INTERNATIONAL JOURNAL OF RESEARCH CULTURE SOCIETY ISSN(O): 2456-6683
Monthly Peer-Reviewed, Refereed, Indexed Journal [Impact Factor: 9.241]
Volume - 8, Issue - 9, September - 2024

DOIs:10.2017/IJRCS/202409019 --:-- Research Paper / Article / Review

COMPARISON OF RANDOM FOREST, LGBM, XGB, AND CATBOOST MODELS IN OBESITY RISK PREDICTION

[1] Nandhini Navaneethan, [2] Vijay Devaraj
[1]Senior Software Engineer, Computer Science Engineering, Anna University, Chennai, India
[2] Software Engineer, Mechanical Engineering, Anna University, Chennai, India
Email – [1]nandhininavaneethan47@gmail.com, [2]vijayvicky1999@gmail.com

Abstract: *The increasing prevalence of obesity presents a significant public health challenge, necessitating effective predictive models to identify at-risk individuals and inform intervention strategies. Provided with a dataset of various individuals, we aim to classify obesity risk considering the following factors: Gender, age, height, weight, family history of overweight, food habits, smoking habits, water consumption, physical activity, alcohol consumption, and mode of transport. By leveraging machine learning models such as Random Forest, LGBM, XGB, and Cat boost, we will consider all these factors and predict the obesity risk factors of the individuals. This analysis involves the processes of data exploration, data visualization, feature selection, data processing, model construction, training, and testing of data. Ultimately, this analysis might be insightful for the public and contribute to the reduction of obesity rates through enhanced predictive capabilities and informed public health strategies.*

Key Words: *Feature Prediction, LGBM, XGB, Cat Boost.*

1. INTRODUCTION :

Obesity is a complex and multifaceted health issue characterized by an excessive accumulation of body fat. It is typically measured using the Body Mass Index (BMI), where a BMI of 30 or above is classified as obese. Obesity is associated with numerous health risks, including heart disease, diabetes, and certain cancers, and it is shaped by a blend of genetic, behavioral, environmental, and socio-economic influences. Data analysis plays a crucial role in understanding obesity by helping researchers and policymakers identify patterns, causes, and potential solutions. By analyzing large datasets, we can uncover trends and correlations that may not be immediately visible, enabling more specific interventions and policies.

Here we are mainly using Behavorial data based on dietary habits, physical activity levels, and other lifestyle factors contributing to obesity. But still, we are considering genetic data which constitutes to Information on genetic predispositions to obesity, which can help in identifying at-risk populations and developing personalized interventions and environmental data which is based on factors such as access to healthy foods, socioeconomic status, and urban planning, which influence obesity rates.

The data collection is made with the help of surveys by using tools like the Behavioral Risk Factor Surveillance System (BRFSS) to collect self-reported data on health-related risk behaviors, chronic medical conditions, along with the use of preventive healthcare services.

The analytical methods involve descriptive statistical methods involving summarizing the basic features of the data, such as mean, median, mode, and standard deviation, to understand the distribution and central tendencies; inferential statistical methods using Techniques like regression analysis and hypothesis testing to infer patterns and relationships within the data; multivariate analysis involving techniques such as factor analysis and cluster analysis to understand the interplay between multiple variables and identify subgroups within the population and machine learning algorithms employing algorithms to predict trends and identify risk factors from large, complex datasets.

2. METHODOLOGY

The following diagram represents the workflow of analysis and model comparison and predictive methodologies.

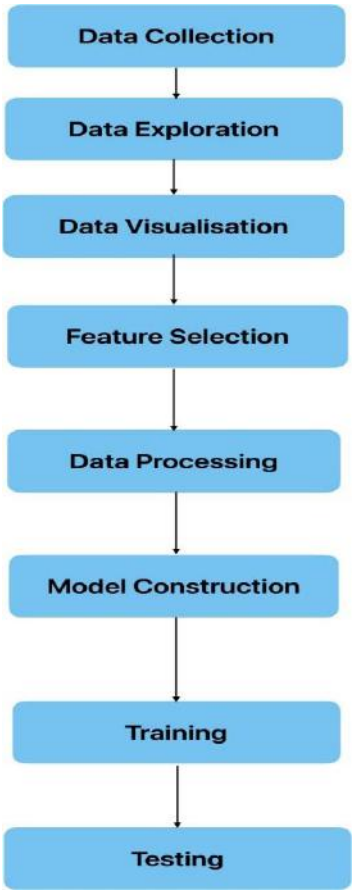


Fig 1: Flow diagram representing various parts of the analysis.

2.1. DATA COLLECTION

The dataset comprises estimates of obesity rates among individuals from Mexico, Peru, and Colombia, aged 14 to 61, featuring a range of dietary habits and physical conditions. Information was collected via an online survey where users participated anonymously. Following data processing, the analysis yielded 17 attributes and a total of 2,111 records.

The attributes related to eating habits are Frequent consumption of high-caloric food (FAVC), Frequency of consumption of vegetables (FCVC), Number of main meals (NCP), Consumption of food between meals (CAEC), Consumption of water daily (CH20), and Consumption of alcohol (CALC). The attributes related to the physical condition are Calories consumption monitoring (SCC), Physical activity frequency (FAF), Time using technology devices (TUE), Transportation used (MTRANS)
variables obtained:
Gender, Age, Height, and Weight.

NObesity values are:
•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

The data contains numerical data and continuous data, so it can be used for analysis based on algorithms of classification, prediction, segmentation, and association. Data is available in CSV format.

2.2. DATA EXPLORATION

The data has been classified based on the following attributes mentioned below:

id
Gender
Age
Height
Weight
family history with overweight
Frequent consumption of high calorie food
Frequency of consumption of vegetables
Number of main meals
Consumption of food between meals
Smoke
Consumption of water daily
Calories consumption monitoring
Physical activity frequency
Time using technology devices
Consumption of alcohol
Transportation used
TARGET

Table 1: List of Attributes for the data.

	id	Gender	Age	Height	Weight	family_history_with_overweight	FAVC	FCVC	NCP	CAEC	SMOKE	CH2O	SCC	FAF	TUE	CALC	MTRANS	NObeyesdad
0	0	Male	24.443011	1.699998	81.669950	yes	yes	2.000000	2.983297	Sometimes	no	2.763573	no	0.000000	0.976473	Sometimes	Public_Transportation	Overweight_Level_II
1	1	Female	18.000000	1.560000	57.000000	yes	yes	2.000000	3.000000	Frequently	no	2.000000	no	1.000000	1.000000	no	Automobile	Normal_Weight
2	2	Female	18.000000	1.711460	50.165754	yes	yes	1.880534	1.411685	Sometimes	no	1.910378	no	0.866045	1.673584	no	Public_Transportation	Insufficient_Weight
3	3	Female	20.952737	1.710730	131.274851	yes	yes	3.000000	3.000000	Sometimes	no	1.674061	no	1.467863	0.780199	Sometimes	Public_Transportation	Obesity_Type_III
4	4	Male	31.641081	1.914186	93.798055	yes	yes	2.679664	1.971472	Sometimes	no	1.979848	no	1.967973	0.931721	Sometimes	Public_Transportation	Overweight_Level_II
5	5	Male	18.128249	1.748524	51.552595	yes	yes	2.919751	3.000000	Sometimes	no	2.137550	no	1.930033	1.000000	Sometimes	Public_Transportation	Insufficient_Weight
6	6	Male	29.883021	1.754711	112.725005	yes	yes	1.991240	3.000000	Sometimes	no	2.000000	no	0.000000	0.696948	Sometimes	Automobile	Obesity_Type_II
7	7	Male	29.891473	1.750150	118.206565	yes	yes	1.397468	3.000000	Sometimes	no	2.000000	no	0.598655	0.000000	Sometimes	Automobile	Obesity_Type_II
8	8	Male	17.000000	1.700000	70.000000	no	yes	2.000000	3.000000	Sometimes	no	3.000000	yes	1.000000	1.000000	no	Public_Transportation	Overweight_Level_I
9	9	Female	26.000000	1.638836	111.275646	yes	yes	3.000000	3.000000	Sometimes	no	2.632253	no	0.000000	0.218845	Sometimes	Public_Transportation	Obesity_Type_III

Fig 2: Sample data overview

The dataset has both testing and training data as well. For training, data has been divided into 18 columns and 20758 rows. The testing data comprises 17 columns and 13840 rows.

Using pandas' methods number of null values, count, number of unique and nonunique values, and minimum and maximum values for each column have been found.

Column Name	count	dtype	nunique	%nunique	%null	min	max
id	20758	int64	20758	100.0	0.0	0	20757
Gender	20758	object	2	0.01	0.0	Female	Male
Age	20758	float64	1703	8.204	0.0	14.0	61.0
Height	20758	float64	1833	8.83	0.0	1.45	1.975663
Weight	20758	float64	1979	9.534	0.0	39.0	165.057269
FHWO	20758	object	2	0.01	0.0	no	yes
FAVC	20758	object	2	0.01	0.0	no	yes
FCVC	20758	float64	934	4.499	0.0	1.0	3.0
NCP	20758	float64	689	3.319	0.0	1.0	4.0
CAEC	20758	object	4	0.019	0.0	Always	no
SMOKE	20758	object	2	0.01	0.0	no	yes
CH2O	20758	float64	1506	7.255	0.0	1.0	3.0
SCC	20758	object	2	0.01	0.0	no	yes
FAF	20758	float64	1360	6.552	0.0	0.0	3.0
TUE	20758	float64	1297	6.248	0.0	0.0	2.0
CALC	20758	object	3	0.014	0.0	Frequently	no
MTRANS	20758	object	5	0.024	0.0	Automobile	Walking
NObeyesdad	20758	object	7	0.034	0.0	Insufficient_Weight	Overweight_Level_II

Fig 3: Attribute Exploration.

The data on various attributes contributing to obesity is compared with gender variables.

Target Distribution with Gender

NObeyesdad	Gender	gender_count	%gender_count	target_class_count	%target_class_count
Insufficient_Weight	Female	1621	0.64	2523	0.12
	Male	902	0.36	2523	0.12
Normal_Weight	Female	1660	0.54	3082	0.15
	Male	1422	0.46	3082	0.15
Obesity_Type_I	Female	1267	0.44	2910	0.14
	Male	1643	0.56	2910	0.14
Obesity_Type_II	Female	8	0.00	3248	0.16
	Male	3240	1.00	3248	0.16
Obesity_Type_III	Female	4041	1.00	4046	0.19
	Male	5	0.00	4046	0.19
Overweight_Level_I	Female	1070	0.44	2427	0.12
	Male	1357	0.56	2427	0.12
Overweight_Level_II	Female	755	0.30	2522	0.12
	Male	1767	0.70	2522	0.12

Fig 4: Target Distribution with Gender

From the above table we can see:

- All people in Obesity_Type_II are male and in Obesity_Type_III all are female.

- Overweight_Level_II consists of 70% male, and Insufficient_Weight consists of more than 60% female.
- From these points we can conclude that gender is important in obesity prediction.

2.3. DATA VISUALISATION

Data visualization transforms raw data into visual representations, making complex information more accessible, understandable, and actionable. In the context of obesity research, effective visualizations can reveal trends, correlations, and outliers that might be missed in traditional data analysis, helping researchers and policymakers to communicate findings more effectively.

Data visualization is an indispensable tool in obesity research, offering a powerful means to understand and communicate the multifaceted aspects of this health issue. By translating data into visual formats, researchers and health professionals can uncover insights, inform policy decisions, and engage the public in meaningful ways. Through innovative visualizations, we can better grasp the complexities of obesity and work towards effective solutions.

First, a graph (bar graph + pie chart) has been made by using seaborn library across target distribution of data based on gender.

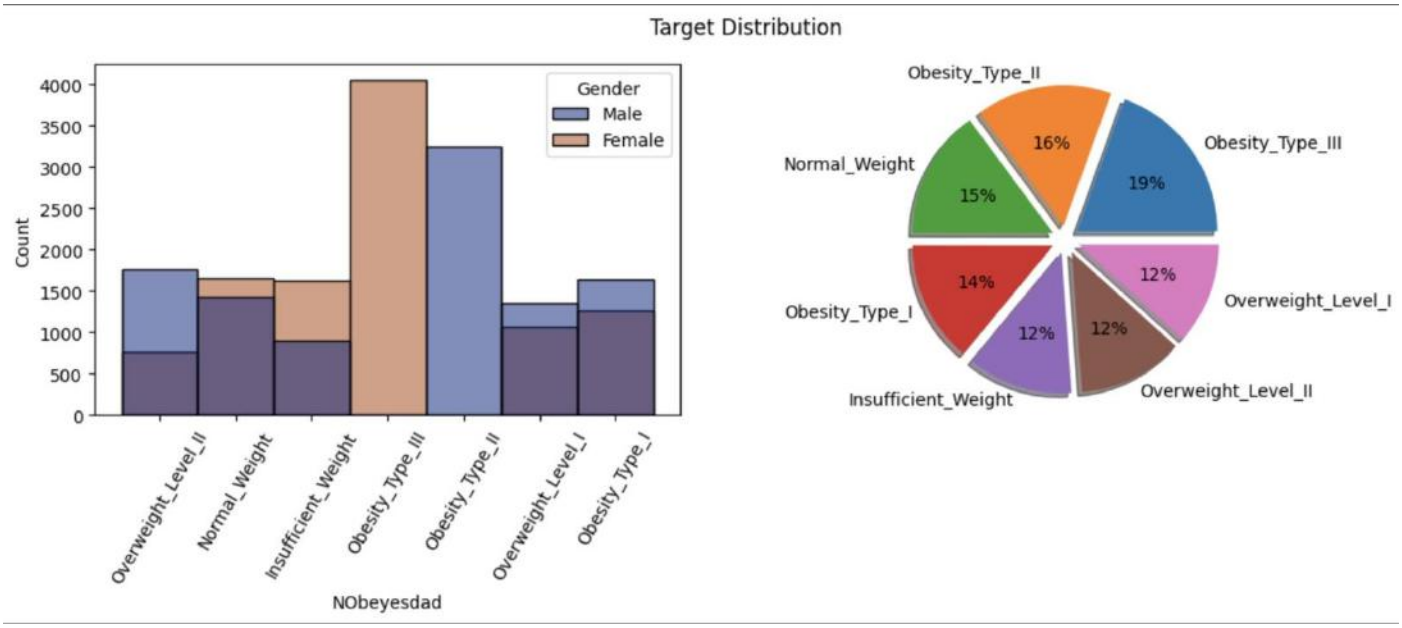


Fig 5: Target Distribution

The following graphical representations have been created for visualizing the data and comparing all the features.

- Individual Numerical Plots
- Individual Categorical Plots
- Numerical Correlation Plots
- Combined Numerical Plots

2.3.1. Individual Numerical Plots

From this graph the following insights can be observed:

- We should Ignore, Female distribution in Obesity_Type_II Class and Male distribution in "Obesity_Type_III". because of very small sample size
- We can see people in the category of Insufficient Weight consume a higher Number of main Meal maybe because to gain weight.
- All individuals categorized as Type III obesity have a vegetable consumption frequency of three.
- Weight, Height and Gender look like the most important features. Weight shows very clear differentiation for diff classes

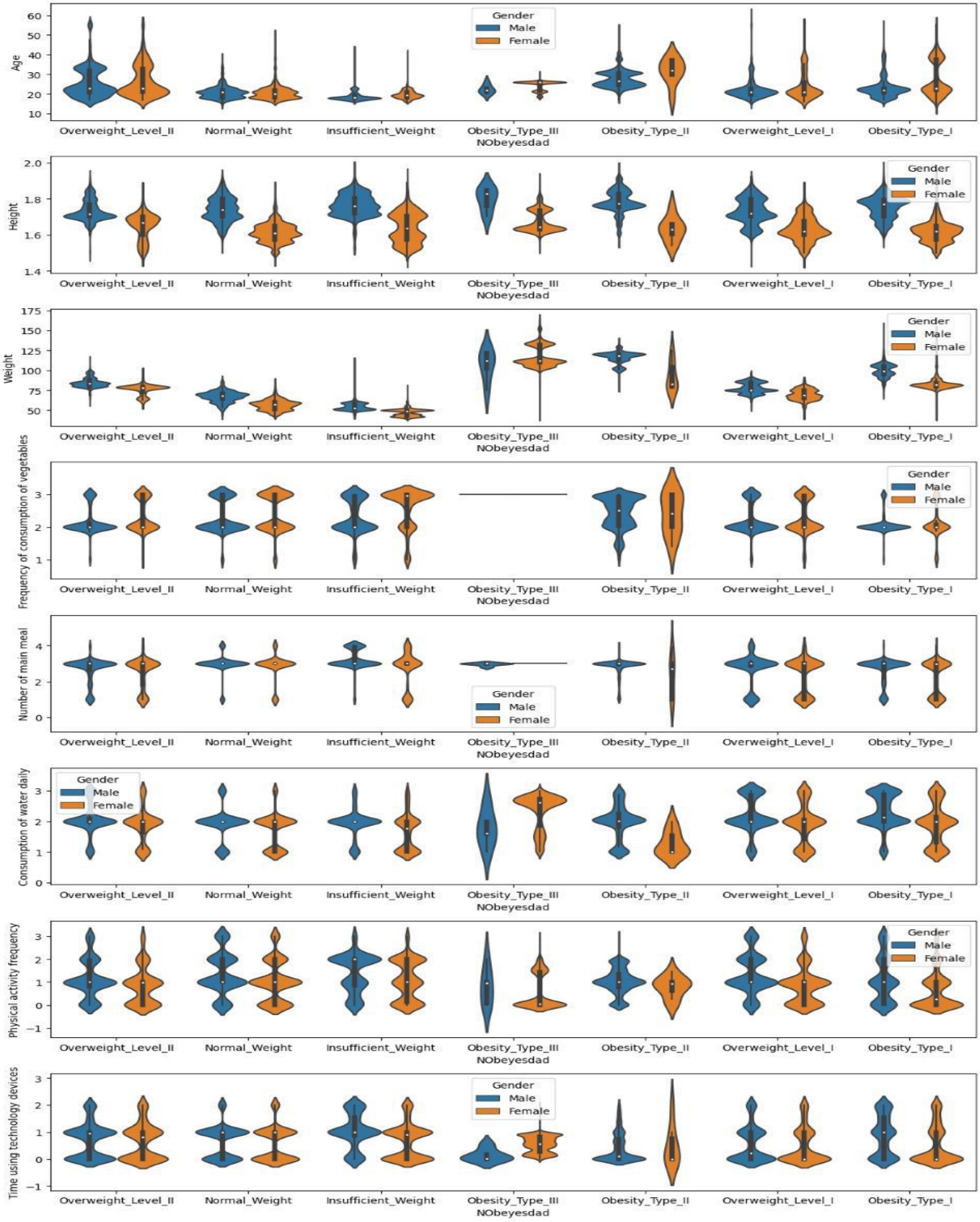
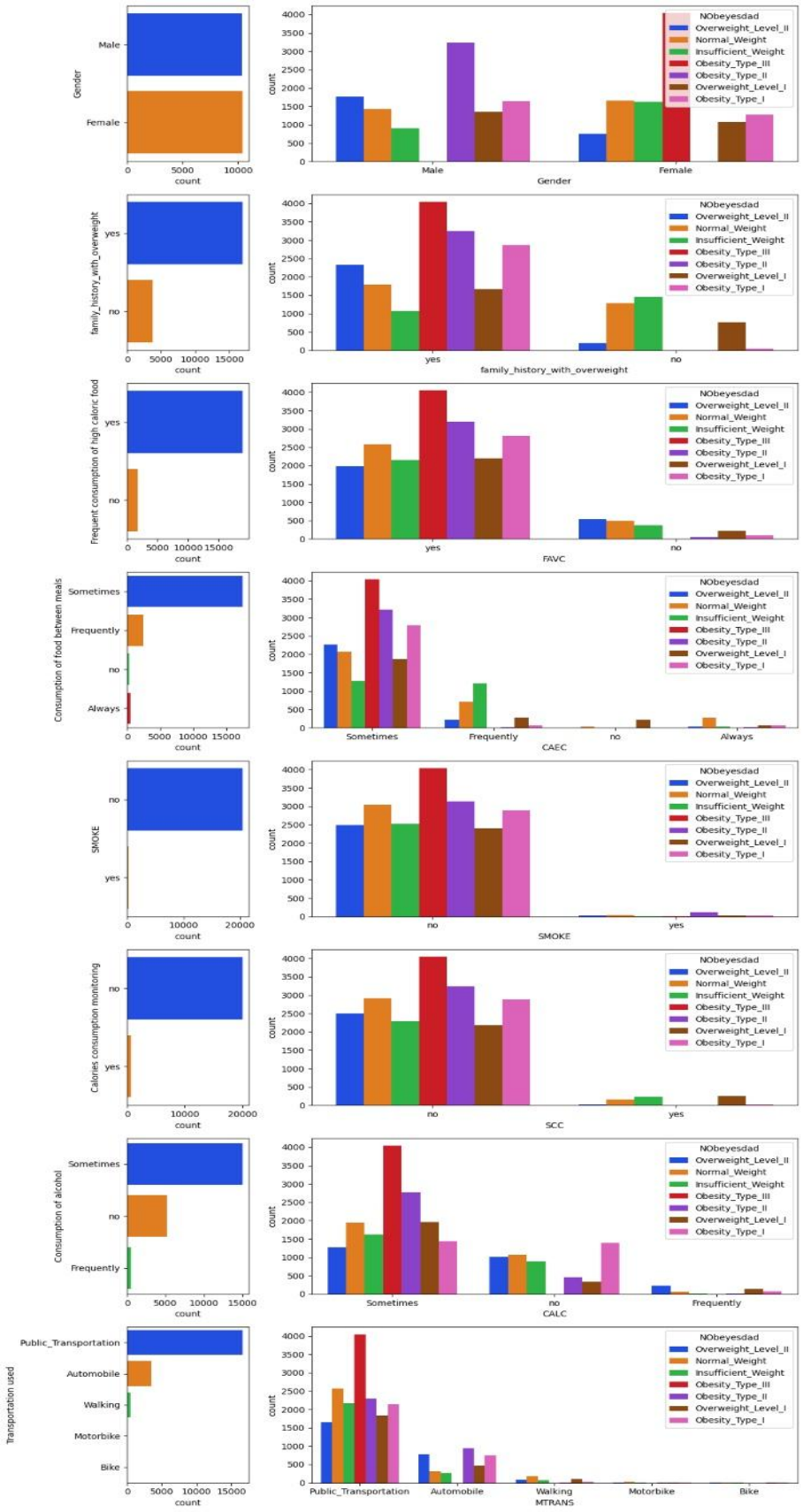


Fig 6: Individual Numerical Plots

2.3.2. INDIVIDUAL CATEGORICAL PLOTS



From the above categorical plot, it is difficult to conclude data has been evenly distributed across various attributes.

2.3.3. NUMERICAL CORRELATION PLOT

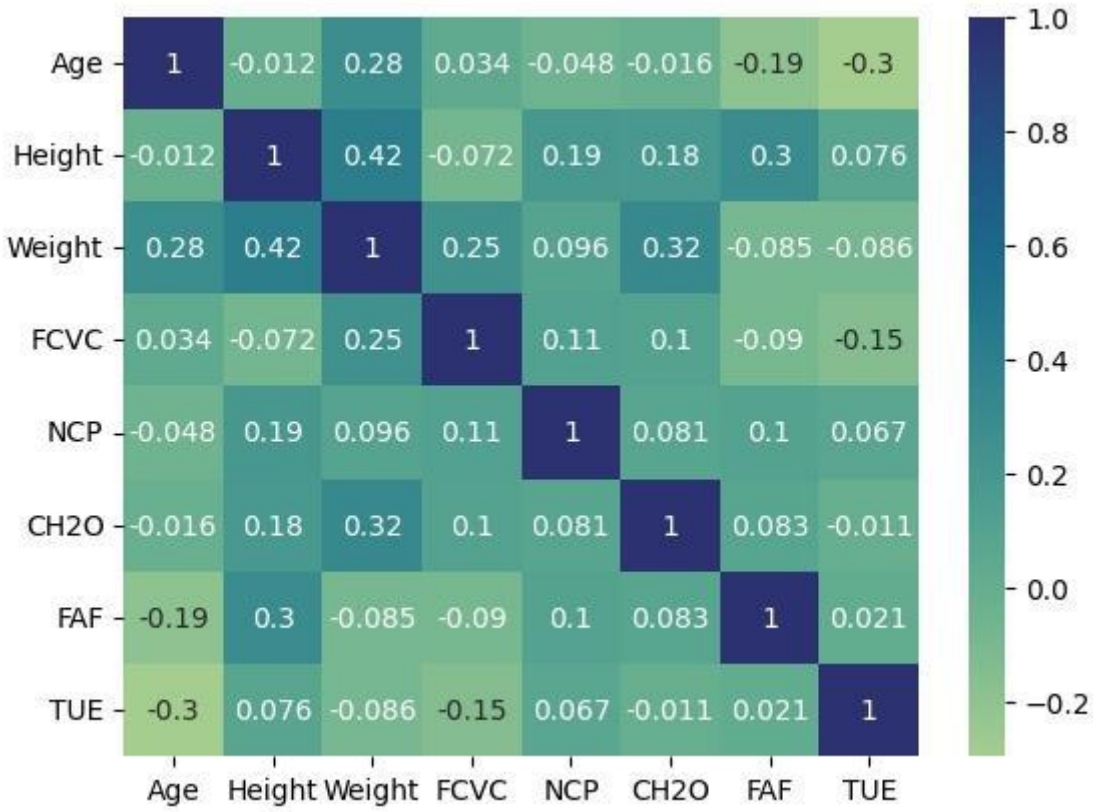


Fig 8: Numerical Correlation Plot

From the above graph we can observe that:

- Height has a positive correlation with Weight, FAF. we will see their combined plots.
- People with higher Weight drink more water.

2.3.4. COMBINED NUMERICAL PLOT

We are using sns.jointplot from the Seaborn library, which creates a multi-panel plot that shows the bivariate relationship between two variables along with their marginal distributions.

A Joint plot is handy for visualising the relationship between two continuous variables and their distributions. Here's how you can create a joint plot using Seaborn.

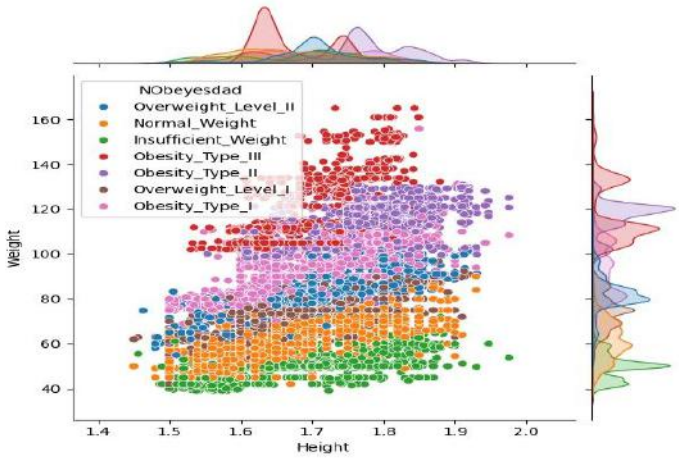


Fig 9: Join plot comparing Height and weight vs Target variable

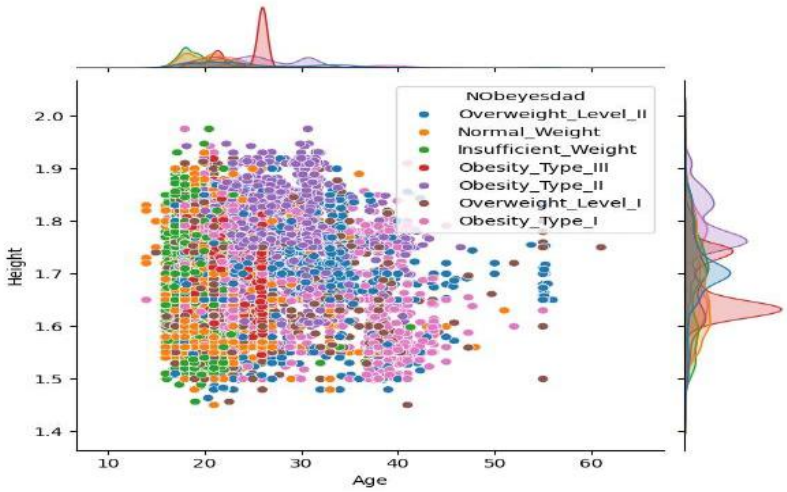


Fig 10: Joint plot comparing height and age vs target variable

2.4 FEATURE SELECTION

For Dimensionality reduction the following 2 methods are used:

1. Principal Component Analysis: Principal Component Analysis (PCA) is a statistical technique used for dimensionality reduction while preserving as much variability as possible in a dataset. It's commonly used in exploratory data analysis and for making predictive models more efficient. PCA transforms the dataset into a new coordinate system, where the highest variance from any projection of the data is represented by the first coordinate, referred to as the first principal component. The second highest variance corresponds to the second coordinate, and this process continues for subsequent components.
2. K-Means Clustering: An unsupervised learning algorithm used to partition data into K distinct clusters based on feature similarity.

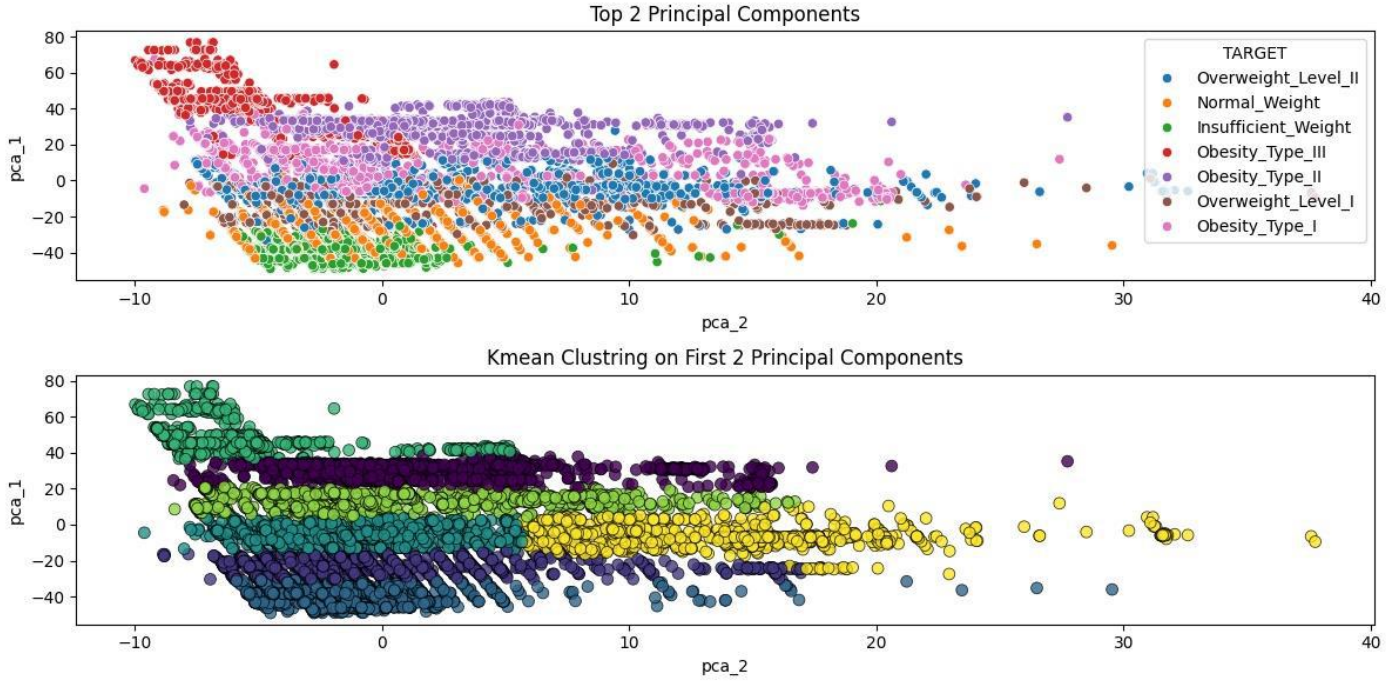


Fig 11

We will establish a function named cross_val_model, which will serve to train and assess all the models used in this notebook. This function will provide three outputs: validation scores, predictions for the validation set, and predictions for the test set.

- **val_scores:** This gives us accuracy score on Validation Data.
- **valid_predictions:** This is an array which stores model predictions on validation set.
- test_predictions: This provides the average predictions for the test set, calculated across the specified number of splits.

Stratified K-Fold cross-validation is a variation of K-Fold cross-validation that ensures each fold is representative of the overall class distribution. This is particularly advantageous for datasets with imbalances, where certain classes are not sufficiently represented. Here's a step-by-step guide to performing Stratified K-Fold cross-validation using Python and scikit-learn.

2.6. MODEL CONSTRUCTION AND TRAINING

Rather than focusing on a single model, it's better to combine predictions from many high performing models.

1. Random Forest Model
2. LGBM Model
3. XGB Model
4. CatBoost Model

2.6.1. RANDOM FOREST MODEL

Random Forest is a versatile and widely used ensemble learning method for classification, regression, and other tasks. It operates by constructing multiple decision trees during training and outputs the mean prediction (regression) or the mode of the classes (classification) of the individual trees.

The following features were considered for constructing the model: 'Gender', 'family_history_with_overweight', 'FAVC', 'CAEC', 'SMOKE', 'SCC', 'CALC', 'MTRANS'.

The testing accuracy for the model is 90.61%.

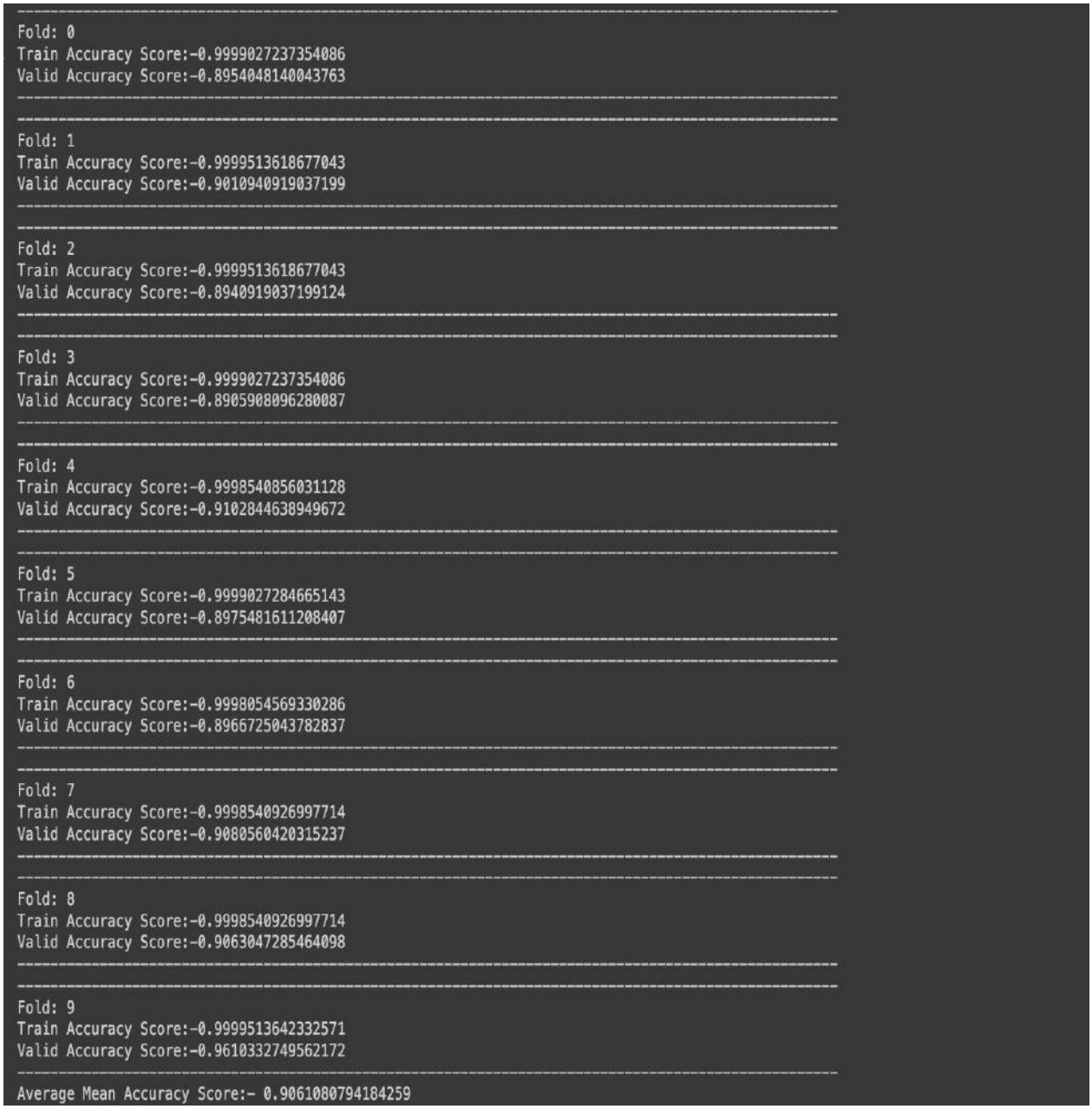


Fig 12 Training using Random Forest

2.6.1. LGBM MODEL

LightGBM (Light Gradient Boosting Machine) is a highly efficient and fast implementation of the gradient boosting framework that uses tree-based learning algorithms. It is designed to be distributed and efficient, making it particularly suitable for large datasets.

The following features were considered for constructing the model: 'Gender', 'family_history_with_overweight', ‘FAVC’, ‘CAEC’, ‘SMOKE’, ‘SCC’, ‘CALC’, ‘MTRANS’.

The testing accuracy for the model is 91.42%.

```
Fold: 0
Train Accuracy Score:-0.9771400778210116
Valid Accuracy Score:-0.9089715536105033
--------------------------------------------------------------------------------------------------------------
--------------------------------------------------------------------------------------------------------------
Fold: 1
Train Accuracy Score:-0.9767509727626459
Valid Accuracy Score:-0.9076586433260394
--------------------------------------------------------------------------------------------------------------
--------------------------------------------------------------------------------------------------------------
Fold: 2
Train Accuracy Score:-0.9776264591439688
Valid Accuracy Score:-0.9059080962800875
--------------------------------------------------------------------------------------------------------------
--------------------------------------------------------------------------------------------------------------
Fold: 3
Train Accuracy Score:-0.9775291828793774
Valid Accuracy Score:-0.9089715536105033
--------------------------------------------------------------------------------------------------------------
--------------------------------------------------------------------------------------------------------------
Fold: 4
Train Accuracy Score:-0.9770428015564202
Valid Accuracy Score:-0.9164113785557987
--------------------------------------------------------------------------------------------------------------
--------------------------------------------------------------------------------------------------------------
Fold: 5
Train Accuracy Score:-0.9779679976654831
Valid Accuracy Score:-0.9076182136602452
--------------------------------------------------------------------------------------------------------------
--------------------------------------------------------------------------------------------------------------
Fold: 6
Train Accuracy Score:-0.9779193618987403
Valid Accuracy Score:-0.9058669001751314
--------------------------------------------------------------------------------------------------------------
--------------------------------------------------------------------------------------------------------------
Fold: 7
Train Accuracy Score:-0.9779193618987403
Valid Accuracy Score:-0.9194395796847635
--------------------------------------------------------------------------------------------------------------
--------------------------------------------------------------------------------------------------------------
Fold: 8
Train Accuracy Score:-0.977676183065026
Valid Accuracy Score:-0.908493870402802
--------------------------------------------------------------------------------------------------------------
--------------------------------------------------------------------------------------------------------------
Fold: 9
Train Accuracy Score:-0.9742230436262828
Valid Accuracy Score:-0.9527145359019265
--------------------------------------------------------------------------------------------------------------
Average Mean Accuracy Score:- 0.91420543252078
```

Fig 13: Training using LGBM Model.

2.6.3. XGB MODEL.

XGBoost (Extreme Gradient Boosting) is a highly efficient and powerful version of the gradient boosting framework. It is engineered for speed and performance, making it a popular choice for handling both regression and classification challenges.

The following features were considered for constructing the model: 'Gender', 'family_history_with_overweight', ‘FAVC’, ‘CAEC’, ‘SMOKE’, ‘SCC’, ‘CALC’, ‘MTRANS’.

The testing accuracy for the model is 91.63%.

```
Fold: 0
Train Accuracy Score:-0.9452821011673151
Valid Accuracy Score:-0.9111597374179431
--------------------------------------------------------------------------------------------------------------
--------------------------------------------------------------------------------------------------------------
Fold: 1
Train Accuracy Score:-0.945136186770428
Valid Accuracy Score:-0.9063457330415755
--------------------------------------------------------------------------------------------------------------
--------------------------------------------------------------------------------------------------------------
Fold: 2
Train Accuracy Score:-0.9449902723735408
Valid Accuracy Score:-0.9080962800875274
--------------------------------------------------------------------------------------------------------------
--------------------------------------------------------------------------------------------------------------
Fold: 3
Train Accuracy Score:-0.9454280155642023
Valid Accuracy Score:-0.9059080962800875
--------------------------------------------------------------------------------------------------------------
--------------------------------------------------------------------------------------------------------------
Fold: 4
Train Accuracy Score:-0.9432392996108949
Valid Accuracy Score:-0.9199124726477024
--------------------------------------------------------------------------------------------------------------
--------------------------------------------------------------------------------------------------------------
Fold: 5
Train Accuracy Score:-0.9460629346821653
Valid Accuracy Score:-0.9128721541155866
--------------------------------------------------------------------------------------------------------------
--------------------------------------------------------------------------------------------------------------
Fold: 6
Train Accuracy Score:-0.946160206215651
Valid Accuracy Score:-0.9106830122591943
--------------------------------------------------------------------------------------------------------------
--------------------------------------------------------------------------------------------------------------
Fold: 7
Train Accuracy Score:-0.9456252127814795
Valid Accuracy Score:-0.9168126094570929
--------------------------------------------------------------------------------------------------------------
--------------------------------------------------------------------------------------------------------------
Fold: 8
Train Accuracy Score:-0.9446524974466223
Valid Accuracy Score:-0.9106830122591943
--------------------------------------------------------------------------------------------------------------
--------------------------------------------------------------------------------------------------------------
Fold: 9
Train Accuracy Score:-0.9407130003404504
Valid Accuracy Score:-0.9610332749562172
--------------------------------------------------------------------------------------------------------------
Average Mean Accuracy Score:- 0.9163506382522121
```

Fig 14: Training using XGB

2.6.4. CATBOOST MODEL.

CatBoost (Categorical Boosting) is a high-performance gradient boosting library developed by Yandex. It is particularly well-suited for handling categorical features and often requires minimal preprocessing.

The following features were considered for constructing the model: 'Gender', 'family_history_with_overweight', ‘FAVC’, ‘CAEC’, ‘SMOKE’, ‘SCC’, ‘CALC’, ‘MTRANS’.

The testing accuracy for the model is 91.21%.

```
Fold: 0
Train Accuracy Score:-0.9478599221789883
Valid Accuracy Score:-0.9050328227571116
----------------------------------------------------------------------------------------------------
----------------------------------------------------------------------------------------------------
Fold: 1
Train Accuracy Score:-0.9498540856031128
Valid Accuracy Score:-0.9054704595185996
----------------------------------------------------------------------------------------------------
----------------------------------------------------------------------------------------------------
Fold: 2
Train Accuracy Score:-0.9500972762645914
Valid Accuracy Score:-0.9024070021881838
----------------------------------------------------------------------------------------------------
----------------------------------------------------------------------------------------------------
Fold: 3
Train Accuracy Score:-0.949124513618677
Valid Accuracy Score:-0.9050328227571116
----------------------------------------------------------------------------------------------------
----------------------------------------------------------------------------------------------------
Fold: 4
Train Accuracy Score:-0.9481517509727626
Valid Accuracy Score:-0.9133479212253829
----------------------------------------------------------------------------------------------------
----------------------------------------------------------------------------------------------------
Fold: 5
Train Accuracy Score:-0.9515587763241088
Valid Accuracy Score:-0.908493870402802
----------------------------------------------------------------------------------------------------
----------------------------------------------------------------------------------------------------
Fold: 6
Train Accuracy Score:-0.9524342201254803
Valid Accuracy Score:-0.9054290718038529
----------------------------------------------------------------------------------------------------
----------------------------------------------------------------------------------------------------
Fold: 7
Train Accuracy Score:-0.9509265113564516
Valid Accuracy Score:-0.9076182136602452
----------------------------------------------------------------------------------------------------
----------------------------------------------------------------------------------------------------
Fold: 8
Train Accuracy Score:-0.9513155974903944
Valid Accuracy Score:-0.9111208406304728
----------------------------------------------------------------------------------------------------
----------------------------------------------------------------------------------------------------
Fold: 9
Train Accuracy Score:-0.9446524974466223
Valid Accuracy Score:-0.957968476357268
----------------------------------------------------------------------------------------------------
Average Mean Accuracy Score:- 0.9121921501301029
```

Fig 15: Training using Catboost.

2.7 MODEL EVALUATION

Ensemble learning involves combining the predictions of multiple machine learning models to produce a more robust and accurate prediction. Ensemble methods such as bagging, boosting, and stacking are widely used in machine learning. In this context, we will emphasize stacking, a technique that entails training several models, referred to as base learners, and then using an additional model, called the meta-learner, to aggregate the predictions from the base models.

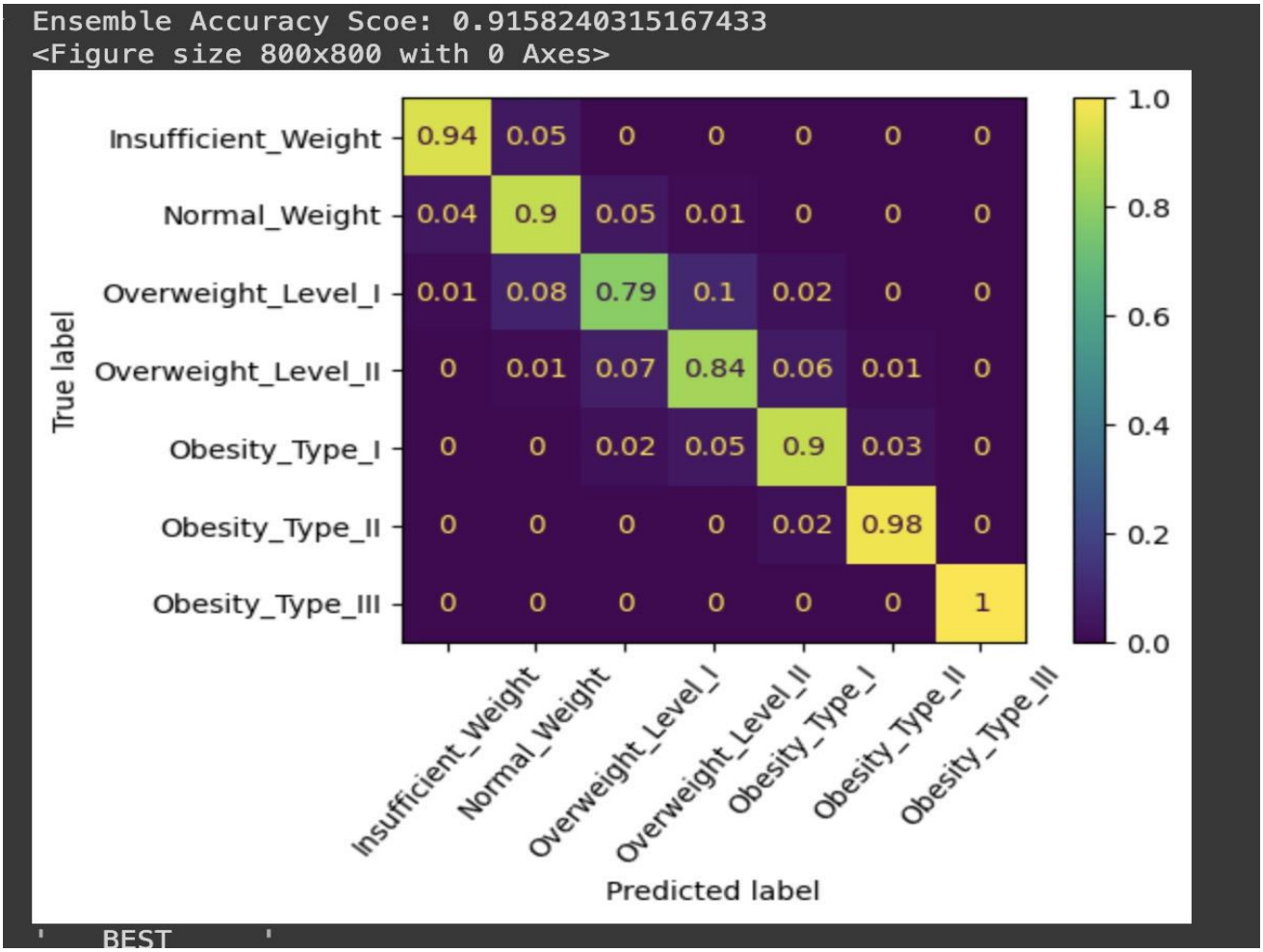


Fig 16: Ensemble Accuracy Score

REFERENCES:

1. Lee, A., Mhurchu, C.N., Sacks, G., Swinburn, B., & Snowdon, W. ; "Predicting Overweight and Obesity in Childhood from Early Feeding Practices" Archives of Pediatrics & Adolescent Medicine (JAMA Pediatrics), 2010.
2. Simmonds, M., Llewellyn, A., Owen, C.G., & Woolacott, N. ; "Predicting Adult Obesity from Childhood Obesity: A Systematic Review and Meta-analysis" Journal: Obesity Reviews , 2016.
3. He, F., & Wu, J. ; "Predicting Childhood Obesity Using Electronic Health Records and Publicly Available Data" , Journal: IEEE Journal of Biomedical and Health Informatics, 2017.
4. Farpour-Lambert, N.J., Aggoun, Y., Marchand, L.M., Martin, X.E., Herrmann, F.R., & Beghetti, M. ; "Predicting Obesity Risk Using Machine Learning Techniques", Journal: Obesity Facts, 2019.
5. Whitaker, R.C., Wright, J.A., Pepe, M.S., Seidel, K.D., & Dietz, W.H. ; "Predicting Obesity in Young Adulthood from Childhood and Parental Obesity", Journal: New England Journal of Medicine, 1997
6. Lytle, L.A., Murray, D.M., Perry, C.L., Eldridge, A.L., & Farbakhsh, K. ; "Predicting Obesity by Assessing Adolescent Behaviors" , Journal: Archives of Pediatrics & Adolescent Medicine (JAMA Pediatrics), 2004