Impact of Lifestyle on Obesity

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

https://github.com/mohammadnusairat/Obesity-Insight-Predictor

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 Support Vectore Classifier SVC 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, SVC, Naive Bayes, Health, Demographic, Preventive Health, Data-Driven Health Insights, Public Health Informatics, Risk**

I. Introduction

Introduction

Obesity is a pressing global health issue that has escalated into a public health crisis over the past few decades. With nearly 1.9 billion adults classified as overweight and over 650 million considered obese, the implications are far-reaching, contributing to an increase in non-communicable diseases such as diabetes and cardiovascular disorders, as well as significant economic burdens on healthcare systems worldwide. Lifestyle factors, including dietary habits, physical activity, and alcohol consumption, play a pivotal role in obesity risk, interacting with demographic attributes such as age, gender, and socioeconomic status. Understanding and addressing these interactions is crucial for effective prevention strategies.

Machine learning presents a powerful approach to tackling this challenge by analyzing complex datasets to uncover patterns and relationships beyond the scope of traditional statistical models. This project focuses on developing a predictive model that assesses obesity risk using a combination of demographic and lifestyle data. By employing Support Vector Classifier and Naive Bayes, the methodology emphasizes both accuracy and interpretability, ensuring the results are actionable for public health initiatives. SVC handles high dimensional data, with strong margins for classification, while Naive Bayes handles categorical data effectively, making these models suitable for personalized obesity risk predictions.

Using a dataset of nearly 20,000 records with 17 features, this study preprocesses data to address missing values, outliers, and feature scaling, ensuring robust model performance. The project not only aims to predict obesity risk but also to provide actionable recommendations for lifestyle modifications, bridging the gap between predictive analytics and practical interventions. By aligning with public health goals, this model has the potential to contribute to personalized health assessments and scalable community health programs, offering a meaningful step toward addressing the global obesity epidemic.

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 SVC 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.

III. Proposed Methodology

This project leverages machine learning techniques to develop a predictive model for assessing obesity risk based on demographic and lifestyle factors. The approach is structured into three primary phases: data preprocessing, model development, and evaluation. SVC and Naive Bayes are the core models used, chosen for their balance between simplicity, interpretability, and predictive power.

A. Data Descriptions

The dataset utilized in this study consists of nearly **20,000 records** with **17 features** encompassing demographic and lifestyle factors. These include variables such as age, gender, height, weight, smoking habits, alcohol consumption, physical activity levels, and dietary preferences. The target variable is a categorical classification of obesity risk, ranging from underweight to obesity levels I, II, and III. The dataset provides a comprehensive basis for analyzing the multifactorial nature of obesity risk.

1) Key Data Characteristics:

- **Source:** Publicly available health data repositories (e.g., UCI Machine Learning Repository).
- **Dimensions:** 20,000 rows and 17 columns, with features distributed across demographic and lifestyle categories.
- **Key Features:**
  - **Demographic:** Age, gender, socioeconomic status
  - **Lifestyle:** Frequency of physical activity, dietary habits (e.g., caloric intake, preferred foods), smoking, and alcohol consumption.
  - **Derived Metrics:** Body Mass Index (BMI), calculated using height and weight for more granular risk classification.

2) Preprocessing Steps: To ensure data quality and prepare it for machine learning, the following preprocessing steps were implemented:

- **Feature Engineering:**
  - Derived new features, such as BMI, to enhance predictive power.
  - Converted units (e.g., height to inches, weight to pounds) for consistency.
- **Categorical Encoding:** Transformed categorical features, such as smoking and dietary preferences, into numerical formats using one-hot encoding.
- **Handling Missing Data:** Missing values were imputed using mean or mode imputation for continuous and categorical variables, respectively.
- **Normalization and Scaling:**
  - Applied normalization to continuous features to ensure compatibility with machine learning algorithms.
  - Standardized variables to have a mean of zero and a standard deviation of one, optimizing model performance.
- **Outlier Detection and Removal:** Identified and removed outliers using interquartile range (IQR) methods to reduce noise and improve model robustness.

3) Data Imbalance: An imbalance was observed in the target classes, with certain obesity risk levels underrepresented. To address this:

- Used resampling techniques to balance the dataset.
- Incorporated evaluation metrics such as F1-score to account for imbalanced class distributions.

B. Experimental Setup

The experimental setup for this project was carefully designed to ensure the development of robust and interpretable machine learning models for obesity risk prediction. It consists of three key stages: programming environment and tools, data preparation, and model training and evaluation. The details of each stage are elaborated below.

1) Programming Environment and Tools: To implement the machine learning pipeline, Python was chosen as the primary programming language due to its versatility and extensive ecosystem of libraries for data analysis and machine learning. The following tools and libraries were utilized:

- **1. Development Environment:** Google Colab is used as the primary interface for writing, testing, and debugging code. It facilitated modular development, visualization, and documentation of the workflow.
- **Libraries for Data Analysis:**
  - **pandas:** Utilized for data manipulation, cleaning, and preprocessing.
  - **NumPy:** Used for numerical computations and handling multidimensional arrays.
- **Machine Learning Libraries: Scikit-learn** is employed for model implementation, training, and evaluation. It provided functionalities for SVC, Naive Bayes, and ensemble methods.
- **Visualization Tools: Matplotlib** and **Seaborn** are Used for visualizing data distributions, feature importance, and model performance metrics.
- **Other Utilities: ColumnTransformer** is enabled streamlined preprocessing of mixed data types (categorical and continuous).
- **2. Data Preparation:** Data preparation was a crucial step in ensuring that the dataset was ready for machine learning. The following preprocessing techniques were applied:
  - **Data Cleaning:**
    - Handled missing values using appropriate imputation methods to maintain data integrity.
    - Identified and removed outliers using statistical methods such as the interquartile range (IQR).
  - **Feature Engineering:**
    - Added a new feature, Body Mass Index (BMI), calculated from height and weight, to enhance the model's predictive ability.
    - Encoded categorical variables such as dietary preferences and smoking habits into numerical formats using one-hot encoding.
  - **Normalization and Scaling:**

    - Normalized continuous variables such as height, weight, and BMI to ensure consistent scaling across features.
    - Applied standard scaling to improve convergence during model training.
- **Model Training:** Two primary models, Support Vector Classifier (SVC) and Naive Bayes, were implemented to predict obesity risk. Additionally, ensemble methods were explored to enhance predictive accuracy. The following steps were carried out:
  - Baseline Models:
    - **Support Vector Classifier (SVC):** This model is used to handle high-dimensional data, with strong margins for classification. Also used for normalizing numerical data and one-hot encoded categorical features.
    - **Naive Bayes:** A probabilistic model particularly effective for handling categorical variables, such as dietary and physical activity patterns.
  - **Ensemble Methods:** Implemented an ensemble combining SVC and Naive Bayes to capture both linear relationships and categorical feature interactions. This approach enhanced predictive accuracy and robustness.
- **Evaluation Metrics:** The models were evaluated using comprehensive metrics to ensure reliability and address the challenges of imbalanced data:
  - **Accuracy:** Measured the proportion of correctly classified samples across all obesity classes.
  - **Precision:** Evaluated the model's ability to correctly identify high-risk individuals while minimizing false positives.
  - **Recall:** Assessed the model's capability to identify all relevant high-risk cases, minimizing false negatives.
  - **F1-Score:** Combined precision and recall into a single metric, particularly useful for imbalanced datasets.
  - **Cross-Validation:** Performed k-fold cross-validation to ensure the generalizability of the models across different subsets of data.
- **Computational Efficiency:** Given the relatively large dataset size (approximately 20,000 records), computational efficiency was a key consideration. The selected models, SVC and Naive Bayes, are lightweight and computationally efficient, enabling rapid training and evaluation. All experiments were conducted on a personal laptop with a standard configuration (Intel i7 processor, 16GB RAM), demonstrating the feasibility of deploying the models in resource-constrained environments.
- **Visualization and Interpretability:** Visualizations played a significant role in understanding the data and interpreting the model results:
  - **Data Exploration:** Histograms and boxplots were used to analyze feature distributions and identify outliers.
  - **Model Performance:** Confusion matrices and bar charts were employed to visualize classification results and metric comparisons.
  - **Feature Importance:** SVC coefficients and Naive Bayes probabilities were visualized to highlight the significance of predictors.

C. Model Evaluation and Results

This section presents the evaluation metrics and results obtained from the implemented machine learning models. The evaluation was designed to assess the predictive power, reliability, and interpretability of the models used for obesity risk classification.

- **Evaluation Metrics:** The following metrics were employed to provide a comprehensive assessment of the models:
  - **Accuracy:** Measures the proportion of correctly classified samples across all obesity risk classes. It provides a high-level overview of model performance.
  - **Precision:** Evaluates the ability of the model to identify true positives while minimizing false positives, crucial for identifying high-risk individuals.
  - **Recall:** Assesses the model's capability to capture true positives, minimizing false negatives, particularly important in identifying individuals who might require immediate intervention.
  - **Cross-Validation:** Applied k-fold cross-validation to ensure the model's generalizability and robustness across different subsets of the dataset.
- **Baseline Models:**
  - **Support Vector Classifier:**
    - Achieved an accuracy of **94.79%** on the test set.
    - Provided interpretable results by highlighting the contribution of each feature to obesity risk.
    - Features such as BMI, physical activity frequency, and dietary habits showed significant contributions to the predictions.
  - **Naive-Bayes:**
    - Achieved an accuracy of **89.59%** but was particularly effective in handling categorical features such as smoking and alcohol consumption habits.
    - Probabilistic outputs provided an intuitive understanding of class likelihoods, making it useful for further analysis.
- **Ensemble Methods:** To improve performance and robustness, an ensemble approach was employed, combining SVC and Naive Bayes models using soft voting:
  - Achieved the highest accuracy of **98%** across the test dataset.
  - Showed improved precision and recall, particularly for underrepresented obesity classes, demonstrating its ability to address data imbalance.

  - The ensemble also achieved the best weighted average when it came to F1-score of **0.98**, reflecting a well-balanced performance.
- **Challenges and Adaptations in Model Evaluation**
  - **Data Imbalance:**
    * Some obesity risk classes were underrepresented, leading to potential biases in model predictions.
    * Solutions:
      · Evaluation metrics such as F1-score and precision-recall curves were prioritized over accuracy.
      · Resampling techniques, such as oversampling of minority classes, were applied during preprocessing.
  - **Feature Importance:** While SVC provided clear coefficients, the interpretability of the ensemble model required additional analysis using permutation importance to quantify the contribution of each feature.
- Visualizations:
  - **Confusion Matrix:** Visualized for all models to highlight misclassification patterns. For the ensemble model, fewer false negatives were observed compared to individual models.
  - **Bar Graphs of Metrics:** Metrics such as accuracy, precision, recall, and F1-score were plotted for comparison across all models.
  - **Feature Importance:** SVC coefficients revealed that BMI, frequency of physical activity, and caloric intake were the strongest predictors of obesity risk.
- **Results Summary:** The results highlight the following:
  - **Support Vector Classifier** excelled in interpretability and offered actionable insights.
  - **Naive Bayes** effectively handled categorical data but slightly underperformed in accuracy compared to SVC.
  - The **ensemble method** provided the best overall performance, achieving high accuracy and balanced metrics, making it the most robust approach for this project.

D. Interpretibility and Insights

A critical goal of this project was to ensure that the predictive models not only deliver accurate obesity risk predictions but also provide actionable insights into the factors driving these predictions. Interpretability is a key component in health-related applications, where understanding the "why" behind a prediction is as important as the prediction itself. By prioritizing models that offer transparency, such as Support Vector Classifier and Naive Bayes, this study bridges the gap between machine learning outcomes and practical, real-world applications.

- **Feature Importance Analysis:**
  - **SVC coefficients:** SVC offers interpretability through its model coefficients, which indicate the relative contribution of each feature to the target variable.
  - **Key Insights:**
    * **BMI** emerged as the strongest predictor of obesity risk, underscoring its established role as a measure of body fat.
    * **Frequency of Physical Activity (FAF)** and **Caloric Intake (CALC)** were also significant predictors, highlighting the importance of an active lifestyle and balanced diet.
    * **Dietary Preferences (FAVC)** and **Water Consumption (CH2O)** showed moderate influence, indicating their role in determining metabolic health.
  - **Actionable Insight:** Individuals can reduce obesity risk by improving physical activity levels, moderating caloric intake, and ensuring adequate hydration.
- **Naive Bayes Probabilities:**
  - Naive Bayes provides probabilistic outputs that intuitively indicate the likelihood of obesity risk given specific feature values.
  - **Key Insights:**
    * Higher probabilities of obesity risk were associated with frequent consumption of high-calorie foods and low levels of physical activity.
    * Probabilities for lower obesity risk were tied to regular exercise and high water intake.
- **Ensemble Methods Interpretability:**
  - **Permutation Importance:**
    * The ensemble method, while effective in balancing accuracy and robustness, required additional techniques to ensure interpretability. Permutation importance was employed to quantify the contribution of each feature across combined models.
    * **Key Insights:**
      · BMI consistently ranked as the top predictor, followed by physical activity levels and dietary patterns, aligning with findings from individual models.
      · Lesser but still notable contributors included smoking habits and alcohol consumption.
  - **Class-Level Insights:**
    * The ensemble model provided refined predictions across all obesity risk classes, particularly improving recall for underrepresented classes like "Underweight" and "Obesity Level III."
    * **Actionable Insight:** The ensemble's ability to generalize better across classes makes it suitable for large-scale public health applications.
- **Visualization of Interpretability:**
  - **Feature importance as bar graphs:**
    * Coefficients and permutation importance were visualized to clearly indicate the relative impact of features.

    - ∗ These graphs provided stakeholders with intuitive representations of which factors influence obesity risk the most.
  - **Probability Distributions:** Naive Bayes probabilities for each class were plotted to demonstrate how lifestyle choices like diet and physical activity shift the likelihood of different obesity risk levels.
- **Practical Implications:** The interpretability of these models enables:
  - **Personalized Health Recommendations:** Individuals can identify which lifestyle changes (e.g., reducing caloric intake or increasing exercise) would have the most significant impact on their health.
  - **Targeted Public Health Strategies:** Policymakers can allocate resources effectively by focusing on factors like improving access to physical activity opportunities or promoting dietary education in high-risk communities.
  - **Scalability and Accessibility:** The clear and interpretable outputs make the models adaptable for use in mobile applications, community health centers, or public health campaigns.
- **Limitation and Future Enhancements:** While the models are interpretable, certain limitations remain:
  - **Interaction Effects:**
    - ∗ We implemented a linear kernel in SVC, which may oversimplify complex interactions between features.
    - ∗ Future work could incorporate advanced explainability methods like SHAP (SHapley Additive exPlanations) to capture these nuances.
  - **Bias in Data Representation:** Feature importance may vary across demographic groups. Extending the dataset to include more diverse populations can provide broader insights.

E. Feasibility and Fallback Plan

The feasibility of this project hinges on its well-defined scope, manageable computational requirements, and robust methodologies. Despite these strengths, potential challenges such as data quality issues, model performance, and resource limitations were considered, and appropriate fallback strategies were designed to ensure project success.

- **Feasibility Assessment:**
  - **Dataset Accessibility:**
    - ∗ The dataset, sourced from publicly available health repositories such as the UCI Machine Learning Repository, provides sufficient diversity and scale with approximately 20,000 records and 17 features.
    - ∗ Features relevant to obesity prediction, such as BMI, dietary habits, and physical activity levels, ensure a strong basis for model development.
  - **Model Selection:** Support Vector Classifier and Naive Bayes were chosen for their computational efficiency and ability to handle categorical and continuous data effectively. These models are lightweight, making them feasible for deployment in resource-constrained environments.
  - **Computational Resources:** The project was implemented on Google Colab, which offers cloud-based computation and GPU support. This setup ensures that even resource-intensive tasks, such as cross-validation and ensemble model training, can be performed efficiently.
- **Anticipated Challenges:**
  - **Data Quality Issues:**
    - ∗ Missing values, outliers, and imbalanced target classes could compromise model accuracy and reliability.
    - ∗ The variability in lifestyle factors (e.g., dietary habits) may introduce noise, affecting the interpretability of results.
  - **Model Performance:** Initial models might fail to meet desired accuracy or generalizability due to non-linear relationships in the data or class imbalances.
  - **Scalability and Deployment:** While the models are computationally efficient, scaling them for real-time applications or large-scale public health initiatives may pose challenges.
- **Fallback Strategies:**
  - **Alternative Modeling Approaches:**
    - ∗ If SVC and Naive Bayes fail to achieve satisfactory performance, ensemble methods such as Random Forest or Gradient Boosting will be explored to capture non-linear relationships and improve predictive power.
    - ∗ Hyperparameter tuning will be conducted to optimize model configurations, including adjusting regularization parameters for support vector classifiers or priors for Naive Bayes.
  - **Data Augmentation and Supplementation:**
    - ∗ To address data quality or variability issues, publicly available datasets with similar demographic and lifestyle features will be incorporated. This could enhance feature diversity and improve model robustness.
    - ∗ Synthetic data generation techniques such as SMOTE (Synthetic Minority Oversampling Technique) will be employed to address imbalances in the target classes.
  - **Evaluation Adjustments:** If accuracy alone does not adequately measure performance, other metrics such as recall, precision, and F1-score will be prioritized to handle class imbalances effectively.
  - **Computational Challenges:** For resource-intensive ensemble methods, computational loads will be managed by leveraging Colab's GPU support and using

efficient libraries like scikit-learn to streamline training and evaluation.

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 SVC 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

ACKNOWLEDGMENT

We would like to express our sincere gratitude to Professor Gu for her invaluable guidance and support for writing this proposal (and future project report). Her expertise and insights greatly contributed to the development and refinement of our methodology. We also wish to thank our teaching assistants, whose assistance has helped greatly with the professor writeup (and future project). Finally, we acknowledge the collaborative efforts of our team members - Ali, Srijita, Blossom, and Mohammad - whose dedication and teamwork made this project possible.