Human-centered Machine Learning 2024 @ UU

INFOMHCML Teaching Team
{d.p.nguyen,h.kaya,y.du,g.sogancioglu}@uu.nl
Utrecht University
Utrecht, the Netherlands

ABSTRACT

Obesity is a significant public health challenge globally, with a rising prevalence in various countries. Understanding the underlying factors contributing to different obesity levels is crucial for developing effective interventions. This project explores various explainability methods applied to different machine learning models to identify key features influencing obesity levels in individuals.

We implemented Decision Trees, Random Forests, LightGBM, RuleFit, Neural Networks, and Partial Dependence Plots to analyze the data. Our approach included plotting feature importances, visualizing model structures, and pruning the depth of trees for the Decision Trees and Random Forests. Additionally, we generated counterfactual explanations to provide insights into how slight changes in input features could lead to different obesity outcomes.

The results indicate that [fill in the key results and findings for each model]. For instance, the Decision Trees revealed that [specific key features] were the most influential factors. Similarly, the Random Forests and LightGBM models identified [other key features]. RuleFit provided a rule-based interpretation, highlighting [specific rules and features]. The Neural Network's analysis showed [specific insights], while the Partial Dependence Plots demonstrated the relationship between [particular features] and obesity levels. Counterfactual explanations highlighted that changes in [specific features] could potentially reduce obesity levels, offering actionable recommendations for individuals and policymakers.

By leveraging these diverse explainability methods, we provide actionable insights into the critical factors contributing to obesity. These findings can inform targeted interventions and policies aimed at improving health outcomes in these populations.

KEYWORDS

machine learning, fairness, explainability

ACM Reference Format:
INFOMHCML Teaching Team. 2024. Human-centered Machine Learning 2024 @ UU. In *Proceedings of Utrecht University (INFOMHCML'2023).* Utrecht University, 6 pages.

1 INTRODUCTION

We need to say the problem statement too.

Obesity is a significant public health challenge globally, with rising prevalence in various countries. Understanding the underlying factors contributing to different obesity levels can aid in developing effective interventions. This project aims to explore various explainability methods with different machine learning models to identify key features influencing obesity levels in individuals from Colombia, Peru, and Mexico. By doing so, we hope to provide actionable insights into the changes needed to improve health outcomes.

INFOMHCML'2023, April 2023, Utrecht, the Netherlands

ACM ISBN xxxxxxx.
https://doi.org/xxxxxxx

2 METHODS

2.1 Dataset Description

. The dataset used for this project is the 'Estimation of obesity levels based on eating habits and physical condition' from the UCI Machine Learning Repository [1]. It includes data on individuals' eating habits, physical conditions, and obesity levels. Each person is classified to one of the seven different obesity levels such as Overweight Level I or Normal Weight. The dataset comprises 17 attributes such as age, gender, family history of overweight, frequency of consumption of high-caloric food, physical activity frequency, and technology usage during meals.

. The dataset variable table can be found on the appendix (2).

2.2 Data Preprocessing

. To enhance the robustness and interpretability of our models, we made a key methodological adjustment in how we handle the target variable and input features. Initially, the project intended to predict obesity levels directly. However, we opted to use Body Mass Index (BMI) as the predictor variable instead. This decision was driven by the need for standardization and consistency. BMI is a widely recognized and standardized measure of body fat, calculated as weight in kilograms divided by height in meters squared (kg/m²). Using BMI allows for a consistent comparison across individuals regardless of their height and weight.

2.2.1 BMI as Predictor. To enhance the interpretability and robustness of our models, we decided to use Body Mass Index (BMI) as the predictor variable instead of the original obesity level categories provided in the dataset. BMI is a widely recognized and standardized measure of body fat, calculated as weight in kilograms divided by height in meters squared. By using BMI, we can maintain consistency in our comparisons across individuals, regardless of their height and weight differences. Consequently, we removed the original weight and height attributes from our input features to avoid redundancy and potential multicollinearity issues.

For our analysis, we categorized BMI into four distinct groups based on widely accepted health guidelines [5]:

- **Underweight**: BMI less than 18.5
- **Normal weight**: BMI between 18.5 and 24.9
- **Overweight**: BMI between 25 and 39.9
- **Obese**: BMI 40 and above

These categories allow us to group individuals into meaningful segments that reflect different health conditions and risk levels associated with their body mass.

2.2.2 Features and Encoding. The dataset includes several binary and categorical attributes that required appropriate encoding to ensure they could be effectively utilized in our models. Binary attributes, such as gender, family history of overweight, and others, were encoded into binary format (e.g., yes/no converted to 1/0). Categorical attributes with multiple levels, such as frequency of alcohol consumption or type of transportation, were transformed into separate binary columns for each category level.

2.3 Models Used

Talk about all the models we used starting from the black box one and their explainability methods (e.g PDP).

2.3.1 Feed-Forward Neural Network. We employed a Multilayer Perceptron (MLP) for the classification task. The MLP architecture consists of an input layer, two hidden layers, and an output layer. The details of each layer are summarized in Table 1.

Layer	Attributes
Input Layer	Receives input features
First Hidden Layer	64 neurons, ReLU activation
Second Hidden Layer	32 neurons, ReLU activation
Output Layer	4 Neurons, softmax activation

Table 1: Attributes of the Multilayer Perceptron (MLP) model.

The model was built using TensorFlow and Keras libraries, and compiled with the Adam optimizer (learning rate of 0.01) and the *sparse_categorical_crossentropy* loss function.

- **Data Preparation**: Features were scaled, and target labels were encoded.
- **Training Configuration**: The model was trained for 50 epochs with a batch size of 32, using 20% of the training data for validation.
- **Compilation**: The model was compiled with the Adam optimizer and *sparse_categorical_crossentropy* loss function.
- **Training Execution**: The model was fitted to the training data, and validation accuracy was monitored to prevent overfitting.

2.3.2 LightGBM. In addition to the Multilayer Perceptron (MLP), we also employed LightGBM, a gradient boosting framework that uses tree-based learning algorithms. LightGBM is categorized as a black-box model due to its complexity and the difficulty in directly interpreting its internal workings. However, it offers several advantages that motivated our choice.

LightGBM is known for its efficiency and speed, particularly with large datasets, making it suitable for our analysis. Furthermore, it provides built-in functions to visualize feature importance based on 'split' and 'gain' metrics. The 'split' importance shows how many times a feature is used in the decision-making process, while the 'gain' importance measures the contribution of a feature to the model's accuracy.

2.3.3 Decision Trees. In our analysis, we also utilized the Decision Tree model. Decision Trees are inherently interpretable, making them an excellent choice as one of our baseline models. This model type provides a clear and straightforward way to understand the relationships between features and the target variable, which is crucial for our goal of identifying key factors influencing obesity levels.

We leveraged the built-in feature importance graphs to identify the most significant features in our dataset. Additionally, we visualized the tree structure to gain deeper insights into the decision-making process of the model. However, the initial decision tree had a high depth due to the large number of features, which made it complex and challenging to interpret.

To address this, we decided to prune the tree to a maximum depth of 5. This pruning significantly increased the interpretability of the model while still maintaining a reasonable level of accuracy. By simplifying the tree, we could better understand and communicate the key factors contributing to obesity, making the Decision Tree an invaluable tool in our analysis.

2.3.4 RuleFit. We also employed RuleFit, which combines the predictive power of ensemble methods with the interpretability of linear models. RuleFit generates a sparse linear model consisting of rules derived from decision trees, offering both robustness and clarity.

The motivation for choosing RuleFit is its ability to produce understandable rules that describe the data, providing actionable insights into factors influencing obesity levels. RuleFit handles high-dimensional data well and highlights feature interactions often missed by linear models alone.

By extracting and fitting rules into a linear model, RuleFit blends interpretability and accuracy, allowing us to visualize and understand complex relationships within the data effectively. This makes RuleFit a valuable tool for analyzing obesity-related factors.

2.3.5 EBM Classifier. Lastly, we utilized the Explainable Boosting Machine (EBM) classifier to obtain both local and global explanations for our predictions. The EBM is a type of Generalized Additive Model (GAM) that combines the interpretability of linear models with the flexibility of decision trees, making it particularly suitable for our analysis.

The motivation for choosing EBM lies in its capability to provide clear insights into model behavior. For local explanations, we examined samples showing the highest prediction errors, allowing us to understand and interpret specific misclassifications. For global explanations, we used EBM to assess overall feature importance, helping us identify the most influential factors affecting obesity levels.

3 RESULTS

3.1 Black-box models

3.1.1 Feed-Forward Neural Network.

Partial Dependence Plot (PDP).

3.1.2 LightGBM.

- The LightGBM achieved an accuracy of 0.87

Feature Importances. To understand the importance of each feature in our dataset, we utilized LightGBM's built-in functions to plot feature importances based on 'split' and 'gain' metrics.

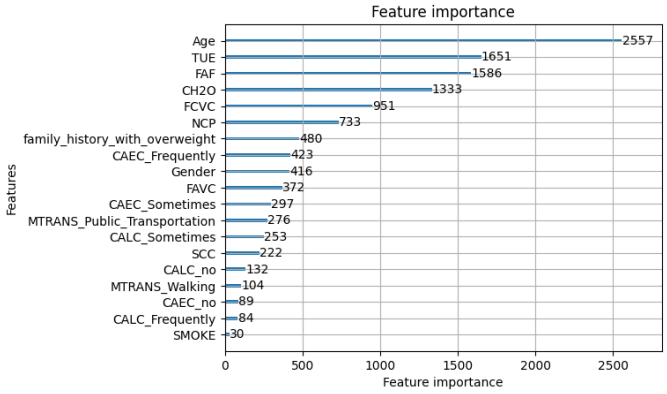

Figure 1: Feature importance based on the number of times a feature is used in the decision-making process (split importance).

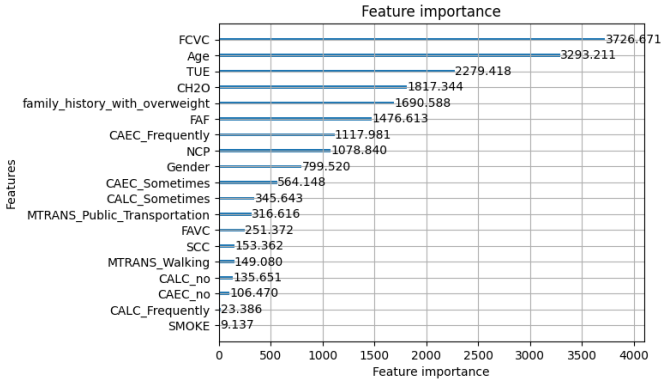

Figure 2: Feature importance based on the contribution of a feature to the model's accuracy (gain importance).

Split Importance. : The plot (Figure 1) shows that 'Age', 'TUE', and 'FAF' are among the most frequently used features in the decision-making process. This indicates that these features are the most important in the structure of the LightGBM model.

Gain Importance. : The plot (Figure 2) reveals that 'FCVC', 'Age', and 'TUE' significantly contribute to the model's accuracy. The higher gain values for these features suggest that they provide substantial information for predicting BMI category levels.

3.1.3 *Random Forest.*

- The Random Forest model achieved an accuracy of 0.886.

The feature importance plot (Figure 3) shows that 'FCVC', 'Age', and 'TUE' are among the most significant features in the model. This indicates that these features play a crucial role in predicting

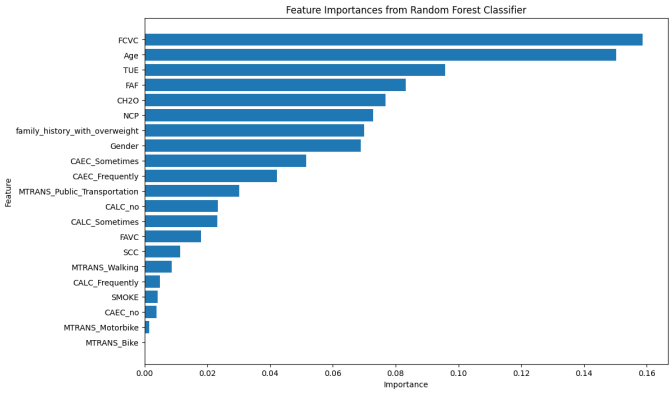

Figure 3: Feature importance from the Random Forest Classifier.

the BMI levels. Other important features include 'FAF', 'CH2O', and 'NCP', which also contribute significantly to the model's predictions. Understanding these feature importances helps us to identify key factors influencing obesity levels and can guide targeted interventions.

3.2 Intrinsically Interpretable Models

3.2.1 *Decision Tree.*

- The Decision Tree model achieved an accuracy of 0.813 with the max depth of 16.

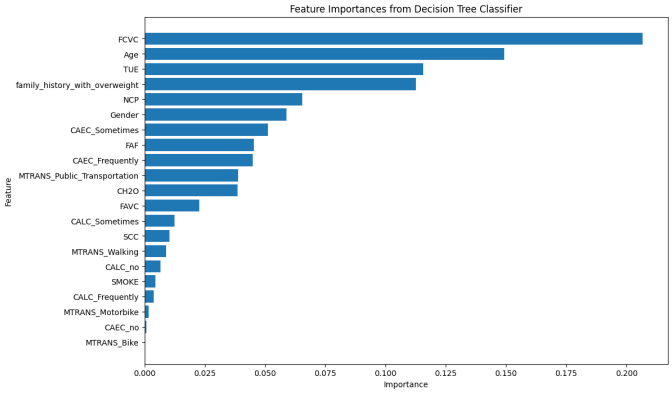

Figure 4: Feature importance from the Decision Tree Classifier.

The feature importance plot (Figure 4) shows that 'FCVC', 'Age', and 'TUE' are among the most significant features in the Decision Tree model. This indicates that these features play a crucial role in predicting obesity levels. Other important features include 'family_history_with_overweight', 'NCP', and 'Gender', which also contribute significantly to the model's predictions.

Pruning. To enhance the interpretability of the Decision Tree, we pruned the tree to a maximum depth of 5. Pruning helps in reducing the complexity of the model and makes it easier to understand the

decision-making process. However, this simplification comes at a slight cost to accuracy. Given the increased interpredability the trade-off is minimal to none.

Experiment Results After Pruning:

- The pruned Decision Tree model achieved an accuracy of 0.801.

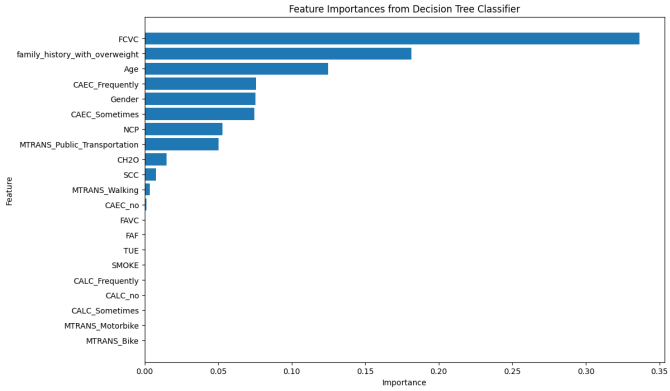

Figure 5: Feature importance from the pruned Decision Tree Classifier (max depth = 5).

The feature importance plot after pruning (Figure 5) reveals that 'FCVC', 'family_history_with_overweight', and 'Age' remain the most significant features. However, their relative importance has been adjusted to reflect the simplified structure of the pruned tree. This pruning process makes the model more interpretable while still maintaining a reasonable level of accuracy, thereby balancing complexity and interpretability.

3.2.2 Rulefit. I think for rulefit we won't say much. I guess we will probably say that because the rules are a lot we don't want it (idk if we can make rules less)

3.2.3 Explainable Boosting Machine (EBM). Something with global and local could be put as paragraphs or something

3.3 Counterfactuals

DiCE library [3]

3.4 Discussion/Conclusion

discuss

3.4.1 Limitations. discuss

4 ETHICAL CONSIDERATIONS

In the development and implementation of our explainable AI models, ethical considerations were taken into account to ensure the integrity and ethical soundness of the project.

One of the major considerations in AI research is that of guaranteeing the privacy of the participants through anonymization of the data. However, since the chosen dataset already provides the data completely anonymized, no further steps were required for this project.

Another important ethical concern is ensuring fairness and eliminating possible biases. Since the dataset used does not include sensitive attributes and only contains attributes that are widely considered to directly affect the probability that an individual is overweight, these concerns are mitigated.

However, the study of counterfactuals throughout the project highlights the need for a thorough investigation by domain experts, given the possibility of producing results that can have a negative impact on a participant's health. For example, when studying counterfactuals, we identified situations where the model recommends the user to start smoking. While this recommendation makes sense in the scope of the project, since smoking can lead to weight loss, it should not be a viable recommendation, given how detrimental this can be to the participant's overall health.

Finally, concerns around transparency are eliminated given the nature of the project, which specifically intends to provide a transparent view of the predictions in order to identify the recommended lifestyle changes that will reduce the risk of obesity and promote healthier living. This approach ensures that users can understand the reasoning behind the AI's recommendations, fostering trust and facilitating informed decision-making.

5 RELATED WORK

In this section, we review previous research related to our project, focusing on the importance of explainability in healthcare, prior usage of the dataset we employed, and methodologies for achieving interpretable machine learning models.

5.1 Importance of Explainability in Healthcare

Explainability in machine learning models is critical, especially in healthcare applications, where decisions can have significant impacts on patient outcomes. Cinà et al. (2022) argue that explainable AI is necessary for healthcare to ensure transparency, trust, and ethical decision-making in clinical settings [2]. Their work emphasizes that healthcare practitioners need to understand the reasoning behind AI-driven decisions to make informed choices and gain confidence in the models' outputs. This insight underscores the importance of developing and employing explainable models in our study.

5.2 Usage of Dataset in Related Research

The dataset used in our project has been previously utilized in other research, which validates its relevance and applicability. For instance, Yagin et al. (2023) employed this dataset to estimate obesity levels using a trained neural network approach optimized by Bayesian techniques[6]. Their study focused on predicting obesity based on physical activity and dietary habits, identifying key factors influencing obesity through feature selection algorithms like chi-square, F-Classify, and mutual information classification. Their results demonstrated high accuracy in obesity prediction, reinforcing the dataset's utility in health-related machine learning applications. This prior usage of the dataset supports its suitability for our analysis and provides a benchmark for evaluating our models' performance.

5.3 Explainability on Decision Trees

The quest for interpretable machine learning models has led to various approaches, including the development of decision trees with short explainable rules. Souza et al. (2022) introduced a method for creating decision trees that balance the trade-off between explainability and depth, aiming to maintain model performance while ensuring that the rules are simple and interpretable [4]. Their technique, which we consider incorporating in our project, offers a potential solution for achieving interpretable models that do not compromise on accuracy. By leveraging their function for evaluating the explainability-depth trade-off, we aim to enhance the interpretability of our models, making them more suitable for practical applications in healthcare.

5.4 Counterfactuals with DiCE library

The use of counterfactual explanations is a growing area of research in the field of explainable AI, providing insights into how slight changes in input data can alter the prediction of a model. In this project, we utilize the DiCE library [3], introduced by Mothilal et al.. Unlike earlier methods, which often produce single or similar counterfactual instances, DiCE focuses on creating multiple, distinct counterfactuals. This diversity is essential for finding varied and actionable pathways to achieve desired outcomes, making the explanations more practical and useful in real-world applications.

6 MEMBER CONTRIBUTIONS

Include one or two paragraphs or a table in which you briefly describe how each project member contributed to each component of the project

The contributions of each project member are described below: TEMPLATE, WE NEED TO CHANGE IT!!!

Alice Smith: Alice led the data collection and preprocessing efforts. She was responsible for cleaning the dataset and performing initial exploratory data analysis. Additionally, Alice implemented the baseline machine learning models and conducted feature engineering.

Bob Johnson: Bob focused on the development of explainable models. He implemented Explainable Boosting Machines (EBMs) and integrated the SHAP (SHapley Additive exPlanations) library for model interpretability. Bob also conducted experiments to evaluate the trade-off between explainability and model performance.

Charlie Brown: Charlie was in charge of the literature review and related work section. He identified relevant research articles, summarized key findings, and integrated these insights into the project report. Charlie also assisted with the final model evaluation and results interpretation.

Diana Green: Diana managed the project documentation and presentation. She coordinated the team meetings, ensured that all deliverables were completed on time, and compiled the final project report. Diana also created visualizations for the model results and contributed to the overall project discussion and conclusions.

REFERENCES

[1] 2019. Estimation of Obesity Levels Based On Eating Habits and Physical Condition . UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5H31Z.

[2] Giovanni Cinà, Tabea Röber, Rob Goedhart, and Ilker Birbil. 2022. Why we do need Explainable AI for Healthcare. arXiv:2206.15363

[3] Ramaravind Kommiya Mothilal, Amit Sharma, and Chenhao Tan. 2019. Explaining Machine Learning Classifiers through Diverse Counterfactual Explanations. *CoRR* abs/1905.07697 (2019). arXiv:1905.07697 http://arxiv.org/abs/1905.07697

[4] Victor Feitosa Souza, Ferdinando Cicalese, Eduardo Laber, and Marco Molinaro. 2022. Decision Trees with Short Explainable Rules. In *Advances in Neural Information Processing Systems*, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (Eds.), Vol. 35. Curran Associates, Inc., 12365–12379. https://proceedings.neurips.cc/paper_files/paper/2022/file/500637d931d4feb99d5cce84af1f53ba-Paper-Conference.pdf

[5] Courtney B Weir and Adam Jan. 2023. *BMI Classification Percentile And Cut Off Points*. StatPearls Publishing, Treasure Island (FL). [Updated 2023 Jun 26].

[6] Fatma Hilal Yagin, Mehmet Gülü, Yasin Gormez, Arkaitz Castañeda-Babarro, Cemil Colak, Gianpiero Greco, Francesco Fischetti, and Stefania Cataldi. 2023. Estimation of Obesity Levels with a Trained Neural Network Approach optimized by the Bayesian Technique. *Applied Sciences* 13, 6 (2023). https://doi.org/10.3390/app13063875

APPENDIX

Table 2: Dataset Variables Table

Variable Name	Role	Type	Demographic	Description	Units	Missing Values
Gender	Feature	Categorical	Gender			no
Age	Feature	Continuous	Age			no
Height	Feature	Continuous				no
Weight	Feature	Continuous				no
family_history_with_overweight	Feature	Binary		Has a family member suffered or suffers from overweight?		no
FAVC	Feature	Binary		Do you eat high caloric food frequently?		no
FCVC	Feature	Integer		Do you usually eat vegetables in your meals?		no
NCP	Feature	Continuous		How many main meals do you have daily?		no
CAEC	Feature	Categorical		Do you eat any food between meals?		no
SMOKE	Feature	Binary		Do you smoke?		no
CH2O	Feature	Continuous		How much water do you drink daily?		no
SCC	Feature	Binary		Do you monitor the calories you eat daily?		no
FAF	Feature	Continuous		How often do you have physical activity?		no
TUE	Feature	Integer		How much time do you use technological devices such as cell phone, videogames, television, computer and others?		no
CALC	Feature	Categorical		How often do you drink alcohol?		no
MTRANS	Feature	Categorical		Which transportation do you usually use?		no
NObeyesdad	Target	Categorical		Obesity level		no