MAKERERE UNIVERSITY

SEMESTER ONE 2024/2025 ACADEMIC YEAR

SCHOOL COMPUTING AND INFORMATICS TECHNOLOGY

DEPARTMENT OF COMPUTER SCIENCE

MASTER OF SCIENCE IN COMPUTER SCIENCE

MCS 7103

MACHINE LEARNING

ASSIGNMENT ONE

KIZZA NAUME NABANJALA

2024/HD05/21931U

2400721931

OBESITY RISK PREDICTION

EXPLORATORY DATA ANALYSIS

Introduction

After carrying out a deep exploratory data analysis on the Obesity risk Prediction dataset. I found out the following.

Obesity, which causes physical and mental problems, is a global health problem with serious consequences. Because of this, new research is needed that examines the factors that influence obesity and how to predict the condition's occurrence according to a few notable factors.

Question:

- **How much impact do eating habits and physical conditions have on obesity?**
  - To answer this question, I looked at the dataset that I chose to give me some quantitative data on each factor and their relation to obesity and potentially predict the level each factor adds to obesity.

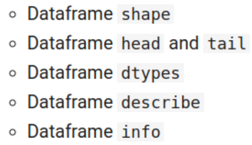
Data selection

- **Why did I choose this dataset and what is the purpose of this analysis?**
  - I chose to use the Obesity Risk Prediction dataset to figure out what factors influence obesity and at what level which I can use to predict its occurrence and potentially its levels. Its purpose is to Estimate Obesity Levels Based On Eating Habits and Physical Condition

Dataset

- **About the dataset**
  - The dataset includes data for the estimation of obesity levels in individuals from the countries of Mexico, Peru, and Colombia, based on their eating habits and physical condition.
- **Extracting the data**
  - For our dataset, I extracted it to be able to use it in my notebook for any analysis that was to be done. To do so, I installed some packages that helped me utilize some python functions for example `pd.read_csv()` that I then used to retrieve the data:
- **Previewing the data**

  - To show the contents of the data, I needed to call the `df` function so that the data is displayed graphically.
  - With that, I was able to show a tabular format of the dataset and show its rows and columns with the values
  - From the previewing of the data, I established that the dataset contains 17 attributes and 2111 records, the records are labeled with the class variable NObeyesdad (Obesity Level), which allows classification of the data using the values of Insufficient Weight, Normal Weight, Overweight Level I, Overweight Level II, Obesity Type I, Obesity Type II and Obesity Type III.
- **Understanding the data**
  - To understand the data in depth, I displayed a detailed description of each column in the dataset explaining what each column and their type of data respectively
  - I then differentiated the data, separating the categorical data from the numerical data by using the `df.select_dtypes()` function and adding the results to the corresponding list i.e. objects in their list and numbers in their list so that I could further analyze with cleaner data.
  - I continued to perform a set of queries on the data to get a more in-depth understanding of its structure. These queries I performed were:



    - Dataframe `shape`
    - Dataframe `head` and `tail`
    - Dataframe `dtypes`
    - Dataframe `describe`
    - Dataframe `info`

  - I used `df.shape` to show the total rows and total columns
  - I used `df.head` to return the first five records
  - I used `df.tail` to view the last five records
  - I used `df.dtypes` to figure out the datatypes of each column
  - I used `df.describe` to show the standard calculations for any numerical values
  - I used `df.info` to show the non-null count and datatype for each column
  - From the above queries, I deduced that:
    - There are 2111 entries, i.e. 2111 rows
    - Each row has a row label (which is the index) with values ranging from 0 to 2110
    - The table has 17 columns, all having a value for each of the rows (all 2111 values are non-null)
    - The columns Gender, family_history_with_overweight, FAVC, CAEC, SMOKE, SCC, CALC, MTRANS, and NObeyesdad consist of textual

data (strings which are also known as objects) and the other columns are numerical data with real numbers (also known as float)

        - There are 8 float data types and 9 object data types

Data wrangling

- To perform any kind of analysis and also figure out what factors from my dataset greatly affect obesity, I had to first make sure that my data was clean enough to be used
- I checked if there are any missing values by running the `df.isnull().sum()` command.
- From my findings, I had no missing data as there were no null values, 0 values, or missing values
- Next, I checked if there were any white spaces in the column titles with the `df.columns()` function
- From the above, I established that the columns need no cleaning or removal of white spaces as they are already clean rough to my satisfaction.
- Next, I checked for any unique columns by running the `df.nunique()` function so that I can use them to check for any duplication. With this, I realized that there was no unique column and this would become an issue if I were to check for duplication. So to fix that I had to create an identifier for the dataset.
- Since I didn't want to change anything in the original data, I copied the dataset into a new one by running the `df.copy()` function and assigned the results to `df_clean`. I then used the new dataset to create a new column ID with sequential numbers as identifiers for each column using the `range(1, len(df_clean) + 1)` and moved the newly created column to the first in order of arrangement.
- I then run `df_clean.nunique()` to check again for a unique column and it returned the new created ID column as the unique column
- Next, I checked for any duplicate values from the dataset with `df_clean.duplicated().sum` and I derived that the data has 0 or no duplicated columns that might need cleaning

Exploratory Data Analysis

- In this section, I used the types of EDA to work out the relation between each column and how they directly or indirectly impact obesity levels.
- To begin with, I asked myself a question
    - Question: **What do I want to achieve from the analysis?**

- I would like to explore how different variables, such as physical activity, and calorie intake, are correlated with obesity levels to better understand risk patterns.
- I also would like to carry out the following:
    - I would like to identify the Key Factors Contributing to Obesity i.e. Identify the lifestyle, physical, and health factors (e.g., age, diet, physical activity) that have the most significant impact on obesity risk.
    - I would like to understand the distribution of key variables like age, weight, and physical activity levels to see how they relate to obesity across different populations.
    - I would also want to identify any outliers or data anomalies that could skew the analysis or provide insights into unusual cases of obesity risk.

1. So I began with the **Univariate Analysis** where I plotted:
    - Histograms
    - KDE
    - Boxplot

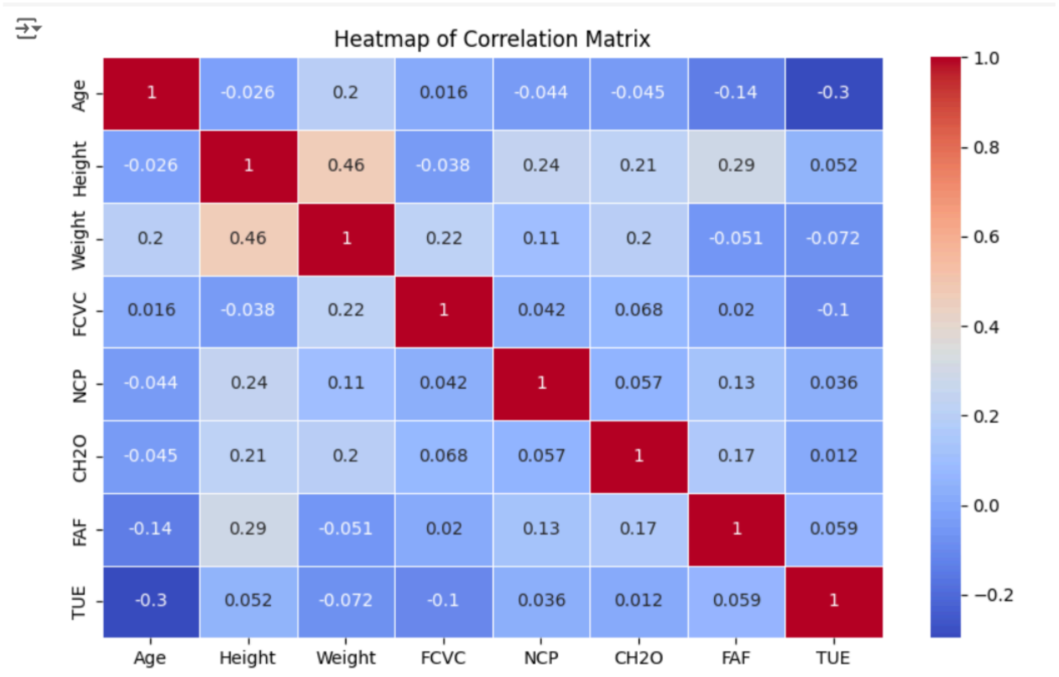
- **Histograms**
    - I plotted histograms for all the individual numerical columns i.e Age, Height, Weight, FCVC, NCP, CH2O, FAF and TUE by looping through the numerical columns list `df_numerical` I had created early on when I separated the categorical data from the numerical data
    - From the above plotting and the histograms, I deduced from the plotted histograms
        - The peak value showed that the highest number of individuals belonged to the age group of 20 from the **Age** histogram
        - The highest number of Individuals recorded were of a height between 1.7 and 1.75 from the **Height** histogram
        - The highest number of individuals who participated recorded a weight of 75 as seen from the **Weight** histogram
        - A large number of participants recorded that they consumed or ate vegetables 3 times a day from the **FCVC** (Vegetable Consumption) histogram
        - The peak value for people who eat 3 main meals a day was recorded as seen from the **NCP** (Number of Main Meals per Day) histogram
        - The **CH20** (Water intake) histogram, showed that many individuals from the dataset take water utmost 2 times a day

        - When it comes to **FAF** (Frequency of physical activity), a large seems to do any physical activity.
        - The same is said for **TUE** (Time Using Technology/Devices) as seen from the TUE histogram

- **Kernel Density Estimation (KDE) graphs**
    - I plotted KDE graphs for all the individual numerical columns by looping through the same numerical columns list `df_numerical`
    - From my findings, I would say i got the same exact results as what the histograms display

- **Box Plots**
    - I also plotted boxplots for all the individual numerical columns i.e Age, Height, Weight, FCVC, NCP, CH2O, FAF and TUE by looping through the same numerical columns list `df_numerical`
    - From the above plotting of boxplots for the individual columns, my findings were as follows:
        - I noticed several outliers in the **Age** and **NCP** (Number of Main Meals per Day) boxplot graphs. Since these two have no direct impact to the level of obesity and are not what am concentrating on to carry out the analysis, I decided to leave them as is

2. Then I went to **Multivariate Analysis** where I plotted the:
    a. Heatmap
    b. scatterplot



- As seen from the above correlation graph, I constructed to help me understand the correlation between the target variable (Obesity Levels) and each of the other features and I noticed the following:
  - **Weight** and **Height**, Correlation = 0.46:
    - There is a strong positive correlation between **Weight** and **Height** (**0.46**). This is understandable from the fact that, generally, taller people tend to weigh more due to a few factors like bone density not to mention but a few from the research I have carried out.
  - **Weight** and **Age**, Correlation = 0.2:
    - From what I see, a moderate positive correlation exists between **Weight** and **Age** (**0.2**), which from my understanding suggests that as people age, their weight tends to increase, though the correlation could be stronger since it does not tend strongly to 1.
  - **Height** and **NCP** (Number of Main Meals per Day), Correlation = 0.24:
    - Because of a noticeable positive correlation (**0.24**) between **Height** and **NCP**, it could suggest that taller individuals tend to consume more meals daily.
  - **Weight** and **FCVC** (Vegetable Consumption), Correlation = 0.22:
    - There is a weak positive correlation (**0.22**) between **Weight** and **FCVC**. This could be due to a variety of factors where individuals with higher

weight may be consuming vegetables but not enough to significantly influence obesity levels or weight.

- **Weight** and **FAF** (Physical Activity Frequency), Correlation = -0.051:
  - Since there is a weak negative correlation (**-0.051**) between **Weight** and **FAF**, this might suggest that increased physical activity might have a small inverse relationship with weight.
- **TUE** (Time Using Technology/Devices) and **Age**, Correlation = -0.3:
  - There is a moderate negative correlation (**-0.3**) between **TUE** (Time Using Technology) and **Age**. This implies that older people spend less time on technological devices compared to younger people.
- **FAF** (Physical Activity Frequency) and **Age**, Correlation = -0.14:
  - A weak negative correlation (**-0.14**) between **FAF** and **Age** might indicate that as people age, they tend to engage in less physical activity.

General Observations:

- Most correlations seem weak or moderate, except for **Height and Weight**, which shows a stronger positive relationship and I considered any feature with a correlation greater than 0.35 to have a strong impact on the target variable.
- **Physical activity (FAF)** has a weak but generally positive relationship with other healthy behaviors like **Water Intake (CH2O)** and **Vegetable Consumption (FCVC)**.