Kaggle Project Report

Challenge: Multi-Class Prediction of Obesity Risk

Members:

Nadia SYLLA
Uyen NGUYEN
Nour ABBOUD
Jihane LAGNAOUI
Ramisa HEIDARI

IMT Atlantique
Bretagne-Pays de la Loire
École Mines-Télécom

25/03/2024

Table of content

I. Introduction

In recent years, machine learning techniques have risen to prominence as potent tools for predictive analytics in healthcare, particularly in forecasting diseases like obesity. Within the realm of obesity and its correlation with cardiovascular health, harnessing multi-class prediction algorithms presents significant promise in pinpointing high-risk individuals based on various features such as age, height, weight, and more. This proactive approach aims to identify characteristics indicative of heightened susceptibility, facilitating the implementation of preventive measures. In this study, our objective is to employ a machine learning framework for a multi-class prediction task to ascertain the specific type of obesity affecting individuals.

II. Data overview

The dataset utilized for this competition was derived from an original collection of data focused on assessing obesity levels among individuals from Mexico, Peru, and Colombia, aged 14 to 61. These individuals had varied dietary habits and physical conditions. The data was gathered through an online survey platform where participants anonymously completed a questionnaire. Subsequently, this information was processed to create a dataset featuring 17 attributes and 1 target variable across 20758 records.

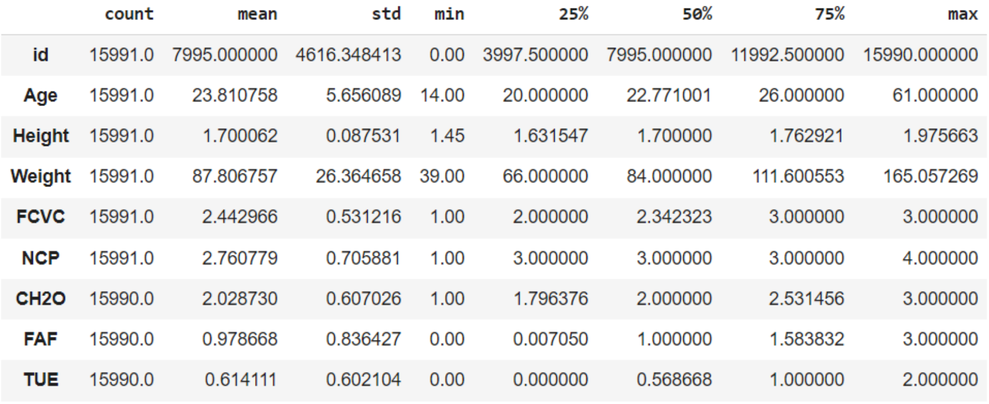
1. Feature Description

- id: Unique identifier for each individual.
- Gender: Gender of the individual (Male/Female).
- Age: Age of the individual.
- Height: Height of the individual in meters.
- Weight: Weight of the individual in kilograms.
- family_history_with_overweight: Whether the individual has a family history of overweight (yes/no).
- FAVC: Frequency of consuming high-caloric food (categorical: yes/no).
- FCVC: Frequency of consuming vegetables (numeric).

- NCP: Number of main meals per day (numeric).
- CAEC: Consumption of food between meals (categorical: no/sometimes/frequently/always).
- SMOKE: Whether the individual smokes (yes/no).
- CH2O: Daily water consumption in liters (numeric).
- SCC: Calories consumption monitoring (categorical: no/sometimes/frequently/always).
- FAF: Physical activity frequency (numeric).
- TUE: Time using technology devices (numeric).
- CALC: Consumption of alcohol (categorical: no/sometimes/frequently/always).
- MTRANS: Mode of transportation (categorical: automobile/bike/motorbike/public transportation/walking).
- **NObeyesdad**: Obesity level classification (target variable).

Note: The target variable is "NObeyesdad," representing the obesity level of the individual, categorized into 7 different levels that are Insufficient_Weight, Normal_Weight, Overweight_Level_I, Overweight_Level_II, Obesity_Type_I, Obesity_Type_II and Obesity_Type_III.

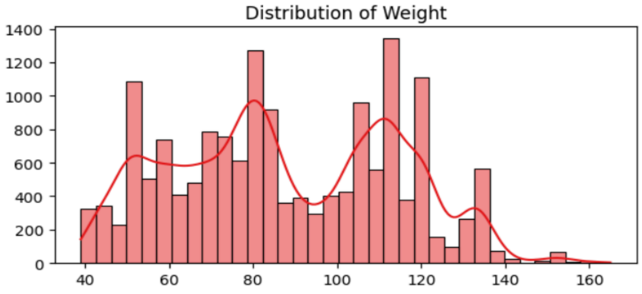
III. Exploratory Data Analysis



	count	mean	std	min	25%	50%	75%	max
id	15991.0	7995.000000	4616.348413	0.00	3997.500000	7995.000000	11992.500000	15990.000000
Age	15991.0	23.810758	5.656089	14.00	20.000000	22.771001	26.000000	61.000000
Height	15991.0	1.700062	0.087531	1.45	1.631547	1.700000	1.762921	1.975663
Weight	15991.0	87.806757	26.364658	39.00	66.000000	84.000000	111.600553	165.057269
FCVC	15991.0	2.442966	0.531216	1.00	2.000000	2.342323	3.000000	3.000000
NCP	15991.0	2.760779	0.705881	1.00	3.000000	3.000000	3.000000	4.000000
CH2O	15990.0	2.028730	0.607026	1.00	1.796376	2.000000	2.531456	3.000000
FAF	15990.0	0.978668	0.836427	0.00	0.007050	1.000000	1.583832	3.000000
TUE	15990.0	0.614111	0.602104	0.00	0.000000	0.568668	1.000000	2.000000

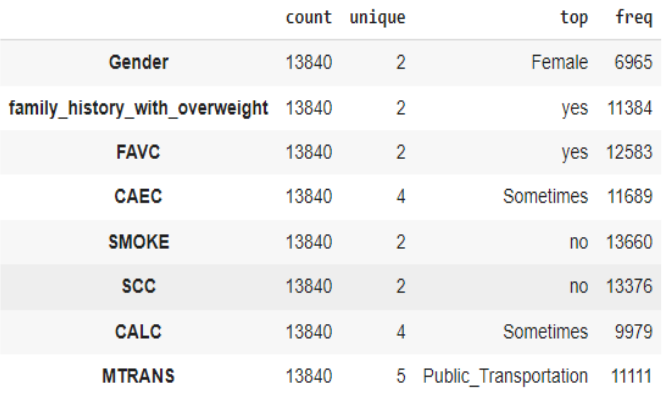
Table 1: Statistic descriptive of the dataset

The descriptive statistics (*cf table 1*) show that there are no missing values or duplicate values in the given dataset. In addition we can clearly say that the age of the individuals in the dataset, with an average of approximately 23.95 years indicating a youthful sample group. The average height of individuals is 1.7 meters, with a range from 1.45 to 1.98 meters. The data indicates a normal height distribution for an adult population with a low standard deviation.



Graphic 1: Distribution of weight

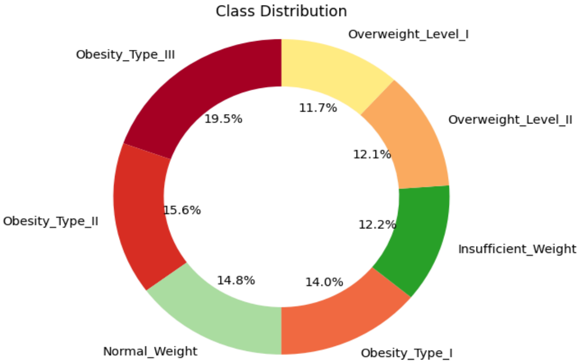
However, individuals have a wide range of weights, from 39 to 165.07 kilograms, with an average of 87.84 kilograms. A large standard deviation suggests substantial variability, which might be correlated with varying obesity levels (*cf graphic 1*).



	count	unique	top	freq
Gender	13840	2	Female	6965
family_history_with_overweight	13840	2	yes	11384
FAVC	13840	2	yes	12583
CAEC	13840	4	Sometimes	11689
SMOKE	13840	2	no	13660
SCC	13840	2	no	13376
CALC	13840	4	Sometimes	9979
MTRANS	13840	5	Public_Transportation	11111

Table 2: Frequent value of each variable

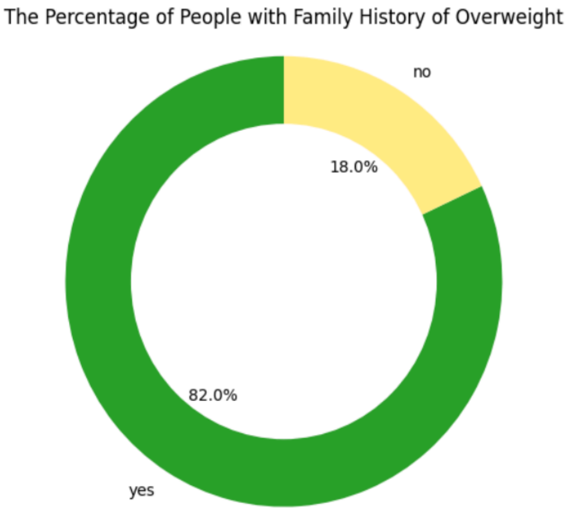
In contrast to the previous summary, the figure below (*cf table 2*) describes the most frequent value in each variable, which expose the nature of two mean variables those who are intuitively take part in gaining weight according to consumption behavior studies are FAVC and CAEC respectively Frequent consumption of high caloric food , Consumption of food between meals.



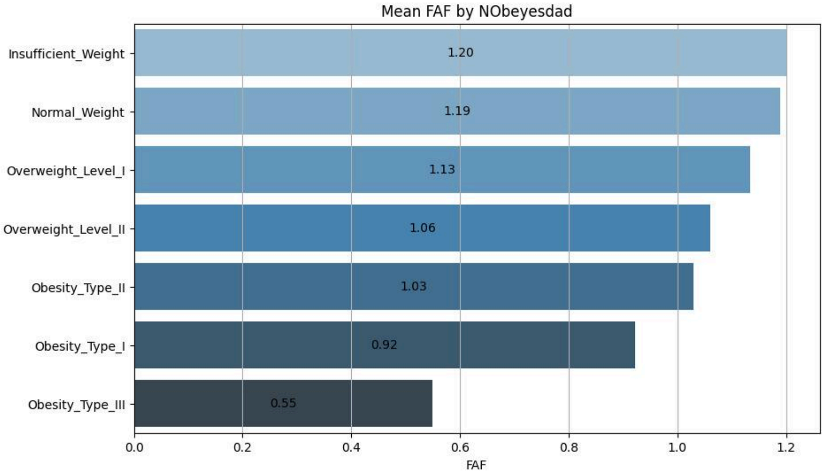
Graphic 2: Class distribution

We also have a balance in the class distribution (*cf graphic 2*). There are 7 classes whose proportion goes from 11.7% to 19.5%. The least common class is Overweight_Level_I and the most common one is Obesity_Type_III.

The donut chart (*cf graphic 3*) shows that 82% of the people in the dataset have a family history of being overweight. Conversely, only 18% of the people in the dataset do not have a family history of being overweight. This information can be important in understanding the prevalence of overweight and obesity within this group and may also be relevant to conclude that a family with such a history is likely to have kids who are going to developpe a bad eating behavior with lack of physical activity to compensate the calories intake in the same way as the grown-ups knowing that dietary is ruled by their habits.



Graphic 3: Family_Historic_overweight feature



The count plot above (*cf graphic 4*) shows us the mean of Physical Activity Frequency of each Class. From this, we can assume that the more you exercise, the more you avoid the risk of being overweight and obese (intuitive).

Graphic 4: Physical Activity Frequency per class

IV. Feature Engineering

1. Label encoding

Since we have several categorical variables, we use label encoding to transform them into numerical variables. Label encoding assigns a unique number to each category.

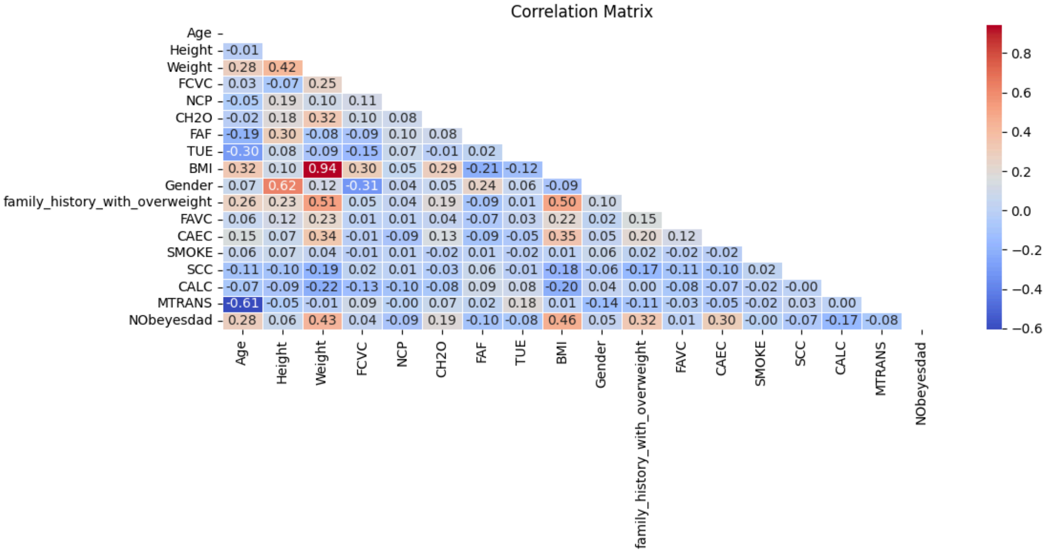
Remark: In the notebook, we evaluated how the encoding model affects the learning rate and results. After trying both One-hot encoding and Label encoding, we made the choice of using label encoding (cf notebook).

2. New BMI variable

We introduced a new variable to the data set , the Body Mass Index (BMI) which stands as a cornerstone in assessing an individual's weight status and overall health. We added this variable as it was calculable (the needed data to do so was available) and we wanted to increase the ability of our model to learn about obesity in general.

$$Body\ Mass\ Index\ (BMI)\ =\ \frac{Weight\ (kg)}{Height^{2}(m)}$$

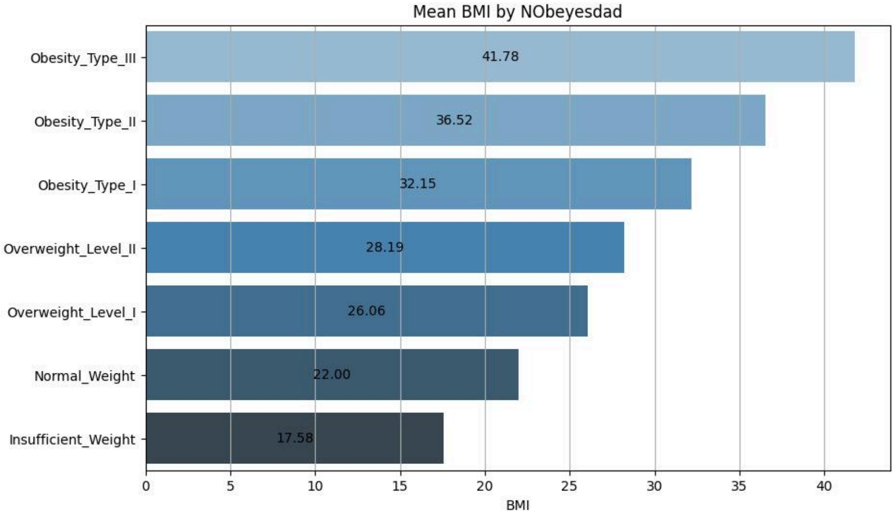
3. Correlation Matrix



	Age	Height	Weight	FCVC	NCP	CH2O	FAF	TUE	BMI	Gender	family_history_with_overweight	FAVC	CAEC	SMOKE	SCC	CALC	MTRANS	NObeyesdad
Age																		
Height	-0.01																	
Weight	0.28	0.42																
FCVC	0.03	-0.07	0.25															
NCP	-0.05	0.19	0.10	0.11														
CH2O	-0.02	0.18	0.32	0.10	0.08													
FAF	-0.19	0.30	-0.08	-0.09	0.10	0.08												
TUE	-0.30	0.08	-0.09	-0.15	0.07	-0.01	0.02											
BMI	0.32	0.10	0.94	0.30	0.05	0.29	-0.21	-0.12										
Gender	0.07	0.62	0.12	-0.31	0.04	0.05	0.24	0.06	-0.09									
family_history_with_overweight	0.26	0.23	0.51	0.05	0.04	0.19	-0.09	0.01	0.50	0.10								
FAVC	0.06	0.12	0.23	0.01	0.01	0.04	-0.07	0.03	0.22	0.02	0.15							
CAEC	0.15	0.07	0.34	-0.01	-0.09	0.13	-0.09	-0.05	0.35	0.05	0.20	0.12						
SMOKE	0.06	0.07	0.04	-0.01	0.01	-0.02	0.01	-0.02	0.01	0.06	0.02	-0.02	-0.02					
SCC	-0.11	-0.10	-0.19	0.02	0.01	-0.03	0.06	-0.01	-0.18	-0.06	-0.17	-0.11	-0.10	0.02				
CALC	-0.07	-0.09	-0.22	-0.13	-0.10	-0.08	0.09	0.08	-0.20	0.04	0.00	-0.08	-0.07	-0.02	-0.00			
MTRANS	-0.61	-0.05	-0.01	0.09	-0.00	0.07	0.02	0.18	0.01	-0.14	-0.11	-0.03	-0.05	-0.02	0.03	0.00		
NObeyesdad	0.28	0.06	0.43	0.04	-0.09	0.19	-0.10	-0.08	0.46	0.05	0.32	0.01	0.30	-0.00	-0.07	-0.17	-0.08	

Graphic 5: Matrix of correlation

According to our correlation matrix (*cf graphic 5*), weight, BMI, family history of overweight, and CAEC (consumption of food between meals) exhibit strong correlations with the target variable. Furthermore, weight and BMI are highly correlated with each other. Hence, it is logical that both weight and BMI would demonstrate strong correlations with the target variable simultaneously.



We can see from the chart below, each class may have the specific mean of BMI while we find the positive correlation between these two variables.

Graphic 5: Mean BMI per each class

V. Models

For this project, we employed two distinct multi-classifier algorithms: LGBM (Light Gradient Boosting Machine) and Random Forest. Below, the summary table outlines the essential characteristics of both models and the rationale behind our choice to utilize them in training our predictive model.

Aspect	**LGBM**	**Random Forest**
Algorithm Type	Gradient Boosting	Ensemble Learning (Bagging)
Performance	Often higher predictive accuracy.	Generally robust and less prone to overfitting.
Training Speed	Faster training speed due to leaf-wise growth	Slower training speed due to growing independent trees

Memory Usage	Lower memory footprint due to histogram-based approach	Higher memory usage due to storing multiple trees
Hyperparameter Tuning	More complex due to larger number of hyperparameters	Simpler due to fewer hyperparameters and less sensitivity
Interpretability	Less interpretable due to complex ensemble structure	More interpretable due to ensemble of decision trees

1. LGBM

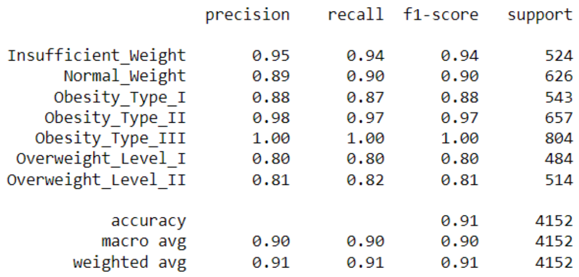
a. Hyperparameter Tuning

We employed Optuna to systematically tune hyperparameters for a LightGBM classifier, aiming to optimize its predictive performance. This involved defining an objective function encompassing various parameters such as learning rate, number of estimators, and regularization terms, and running 50 optimization trials using a Tree-structured Parzen Estimator sampler to identify the best parameter combination maximizing accuracy on the testing data. The resulting best parameters were then utilized to configure the final LightGBM classifier model for improved performance. The best parameters selected by the algorithm were:

{'learning_rate': 0.029360941280932137, 'n_estimators': 512, 'lambda_l1': 0.012912879986094793, 'lambda_l2': 0.03144935519184769, 'max_depth': 13, 'colsample_bytree': 0.3234407387824524, 'subsample': 0.8706657559633919, 'min_child_samples': 42}

b. Results

The classification report provides insights into the performance of our model, which achieved an overall accuracy of **91.02%** on the test dataset. Notably, it demonstrates strong precision, recall, and F1-score for Obesity_Type_II and Obesity_Type_III categories, indicating robust predictions in these classes. However, challenges are observed in accurately classifying Overweight_Level_I, Overweight_Level_II, Obesity_Type_I, and Insufficient_Weight categories, with slightly lower precision, recall, and F1-scores. Overall, the model showcases satisfactory performance across various weight categories, with room for improvement in specific classifications to enhance its predictive capabilities further.

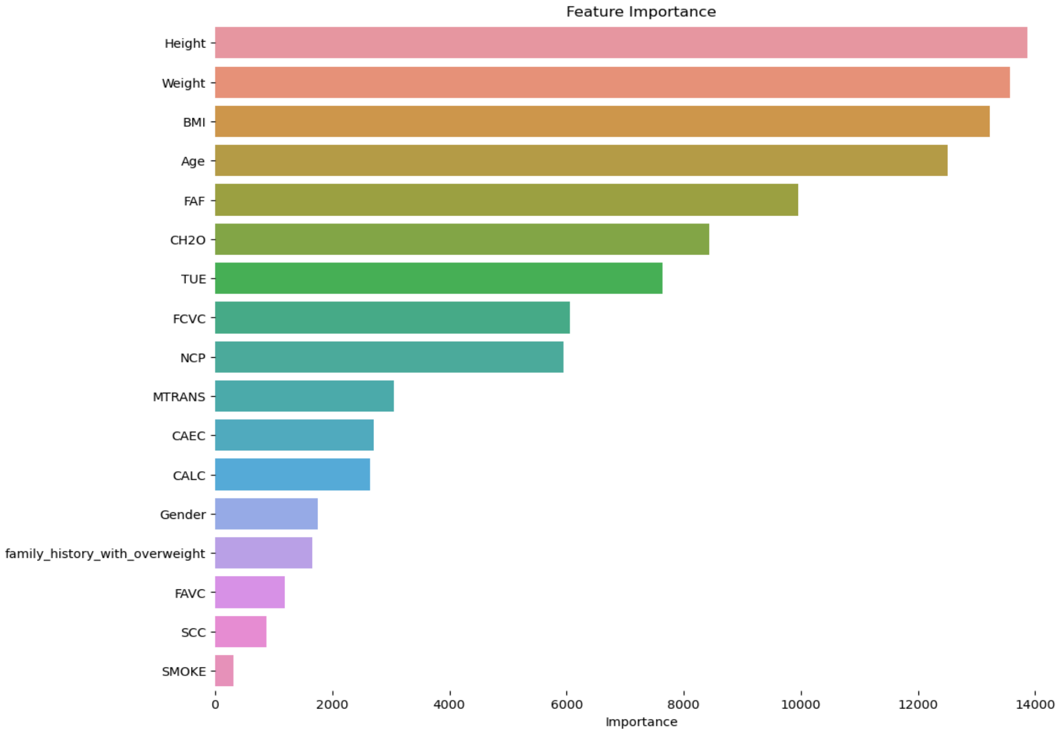


```
                     precision    recall  f1-score   support

Insufficient_Weight       0.95      0.94      0.94       524
      Normal_Weight       0.89      0.90      0.90       626
     Obesity_Type_I       0.88      0.87      0.88       543
    Obesity_Type_II       0.98      0.97      0.97       657
   Obesity_Type_III       1.00      1.00      1.00       804
 Overweight_Level_I       0.80      0.80      0.80       484
Overweight_Level_II       0.81      0.82      0.81       514

           accuracy                           0.91      4152
          macro avg       0.90      0.90      0.90      4152
       weighted avg       0.91      0.91      0.91      4152
```

Table 3: Classification score reports of LGBM model

c. Features importance

We computed the importance of each feature and visualized the results in a bar plot. For the LGBM Model with label encoding, the height , the weight and the BBMI are the most important.

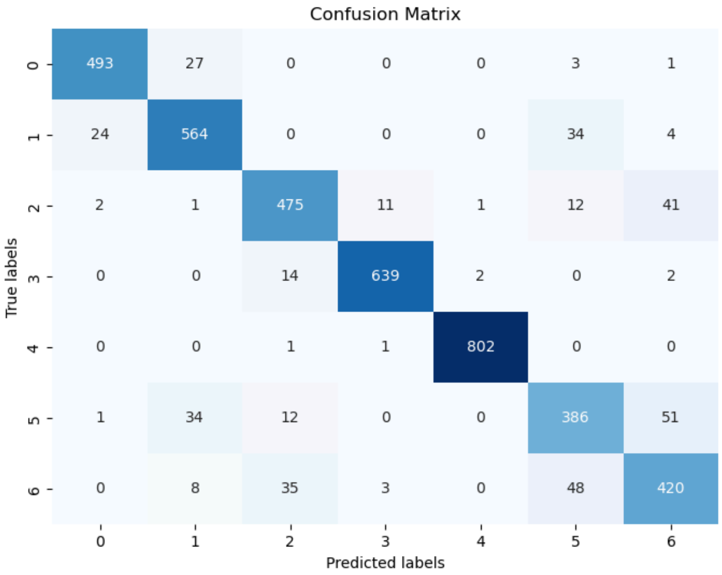


Graphic 6: Features importance in LGBM model

d. Confusion Matrix

Based on the visualization of our model's predictions, we observe that it effectively distinguishes between Obesity_Type_II and Obesity_Type_III classifications. However, it encounters challenges in accurately labeling Overweight_Level_I, Overweight_Level_II, and Obesity_Type_I categories. While the model demonstrates notable accuracy in certain classifications, further refinement is required to enhance its performance across all weight categories. The labels in plot are as followed:

- 0= Insufficient_Weight
- 1= Normal_Weight
- 2= Obesity_Type_I
- 3= Obesity_Type_II
- 4= Obesity_Type_III
- 5= Overweight_Level_I
- 6= Overweight_Level_II



True labels \ Predicted labels	0	1	2	3	4	5	6
0	493	27	0	0	0	3	1
1	24	564	0	0	0	34	4
2	2	1	475	11	1	12	41
3	0	0	14	639	2	0	2
4	0	0	1	1	802	0	0
5	1	34	12	0	0	386	51
6	0	8	35	3	0	48	420

Graphic 7: Confusion matrix of LGBM model

2. Random Forest

a. Hyperparameter Tuning

To optimize our random forest classifier for accuracy, we employed grid search, a systematic technique to find the best hyperparameters. These parameters include the number of trees in the forest, features to consider for splitting, maximum tree depth, and more. Utilizing a parameter grid and 5-fold cross-validation, we exhaustively explored various combinations to find the most effective setup. The final model was trained with the identified optimal hyperparameters. Here are the best parameters obtained:

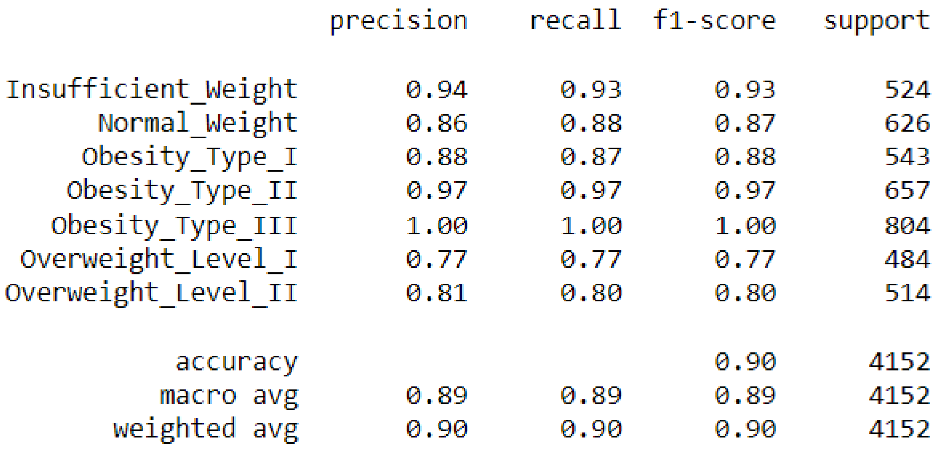
- Number of trees (n_estimators): 150
- Maximum features (max_features): 'sqrt'
- Maximum depth (max_depth): 15
- Minimum samples split (min_samples_split): 2
- Minimum samples leaf (min_samples_leaf): 2
- Bootstrap: False

These parameters were selected to maximize the classifier's accuracy on our dataset:

{'bootstrap': False, 'max_depth': 15, 'max_features': 'sqrt', 'min_samples_leaf': 2, 'min_samples_split': 2, 'n_estimators': 150}

b. Results

The Random Forest classification yielded an overall accuracy of 90.05% on the test dataset comprising 4152 samples. Notably, the model demonstrates strong precision, recall, and F1-score for Obesity_Type_II and Obesity_Type_III categories, indicating robust predictions in these classes. However, it encounters challenges in accurately classifying Overweight_Level_I and Overweight_Level_II, with slightly lower precision, recall, and F1-scores. Overall, the model showcases satisfactory performance across various weight categories, with potential for further refinement to enhance its predictive capabilities, particularly in specific classifications.

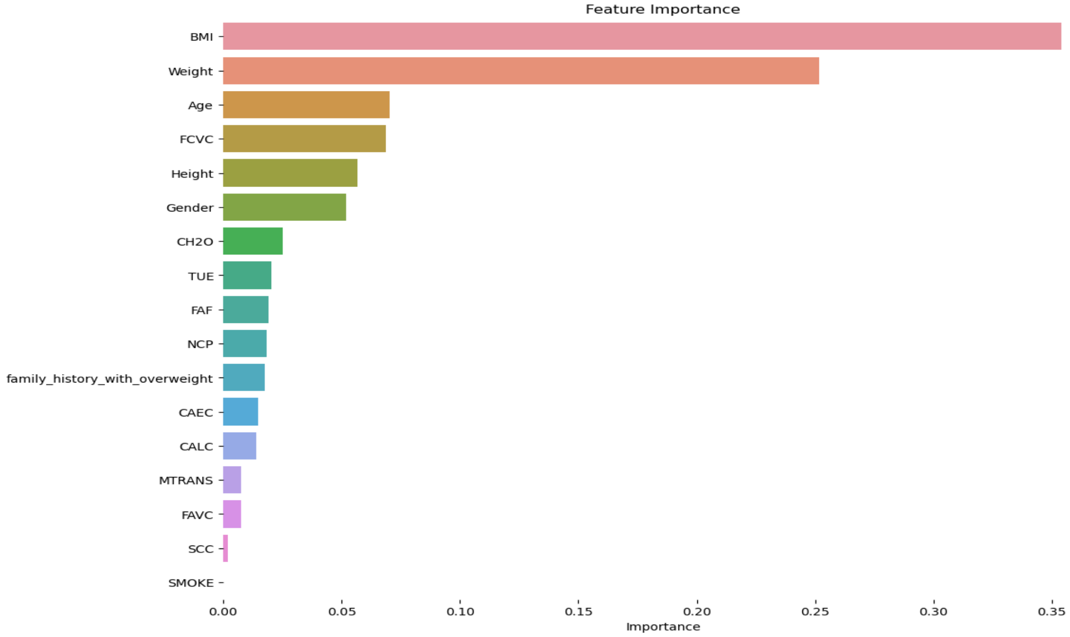


```
                     precision    recall  f1-score   support

 Insufficient_Weight       0.94      0.93      0.93       524
       Normal_Weight       0.86      0.88      0.87       626
      Obesity_Type_I       0.88      0.87      0.88       543
     Obesity_Type_II       0.97      0.97      0.97       657
    Obesity_Type_III       1.00      1.00      1.00       804
  Overweight_Level_I       0.77      0.77      0.77       484
 Overweight_Level_II       0.81      0.80      0.80       514

            accuracy                           0.90      4152
           macro avg       0.89      0.89      0.89      4152
        weighted avg       0.90      0.90      0.90      4152
```

Table 4: Classification score reports of Random Forest model

c. Features importance

We computed the importance of each feature and visualized the results in a bar plot. For this model, the BMI and Weight and age are the most important features for the model.

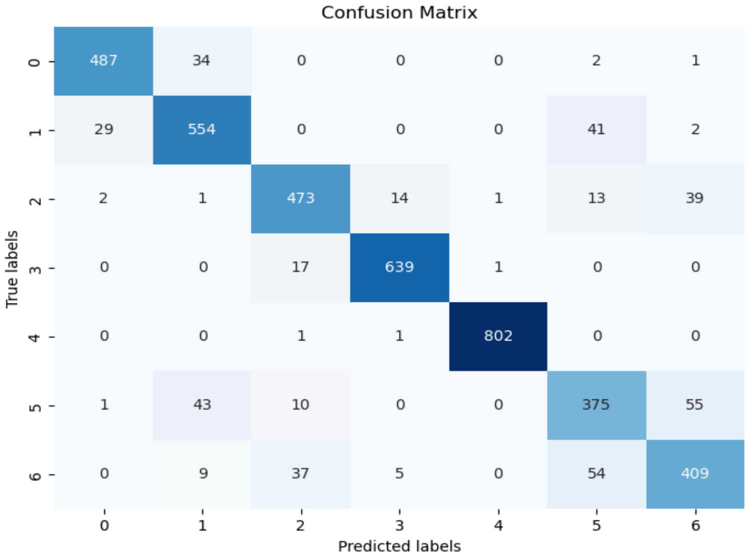


Graphic 8: Features importance in RF model

d. Confusion Matrix

Based on the confusion matrix visualization for the Random Forest model, it effectively predicts Obesity_Type_II and Obesity_Type_III categories. However, it struggles with accurate labeling of Overweight_Level_I, Overweight_Level_II, Obesity_Type_I, and Insufficient_Weight classifications. Despite these challenges, the model demonstrates overall proficiency in predicting certain weight categories, highlighting areas where further refinement may be needed for improved performance.The labels in plot are as followed:

- 0= Insufficient_Weight
- 1= Normal_Weight
- 2= Obesity_Type_I
- 3= Obesity_Type_II
- 4= Obesity_Type_III
- 5= Overweight_Level_I
- 6= Overweight_Level_II



	0	1	2	3	4	5	6
0	487	34	0	0	0	2	1
1	29	554	0	0	0	41	2
2	2	1	473	14	1	13	39
3	0	0	17	639	1	0	0
4	0	0	1	1	802	0	0
5	1	43	10	0	0	375	55
6	0	9	37	5	0	54	409

Graphic 9: Confusion Matrix of the RF model

VI. Comparison of Models

When comparing the results of the LightGBM and Random Forest classifiers, we observe that the LightGBM model outperforms the Random Forest model in terms of overall accuracy, with 91.02% compared to 90.05%, respectively. Notably, both models demonstrate strong precision, recall, and F1-scores for Obesity_Type_II and Obesity_Type_III categories, indicating robust predictions in these classes.

However, our LightGBM model may be susceptible to overfitting, where it learns to memorize the training data rather than generalize well to unseen data. In the case of this challenge, we cannot evaluate the overfitting of the model as the given training data is very small.

VII. Conclusion

While both models showcase satisfactory performance in predicting weight categories, the LightGBM model demonstrates a slight edge over the Random Forest model in terms of overall accuracy and performance metrics. However, it's essential to consider other factors such as computational efficiency and interpretability when selecting the most suitable model for deployment.