Obesity Insight Predictor

Predicting Obesity Risk Based on Lifestyle and Demographic Factors Using Machine Learning

1st Srijita Banerjee
Computer Science, College of Engineering
University of Illinois at Chicago
Chicago, United States
sbane4@uic.edu

2nd Blossom Egbuonu
Computer Science, College of Engineering
University of Illinois at Chicago
Chicago, Illinois
begbu2@uic.edu

3rd Muhammad Ali Khurram
Computer Science, College of Engineering
University of Illinois at Chicago
Chicago, United States
mkhur4@uic.edu

4th Mohammad Nusairat
Computer Science, College of Engineering
University of Illinois at Chicago
Chicago, United States
mnusa2@uic.edu

***Abstract*—This project develops a predictive model to assess obesity risk based on demographic and lifestyle factors, such as diet and physical activity. Using logistic regression and Naive Bayes, our approach highlights key predictors and provides actionable recommendations to reduce obesity risk. This model offers practical insights to support personal and public health initiatives."**

***Index Terms*—Obesity, Prediction, Lifestyle, Machine Learning, Logistic Regression, Naive Bayes, Health, Demographic, Preventive Health, Data-Driven Health Insights, Public Health Informatics, Risk**

I. Introduction

Obesity poses serious health risks globally; this project develops a machine learning model using demographic and lifestyle data to predict obesity risk and provide actionable insights for prevention.

II. Related Work

A. Work that already exists

Previous studies have explored the link between lifestyle factors—such as diet, physical activity, smoking, and alcohol consumption—and obesity risk. Traditional statistical models have been used to analyze these factors, but they often rely on a limited set of predictors, reducing the scope for personalized recommendations. Recent advancements in machine learning, particularly logistic regression and Naive Bayes, have improved the interpretability and accuracy of health risk predictions by incorporating a wider range of variables. However, most models lack actionable insights for lifestyle modification, focusing instead on classification without practical recommendations. Our approach seeks to fill this gap by integrating diverse demographic and lifestyle data to create a model that not only predicts obesity risk but also offers tailored guidance on risk reduction, thus contributing to personal and public health efforts.

Identify applicable funding agency here. If none, delete this.

III. Proposed Methodology

For this project, we will use logistic regression and Naive Bayes as our primary modeling methods due to their simplicity, interpretability, and suitability for categorical and continuous data. Logistic regression will allow us to interpret the influence of each lifestyle factor on obesity risk, while Naive Bayes offers a robust, probabilistic approach, especially effective for handling categorical variables like dietary and exercise habits. These models are computationally efficient, making them feasible to train and deploy within the project timeline, even with a dataset of nearly 20,000 records. By using a combination of these models, we aim to achieve both accurate predictions and actionable insights, ensuring our approach remains interpretable and practical for real-world applications.

A. Data Descriptions

The primary dataset includes nearly **20,000 records** with **17 features** covering **demographic and lifestyle factors** relevant to obesity, such as **age, gender, height, weight, smoking, alcohol consumption, dietary preferences, and exercise habits**. Our goal is to use these features to predict obesity risk. Initially, we will clean the data by handling missing values, removing outliers, and performing feature engineering, such as normalizing continuous variables and encoding categorical ones. After cleaning, we will split the data for training and testing, with cross-validation to ensure robust model performance. This preparation will make the data suitable for machine learning and keep the project on schedule.

If logistic regression and Naive Bayes models do not achieve the desired predictive accuracy or interpretability, we will implement a fallback plan that involves exploring alternative models and additional data preprocessing techniques. We will consider introducing ensemble methods, such as a simple decision tree or Random Forest, to capture non-linear relationships and feature interactions that may improve model performance.

Additionally, if limited data quality or variability in lifestyle factors poses challenges, we may supplement our dataset with publicly available health data to enrich feature diversity. This fallback approach ensures we maintain model reliability and actionable insights, adapting our strategy if initial results fall short of project goals.

B. Experimental Setup

- **Programming Language**: Python, chosen for its versatility and extensive libraries for data science and machine learning.
- **Libraries:**
  - **scikit-learn**: For implementing logistic regression and Naive Bayes models, offering simplicity and efficiency.
  - **pandas** and **NumPy**: To handle data preprocessing, including missing value imputation, feature normalization, and categorical encoding.
  - **Matplotlib** and **Seaborn**: For visualizing results, model performance, and feature importance.
- **Development Environment**: Jupyter Notebook, enabling modular code development, easy visualization, and iterative tuning of models.

C. Model Evaluation and Results

We will use performance metrics including accuracy, precision, recall, and F1-score. These metrics provide a comprehensive assessment of the model's predictive power and reliability in identifying individuals at varying levels of obesity risk.

- **Accuracy** will gauge the overall performance of the model in correctly predicting obesity risks across all classes.
- **Precision** and **recall** will be especially important in assessing the model's ability to correctly identify high-risk individuals, minimizing false positives and false negatives.
- **F1-score** will offer a balanced measure of precision and recall, particularly useful in handling any class imbalance within the dataset.
- Following baseline testing, we will conduct **cross-validation** to ensure the model's robustness and generalizability. By fine-tuning hyper-parameters, we aim to maximize the model's performance metrics, providing a reliable obesity risk prediction tool.

D. Interpretibility and Insights

Our model's interpretability is a core aspect, aimed at providing clear insights into how specific lifestyle and demographic factors contribute to obesity risk. By using **logistic regression**, we can identify the individual impact of each variable, such as diet, physical activity, smoking, and alcohol consumption, on obesity likelihood. The probabilistic nature of **Naive Bayes** further aids in managing categorical data, enabling an intuitive breakdown of risk associated with various lifestyle patterns.

The resulting insights will not only classify obesity risk but also reveal actionable factors, helping individuals understand which lifestyle changes—such as increased physical activity or dietary adjustments—can most effectively reduce their obesity risk. This interpretive approach supports public health initiatives by offering a framework for personalized health recommendations and broader community health strategies.

E. Feasibility and Fallback Plan

- **Project Timeline**: Structured over one month with designated phases for data processing, model training, and evaluation, ensuring completion within the available time and resources.
- **Primary Models**: Logistic regression and Naive Bayes are chosen for their computational efficiency, allowing us to work with a dataset of nearly 20,000 records without straining resources.
- **Fallback Plan**:
  - **Alternative Models**: If initial results are unsatisfactory, we will explore ensemble methods like decision trees or Random Forest to capture non-linear relationships and improve model performance.
  - **Supplemental Data**: Should limited data variability or feature diversity affect results, we will enhance the dataset with publicly available health data to increase robustness.
- **Outcome Assurance**: These steps ensure reliable and actionable insights even if the primary models do not meet performance expectations.

F. Our Project Timeline

- **Week 1**:
  - Data Preprocessing and Cleaning
  - Team members: Ali, Srijita
  - Tasks: Handle missing values, detect outliers, encode categorical features
- **Week 2**:
  - Model Setup and Baseline Testing
  - Team member: Mohammad
  - Tasks: Tasks: Implement logistic regression and Naive Bayes models, conduct initial testing
- **Week 3**:
  - Model Fine-Tuning and Cross-Validation.
  - Team members: Blossom, Ali
  - Tasks: Hyperparameter tuning, improve model accuracy, evaluate performance metrics
- **Week 4**:
  - Model Evaluation and Final Report
  - Team members: Srijita, Blossom
  - Tasks: Final model evaluation, ensure interpretability, generate final report with actionable insights