# Machine Learning

## Obesity in Focus: data for healthier lives

Obesity type prediction project

### Group14

**Henry Kenneth Lewis – 20222002**

**Yan Sidoryk – 20222004**

**Lowie Edgard E. De Wever – 20231733**

**Phat Ma – 20222001**

**NOVA Information Management School**
**Instituto Superior de Estatística e Gestão de Informação**
Universidade Nova de Lisboa

# ABSTRACT

Obesity is a growing global health concern, requiring accurate models to support prevention and management. This study predicts seven obesity levels, ranging from Insufficient Weight to Obesity Type III, and using a plethora of behavioral and demographic data like age, exercise, caloric intake, and alcohol consumption. We developed machine learning models optimized for performance, and evaluated by the macro F1-score. This project used a dataset containing 1,611 rows and 21 columns. The initial step involved analyzing the data to gain insights and creating visualizations to identify patterns and relationships, then categorical columns were encoded, some numeric columns were standardized, missing values were imputed, and outliers were removed to ensure data quality. New features were formed based on existing columns and insights from external research, all aiming to enhance model performance. Our analysis found that Obesity Type I was the most common category, while Insufficient Weight was the least. Obesity Type III occurred only in females, and Obesity Type II only in males. Females were more likely to have Insufficient Weight, possibly due to societal pressures favoring slimness, while males were more likely to fall into Overweight Level III. Most participants were 20–40, and higher obesity levels correlated with lower physical activity, as expected. BMI showed a strong correlation with all classifications, confirming its predictive importance. Weekly physical activity had the most missing values, requiring careful preprocessing. Logistic Regression, Gradient Boosting, and Random Forest were the top-performing models, with their voting ensemble achieving the best macro F1-score. We believe this study successfully demonstrated the use of machine learning models to predict various obesity levels using behavioral and demographic features. The combination of Logistic Regression, Gradient Boosting, and Random Forest in a voting ensemble proved to be the most effective approach. Key insights included strong correlations between BMI and obesity classifications, as well as the impact of physical activity on health outcomes. Gender-specific trends and age distributions highlighted numerous potential sociocultural and biological factors influencing obesity levels that could be further explored.

# KEYWORDS

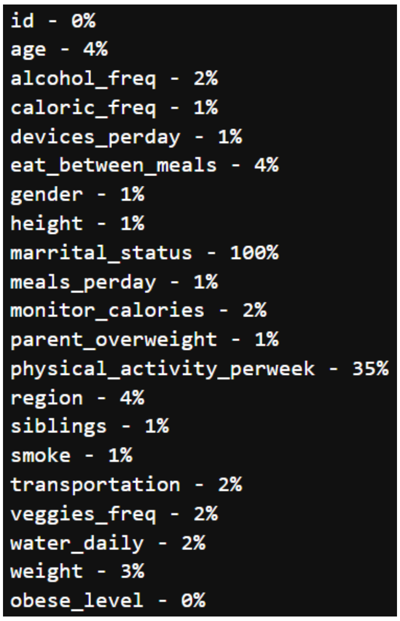
Keyword 1; Obesity Prediction Keyword 2; Machine Learning Keyword 3; Demographic Features
Keyword 4; Physical Activity Keyword 5; Health Data Analytics Keyword 6; Ensemble Learning

## INTRODUCTION

Obesity is a major global health issue and is often linked to serious conditions like heart disease, diabetes, and high blood pressure. Its increasing prevalence makes it even more important to try and develop tools/models that can predict and help manage obesity more effectively. Understanding the key factors behind obesity can aid and allow healthcare professionals to create better strategies for prevention and treatment. This study aims to predict obesity levels across seven categories, from Insufficient Weight to Obesity Type III, using 20 various behavioral and demographic data such as age, weight, height, physical activity, caloric intake, and alcohol consumption. By applying machine learning techniques, we aim to create an accurate model that can help healthcare professionals better classify obesity types, and be able to provide the best advice/solutions based on the results.

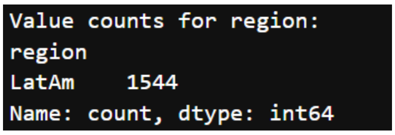
## DATA EXPLORATION

Before training and testing any models, we had to look over the whole dataset in search of patterns/correlations that may give us insights into the potential causes for each classification, so we can then better train our model in a way that fits our findings, here is what we found:



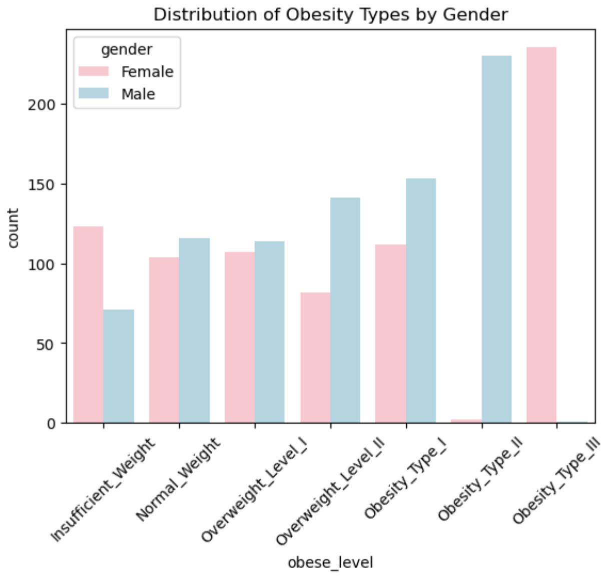
```
id - 0%
age - 4%
alcohol_freq - 2%
caloric_freq - 1%
devices_perday - 1%
eat_between_meals - 4%
gender - 1%
height - 1%
marrital_status - 100%
meals_perday - 1%
monitor_calories - 2%
parent_overweight - 1%
physical_activity_perweek - 35%
region - 4%
siblings - 1%
smoke - 1%
transportation - 2%
veggies_freq - 2%
water_daily - 2%
weight - 3%
obese_level - 0%
```

We saw that the dataset had some missing values across multiple features. For categorical columns without progression, we used mode imputation, while for numeric and progressive categorical columns, we encoded the categorical columns based on the progression, and applied KNN imputation. The most missing values were found in the physical activity per week column (35%), and we found that marital status had 100% missing values, so we decided to drop this column entirely as we also believe that it wouldn't have been a great predictor anyways.

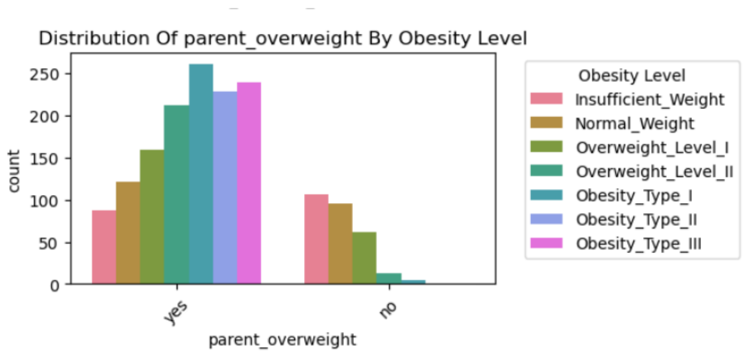


```
Value counts for region:
region
LatAm     1544
Name: count, dtype: int64
```

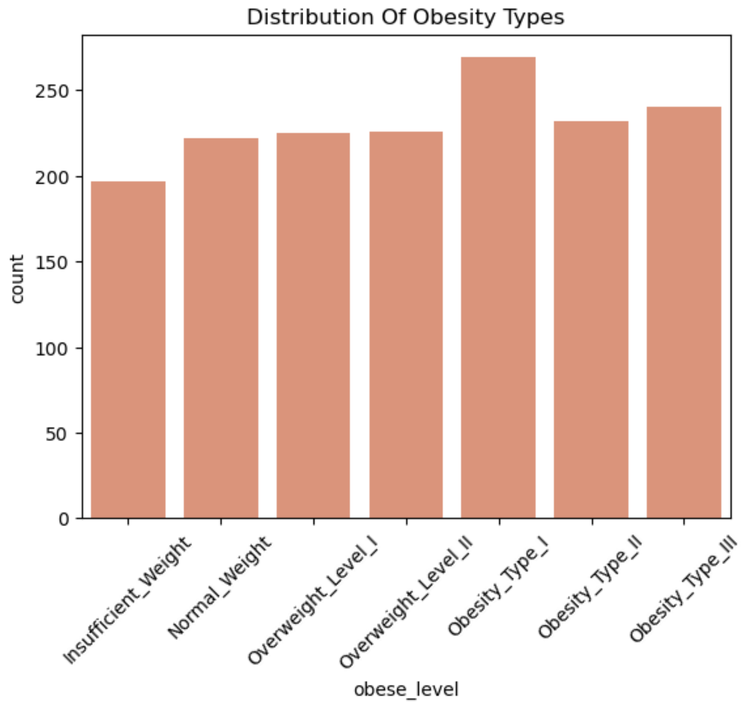
We also found that the region column was the same for each entry, so we removed that from the dataset as it would add no insights to our models.



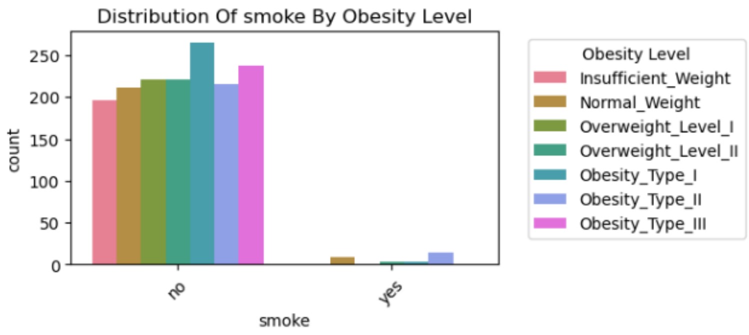
Upon examining the gender distributions, we found that some categories were fairly balanced. However, Obesity Type II and Obesity Type III showed a strong gender skew, with a clear predominance of males and females, respectively. We also observed that more females were found to be suffering with insufficient weight, while more males were categorized in Overweight Level II



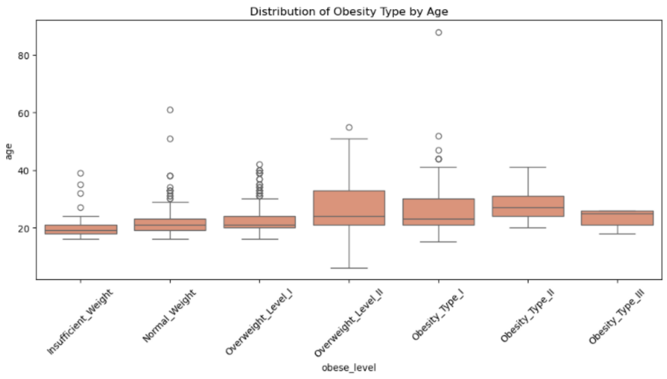
This visualization clearly shows that children of overweight parents are more likely to be overweight. In contrast, when parents are not overweight, their children are far more likely to have a normal weight or insufficient weight.



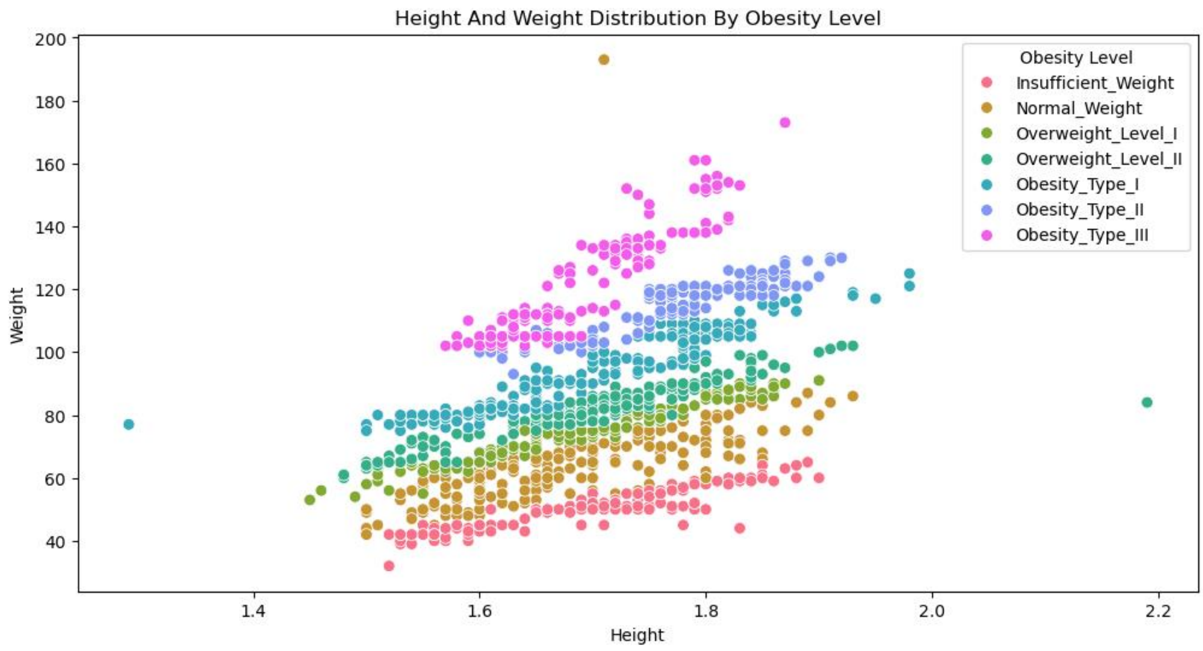
We saw that most of our entries fell in the obesity type I category, this suggested that we should pay special attention to this category and ensure that our model has good performance in this class. We also see that insufficient weight is the least prominent which could make it harder to predict as we have fewer samples to train our model on.



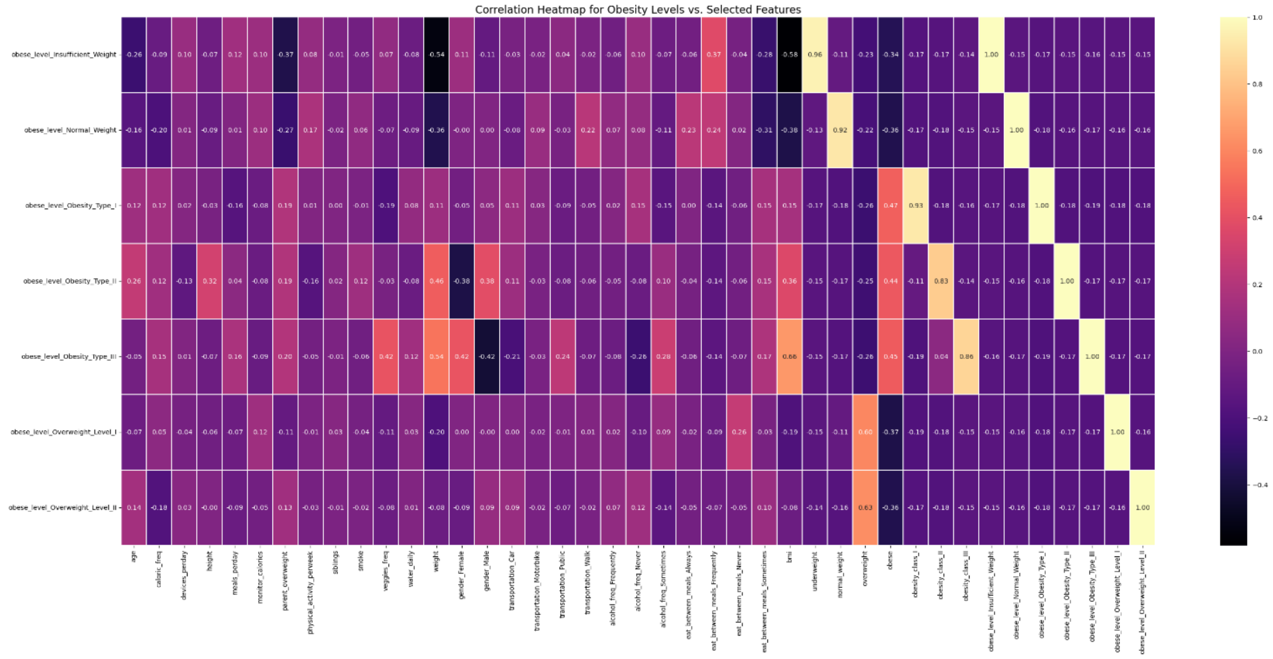
We saw that that most individuals in the dataset did not smoke. However, Obesity Type II was noticeably more common among those who did. This correlation caused our decision to keep smoking as a feature in the model.



The age distributions across all classifications were fairly similar, with some outliers present in each. We chose to retain these outliers, as age does not appear to be a major determining factor, and standardization helps to mitigate their impact in negatively affecting the model.



The weight and height distributions for each category aligned closely with expectations, reflecting BMI's role as a key factor in obesity classification. The categories were generally well clustered, although we identified notable outliers, including an individual weighing nearly 200kg in the normal weight category which we removed due to the large logical imbalance. Despite the presence of height outliers on both extremes, we chose to retain them as they improved model performance. Standardizing the height feature balanced their influence, while the additional data contributed positively to the model's training process.



We used a correlation matrix, along with other feature selection techniques, to ensure that the defined obesity classes and our BMI column were strongly correlated with the target variables. Additionally, this visualization helped us identify individual features to keep or remove, based on the results from our confusion matrix.

# METHODOLOGY

### Model Assessment Strategy

To evaluate the performance of our models, we used the train-test split for the initial scoring of all tested models. We chose this approach to increase the speed of the outputs, as we were working with complex models and wanted to be able to test new ideas and evaluate results quickly. While tuning individual models, we also checked cross-validation scores to evaluate their individual performances after optimization. For the final voting classifier, we used a learning curve with cross-validation to analyze variance and check for overfitting. Additionally, we assessed performance using the F1-score to ensure consistency with the competition's scoring system, supported by a classification report and a confusion matrix. These tools helped identify areas where the model required refinement, both overall and for individual class labels.

### Data Preprocessing

Data preprocessing involved multiple steps to prepare the dataset for modeling:

- Outlier Removal: The normal weight outlier was removed to improve model performance.
- Column Removal:
  - The region column was dropped as it contained only one unique value.
  - The marital status column was dropped due to being completely empty.
- Missing Data Handling:
  - Categorical columns without progression were imputed using the most common value.
  - Numerical and progression-encoded columns were imputed using KNN imputation.
- Encoding:
  - Categorical columns with clear progression were encoded on a scale from 0 to their respective maximum values.
  - Binary columns (e.g., yes/no) were converted to 1s and 0s.
  - Categorical columns imputed with the most common value were one-hot encoded.
- Feature Engineering:
  - A BMI column was added to provide a calculated feature for categorization and prediction.
  - Seven custom obesity classes, based on BMI ranges, were created and incorporated into the dataset.

**Feature Selection Strategy**

To identify the most relevant features, we applied:

- Recursive Feature Elimination (RFE) to identify top-performing predictors.
- Mutual Information Classification to measure the dependency between features and the target variable.
- A correlation matrix to detect and manage multicollinearity among features.

These techniques ensured that only the most impactful features were retained, improving model interpretability and performance. We also looked at the feature importances of the final random forrest and gradient boosting models to see the best perfoming models in the voting classifier (as you can't directly show that using feature importances, which is the same with the logistic regression model)

**Predictive Algorithms**

We selected a voting classifier as our final model, combining three algorithms:

- Gradient Boosting Classifier
- Random Forest Classifier
- Logistic Regression

These models were selected based on their strong individual macro f1 scores compared to other tested algorithms. By combining them in a voting classifier, we leveraged the strengths of each model, achieving an ensemble with superior overall performance. We also experimented with different weights before deciding on using ones that favored the gradient boosting model, as this yelled us better results.

**Model Optimization**

To fine-tune our models, we undertook a systematic optimization process:

- RandomizedSearchCV was used to identify a broad range of optimal hyperparameters.
- GridSearchCV further refined the hyperparameters based on the results from the randomized search.
- Voting Classifier Weights, We adjusted the voting model's weights to favor the Gradient Boosting Classifier, as it consistently delivered better results.

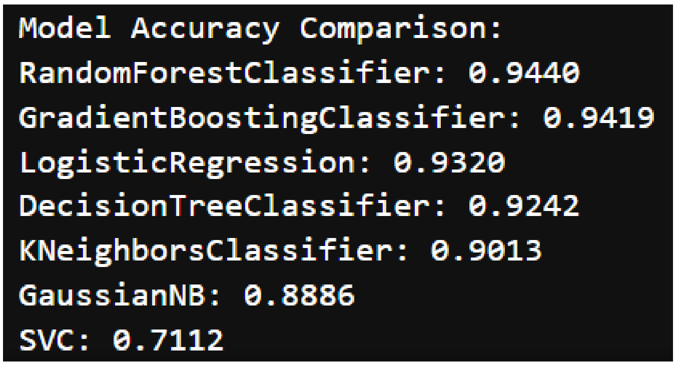
These efforts ensured that each individual model within the ensemble, as well as the overall voting classifier, was fully optimized for performance.

## RESULTS

The results of this study highlight the effectiveness of machine learning in predicting obesity levels based on the provided dataset. Below, we will summarize the key findings and provide supporting statistics, tables, and figures to present the data concisely.

### Initial Model Performance Evaluation

The initial evaluation of individual models using the train-test split of the training data demonstrated each algorithm's relative performance. The table below displays the F1 scores for the tested models, where they are all ranked by their macro F1 score on the test set.



```
Model Accuracy Comparison:
RandomForestClassifier: 0.9440
GradientBoostingClassifier: 0.9419
LogisticRegression: 0.9320
DecisionTreeClassifier: 0.9242
KNeighborsClassifier: 0.9013
GaussianNB: 0.8886
SVC: 0.7112
```

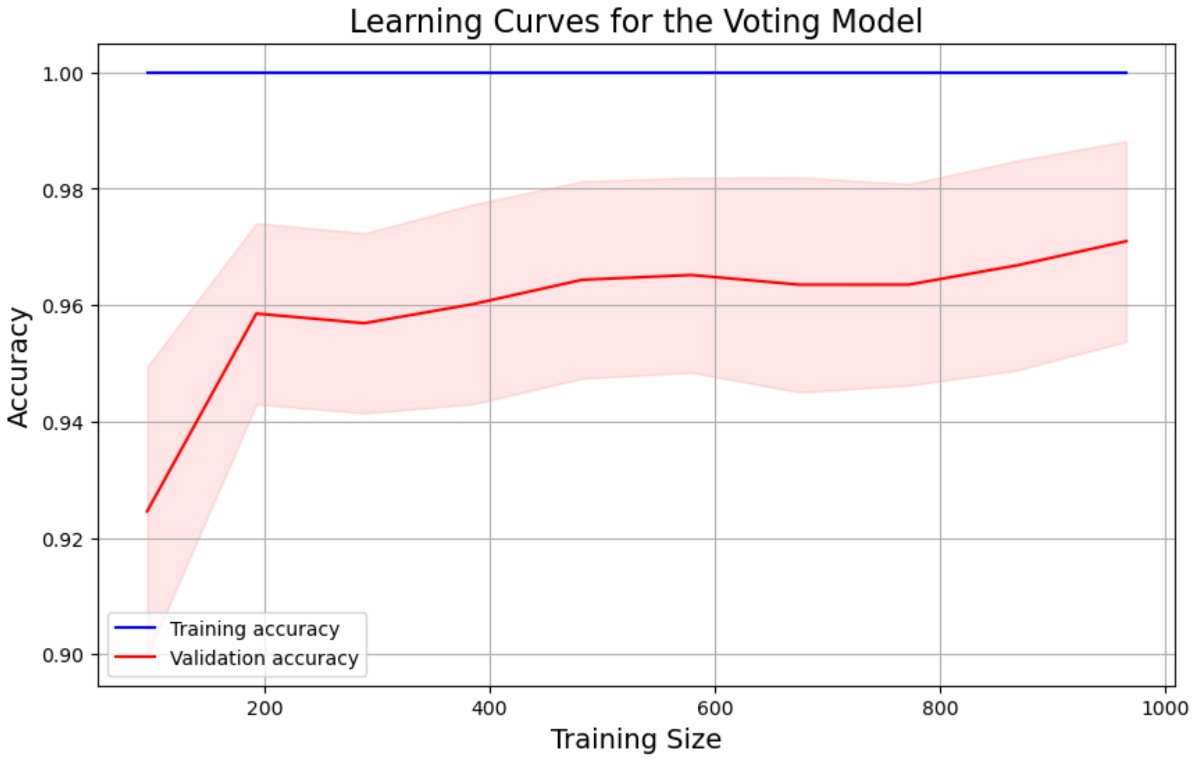
From this analysis, we see that the RandomForrestClassifier emerged as the best-performing individual model without tuning, followed closely by the GradientBoostingClassifier, with the LogisticRegression model finishing the top three. These are the models we selected to move forward with to tune and eventually group together in an ensable model.

### Voting Classifier Performance

After selecting the top-performing models, we built the final voting classifier using GradientBoostingClassifier, RandomForestClassifier, and LogisticRegression, with weights [2,1,1]. These weights favor the GradientBoostingClassifier due to its strong individual performance shown in cross-validation scores after tuning. We used a learning curve with cross-validation (shown below) to evaluate the voting model's performance and check for overfitting and variance. The curve plots training accuracy (blue) and validation accuracy (red) against training set size, with shaded areas showing score variability.

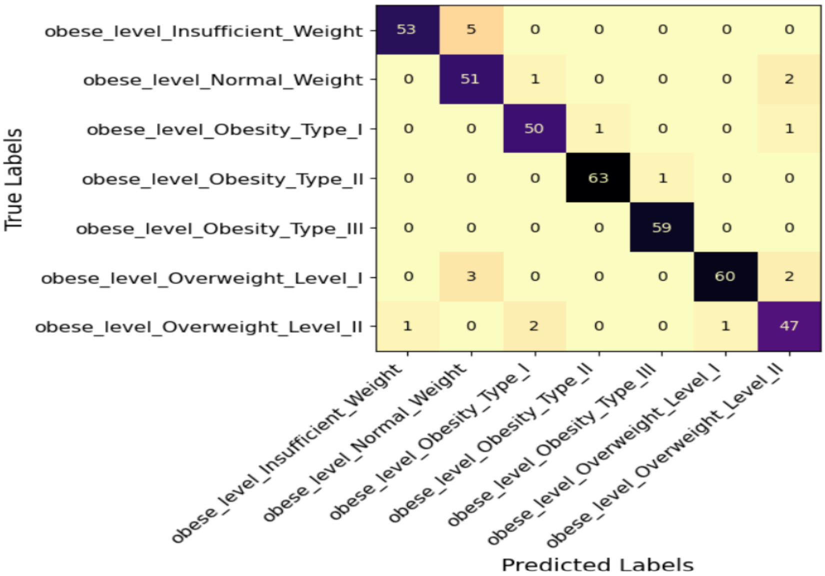
Key findings:

- Training accuracy stays consistently high, which could suggest strong overfitting, but this can be expected for complex models in this situation
- Validation accuracy improves with training size and eventually aligns closely with training accuracy, showing good generalization.
- The small variance in scores across cross-validation folds confirms consistent predictions.



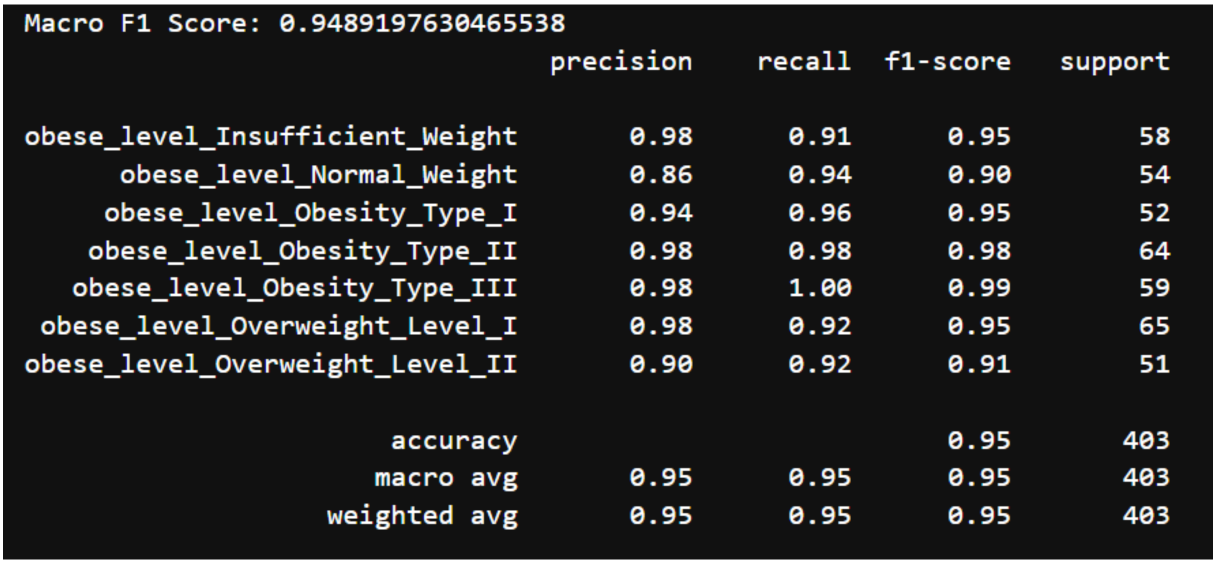
### Confusion Matrix and Class-Specific Insights

To better understand the performance across individual obesity categories, we used a confusion matrix and classification report. The macro F1 score achieved by the voting classifier was **0.95** on the test set. We generated a ConfusionMatrixDisplay to gain a better understanding on where our model is making incorrect predictions.



From this display, we observe that most of our misclassifications occur when predicting Normal Weight. However, these errors typically result in predictions of Insufficient Weight or Overweight Level I, which are neighboring classifications to Normal Weight. This pattern is expected given the small training set, as neighboring classes often share overlapping features. Especially in this instance.

We also produced a classification report to gain a deeper understanding on the predictive capabilities of our final model on each obesity type. As shown below.



```
Macro F1 Score: 0.9489197630465538
                                precision    recall  f1-score   support

obese_level_Insufficient_Weight       0.98      0.91      0.95        58
      obese_level_Normal_Weight       0.86      0.94      0.90        54
     obese_level_Obesity_Type_I       0.94      0.96      0.95        52
    obese_level_Obesity_Type_II       0.98      0.98      0.98        64
   obese_level_Obesity_Type_III       0.98      1.00      0.99        59
  obese_level_Overweight_Level_I       0.98      0.92      0.95        65
 obese_level_Overweight_Level_II       0.90      0.92      0.91        51

                        accuracy                          0.95       403
                       macro avg       0.95      0.95      0.95       403
                    weighted avg       0.95      0.95      0.95       403
```

From the results of the report, we see that:

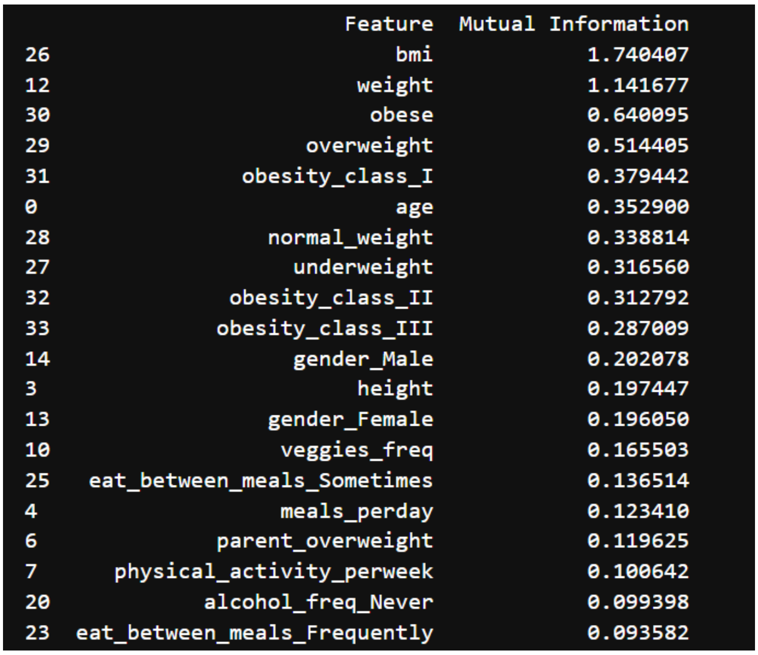
**-** Insufficient Weight, Overweight Level I, Obesity Type I, Obesity Type II, and Obesity Type III achieved particularly high scores, with F1-scores near or above 0.95, reflecting great classification performance in these categories.

- Normal Weight shows slightly lower precision (0.86) compared to other classes, with misclassifications occurring in neighboring categories, as we saw in the display before, such as Insufficient Weight or Overweight Level I. This could be due to the low amount of data we have for Normal Weight.

- Overweight Level II shows underperformance too in comparison to the overall score, we see that this class often got confused with other obese/overweight levels, but also insufficient weight in one instance, implying that it shares some similarities with a lot of the different obesity types.

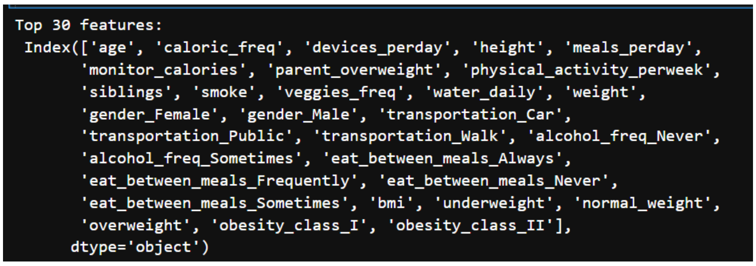
**Feature Importance Analysis**

We used a plethora of feature importance metrics to select the optimal features to use to maximize performance such as: recursive feature elimination (RFE), feature importances of tuned models, and mutual information analysis which all revealed that the most important features for prediction included BMI, weight, meals per day, exercise frequency, and caloric intake. They also all show that the obesity classifications we made are very important predictors. These findings align with domain knowledge about obesity factors.



```
                          Feature  Mutual Information
26                            bmi            1.740407
12                         weight            1.141677
30                          obese            0.640095
29                     overweight            0.514405
31                obesity_class_I            0.379442
0                             age            0.352900
28                  normal_weight            0.338814
27                    underweight            0.316560
32               obesity_class_II            0.312792
33              obesity_class_III            0.287009
14                    gender_Male            0.202078
3                          height            0.197447
13                  gender_Female            0.196050
10                   veggies_freq            0.165503
25     eat_between_meals_Sometimes            0.136514
4                     meals_perday            0.123410
6                parent_overweight            0.119625
7         physical_activity_perweek            0.100642
20               alcohol_freq_Never            0.099398
23    eat_between_meals_Frequently            0.093582
```

Figure shows the top 20  mutual information findings between the features and the targets.



```
Top 30 features:
 Index(['age', 'caloric_freq', 'devices_perday', 'height', 'meals_perday',
       'monitor_calories', 'parent_overweight', 'physical_activity_perweek',
       'siblings', 'smoke', 'veggies_freq', 'water_daily', 'weight',
       'gender_Female', 'gender_Male', 'transportation_Car',
       'transportation_Public', 'transportation_Walk', 'alcohol_freq_Never',
       'alcohol_freq_Sometimes', 'eat_between_meals_Always',
       'eat_between_meals_Frequently', 'eat_between_meals_Never',
       'eat_between_meals_Sometimes', 'bmi', 'underweight', 'normal_weight',
       'overweight', 'obesity_class_I', 'obesity_class_II'],
      dtype='object')
```

Figure shows the top 30 features found using the RFE (recursive feature elimination)

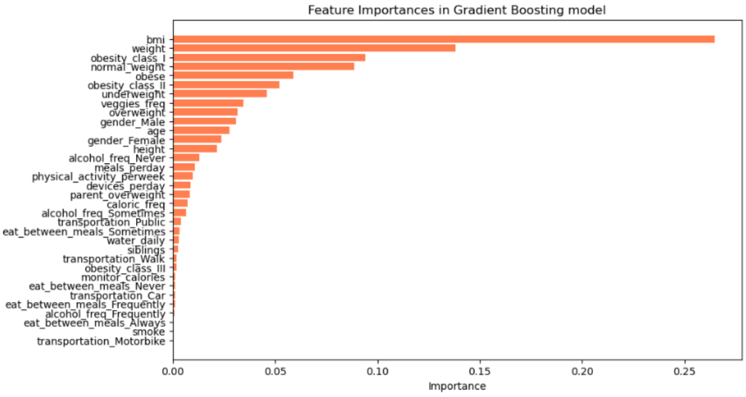


Figure shows feature importances of the tuned GradientBoostingClassifier

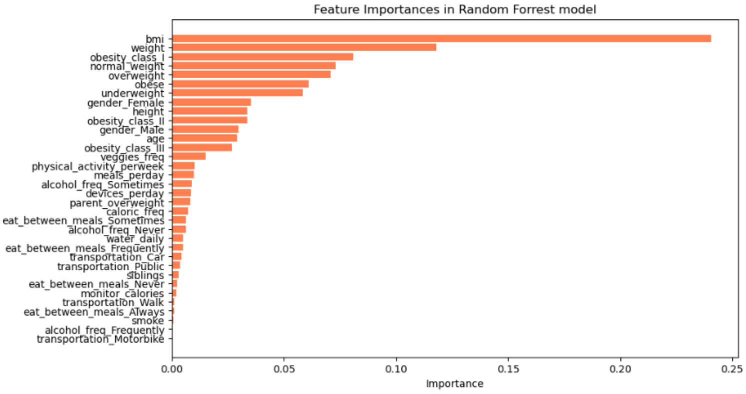


Figure shows feature importances of the tuned RandomForrestClassifier

## DISCUSSION

The findings of this study demonstrate the effectiveness of using a machine learning model for predicting obesity levels. We used the model to achieve a high Macro F1-Score of 0.95, indicating accurate, balanced performance across all classes, including underrepresented ones. It aligns perfectly with prior research into the topic, as all of the high-performing features used in our model are often cited as being highly correlated factors in cases of obesity.

Notably, the learning curve analysis highlighted that while the training accuracy remained stable and high, the validation accuracy steadily improved with increased training size, and was still seen to be following an upward trend at the end of the testing. This behavior suggests that the model is well-generalized and capable of learning from more data without overfitting and that if it were trained on a larger dataset it could have been even more accurate.

Despite these successes, some limitations were observed. Misclassifications occurred primarily for "Normal Weight," often being confused with neighboring classes like "Insufficient Weight" or "Overweight Level I." These errors are to be expected as these classifications are so close together in terms of correlated features and BMI results.

## CONCLUSION

To conclude, our findings align closely with established research on obesity, especially regarding the impact of physical activity, caloric intake, and BMI. For instance, studies such as the UK Biobank research demonstrate the critical role of physical activity in reducing obesity risk, even for those with genetic predispositions. This aligns with our model's identification of physical activity being one of the most useful features when predicting the obesity type of an individual. The ability to capture these relationships through machine learning validates the model’s accuracy and relevance.

As obesity is a heavily researched and well-understood field, our results were consistent with what we expected from the start of this project. The high performance of our model, with a macro F1-score of 0.95, underlines how effective machine learning tools can be in predicting obesity classifications. This highlights a significant opportunity for healthcare professionals to incorporate predictive models like ours into their normal routines. By doing so, they could provide specially tailored advice and interventions to individuals based on their predicted classifications, addressing specific risk factors like a lack of physical exercise, excessive caloric intake, or daily water intake. Moreover, the potential for integrating these models online opens up so many avenues for accessible, remote consultations. Individuals could receive accurate obesity classifications and personalized recommendations without the need for in-person testing. This would allow a lot more people to seek help and advice that otherwise wouldn't have been able to.

The widespread implementation of models like ours could be used to transform public health efforts, offering scalable, accessible, and cost-effective solutions to combat obesity. By healthcare professionals having access to these tools, we can move toward proactive, data-driven approaches that can help empower individuals to make healthier choices and reduce the long-term mental and physical risks associated with obesity-related conditions.

## REFERENCES

[1] Beenish Masood [A,✉], Myuri Moorthy [B]. Causes of obesity: a review, 2023. https://pmc.ncbi.nlm.nih.gov/articles/PMC10541056/

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa