Estimation of Obesity Levels Based on Eating Habits and Physical Condition

Inamullah Mohammad[1] and Bala Nagaraju Narra[1]

[1]Data Science & Analytics, University of Oklahoma, Norman, OK, United States

Abstract

Obesity is a major public health issue globally, associated with various chronic conditions that necessitate early intervention. This project aims to contribute to the efforts against this challenge by developing machine learning models capable of predicting obesity levels from detailed data on individuals' eating habits and physical conditions. Our data, sourced from the UCI repository, includes 2111 entries from Mexico, Peru, and Colombia, and combines both real and synthesized information to ensure comprehensive coverage and reliability. The study will deploy five predictive models: Logistic Regression, Random Forest, Support Vector Machines, Gradient Boost, and Neural Networks, with an emphasis on achieving high accuracy through hyperparameter tuning and feature engineering. The project's goal is to advance public health by providing tools for early obesity detection and promoting preventive health measures.

1 Introduction

Obesity is a rapidly growing global public health challenge.[1] A BMI (body mass index) of 30 kg/m^2 or higher is used to define obesity in individuals. Over the past three decades, the global commonness of obesity has increased by 27.5% among adults and 47.1% among children.[2] Individuals afflicted with obesity face increased risks of developing a variety of comorbid conditions. These include cardiovascular diseases, gastrointestinal disorders, type 2 diabetes, joint and musculoskeletal disorders, respiratory complications, and psychological conditions. Collectively, these comorbidities can substantially impair daily functioning and elevate the risk of mortality.[2]

Body mass index (BMI) serves as a primary metric for assessing obesity within populations. Alternative measures include waist-to-hip ratio, percentage of body or visceral fat, and waist circumference. BMI is determined through mathematical calculations utilizing height and weight to assess an individual's health status. This measurement is commonly employed to evaluate the risk of chronic diseases such as diabetes, hypertension, depression, and cancer.[3]

Machine learning comprises a robust suite of algorithms capable of characterizing, adapting, learning from, predicting, and analyzing data. This technology enhances our understanding of obesity and significantly improves our predictive accuracy. Consequently, the application of machine learning in obesity research has seen a notable increase [4]. SMOTE-NC (Synthetic Minority Oversampling Technique for Nominal and Continuous data) is an extension of the original SMOTE algorithm, designed to handle datasets that include both nominal (categorical) and continuous

features. This method is particularly useful in scenarios where data classes are imbalanced. It generates synthetic samples by interpolating between minority class samples in the feature space, helping to balance the dataset and improve the performance of classifiers.[5]

2 Related Work

Several studies have explored the application of machine learning for predicting obesity rates. Buani and Nuraeni (2023) investigated the effectiveness of various machine learning models, including XGB Classifier, SVM, Random Forest, Naive Bayes, K-NN, Logistic Regression, and Decision Tree Classifier.[6] Their findings suggest that the XGB Classifier achieved the highest accuracy (96%) in predicting obesity, demonstrating its potential for handling complex health datasets.

Choudhuri (2022) delved into the use of hybrid models, combining Extremely Randomized Trees, Multilayer Perceptron, and XGBoost. This approach combined supervised and unsupervised learning, achieving improved accuracy in predicting obesity levels from a dataset containing information on eating habits and physical condition.[7] Feature engineering techniques were employed extensively to address challenges associated with data diversity and imbalance.

Interestingly, all these studies, including the current research, utilize the same dataset from the UCI Machine Learning Repository. This dataset has been a focal point for developing predictive models due to its comprehensive collection of variables related to eating habits and physical conditions, which are crucial for understanding and predicting obesity levels.

The continuous development of both obesity as a public health concern and machine learning technologies motivates further research in this field. Each study contributes to a growing knowledge base that can adapt to evolving population health dynamics and advancements in data collection and analysis. Refining predictive models can significantly aid in early detection and prevention strategies, ultimately mitigating the global burden of obesity on individuals and healthcare systems. Additionally, this line of research paves the way for personalized healthcare solutions, leveraging data-driven insights to cater to individual needs and enhance the effectiveness of interventions.

3 Data Description

The dataset utilized in this study is sourced from the UCI Machine Learning Repository and comprises 2111 records with 17 attributes. According to the introductory paper, the data was initially collected through a webpage survey designed to assess participants' eating habits and various physical condition attributes. The original data exhibited class imbalance, prompting the use of the Weka tool and the SMOTE filter to generate synthetic data. Approximately 77% of the data is synthetic, while the remaining 23% consists of the original data. Since this is a augmented dataset, it does not have any missing values.[8]

The description of each attribute is as follows:[9]

1. **Gender** : Biological sex of the person. *value will be either Male or Female*
2. **Age**: Individual's age. *value in years*
3. **Height**: Individual's height. *value in meters*

4. **Weight**: Individual's weight. *value in pounds*
5. **Family history of overweight** : Whether the individual has any family member who is obese. *value will be either yes or no*
6. **FAVC** : Whether the individual often eats high-calorie food. *value will be either yes or no*
7. **FCVC** : How often the individual eats vegetables? *values are (1 = never, 2 = sometimes, 3 = always)*
8. **NCP** : How many main meals the individual has daily? *values are (1 = between one and two, 2 = three, 3 = more than three, 4 = no answer)*
9. **CAEC** : How often the individual eats food between meals? *values are (1 = no, 2 = sometimes, 3 = frequently, 4 = always)*
10. **SMOKE** : Whether the individual smokes or not. *value will be either yes or no*
11. **CH2O** : How much water the individual drinks daily? *values are (1 = less than a liter, 2 = between 1 and 2 L, 3 = more than 2 L)*
12. **SCC** : Whether the individual consistently keeps track of their caloric intake. *value will be either yes or no*
13. **FAF** : How often the individual does physical activity? [1] *(1 = never, 2 = once or twice a week, 3 = two or three times a week, 4 = four or five times a week)*
14. **TUE** : How long the individual uses electronic devices? *(0 = none, 1 = less than an hour, 2 = between one and three hours, 3 = more than three hours)*
15. **CALC** : How often the individual drinks alcohol? *values are (1 = no, 2 = sometimes, 3 = frequently, 4 = always)*
16. **MTRANS** : What kind of transportation the individual uses? *values in (automobile,motorbike, bike, public transportation, walking)*
17. **NObeyesdad** : Level of obesity according to body mass index. *values in (insufficient weight, normal weight, overweight level I, overweight level II, obesity type I, obesity type II, obesity type III)*

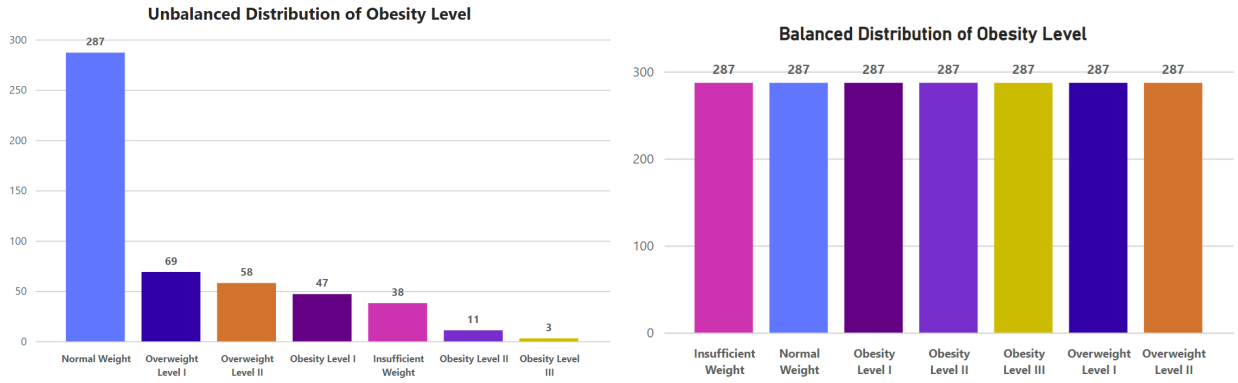
In this study, the attribute 'NObeyesdad' is designated as our target variable. We will predict the obesity level using the predictors provided, framing this as a multi-class classification problem. The multi-class nature of 'NObeyesdad' allows us to categorize individuals across a spectrum of obesity levels, from insufficient weight to extreme obesity (obesity type III).

4 Data Preprocessing

4.1 Removing Incorrect Data

Upon detailed examination of the dataset, it was observed that the data generated using the Weka tool and the SMOTE filter was incorrectly processed. Specifically, categorical variables were treated

[1] In the actual data, values for this column are (0, 1, 2, 3). Since there was no occurrence of the value 4, we assumed the mapping to be *(0 = never, 1 = once or twice a week, 2 = two or three times a week, 3 = four or five times a week).*



(a) Imbalanced Distribution of Obesity Levels

(b) Balanced Distribution of Obesity Levels

Figure 1: Bar plots illustrating Obesity Level distribution before and after applying the SMOTE-NC Filter. The imbalance in class distribution is completely eliminated in the augmented data.

as continuous, resulting in values being represented in decimals, which is inappropriate for such data types. After removing all rows containing the inaccurately processed data, we were left with 513 rows of correct data.[2]

4.2 Dealing with class imbalances

In the remaining data, a noticeable imbalance was observed in the Obesity Levels, as illustrated in Fig. 1a. This imbalance could lead to bias in the machine learning models if not addressed. To mitigate this issue and ensure a balanced representation of Obesity Levels, we intend to generate synthetic data.

To achieve this, we will utilize SMOTE-NC, an extension of the Synthetic Minority Over-sampling Technique (SMOTE) designed to handle both numeric and categorical data. This method is particularly suited to our dataset, which contains a mix of these data types.

Prior to the application of SMOTE-NC, it was essential to encode all categorical columns to numerical formats, except for the class label 'NObeyesdad'. This encoding step is crucial as it prepares the data for effective integration with the synthetic data generated by SMOTE-NC, ensuring the models are trained on a balanced and cohesive dataset.

After applying the SMOTE-NC filter, the original data was augmented with synthetic data, resulting in a total of 2,009 records. Each class now has an equal number of rows (Fig. 1b), effectively balancing the dataset for more accurate model training.

[2]We removed all rows in which the categorical features had decimal values. Among the remaining records, it is possible that some synthetic records are indistinguishable from the original data.

4.3 Data Transformation

4.3.1 Data Standardization

Data standardization is a key step in machine learning that helps ensure every feature has an equal impact on the analysis. This procedure involves calculating the average and standard deviation for each feature, then adjusting each data point by subtracting the average and dividing by the standard deviation. This adjustment centers the data around zero and normalizes the spread of values to a standard deviation of one.

In this study, we used the `StandardScaler` class from the `Scikit-learn` library to standardize the numerical columns, specifically Age, Height, and Weight.

4.3.2 Data Encoding

In this study, data encoding is streamlined as most categorical columns are already numerically encoded through SMOTE-NC, which is necessary for handling both nominal and continuous data. The only exception is the target column, which remains unencoded. To address this, we use the `LabelEncoder` from `Scikit-learn`, which transforms each unique class label in the target column into a corresponding integer.

4.4 Correlation Matrix

The correlation matrix heatmap (Fig. 2) in our dataset reveals important relationships between different variables, providing insights into how various factors may interact. For example, it shows a moderate positive correlation between gender and height, suggesting that height varies by gender. This can be particularly relevant for studies focusing on body size differences across genders, highlighting how physical attributes can influence health outcomes.

Additionally, there is a strong link between having a family history of overweight and frequent consumption of high-caloric foods. This underscores the influence of both genetic and lifestyle factors in the development of obesity. Understanding these connections is crucial for targeting interventions and creating strategies that address both inherited and behavioral aspects of obesity prevention.

However, due to the complexity of interactions between these variables, we decided to employ an additional method for feature selection to refine our model's accuracy further and ensure robustness in our predictive analysis. This method will be discussed in the upcoming sections, as it helped us effectively identify the most significant features for our models, enhancing both the accuracy and reliability of our predictions in addressing public health concerns.

5 Methodologies

In this section, we discuss the methodologies used in this project which are adapted from previous research utilizing the same dataset for multi-class classification of obesity. These methodologies are

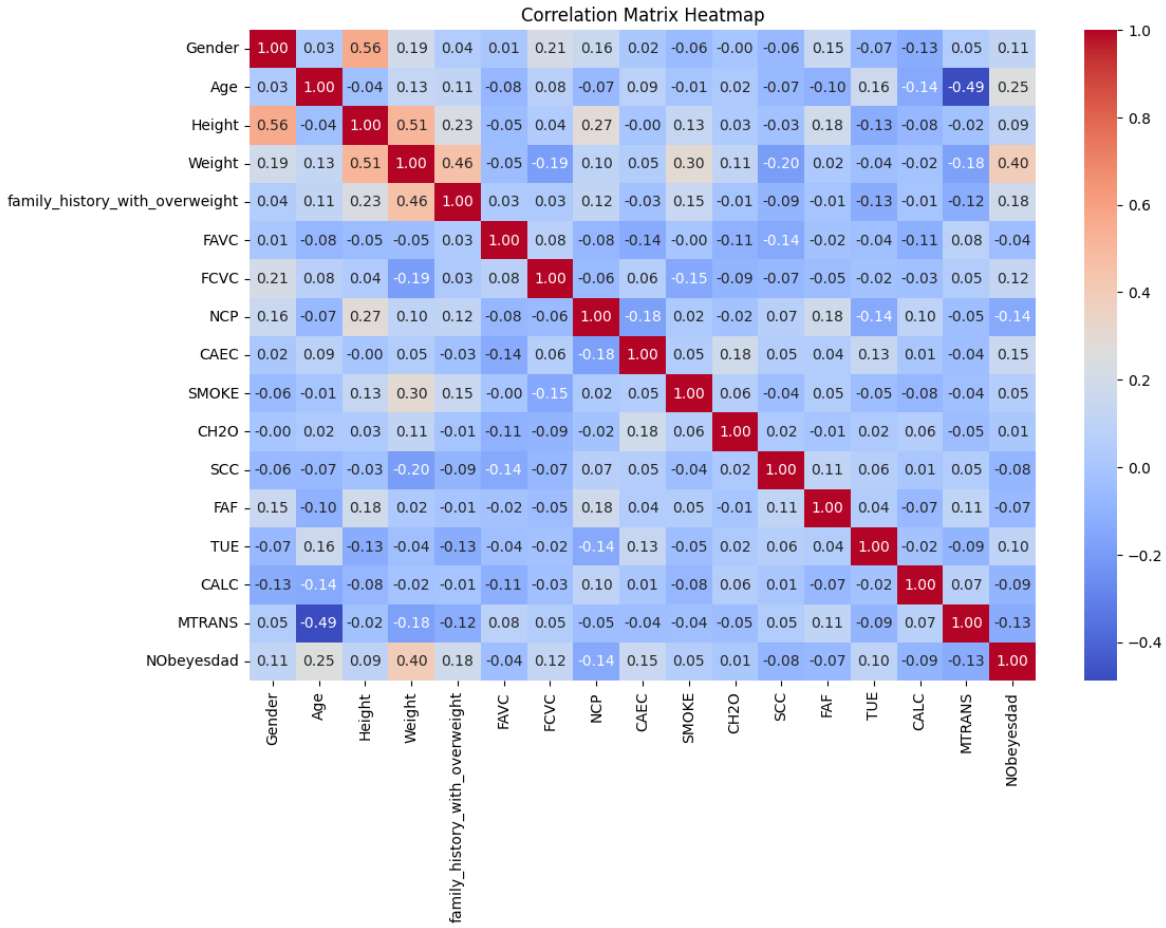


Figure 2: Correlation Matrix Heatmap illustrating the relationships between various health and lifestyle variables within the dataset.

important for understanding analytical framework and guiding the application of machine learning techniques to our data.

5.1 Feature selection: ANOVA F-Test

In the preliminary phase, we utilized the ANOVA F-test for feature selection, a statistical method that helps identify significant predictors by comparing variances across multiple classes of obesity. This approach pinpoints the features that hold the greatest discriminative power for our classification task, ensuring that the models are trained on variables that most effectively distinguish between the classes.

By focusing on these key features, we improve the efficiency and accuracy of our models. This streamlined approach not only enhances the performance of our predictive models but also aids in understanding the underlying factors of obesity. Such insights are valuable for developing targeted healthcare interventions and strategies.

5.2 Predictive Models

5.2.1 Logistic Regression

Logistic Regression, especially in its multinomial form, is highly effective for multi-class classification tasks such as predicting different obesity levels in our study. This model is capable of handling multiple categories smoothly by estimating the probabilities that a given input belongs to each class. This ability makes it particularly suitable for medical and health-related problems where understanding the probability of various outcomes is crucial.

In our research, multinomial logistic regression's use of the softmax function allows it to manage the various categories of obesity effectively. It estimates the likelihood of each category with a logistic function, ensuring that the output probabilities are well-calibrated and easy to interpret. This makes the model not only powerful in predicting but also helpful in analyzing the importance of different predictors. Its robustness against overfitting, thanks to techniques like L1 regularization, ensures that our model remains accurate and reliable even when dealing with complex, high-dimensional data.

5.2.2 Gradient Boost

Gradient Boosting is a powerful ensemble technique that builds models sequentially, with each new model correcting errors made by the previous ones. This approach is particularly useful for our obesity classification task as it allows for incremental improvements, making the model increasingly accurate with each step. GB uses decision trees as base learners, which can handle both numerical and categorical data efficiently, making it versatile for our diverse dataset.

Gradient Boosting is beneficial for our dataset because it combines multiple weak learners to form a strong learner, ensuring robust predictions. The method's ability to focus on mistakes and misclassifications from earlier trees allows it to improve where it is most needed, potentially leading to better performance on complex classification tasks like predicting different obesity levels, where subtle distinctions between classes can be crucial.

5.2.3 Support Vector Machine

Support Vector Machine (SVM) is renowned for its effectiveness in high-dimensional spaces, which is ideal for our study that involves multiple predictors of obesity. SVM works by finding the hyperplane that best divides the classes with the maximum margin, thus ensuring that the model is not only accurate but also generalizable to new data. This is particularly important in medical datasets where precision is critical.

For our research on obesity, SVM's ability to handle nonlinear relationships through the use of kernel functions makes it an excellent choice. It can model complex relationships between features and obesity levels, which are not necessarily linear, enhancing the model's ability to differentiate between closely spaced outcomes effectively.

5.2.4 Random Forest

Random Forest is an ensemble learning method known for its high accuracy, robustness, and ease of use. By constructing multiple decision trees and aggregating their predictions, RF reduces the risk of overfitting, making it reliable for our comprehensive dataset on obesity. It's capable of handling a large number of input variables without variable deletion, which is advantageous for analyzing the numerous factors influencing obesity.

Random Forest's strength lies in its ability to provide importance scores for each feature, giving insights into which factors are most influential in predicting obesity levels. This feature is invaluable for identifying key predictors and understanding the underlying patterns in obesity prevalence, assisting in targeted intervention strategies.

5.2.5 Multi-Layered Perceptron

Multi-Layer Perceptron (MLP) is a type of neural network suitable for complex pattern recognition, which occurs in multifactorial diseases like obesity. MLP can approximate any non-linear function, which is essential for modeling the intricate interactions between different predictors of obesity. Its structure with multiple hidden layers enables the model to learn deep representations at various abstraction levels, capturing subtle nuances in the data.

MLP's flexibility in architecture design allows it to be tailored specifically for the task at hand, potentially improving its effectiveness in distinguishing between different classes of obesity. The ability of MLP to learn from large datasets and its effectiveness in handling intricate features make it a potent tool for our predictive analysis, ensuring comprehensive learning from the diverse factors presented in the dataset.

5.3 Model Optimization Techniques

Model optimization is critical for achieving high-performance outcomes. This involves a series of processes aimed at refining the model to enhance its accuracy and efficiency in predicting outcomes. This section delves into two pivotal components of model optimization: Cross Validation and Hyperparameter Tuning, each playing a vital role in enhancing the model's predictive power and reliability in real-world scenarios.

5.3.1 Cross Validation

Cross-validation is a robust statistical technique used to evaluate how well a predictive model performs. It involves partitioning the data into subsets, training the model on some subsets (training set) and testing it on others (validation set). This method helps to mitigate overfitting, ensuring that the model's ability to generalize to new data isn't compromised by its performance on just one set of data. Common methods include k-fold cross-validation, where the data is divided into k subsets and the model is evaluated k times, each time with a different subset as the test set and the others as the training set.

5.3.2 Hyperparameter Tuning

Hyperparameter tuning is an optimization process to find the most effective parameters for a predictive model, which are not directly learned from the training process. These parameters, such as the number of decision trees in a random forest or the learning rate of a neural network, significantly influence model performance. Techniques like grid search, where a range of hyperparameter values are systematically tested, or random search, which samples parameter settings at random, are used to find the optimal values. Effective hyperparameter tuning can dramatically enhance the performance of a model, particularly in complex datasets where the right settings are not intuitively obvious.

5.4 Performance Metrics

To effectively measure the performance of classification models, several metrics are commonly used: accuracy, recall, precision, and F1 score. These metrics are derived from the confusion matrix, which is a table used to describe the performance of a classification model on a set of test data for which the true values are known.

5.4.1 Confusion Matrix

A confusion matrix is a specific table layout that allows visualization of the performance of an algorithm. It includes four different combinations of predicted and actual values:

- **True Positives (TP):** The cases in which the model correctly predicted the positive class.
- **True Negatives (TN):** The cases in which the model correctly predicted the negative class.
- **False Positives (FP):** The cases in which the model incorrectly predicted the positive class (a type I error).
- **False Negatives (FN):** The cases in which the model incorrectly predicted the negative class (a type II error).

5.4.2 Accuracy

This measures the overall correctness of the model and is defined as the ratio of correctly predicted observations to the total observations.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

5.4.3 Recall

Recall, also known as sensitivity or True Positive Rate measures the ratio of correctly predicted positive observations to the all observations in actual class - yes. It is crucial for models where missing positive instances (FN) is costly.

$$\text{Recall} = \frac{TP}{TP + FN}$$

5.4.4 Precision

This assesses the ratio of correctly predicted positive observations to the total predicted positives. It is vital when the cost of a false positive is high.

$$\text{Precision} = \frac{TP}{TP + FP}$$

5.4.5 F1 Score

This is the weighted average of Precision and Recall. Therefore, this score takes both false positives and false negatives into account. It is particularly useful when the class distribution is uneven.

$$F1 \text{ Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

6 Experiments and Results

In our experimental setup, we selected 70% of the dataset for training the models and remaining 30% was used for testing their performance. This division was chosen to ensure a robust training process and maintaining substantial test set to evaluate the generalizability of the model across unseen data.

A comprehensive grid search strategy was used to optimize model hyperparameters and feature selection, ensuring robust predictive performance across multiple classifiers. We have utilized pipelines for logistic regression, support vector machine, gradient boosting, random forest, and neural networks. Each of these models were configured with a range of hyperparameters. These include regularization strengths, kernel types, learning rates and others (Table. 1). The selection of features was performed using the `SelectKBest` method with criteria from `ANOVA F-test`, varying the number of features to determine the optimal subset for classification.

Each model's hyperparameters were tuned through `GridSearchCV`, using cross-validation with 5 folds to evaluate model accuracy. This approach allowed us to assess the influence of different hyperparameters and feature sets on the classification accuracy, thereby identifying the most effective configurations of our dataset.

The optimal hyperparameters and features selected for each model demonstrated significant influence on model performance (illustrated in Table. 2). For Logistic Regression, a regularization strength (C) of 4.0, L1 penalty, and 'saga' solver were found effective with features including Gender, Age, and dietary habits. The Support Vector Machine utilized a high regularization value (C=100) and a linear kernel, enhancing accuracy with a similar feature set. Gradient Boosting and Random Forest models leveraged deep trees and multiple estimators, respectively, emphasizing family history and dietary inputs. Finally, the Multi-Layered Perceptron, with 'relu' activation and dual 50-node layers, confirmed the importance of lifestyle features in predicting outcomes. These configurations were tailored to maximize accuracy while addressing the specific characteristics of our obesity-related dataset.

The performance metrics of our models provide an overview of each algorithm's ability to classify ability to classify obesity accurately (Table. 3). Support Vector Machine exhibited high perfor-

Classifier	**Hyperparameter**	**Value Range or Type**	**Feature Count**
LogisticRegression	C	0.25-4	1 - 10 features
	max_iter	800-1000 with step value 50	
	penalty	l1, l2, elasticnet	
	multi_class	ovr, multinomial	
	solver	saga	
SVC	C	logspace(-3, 2, 6)	1 - 10 features
	kernel	rbf, linear	
	gamma	scale, auto	
GradientBoostingClassifier	n_estimators	100, 200, 300	5 - 10 features
	learning_rate	0.01, 0.1, 0.2	
	max_depth	3, 5, 7	
RandomForestClassifier	n_estimators	100, 200, 300, 400	1 - 10 features
	min_samples_split	2, 5, 10	
	min_samples_leaf	1, 2, 4	
	max_depth	None, 10, 20, 30	
	max_features	auto, sqrt, log2	
MLPClassifier	hidden_layer_sizes	(50), (100), (50, 50), (100, 50)	1 - 10 features
	activation	tanh, relu	
	solver	sgd, adam	
	max_iter	800-1000 with 50 step value	
	learning_rate_init	0.001-0.01	

Table 1: Summary of hyperparameters and feature counts for different classifiers used in the study.

Classifier	Hyperparameter	Features selected
LogisticRegression	C = 4.0 max_iter = 1000 penalty = 'l1' multi_class = 'multinomial' solver = 'saga'	Gender, Age, Height, Weight, family_history_with_overweight, FCVC, SMOKE
SVC	C = 100 kernel = 'linear' gamma = 'scale'	Gender, Age, Height, Weight, family_history_with_overweight, FCVC, NCP, SMOKE
GradientBoostingClassifier	n_estimators = 200 learning_rate = 0.2 max_depth = 5	Gender, Age, Height, Weight, family_history_with_overweight, FCVC, NCP, CAEC, SMOKE
RandomForestClassifier	n_estimators = 300 min_samples_split = 2 min_samples_leaf = 1 max_depth = None max_features = 'log2'	Gender, Age, Height, Weight, family_history_with_overweight, FCVC, NCP, SMOKE
MLPClassifier	hidden_layer_sizes = (50, 50) activation = 'relu' solver = 'sgd' max_iter = 800 learning_rate_init = 0.01	Gender, Age, Height, Weight, family_history_with_overweight, FCVC, NCP, SMOKE

Table 2: Summary of Models, Hyperparameters, and Features Selected

Table 3: Model Performance Metrics

Model	**Accuracy**	**Precision**	**Recall**	**F1 Score**
Logistic Regression	98.01	98.04	98.01	98.00
Support Vector Machine	99.17	99.18	99.17	99.17
Gradient Boost	96.85	96.85	96.85	96.84
Random Forest	95.36	95.42	95.36	95.34
Multi-Layered Perceptron	98.18	98.21	98.18	98.17

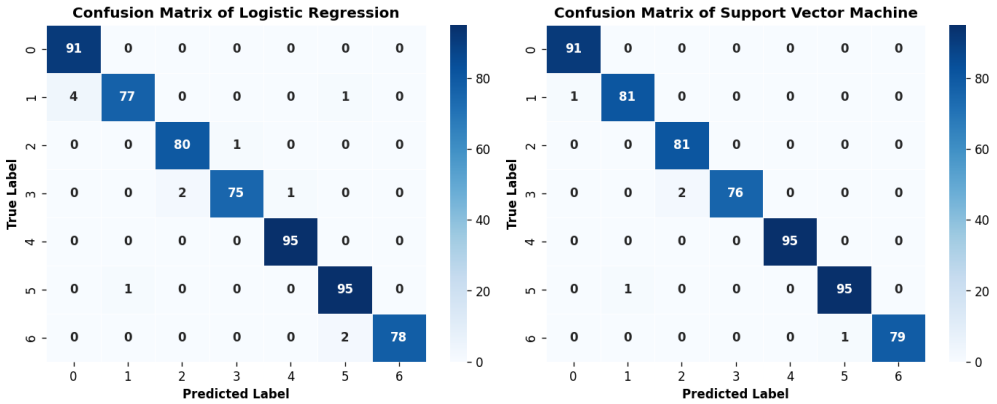
mance across all metrics, achieving near-perfect scores with accuracy and an F1 score of 99.17%. Logistic Regression and the Multi-Layered Perceptron also demonstrated high effectiveness, with F1 scores around 98%, indicating robust precision and recall. In contrast, Gradient Boost and Random Forest showed relatively lower performance, with F1 scores of 96.84% and 95.34% respectively. This variance underscores the impact of model choice and feature selection on predictive accuracy and reliability in classifying complex health conditions like obesity.

If we look at the confusion matrices of each classifier (Fig. 3), the Support Vector Machine (SVM) stands out as the most effective classifier, showing superior performance in accurately predicting all obesity levels with high precision and fewer errors. Conversely, the Random Forest model appears less adept for this task, as evidenced by its higher rate of misclassifications and less consistent accuracy across the various classes.

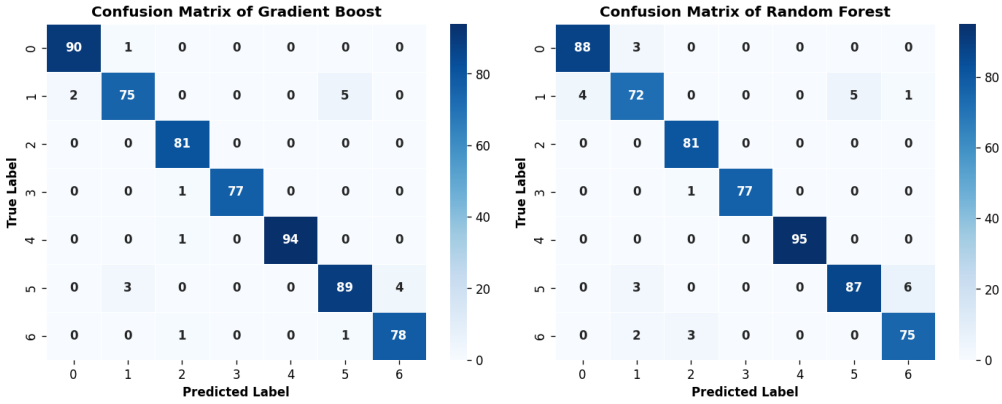
7 Discussion

Our research aimed at predicting obesity levels based on dietary habits and physical conditions has shown promising results, confirming many of our initial hypotheses. We utilized various machine learning models, finding that especially the Support Vector Machine (SVM) performed exceptionally well, achieving about 99.17% in metrics like accuracy, precision, recall, and the F1 score. Logistic Regression and the Multi-Layered Perceptron also demonstrated commendable performances, making them suitable for obesity prediction. In contrast, models like Gradient Boost and Random Forest exhibited slightly lower effectiveness, highlighting a potential trade-off between model complexity and interpretability.

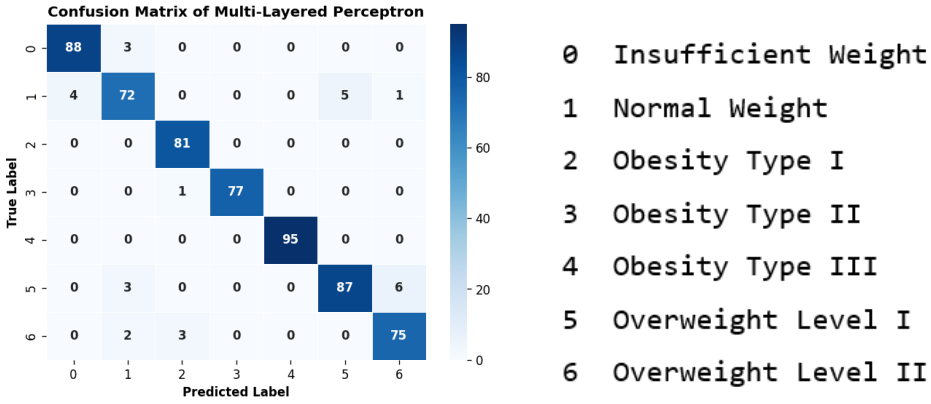
The analysis shows that specific variables—such as family history of obesity, dietary habits, and physical activities—play significant roles in predicting obesity. This aligns with existing medical insights, emphasizing the complex interactions between genetic and lifestyle factors in determining obesity levels. Our study illustrates the potential of machine learning to enhance medical diagnostics and interventions for obesity, a major global health challenge. By identifying at-risk individuals early, the predictive models we developed could enable more timely and effective healthcare interventions.



(a) Logistic Regression (b) Support Vector Machine



(c) Gradient Boost (d) Random Forest



(e) Multi-Layered Perceptron (f) Obesity Level Legend

Figure 3: Confusion Matrix of predicted data for each respective models. (f) Obesity Levels representation in the confusion matrices.

7.1 Impact on Global Health Outcomes

The application of these predictive models could significantly impact global health by aiding early detection and intervention efforts, potentially reducing the rising trends of obesity and its related health issues. This ability to predict obesity levels accurately underscores the importance of machine learning in public health strategies, offering a powerful tool for healthcare providers to manage and prevent obesity effectively.

7.2 Importance of Feature Selection

Our findings also underscore the critical role of feature selection in enhancing the accuracy of predictive models. The SVM's high performance can be attributed to its effective use of crucial features such as family history, specific dietary habits, and lifestyle choices like smoking. These factors are strongly linked to obesity, suggesting that they should be prioritized in future models to improve prediction accuracy. Understanding these relationships helps in developing robust models that can adapt to various data properties and meet different analytical needs.

7.3 Limitations

Our study has shown significant potential in predicting obesity levels using machine learning models, but it does have some limitations. Firstly, the dataset we used mainly came from Mexico, Peru, and Colombia, which may not represent global populations adequately. This raises concerns about the generalizability of our findings to other regions. Secondly, a substantial part of our data was synthetic, created to balance class distribution. While helpful for model training, this synthetic data may not accurately reflect real-world data nuances, potentially affecting the model's performance outside of a controlled setting. Moreover, the correlation matrix revealed complex interactions between variables, indicating that our current models and feature engineering methods might need enhancement to capture these dynamics more effectively.

7.4 Future Work

To address the limitations identified, future research should aim to integrate larger and more diverse datasets. This expansion would enhance the robustness and generalizability of our models across different populations and provide deeper insights into obesity's influencing factors. Employing advanced machine learning techniques, such as deep learning, could also prove advantageous. These methods are adept at discerning more intricate patterns within data, potentially leading to higher prediction accuracy. Moreover, it is crucial to test these models in real-world settings. Applying the models in clinical environments or through longitudinal studies would help validate their effectiveness and utility, confirming their potential for practical use in healthcare to assist in early obesity detection and intervention efforts.

References

[1] S. M. Fruh, "Obesity: Risk factors, complications, and strategies for sustainable long-term weight management," *J Am Assoc Nurse Pract*, vol. 29, no. S1, S3–S14, 2017. DOI: 10.1002/2327-6924.12510.

[2] C. M. Apovian, "Obesity: Definition, comorbidities, causes, and burden," *Am J Manag Care*, vol. 22, no. 7 Suppl, s176–s185, 2016.

[3] D. Khanna, C. Peltzer, P. Kahar, and M. S. Parmar, "Body mass index (bmi): A screening tool analysis," *Cureus*, vol. 14, no. 2, e22119, 2022. DOI: 10.7759/cureus.22119.

[4] K. W. DeGregory *et al.*, "A review of machine learning in obesity," *Obes Rev*, vol. 19, no. 5, pp. 668–685, 2018. DOI: 10.1111/obr.12667.

[5] R. Blagus and L. Lusa, "Smote for high-dimensional class-imbalanced data," *BMC Bioinformatics*, vol. 14, p. 106, 2013. [Online]. Available: https://doi.org/10.1186/1471-2105-14-106.

[6] D. C. P. Buani and N. Nuraeni, "Application of xgb classifier for obesity rate prediction," *Jurnal Riset Informatika*, vol. 6, no. 1, pp. 1–6, 2023. DOI: 10.34288/jri.v6i1.260.

[7] A. Choudhuri, "A hybrid machine learning model for estimation of obesity levels," *medRxiv*, 2022. DOI: 10.1101/2022.08.17.22278905.

[8] F. M. Palechor and A. de la Hoz Manotas, "Dataset for estimation of obesity levels based on eating habits and physical condition in individuals from colombia, peru and mexico," *Data in Brief*, vol. 25, p. 104 344, 2019, ISSN: 2352-3409. DOI: https://doi.org/10.1016/j.dib.2019.104344. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2352340919306985.

[9] H. G. G. Bag, F. H. Yagin, Y. Gormez, *et al.*, "Estimation of obesity levels through the proposed predictive approach based on physical activity and nutritional habits," *Diagnostics*, vol. 13, p. 2949, 2023. DOI: 10.3390/diagnostics13182949. [Online]. Available: https://doi.org/10.3390/diagnostics13182949.