UNIVERSITI TEKNOLOGI MALAYSIA

FACULTY OF COMPUTING

SEMESTER 1, SESSION 2025/2026

PROJECT PROGRESS 3

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 SHAIFUDDIN	A25CS0109

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

GROUP **:** GROUP 07

TABLE OF CONTENTS

1.0 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.

2.0 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
```

3.0 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
```

4.0 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

```python
# 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.

5.0 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:

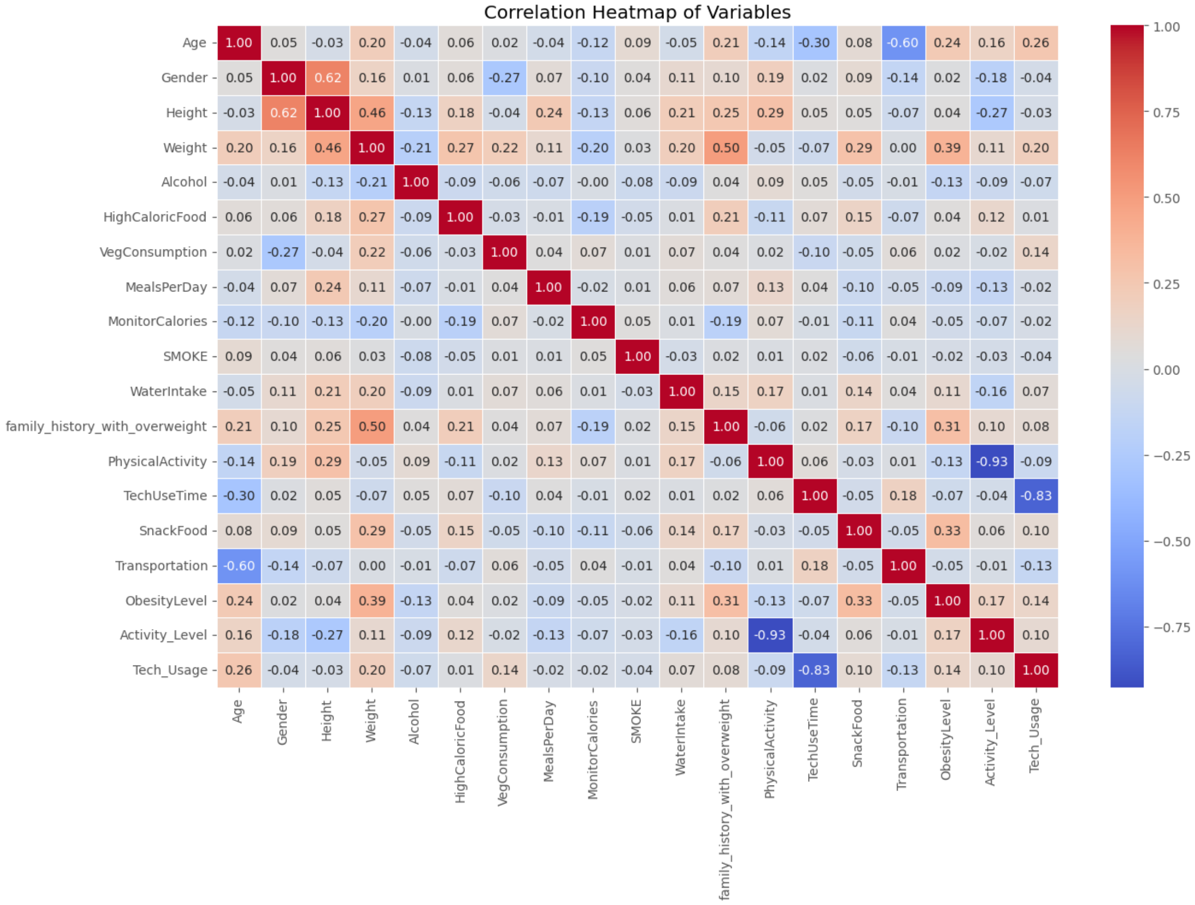
- Values close to **+1** indicate strong positive correlation
- Values close to **−1** indicate strong negative correlation
- Values close to **0** indicate a weak or no correlation

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



	Age	Gender	Height	Weight	Alcohol	HighCaloricFood	VegConsumption	MealsPerDay	MonitorCalories	SMOKE	WaterIntake	family_history_with_overweight	PhysicalActivity	TechUseTime	SnackFood	Transportation	ObesityLevel	Activity_Level	Tech_Usage
Age	1.00	0.05	-0.03	0.20	-0.04	0.06	0.02	-0.04	-0.12	0.09	-0.05	0.21	-0.14	-0.30	0.08	-0.60	0.24	0.16	0.26
Gender	0.05	1.00	0.62	0.16	0.01	0.06	-0.27	0.07	-0.10	0.04	0.11	0.10	0.19	0.02	0.09	-0.14	0.02	-0.18	-0.04
Height	-0.03	0.62	1.00	0.46	-0.13	0.18	-0.04	0.24	-0.13	0.06	0.21	0.25	0.29	0.05	0.05	-0.07	0.04	-0.27	-0.03
Weight	0.20	0.16	0.46	1.00	-0.21	0.27	0.22	0.11	-0.20	0.03	0.20	0.50	-0.05	-0.07	0.29	0.00	0.39	0.11	0.20
Alcohol	-0.04	0.01	-0.13	-0.21	1.00	-0.09	-0.06	-0.07	-0.00	-0.08	-0.09	0.04	0.09	0.05	-0.05	-0.01	-0.13	-0.09	-0.07
HighCaloricFood	0.06	0.06	0.18	0.27	-0.09	1.00	-0.03	-0.01	-0.19	-0.05	0.01	0.21	-0.11	0.07	0.15	-0.07	0.04	0.12	0.01
VegConsumption	0.02	-0.27	-0.04	0.22	-0.06	-0.03	1.00	0.04	0.07	0.01	0.07	0.04	0.02	-0.10	-0.05	0.06	0.02	-0.02	0.14
MealsPerDay	-0.04	0.07	0.24	0.11	-0.07	-0.01	0.04	1.00	-0.02	0.01	0.06	0.07	0.13	0.04	-0.10	-0.05	-0.09	-0.13	-0.02
MonitorCalories	-0.12	-0.10	-0.13	-0.20	-0.00	-0.19	0.07	-0.02	1.00	0.05	0.01	-0.19	0.07	-0.01	-0.11	0.04	-0.05	-0.07	-0.02
SMOKE	0.09	0.04	0.06	0.03	-0.08	-0.05	0.01	0.01	0.05	1.00	-0.03	0.02	0.01	0.02	-0.06	-0.01	-0.02	-0.03	-0.04
WaterIntake	-0.05	0.11	0.21	0.20	-0.09	0.01	0.07	0.06	0.01	-0.03	1.00	0.15	0.17	0.01	0.14	0.04	0.11	-0.16	0.07
family_history_with_overweight	0.21	0.10	0.25	0.50	0.04	0.21	0.04	0.07	-0.19	0.02	0.15	1.00	-0.06	0.02	0.17	-0.10	0.31	0.10	0.08
PhysicalActivity	-0.14	0.19	0.29	-0.05	0.09	-0.11	0.02	0.13	0.07	0.01	0.17	-0.06	1.00	0.06	-0.03	0.01	-0.13	-0.93	-0.09
TechUseTime	-0.30	0.02	0.05	-0.07	0.05	0.07	-0.10	0.04	-0.01	0.02	0.01	0.02	0.06	1.00	-0.05	0.18	-0.07	-0.04	-0.83
SnackFood	0.08	0.09	0.05	0.29	-0.05	0.15	-0.05	-0.10	-0.11	-0.06	0.14	0.17	-0.03	-0.05	1.00	-0.05	0.33	0.06	0.10
Transportation	-0.60	-0.14	-0.07	0.00	-0.01	-0.07	0.06	-0.05	0.04	-0.01	0.04	-0.10	0.01	0.18	-0.05	1.00	-0.05	-0.01	-0.13
ObesityLevel	0.24	0.02	0.04	0.39	-0.13	0.04	0.02	-0.09	-0.05	-0.02	0.11	0.31	-0.13	-0.07	0.33	-0.05	1.00	0.17	0.14
Activity_Level	0.16	-0.18	-0.27	0.11	-0.09	0.12	-0.02	-0.13	-0.07	-0.03	-0.16	0.10	-0.93	-0.04	0.06	-0.01	0.17	1.00	0.10
Tech_Usage	0.26	-0.04	-0.03	0.20	-0.07	0.01	0.14	-0.02	-0.02	-0.04	0.07	0.08	-0.09	-0.83	0.10	-0.13	0.14	0.10	1.00

Figure 5.1: Correlation Heatmap of Variables