Exploring Personalized Blood Pressure Management through Predictive Modeling: Incorporating Demographic and Dietary Factors

A Preprint

Alexander Hawthorne
University of Michigan
Ann Arbor, MI
hawthoal@umich.edu

Evelyn Paskhaver
University of Michigan
Ann Arbor, MI
evelynpa@umich.edu

Jeanne Yang
University of Michigan
Ann Arbor, MI
jeayang@umich.edu

Li Yuan
University of Michigan
Ann Arbor, MI
leeyuan@umich.edu

December 3, 2023

1 Introduction

Hypertension, a prevalent and frequently asymptomatic health condition, persists as a global health concern, significantly contributing to the burden of cardiovascular diseases (Forouzanfar et al. 2016). Contemporary research on hypertension highlights the multifaceted nature of its determinants, necessitating a comprehensive approach to prediction and management. A seminal study by Iqbal et al. (2021) emphasizes the significance of demographic factors in predicting hypertension prevalence, underscoring the need for nuanced models that account for individual characteristics. Additionally, the work of Johnson et al. (2009) advocates for a personalized approach, accentuating the substantial influence of dietary habits on blood pressure regulation.

Against the backdrop of recent advancements in data science and machine learning, this report aims to solve two questions. Firstly, is it possible to predict people's blood pressure levels only using their demographic information and daily dietary intakes? By initiating a meticulous investigation into the predictive modeling of blood pressure categories with demographic and dietary intake data, we want to provide a new way for personal blood pressure management and think about possible applications of statistical models to achieve this goal.

Secondly, what factors affect people's blood pressures the most? In addition to the blood pressure category prediction, we also want to explore what factors affect people's blood pressures the most, which can provide people with precise insights and strategies on how to prevent from getting hypertension. This aligns with the evolving landscape of precision medicine in cardiovascular health.

2 Data

2.1 Overview

NHANES employs a complex, multistage probability design for sampling the civilian, noninstitutionalized population in the U.S. In 2017-2018, 16,211 persons were selected from 30 survey locations, with 9,254 completing interviews and 8,704 undergoing examinations. Each participant has a unique identification number `SEQN`. To ensure representation, materials were translated into various languages, and cultural competency training was provided to staff (Centers for Disease Control and Prevention 2020a).

According to the Centers for Disease Control and Prevention (2023), NHANES field operations were suspended in March 2020 because of COVID-19. Consequently, data collection for the NHANES 2019-2020 cycle was

incomplete, rendering it non-nationally representative. In response to the disruption caused by the COVID-19 pandemic, we only use those data collected in the 2017-2018 cycle to ensure the study's relevance and generalizability to the U.S. civilian non-institutionalized population.

In the context of this study, the most important data we selected is the examination data of blood pressure (`BPX_J`), which "provides data for three consecutive blood pressure (BP) measurements and other methodological measurements to obtain an accurate BP. Heart rate or pulse, depending on age, are also reported" (Centers for Disease Control and Prevention 2020b). This data contains 4 readings of systolic blood pressure and 4 readings of diastolic blood pressure for each participant. In order to create a response variable about blood pressure level (`BPXLEVEL`), we first average the 4 readings of systolic blood pressure and diastolic blood pressure of each participant respectively. Then we follow the definition of blood pressure levels provided by Centers for Disease Control and Prevention (2021) to divide our average systolic blood pressure and diastolic blood pressure into two blood pressure levels shown in table 1.

The reason why we want to put elevated and hypertension into the risky category is inspired by Volpe, Gallo, and Tocci (2018), who showed the effectiveness of early control on blood pressure. This approach reflects our commitment to prioritizing the identification of individuals at risk, emphasizing early intervention and control measures. By consolidating elevated and hypertension classifications into a unified risk category, we aim to streamline the focus on proactive health management and treatment for those exhibiting possible hypertension signals.

Table 1: Blood Pressure Levels Divided by Systolic and Diastolic Blood Pressure

Blood Pressure Levels	Systolic Blood Pressure		Diastolic Blood Pressure
Normal (`BPXLEVEL` = 0)	< 120 mmHg	and	< 80 mmHg
Risky (`BPXLEVEL` = 1)	≥ 120 mmHg	or	≥ 80 mmHg

After getting the blood pressure levels (`BPXLEVEL`), we merged two other data from the NHANES based on those participants' unique identification number `SEQN`.

First is the Demographic Variables (`DEMO_J`) data, which "provides individual, family, and household-level information" (Centers for Disease Control and Prevention 2020c). Second is the Dietary Interview - Total Nutrient Intakes (`DR1TOT_J`) data, which contains "detailed dietary intake information from NHANES participants. The dietary intake data are used to estimate the types and amounts of foods and beverages (including all types of water) consumed during the 24-hour period prior to the interview (midnight to midnight), and to estimate intakes of energy, nutrients, and other food components from those foods and beverages" (Centers for Disease Control and Prevention 2020d).

By merging data, selecting relevant predictors, and removing some of blank data entries, we got a curated data frame with 6,125 observations and 16 variables. Detail about these 16 variables are shown in table 2.

Table 2: Descriptive Summary of Variables in the Curated Dataset

Variable	Description	Source
`BPXLEVEL`	The blood pressure categories, our response variable. There are two levels, normal is coded as 0 (3,131 observations) and risky is coded as 1 (2,994 observations).	`BPX_J`
`BPXPLS`	Each participant's number of pulses in 60 seconds.	`BPX_J`
`DR1TCARB`	Amount of Carbohydrate intake during the previous day. The unit is gm.	`DR1TOT_J`
`DR1TPROT`	Amount of Protein intake during the previous day. The unit is gm.	`DR1TOT_J`
`DR1TFIBE`	Amount of Dietary Fiber intake during the previous day. The unit is gm.	`DR1TOT_J`
`DR1TTFAT`	Amount of Fat intake during the previous day. The unit is gm.	`DR1TOT_J`
`DR1TCHOL`	Amount of Cholesterol intake during the previous day. The unit is mg.	`DR1TOT_J`
`DR1TRET`	Amount of Retinol intake during the previous day. The unit is mcg.	`DR1TOT_J`
`DR1TVARA`	Amount of Vitamin A as retinol activity equivalents intake during the previous day. The unit is mcg.	`DR1TOT_J`
`DR1TACAR`	Amount of Alpha-carotene intake during the previous day. The unit is mcg.	`DR1TOT_J`
`DR1TBCAR`	Amount of Beta-carotene intake during the previous day. The unit is mcg.	`DR1TOT_J`
`DR1TCRYP`	Amount of Beta-cryptoxanthin intake during the previous day. The unit is mcg.	`DR1TOT_J`

Variable	Description	Source
DR1TLYCO	Amount of Lycopene intake during the previous day. The unit is mcg.	DR1TOT_J
RIAGENDR	Gender of each survey participant. There are two levels, male is coded as 1 (3,007 observations) and female is coded as 2 (3,118 observations).	DEMO_J
RIDAGEYR	Age in years of each participant at the time of screening. Individuals 80 and over are topcoded at 80 years of age.	DEMO_J
RIDRETH3	Race of each survey participant. There are six levels, Mexican American is coded as 1 (875 observations), Other Hispanic is coded as 2 (533 observations), Non-Hispanic White is coded as 3 (2,119), Non-Hispanic Black is coded as 4 (1,443 observations), Non-Hispanic Asian is coded as 6 (765 observations), and Other Race - including multi-Racial - is coded as 7 (390 observations).	DEMO_J

We included DR1TCARB, DR1TPROT, DR1TTFAT, DR1TCHOL as our predictors since carbohydrates, proteins, fats, and cholesterol are macronutrients according to United States Department of Agriculture (2022), and a study conducted by Miller, Erlinger, and Appel (2006) showed that "a carbohydrate-rich diet that emphasizes fruits, vegetables, and low-fat dairy products and that is reduced in saturated fat, total fat, and cholesterol substantially lowered blood pressure."

Moreover, another research conducted by Chen et al. (2002) showed Vitamin A was "positively and significantly associated with both systolic and diastolic BP," whereas Alpha-carotene and Beta-carotene were "was inversely related to both systolic and diastolic BP." Respectively, Rezaei kelishadi et al. (2022) and Zhu et al. (2022) showed that Lycopene and Beta-cryptoxanthin are inversely related to hypertension, meaning that these to micronutrients can lower the risk of getting hypertension. Because of these authoritative studies, we included DR1TRET, DR1TVARA, DR1TACAR, DR1TBCAR, DR1TCRYP, and DR1TLYCO in this study. We expect these predictors to play a critical role in the modeling process and achieve high prediction accuracy.

2.2 Visualization

In this section, we present two scatterplot matrices that provide a comprehensive visual exploration of the dataset. The first matrix focuses on demographic information, offering insights into the relationships and distributions among key demographic variables. The second matrix encompasses macronutrient intakes against the blood pressure level. These visualizations aim to reveal potential patterns, correlations, and trends within the dataset.

Figure 1: Scatterplot Matrices of Blood Pressure Level Against Other Variables

In Figure 1.1, we utilized a color-coded scheme to represent different blood pressure levels: red for normal (BPSLEVEL = 0), and blue for risky (BPSLEVEL = 1). By examining the relationship between blood pressure

levels (BPXLEVEL) and gender (RIAGENDR), noteworthy patterns emerge. The plot reveals a higher prevalence of elevated blood pressure and hypertension among male participants compared to their female counterparts.

Further exploration of blood pressure levels against age (RIAGEYR) exposes intriguing insights. The distributions indicate a skewed pattern, with individuals younger than 20 predominantly exhibiting normal blood pressure levels. However, a concerning trend is observed among those around 60 years old, who are more likely to have elevated or hypertension. Age emerges as a potential influential factor for predicting blood pressure levels in future models.

Analyzing blood pressure levels against race (RIDRETH3) unveils distinct prevalence rates. Non-Hispanic White individuals demonstrate the highest incidence of hypertension, followed by Non-Hispanic Black, Mexican American, and Non-Hispanic Asian individuals. Other Hispanic and other races (including multi-racial), exhibit the lowest elevated or hypertension cases.

Figure 1.2 presents a scatterplot matrix investigating the potential impact of macronutrients intake, including carbohydrates, proteins, fats, and cholesterol, on blood pressure levels. Histograms of macronutrients distributions across all three blood pressure levels reveal right-skewed patterns.

Notably, the analysis highlights substantial correlations among macronutrients variables. The highest correlation is observed between protein intake (DR1TPROT) and dietary fat intake (DR1TTFAT), reaching 0.729. Additional pairs, such as protein intake (DR1TPROT) and cholesterol intake (DR1TCHOL) with a correlation of 0.684, indicate potential multicollinearity among predictor variables. This observation prompts caution when employing certain parametric modeling methods, which may be sensitive to multicollinearity issues.

These findings lay the groundwork for a nuanced understanding of the dataset and underscore the importance of considering demographic and nutritional factors in predicting blood pressure levels. Subsequent sections will delve deeper into statistical analyses and modeling techniques to derive actionable insights from the presented visualizations.

3 Methods

3.1 Data Preparation (One-hot encoding and train test sets spliting)

We employed one-hot encoding to convert categorical variables, RIAGENDR (2 levels), and RIDRETH3 (6 levels), and after applying one-hot encoding to these categorical predictors, our data frame has 20 columns of predictors in total. In addition to one-hot encoding, we randomly selected 80% of our data (4,900 observations) as the training set without replacement and 20% of the data (1,225 observations) as the testing set. In this way, we can evaluate our models objectively.

3.2 Logistic Regression with Lasso Regularization and 10-Fold Cross Validation as a Baseline

We chose logistic regression with Lasso regularization for the baseline model because of its efficacy in handling diverse data characteristics, and its suitability in this case particularly for binary classification between Normal (BPXLEVEL = 0) and Risky (BPXLEVEL = 1) blood pressure levels as classes.

Lasso regularization is also important to help prevent overfitting and control model complexity, potentially deal with the multicollinearity issue; with the lasso penalty, we will be able to identify which predictors are the most influential in predicting the blood pressure class. Both will benefit the interpretability of the model, by focusing on the most influential predictor variables through regularization, and the interpretability of logistic regression makes it easier to understand the impact of each predictor variable.

To find the best lasso penalty term (λ) and rigorously evaluate our model, we implemented a 10-fold cross-validation strategy. This approach provides a robust estimate of the model's performance.

Moreover, the model's performance is systematically assessed by test accuracy primarily, Area Under the Curve (AUC) will also be calculated and compared in case the later models' test accuracies have ties.

3.3 XGBoost Model with 10-Fold Cross Validation

Extreme Gradient Boosting (XGBoost) is a powerful and flexible ensemble learning method recognized for its predictive strength. In order to avoid issues brought by multicollinearity, XGBoost is considered in this study to be expected to achieve a high test accuracy.

XGBoost constructs a sequence of decision trees, leveraging their collective predictions to enhance accuracy and generalize to new data. Incorporating regularization terms, it effectively manages model complexity, preventing overfitting and ensuring robustness - another crucial factor makes us decide to use XGBoost for accurate and generalized predictions.

In our analysis, we employ XGBoost to improve upon the performance achieved by the baseline logistic regression model with Lasso regularization. We will assess the performance of the XGBoost model based on test accuracy. In case the accuracy of XGBoost model is the same as that of logistic regression model, we compare the area under the curve (AUC) to see which one is better. We employ 10-fold cross-validation to ensure robust estimation of these metrics across different data subsets, enhancing model reliability and generalizability.

3.4 Feature Selection with XGBoost Feature Importance

One distinctive feature of XGBoost is its ability to provide valuable insights into feature importance. This is achieved through the computation of importance scores assigned to each predictor, utilizing Gain as the metric, which represents the improvement in accuracy attributed to a specific feature across the model's trees (XGBoost Developer 2022).

Leveraging the XGBoost-derived importance scores to determine the most influential predictors is an important advantage that XGBoost provides. The incorporation of these scores and the resultant focus on only the critical predictors can enhance the predictive performance of our model, providing improved accuracy and AUC scores. More importantly, it can help us extract the most influential predictors.

4 Results

4.1 Logistic Regression Model with Lasso Regularization

The logistic regression model was trained with various lasso regularization strengths, spanning a range from low to high values, using a 10-fold cross-validation strategy. The model's misclassifiation error rate was documented for each regularization strength (λ).

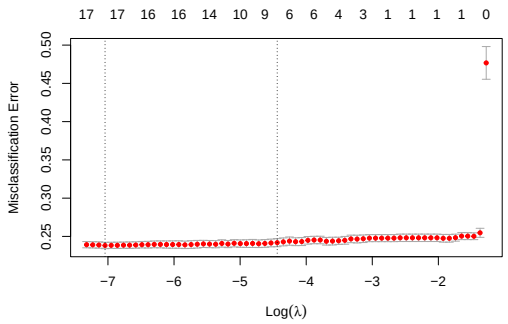

Figure 2: 10-Fold Cross Validation Findinig the Best Lasso Penalty Term

Figure 2 helps identify the optimal λ - the lasso penalty term . We aim to get the λ which minimizes the misclassification rate. The smallest misclassification rate exists when $\lambda = 8.7267008 \times 10^{-4}$ and 17 predictors were selected by the lasso penalty, which is indicated by the vertical dash line on the left side of the plot.

The selected predictors are shown in table 3 below. We can see that race Non-Hispanic White (`RIDRETH33`), Other Hispanic (`RIDRETH32`), and Non-Hispanic Asian (`RIDRETH36`) has the smallest coefficients, meaning people in these races are less likely to have risky blood pressure. Moreover, we can see the increase of age (`RIDAGEYR`) and men (`RIAGENDR1`) are more likely to have risky blood pressure, which is the same as we observed in the visualization part.

The mathematical representation of this logistic regression model is as follows:

Table 3: Coefficients of the Logistic Regression Model with Lasso Penalty

Predictor	Coefficient	Predictor	Coefficient	Predictor	Coefficient
Intercept	-3.6586002	DR1TCHOL	0.0005090	RIDRETH34	0.1951030
BPXPLS	0.0090795	RIAGENDR1	0.5101546	RIDRETH36	**-0.1076511**
DR1TCARB	0.0004857	RIAGENDR2	-4.75×10^{-13}	DR1TRET	-0.0003336
DR1TPROT	0.0005661	RIDAGEYR	0.0655914	DR1TVARA	-0.0000066
DR1TFIBE	-0.0076363	RIDRETH32	**-0.2841042**	DR1TACAR	0.0000315
DR1TTFAT	0.0006496	RIDRETH33	**-0.3780946**	DR1TLYCO	-0.0000030

$$\Pr(\texttt{BPXLEVEL} = 1|X) = \frac{e^{\mathbf{X}\hat{\beta}}}{1 + e^{\mathbf{X}\hat{\beta}}},$$

where

$$\mathbf{X} = \begin{bmatrix} 1 & \texttt{BPXPLS}_1 & \texttt{DR1TCARB}_1 & \cdots & \texttt{DR1TLYCO}_1 \\ 1 & \texttt{BPXPLS}_2 & \texttt{DR1TCARB}_2 & \cdots & \texttt{DR1TLYCO}_2 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & \texttt{BPXPLS}_n & \texttt{DR1TCARB}_n & \cdots & \texttt{DR1TLYCO}_n \end{bmatrix}_{n\times 18}, \hat{\beta} = \begin{bmatrix} -3.6048228 \\ 0.0090795 \\ \vdots \\ -0.0000030 \end{bmatrix}_{18\times 1}.$$

Note the n represents the number of observations of the input dataset. For example, for our training set, the dimension of $\mathbf{X}$ is $4,900 \times 18$, for the testing set, the dimension of $\mathbf{X}$ is $1,225 \times 18$.

In our logistic regression model, we achieved a test accuracy of 74.4489796% and the Area Under the Curve (AUC) was calculated as 0.7442811, demonstrating the model's robust discriminative ability across blood pressure classes. These results establish a baseline for our future modeling efforts, showcasing the effectiveness of the logistic regression approach in capturing and understanding underlying patterns within the dataset.

4.2 XGBoost Model

The XGBoost model was trained using Extreme Gradient Boosting with exact tree method, a powerful ensemble learning method. The following hyperparameters were modified and utilized in the model:

- Learning Rate (eta): 0.005
- Subsample: 0.75
- Column Subsample: 0.8
- Maximum Depth: 10
- Number of Trees (Rounds): 35

With these hyperparameters, we used 10-fold cross-validation to get a test accuracy of 74.9387755% and a test AUC of 0.754929. The test accuracy is greater by 0.4897959 percentage points than that of logistic regression model with lasso regularization, and the test AUC is 0.0106479 higher than that of logstic regression model. These indicate that the XGBoost model is slightly better than logstic regression model with lasso regularization when classifying the blood pressure levels.

4.3 XGBoost Model with Selected Predictors

Figure 3 shows the Gain scores of the predictors used in the XGBoost model. A higher bar (higher Gain score) represents more important the predictor is. Notably, key predictors such as RIDAGEYR (age), DR1TTFAT (dietary fat intake), and DR1TCHOL (dietary Cholesterol intake) emerged as significant contributors to the predictive power of the model.

To enhance our model and identify the most influential predictors, we conducted a systematic feature selection process. Starting with the least important predictor based on the Gain score, we iteratively eliminated predictors, assessing the impact on test accuracy and AUC. This stepwise approach allowed us to identify a

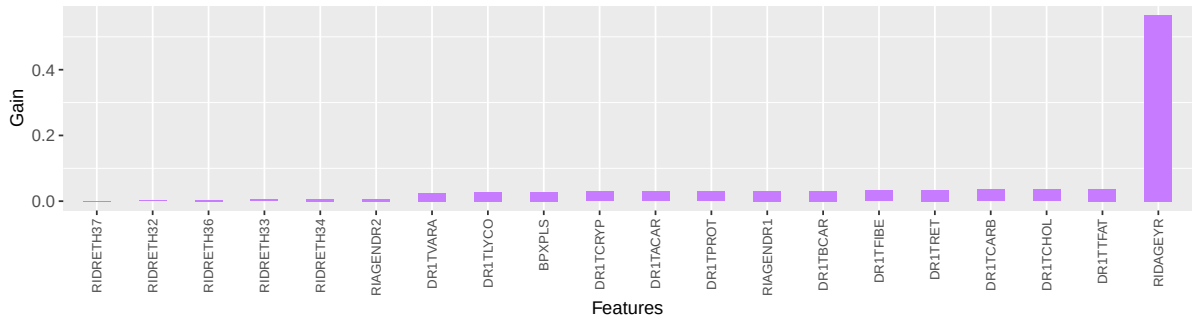

Figure 3: Bar Plots of Gain Score of Each Feathure in the XGBoost Model

subset of predictors maintaining optimal predictive performance. Throughout this process, we retained the same hyperparameters as the original XGBoost model, employing 10-fold cross-validation at each step.

The results of this feature selection journey revealed a compelling trade-off between the number of predictors and predictive accuracy. Significantly, in figure 4, the model showcased a remarkable test accuracy of 74.53061% and an AUC of 0.7501558 even with just the top 11 most important predictors. This underscores the efficiency of the selected predictors in encapsulating crucial information for the accurate prediction of health outcomes. As indicated by the red dash lines in figure 5, the model achieved the highest test accuracy of 75.42857% and the highest test AUC of 0.7613304 with the top 18 most important predictors.

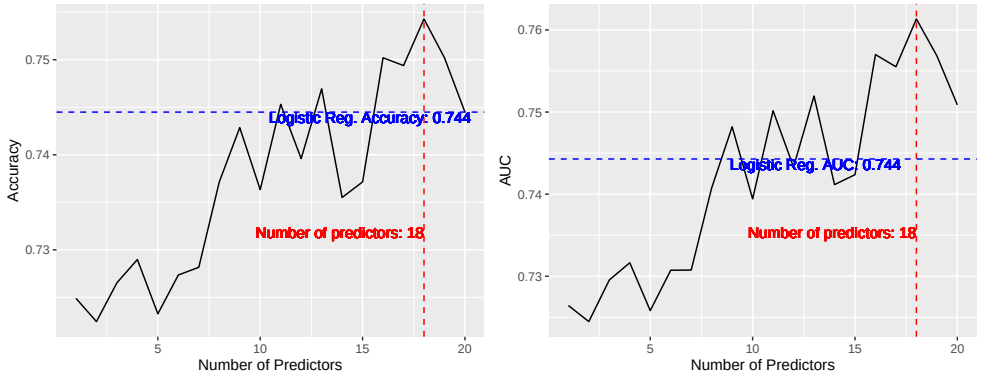

Figure 4: XGBoost Model Test Accuracy and AUC from 1 Predictor to 88 Predictors

Table 4 is a summary table of the test accuracies and test AUC of the logistic regression model with lasso regularization, XGBoost model with all predictors, and XGBoost model with reduced features. Obviously, the XGBoost model using 18 predictors performs the best in both test accuracy and test AUC. Our focus lies on the accuracy and AUC metrics, and, based on these, the model with the top 18 most important predictors stands out as the preferred choice.

Table 4: Summary of Three Models' Test Accuracies and AUCs

Model	Test Accuracy	Test AUC
Logistic Regression with Lasso	74.44898%	0.74428
XGBoost Using 20 Predictors	74.93878%	0.75493
XGBoost Using 18 Predictors	75.42857%	0.76133

In addition to the high test accuracy and AUC brought by the reduced XGBoost model, among these 18 selected predictors, there are 15 predictors selected by both the logistic model and the XGBoost model. These predictors are `RIDAGEYR`, `DR1TTFAT`, `DR1TCHOL`, `DR1TCARB`, `DR1TRET`, `DR1TFIBE`, `RIAGENDR1`, `DR1TPROT`,

DR1TACAR, BPXPLS, DR1TLYCO, DR1TVARA, RIAGENDR2, RIDRETH34, RIDRETH33, RIDRETH36, RIDRETH32. Since both models selected these predictors, indicating the importance and effectiveness of these predictors on predicting blood pressure levels. The other 3 predictors selected by the XGBoost model but not the logistic model are DR1TBCAR, DR1TCRYP, RIDRETH37.

Our systematic approach to feature selection not only fine-tuned the model but also provided insightful perspectives on the pivotal factors influencing its predictive power. This enhanced interpretability contributes to a more robust and effective health outcome prediction system.

5 Conclusions

In this report we successfully developed a statistical model to predict people's blood pressure levels with 75.43% test accuracy, possibly providing a new way for personal blood pressure management. Moreover, through the Lasso penalty and the XGBoost feature importance scores, we were able to determine the 15 most influential factors in our data that affect people's blood pressure levels.

Despite the advancements made in developing predictive models, it's crucial to acknowledge certain limitations. One prominent drawback is the challenge of multicollinearity in our dataset. Given the degree of correlation found between the variables, caution is warranted regarding attributing individuals impacts to any single variable. We might consider the method of logistic regression is inappropriate and might produce misleading result in this study. Thus, in future refinements, further feature selection or regularization techniques may be explored to mitigate the effects of multicollinearity and enhance the robustness of the predictive models.

Another significant drawback is the low accuracy of 75.43%, particularly problematic in health-related predictions where precision is crucial. Inaccuracies can undermine the reliability of personalized recommendations and interventions, impacting effective health management. Exploring advanced machine learning techniques, including deep learning methods like neural networks, holds promise in uncovering complex data relationships and improving overall predictive accuracy.

An exciting application of our predictive models lies in the integration with health apps, such as Apple Health, Samsung Health, and so on. Implementing our models in these platforms could empower individuals to receive personalized daily blood pressure suggestions based on their recorded dietary intakes, known health conditions, and demographic information. This practical application could serve as a proactive tool for users to manage their health more effectively, offering real-time insights and guidance. The envisioned integration with health apps presents an exciting avenue for translating our research into actionable insights, fostering proactive health management among individuals.

6 Contributions

Alexander Hawthorne: Led the development of the introduction and conclusion sections, ensuring a cohesive and impactful narrative for the report.

Evelyn Paskhaver: Took charge of the data description, visualization, and analysis results sections, providing a comprehensive and insightful representation of the dataset and its analysis.

Jeanne Yang: Led the Methodology section, ensuring a robust and well-documented approach to the analysis and findings.

Li Yuan: Played a pivotal role in the report's initial draft, providing foundational content and offering valuable insights into the structure and management of material throughout the writing and coding processes.

7 Reproducibility

To reproduce the analyses, please download R Markdown attached to this CANVAS submission. If you want to reproduce the whole report, including the style and reference files, please download the source code from the following link:

`https://lygitdata.github.io/STATS-415-Project/code/download.zip`

Simply unzip the file, open the R Markdown file in your RStudio, install necessary R libraries from CRAN if not already installed, and knit the R Markdown should successfully generate this report.

Reference

Centers for Disease Control and Prevention. 2020a. "NHANES 2017-2018 Overview." wwwn.cdc.gov. https://wwwn.cdc.gov/nchs/nhanes/continuousnhanes/overview.aspx?BeginYear=2017.

———. 2020b. "2017-2018 Data Documentation, Codebook, and Frequencies Blood Pressure (BPX_j)." wwwn.cdc.gov. https://wwwn.cdc.gov/Nchs/Nhanes/2017-2018/BPX_J.htm.

———. 2020c. "2017-2018 Data Documentation, Codebook, and Frequencies Demographic Variables and Sample Weights (DEMO_j)." wwwn.cdc.gov. https://wwwn.cdc.gov/Nchs/Nhanes/2017-2018/DEMO_J.htm.

———. 2020d. "2017-2018 Data Documentation, Codebook, and Frequencies Dietary Interview - Total Nutrient Intakes, First Day (DR1TOT_j)." wwwn.cdc.gov. https://wwwn.cdc.gov/Nchs/Nhanes/2017-2018/DR1TOT_J.htm.

———. 2021. "Facts about Hypertension." Centers for Disease Control; Prevention. https://www.cdc.gov/bloodpressure/facts.htm.

———. 2023. "2019-2020 Examination Data - Continuous NHANES." wwwn.cdc.gov. https://wwwn.cdc.gov/nchs/nhanes/continuousnhanes/default.aspx?BeginYear=2019.

Chen, Jing, Jiang He, Lee Hamm, Vecihi Batuman, and Paul K. Whelton. 2002. "Serum Antioxidant Vitamins and Blood Pressure in the United States Population." *Hypertension* 40 (December): 810–16. https://doi.org/10.1161/01.hyp.0000039962.68332.59.

Forouzanfar, Mohammad H, Ashkan Afshin, Lily T Alexander, H Ross Anderson, Zulfiqar A Bhutta, Stan Biryukov, Michael Brauer, et al. 2016. "Global, Regional, and National Comparative Risk Assessment of 79 Behavioural, Environmental and Occupational, and Metabolic Risks or Clusters of Risks, 1990–2015: A Systematic Analysis for the Global Burden of Disease Study 2015." *The Lancet* 388 (October): 1659–1724. https://doi.org/10.1016/s0140-6736(16)31679-8.

Iqbal, Afrin, Karar Zunaid Ahsan, Kanta Jamil, M. Moinuddin Haider, Shusmita Hossain Khan, Nitai Chakraborty, and Peter Kim Streatfield. 2021. "Demographic, Socioeconomic, and Biological Correlates of Hypertension in an Adult Population: Evidence from the Bangladesh Demographic and Health Survey 2017–18." *BMC Public Health* 21 (June). https://doi.org/10.1186/s12889-021-11234-5.

Johnson, Rachel K., Lawrence J. Appel, Michael Brands, Barbara V. Howard, Michael Lefevre, Robert H. Lustig, Frank Sacks, Lyn M. Steffen, and Judith Wylie-Rosett. 2009. "Dietary Sugars Intake and Cardiovascular Health." *Circulation* 120 (September): 1011–20. https://doi.org/10.1161/circulationaha.109.192627.

Miller, Edgar R., Thomas P. Erlinger, and Lawrence J. Appel. 2006. "The Effects of Macronutrients on Blood Pressure and Lipids: An Overview of the DASH and Omniheart Trials." *Current Atherosclerosis Reports* 8 (November): 460–65. https://doi.org/10.1007/s11883-006-0020-1.

Rezaei kelishadi, Mahnaz, Omid Asbaghi, Behzad Nazarian, Fatemeh Naeini, Mojtaba Kaviani, Sajjad Moradi, Gholamreza Askari, Mojgan Nourian, and Damoon Ashtary-Larky. 2022. "Lycopene Supplementation and Blood Pressure: Systematic Review and Meta-Analyses of Randomized Trials." *Journal of Herbal Medicine* 31 (February): 100521. https://doi.org/10.1016/j.hermed.2021.100521.

United States Department of Agriculture. 2022. "Macronutrients | National Agricultural Library." Usda.gov. https://www.nal.usda.gov/human-nutrition-and-food-safety/food-composition/macronutrients.

Volpe, Massimo, Giovanna Gallo, and Giuliano Tocci. 2018. "Is Early and Fast Blood Pressure Control Important in Hypertension Management?" *International Journal of Cardiology* 254 (March): 328–32. https://doi.org/10.1016/j.ijcard.2017.12.026.

XGBoost Developer. 2022. "Understand Your Dataset with XGBoost — Xgboost 2.0.2 Documentation." xgboost.readthedocs.io. https://xgboost.readthedocs.io/en/stable/R-package/discoverYourData.html#build-the-feature-importance-data-table.

Zhu, Xu, Mengshaw Shi, Hui Pang, Iokfai Cheang, Qingqing Zhu, Qixin Guo, Rongrong Gao, et al. 2022. "Inverse Association of Serum Carotenoid Levels with Prevalence of Hypertension in the General Adult Population." *Frontiers in Nutrition* 9 (September). https://doi.org/10.3389/fnut.2022.971879.