# Estimating Obesity Levels in Individuals *(Using Machine Learning Techniques)*

A Machine Learning Approach to Predicting Obesity Levels Based on Lifestyle and Health Data

Group 11

Amina Danlami, Sara Polley Farjoodi, Ben Earl-Kinley, Victor Isaac Adeyeye, Ethan Sadler, Abdinasir Ali

CP322-A - Machine Learning

Yang Liu

Wilfrid Laurier University

December 8, 2024

# Introduction

### What is the problem?

Obesity has become a significant global issue, and the numbers continue to rise. It's not just about weight; obesity is closely linked to serious health problems like diabetes, heart disease, and even cancer. The ability to identify the factors contributing to obesity and predict obesity levels in people can help healthcare systems provide specific advice and preventative measures to individuals who are at risk. This proactive approach could reduce the long-term burden on healthcare services and improve overall health outcomes.

The goal of this project is to estimate obesity levels in individuals by analyzing personal, lifestyle, and health-related factors from a dataset. By focusing on these elements, the aim is to develop a model that can predict a person's obesity level, providing valuable insights without the need for expensive or time-consuming medical tests.

Our dataset contains data from three countries: Mexico, Peru, and Colombia. It includes 17 features related to individuals' daily habits, physical conditions, and health histories. Some key features in the dataset include age, weight, height, family history of being overweight, frequency of eating high-calorie foods, and physical activity levels. Using this information, individuals will be classified into one of seven obesity levels: Insufficient Weight, Normal Weight, Overweight Level I, Overweight Level II, Obesity Type I, Obesity Type II, and Obesity Type III. A significant portion of the dataset is synthetic, generated to balance the data using advanced techniques, while the rest was collected directly from people via surveys. This blend of synthetic and real data ensures a diverse dataset that reflects a wide range of obesity levels across different lifestyles and conditions.

### Why can't any of the existing techniques effectively tackle this problem?

Traditional methods for predicting obesity levels face several challenges. These techniques often rely heavily on direct clinical data, such as blood tests or other medical diagnostics. While these methods are effective, they can be expensive, time-consuming, and not easily accessible to everyone. Also, they typically require individuals to visit healthcare facilities, which can delay early identification and intervention.

Machine learning techniques can help analyze how lifestyle choices, personal traits, and family history are connected to predict obesity levels more accurately and easily. The aim is to develop a reliable model capable of quickly and efficiently estimating a person's obesity level using easily accessible data, eliminating the need for extensive clinical tests. This model will help to identify at-risk individuals early on to prevent extensive health issues. By comparing several machine learning models including Logistic Regression, Decision Trees, k-nearest Neighbors (kNN), and others, this paper identifies the most accurate, precise, and practical solution for estimating obesity levels.

### What is the intuition behind the technique that you have developed?

For our specific problem, the k Nearest Neighbors algorithm and Decision Tree Classifier were chosen as the primary models. kNN was chosen for its simplicity and effectiveness in

classification tasks since it compares an individual's data to their closest neighbors in the dataset. The Manhattan distance metric was used to optimize kNN, achieving strong accuracy and precision. However, the Decision Tree outperformed kNN and other models, achieving the highest precision of 99.37% while maintaining interpretability, making it a standout choice for healthcare applications.

By using these machine learning techniques and comparing them against established models like Logistic Regression and Random Forest, our model not only aims to predict obesity levels accurately but also provides a framework for applying machine learning to other health-related challenges. The findings from this work emphasize the potential of data-driven methods to revolutionize preventive healthcare strategies.

## Techniques to Tackle the Problem

Brief review of previous work concerning this problem

Predicting obesity levels using machine learning models has been an area of focus for researchers due to the global prevalence of obesity and its health implications. Some of the insights from past studies are as follows:

**Data-Driven Approaches:**

- Past research has shown that lifestyle and behavioral data (e.g., dietary habits, physical activity levels, family history) provide critical insights into obesity risk factors.
- Datasets that combine real and synthetic data are used to address imbalance and sparsity issues.

**K-Nearest Neighbors (kNN):**

- Widely used in healthcare for classification tasks due to its simplicity and effectiveness in comparing individuals based on feature proximity.
- Studies highlight the use of distance metrics like Manhattan or Euclidean to optimize performance.

**Decision Trees:**

- Decision Trees are effective in handling categorical and numerical data and provide interpretable insights.
- They have been applied successfully to segment individuals based on BMI-related features.

**Other Models Found:**

- Logistic regression and Naive Bayes have been explored but are often constrained by feature independence assumptions or linear separability issues.
- Deep learning has shown promise for large-scale datasets but requires high computational resources.

## Elaboration of the technique developed

The implemented technique for classifying obesity levels in individuals involves a machine learning pipeline including data preprocessing, feature selection, model implementation, and evaluation. Data preprocessing began with handling missing values, where the dataset was checked and confirmed to have no missing entries. Outliers in numerical features such as Age, Height, and Weight were removed using the Interquartile Range (IQR) method to reduce noise and improve data quality. Categorical variables, including family_history_with_overweight, MTRANS, and NObeyesdad, were encoded using one-hot encoding to convert them into binary variables suitable for machine learning models. In addition, the obesity levels were consolidated into a binary classification (obese = True/False) to simplify the target variable and enhance model performance. In total, 829 people were classified as obese, and 291 people were classified as not obese in this adjusted target variable.

For feature selection, the most relevant features were identified based on domain knowledge and exploratory data analysis (EDA). These features included numerical variables like Height, Weight, and FAF (frequency of physical activity), along with encoded binary columns derived from categorical variables such as transportation modes and family history of obesity. This selection ensured the inclusion of features with the highest predictive value.

The k-nearest Neighbors (kNN) algorithm and Decision Tree Classifier were utilized for model implementation. For kNN, both Manhattan and Euclidean distance metrics were evaluated using 10-fold cross-validation, with Manhattan distance demonstrating better accuracy due to its emphasis on feature-wise differences. The optimal number of neighbors (k) was determined to be 3, achieving a test set accuracy of 94.6% and a precision of 94.7%. The Decision Tree Classifier was trained and validated using an 80-20 train-test split, outperforming kNN with an accuracy of 96.9% and a precision of 99.4%. The Decision Tree's performance was visualized using a confusion matrix, which highlighted its effectiveness in handling mixed feature types and capturing complex relationships within the data.

The models were evaluated using accuracy and precision metrics, with the Decision Tree proving to be the best-performing model due to its higher scores. The machine learning pipeline effectively classified individuals into obese and non-obese categories using features related to health, lifestyle, and habits. This interpretable approach to obesity classification demonstrated the strength of the chosen models, particularly the Decision Tree.

## Description of the existing techniques that will be used for comparison

To analyze and compare the developed approach for estimating obesity levels, the following existing machine learning techniques, relevant to health and lifestyle datasets, are considered as benchmarks.

**Logistic Regression:**

Logistic regression is a statistical model used for binary classification. It predicts probabilities for classes based on a linear combination of input features. As a baseline model, logistic regression is simple and interpretable, making it suitable for identifying relationships between health and lifestyle features and obesity levels. However, its assumption of linear relationships between features and the target variable limits its applicability to more complex patterns.

**Naive Bayes:**

Naive Bayes is a probabilistic model that assumes feature independence, making it computationally efficient and effective for categorical data. While the independence assumption is unrealistic for correlated features (e.g., dietary habits and physical activity), Naive Bayes is a useful benchmark due to its simplicity and ability to handle categorical data.

**Random Forest:**

Random Forest is an ensemble learning method that constructs multiple decision trees and aggregates their predictions. It is robust against overfitting and works well with mixed data types. It can serve as a stronger alternative to single decision trees by capturing complex interactions between health and lifestyle features. It provides insights into the importance of features, which is valuable in healthcare applications.

**Support Vector Machines (SVM):**

SVMs classify data by finding a hyperplane that maximizes the margin between classes. With kernel functions, they can model non-linear relationships. They are effective for datasets with clear class boundaries. They can be used to compare the linear and non-linear capabilities of kNN and decision trees in the context of obesity estimation.

**Weighted k-Nearest Neighbors (kNN):**

A variant of kNN where neighbors closer to the test point have a greater influence on classification decisions. Weighted kNN provides a natural extension to the standard kNN approach. It allows for comparison to evaluate whether emphasizing closer neighbors improves classification performance for health and lifestyle data.

By comparing the developed kNN and Decision Tree models with these techniques, we can evaluate their performance in estimating obesity levels both empirically—using metrics like accuracy, precision, and F1-score—and theoretically, in terms of complexity, interpretability, and suitability for health-related datasets.

## Evaluation

**- Evaluation Goals**

The objective of this evaluation is to analyze the performance of K-Nearest Neighbors (kNN) and Decision Trees to estimate obesity levels compared to the baseline models: Logistic Regression, Naïve Bay, Random Forest, and Support Vector Machine. Evaluations based on precision, recall, F1 score, and ROC-AUC are also taken into account.  It also considers the interpretability, complexity, and suitability of models for healthcare data.

- **Evaluation Metrics**

The following metrics were used:

- Accuracy: Percentage of correctly classified obesity levels.
- Precision: Proportion of true positive predictions out of all positive predictions.
- Recall: Proportion of true positive predictions out of all actual positive cases

- F1-Score: Harmonic mean of precision and recall, balancing false positives and false negatives
- ROC-AUC: Measures model ability to distinguish between classes.

**- Results**

We evaluated the following machine learning models for estimating obesity levels in individuals: k-nearest Neighbors (kNN), Decision Trees, Random Forest, Support Vector Machines (SVM), Logistic Regression, and Naive Bayes. Logistics and innocent base Models are evaluated based on their accuracy and precision. Models were evaluated based on accuracy and precision.

| **Model** | **Accuracy** | **Precision** |
|---|---|---|
| K-Nearest Neighbors **(**kNN) | 0.9375 | 0.9521 |
| Decision Tree | 0.9643 | 0.9937 |
| Random Forest | 0.9509 | 0.9873 |
| Support Vector Machines (SVM) | 0.9464 | 0.9581 |
| Logistic Regression | 0.9732 | 0.9877 |
| Naive Bayes | 0.8661 | 0.9592 |

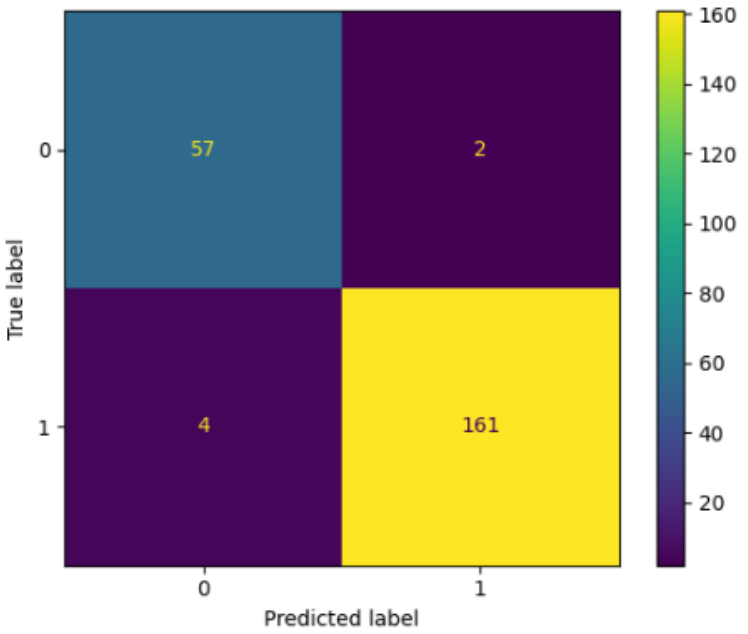
Logistic Regression achieved the highest accuracy with **97.32%**
Decision Tree had the highest precision with **99.37%**
Naive Bayes had the lowest accuracy with **86.61%** but maintained a reasonable precision of **95.92%**

**- Analysis**

Empirical Observations:

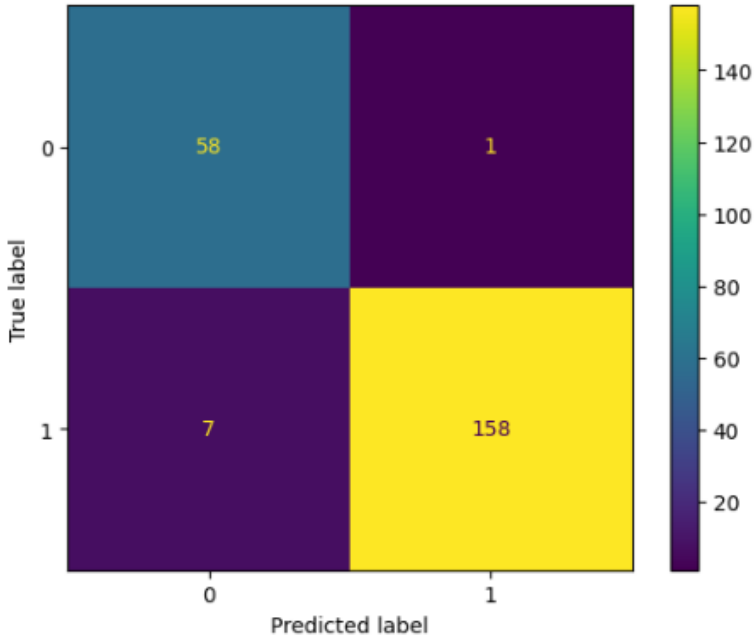
1. **Logistic Regression:** Its simplicity and ability to handle linear relationships effectively provided the highest accuracy. The model performed well due to the dataset's balance after preprocessing and its ability to capture clear boundaries between obesity classes. Precision was also strong at 98.77%, indicating reliable predictions with minimal false positives.

Logistic Regression Accuracy: 0.9732142857142857
Logistic Regression Precision: 0.9877300613496932




2. **Decision Tree:** This excelled with a high accuracy of 96.43% and the best precision of 99.37%. This can be attributed to their ability to handle non-linear relationships and mixed data types. They also provided interpretable insights into feature importance, making them valuable for healthcare applications.

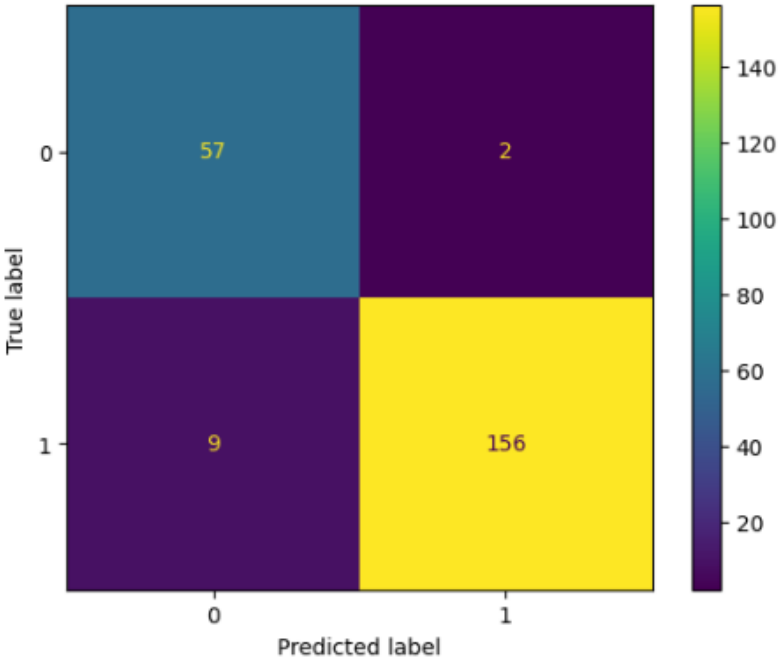
The accuracy of the data is 0.9642857142857143
The precision of the data is 0.9937106918238994




3. **Random Forest:** Random Forest delivered a high accuracy of 95.09% and a high precision of 98.73% by combining multiple decision trees. Its performance was slightly

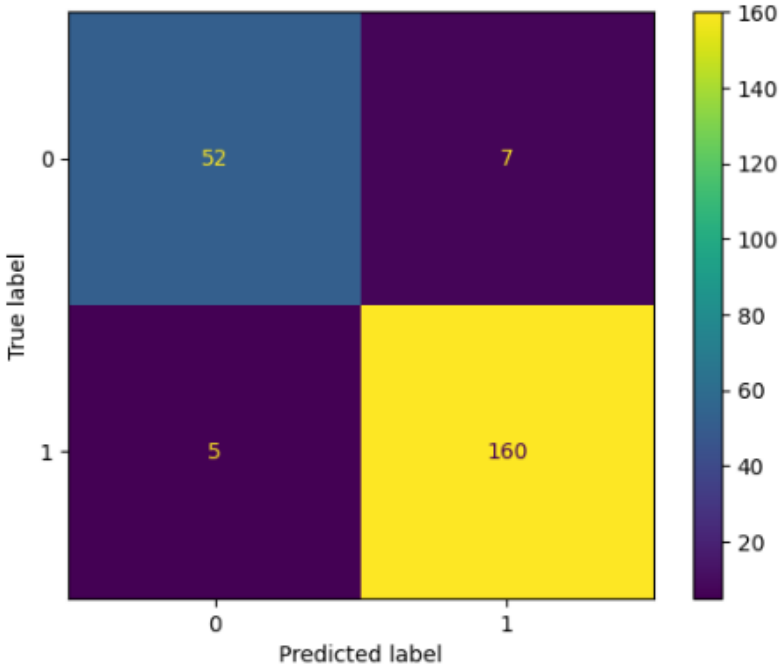
lower than logistic regression. This is because there is potential overfitting on specific obesity classes.

```
Random Forest Accuracy: 0.9508928571428571
Random Forest Precision: 0.9873417721518988
```




4. **Support Vector Machines** (SVM): Support Vector Machines (SVM) show balanced performance. It has an accuracy of 94.64% and a precision of 95.81%. It effectively modeled non-linear relationships using kernel functions. But fell short compared to ensemble methods like Random forest.

```
SVM Accuracy: 0.9464285714285714
SVM Precision: 0.9580838323353293
```

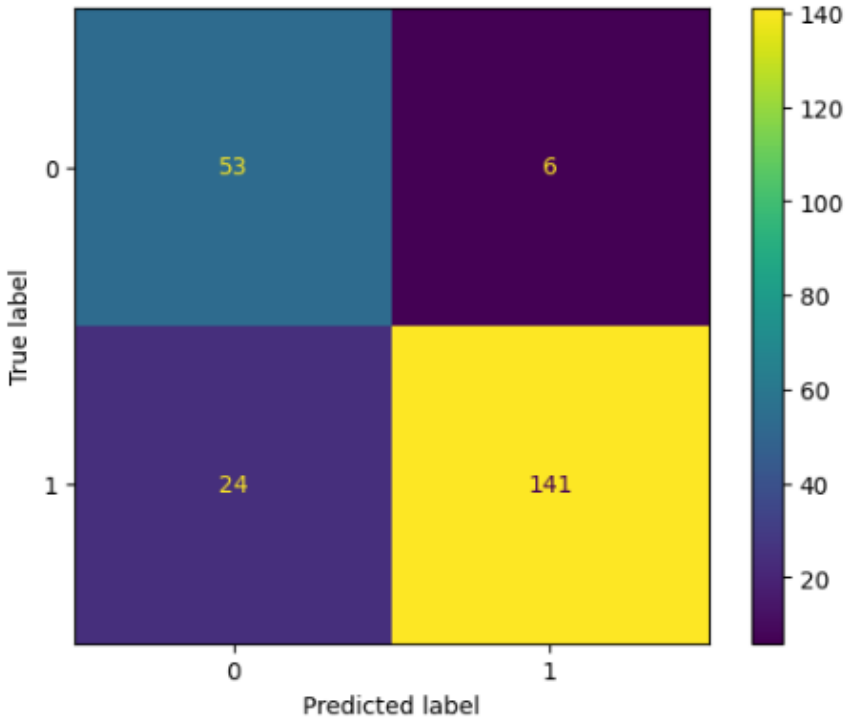


5. **k-Nearest Neighbors** (kNN): kNN achieved good accuracy of 93.75% and precision of 95.21% with Manhattan distance and an optimal neighbor count of 3. It's simple which makes it easy to implement, but it relies on distance metrics which makes it difficult to apply in high dimensional data.

```
The optimal k value is 4, with manhattan as the metric.
The accuracy of the dataset with 4 neighbors is 0.9375
The precision of the dataset with 4 neighbors is 0.9520958083832335
```

6. **Naive Bayes**: Naive Bayes performed the worst, with an accuracy of 86.61% but a decent precision of 95.92%. The independence assumption limited its ability to model correlated features effectively.

```
Naive Bayes Accuracy: 0.8660714285714286
Naive Bayes Precision: 0.9591836734693877
```




- **Evaluation Conclusion**

The evaluation showed that logistic regression was the best model to estimate obesity levels. It has the highest accuracy of 97.32% and high precision of 98.77%, with computational efficiency and interpretability. This makes it suitable for healthcare applications. However, where it is necessary to reduce false positives. The decision tree has the highest precision of 99.37%, resulting in lower misclassifications of obesity levels.

## Conclusion

Our paper highlights the importance of using machine learning tools to address a critical global health challenge, obesity. By developing a model to predict obesity levels based on lifestyle, health, and personal traits, we can provide accessible, efficient, and accurate tools to assist with

early intervention efforts. Among the models evaluated, the Logistic Regression model achieved the highest accuracy, emphasizing its ability to identify clear boundaries within the dataset. But in regards to precision, the decision tree model did the best. It gave a diverse prediction with fewer false positives.

Machine learning methods like K-Nearest Neighbours and Random Forest also demonstrated strong performances, showing their usefulness in handling complex data. K-Nearest Neighbours, which is the simplest of the models we evaluated, demonstrated an intuitive approach to classification while Random Forest's ensemble approach effectively took into account nuanced patterns. Support Vector Machines and Naive Bayes, which are less dominant, highlighted the trade-offs between computational efficiency and accuracy when dealing with non-linear and independent feature assumptions.

By testing all of these models, our paper highlights the significance of tailored approaches in healthcare. Decision Trees and Logistic Regression with their easy interpretability and high precision stand out as the best choices for real-world applications. These models' ability to integrate easily with healthcare systems and produce useful insights can drive meaningful improvements in obesity management and prevention.

Lastly, the findings confirm that machine learning models are not just tools for prediction but triggers for change. With specific and detailed implementation, they can transform data in decisions, empowering healthcare providers to combat obesity more efficiently. As technology continues to advance, these innovative approaches will remain important in tackling other world health issues.

## References


A;, P.-A. M. (n.d.). *Obesity, metabolic health and OMICS: Current status and future directions*. World journal of diabetes. https://pubmed.ncbi.nlm.nih.gov/33889288/

Alsareii, S. A., Shaf, A., Ali, T., Zafar, M., Alamri, A. M., AlAsmari, M. Y., Irfan, M., & Awais, M. (2022, September 10). *IOT framework for a decision-making system of obesity and overweight extrapolation among children, youths, and adults*. Life (Basel, Switzerland). https://pmc.ncbi.nlm.nih.gov/articles/PMC9500775/

Jeon, J., Lee, S., & Oh, C. (2022, December 6). *Age-specific risk factors for the prediction of obesity using a machine learning approach*. Frontiers. https://www.frontiersin.org/journals/public-health/articles/10.3389/fpubh.2022.998782/full

Mehrparvar, F. (2024, April 7). *Obesity levels*. Kaggle. https://www.kaggle.com/datasets/fatemehmehrparvar/obesity-levels

Osadchiy, V., Bal, R., Mayer, E. A., Kunapuli, R., Dong, T., Vora, P., Petrasek, D., Liu, C., Stains, J., & Gupta, A. (2023, April 4). *Machine learning model to predict obesity using gut metabolite and brain microstructure data*. Nature News. https://www.nature.com/articles/s41598-023-32713-2

Putri, A. I., Husna, N. A., Cia, N. M., Arba, M. A., Aisyi, N. R., Pramesthi, C. H., & Irdayusman, A. S. (2024). Implementation of K-nearest neighbors, naïve Bayes classifier, support vector

machine and decision tree algorithms for obesity risk prediction. *Public Research Journal of Engineering, Data Technology and Computer Science*, *2*(1), 26–33. https://doi.org/10.57152/predatecs.v2i1.1110

World Health Organization. (n.d.). *Obesity and overweight*. World Health Organization. https://www.who.int/news-room/fact-sheets/detail/obesity-and-overweight