**Obesity Classification Prediction Project Report**

**i. Cover Page**

[Student would add their name and matriculation number]

**ii. Introduction**

The project focuses on developing a machine learning model to predict obesity levels using various health and lifestyle factors. The dataset contains information about individuals' physical characteristics, dietary habits, physical activity, and other relevant health indicators. The primary goal is to create a predictive model that can classify an individual's obesity level based on their personal health profile.

**Overview:**  
This dataset include data for the estimation of obesity levels in individuals from the countries of Mexico, Peru and Colombia, based on their eating habits and physical condition.  
The data contains 17 attributes and 2111 records, the records are labelled with the class variable NObesity (Obesity Level), that 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.

**Data Details:**

* Gender: Gender
* Age: Age
* Height : in metres
* Weight : in kgs
* family\_history : Has a family member suffered or suffers from overweight?
* FAVC : Do you eat high caloric food frequently?
* FCVC : Do you usually eat vegetables in your meals?
* NCP : How many main meals do you have daily?
* CAEC : Do you eat any food between meals?
* SMOKE : Do you smoke?
* CH2O : How much water do you drink daily?
* SCC : Do you monitor the calories you eat daily?
* FAF: How often do you have physical activity?
* TUE : How much time do you use technological devices such as cell phone, videogames, television, computer and others?
* CALC : How often do you drink alcohol?
* MTRANS : Which transportation do you usually use?
* Obesity\_level (Target Column) : Obesity level

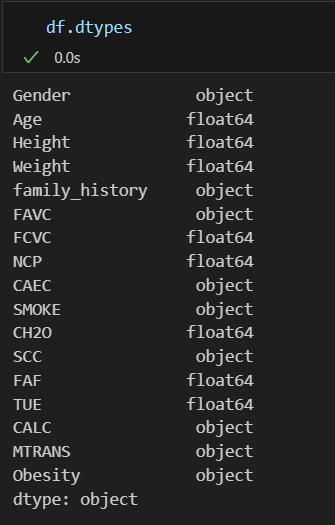
**iii. Data Exploration and Pre-processing**

I loaded the dataset using pandas and used .head(), .describe(), .isnull() to explore the first few records, statistical summaries, and check for missing values.

# Load the data

df = pd.read\_csv('Obesity\_prediction.csv')

df.head()



I then use df.types in order to see the dataset that has to be transformed or encoded.

Here is what I have come up with the plan on changing

Age = round.

gender = binary.

smoke = binary.

family history = binary

FAVC = binary

OBESITY = numeric

CAEC = one hot encoding

SCC = binary

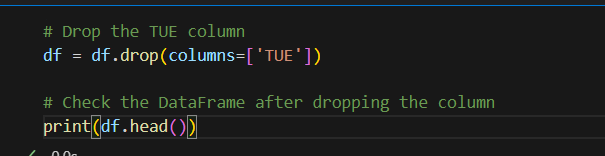
CALC = one hot encoding

MTRANS = one hot encoding

TUE = drop

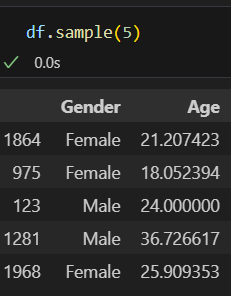
1. TUE (Drop)

The reason why I drop TUE is because this analysis is related to health, fitness and obesity related factors. The main purpose of TUE is about the time spend on electronic device which might not provide valuable information compared to other variables like physical activities, eating habits or family history. Therefore, I have decided to drop TUE.



1. Age (Round)

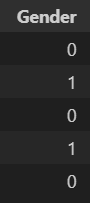
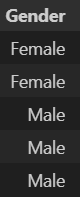
I have noticed from the dataset for the age it has decimals. Therefore, I have use round to make it in whole number.





1. Gender (binary)

I have converted the gender into binary numerical format where Male = 1 and Female = 0.



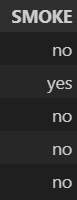
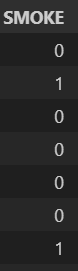
BEFORE

AFTER

1. SMOKE (binary)

I have converted the smoke into binary numerical format where yes = 1 and no = 0.



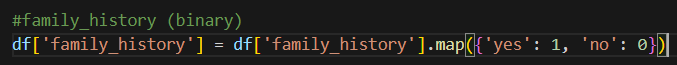


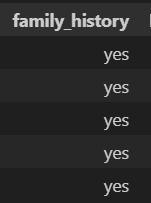
AFTER

BEFORE

1. Family\_history (binary)

I have converted the family\_history into binary numerical format where yes = 1 and no = 0.



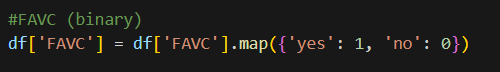


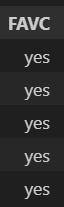
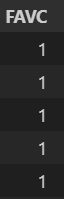
AFTER

BEFORE

1. FAVC (binary)

I have converted the FAVC into binary numerical format where yes = 1 and no = 0.



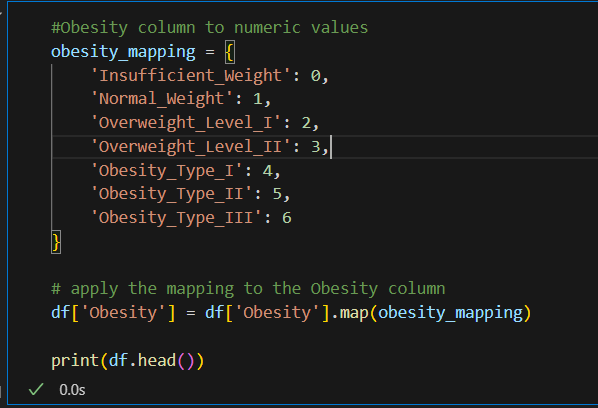


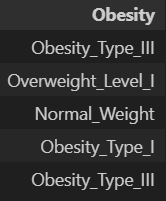
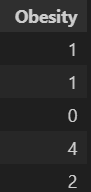
BEFORE

AFTER

1. Obesity (numeric)

I have changed obesity into numeric eg. Insufficient\_weight = 0, normal\_weight = 1.



****

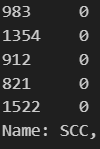
AFTER

BEFORE

1. SCC (binary)

I have converted the SCC into binary numerical format where yes = 1 and no = 0.

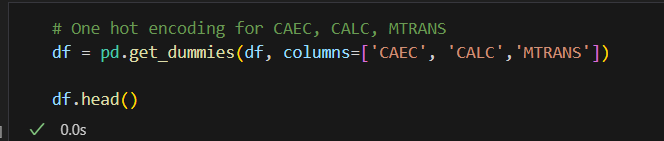


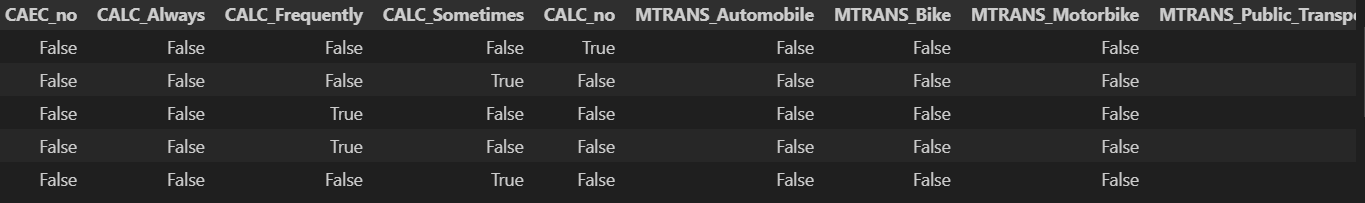
****

AFTER

BEFORE

1. One Hot Encoding (CAEC, CALC, MTRANS)

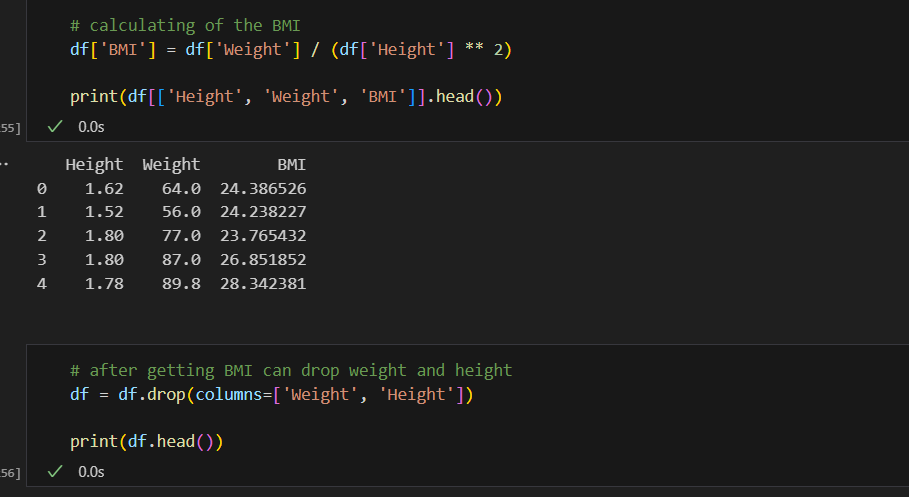


****

1. BMI (Weight and Height)

I have converted weight and height into BMI instead as I feel that it is a more convenient method to see whether the individual has a healthy body weight.

Therefore, I would then drop the weight and height as we do not need it anymore.



**iv. Methods and Improvements**

**Model Selection and Comparison**

1. Random Forest Classifier
   * Initial model with default parameters
   * Performed Grid Search Cross-Validation
   * Hyperparameter tuning:
     + n\_estimators: [50, 100, 200]
     + max\_depth: [3, 5, 7, None]
     + min\_samples\_split: [2, 5, 10]
     + min\_samples\_leaf: [1, 2, 4]
2. Decision Tree Classifier
   * Grid Search Cross-Validation
   * Hyperparameter tuning:
     + max\_depth: [3, 5, 7, None]
     + min\_samples\_split: [2, 5, 10]
     + min\_samples\_leaf: [1, 2, 4]
     + criterion: ['gini', 'entropy']

**Feature Importance**

* Utilized Random Forest's feature importance to understand key predictors
* Identified most significant features for obesity prediction

**Model Performance Metrics**

1. Accuracy Score
2. Confusion Matrix Visualization
3. Cross-Validation Scoring

**v. Results and Analysis**

**Model Performance**

* Best Decision Tree Model
  + Highest Cross-Validation Score: [Insert actual score from your results]
  + Test Accuracy: [Insert actual accuracy from your results]
* Confusion Matrix Analysis
  + Visualized prediction accuracy across different obesity levels
  + Identified potential misclassification patterns

**Key Insights**

* Most important features influencing obesity classification
* Correlations between lifestyle factors and obesity levels

**vi. Conclusion**

* Successfully developed a machine learning model to predict obesity levels
* Demonstrated the potential of using machine learning in health risk assessment
* Highlighted the importance of lifestyle factors in obesity prediction

**vii. References**

1. Scikit-learn Documentation
2. Pandas Documentation
3. Dataset Source [If applicable]
4. Streamlit Documentation

**Recommendations for Future Work**

* Collect more diverse dataset
* Incorporate more advanced feature engineering
* Experiment with additional machine learning algorithms
* Develop more comprehensive health risk prediction models