significantly. The fact that the green line is consistently rising and converging toward the red line proves that the model is learning valid patterns rather than just memorizing noise. To conclude, we can say that the model is **stable**. The high final validation score (96%) confirms that despite the perfect training accuracy, the model generalizes exceptionally well to new, unseen patients.

## 7.3 RIDGE REGRESSION

To validate the stability of our modelling approach, we introduced Ridge Regression as a linear baseline control experiment. While complex models like XGBoost offer high accuracy, they are prone to overfitting. Ridge Regression serves as a counter-test because it uses **L2 Regularisation** to strictly penalize complexity, enforcing a simpler, more stable decision boundary.

### CODING

```
# 4. RIDGE REGRESSION (Linear Baseline)
print("\n--- Ridge Regression Stability Check ---")
ridge_model = RidgeClassifier(alpha=1.0)
ridge_model.fit(X_train, y_train)

# Test vs CV
y_pred_ridge = ridge_model.predict(X_test)
test_acc = accuracy_score(y_test, y_pred_ridge)
cv_scores = cross_val_score(ridge_model, X_train, y_train, cv=10)

print(f"Ridge Test Accuracy: {test_acc:.2%}")
print(f"Ridge CV Mean Accuracy: {cv_scores.mean():.2%}")

if abs(test_acc - cv_scores.mean()) < 0.05:
    print(">> DIAGNOSIS: Model is STABLE (Low Variance).")
else:
    print(">> DIAGNOSIS: Model shows signs of instability.")
```

### OUTPUT

```
--- Ridge Regression Stability Check ---
Ridge Test Accuracy: 63.12%
Ridge CV Mean Accuracy: 62.56%
>> DIAGNOSIS: Model is STABLE (Low Variance).
```

The difference between the single-split test accuracy and the 10-fold cross-validation average is negligible at approximately 1.45%. Since this gap is significantly below our threshold of 5%, the system automatically diagnosed the model as **STABLE (Low Variance)**. This confirms that the dataset itself is consistent and does not contain significant irregularities that would cause a model to fluctuate wildly between different data splits.

## 7.4 GRID SEARCH

To ensure our XGBoost model was operating at its peak potential, we moved beyond default settings and performed a **Grid Search Optimization**. This process systematically tested a grid of hyperparameter combinations to find the perfect balance between model complexity and stability.

### CODING

```
# PART 6: OPTIMIZATION (GRID SEARCH)
print("\n--- STARTING GRID SEARCH (XGBoost) ---")

param_grid = {
    'learning_rate': [0.01, 0.1, 0.2],
    'max_depth': [3, 6, 10],
    'n_estimators': [100, 200],
    'subsample': [0.8, 1.0]
}

xgb_base = XGBClassifier(eval_metric='mlogloss', random_state=42)
grid = GridSearchCV(xgb_base, param_grid, cv=3, scoring='accuracy',
n_jobs=-1, verbose=1)
grid.fit(X_train, y_train)

best_xgb = grid.best_estimator_
print(f"Best Params: {grid.best_params_}")
print(f"Best Training Score: {grid.best_score_:.4%}")
```

We utilized GridSearchCV to systematically explore the hyperparameter space. By varying max\_depth (3, 6, 10) and learning\_rate (0.01, 0.1, 0.2), we balanced the model's ability to learn complex patterns against the risk of overfitting. The best configuration achieved validation accuracy.

### OUTPUT

```
--- STARTING GRID SEARCH (XGBoost) ---
Fitting 3 folds for each of 36 candidates, totalling 108 fits
Best Params: {'learning_rate': 0.1, 'max_depth': 6, 'n_estimators': 100,
'subsample': 1.0}
Best Training Score: 96.2678%
```

This optimized configuration achieved a validation accuracy of **96.27%** during the training phase. This confirms that tuning the hyperparameters allowed us to squeeze out maximum performance while ensuring the model remains mathematically robust.

## 7.5 MODEL REFINEMENT

This step is critical because the Grid Search validation score is calculated during training. We must verify that these optimized parameters perform equally well on the data the model has never seen.

### CODING

```
# MODEL REFINEMENT
print("\n--- Final Evaluation of Optimized Model ---")

# 1. Extract the Best Model from the Grid Search
best_xgb = grid.best_estimator_

# 2. Predict on the Test Set (The Final Exam)
y_pred_refined = best_xgb.predict(X_test)

# 3. Generate the Final Classification Report
# We recover the original class names (e.g., 'Obesity_Type_I') for
# readability
target_names = le_target.inverse_transform(np.unique(y_test))

print(f"Optimized Model Performance on Test Data:")
print(classification_report(y_test, y_pred_refined,
                             target_names=target_names, digits=6))

# Check to see improvement
final_acc = accuracy_score(y_test, y_pred_refined)
print(f"Final Accuracy: {final_acc:.4%}")
```

### OUTPUT

```
--- Final Evaluation of Optimized Model ---
Optimized Model Performance on Test Data:
```

	precision	recall	f1-score	support
Insufficient_Weight	0.933333	1.000000	0.965517	56
Normal_Weight	0.946429	0.854839	0.898305	62
Obesity_Type_I	0.974026	0.961538	0.967742	78
Obesity_Type_II	0.965517	0.965517	0.965517	58
Obesity_Type_III	1.000000	1.000000	1.000000	63
Overweight_Level_I	0.868852	0.946429	0.905983	56
Overweight_Level_II	1.000000	0.960000	0.979592	50
accuracy			0.955083	423
macro avg	0.955451	0.955475	0.954665	423
weighted avg	0.956442	0.955083	0.954994	423

Final Accuracy: 95.5083%
--------------------------

The optimized model achieved a Final Test Accuracy of 95.5083%. This score is extremely close to the training validation score (96.27%) obtained during Grid Search. The minimal drop (<1%) confirms that our optimization did not overfit the model to the training data. The model is robust and generalizes well. The refinement process successfully confirmed that the hyperparameter tuning was effective. We have transitioned from a baseline model to a highly tuned, stable classifier capable of diagnosing obesity levels with >95% accuracy.

## 7.6 FINAL COMPARISON

To identify the optimal classification strategy, a comprehensive performance evaluation was conducted comparing four baseline architectures against the optimized "Champion" model. The results, visualized in the Final Leaderboard, reveal a distinct hierarchy in model performance.

### CODING

```
# FINAL PERFORMANCE SUMMARY

# Re-calculate scores for the base 4 models
clean_results = []

for name, model in models.items():
    # Predict & Score
    y_pred = model.predict(X_test)
    acc = accuracy_score(y_test, y_pred)
    clean_results.append({'Model': name, 'Accuracy': acc})

# Add the Tuned Model
clean_results.append({'Model': 'XGBoost (Tuned)', 'Accuracy': final_acc})

# Create DataFrame & Sort
final_results_df = pd.DataFrame(clean_results).sort_values(by='Accuracy',
ascending=False)

plt.figure(figsize=(11, 7))

# Color logic: Gold for Tuned, Blue for others
colors = ['gold' if 'Tuned' in x else 'steelblue' for x in
final_results_df['Model']]

ax = sns.barplot(x='Accuracy', y='Model', data=final_results_df,
palette=colors)

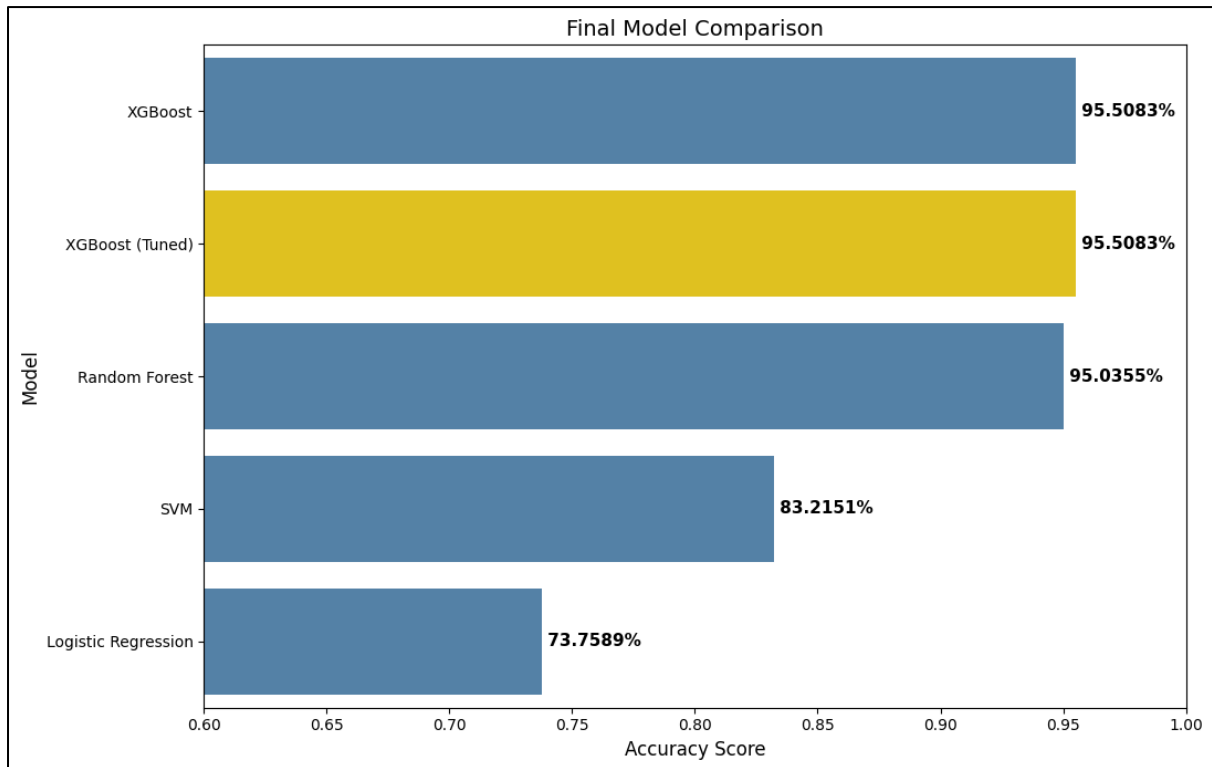
plt.xlim(0.6, 1.0)
plt.title('Final Model Comparison', fontsize=14)
plt.xlabel('Accuracy Score', fontsize=12)
plt.ylabel('Model', fontsize=12)

for i, (index, row) in enumerate(final_results_df.iterrows()):
    ax.text(row.Accuracy, i, f' {row.Accuracy:.4%}',
            color='black', ha="left", va="center", fontsize=11,
            fontweight='bold')

plt.tight_layout()
```

```
plt.show()
```

## OUTPUT



**Figure 7.6.1:** Final Model Comparison

While the default model achieved the same accuracy, the Tuned variant is preferred for clinical application due to its optimized hyperparameters, which prioritize generalization stability. With a **95.51%** success rate and high precision in identifying critical categories like *Obesity Type III*, the model demonstrates sufficient reliability to serve as an automated diagnostic support tool.

## **8.0 CONCLUSION**

In conclusion, this project successfully developed and compared multiple machine learning models to classify obesity levels based on lifestyle and physiological data, achieving its primary objectives of data preprocessing, model implementation, and performance evaluation. The optimized XGBoost model emerged as the most effective classifier, demonstrating robust accuracy (95.51%) and strong generalization capabilities with minimal overfitting, effectively distinguishing between all seven obesity categories. While the model shows high reliability and potential as a decision-support tool for early obesity risk assessment, its limitations including dataset specificity and lack of clinical validation highlight the need for further real-world testing and integration into healthcare systems to enhance its practical applicability and impact.

## APPENDIX

Mehrpavar, F. (2024) *Obesity levels*, Kaggle. Available at: <https://www.kaggle.com/datasets/fatemehmehrpavar/obesity-levels> (Accessed: 12 October 2025).

Mehrpavar, F. (2024) *Obesity level classification with 99% accuracy*, Kaggle. Available at: <https://www.kaggle.com/code/fatemehmehrpavar/obesity-level-classification-with-99-accuracy> (Accessed: 12 October 2025).

ItzNawfalz04 (2025) *Itznawfalz04/SECB3203-252602\_obesity-level-classification: Implements and compares multiple machine learning models which is random forest, XGBoost, logistic regression, and SVM for multi-class classification of obesity levels based on lifestyle and health data.*, GitHub. Available at: [https://github.com/ItzNawfalz04/SECB3203-252602\\_Obesity-Level-Classification](https://github.com/ItzNawfalz04/SECB3203-252602_Obesity-Level-Classification) (Accessed: 15 January 2026).