Machine Learning Project in Python

ANALYSIS OF THE DATASET: ESTIMATION OF OBESITY LEVELS BASED ON EATING HABITSAND PHYSICAL CONDITION

LOLIETT VALDES CASTILLO

2024

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1. INTRODUCTION

Obesity has become a global public health crisis, affecting around 30% of the world's population, with a threefold increase in Latin America over the past five decades (Lin and Li, 2021, Moschonis and Trakman, 2023, Safaei et al., 2021, The Lancet Regional Health –, 2023). The rising prevalence of obesity has severe implications for individual health and contributes to various associated conditions such as diabetes, cardiovascular diseases, and mental health disorders. This trend burdens individuals and strains healthcare systems globally (Lin and Li, 2021, Moschonis and Trakman, 2023).

The body mass index (BMI) classification, endorsed by the World Health Organization, is a crucial metric for assessing weight status and associated health risks. A BMI ≥ 30.0 is indicative of obesity. However, the complex nature of obesity, involving genetic, epigenetic, lifestyle, and environmental factors, poses challenges for effective prevention and intervention. Recent research even explores the impact of the COVID-19 pandemic and lockdowns on obesity risk (Lin and Li, 2021, Safaei et al., 2021).

Assumptions of independence and linearity limit traditional statistical approaches in obesity research. Machine learning (ML) methods, including unsupervised and supervised techniques like Decision Tress and Naïve Bayes, emerge as promising alternatives (Ferdowsy et al., 2021, Thamrin et al., 2021, Chatterjee et al., 2020). These ML methods allow for a more comprehensive analysis of multifactorial influences on obesity, going beyond the constraints of classical statistical models.

This study assesses the efficacy of unsupervised and supervised ML methods using data from an online survey on eating habits and physical conditions in Colombia, Mexico, and Peru (Palechor and Manotas, 2019). The goal is to evaluate the performance of unsupervised and supervised models in predicting obesity and to compare their effectiveness.

2. EXPLORATORY DATA ANALYSIS

2.1 INTRODUCTION TO THE DATASET

Data for this study were collected through an anonymous online survey (Palechor and Manotas, 2019), available in the UCI Machine Learning Repository (UCI Machine Learning Repository, 2019). With 16 questions, the survey aimed to capture information about participants' living habits and demographic details. The dataset included responses from 485 individuals in Colombia, Mexico, and Peru, ranging from 14 to 61 years, and was intended for a machine-learning project.

In a subsequent research phase, each individual's Body Mass Index (BMI) was determined using Equation (1).

$$\text{BMI} = \frac{\text{Weight(kg)}}{\text{Height}^2\ \text{(meters)}} \quad (1)$$

Based on these results, individuals were categorised and assigned obesity levels (Table 1) following WHO and Mexican Normativity guidelines (Palechor and Manotas, 2019):

TABLE 1: BMI CLASSIFICATION

Underweight	Less than 18.5
Normal	18.5 to 24.9
Overweight	25.0 to 29.9
Obesity I	30.0 to 34.9
Obesity II	35.0 to 39.9
Obesity III	Higher than 40

When identifying an imbalance in the distribution of obesity levels, researchers used the Weka tool and Synthetic Minority Over-sampling Technique (SMOTE) filter to generate synthetic data. This process increased the dataset by 70%, resulting in 2111 records and a balanced representation across different obesity levels (Palechor and Manotas, 2019). The dataset comprised 17 features: 16 unlabelled and one labelled "NObeyesdad." Nine features were objects, and eight were float64 in Python (Table 2).

TABLE 2: ORIGINAL COLUMN NAMES, SURVEY QUESTIONS, CATEGORIES AND DATATYPE IN DATASET

Index	Column	Questions and possible answers in the survey	Data type and categories in dataset
0	Gender	What is your gender? ‣ Female ‣ Male	Object: ‣ Female ‣ Male
1	Age	What is your age?	float64
2	Height	What is your height?	float64
3	Weight	What is your weight?	float64
4	family_history_ with_overweight	Has a family member suffered or suffered from being overweight? ‣ Yes ‣ No	Object: ‣ Yes ‣ no
5	FAVC	Do you eat high-caloric food frequently? ‣ Yes ‣ No	Object: ‣ Yes ‣ no
6	FCVC	Do you usually eat vegetables in your meals? ‣ Never ‣ Sometimes ‣ Always ‣ Between 1 y 2	float64
7	NCP	How many main meals do you have daily? ‣ Between 1 y 2 ‣ Three ‣ More than three	float64
8	CAEC	Do you eat any food between meals? ‣ No ‣ Sometimes ‣ Frequently ‣ Always	Object: ‣ no ‣ Sometimes ‣ Frequently ‣ Always
9	SMOKE	Do you smoke? ‣ Yes ‣ No	Object: ‣ Yes ‣ No
10	CH2O	How much water do you drink daily? ‣ Less than a liter ‣ Between 1 and 2 L ‣ More than 2 L	float64
11	SCC	Do you monitor the calories you eat daily? ‣ Yes ‣ No	Object: ‣ Yes ‣ No
12	FAF	How often do you have physical activity? ‣ I do not have ‣ 1 or 2 days ‣ 2 or 4 days ‣ 4 or 5 days	float64
13	TUE	How often do you use technological devices? ‣ 0-2 hours ‣ 3-5 hours ‣ More than 5 hours	float64
14	CALC	How often do you drink alcohol? ‣ I do not drink ‣ Sometimes ‣ Frequently ‣ Always	Object: ‣ no ‣ Sometimes ‣ Frequently ‣ Always
15	MTRANS	Which transportation do you usually use? ‣ Automobile ‣ Motorbike ‣ Bike ‣ Public Transportation ‣ Walking	Object: ‣ Automobile ‣ Motorbike ‣ Bike ‣ Public_Transportation ‣ Walking
16	NObeyesdad Target variable	‣ Underweight Less than 18.5 ‣ Normal 18.5 to 24.9 ‣ Overweight 25.0 to 29.9 ‣ Obesity I 30.0 to 34.9 ‣ Obesity II 35.0 to 39.9 ‣ Obesity III Higher than 40	Object: ‣ Insufficient_Weight ‣ Normal_Weight ‣ Overweight_Level_I ‣ Overweight_Level_II ‣ Obesity_Type_I ‣ Obesity_Type_II ‣ Obesity_Type_III

2.2 DATA EXPLORATION

2.2.1 Method

Data exploration involved checking the dataset's head and tail for initial insights. Entries and columns were reviewed, checking for missing values and examining variable data types. Unusual data was investigated through unique values, descriptive statistics, and visualisation.

2.2.2 Preliminary findings

- The data does not present null values.
- The data does not present unusual values in the variables Gender, Height, Weight, family_history_with_overweight, FAVC, CAEC, SMOKE, SCC, CALC, and MTRANS.
- The variables FCVC, NCP, CH2O, FAF, and TUE, identified as categorical by the researchers (Palechor and Manotas, 2019), were presented as decimal values in the available dataset (Table 3). Decimal values may have been generated through the SMOTE. This method is not recommended for mixed data types, particularly when the data includes categorical features, as it may struggle with multi-categorical features. The decimal values in the synthetic data could stem from interpolation between existing instances and the random factors in this process, resulting in decimal values in the newly created synthetic samples.

TABLE 3: FREQUENCY OF CATEGORICAL VARIABLES WITH DECIMALS

	FCVC	NCP	CH2O	FAF	TUE
0-0.999999	0	0	0	1011	1415
1-1.999999	202	395	769	724	587
2-2.999999	1257	285	1180	301	109
3-3.999999	652	1362	162	75	0
4+	0	69	0	0	0

- The researchers propose six categories for classification based on BMI values (Palechor and Manotas, 2019). However, in the dataset, the overweight category is subdivided into two subcategories (overweight level I and level II), resulting in the target variable (NObeyesdad) having seven categories instead of the initially described six. The limits of these subcategories are not specified in the literature or by the authors (NHS Quality Improvement Scotland, 2010, Palechor and Manotas, 2019, WHO, 2022).
- The BMI ranges described are intended for adults (age ≥ 19), and the dataset includes entries with age < 19 (de Onis et al., 2007, WHO, 2022).

2.2.3 Descriptive statistics

The mean age was 24 years, with a minimum of 14 and a maximum of 61. Only ten entries had an age > 50. The mean height and weight were 1.70 m and 87 kg, respectively. Males were slightly more represented in the dataset (males = 1068, females = 1043). 81.76% of the sample had a family history of overweight (FHO), 88.39% frequently consumed high-caloric food (FAVC), 83.61% ate food between meals sometimes (CAEC), 97.92% did not smoke (SMOKE), 95.45% did not monitor daily calorie intake (SCC), 66.36% consumed alcohol sometimes (CALC), and 74.85% used public transport as their main transportation (MTRANS) (Figure 1).

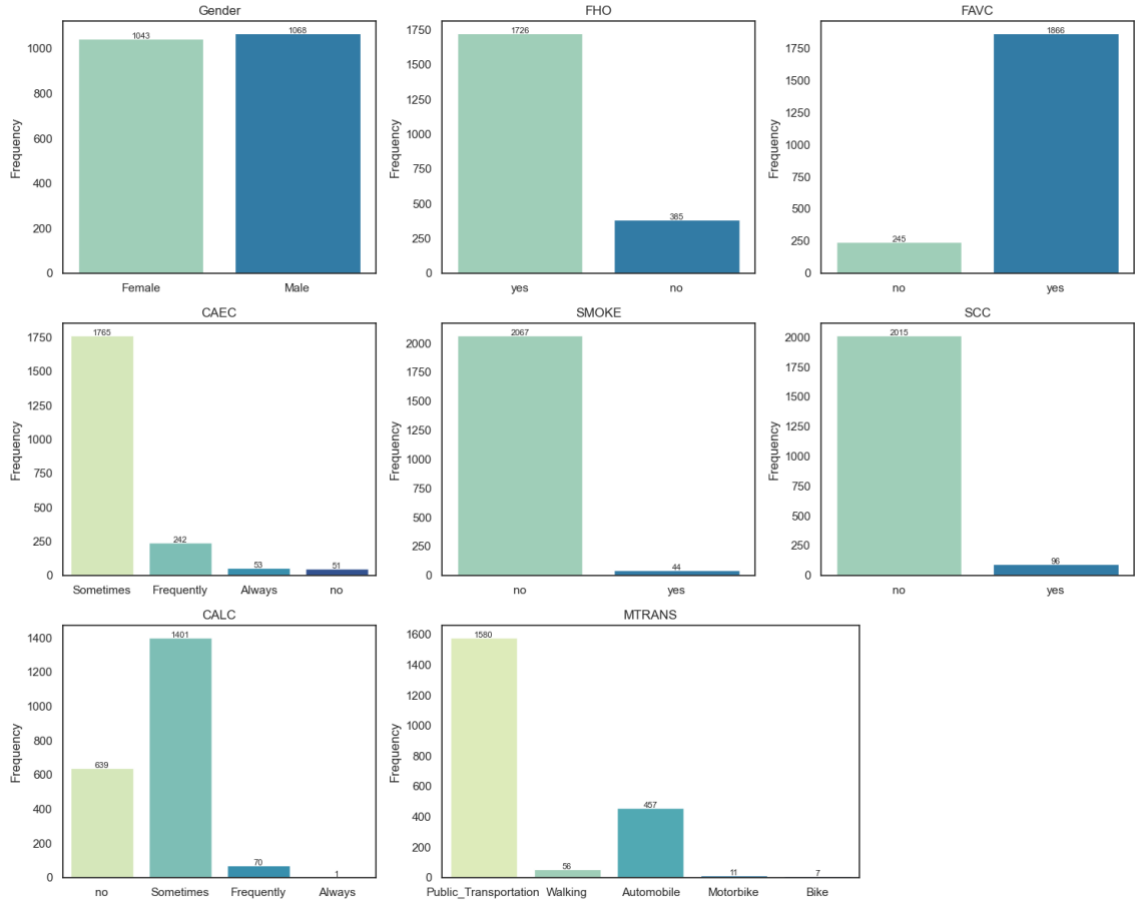


FIGURE 1. FREQUENCY OF EACH CATEGORY FOR CATEGORICAL VARIABLES

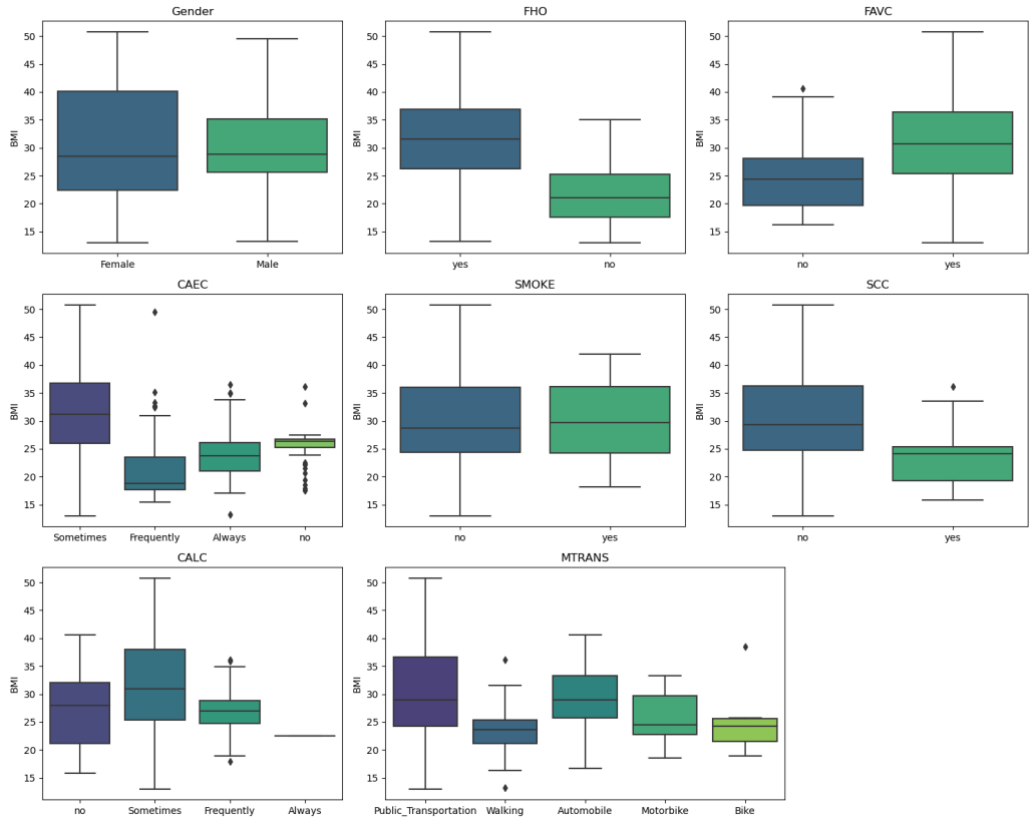


FIGURE 2. CHARACTERIZATION OF EACH CATEGORY ACCORDING TO ITS BMI VALUE

Males and females exhibited similar BMI averages (Figure 2), with females displaying a wider range and higher maximum values. Lower BMI values were evident in individuals without a family history of being overweight (FHO), those who did not consume high-caloric food (FAVC), individuals who frequently snacked between meals (CAEC), those who monitored calories (SCC), those who frequently consumed alcohol (CALC), and those who used walking as their primary mode of transportation (MTRANS). Potential outliers were noted in FAVC, SCC, CALC, MTRANS, and notably in CAEC.

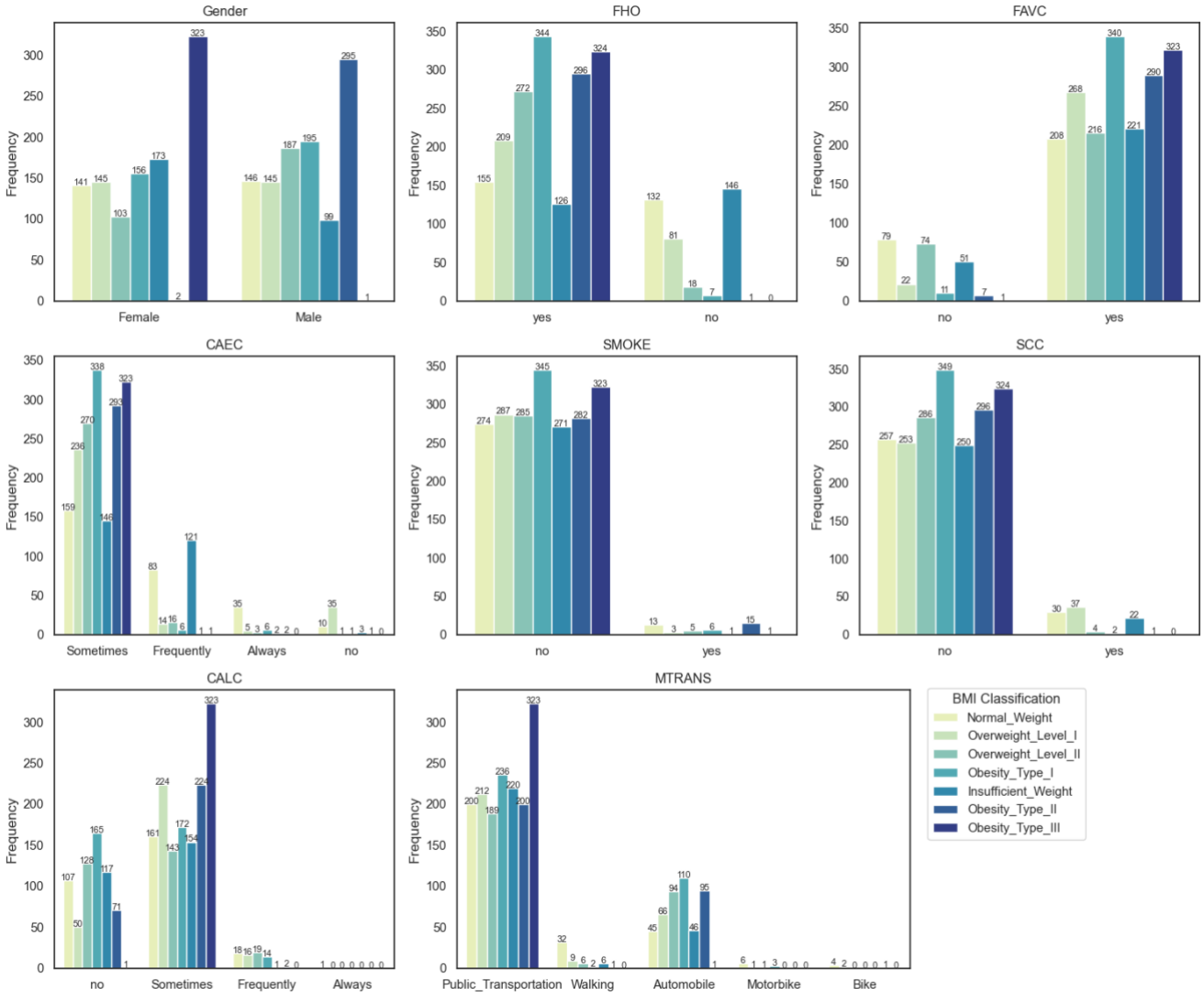


FIGURE 3. FREQUENCY OF EACH CATEGORY BY BMI CLASSIFICATION

Some categories were poorly represented (Figure 3), such as "always" and "no" in CAEC, "yes" in Smoke and SCC, "always" and "frequently" in CALC, and "walking", "motorbike", and "bike" in MTRANS. This posed challenges for ML models in accurately predicting outcomes for these underrepresented categories. Also, some well-represented categories had very low representation of different BMI classifications, as in the case of Obesity_Type II in "females", and in FHO and FAVC where the answer was "no", and Obesity_Type III in "males", and in FHO, FAVC, and CALC where the answer was "no".

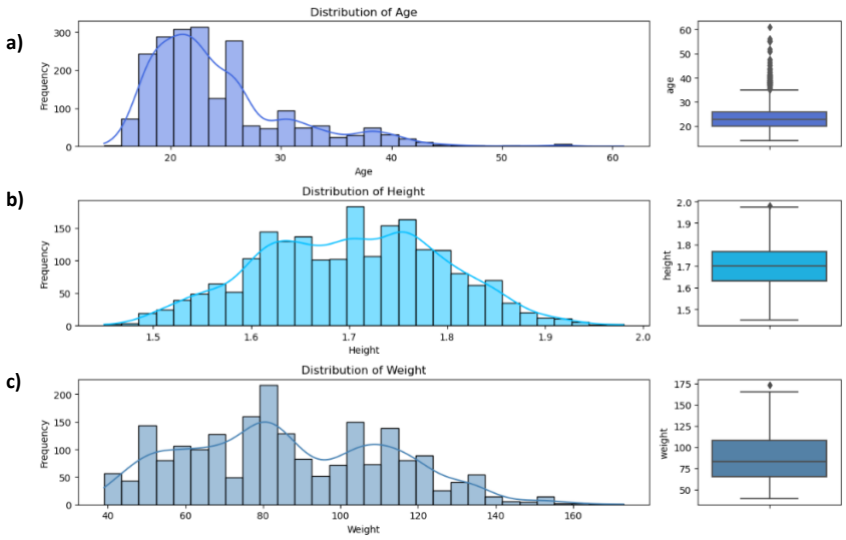


FIGURE 4. DISTRIBUTION OF AGE (a), HEIGHT (b) AND WEIGHT (c)

The height and weight roughly followed a normal distribution (Figure 4). However, age was skewed with a long right tail, indicating a non-uniform distribution. Outliers were observed in all variables, emphasising potential challenges for ML models in handling extreme values.

The variable weight was highly correlated with BMI (Figure 5). This is expected, as it is precisely the weight and height required to calculate BMI and make the consequent classification into one of the categories. Accordingly, the categorical variable created by the researchers based on BMI values (NObeyesdad) will be influenced by this fact.

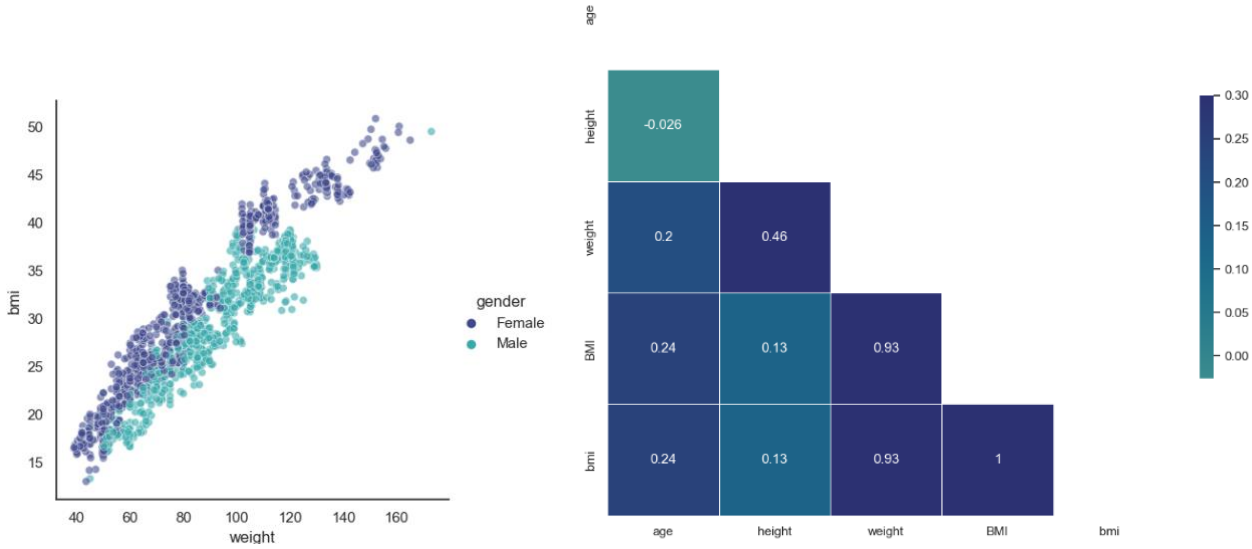


FIGURE 5. EXPLORING CORRELATIONS

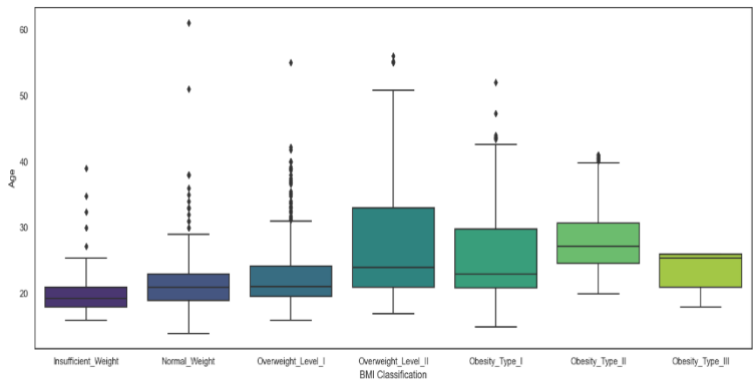


FIGURE 6. FREQUENCY OF EACH CATEGORY BY BMI

In all BMI categories (NObeyesdad), 75% of individuals had an age < 30 years, except in the Overweight Level II group, where the range was wider (Figure 6). The highest median age was observed in the Obesity Type II group, while the lowest was in the Insufficient Weight group. Potential outliers were evident in all groups except for Obesity Type III.

3. UNSUPERVISED METHOD

3.1 PREPARING DATA

3.1.2 CLEANING AND TRANSFORMING THE DATASET

A copy of the original dataset (oh_raw) was created and stored in a two-dimensional 2111 x 17 Pandas Dataset called oh. The modifications and transformations described below were applied:

- Following the PEP8 Guide (Van Rossum, 2001), the variables were renamed as follows:

  - The variable family_history_with_overweight was renamed as fho.

  - All category names were changed to lowercase for the remaining variables.

  - The name of the variable NObeyesdad was changed to a more meaningful name: bmi_class.

- Considering the BMI classification accepted in the literature (NHS Quality Improvement Scotland, 2010, Palechor and Manotas, 2019, WHO, 2022), the categories Overweight_Level_I and Overweight_Level_II were merged as overw (Table 4).
- The rest of the categories in the bmi_class (previously named NObeyesdad) were renamed (Table 4) as follows:

TABLE 4: RENAMING CATEGORIES IN bmi_class

Original name	Modification
Insufficient_Weight	under
Normal_Weight	normal
Overweight_Level_I	overw
Overweight_Level_II	overw
Obesity_Type_III	obes_iii
Obesity_Type_I	obes_i
Obesity_Type_II	obes_ii

- Given the high correlation between weight and BMI, the decision is made to eliminate the weight variable.
- Considering that the variable age is skewed, it was decided to apply a natural logarithm. After this transformation, the variable roughly follows a normal distribution.
- Since the ranges for classification according to the BMI used in this studio apply to adults (age ≥ 19), and there is a different scale for categorising children into various BMI categories (de Onis et al., 2007, WHO, 2022), 351 rows where the age < 19 were excluded.
- As a result of the descriptive statistics and visualisation, the entries with age > 60 were considered outliers, and one rows were dropped for this reason.
- The poorly represented categories in mtrans, calc and caec are merged (Table 5).

TABLE 5: MERGING CATEGORIES IN mtrans

Variable	Original name	New merge category
mtrans	▸ Walking ▸ Bike	active_travel
	▸ Motorbike ▸ Automobile	private
caec and calc	▸ Frequently ▸ Always	Frequently

- In mtrams Public_Transportation changed to public.
- Different strategies were applied to handle the variables FCVC, NCP, CH2O, FAF, and TUE, originally categorical but presented as decimal values in the available dataset (Table 6). Even when it can result in a loss of information, it was decided to choose the change number five and drop these features.

TABLE 6: CHANGING TESTED FOR FEATURES FCVC, NCP, CH2O, FAF, AND TUE

Change proposed	Consequence
1. Delete all the rows with decimal values.	Important information will be lost.
2. Leave all the data as it is presented and escalated later.	In all instances, the variables are ordinal. Due to the limited information, it is impossible to ascertain the original category for each entry. The encoding process is unknown, preventing the assignment of specific categories to the present numbers.
3. Convert all the values to integers, and discard the decimal part.	
4. Round and convert to integer.	
5. Do not consider these features and delete all these columns.	Important features may be lost, resulting in a loss of information.

A two-dimensional 1759 x 12 Pandas dataset was obtained at the end of this step. One of the columns is the logarithmic transformation of age (age_l), ten are attributes and one output.

3.1.3 PREPARING DATA FOR UNSUPERVISED METHOD

Because the original data had a labelled variable (NObeyesdad), the attributes were separated from the output variable as part of the data preparation process for applying an unsupervised method. In oh_atr, all the attributes used for ML and modified in previous phases were stored (gender, age, age_l, height, fho, favc, caec, smoke, scc, calc, and mtrans), resulting in a two-dimensional data frame (1759 x 11). On the other hand, the oh_out series was obtained, containing the output variable (bmiclass).

The nominal variables gender, fho, favc, smoke, scc, mtrans in ohx were encoded using LabelEncoder from the Scikit-learn library. The ordinal variables, caec and calc, were encoded using the OrdinalEncoder from the Scikit-learn library. The output (bmiclass) was also encoded as an ordinal variable like it was done for the ordinal variables in ohx (OrdinalEncoder). New columns were created to store the codes (Table 7).

TABLE 7: ORDINAL CODIFICATION OF THE OUTPUT

Attribute	Category	Encoded	
NOMINAL VARIABLES:			
▸ fho ▸ favc ▸ smoke ▸ scc	No	0	
	Yes	1	
▸ mtrans	active_travel	0	
	Private	1	
	Public	2	
ORDINAL VARIABLES:			
▸ caec ▸ calc	No	0	
	Sometimes	1	
	Frequently	2	
Output	**Category**	**Encoded**	**BMI classification**
bmiclass	under	0	BMI < 18.5 - Underweight
	normal	1	18.5 < BMI < 24.9 - Normal
	overw	2	25.0 < BMI < 29.9 - Overweight I
	obes_i	3	30.0 < BMI < 34.9 - Obesity I
	obes_ii	4	35.0 < BMI < 39.9 - Obesity II
	obes_iii	5	BMI > 40 Obesity III

Both the age and age_l variables were included as features and tested separately in the AHC to assess their impact on the clustering results. The StandardScaler from the Scikit-learn library standardised the relevant columns in the dataset ohx. The standardised values were then organised into an ohx (1759 x 10) new data frame. Simultaneously, the encoded values of ohy were transformed into an array of integers (ohy_out), serving as target variables in the subsequent machine-learning model. The number of rows, columns, and unique values were defined and stored in the n_samples, n_features, and n_digits variables for the ML method.

3.2 APPLYING THE MODEL

An agglomerative hierarchical clustering (AHC) method was used with various distance metrics, including Euclidean, Manhattan, and cosine. Different linkage strategies were also tested: ward, complete, and average. The results were assessed using Silhouette, Completeness, and Homogeneity scores. The combination yielding the maximum values for these three scores was considered the optimal choice. This evaluation process was integral in selecting the most effective configuration for the AHC.

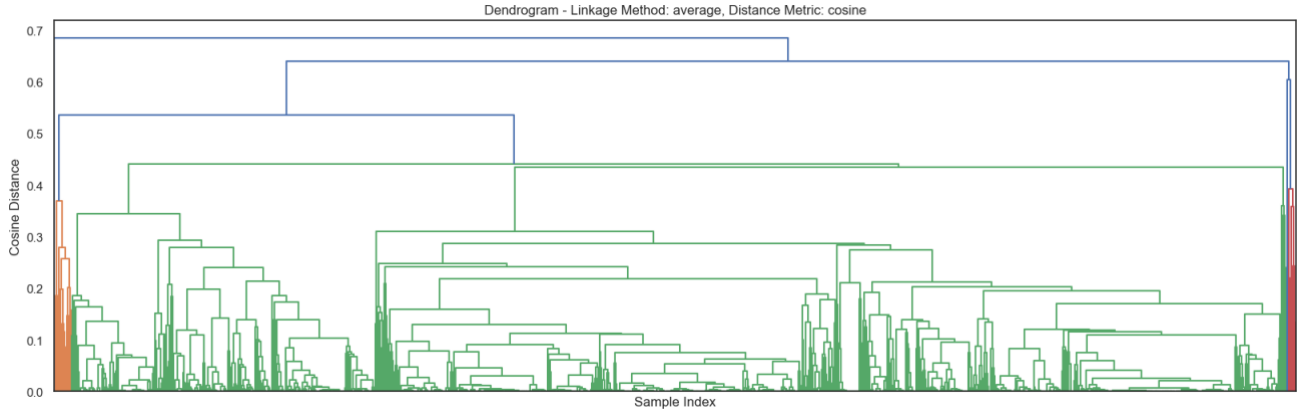
AHC was preferred over K-means for mixed data due to the limitations posed by K-means in handling categorical variables. Additionally, K-Means assigned continuous centroids, posing challenges in interpreting cluster centres when dealing with a combination of continuous and categorical features. In contrast, AHC provided a more versatile approach, accommodating mixed data types and mitigating K-Means limitations in diverse variable types.

3.3 AHC METHOD' RESULT

The maximum scores were obtained with Ward or Average as linkage and Euclidean and cosine as distance (Table 8). With the variable age, without transformation, with higher completeness, homogeneity and Silhouette scores. However, in all the scenarios tested, the scores were very low.

Choosing average linkage and distances, such as cosine, was a strategic approach when confronted with noisy data, outliers, and sparse categories, as observed in this scenario. These methods prioritised overall trends, demonstrating resilience to individual data points. Averaging and distance calculations based on broader patterns minimised the impact of noise. Additionally, cosine distance, relying on angles, remained less affected by the magnitude of outliers. This strategic selection was particularly apt for scenarios where some categories had limited entries, as these methods considered overarching patterns and relationships rather than relying on specific data points.

a)



b)

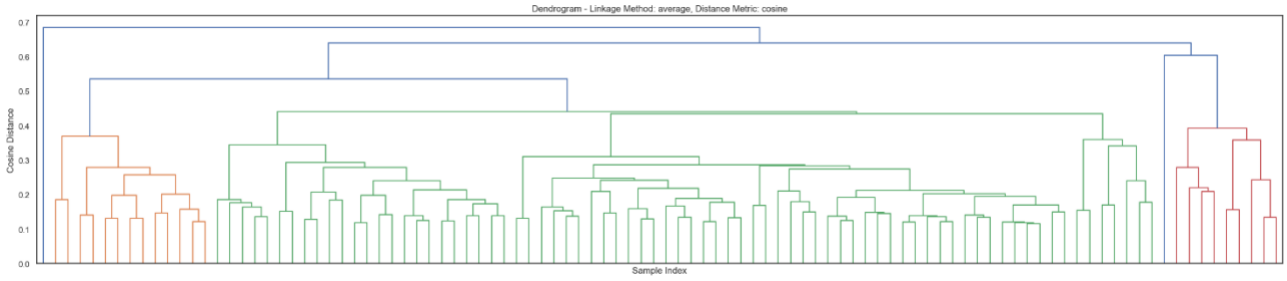


FIGURE 7. AHC DENDROGRAM (a) and AHC DENDROGRAM TRUNCATE MODE (b)

TABLE 8: EVALUATING AHC

Score	Max value	Metric	Linkage
For age without transformation			
Completeness	0.198	Euclidean	Ward
Homogeneity	0.191	Euclidean	Ward
Silhouette	0.215	Cosine	Average
For age with natural logarithm transformation			
Completeness	0.189	Manhattan	Complete
Homogeneity	0.170	Manhattan	Complete
Silhouette	0.213	Cosine	Average

4. SUPERVISED METHOD

4.1 PREPARING DATA

A duplicate of the raw data was generated, and the Age variable was discretised into bins, with values capped and stored in a new variable labelled age_str. Subsequently, a stratified train-test split was executed based on these age groups. Following the split, the age_str column was removed from both the training and testing subsets, having served its purpose in the stratified division. The resulting training subset, denoted as oh_train, encompassed 70% of the original data (n = 1477), while the testing subset, referred to as oh_test (n = 634), comprised 30% of the initial dataset.

The dataset underwent visual exploration and descriptive statistical analysis. Following the procedures outlined in Section 3.1 of this study, uniform modifications, codification, and scaling of numerical variables (age) were applied to the training and test subsets. This approach aimed at ensuring methodological coherence by replicating the specified adjustments consistently across both subsets.

4.2 FITTING THE MODEL

The study involved fitting and evaluating four ML models: K-nearest neighbours (KNN) with uniform weights, K-nearest neighbours with distance weights (KNN-distance), Decision Trees (DTs), and Random Forest Classifier (RFC). Each model underwent training using the designated training datasets (ohx_train for features and ohy_train for labels).

The four classifiers were evaluated and validated using k-fold cross-validation, with k = 10. The resulting classification report and confusion matrix provided insights into precision, recall, F1-score, cross-validated accuracies, and standard deviation. This thorough evaluation and validation process aimed to inform decision-making regarding model selection by comprehensively understanding each classifier's capabilities.

4.3 AHC METHOD' RESULT

4.3.1 Training Subset Results

In the unsupervised method, the variable age was tested with and without a logarithm. The logarithmic transformation showed no improvement in the evaluated parameter. No notable differences existed among all the classifiers evaluated, with values in the range of 0.69 to 0.76 and lower standard deviation for all the considered scores (SD $\leq$ 0.06). The RFC obtained the highest values (Table 9) without the logarithmic transformation of age. With this model, and considering the included risk factor, 76% of the persons were classified correctly (accuracy = 0.76) in the corresponding BMI category, were detected correctly (recall = 0.76), and were identified as being in a class (precision = 0.76). The balance between precision and recall, F1, was equal to 75%.

Regarding the outcomes of the confusion matrix (Figure 8), the RFC demonstrated strong performance, making numerous accurate predictions across all classes. However, class 2 exhibited a relatively higher

misclassification, particularly with class 1 and class 3. Class 5 demonstrated the best performance, with the lowest misclassification rate compared to the other classes. This pattern and differences in the classes' performance were similar for all the classifiers tested.

TABLE 9: EVALUATING AHC

Classifier	Accuracy	Precision	Recall	F1
KNN - uniform weights	0.74 SD=0.04	0.74 SD=0.04	0.74 SD=0.04	0.73 SD=0.04
KNN - distance weights	0.75 SD=0.04	0.75 SD=0.04	0.75 SD=0.04	0.75 SD=0.04
DTs	0.71 SD=0.04	0.73 SD=0.04	0.71 SD=0.04	0.71 SD=0.04
RFC	**0.76** SD=0.03	**0.76** SD=0.04	**0.76** SD=0.03	**0.75** SD=0.04

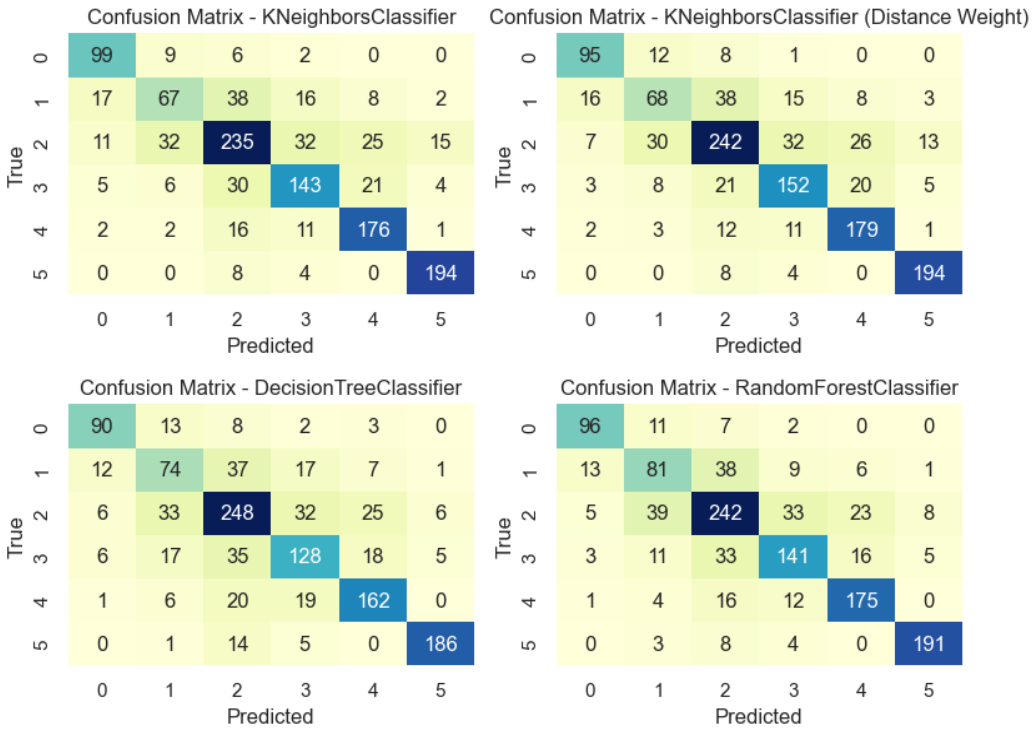


FIGURE 8. CONFUSION MATRIX FOR EACH CLASSIFIER

4.3.2 Test Subset Results

The model obtained higher values in the test subset than in the training subset (Figure 9) with accuracy = 0.74 and values for precision, recall and F1 between 0.53 (recall for class 1) and 0.94 (recall for class 5). Particularly in class 1 (underweight) and class 5 (obesity III), the precision (class 1 = 0.80, class 5 = 0.91), recall (class 1 = 0.84, class 5 = 0.94), and F1 scores (class 1 = 0.82, class 5 = 0.93) are above 80%. These classes are the extreme of the categories analysed.

In the confusion matrix (Figure 10), as in the training subset, class 2 has relatively higher misclassification than the rest, particularly with classes 3 and 4. Class 5 has the lowest misclassification rate compared to the other classes.

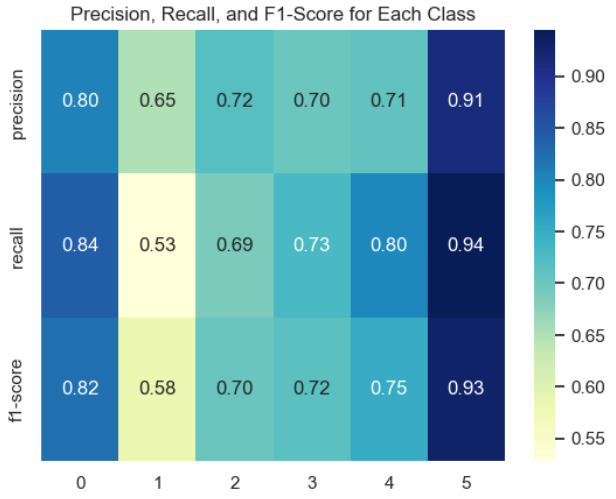


FIGURE 9. SCORES IN TEST SUBSET

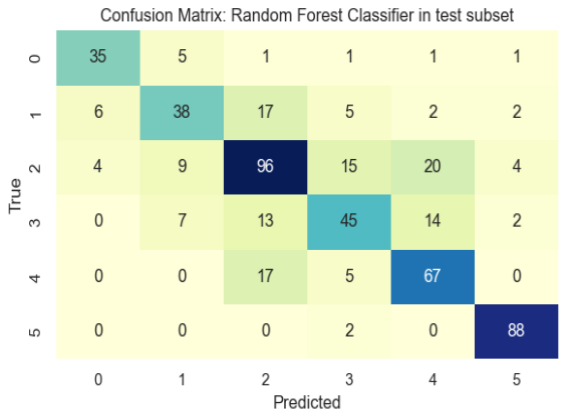


FIGURE 10. CONFUSION MATRIX FOR TEST SUBSET

5. DISCUSSION

AHC indicates inadequate separation of six clusters, with lower scores for completeness, homogeneity and Silhouette. It is ineffective for BMI classification but valuable for exploration. Conversely, RFC successfully detected in the test subset above 70% of cases of obesity I and II and above 90% of obesity III, considering individuals aged between 19 and 60 years old. The result for obesity III and underweight (above 80%) suggests it could be a good model for classifying individuals at risk of obesity or underweight.

However, as class 1, some classes show high misclassification rates and lower recall, F1 scores and precision. Then, some individuals with a normal weight can be misclassified. Therefore, there is room for improvement. Further analysis may involve exploring feature importance and tuning hyperparameters. Adding FCVC, NCP, CH2O, FAF, and TUE is crucial, significantly impacting the result. Additional research is needed to obtain the original dataset, address inconsistencies, and consider these risk factors.

Survey design presents issues surpassing this report's scope, with poorly represented categories. Predominantly, individuals below 35 (75%) impact overall results. Suboptimal data quality significantly

influences the outcome. A more extensive survey with a redesigned questionnaire could positively impact a more effective ML model.

Future analysis could explore using the model with two categories: not at risk and at risk. AHC and RFC suggest the potential for effective risk detection in this scenario. Data engineering tools could include directly calculating BMI and using it as the output variable instead of categorical classification.

6. REFLECTIONS

The development of this project underscored the importance of data quality as a crucial factor. Indeed, the majority of the work focused on that phase. I would have liked to continue making some transformations and trying different options to enhance the method. Still, it is a time-consuming process, and the timeline for this project is limited. Discarding the five risk factors was the most challenging decision. However, I could not find a way to link the presented values with the original categories. I feared manipulating the outcome beyond appropriate and I chose a conservative approach. Nevertheless, I know this decision had a significant impact on the result.

7. CONCLUSION

This dataset provides valuable insights into BMI-influencing factors such as family history, dietary habits, snacking behaviour, calorie monitoring, and transportation choices. However, challenges, including underrepresented groups and damaged variables, emphasise the need for careful consideration in data collection and modelling development. Quality data, a comprehensive understanding of data characteristics, and addressing underrepresented categories and skewed distributions are crucial for designing robust models with broad applicability.

By leveraging ML techniques, this research aims to enhance our understanding of the complex interplay of risk factors contributing to obesity and contribute to the formulation of more effective public policies for prevention and intervention. The study addresses the urgent need for comprehensive strategies to tackle the escalating global obesity crisis.

8. REFERENCES

CHATTERJEE, A., GERDES, M. W. & MARTINEZ, S. G. 2020. Identification of Risk Factors Associated with Obesity and Overweight—A Machine Learning Overview. *Sensors,* 20**,** 2734.

DE ONIS, M., ONYANGO, A. W., BORGHI, E., SIYAM, A., NISHIDA, C. & SIEKMANN, J. 2007. Development of a WHO growth reference for school-aged children and adolescents. *Bull World Health Organ,* 85**,** 660-7.

EUROPE, W. R. O. F. 2022. WHO European Regional Obesity Report. *In:* WHO (ed.).

FERDOWSY, F., RAHI, K. S. A., JABIULLAH, M. I. & HABIB, M. T. 2021. A machine learning approach for obesity risk prediction. *Current Research in Behavioral Sciences,* 2**,** 100053.

LIN, X. & LI, H. 2021. Obesity: Epidemiology, Pathophysiology, and Therapeutics. *Frontiers in Endocrinology,* 12.

MOSCHONIS, G. & TRAKMAN, G. L. 2023. Overweight and Obesity: The Interplay of Eating Habits and Physical Activity. *Nutrients,* 15**,** 2896.

PALECHOR, F. M. & MANOTAS, A. D. L. H. 2019. Dataset for estimation of obesity levels based on eating habits and physical condition in individuals from Colombia, Peru and Mexico. *Data in Brief,* 25**,** 104344.

REPOSITORY, U. M. L. 2019. *Estimation of obesity levels based on eating habits and physical condition* [Online]. [Accessed 20/02/2024].

SAFAEI, M., SUNDARARAJAN, E. A., DRISS, M., BOULILA, W. & SHAPI'I, A. 2021. A systematic literature review on obesity: Understanding the causes & consequences of obesity and reviewing various machine learning approaches used to predict obesity. *Computers in Biology and Medicine,* 136**,** 104754.

SCOTLAND, N. Q. I. 2010. Management of obesity: A national clinical guideline.

THAMRIN, S. A., ARSYAD, D. S., KUSWANTO, H., LAWI, A. & NASIR, S. 2021. Predicting Obesity in Adults Using Machine Learning Techniques: An Analysis of Indonesian Basic Health Research 2018. *Frontiers in Nutrition,* 8.

THE LANCET REGIONAL HEALTH –, A. 2023. The coexistence of obesity and hunger in Latin America and the Caribbean. *The Lancet Regional Health,* 28**,** 100653.

VAN ROSSUM, G. W., BARRY; COGHLAN, ALYSSA. 2001. *PEP 8 – Style Guide for Python Code* [Online]. Available: https://peps.python.org/pep-0008/ [Accessed 20/02/2024 2024].

APPENDIX

Python version 3.11.5

Build Information: September 11, 2023, at 13:26:23

Environment: Anaconda, Inc.

Modules and packages

- Pandas (pd) 2.0.3
- Numpy (np) 1.24.3
- Matplotlib 3.7.2
- Seaborn (sns) 0.12.2
- Sklearn 1.3.0
- SciPy version: 1.11.1
- Math