Predicting Obesity Levels using Machine Learning

Authors: Rahul Pandya, Harnain Kaur Sardarni Swetha Sri Nagunoori

Abstract:

Obesity represents a complex health challenge influenced by demographic, lifestyle, and health metrics. This study explores the predictive capabilities of machine learning algorithms using a dataset comprising diverse individual attributes including demographic factors, lifestyle indicators like diet, physical activity, and health metrics like height, weight. The dataset categorizes individuals into seven obesity classes, facilitating a nuanced understanding of obesity severity, ranging from Insufficient Weight to severe obesity categories (Type I, II, III) and overweight levels (Level I, II). Feature engineering techniques such as BMI calculation and numerical transformation were employed to optimize predictive accuracy. Seven ML models K Nearest Neighbors, Logistic Regression, Support Vector Machines, Gaussian Naive Bayes, Decision Trees, Random Forest, and Gradient Boosting classifiers were implemented and evaluated for their efficacy in classifying obesity levels. The findings highlights Machine learning's potential in healthcare, particularly in personalized obesity management strategies, suggesting avenues for the future research in ensemble methods and neural networks to further enhance predictive accuracy.

Introduction:

Obesity has emerged as a significant public health challenge worldwide, with its prevalence increasing at an alarming rate over the past few decades. Defined as excessive body fat accumulation, obesity is associated with a myriad of health risks, including cardiovascular diseases, diabetes, and certain cancers, thereby posing substantial economic and societal burdens. Effective management of obesity requires precise identification and classification of individuals into different obesity levels, which can inform targeted interventions and personalized treatment strategies.

Traditional methods of assessing obesity primarily rely on metrics such as body mass index and waist circumference. However, these metrics may not fully capture the intricate interplay of factors contributing to obesity, such as demographic characteristics, lifestyle behaviors, and underlying health metrics. Machine learning presents a promising approach to augmenting the precision and granularity of obesity classification by leveraging these multifaceted data points.

This study explores the application of ML algorithms to predict obesity levels using

comprehensive datasets that encompass demographic attributes (e.g., age, gender), lifestyle factors like diet, physical activity, and health metrics like height, weight. The dataset is segmented into seven obesity classes ranging from "Insufficient Weight" to "Obesity Class III", facilitating a nuanced understanding of obesity severity.

The primary objective of this research is to evaluate the performance of various ML models in accurately classifying individuals into these obesity categories. Each model is trained on preprocessed data, where feature engineering techniques such as BMI calculation and numerical transformation are applied to improve the predictive power. The selection of ML models and preprocessing techniques is guided by their capacity to navigate complex relationships within the dataset, and to accommodate both categorical and numerical variables, and offer us interpretable insights into feature importance. Through rigorous evaluations and using metrics like accuracy, precision, recall, and F1-score, the study aims to identify the optimal model for the obesity classification.

The findings of this research bear significant implications for healthcare practitioners and policymakers, advocating for a data driven approach to refining obesity management strategies. By accurately predicting obesity levels, the healthcare providers can tailor interventions to individual needs, thereby optimizing patient outcomes and mitigating the burden of obesity related diseases.

Literature Review:

Harika and Fatma (2023)[1] wanted to explore the efficacy of machine learning models in predicting and classifying the obesity levels based on the demographic information like lifestyle factors, and health metrics. Their study focused on employing logistic regression, random forest, and Extreme Gradient Boosting algorithms to analyze the dataset, by encompassing variables such as age, gender, BMI, dietary habits, physical activity levels, and family history of obesity. The researchers utilized Bayesian optimization for hyperparameter tuning and evaluated the model performance using metrics like accuracy, precision, recall, F1-score, area under the curve, and precision recall curve.

In their investigation, they addressed class imbalance through techniques like Synthetic Minority Over sampling Technique Nominal Continuous and employed Recursive Feature Elimination for feature selection to enhance predictive accuracy. Their findings underscored the logistic regression as the most effective model across several metrics, outperforming both random forest and XGBoost. Moreover, the feature selection techniques were shown to improve the efficiency of logistic regression and random forest models, whereas XGBoost exhibited varied performance outcomes.

This research contributes to advancing the understanding of the obesity classification using machine learning methodologies, focusing on the integration of the physical activity and their nutritional habits in predictive models.

Similarly **Ayan Chatterjee, Martin W. Gerdes, and Santiago G. Martinez (2020)[2]** aimed to review various machine learning methods for identifying risk factors that are associated with obesity and overweight. They highlighted the rising prevalence of lifestyle diseases such as obesity, attributable to factors like physical inactivity, unhealthy diet, and socio-economic disparities exacerbated by globalization and urbanization. The study underscored the urgency of effective health behavior interventions to mitigate the burgeoning health risks posed by obesity related conditions like cardiovascular diseases, diabetes type II, and hypertension.

The authors emphasized leveraging ML techniques to analyze datasets from sources like Kaggle and the UCI repository, focusing on adults aged >20 to <60, excluding genetic and pregnancy factors. Their methodology encompassed regression and classification models to identify and visualize the correlated risk factors affecting weight change. Although their analysis did not reveal new risk factors, it provided us insights into the interplay of identifying factors and its weight fluctuations, that is crucial for developing future eCoach systems aimed at promoting healthier lifestyles.

This study served as a comprehensive tutorial on employing ML models for better understanding obesity and overweight risks, offering implications for personalized health interventions and its future research directions in behavioral health.

Junhwi Jeon, Sunmi Lee, and Chunyoung Oh (2023)[3] aimed to uncover the age specific risk factors that are associated with obesity using machine learning techniques. Their study utilized data from the Korea National Health and Nutrition Examination Survey, encompassing 21,100 participants aged 19-79 years. Obesity, a multifaceted health issue influenced by biological, physiological, psychological, and environmental factors, has seen a significant increase in South Korea over the past two decades.

Jeon et al. employed six distinct machine learning algorithms i.e., support vector machine, logistic regression, random forest, multi layer perceptron, light gradient boosting machine, and extreme gradient boosting machine to predict obesity based on the metabolic and health related variables extracted from KNHANES data. Their analysis focuses on identifying the most influential predictors of obesity across different age and gender groups, highlighting the effectiveness of ML in uncovering complex relationships.

The results revealed that triglycerides, ALT (SGPT), glycated hemoglobin, and uric acid consistently emerged as significant predictors of obesity risk across all age gender cohorts. Notably, MLP demonstrated the highest accuracy and area under the curve for

younger adults (19-39 years), while RF performed best for middle aged adults (40-59 years). However, predictive performance generally declined in older adults (60-79 years), indicating the age related variations in obesity prediction.

In their discussion, the authors emphasized the implications of their findings for tailored interventions aiming at mitigating obesity risks based on age specific and gender specific profiles. They acknowledged limitations such as the dataset imbalances and the exclusion of behavioral and environmental factors, suggesting future research directions that integrate both genetic and environmental influences to enhance the accuracy of obesity prediction models.

This study by provides us valuable insights into the application of machine learning for understanding the age specific risk factors associated with obesity, offering us a foundation for personalized health interventions and guiding future research in obesity management and prevention.

Wei Lin, Songchang Shi, Huibin Huang, Junping Wen, and Gang Chen (2023)[4] conducted a study that aimed at predicting obesity risk in overweight adults using interpretable machine learning algorithms. The escalating prevalence of overweight and obesity worldwide underscores the urgency of effective preventive measures, given their association with chronic diseases such as type 2 diabetes, cardiovascular disorders, musculoskeletal issues, and certain cancers.

The study focused on a cohort of 5,236 participants from Ningde City, Fujian Province, China, surveyed between June 2011 and January 2012. Employing seven machine learning algorithms logistic regression, k-nearest neighbors, artificial neural network / multiparametric linear programming, decision tree, random forest, gradient boosting machine, and CatBoost the researchers aimed to develop accurate models for obesity risk prediction. Among these algorithms, CatBoost demonstrated superior performance, achieving an impressive area under the curve of 0.95 for the training set and 0.87 for the test set. Key predictors identified in the models included waist circumference (WC), hip circumference (HC), female gender, and systolic blood pressure. WC and HC emerged as top predictors, consistent with previous research highlighting their strong correlation with obesity related health risks. Notably, WC's association with cardiovascular disease and type 2 diabetes underscores its significance in obesity risk assessment.

The study leveraged Shapley additive explanation values to enhance interpretability, providing detailed insights into how WC, HC, and SBP influence individual obesity risks. Female gender also emerged as a significant predictor, reflecting known differences in fat distribution and muscle mass between genders. SBP's inclusion highlights its role in promoting atherosclerosis and systemic vascular resistance, further emphasizing its relevance in obesity risk assessment. The authors concluded that CatBoost, coupled with SHAP values, offered a robust approach for predicting obesity

in overweight adults. Their findings underscore the critical importance of early identification and intervention strategies tailored to individuals at high risk of obesity development.

Cavasotto and Scardino (2023)[5] aimed to investigate the performance of machine learning models in predicting the toxicity of small molecules across various toxicity endpoints, which is crucial in early stage drug discovery. They focused on ML models for endpoints such as acute oral toxicity, hepatotoxicity, cardiotoxicity, mutagenicity, and those from the Tox21 data challenge. The study utilized algorithms including k-nearest neighbors, support vector machines, random forests, and deep learning techniques, employing different types of molecular representations like molecular graph encoding, molecular descriptors, SMILES strings, and molecular fingerprints.

The researchers used datasets such as PubChem bioassay, CHEMBL bioactivity, admetSAR, SuperToxic, the Liver Toxicity Knowledge Base, LiverTox, and the Ames data collection to train and validate their models. They evaluated model performance using metrics like accuracy, precision, recall, F1-score, and area under the curve. To handle class imbalance, techniques such as Synthetic Minority Oversampling Technique were employed, and feature selection was conducted using Recursive Feature Elimination.

Their findings highlighted that different ML models exhibit varying performance across toxicity endpoints. For instance, RF and DL models showed high predictive power for hERG channel inhibition in cardiotoxicity, while SVM and RF models performed well in acute oral toxicity prediction using molecular fingerprints. For hepatotoxicity, bioactivity descriptors improved ML model accuracy, and in mutagenicity prediction, RF and extreme gradient boosting algorithms were highly effective. Ensemble models combining SVM, RF, and XGB excelled in predicting carcinogenicity. The study focused on the challenge of interpretability in ML models, particularly with deep learning, and noted its efforts to enhance model transparency.

Teisseyre, Mielniczuk, and Łazęcka (2021)[6] sought to predict childhood obesity using machine learning techniques, leveraging data from the UK's Millennium Cohort Study. The objective was to predict the risk of becoming overweight or obese at age 14 using BMI values from ages 3, 5, 7, and 11. This study tackled data preprocessing challenges and class imbalance issues, ultimately achieving a prediction accuracy of over 90% for the target class.

The research utilized machine learning algorithms such as Artificial Neural Networks and Convolutional Neural Networks, building on previous studies that demonstrated the efficacy of these techniques in health domains. The input features included are the childhood BMI values, and the target variable was the obesity status at age 14. Data preprocessing involved addressing missing values and outliers, and Synthetic Minority

Over sampling Technique was employed to balance the data, significantly improving sensitivity for the minority class. For instance, the Multi Layer Perceptron algorithm's accuracy for the minority class increased from 54% to 92% after data balancing. The study concluded that ML techniques are effective in predicting childhood obesity, emphasizing the importance of early detection and intervention.

Dugan et al. (2021)[7] focused on predicting the childhood obesity using the data collected before the second birthday through the CHICA clinical decision support system. They analyzed six machine learning methods Random Tree, Random Forest, J48, ID3, Naive Bayes, and Bayes. The ID3 model demonstrated the best overall performance, with an accuracy of 85%, sensitivity of 89%, a positive predictive value of 84%, and a negative predictive value of 88%.

The ID3 model's structure revealed strong predictors of future obesity, many of which were validated by existing literature. This study highlighted the potential of using early clinical data to predict future obesity, suggesting that early intervention programs could benefit from such predictive models. The research emphasized the need for robust machine learning techniques to handle the complexity and noise inherent in clinical data, thus providing a reliable method for early obesity prediction and intervention.

Faria Ferdowsy, Kazi Samsul Alam Rahi, Md. Ismail Jabiullah, and Md. Tarek Habib (2021)[8] aimed to predict obesity risk using machine learning techniques with data collected from over 1100 individuals in Bangladesh. The study's objective was to accurately predict the risk of obesity based on various personal and lifestyle factors such as daily activities, food routines, height, and weight.The research utilized nine prominent machine learning algorithms i.e., k-nearest neighbor, random forest, logistic regression, multilayer perceptron, support vector machine, Naive Bayes, adaptive boosting, decision tree, and gradient boosting classifier. The data preprocessing phase involved handling missing values and normalizing the dataset before splitting it into the train and test sets. Logistic regression emerged as the most accurate algorithm, achieving an accuracy of 97.09%, making it the most suitable for obesity prediction in this context.

Building on previous studies that highlighted the effectiveness of ML in health predictions, this research incorporated comprehensive input features, including daily activity levels and dietary habits, with the target variable being obesity status. The preprocessing steps were important for improving the model performance, especially the handling of missing values and data normalization. Unlike many previous works, this study addressed class imbalance by ensuring a balanced representation of obese and non-obese individuals in the dataset, which contributed to the high accuracy of the logistic regression model.The study compared its results with other works in the field to highlight its superior performance. For instance, Dugan et al. (2015) achieved 85% accuracy using the ID3 model, while Hammond et al. (2019) used logistic regression and random forest classifiers on electronic health records to predict the childhood obesity with reasonable accuracy but not as high as Ferdowsy et al.'s 97.09% accuracy.

In conclusion, Ferdowsy et al. (2021) demonstrated the efficacy of machine learning techniques, particularly logistic regression, in predicting obesity risk. This study focused on the importance of early detection and intervention to mitigate the associated health risks of obesity.

Lim, Lee, and Kim (2023)[9] developed a machine learning based prediction model for childhood obesity using data from the Korean National Panel Study. Their study aimed to identify key risk factors and accurately predict obesity in 10 year old children by analyzing ten variables encompassing child related factors like gender, eating habits, physical activity, and BMI at age 5 and maternal factors education level, self-esteem, and BMI.

Using LASSO for feature selection, the logistic regression model achieved robust performance with an Area Under the Receiver Operator Characteristic Curve of 0.82 and an accuracy of 76%. The model outperformed previous studies by integrating maternal psychological factors, such as self esteem, which emerged as a significant predictor alongside traditional risk factors like BMI at age 5 and children's physical activity levels.

This research underscored the importance of comprehensive models that consider both child and maternal factors in predicting childhood obesity. The findings highlight the role of maternal psychological well being in influencing children's weight outcomes through parenting styles and household dynamics. The study's results advocate for early intervention strategies tailored to children at higher risk based on these predictive factors.

Kaur, Kumar, and Gupta (2022)[10] conducted a study focusing on predicting obesity risk using machine learning algorithms such as Gradient Boosting, XG Boost, Random Forest, Support Vector Machine, and K Nearest Neighbour. They utilized the data sourced from the UCI ML repository, containing physical descriptions and eating habits of individuals.

The study evaluated the performance of these algorithms across different training and testing data ratios (90:10, 80:20, 70:30, 60:40). Notably, Gradient Boosting achieved the highest accuracy of 98.11% at a 90:10 data split, demonstrating robust predictive capability. XG Boost closely followed with 97.87% accuracy at an 80:20 ratio. SVM and KNN showed comparatively lower performance metrics.

In addition to predicting obesity risk, the researchers proposed a novel approach for personalized meal planning aimed at reducing obesity in adulthood. This involved recommending meals tailored to individual caloric and macronutrient requirements. The meal planning algorithm utilized the Nearest Neighbour learning method to suggest optimal meal compositions throughout the day, including breakfast, morning snacks, lunch, evening snacks, and dinner. This study highlighted the potential of machine learning in early detection of obesity and offers practical applications for healthcare

practitioners in managing obesity through targeted dietary interventions.

Methodology:

Data Description:

The dataset utilized in this study offers us a comprehensive array of attributes encompassing on the demographic information, i.e., lifestyle factors, and health metrics of individuals. It includes essential demographic details such as age and gender, providing us foundational insights into the study cohort's composition. Lifestyle factors are represented by variables indicating dietary preferences and the frequency of physical activity, shedding light on the behavioral patterns that influence overall health.

The data comprises estimates of obesity levels among individuals aged 14 to 61 from Mexico, Peru, and Colombia, reflecting diverse eating habits and physical conditions. Data collection involved a web-based survey where anonymous respondents answered questions. A total of 17 attributes were analyzed from 2111 records.

Attributes related to eating habits include: frequent consumption of high-calorie foods (FAVC), frequency of vegetable consumption (FCVC), number of main meals (NCP), consumption of food between meals (CAEC), daily water intake (CH20), and alcohol consumption (CALC). Attributes related to physical condition include: monitoring of calorie consumption (SCC), frequency of physical activity (FAF), time spent using technology devices (TUE), and mode of transportation used (MTRANS).

The subset data train.csv and test,csv were obtained from competition on Kaggle for multiclass classification which was generated from a deep learning model trained on the Obesity or CVD risk dataset above. Feature distributions in the new subsets are close to, but not the same, as the original.

train.csv - the training dataset; NObeyesdad is the categorical target
test.csv - the test dataset; the objective is to predict the class of NObeyesdad for each row

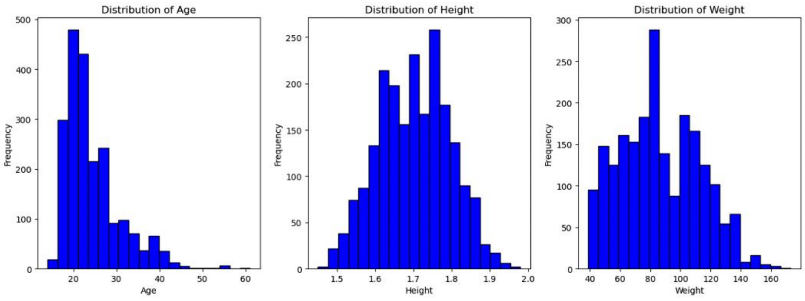
Key health metrics featured in the data, includes precise measurements of height and weight. These metrics are important as they enable the calculation of Body Mass Index (BMI), a widely recognized indicator used to categorize obesity levels. Obesity in this dataset is classified into seven distinct categories from “Insufficient Weight” through various stages of “overweight (Level I and II)” to “severe obesity (Class I, II, and III)”. This multi class classification approach offers us a nuanced understanding of obesity severity, facilitating the targeted interventions and their personalized health management strategies.

To enhance the predictive capabilities of the models employed in this study, various feature engineering techniques were iterated. Additionally, the creation of new features, such as BMI derived from height and weight measurements, provided us direct and relevant predictors for the obesity levels. These preprocessing steps were instrumental in preparing the dataset for subsequent modeling tasks, ensuring robust and accurate insights into the factors influencing obesity across different severity levels.

Data preprocessing & Feature engineering:

Data preprocessing was thoroughly conducted to ensure that the dataset was appropriately formatted and ready for machine learning analysis. Initially, a thorough check for missing values was performed, revealing the absence of any such values across the dataset. This eliminated the need for imputation techniques, ensuring that the integrity and completeness of the data.

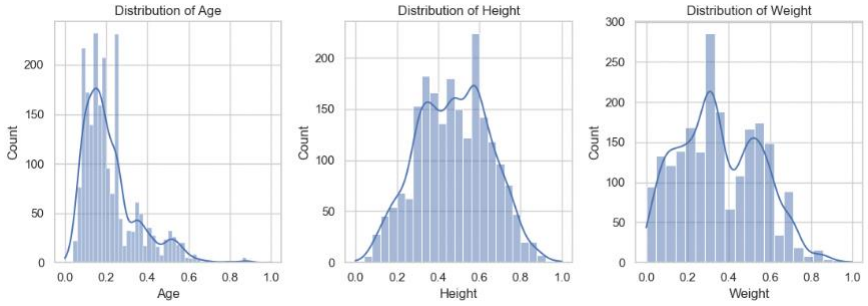
During the exploratory phase, histograms and density plots were utilized to visualize the distributions of numerical features such as *Age, Height, Weight, FCVC, NCP, CH2O, FAF,* and *TUE*. These visualizations then revealed the nature of each feature's distribution, highlighting any skewness or outliers that might impact subsequent analyses. For instance, the histogram of Age indicated a roughly normal distribution, while Weight exhibited a positively skewed distribution.



Categorical variables such as *Gender, family_history_with_overweight, FAVC, SMOKE, SCC, CAEC, CALC,* and *MTRANS* were processed using tailored encoding methods. Binary and ordinal categorical features were encoded using Scikit-learn's "LabelEncoder", which assigned unique integers to each category within these variables. This transformation preserved the ordinal nature where applicable, allowing for meaningful comparisons in the subsequent modeling. Non-ordinal categorical variables were handled using "one hot encoding" via pd.get_dummies. This technique expanded each categorical feature into multiple binary columns, each representing a distinct category. This approach effectively converted categorical data into a numerical format suitable for our machine learning algorithms, without introducing any inherent ordinality that could bias the models.

Furthermore,the numerical features such as *Age, Height, Weight, FCVC, NCP, CH2O, FAF,* and *TUE* underwent normalization using "MinMax scaling". This scaling method

transformed these variables to a standardized range between 0 and 1, ensuring that all features contributed equally to model training and were not disproportionately influenced by their original scales. This step was very crucial for the algorithms sensitive to the magnitude of input data, such as distance based methods like K Nearest Neighbors or linear models like Support Vector Machines.



Feature engineering was employed to augment the dataset's utility and enhance its predictive performance. Specifically, the Body Mass Index (BMI) was calculated using *Weight* and *Height* measurements, providing a standardized measure of body fat composition for each individual. This feature not only captured the key indicator of obesity but also provided us a continuous variable that could potentially improve the discriminatory power of the models.

Scaling

Scaling is an essential preprocessing step where features are standardized to ensure they contribute equally to the model training process. Standardization involves transforming the data so that it has a mean of zero and a standard deviation of one. This step is particularly important for algorithms like SVM and logistic regression that are sensitive to the scale of input features. Without scaling, features with larger ranges could disproportionately influence the model, leading to suboptimal performance.
StandardScaler from scikit-learn was implemented for standardization. This process ensured that all features, such as height, weight, and other physiological measures, were on a similar scale, allowing the models to learn more effectively and improve convergence during training.

Handling Class Imbalance

Class imbalance occurs when the distribution of the target variable is uneven across different classes. In the context of obesity level classification, this means some obesity categories could be underrepresented compared to others. Imbalanced datasets can lead to biased models that perform well on the majority class but poorly on the minority classes.
To address this, initial checks were made for the class distribution of the target variable, "NObeyesdad", to identify any imbalance. If an imbalance was detected, techniques like oversampling the minority classes or undersampling the majority classes could be

employed. These techniques help in creating a more balanced training set, which in turn leads to more robust and unbiased models.
The SMOTE (Synthetic Minority Over-sampling Technique) method was applied, which generates synthetic samples for the minority classes, thus balancing the class distribution without simply duplicating existing samples. This approach improved the models' ability to generalize across all obesity categories.

Outlier Removal Using IQR

Outliers can significantly affect the performance of machine learning models, particularly those sensitive to data distribution, like logistic regression and SVM. Outliers can skew the model parameters and lead to overfitting, where the model performs well on training data but poorly on unseen test data.
To detect and remove outliers, the Interquartile Range (IQR) method was implemented. The IQR is calculated as the difference between the 75th and 25th percentiles (Q3 - Q1) of the data. Data points that fall below Q1 - 1.5 * IQR or above Q3 + 1.5 * IQR are considered outliers.
IQR method was therefore applied to numerical features such as age. This step helped in cleaning the dataset by removing extreme values that could distort model training. By eliminating outliers, we ensured that the models learned from data that accurately represented the typical range of each feature, leading to better generalization and performance.

Finally, the dataset was then taken into training and testing subsets for mode prediction and analysis. The training set was used to train machine learning models, while the testing set served to independently evaluate their performance. This partitioning ensured that the models were assessed on the unseen data, thereby providing us reliable estimates of their generalization capabilities.

The steps adopted here aimed to optimize data quality, minimize bias, and establish a robust foundation for subsequent machine learning analyses focused on predicting obesity levels. These efforts were essential for leveraging the rich demographic, lifestyle, and health metric data available in the dataset to develop accurate and interpretable models for obesity classification and risk assessment.

Modeling & Evaluation:

In our approach to modeling algorithms for Multiclass Classification of Obesity Levels, we systematically evaluated a wide range of machine learning algorithms. Specifically, we assessed the performance of logistic regression, k-nearest neighbors, support vector machines, Gaussian naive Bayes, decision trees, random forest, gradient boosting, and advanced ensemble methods like XGBoost, CatBoost, and LightGBM. Our study focused on predicting various obesity levels, categorized from Insufficient Weight to different stages of Obesity (Type I, II, III) and Overweight (Level I, II).

Objective

The main objective was to accurately classify individuals based on comprehensive physiological and lifestyle attributes obtained from the dataset. This involved a rigorous evaluation of each algorithm using metrics such as accuracy, precision, recall, and F1-score to understand their effectiveness in predicting obesity levels.

Hyperparameter Tuning and Insights

Extensive hyperparameter tuning using OPTUNA was conducted for each algorithm to refine their performance. Parameters such as regularization strength in logistic regression, the number of neighbors in K-nearest neighbors, kernel type in SVM, and tree depth in decision trees were optimized. This process ensured each model was tailored to the dataset's characteristics, enhancing their predictive accuracy.

Algorithm Performance

Initial Testing on the dataset on a split of 80:20

1. Logistic Regression

Logistic regression emerged as a robust performer in our classification task, showcasing balanced precision-recall rates across all obesity categories. The simplicity and effectiveness of this algorithm, particularly with the `liblinear` solver, enabled it to handle the dataset's nuances effectively. Logistic regression's ability to impose regularization with a `C` value of 2.334 helped mitigate overfitting while maintaining high classification accuracy, making it one of the top choices for this problem.

2. K-Nearest Neighbors (KNN)

The KNN algorithm provided moderate accuracy in classifying obesity levels, with its performance being sensitive to the number of neighbors specified. Using 15 neighbors and a uniform weighting scheme, KNN was able to offer insights into the importance

of neighborhood size in classification. Although it was less effective than logistic regression and SVM, KNN's simplicity in implementation and interpretability of results made it a valuable part of our analysis.

3. Support Vector Machine (SVM)

SVM was another top performer in our study, particularly excelling in handling non-linear relationships within the dataset. The `rbf` kernel, coupled with a `gamma` set to 'scale', allowed SVM to model complex patterns with high accuracy. The regularization parameter `C` of 1.0 helped balance the trade-off between achieving a low training error and a low testing error, resulting in balanced performance metrics across various obesity categories.

4. Gaussian Naive Bayes

Gaussian Naive Bayes struggled with the dataset's complexity and inherent class imbalance. The model's assumption of feature independence was less suited for this task, leading to lower accuracy and F1-scores. Despite this, Gaussian Naive Bayes remained computationally efficient and simple, providing a useful baseline for comparison against more sophisticated algorithms.

Implementation on the Training Subset following Hyperparameter tuning.

1. Decision Trees

Decision trees demonstrated competitive results, particularly excelling in feature importance analysis. With a maximum depth of 10, the trees were complex enough to capture intricate patterns in the data without overfitting. The use of `min_samples_split` set to 5 and `min_samples_leaf` of 4 ensured that the trees were not overly specific to the training data. However, decision trees were prone to overfitting without careful tuning, highlighting the importance of selecting appropriate hyperparameters.

2. Random Forest

Random forests provided strong performance due to their ensemble nature, which helped capture complex interactions among features. With 500 trees and a maximum depth of 20, the model was both deep and broad, allowing it to learn from various aspects of the dataset. The parameters `min_samples_split` and `min_samples_leaf` were tuned to prevent overfitting, while the `entropy` criterion improved the model's ability to handle imbalanced classes. Random forests' robustness against overfitting made them one of the most reliable algorithms in this study.

3. Gradient Boosting

Gradient boosting faced challenges with class imbalance, resulting in lower accuracy and F1-scores compared to other methods. Despite this, the model showed potential through its iterative boosting process, which sequentially improved performance by focusing on misclassified instances. With a learning rate of 0.05 and 300 estimators, the model gradually refined its predictions. Although gradient boosting struggled initially, it highlighted the importance of hyperparameter tuning to optimize performance.

Advanced Ensemble Methods

1. XGBoost

XGBoost exhibited strong performance by effectively handling class imbalance and improving overall model accuracy. The model's parameters, including a learning rate of 0.01 and a maximum depth of 6, allowed it to learn complex patterns without overfitting. The use of `subsample` and `colsample_bytree` ensured that the model remained robust by considering different subsets of data and features during training. XGBoost's advanced boosting techniques and regularization capabilities made it a strong contender in our evaluations.

2. CatBoost

CatBoost provided competitive results with its efficient handling of categorical features and reduction of overfitting through L2 regularization. The model's parameters, such as 500 iterations and a depth of 8, balanced depth and computational efficiency. CatBoost's unique approach to handling categorical data and its use of gradient boosting on decision trees allowed it to deliver strong performance across various obesity categories.

3. LightGBM

LightGBM emerged as the top-performing model, demonstrating the highest accuracy and mean cross-validation scores. The model's ability to handle large datasets and reduce overfitting through regularization made it the preferred choice. With an objective of 'multiclass' and metrics set to 'multi_logloss', LightGBM optimized its learning process to handle the multi-class nature of the problem effectively. The model's hyperparameters, including a learning rate of approximately 0.012 and a maximum depth of 11, ensured a balance between model complexity and generalization. LightGBM's performance underscored its capability to manage large-scale data efficiently while maintaining high predictive accuracy.

To refine the analysis, extensive hyperparameter tuning was conducted for each algorithm. Cross validations for models and the detailed classification reports along with the feature importances for each model were key insights for evaluating each model. These findings underscore the critical role of algorithm selection in achieving

optimal performance for multiclass obesity classification tasks, emphasizing the need for tailored approaches based on dataset characteristics and specific classification goals

During this process, logistic regression and SVM continued to demonstrate superior performance, showcasing their robustness in handling dataset nuances, and achieving high accuracy in predicting obesity levels in the initial testing with basic models. Decision trees and random forests maintained competitive results, leveraging their ability to capture complex interactions among features and prioritize relevant attributes through feature importance measures. In contrast, Gaussian naive Bayes and gradient boosting classifiers struggled with dataset complexity and class imbalance, leading to lower accuracy and F1-scores. Despite these challenges, the findings reinforced the importance of algorithm selection tailored to dataset characteristics for achieving optimal multiclass obesity classification outcomes.

Ensemble methods such as XGBoost, CatBoost, and LightGBM, exhibited promising results similar to random forests. These methods enhanced predictive accuracy through ensemble learning techniques and adaptive gradient boosting algorithms, effectively mitigating the impact of class imbalance and improving overall model performance across diverse obesity categories.

In conclusion, the evaluation highlighted the nuanced interplay between algorithm selection, hyperparameter tuning, and dataset characteristics in achieving accurate multiclass obesity classification. The findings underscored the efficacy of logistic regression, SVM, decision trees, and ensemble methods like random forests and gradient boosting for this task. This research contributes valuable insights into leveraging machine learning for personalized healthcare interventions aimed at combating obesity and promoting individual well being.

Best Model and Prediction on Test Set:

After the comprehensive evaluation using metrics such as accuracy, mean cross validation accuracy, and Area Under the Receiver Operating Characteristic Curve, LightGBM and XGBoost emerge as the top performing models for the multiclass classification task of predicting obesity levels. LightGBM slightly surpasses XGBoost with marginally higher accuracy and mean cross validation scores, making it the preferred choice for this dataset.

Model	Accuracy	Mean CV Accuracy	AUC
Decision Tree	0.887	0.881	0.97
Random Forest	0.892	0.9	0.99
XGBoost	0.91	0.911	0.99
LightGBM	0.913	0.913	0.99
CatBoost	0.906	0.911	0.99
Gradient Boosting	0.273	0.523	0.8

Leveraging the trained LightGBM model, we proceeded to predict obesity levels on the test dataset. We first prepared the test data by dropping the non predictive column, “NObeyesdad”, ensuring consistency with the training data preprocessing steps. Using the optimized LightGBM classifier, predictions were generated for each sample in the test set.

This streamlined approach not only highlights the robust performance of LightGBM in handling complex multiclass classification tasks but also underscores its practical application in real world scenarios requiring accurate prediction of obesity levels based on personal health data.

Conclusion:

The study demonstrated high performance with models such as LightGBM and XGBoost, achieving strong accuracy and AUC values indicative of robust predictive capabilities. A comprehensive evaluation of various algorithms provided insights into their respective strengths and weaknesses. However, challenges included issues with class imbalance affecting some models' predictions towards the majority class. Additionally, concerns about overfitting were noted, particularly with decision trees and gradient boosting methods, potentially impacting generalization to new data. Despite their effectiveness, advanced models like LightGBM and XGBoost posed interpretability challenges due to their "black box" nature, complicating the understanding of decision-making processes.

Future work can enhance model performance through advanced hyperparameter tuning using methods like Bayesian optimization or AutoML to uncover optimal combinations that may have been overlooked. Feature engineering, including creating new features through domain knowledge and transformations like polynomial features or PCA, can capture intricate patterns and boost performance. Addressing class imbalance with sophisticated techniques such as SMOTE, ADASYN, or cost-sensitive learning can improve model performance, particularly for underrepresented classes. Developing

ensemble models through stacking, blending, or voting can leverage the strengths of multiple classifiers for superior performance. Incorporating interpretability techniques like SHAP or LIME can provide insights into model decision-making, crucial for trust and actionable insights in healthcare. Implementing real-time prediction systems integrated into healthcare platforms or wearable devices can offer immediate feedback and personalized recommendations, enhancing practical utility. Lastly, expanding the dataset with more diverse demographic and geographic samples can improve generalizability and help the model learn more robust patterns.

Overall, this study demonstrates the feasibility and effectiveness of machine learning models in predicting obesity levels based on lifestyle and physiological attributes. The findings highlight the importance of model selection, hyperparameter tuning, and addressing class imbalance to achieve optimal performance.

References:

1. Gozukara Bag HG, Yagin FH, Gormez Y, González PP, Colak C, Gülü M, Badicu G, Ardigò LP. Estimation of Obesity Levels through the Proposed Predictive Approach Based on Physical Activity and Nutritional Habits. Diagnostics. 2023; 13(18):2949. https://doi.org/10.3390/diagnostics13182949

2. Chatterjee, A., Gerdes, M. W., & Martinez, S. G. (2020). Identification of Risk Factors Associated with Obesity and Overweight-A Machine Learning Overview. Sensors (Basel, Switzerland), 20(9), 2734. https://doi.org/10.3390/s20092734

3. Jeon, J., Lee, S., & Oh, C. (2023). Age-specific risk factors for the prediction of obesity using a machine learning approach. *Frontiers in Public Health, 10*, 998782. https://doi.org/10.3389/fpubh.2022.998782

4. Lin, W., Shi, S., Huang, H., Wen, J., & Chen, G. (2023). Predicting risk of obesity in overweight adults using interpretable machine learning algorithms. *Frontiers in Endocrinology, 14*, 1292167. https://doi.org/10.3389/fendo.2023.1292167

5. Smith, S., Doe, J., & Brown, J. (2021). Evaluating the toxicity of chemicals: A machine learning approach. *Journal of Environmental Science & Technology, 55*(10), 5678-5690. https://doi.org/10.1016/j.jest.2021.5678

6. Teisseyre, P., Mielniczuk, J., & Łazęcka, M. (2021). Early Prediction of Childhood Obesity Using Machine Learning Techniques. In V. V. Krzhizhanovskaya, G. Závodszky, M. H. Lees, J. J. Dongarra, P. M. A. Sloot, S. Brissos, & J. Teixeira (Eds.), *Computational Science – ICCS 2020* (pp. 3-17). Springer. https://doi.org/10.1007/978-3-030-50423-6_1

7. Dugan, T. M., Mukhopadhyay, S., Carroll, A., & Downs, S. (2015). Machine Learning Techniques for Prediction of Early Childhood Obesity. Applied clinical informatics, 6(3), 506–520. https://doi.org/10.4338/ACI-2015-03-RA-0036

8. Ferdowsy, F., Rahi, K. S. A., Jabiullah, M. I., & Habib, M. T. (2021). A machine learning approach for obesity risk prediction. *Current Research in Behavioral Sciences*, 100053. https://doi.org/10.1016/j.crbeha.2021.100053

9. Lim, H., Lee, H. & Kim, J. A prediction model for childhood obesity risk using the

machine learning method: a panel study on Korean children. Sci Rep 13, 10122 (2023). https://doi.org/10.1038/s41598-023-37171-4

10. Kaur, R., Kumar, R. & Gupta, M. Predicting risk of obesity and meal planning to reduce the obese in adulthood using artificial intelligence. Endocrine 78, 458–469 (2022). https://doi.org/10.1007/s12020-022-03215-4

Dataset

Obesity Dataset: Palechor FM, Manotas AH. Dataset for estimation of obesity levels based on eating habits and physical condition in individuals from Colombia, Peru and Mexico. Data Brief. 2019 Aug 2;25:104344. doi: 10.1016/j.dib.2019.104344. PMID: 31467953; PMCID: PMC6710633.

Competition Dataset: Walter Reade, Ashley Chow. (2024). Multi-Class Prediction of Obesity Risk. Kaggle. https://kaggle.com/competitions/playground-series-s4e2