PREDICTING HYPERTENSION RISK AMONG U.S. ADULTS:

A MACHINE LEARNING APPROACH USING ALCOHOL INTAKE, BMI, AND SOCIODEMOGRAPHIC FACTORS

By: Zheng Luo

Bachelor of Science in Applied Mathematics, University of Liverpool, 2023

Thesis Advisor: David Abramson, PhD

A Master’s thesis submitted in partial fulfillment of the requirements for the degree of Master of Science

Department of Biostatistics

School of Global Public Health, New York University

May 2025

**Abstract**

*Introduction*

Hypertension continues to affect nearly half of U.S. adults, with risk patterns shaped by lifestyle behaviors and structural inequalities. Although alcohol intake, body mass index (BMI), sleep quality, and income level have each been associated with blood pressure outcomes, few studies have employed contemporary machine learning techniques to assess their relative contributions to hypertension risk. This study applied machine learning to enhance predictive performance and identify the most influential factors associated with hypertension among U.S. adults.

*Methods*

Data were collected from the National Health and Nutrition Examination Survey (NHANES) from 2017 through March 2020, focusing on adults aged 18 years and older (n = 6,175). Predictor variables included frequency of alcohol consumption, BMI, self-reported sleep problems, income category, and demographic characteristics. Support vector machine (SVM), random forest, and XGBoost models were trained using 7:3 partitioning and five-fold cross-validation. SHAP values were used to interpret the relative importance of predictors.

*Results*

SVM showed the highest performance (AUC = 0.997), followed by Random Forest (0.730) and XGBoost (0.728). The SHAP interpretation showed that age, BMI and alcohol consumption were the most influential predictors. Sleep disorders and low income also slightly increased risk. The models consistently identified subgroups that predicted disproportionate risk of hypertension, such as older adults and heavy drinkers.

*Discussion*

Combining machine learning with interpretable methods such as SHAP yielded important characteristics, providing a practical approach to identifying at-risk populations and understanding the multifactorial triggers of hypertension. These findings support more targeted public health efforts that target behaviors and environments.

***Keywords****:* Alcohol intake, Hypertension, BMI, sleep disorder, family income

*Word count of abstract: 250*

*Word count of main text: 4,350*

*Number of tables: 2*

*Number of figures: 5*

**Introduction**

High blood pressure (HBP), also known as hypertension, represents the most significant global cardiovascular risk factor. As noted by Cutler et al., hypertension is responsible for roughly half of all coronary heart disease cases and nearly two-thirds of cerebrovascular disease cases worldwide.1 These figures highlight the substantial burden this condition imposes—not only on individuals but also on healthcare systems around the world. Its asymptomatic nature and widespread prevalence make hypertension especially difficult to manage on a population level. Despite substantial progress in public health interventions and management strategies, the prevalence of hypertension continues to rise, particularly among adults in the United States. According to the Centers for Disease Control and Prevention (CDC), nearly half of U.S. adults have hypertension or are taking medication to manage it, with significant racial and socioeconomic disparities in its incidence and outcomes.1,2 This persistent upward trend suggests that existing strategies may be insufficiently tailored to address the complex and multifactorial nature of hypertension risk.

A 2024 dose-response meta-analysis by Cecchini et al. confirmed a dose-dependent relationship between alcohol consumption and hypertension risk, finding that intake of ~48 g/day was associated with a 33% higher risk compared to a 12 g/day reference group.3 There is no clear safety threshold, especially in men, and even moderate alcohol consumption is associated with an elevated risk of hypertension.4 Other studies observed relative risks of ~1.6 at 50 g/day and ~2.5 at 100 g/day in men and even higher in women.² Taken together, these findings emphasize that alcohol’s effect on blood pressure is not limited to heavy drinking but extends even to commonly accepted “moderate” levels, challenging public perceptions of alcohol safety. Although early observational studies suggested a J-shaped curve with potential cardioprotective effects of light drinking, emerging Mendelian randomization analyses of Abdelhady et al. have refuted this, indicating that even low alcohol intake can increase blood pressure.5 This evolving understanding underscores the need for more rigorous causal inference methodologies in research examining the links between lifestyle factors and disease outcomes.

The relationship between alcohol intake and hypertension is significantly shaped by other modifiable factors, such as body mass index, sleep quality, and socioeconomic status. For example, sleep disorders—particularly chronic insomnia and obstructive sleep apnea (OSA)—have been linked to significantly higher hypertension risk.6,7,8,9 Sleep health is increasingly recognized as a key determinant of cardiovascular outcomes, yet it remains under-addressed in many clinical and public health settings. Kaveh Hosseini’s 2024 meta-analysis of over one million individuals found that sleeping less than 7 hours per night was associated with a significantly higher incidence of hypertension, especially among women.8 Additionally, short sleep duration and circadian disruption (e.g., from shift work) can amplify the effects of other risk factors like alcohol or stress.7 These findings underscore the importance of considering sleep not as an isolated factor but as a dynamic modifier of behavioral and environmental exposures. Sleep disturbances may exacerbate the risk of hypertension, particularly in the presence of high levels of alcohol consumption.

Body mass index (BMI) is a well-established and consistently strong predictor of hypertension, frequently emerging as one of the most influential risk factors in both conventional statistical analyses and machine learning models.9 Elevated BMI is associated with greater blood pressure burden across all age groups. Considering the global obesity epidemic, targeting BMI reduction remains a critical component of hypertension prevention efforts, particularly among young and middle-aged adults.

Socioeconomic status (SES), particularly family income, also plays a moderating role. Individuals with lower income tend to face higher levels of chronic stress, limited access to healthcare, and more barriers to healthy behaviors such as alcohol moderation or sleep hygiene.2,10,11,12 Socioeconomic status (SES) influences not only individual health behaviors but also exposure to structural and environmental risk factors, thereby perpetuating health disparities across generations. These social determinants contribute to a disproportionately high burden of uncontrolled hypertension among low-income populations. For instance, one analysis demonstrated that individuals living below the federal poverty line faced a significantly elevated risk of inadequate blood pressure control.12 Addressing hypertension effectively thus requires a multidimensional approach that incorporates both biological and social risk factors.

To capture the complex interactions among alcohol intake, sleep quality, BMI, and family income, recent studies have turned to machine learning (ML) for hypertension risk prediction.13 Machine learning algorithms including XGBoost, Random Forest, and Support Vector Machines offer the ability to model nonlinear relationships and interactions that are often missed by traditional regression approaches.13

These methods offer a robust framework for uncovering hidden patterns in large-scale health data, making them particularly well-suited for personalized risk assessment. In predictive modeling, machine learning (ML) techniques have consistently demonstrated superior performance compared with conventional statistical approaches, with area under the curve (AUC) values exceeding 0.90 in certain population-based datasets.13 Furthermore, interpretable ML methods, such as SHapley Additive exPlanations (SHAP), enable researchers to quantify the relative contribution of individual predictors to model outputs, thereby improving model transparency and clinical applicability.14 The ability to interpret model outputs is especially valuable in clinical decision-making, where explainability can support physician trust and patient communication. These techniques are particularly valuable for identifying high-risk subpopulations, such as younger individuals with elevated alcohol consumption and poor sleep quality, who may remain undetected by conventional modeling approaches.15

*Research Gap and Objectives*

While prior studies have examined the individual effects of alcohol intake, sleep quality, BMI, and family income on hypertension, few have investigated their combined impact using advanced ML methods. Despite its potential, the application of SHAP analysis to interpret complex risk associations remains limited in current literature. Given the multifactorial nature of hypertension, there is a pressing need for integrative modeling approaches that incorporate both behavioral and structural determinants in a manner that is both flexible and interpretable. This study addresses these gaps by applying machine learning algorithms to predict hypertension risk among U.S. adults using data from the National Health and Nutrition Examination Survey (NHANES).

The central research question guiding this study is: How accurately can machine learning models predict the risk of hypertension among U.S. adults using alcohol intake, BMI, sleep quality, and sociodemographic factors?

The specific objectives are to:

The objectives of this study are twofold: (1) to develop and evaluate multiple machine learning models for predicting hypertension risk based on alcohol intake, BMI, sleep quality, family income, and other sociodemographic factors; and (2) to assess the predictive performance of these models using accuracy, sensitivity, specificity, and area under the receiver operating characteristic curve (AUC) estimates.

This study develops machine learning models to predict hypertension risk based on lifestyle behaviors and sociodemographic factors, using data from a nationally representative sample. To enhance model interpretability, SHapley Additive exPlanations (SHAP) analysis is employed to evaluate the relative importance of key predictors and to explore potential moderating effects. By leveraging advanced machine learning techniques and interpretability methods, this study aims to generate novel insights into the complex interplay between behavioral and socioeconomic factors in hypertension risk, thereby informing targeted and data-driven public health interventions.

**Methods**

*Data Source*

The data for this study were obtained from the 2017 to March 2020 Pre-pandemic National Health and Nutrition Examination Survey (NHANES), a cross-sectional survey designed and conducted by the Centers for Disease Control and Prevention (CDC). Field operations for the NHANES program were discontinued in March 2020 because of the COVID-19 pandemic, resulting in incomplete data collection for the 2019 - 2020 cycle. Consequently, the data gathered from 2019 through March 2020 alone were insufficient to be nationally representative. To address this limitation, the analytical dataset was constructed by merging the study variables with data from the NHANES 2017–2018 cycle, resulting in a combined, nationally representative sample that captures the pre-pandemic period from 2017 to March 2020. NHANES collected health-related data through self-reported questionnaires, physical examinations, and laboratory tests, thereby integrating subjective and objective measures of individual health status. Employing a complex, multistage probability sampling design, NHANES ensured that the obtained data were nationally representative across diverse demographic subgroups, including age, sex, race/ethnicity, and geographic regions.

*Study Design*

During this approximately three-year period, 15,560 participants aged 12 years and older were enrolled in the NHANES survey. Demographic information was documented for all participants, and the majority completed detailed questionnaires covering sensitive personal information such as sleep disorders and alcohol consumption. Based on comprehensive individual-level data, this study applied machine learning methods—such as XGBoost and logistic regression—to estimate the risk of hypertension among U.S. adults, incorporating demographic, behavioral, and lifestyle-related predictors.

*Participants and Procedures*

In the sample of 15,560 participants, all participants completed an interview regarding the Family’s monthly poverty level. Of these, 8,965 participants responded to an interview on alcohol consumption, yielding a response rate of 57.6%, while 10,195 participated in a sleep quality interview (response rate = 65.5%). Furthermore, 14,300 subjects underwent physical measurements, which were collected by trained health technicians at the Mobile Examination Center (MEC) with the assistance of recorders. The physical measurement protocol, tailored based on the participants’ ages as determined during the screening interview, included assessments of height, weight, and waist circumference, among other metrics. Additionally, 11,656 subjects underwent a standardized blood pressure assessment; following a five-minute seated rest, three consecutive blood pressure readings were obtained using a digital upper arm monitor (Omron HEM–907XL) at 60-second intervals, providing the data used for blood pressure analysis. Within the data, a proportion of missing values was observed. Specifically, codes 7 or 77 and 9 or 99 were used to indicate responses of "refused" and "do not know," respectively, while missing values were denoted by a period ("."). The NHANES was reviewed and approved by NCHS IRB. The proposed secondary data analysis was exempt from the New York University Institutional Review Board. Data were collected via face-to-face interviews with adult household members. Medical assessments, laboratory tests, and physiological measurements were systematically conducted by trained healthcare professionals as part of the data collection process.

*Measures*

To minimize the potential biases associated with imputation techniques such as nearest neighbor or multiple imputation—particularly in situations where missingness may not be random—the study employed listwise deletion, excluding participants with missing data on key analytic variables. Consequently, the final analytic sample included 6,175 participants. Although listwise deletion assumes data are missing at random (MAR), and some missingness in the NHANES dataset arises from planned skip patterns, a bias analysis revealed significant differences between included and excluded participants in gender, race/ethnicity, BMI, family income, and sleep problems (all p < 0.001), with no significant difference observed in age. These findings suggest that selection bias may be present and should be considered when interpreting the generalizability of the results. Nonetheless, given the risk of introducing additional biases through imputation in the presence of potential nonrandom missingness, listwise deletion was deemed an appropriate and conservative approach for the primary analysis.

Researchers assessed each participant’s blood pressure by obtaining three repeated measurements of both systolic and diastolic values. They then calculated the final blood pressure as the average of the three systolic and diastolic readings to represent the participant’s blood pressure status at the measurement time. Systolic blood pressure values ranged from 52 mmHg to 225 mmHg, while diastolic blood pressure values ranged from 28 mmHg to 151 mmHg. Missing values were indicated by a period (".").

According to the World Health Organization (WHO), hypertension is diagnosed when systolic blood pressure is ≥140 mmHg and/or diastolic blood pressure is ≥90 mmHg, measured on at least two days. Based on this definition and given the NHANES protocol of obtaining multiple measurements during a single visit, the researchers operationalized hypertension using the average of the three readings. A new binary variable was created to indicate hypertension status: participants were classified as hypertensive if their mean systolic blood pressure was ≥140 mmHg and/or mean diastolic blood pressure was ≥90 mmHg. This binary outcome variable (hypertension) was generated in R using a logical condition across averaged systolic and diastolic values.

Alcohol use was assessed via interviewer-administered questionnaires. Participants were asked: “During the past 12 months, about how often did you have 4/5 (4 for women, 5 for men) or more drinks of any alcoholic beverage?” Responses were coded into the following categories: 0 = Never in the last year, 1 = Every day, 2 = Nearly every day, 3 = 3 to 4 times a week, 4 = 2 times a week, 5 = Once a week, 6 = 2 to 3 times a month, 7 = Once a month, 8 = 7 to 11 times in the last year, 9 = 3 to 6 times in the last year, 10 = 1 to 2 times in the last year, 77 = Refused, 99 = Do not know, . = Missing.

For analysis, these responses were recoded into a three-level categorical variable representing alcohol use frequency: “Heavy drinker” (codes 1–3), “Moderate drinker” (codes 4–6), and “Light drinker” (codes 7–10). Responses coded as 0 (never in the last year), 77 (refused), 99 (do not know), and missing were excluded from the analysis. This recoding was implemented using case\_when() in R, which ensured that each group had clear behavioral thresholds. Frequencies outside the drinking categories were filtered to avoid misclassification bias.

Body Mass Index (BMI) was computed by NHANES using measured height and weight, following the formula: weight (kg) divided by height squared (m²). The resulting values ranged from 11.9 to 92.3 kg/m². BMI was treated as a continuous numeric variable in the analysis. Values with NA were dropped prior to analysis, which was consistent with the listwise deletion strategy.

Family economic status was assessed using the INDFMMPC variable from the NHANES dataset. This variable reflects the ratio of a family's monthly income to the federal poverty guidelines, where the poverty threshold is specifically adjusted for household size, survey year, and state of residence. The federal poverty guidelines increase with the number of individuals in the household to reflect the greater financial needs of larger families. Participants were categorized based on this ratio into three groups: (1) ≤130% of the poverty line, (2) 130–185% of the poverty line, and (3) >185% of the poverty line. Values corresponding to refusal, unknown, or missing responses (7, 9, or .) were excluded. For analytic purposes, the categories were relabeled as Low, Medium, and High income. This classification is consistent with socioeconomic standards commonly used in public health research.

Sleep quality was assessed via the question: “Have you ever been told by a doctor that you have trouble sleeping?” Responses were coded as: 1 = Yes, 2 = No, 7 = Refused, 9 = Do not know, . = Missing. This variable was treated as a binary categorical (yes/no), excluding non-informative responses. Responses were recoded into a factor variable in R to enable direct modeling without dummy transformation.

Age was treated as a numeric variable. Participants younger than 18 years or older than 99 years were excluded. A new categorical age group variable was created as follows: 18–39 years, 40–59 years, and 60 years and above. The age grouping was implemented using mutate() and case\_when(), enabling downstream stratified analysis.

Other demographic variables, including sex and race/ethnicity, were retained as coded in the original NHANES dataset without further recoding or modification.

*Statistical analysis*

Descriptive statistics were computed to summarize the key characteristics of the study population. Descriptive statistics were computed without applying NHANES sample weights. Analyses summarized characteristics within the analytic sample and were not intended to produce nationally representative estimates. Table 1 presents an overview of all relevant covariates in the full analytic sample. Categorical variables, including age group, gender, race/ethnicity, family income, sleep quality, alcohol consumption category, and hypertension status, were summarized as frequencies and percentages. Continuous variables, such as age and body mass index (BMI), were summarized using means and standard deviations. All categorical variables were treated as factors in the analysis.

To examine bivariable associations between covariates and hypertension, stratified descriptive statistics were generated using the tbl\_summary() function from the gtsummary package. Table 2 presents means and standard deviations for continuous variables, and frequencies with percentages for categorical variables stratified by hypertension status.

Statistical comparisons between groups were performed using independent samples t-tests for continuous variables (age and BMI) and chi-square tests for categorical variables (e.g., age group, gender, race/ethnicity, income, sleep quality, and drinking category). Missing data were excluded from the analysis of each variable.

The dataset was initially partitioned into a training set and a testing set in a 7:3 ratio. The training set was used to develop the models, while the testing set served as an independent subset for evaluating model performance. Five-fold cross-validation (k = 5) was applied within the training set to improve model generalizability further and reduce the risk of overfitting. Three supervised machine learning algorithms, namely Support Vector Machine (SVM), random forest, and XGBoost, were implemented. Model performance was evaluated based on classification accuracy and the area under the receiver operating characteristic curve (AUC). Sensitivity and specificity were additionally examined to provide a more complete assessment of model performance, with sensitivity reflecting the ability to correctly identify true hypertension cases and specificity reflecting the ability to correctly identify non-hypertensive individuals. The model demonstrating the highest predictive performance was selected for identifying potential risk factors for hypertension, including gender, age, race/ethnicity, current alcohol consumption status, and body mass index (BMI), which have been consistently reported in prior literature.

*Software*

All statistical analyses were performed in R version 4.4.3. Data cleaning and preprocessing were conducted using the dplyr, forcats, and ggplot2 packages. Missing data was handled through case-wise deletion, and categorical recoding was implemented using logical filtering and mapping. The dataset was constructed from multiple NHANES files (2017–2020), including demographic, alcohol, blood pressure, BMI, income, and sleep quality variables, and merged using inner\_join().

Machine learning models—including eXtreme Gradient Boosting (XGBoost), Random Forest, and Support Vector Machine (SVM)—were developed using the caret, xgboost, randomForest, and e1071 packages. Hyperparameter tuning was performed using 5-fold cross-validation with trainControl() and model evaluation was based on ROC and AUC metrics.

To address the class imbalance, weighted loss functions and SMOTE (Synthetic Minority Over-sampling Technique) from the smotefamily package were applied. One-hot encoding was executed using dummyVars() from the caret package.

The model interpretation was enhanced through SHAP (SHapley Additive Explanations) using the fastshap, iml, and SHAPforxgboost packages, allowing for a post hoc explanation of variable importance. Permutation-based importance analysis was conducted using varImp() and visualized with ggplot2.

All analyses were conducted locally in a secure academic environment. All machine learning models, corresponding code, and datasets, are publicly available on GitHub: https://github.com/ZhengLuo25/Hypertension\_ML\_Models.

**Results**

*Descriptive Statistics*

A total of 6,175 participants were included in the analysis after excluding missing values and outliers. Table 1 presents the univariable analysis of key characteristics, including age, gender, race, BMI, family income, sleep quality, and alcohol intake. The average age of participants was 49.54 years (SD = 17.93), with 51.1% identified as male and 48.9% as female. Most participants were non-Hispanic White (39.0%), followed by non-Hispanic Black (25.7%) and Mexican American (11.2%). The mean BMI was 30.07 (SD = 7.51), with 30.8% classified as having low family income and 30.3% reporting sleep problems. Regarding alcohol intake, 35.4% were light drinkers, 29.5% moderate drinkers, 14.0% heavy drinkers, and 21.0% non-drinkers.

*Bivariable Associations and Hypertension Status*

Table 2 presents the bivariable associations between hypertension status and key sociodemographic characteristics and covariates. The overall prevalence of hypertension was 21.2%. Participants aged 60 years and older had a significantly higher prevalence of hypertension (58% vs. 11% in the 18-39 age group, p < 0.001). Male participants had a higher prevalence of hypertension compared to females (55% vs. 45%, p = 0.005). Non-Hispanic Black participants showed a higher prevalence of hypertension (37%) compared to non-Hispanic White and Mexican American participants (p < 0.001). BMI was significantly higher in participants with hypertension (31.0 vs. 30.0, p < 0.001), and lower family income was also significantly associated with hypertension status (p = 0.027).

*Model Performance and Evaluation*

Three machine learning models were implemented to predict hypertension: XGBoost, Random Forest, and Support Vector Machine (SVM). The model development process involved iterative testing and optimization. Simpler models, such as logistic regression and Lasso, were initially explored to establish baseline performance. However, because the initial models were unable to capture the complex nonlinear relationships among predictors, tree-based and kernel-based methods were subsequently employed to provide greater modeling flexibility. Researcher found XGBoost particularly promising during early tuning phases but struggled with overfitting when using default parameters. Fine-tuning regularization terms and adjusting class weights became critical for balancing sensitivity and specificity.

As the analysis progressed, it was observed that model performance varied not only across different algorithms but also depending on how categorical variables were encoded. Consequently, one-hot encoding was replaced with label encoding in some experiments, which improved the stability of the XGBoost model. The analytic process involved iterative refinement, including repeated adjustments to model diagnostics, visualizations, and hyperparameter tuning.

XGBoost Model: The XGBoost model achieved an AUC of 0.728 on the test set, with a sensitivity of 0.91 and a specificity of 0.42. Variable importance analysis indicated that age and BMI were the strongest predictors, followed by race, gender, and alcohol intake.

Random Forest Model: The Random Forest model, optimized using 5-fold cross-validation, achieved an AUC of 0.730. Like XGBoost, age, BMI, and race emerged as key predictors of hypertension.

SVM Model: The SVM with radial basis function (RBF) kernel demonstrated the highest performance, with an AUC of 0.997, sensitivity of 0.96, and specificity of 0.98. Results revealing the importance results of permutation revealed that BMI, age, and alcohol intake contributed most to the model's predictive performance.

*Feature Importance and SHAP Analysis*

SHAP (SHapley Additive exPlanations) values were computed to interpret the XGBoost and SVM models. The SHAP analysis revealed that age, BMI, and alcohol intake categories exerted the most significant influence on hypertension predictions. Specifically, moderate and heavy drinkers showed a higher probability of developing hypertension, while light drinkers and non-drinkers had lower predicted risk. Sleep problems and low family income also moderately contributed to increased hypertension risk.

In addition, a permutation-based variable importance analysis provided further insight into which predictors most strongly influenced model performance (Figure 1). Systolic and diastolic blood pressure were the most influential features, followed by age and age group categories. Alcohol-related variables (e.g., heavy drinker, non-drinker, light drinker) and sleep problems showed relatively lower importance. These findings align with clinical understanding and reinforce that blood pressure measurements and age remain the dominant predictors of hypertension. At the same time, lifestyle factors such as drinking and sleep contribute more modestly to model predictions.

SHAP dependence plots further revealed nonlinear marginal effects for key predictors (Figure 2 - 5). For systolic blood pressure (SBP), the SHAP values followed an S-shaped pattern: relatively flat below 125 mmHg, a steep increase above 130 mmHg, and a plateau around 160 mmHg (Figure 2), suggesting a rapid escalation in predicted risk across the threshold range. Diastolic blood pressure (DBP) also showed a nonlinear contribution, with SHAP values rising sharply above 90 mmHg (Figure 3). For BMI, the SHAP values increased steadily until around 40 kg/m² and then flattened or slightly declined (Figure 4), indicating diminishing marginal risk at extreme obesity levels. Age exhibited a monotonic relationship, with SHAP values gradually increasing with age (Figure 5), reflecting the consistent role of older age in predicting hypertension. These patterns are consistent with known clinical thresholds and demonstrate the model’s ability to learn meaningful nonlinear relationships between predictors and hypertension risk. These patterns are consistent with known clinical thresholds and demonstrate the model’s ability to learn meaningful nonlinear relationships between predictors and hypertension risk.

**Discussion**

*Summary of Key Findings*

This study investigated the relationship between alcohol intake and the risk of hypertension among U.S. adults using NHANES data from 2017–2020. The findings suggest that age, BMI, and alcohol intake categories were the most influential predictors of hypertension. The XGBoost and SVM models demonstrated high predictive performance, with AUCs of 0.728 and 0.997, respectively. SHAP analysis further highlighted the contribution of drinking patterns and demographic factors to the likelihood of developing hypertension.

*Interpretation of Model Results*

The support vector machine (SVM) model outperformed other models, achieving a nearly perfect AUC of 0.997, indicating that it effectively captured complex nonlinear relationships within the data. However, the XGBoost and Random Forest models also demonstrated strong predictive performance, emphasizing the overall robustness of the predictive framework.

Notably, the superior performance of the SVM model was initially unexpected. Tree-based models such as XGBoost and Random Forest were anticipated to perform better due to their widespread application in health prediction research. Following the application of the Synthetic Minority Over-sampling Technique (SMOTE) to balance the training data and subsequent kernel parameter tuning, the SVM model rapidly surpassed the other approaches. This finding highlights the importance of empirical model evaluation rather than relying on prevailing trends when selecting modeling strategies.

Feature importance analysis across all models consistently identified age, body mass index (BMI), and alcohol intake as the top contributors to hypertension risk prediction. To further interpret model behavior, SHAP (SHapley Additive exPlanations) analyses were conducted. Visualization of individual SHAP values revealed how risk estimates shifted depending on combinations of age, BMI, and alcohol use patterns, providing clinically relevant insights beyond traditional performance metrics. These findings reinforce the notion that predictive accuracy must be accompanied by interpretability to inform actionable public health strategies.

*Alignment with Previous Literature*

The findings align with existing literature, which consistently identifies age and BMI as significant risk factors for hypertension. Studies have also highlighted the complex role of alcohol intake in modulating blood pressure. While moderate alcohol consumption is sometimes associated with lower cardiovascular risk, heavy alcohol consumption has been shown to elevate blood pressure and increase the likelihood of hypertension. This study supports these findings by demonstrating that heavy drinkers had a higher risk of hypertension compared to light drinkers and non-drinkers.

*Possible Mechanisms and Implications*

Several biological mechanisms may explain the observed associations. Chronic alcohol consumption has been shown to impair endothelial function, promote inflammation, and increase oxidative stress, all of which contribute to elevated blood pressure. Moreover, individuals with higher BMI and older age often exhibit increased arterial stiffness and vascular remodeling, further predisposing them to hypertension.

*Strengths and Limitations*

This study leveraged a large, nationally representative dataset (NHANES) with rigorous data collection protocols, enhancing the generalizability of the findings. Using machine learning models allowed for robust predictions and comprehensive evaluation of variable importance. However, it is important to acknowledge several limitations:

Residual Confounding: Despite adjusting for key demographic and lifestyle factors, unmeasured confounders may have influenced the results.

Cross-Sectional Design: The study's cross-sectional design precludes causal inference, limiting the ability to establish temporality between alcohol intake and hypertension.

Class Imbalance: Despite using SMOTE and class balancing techniques, class imbalance in the outcome variable may have influenced model performance.

*Significance and Future Directions*

The findings underscore the importance of considering alcohol intake patterns, BMI, and age when assessing hypertension risk. Future research should explore longitudinal data to establish causal relationships and investigate potential gene-environment interactions that may modify the observed associations. Additionally, intervention studies targeting modifiable risk factors such as alcohol consumption and BMI could provide valuable insights for hypertension prevention and management.

**References**

1. Cutler JA, Sorlie PD, Wolz M, Thom T, Fields LE, Roccella EJ. Trends in hypertension prevalence, awareness, treatment, and control rates in United States adults between 1988–1994 and 1999–2004. Hypertension. 2008;52(5):818-827. doi:10.1161/HYPERTENSIONAHA.108.113357

2. Leng B, Jin Y, Li G, Chen L, Jin N. Socioeconomic status and hypertension: a meta-analysis. *J Hypertens*. 2015;33(2):221-229. doi:10.1097/HJH.0000000000000428

3. Cecchini M, Filippini T, Whelton PK, et al. Alcohol Intake and Risk of Hypertension: A Systematic Review and Dose-Response Meta-Analysis of Nonexperimental Cohort Studies. Hypertension. 2024;81(8):1701-1715. doi:10.1161/HYPERTENSIONAHA.124.22703

4. Vacca A, Bulfone L, Cicco S, et al. Alcohol Intake and Arterial Hypertension: Retelling of a Multifaceted Story. Nutrients. 2023;15(4):958. Published 2023 Feb 15. doi:10.3390/nu15040958

5. Algharably EA, Meinert F, Januszewicz A, Kreutz R. Understanding the impact of alcohol on blood pressure and hypertension: From moderate to excessive drinking. Kardiol Pol. 2024;82(1):10-18. doi:10.33963/v.kp.98704

6. Jarrin DC, Alvaro PK, Bouchard MA, Jarrin SD, Drake CL, Morin CM. Insomnia and hypertension: A systematic review. Sleep Med Rev. 2018;41:3-38. doi:10.1016/j.smrv.2018.02.003

7. Belloir J, Makarem N, Shechter A. Sleep and Circadian Disturbance in Cardiovascular Risk. Curr Cardiol Rep. 2022;24(12):2097-2107. doi:10.1007/s11886-022-01816-z

8. Hosseini K, Soleimani H, Tavakoli K, et al. Association between sleep duration and hypertension incidence: Systematic review and meta-analysis of cohort studies. PLoS One. 2024;19(7):e0307120. Published 2024 Jul 15. doi:10.1371/journal.pone.0307120

9. Chang JL, Goldberg AN, Alt JA, et al. International Consensus Statement on Obstructive Sleep Apnea. Int Forum Allergy Rhinol. 2023;13(7):1061-1482. doi:10.1002/alr.23079

10. Chaturvedi A, Zhu A, Gadela NV, Prabhakaran D, Jafar TH. Social Determinants of Health and Disparities in Hypertension and Cardiovascular Diseases. Hypertension. 2024;81(3):387-399. doi:10.1161/HYPERTENSIONAHA.123.21354

11. Metlock FE, Hinneh T, Benjarsiran C, Alharthi A, Ogungbe O, Turkson-Ocran R-AN, Himmelfarb CR, Commodore-Mensah Y. Impact of social determinants of health on hypertension outcomes: A systematic review. Hypertension. 2024;81(8). doi:10.1161/HYPERTENSIONAHA.123.22571.

12. Anstey DE, Christian J, Shimbo D. Income inequality and hypertension control. J Am Heart Assoc. 2019;8(15):e013636. doi:10.1161/JAHA.119.013636.

13. Estiko R, Ishak R, Rijanto E, Juwana YB, Juzar DA, Widyantoro B. Hypertension prediction models using machine learning with easy-to-collect risk factors: a systematic review. J Hypertens. 2024;42(suppl 2):e19. doi:10.1097/01.hjh.0001027072.19895.81.

14. Cai A, Zhu Y, Clarkson SA, Feng Y. The Use of Machine Learning for the Care of Hypertension and Heart Failure. JACC Asia. 2021;1(2):162-172. Published 2021 Sep 21. doi:10.1016/j.jacasi.2021.07.005

15. Islam SMS, Talukder A, Awal MA, et al. Machine Learning Approaches for Predicting Hypertension and Its Associated Factors Using Population-Level Data From Three South Asian Countries. Front Cardiovasc Med. 2022;9:839379. Published 2022 Mar 31. doi:10.3389/fcvm.2022.839379

**Table and Figures**

Table 1. Demographic Characteristics analysis of alcohol intake and hypertension, key variables among U.S. adults, National Health and Nutrition Examination Survey (NHANES), United States, 2017-2020, n = 6175

|  |  |
| --- | --- |
|  | N (%)[[1]](#footnote-1) |
| age (mean (SD\*) IQR\*\*) | 49.54 (17.93) 30 |
| age (%)  18-39  40-59  60+ | 2,033 (32.9)  1,986 (32.2)  2,156 (34.9) |
| Gender (%) |  |
| Male | 3,154 (51.1) |
| Female | 3,021 (48.9) |
| race (%)  Mexican American | 692 (11.2) |
| Other Hispanic | 607 (9.8) |
| Non-Hispanic White | 2,408 (39.0) |
| Non-Hispanic Black | 1,590 (25.7) |
| Non-Hispanic Asian  Other Race  BMI (mean (SD) IQR)  Family monthly poverty level category (%)  Low  Medium  High    Ever told doctor had trouble sleeping (%)  Yes  No  Main Feature (Drinking status)  drinking category (%)  Heavy drinker  Light drinker  Moderate drinker  None-drinker  Main Outcome  hypertension (%)  No  Yes | 5,53 (9.0)  3,25 (5.3)  30.07 (7.51) 9.1  1,899 (30.7)  9,55 (15.5)  3,321 (53.8)  1,872 (30.3)  4,303 (69.7)  865 (14.0)  2,189 (35.4)  1,822 (29.5)  1,299 (21.1)  4,867 (78.8)  1,308 (21.2) |

SD\* = standard deviation

IQR\*\* = interquartile range

Table2. Bivariable associations between alcohol intake and hypertension, and key demographic characteristics and other features among U.S. adults, National Health and Nutrition Examination Survey (NHANES), United States, 2017–2020, n=6175

| **Key Characteristics** | **No Hypertension (n = 4,867)** | **Yes Hypertension (n = 1,308)** | **p-value** |
| --- | --- | --- | --- |
| **Age in years (Mean ± SD)** | 47 (18) | 60 (15) | <0.001 |
| **Age group** |  |  | <0.001 |
| 18–39 | 1,895 (39%) | 138 (11%) |  |
| 40–59 | 1,577 (32%) | 409 (31%) |  |
| 60+ | 1,395 (29%) | 761 (58%) |  |
| **Gender** |  |  | 0.005 |
| Male | 2,440 (50%) | 714 (55%) |  |
| Female | 2,427 (50%) | 594 (45%) |  |
| **Race** |  |  | <0.001 |
| Mexican American | 580 (12%) | 112 (8.4%) |  |
| Other Hispanic | 496 (10%) | 111 (8.3%) |  |
| Non-Hispanic White | 1,941 (40%) | 467 (36%) |  |
| Non-Hispanic Black | 1,107 (23%) | 483 (37%) |  |
| Non-Hispanic Asian | 471 (9.5%) | 82 (6.2%) |  |
| Other Race | 272 (5.5%) | 53 (4.1%) |  |
| **BMI (Mean ± SD)** | 30 (8) | 31 (7) | <0.001 |
| **Family monthly poverty level** |  |  | 0.027 |
| Low | 1,496 (31%) | 403 (31%) |  |
| Medium | 723 (15%) | 232 (18%) |  |
| High | 2,648 (54%) | 673 (51%) |  |
| **Ever told had sleep problems** |  |  | 0.025 |
| Yes | 1,442 (29.6%) | 430 (32.9%) |  |
| No | 3,425 (70.4%) | 878 (67.1%) |  |
| **Drinking category** |  |  | <0.001 |
| Heavy drinker | 633 (13%) | 232 (18%) |  |
| Light drinker | 1,815 (37%) | 374 (29%) |  |
| Moderate drinker | 1,492 (31%) | 330 (25%) |  |
| None-drinker | 927 (19%) | 372 (28%) |  |

**Note**: Values are presented as n (%) unless otherwise indicated.  
**Statistical test**: Pearson’s Chi-squared test is used for categorical variables, and the t-test is used for continuous.

Figure 1. Permutation-based variable importance plot for hypertension prediction using XGBoost model.

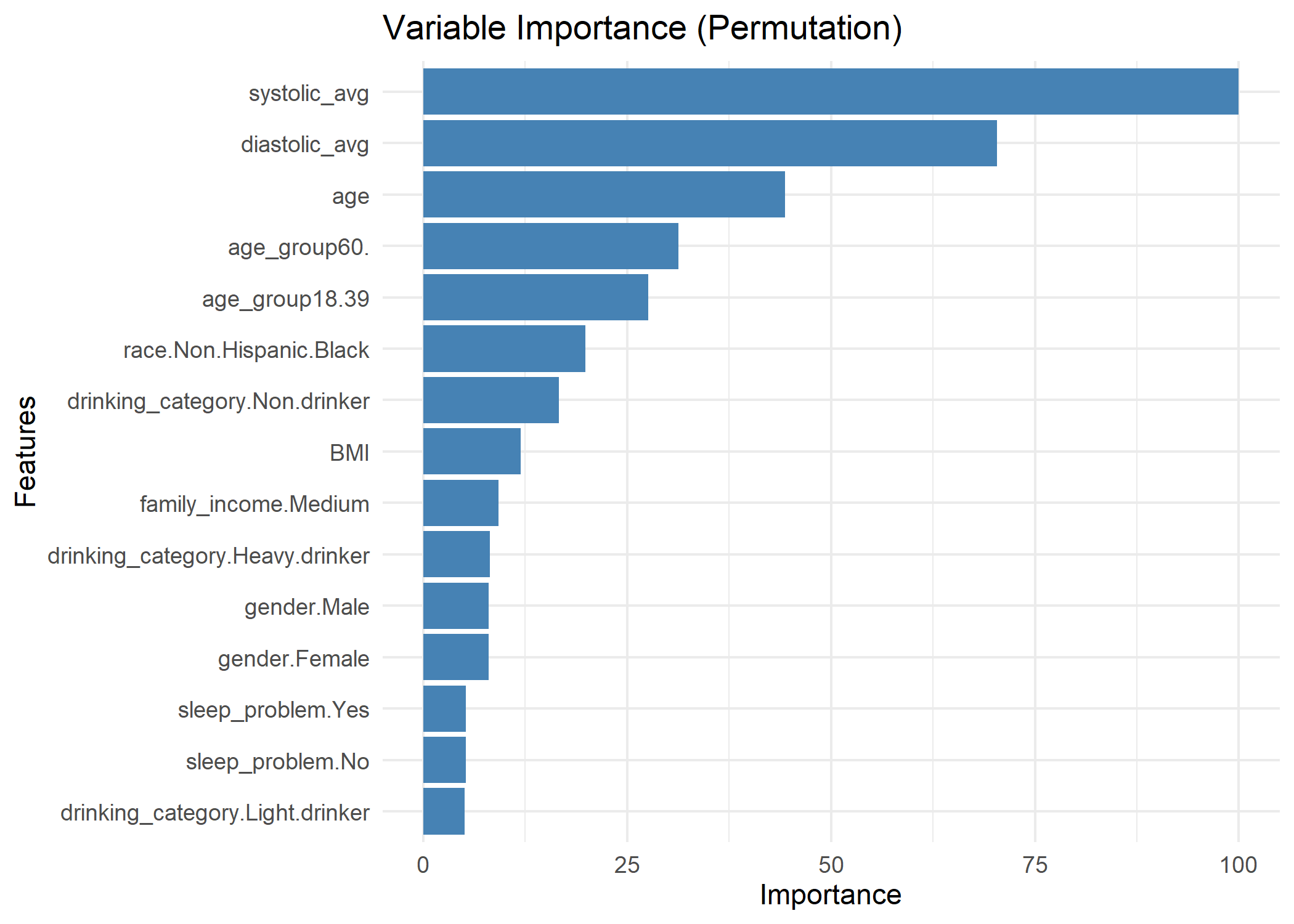


Figure 2. SHAP dependence plot for systolic blood pressure (SBP)

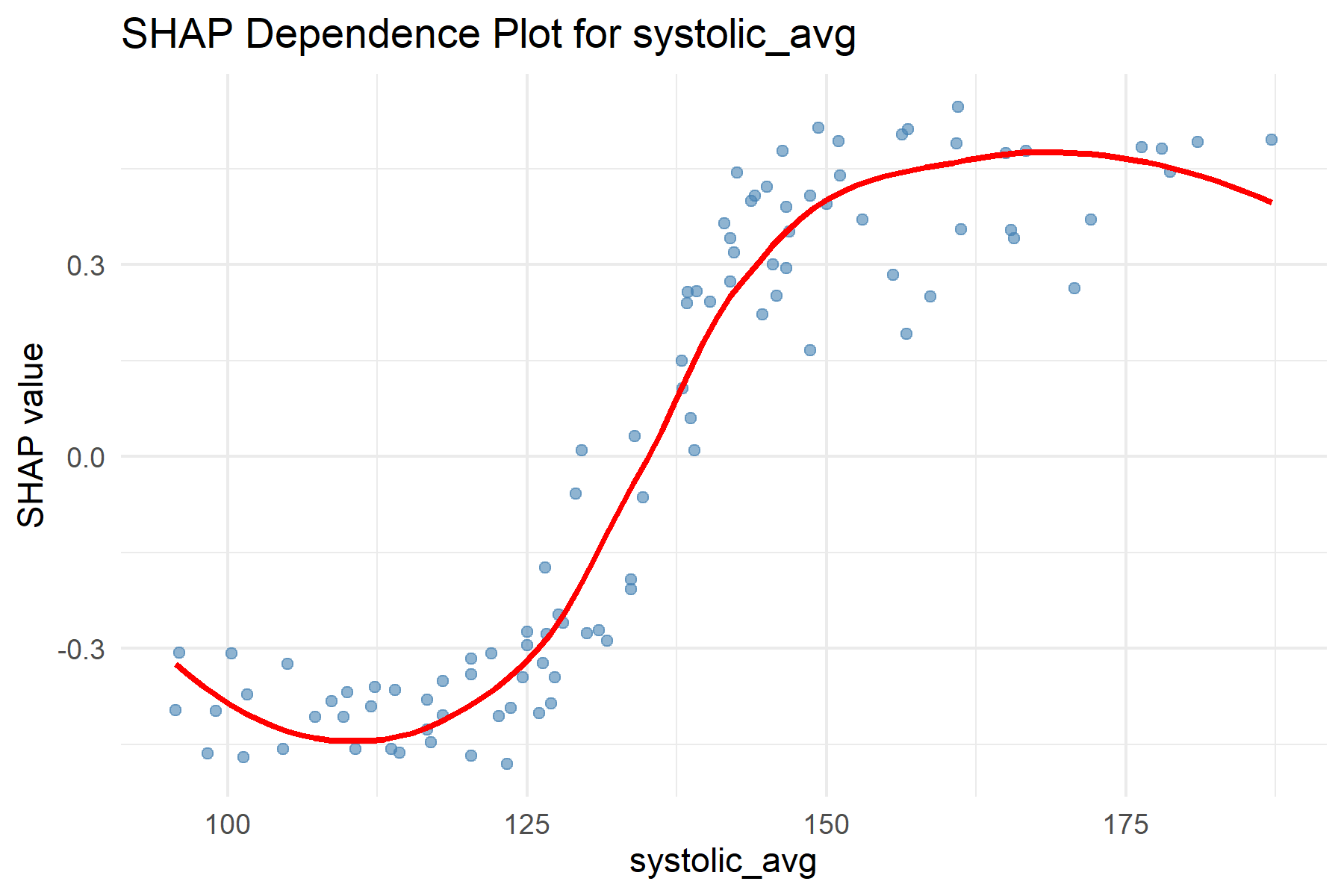


Figure 3. SHAP dependence plot for diastolic blood pressure (DBP)

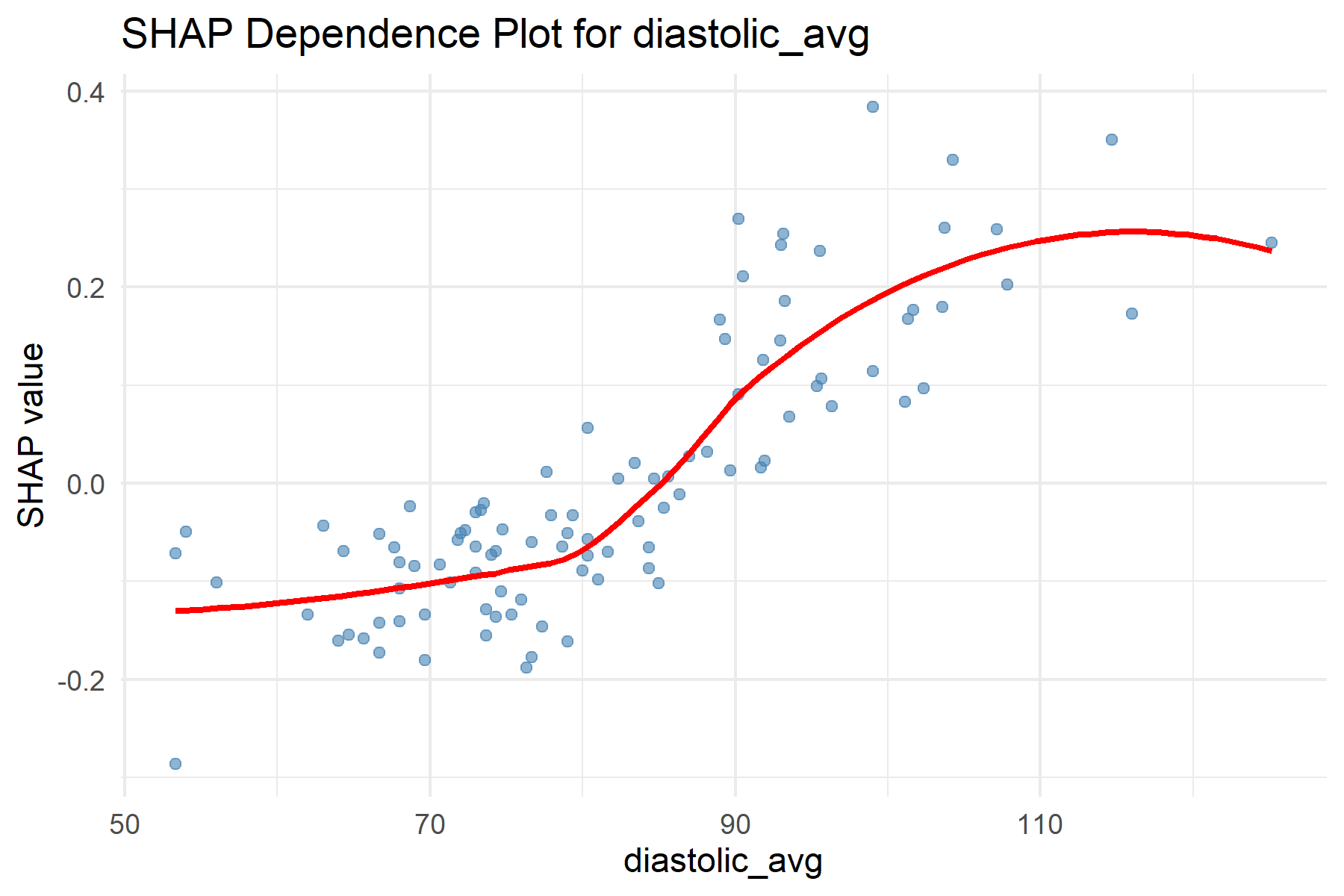


Figure 4. SHAP dependence plot for body mass index (BMI)

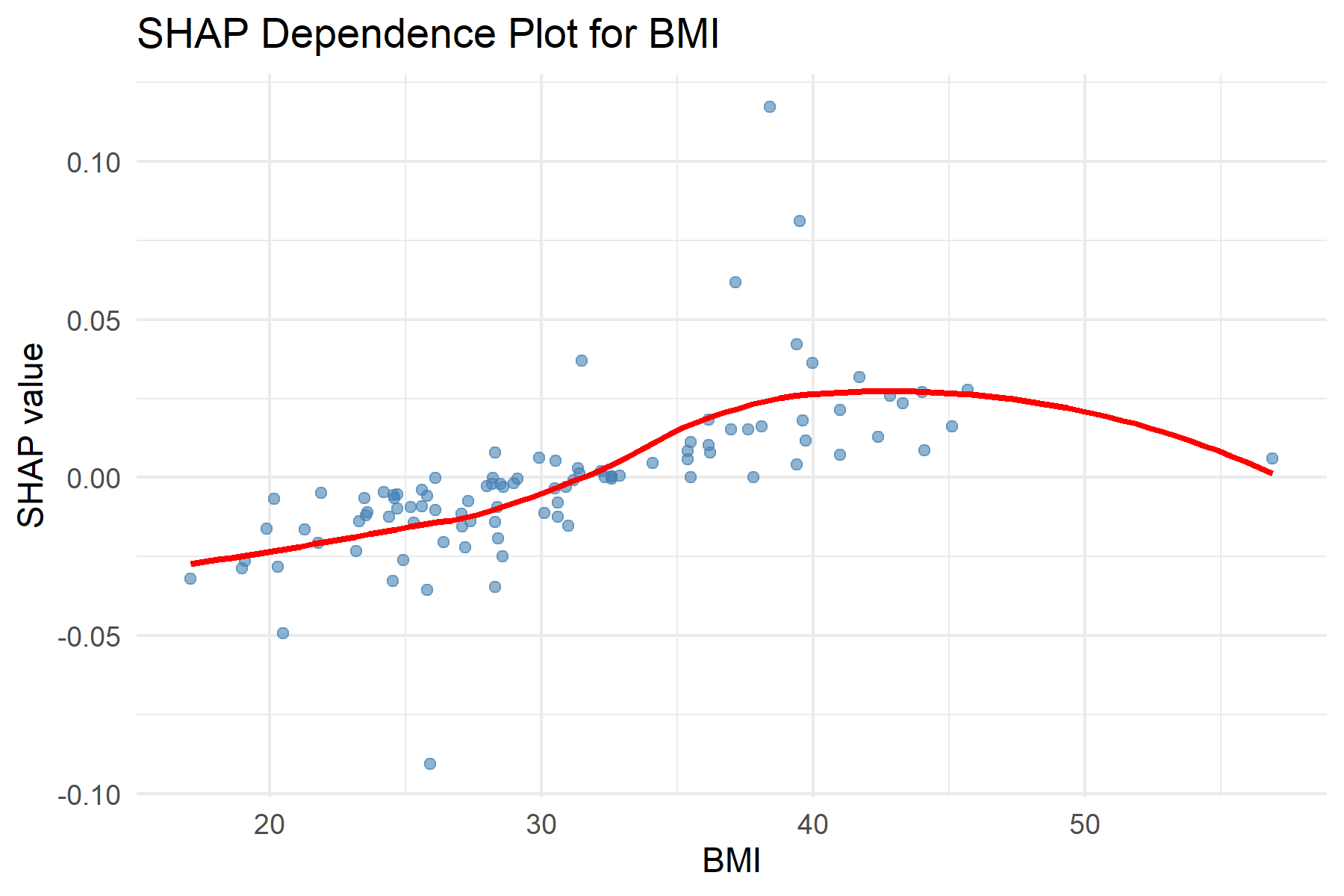
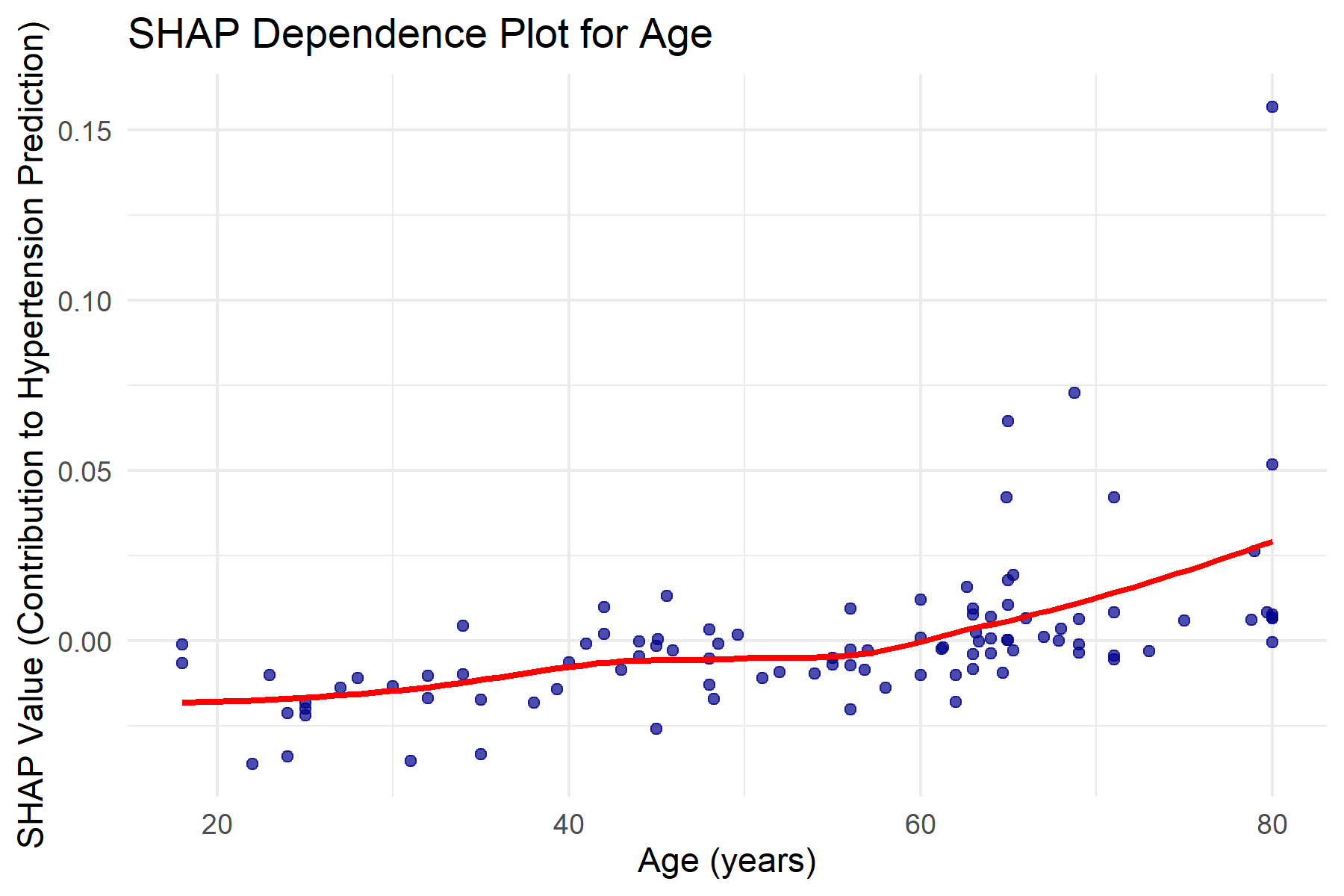


Figure 5. SHAP dependence plot for age



1. Unless otherwise indicated [↑](#footnote-ref-1)