Modeling Health Risk Probabilities of Obesity and Hypertension Using Machine Learning

Author: Ramón Moreno Ríos
Advisor: Nelliud Torres Batista, DBA
Master of Engineering in Computer Engineering
Graduate Project EXPO, May 2026

Abstract

This study investigates the use of machine learning techniques to model health risk probabilities using data from the National Health and Nutrition Examination Survey 2017–2018. The objective is to evaluate the predictive performance of multiple models across health outcomes, including hypertension, stroke, kidney disease, and sleep disorders, with a focus on the relationship between obesity and hypertension. Variables such as body mass index, waist circumference, age, gender, and blood pressure measurements were used to develop predictive models. The results indicate that hypertension prediction achieved the highest performance, while stroke and kidney disease showed moderate accuracy. Sleep-related outcomes showed reduced predictive power due to limited feature representation. Additionally, the analysis confirmed a strong relationship between obesity and hypertension, as individuals with higher body mass index exhibited increased blood pressure levels.

Introduction

High blood pressure and obesity are increasingly prevalent chronic health issues that increase the risk of cardiovascular and metabolic complications [1]–[3]. The body mass index is used in the United States to categorize the severity of obesity, which is still very common in adults [1]. High blood pressure, also known as hypertension, is frequently referred to as a "silent" illness since many people may not have symptoms, but it is connected to significant consequences if it is not diagnosed and treated [2], [3].

Large-scale population health studies enable this type of modeling using standardized measurements collected across diverse participants. The National Health and Nutrition Examination Survey combines interviews with physical examinations and laboratory testing, providing measured body mass index, repeated blood pressure readings, and outcome-related variables.

Background

Obesity is a major global public health challenge, affecting millions across all age groups. It is defined as excessive body fat that poses health risks and is commonly measured using body mass index. The World Health Organization reports that obesity rates have increased significantly over recent decades, contributing to higher rates of chronic diseases such as cardiovascular disease, diabetes, and other metabolic conditions [4].

Hypertension, or high blood pressure, is another serious health condition that is closely related to obesity. It is considered a major risk factor for cardiovascular diseases, including heart attacks and stroke. Elevated blood pressure places additional strain on the cardiovascular system and can cause long-term health complications if it is not properly diagnosed and treated [2], [3].

The relationship between obesity and hypertension is important because excess body weight can influence several physiological factors connected to blood pressure regulation. Clinical studies show that individuals with higher body mass index are significantly more likely to develop hypertension, indicating a direct relationship between excess body weight and elevated blood pressure [5].

Machine learning provides a data-driven approach for analyzing complex health datasets. By evaluating multiple variables at the same time, machine learning models can estimate health risk probabilities rather than relying only on single-threshold screening methods. This makes it useful for studying obesity-related conditions such as hypertension, kidney risk, stroke, and sleep-related outcomes.

Problem

Obesity and hypertension are extensively researched health conditions; however, accurately assessing individual risk remains challenging. Conventional clinical methods often rely on fixed thresholds, such as body mass index categories and blood pressure cutoffs, but these methods may not fully capture the combined effects of age, body composition, blood pressure, and other health indicators. Predictive modeling across multiple health outcomes also presents challenges because not all conditions are represented equally in the available data. While hypertension is closely associated with measurable variables such as body mass index and blood pressure, outcomes such as kidney disease, stroke, and sleep disorders may depend on additional biological and behavioral factors that are not fully captured in the dataset.

Methodology

This study utilizes data from the National Health and Nutrition Examination Survey 2017–2018 dataset, which provides demographic, clinical, examination, and laboratory measurements for a representative sample of the United States population [6]. Multiple data files were merged using the participant identifier to create a unified dataset containing variables related to obesity, hypertension, and additional health outcomes.

The selected features included body mass index, waist circumference, age, gender, systolic blood pressure, diastolic blood pressure, cholesterol measurements, and glucose-related variables. Binary outcome variables were defined for hypertension, stroke, kidney risk, and sleep-related risk using clinical thresholds and questionnaire responses from the dataset.

Data preprocessing was completed using a structured machine learning pipeline. Missing numerical values were handled using median imputation, while categorical values were handled using most-frequent imputation. Numerical features were standardized to ensure consistent input ranges, and categorical variables such as gender were transformed using one-hot encoding.

Several machine learning models were implemented to evaluate predictive performance. Logistic Regression was used as the baseline model because it is interpretable and effective for binary classification tasks [7]. A Neural Network was also used to explore possible non-linear relationships among variables, and a Random Forest classifier was tested as an ensemble-based comparison model.

Model performance was evaluated using stratified cross-validation to maintain class balance across folds. The primary evaluation metric was the area under the receiver operating characteristic curve, while precision, recall, and F1-score were also examined to evaluate classification performance. Threshold tuning was applied when needed to improve the balance between precision and recall for imbalanced outcomes.

Results and Discussion

The performance of machine learning models varied across health outcomes, with hypertension exhibiting the highest predictive accuracy among all targets.

Figure 1 demonstrates that the logistic regression model achieved the strongest performance for hypertension prediction, with a ROC-AUC of approximately 0.80–0.82, indicating a strong ability to distinguish positive from negative cases. In contrast, models for stroke and kidney disease achieved moderate performance, with ROC-AUC values ranging from approximately 0.72 to 0.78. Sleep-related risk prediction yielded substantially lower performance, with ROC-AUC values near 0.58 to 0.60.

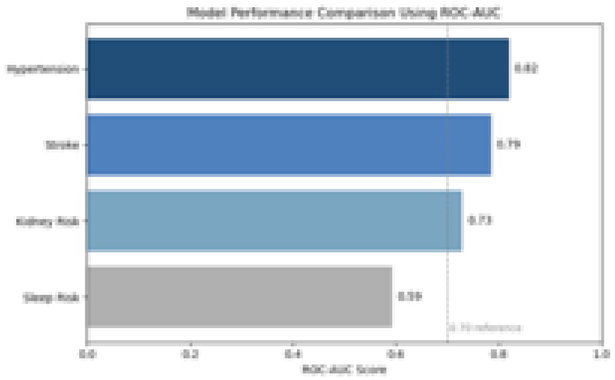

Figure 1
Model Performance Comparison Using ROC-AUC

The superior performance in hypertension prediction is attributable to the direct availability of highly relevant clinical features, particularly systolic and diastolic blood pressure measurements, as well as obesity-related indicators such as body mass index and waist circumference. Conversely, outcomes such as stroke, kidney disease, and sleep disorders are influenced by more complex and multifactorial processes that are not fully captured by the available dataset features.

The relationship between obesity and hypertension was analyzed using both graphical and statistical approaches. Figure 2 illustrates the relationship between body mass index and systolic blood pressure, showing a general upward trend as body mass index increases. The results demonstrate that individuals with higher body mass index values tend to exhibit increased systolic blood pressure, indicating a positive association between obesity and hypertension risk.

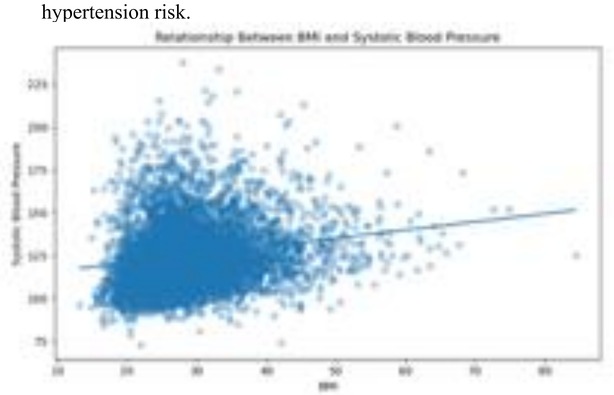

Figure 2
Relationship Between BMI and Systolic Blood Pressure

Conclusions

This study showed that machine learning techniques can effectively model health risk probabilities using National Health and Nutrition Examination Survey 2017–2018 data. Hypertension had the highest predictive performance, likely because it is linked to clear clinical features like blood pressure and obesity-related indicators. On the other hand, predictions for stroke, kidney disease, and sleep disorders were less accurate, which may be due to the complex and varied causes of these conditions.

The analysis also confirmed a strong link between obesity and hypertension. Both the graphs and statistics showed that people with higher body mass index and waist circumference are more likely to have high blood pressure and a greater risk of hypertension.

Future Work

Future research should aim to improve model performance by incorporating additional variables that capture behavioral, environmental, and longitudinal health factors currently absent from the dataset. Including lifestyle-related data such as diet, physical activity, sleep quality, and medical history may substantially enhance the predictive capability of models, especially for complex conditions such as stroke, kidney disease, and sleep disorders. In addition, exploring advanced machine learning techniques, such as ensemble methods and deep learning architectures, may facilitate the identification of non-linear relationships that traditional models do not fully capture.

Acknowledgements

I would like to thank Nelliud Torres Batista, for his guidance and support throughout this project. I also thank the Polytechnic University of Puerto Rico and the Electrical & Computer Engineering and Computer Science Department for providing the academic resources and structure needed to complete this research. Special thanks to the National Center for Health Statistics for providing access to the National Health and Nutrition Examination Survey dataset used in this study.

References

[1] Centers for Disease Control and Prevention. (2024, May 14). Adult Obesity Facts [Online]. Available: https://www.cdc.gov/obesity/adult-obesity-facts/index.html.

[2] World Health Organization. (2025, September 25). Hypertension [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/hypertension.

[3] National Center for Health Statistics. (2024, August 7). Hypertension—Health, United States: Sources and Definitions [Online]. Available: https://www.cdc.gov/nchs/hus/topics/hypertension.htm.

[4] World Health Organization. (2024). Obesity and Overweight [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/obesity-and-overweight.

[5] J. Hall et al., "Obesity-induced hypertension: interaction of neurohumoral and renal mechanisms," Circulation Research, vol. 116, no. 6, pp. 991–1006, 2015.

[6] National Center for Health Statistics, "National Health and Nutrition Examination Survey: Overview 2017–2018," Centers for Disease Control and Prevention, Hyattsville, MD, 2020.

[7] T. Hastie, R. Tibshirani, and J. Friedman, "Supervised learning," in The Elements of Statistical Learning, 2nd ed. New York, NY, USA: Springer, 2009, ch. 2, pp. 9–41.