Predicting Hypertension Using Environmental and Heritable Risk Factors

By: Elizabeth Hora, Isaac Opiyo, Jiayi Zhang, Ziqi Zhou

DATA-200: Foundations of Data Analytics

Professor Batorsky and Professor Georgalis

22 December 2022

A. Abstract

Hypertension (informally known as high blood pressure), is one of the most common sources of Cardiovascular Disease (CVD), the leading cause of death in the United States. Therefore, it is important to analyze which environmental and/or genetic factors may lead to hypertension in order to identify individuals that may need treatment. We studied the factors measured in the National Health and Nutrition Examination Survey (NHANES) longitudinal study. NHANES has laboratory results, questionnaire answers, and examination results, which could all be useful predictor variables for hypertension. However, to eliminate unnecessary predictors in the model, a series of Forward Stepwise Regression models were used to iteratively add statistically significant predictor variables. Classification models that predicted if patients have regular blood pressure or hypertension based on a set of predictor variables were run to better understand the contributing factors for hypertension. The Logistic Regression and Random Forest models identified inherent characteristics of the survey participants such as age and BMI as having the most impact on their status of hypertension. K Nearest Neighbors is the best-performing model for predicting hypertension.

B. Introduction

An individual is diagnosed with hypertension when their systolic pressure is at or above 140 mmHg or the diastolic pressure is at or above 90 mmHg. Hypertension is one of the most common sources of Cardiovascular Disease (CVD), the leading cause of death in the United States (Husain et al., 2014). It is known that the risk of hypertension increases with age, and other hypothesized genetic and environmental risk factors have been identified as well (Benetos, Athanase et al., 2019). Therefore, it is important to analyze which environmental and/or genetic factors may lead to hypertension in order to prevent and treat CVD.

One approach to better understand hypertension and its deterministic factors is to run a classification model that identifies if patients have regular blood pressure or hypertension. The National Health and Nutrition Examination Survey (NHANES) longitudinal study collects data on several thousand Americans entailing medical examinations, laboratory results, and questionnaire responses over several decades. From the available data, we obtained a cross-section of about 6,000 Americans in 2016 and focused on studying the effects on hypertension based on several associated risk factors such as patient demographics, socioeconomic status, gender, level of alcohol consumption, and other dietary information. To eliminate unnecessary predictors in the model, a series of Forward Stepwise Regression models were used to iteratively add important predictor variables. Different classification methods such as Logistic Regression, K-Nearest Neighbors (KNN) and Random Forest were applied to predict whether a given individual will have hypertension. With data collected from the longitudinal study, we want to compare the risk factors that have the greatest impact on hypertension. We hypothesize that inherited factors (e,g., age, gender, and race) are more likely to determine hypertension than environmental factors such as drinking, smoking or people's dietary habits. To test this hypothesis, we applied our classification models to the NHANES 2015-2016 dataset and evaluated the model performance based on the predictors included.

C. Literature Review

From the literature on hypertension, there are two types of studies commonly conducted: longitudinal and cross-sectional studies. Longitudinal studies follow individuals over several years, taking the same measurements to track change in an individual. Cross-section analyses look at a specific time within the longitudinal study to see any trends that exist at that particular time across different individuals. A longitudinal study can have several cross-section analyses within it to test if trends found in one time period are maintained over a longer period of time. It can be challenging to compare longitudinal/cross-section studies with one another because of uncontrollable factors. However, even with these uncontrolled factors, general hypertension trends from different geographic populations (e.g., Brazil and New York) can still be observed and studied.

One large dietary contributor to hypertension is alcohol consumption. Alcohol is a part of many cultures and has been consumed for thousands of years. Lian et al. (1915) were the first to establish a link between alcohol consumption and blood pressure. Studies often categorize participants as having no, moderate, or excessive drinking habits. While the exact boundaries between these categories are not easily defined, one study defines moderate drinking for men

under 65 years old as 2 drinks per day, 1 drink per day for men over 65 years, and 1 drink per day for women of all legal ages (Husain et al., 2014). People who drink in excess of that are therefore classified as excessive drinkers. In the United States alone, 20 million people are affected by alcoholism, and roughly 100 thousand lives are lost due to the effects of alcohol per year (Husain et al., 2014). Since currently there is no medicine that can lower blood pressure to safer levels for all patients, three current strategies are used to help counteract alcohol-induced hypertension: reduction of alcohol consumption over a safe time period to minimize withdrawal symptoms, increased exercise activity, and use of available medicines that can help lower blood pressure to safer levels for the patient on an individual level (Husain et al., 2014). One cross-section study hinted at there being a health benefit to consuming alcohol along with a meal rather than just drinking (Beilin and Puddey, 2006).

One example of a longitudinal study is the Estudo Longitudinal de Saúde do Adultos (ELSA) of 7,000 active and retired civil servants from 2008 to 2010 in Brazil. Santana et al. (2018) found that some cultural risk factors for hypertension such as smoking declined, while other risk factors such as obesity and alcohol consumption increased. In addition, abdominal obesity was observed at a statistically significantly higher level in those with higher blood pressure compared to those with normal blood pressure (Santana et al., 2018). They also found that compared to women, men are three times more likely to exhibit excessive drinking habits as well as consume 80% more alcohol on average, with a difference between consumption being more pronounced in older populations (Santana et al., 2018). Alcohol consumption was only associated with hypertension in women that drank excessively, with no such association detected in women who drink moderate amounts of alcohol. These findings suggest that men could be at higher risk for alcohol-induced hypertension than women (Santana et al., 2018).

In another longitudinal study of Black and Non-Black adults in Erie County, New York with cross-sections in 1986, 1989, and 1993, Russell et al. (1999) observed differences in both Black and Non-Black as well as Female and Male participants. They conceptualized hypertension and its risk factors using a Social Learning Model framework: the idea that personal, environmental, and behavioral factors interact and shape outcomes. They saw alcohol consumption and stress as having a confounding relationship when related to elevated blood pressure since stress can cause hypertension as well as provoke an individual to increase their alcohol consumption to cope with the stress, also causing blood pressure to increase (Russell et al., 1999). They observed that Black men compared to Non-Black men engaged in avoidance coping or using alcohol to temporarily block out a problem they feel they cannot immediately solve. Consistent with the literature, daily drinkers had higher blood pressure than weekly drinkers (Russell et al.).

What we consume plays a vital role in our health outcomes. Thus, it is critical to examine the effect of dietary patterns on blood pressure. Physiologically, blood pressure is influenced by urine and sodium excretion through the Na/K (Sodium/Potassium) reuptake mechanism located in the renal tubule in the kidney (Johnston and Pollock, 2018). Excess sodium intake and insufficient potassium intake have been shown to result in hypertension (Yatabe et al., 2017). To observe the joint effects of sodium intake and potassium intake, the use of the sodium-to-potassium ratio, called the Na/K ratio, has been proposed in various studies (Imamura et al., 2022). The Na/K ratio has been reported to show a stronger association with blood pressure than with sodium or potassium alone (Perez and Chang, 2014). However, according to

the CDC (2022), most Americans eat too little potassium and too much sodium as a result of reliance on processed food diets coming from packaged and restaurant food. This indicates a dietary risk in the population. Potassium and fiber rich foods such as bananas, oranges, melons, cooked spinach and broccoli, potatoes, and sweet potatoes are known to have dietary benefits on blood pressure (Weaver, 2013). Therefore, focusing on the various nutrients individuals consume can be used to study risk-related factors of blood pressure.

There are many unanswered questions surrounding whether the environment or genetics are more responsible for how individuals respond to risk factors for alcohol consumption and hypertension. Unfortunately, there are documented instances that racial and ethnic groups receive different levels of treatment (Gu et al., 2017) (Williams and Rucker, 2000) (Sabin, 2022). The "firewater myth" hypothesized that Native Americans lack the genetics needed to restrain from engaging in excessive drug consumption (e.g., alcohol, tobacco, etc.) (Ehlers and Gizer, 2014). This myth ignores the genetic differences between tribes as well as the degree of integration with other cultures. While it is true that withdrawal symptoms can be heritable, the myth that all Native Americans lack a form of self control has yet to be supported by evidence, and so it is no longer an acceptable position to maintain in the field of medicine. Instead, it is more likely that the extreme socioeconomic position of Native Americans and their lack of resources are responsible for the exposure to risk factors at early ages and the lack of medical attention to help those individuals adjust their habits (Ehlers and Gizer, 2014). Therefore, while genetic components might play a role in how an individual can respond to risk factors, this can be overshadowed by a crippling economic situation that could potentially override whatever genetic condition an individual has.

Advancements in machine learning have improved predictions of hypertension. Yiming Li, Sanjiv J. Shah et al. (2021) utilized both environmental variables and genetic factors to predict the subtypes of hypertension. Their study used the Hypertension Genetic Epidemiology Network (HyperGEN) data with genotype information of 911 African American participants and 1,171 European American participants. The subtypes of hypertension were defined as 'non-hypertensive', 'mild hypertensive', and 'severe hypertensive' according to the clinic blood pressure measurements of participants and the number of medications they were taking. Different variables were used to fit, train and evaluate the performance of these models, including four covariates (age, sex, race, and whether the participant ever smoked cigarettes) known to be relevant to having hypertension, and multiple subsets of genetic predictors. These genetic predictors are SNPs (Single-Nucleotide Polymorphisms, the most common type of genetic variation). They were filtered using different methods, resulting in the number of genetic predictors ranging from none up to 16,227 included along with the covariates. The classification models included in the study were multinomial logistic regression, multi-layer perceptron, and ScanMap. The ScanMap filter with the Exac filter had the best model performance. According to their findings, age, race, and filtered SNPs are the most important variables to predict hypertension. Even though genetic information is not available in the datasets we are using, this research ascertained the significance of age and race variables when predicting hypertension, which is why we include these heritable predictors in our model.

While regular hypertension affects arteries throughout the body, pulmonary hypertension affects arteries in the lungs specifically. To study this particular type of hypertension, Alice Le Brigant et al. (2021) used 204 shapes data of right ventricles (RV) of the heart extracted from

3d-echocardiographic sequences in their analysis because cardiac shape deformations can indicate the likelihood of survival in pulmonary hypertension. They proposed using information geometry tools to classify histograms derived from the medical data, and the data could be represented in the space of beta distributions after fitting the histograms. To be specific, they applied Fisher information metric to compare and classify the histograms with the K-means algorithm, and this method was able to identify healthy control subjects and diseased subjects accurately.

In addition to deciding which data to incorporate into the prediction model, Mohammad Kachuee et al. (2019) pointed out the importance of considering the monetary and non-monetary costs associated with data acquisition. They proposed a cost-sensitive and context-aware approach to address these issues when carrying out classification tasks. In this study, monetary costs refer to the payment required to collect the data, and non-monetary costs include patient discomfort experienced in the process of medical procedures or the potential privacy issues of collecting personal medical or health data. With a cost-sensitive approach, a balance between the predicted results and monetary or non-monetary data acquisition costs can be achieved. They used the NHANES data gathered between 1999 to 2016 for hypertension classification. In regard to data acquisition costs, they conducted a survey to find out 108 people's opinions about the inconvenience of acquiring certain data using the Amazon Mechanical Turk framework in the United States. The results show that higher classification accuracies are often associated with higher acquisition costs, which means the datasets that generate high prediction accuracy are difficult to gain access to. Thus, it is important to take the accessibility of datasets into account when building prediction models.

D. Data Description and Visualization

The NHANES longitudinal study collects data on several thousand Americans entailing medical examinations, laboratory results, and questionnaire responses about dietary habits over several decades. Of the data available, we obtained a cross-section of complete data for 6,361 Americans in 2015-2016 and studied the effects on hypertension based on several associated risk factors such as patient demographics, their socioeconomic status, gender, level of alcohol consumption, and other dietary information.

We explored the variables identified in the literature as having an impact on hypertension. The first variable we examined was how the participant age was related to their systolic and diastolic blood pressure readings. From Figure 3, the distribution of the systolic blood pressure widens and appears to increase for older compared to younger participants. The overall relationship between age and hypertension is the same for different gender groups, with males on average having a higher systolic blood pressure (the regression equation has a greater intercept but a more gradual slope). However, regardless of the gender of the individual, the risk of hypertension increases with age. Therefore, the age of a participant will likely play an important role in the prediction of hypertension.

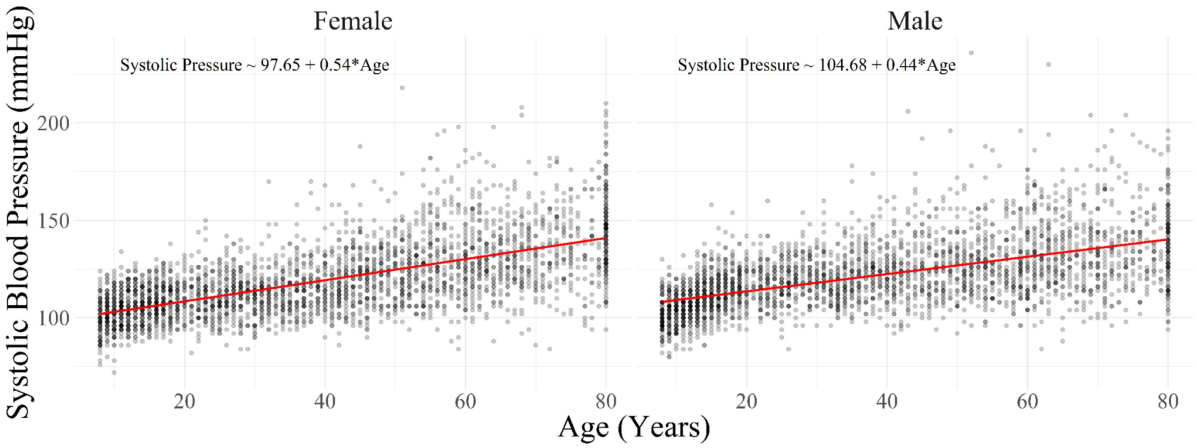

Figure 1: This faceted plot by gender illustrates the Linear Regression of hypertension explained by age. A transparency was applied to the points, with darker regions corresponding to more data points. The slopes of both Linear Regression equations are positive (with the slope for women being steeper), meaning that elders are more likely to be diagnosed with hypertension.

Another possible predictor variable for hypertension we explored was the alcohol consumption of the participants. Studies taken in New York and Brazil found alcohol consumption to be an important factor in their hypertension diagnosis, along with observable drinking habit differences for both genders studied (Russell et al., 1999)(Santana et al., 2018). We see that there are observably different drinking habits for men and women in the NHANES dataset (Figure 2). There does not appear to be a relationship between alcohol consumption and hypertension (Figure 2). However, separating the effects on hypertension due to biological differences in gender will be difficult to parse from behavioral differences (Figure A1). We also checked the relationship between hypertension and sodium intake, but we found no appreciable link (Figure A2).

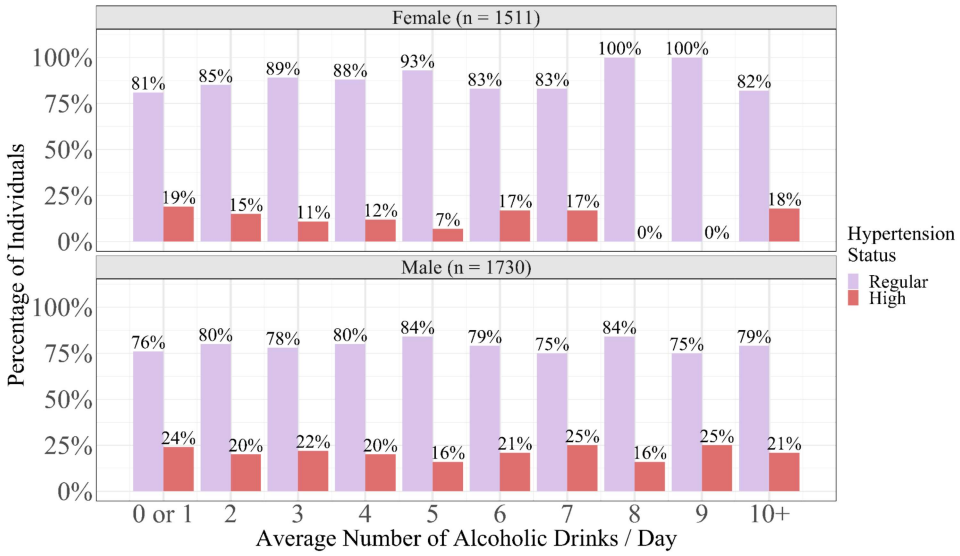

Figure 2: This histogram illustrates the interaction between gender and the number of alcoholic beverages each individual drinks per day, which are both related to hypertension. The percentage of individuals by hypertension diagnosis for each level of drinking is listed at the top of each bar.

While most of the data were in a ready-to-use format, we made some modifications to the dietary information and medication datasets before being able to use those data as predictors in our hypertension models. For dietary information, we grouped by the participant and found the total consumption of a particular type of nutrient. For each participant's current prescription medication, rather than having a specific medication count as a predictor, we used the provided International Classification of Diseases, Tenth Revision (ICD-10-CM) codes provided by NHANES. Codes such as I10 (Essential - primary - hypertension) and I10.P (prevent hypertension) are grouped primarily by the first letter of the code (I referring to heart medication). We opted to group medications by their overall intended purpose (the first letter of the code) and call them sections, with "Section I" corresponding to the heart medications, for example. We also grouped the data by the participant and found the number of medications each individual was prescribed. Most individuals regardless of hypertension status were not on a medication, but some were on multiple. The relationship between the number of prescription medications an individual takes and their systolic blood pressure is shown in Figure 3. The median number of medications an individual takes is greater in those diagnosed with hypertension compared to those with regular blood pressure. Moreover, the length of the box for hypertension is longer than the one for non-hypertension, indicating that there is a greater spread in the number of medications these individuals take.

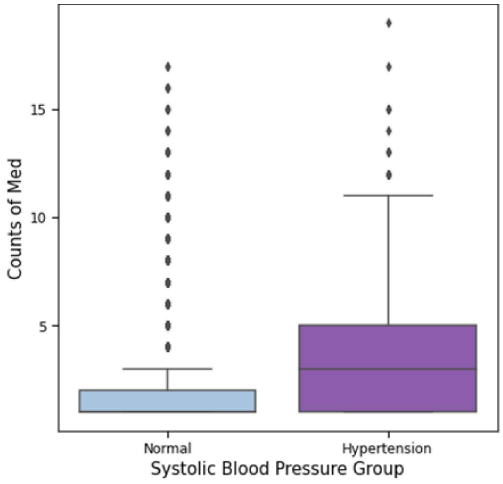

Figure 3: These box and whisker plots compare the counts of medicine for those with and without hypertension. The median number of prescribed medications for individuals with hypertension is greater than those with regular blood pressure. The interquartile range of hypertension samples was much higher than the non-hypertension group, indicating that there is a wider range of the number of medications that individuals with hypertension take.

The results from the data visualizations we conducted display the potential predictor variables needed to accurately predict hypertension in individuals. Although there are observable trends in the data (Figures 1-3), we still need to use a variety of models to understand the impact different predictor variables have in predicting hypertension.

E. Model

In our project, once the data were cleaned, we then ran a series of Forward Stepwise Regression models to select the most important independent variables in predicting hypertension, the dependent variable. We used the 10 best predictors from this feature selection process as well as other predictors we thought that could potentially explain hypertension based on our research. We used several models to predict hypertension: Logistic Regression, K-Nearest Neighbors, and Random Forest. We used 8-fold cross-validation to select the best models to test. We first split our data into training and testing sets. We set the folds equal to 8 to divide all training data into 8 parts. The model randomly selected each part as a validation set eight times, with the remaining non-selected data as the training data. Finally, the average accuracy was calculated over each set of folds (equation shown below). The number of model iterations increases with the number of folds, but there is a greater chance that the results are robust.

$$CV_n = \frac{1}{n}\sum_{i=1}^{n} Accuracy_i \text{ , } with\ n = number\ of\ folds$$

E.1. Forward Stepwise Regression: The goal of the Forward Stepwise Regression process was to select features that were most important in predicting hypertension. The stepwise regression function in R is limited in that it cannot produce a categorial prediction. The classification of whether someone has high or regular blood pressure is based on a combination of systolic and diastolic pressure, which are both continuous values. The stepwise regression can only make predictions for a single continuous variable, so we chose to make predictions for the systolic pressure since that is the factor that most commonly results in a hypertension diagnosis. Another challenge associated with Forward Stepwise Regression is the number of models increases with the number of predictors. Having over 500 predictors is not a feasible problem for Forward Stepwise Regression to attempt. As shown in Figure 4, instead of loading all predictors at once, we ran a Forward Stepwise Regression model on each set of related observations (e.g., demographic survey, alcohol questionnaire, physical examination measurements).

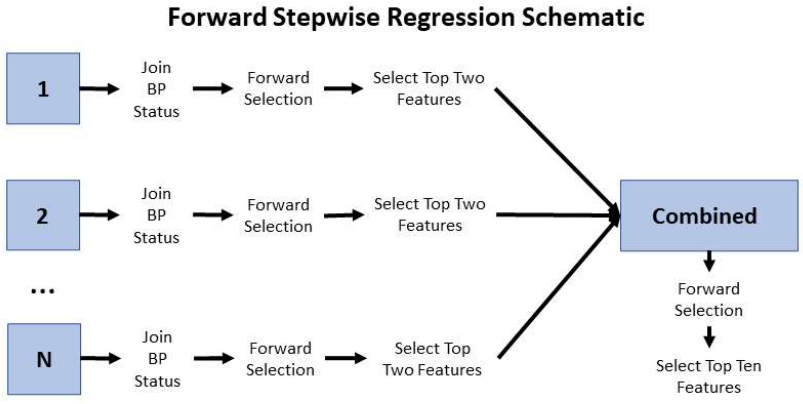

Figure 4: This schematic represents the Forward Selection process. The goal of this process was to extract the most important predictor variables from N different groups of potential predictor variables without having domain knowledge. The box named Combined represents a data table with 2N + 1 columns (2 predictors per group of observations + systolic blood pressure).

A key limitation to this process is that the Forward Stepwise Regression uses a Linear Regression as its model, which makes the assumption that the relationship between the predictor variables and the outcome is linear. The top two predictors from each set of related observations were then combined to have about 30 potential predictors. The best 10 predictors from these were then determined to have the most impact on predicting systolic pressure (see Table A1 for more detail on each predictor).

After running the Forward Stepwise Regression models, we found that the following variables were the most important in the order of: age at the time of screening, Body Mass Index (BMI- the ratio of weight to height), gender, the hours spent on the computer in the past month, the total caffeine intake per day, whether the participant will qualify for food stamps in the next month, the number of prescribed medications, whether they take heart medication, whether they take eye medication, and if they are diabetic. After ensuring the data were ready to use for modeling, we began an exploratory data analysis. The first step we did was to produce a correlogram of variables that we gathered from the literature as potential predictors of hypertension (Figure 5). The correlation is a measure of how two variables are linearly related, with changes in one variable explaining changes in the other. The magnitude of the correlation is related to the strength of the relationship, with a magnitude of 1 being a very strong relationship. Since there are potentially several important predictors for hypertension, we chose to visualize the correlations between sets of variables in the form of a correlogram color-coded by the strength and magnitude of each correlation. One concern for modeling is autocorrelation, which is the amplification of the importance of a particular predictor variable by incorporating another correlated variable that negatively affects the model's predictions.

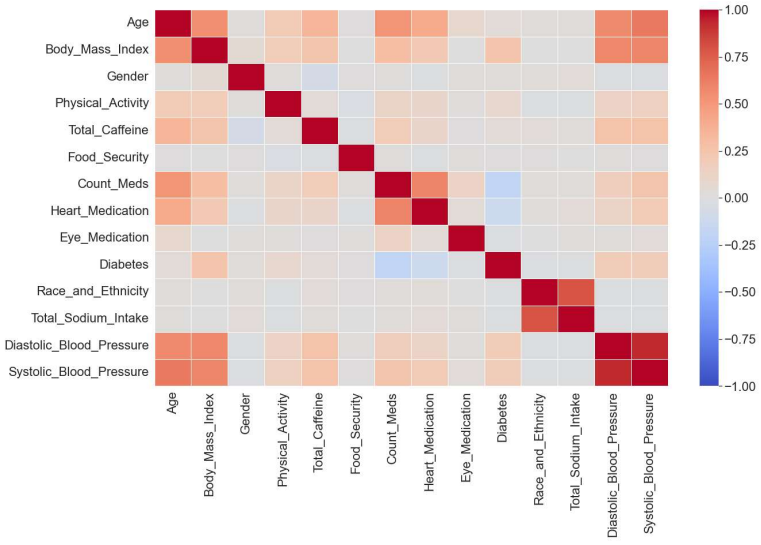

Figure 5. This correlogram illustrates the correlation of potential predictor variables to each other and the variables used to diagnose hypertension. The red squares stand for the strong positive correlations, the blue squares represent strong negative correlations, and the gray squares indicate no correlation. By observing the color display, age has moderately a strong correlation with multiple predictor variables, including BMI, the number of medications, total caffeine intake, and whether the person is diabetic.

E.2. Logistic Regression: In the logistic regression, we used variables that are statistically significant with regard to hypertension as predictors and a binary variable indicating whether a person has hypertension as the response. If a person has more than a 50% likelihood of having systolic blood pressure above 140 or diastolic blood pressure above 90, then this person would be classified as having hypertension. The probability is calculated using the equation below, with n denoting the number of predictors in the model and β denoting the coefficients of corresponding predictors.

$$f(X) = \frac{e^{\beta_0+\sum_{i=1}^{n}\beta_i x_i}}{1+e^{\beta_0+\sum_{i=1}^{n}\beta_i x_i}} = \begin{cases} > 0.5, & hypertensive \\ \leq 0.5, & non-hypertensive \end{cases}$$

E.3. K-Nearest Neighbors: The KNN algorithm is used as a classification model for predicting hypertension. It is an example of instance-based learning, which are algorithms that classify the testing data by directly comparing against the training data stored in memory. The most basic form of the KNN modeling process starts by calculating the distance within the testing data and training data, and then sorts the calculated distances in increasing order. Next, the K neighboring points that have comparably smaller distances with target features are selected and their frequency of occurrence is determined at the same stage. The majority class of the nearest neighboring points determines the classification of the testing data points. This process is repeated for each testing data point.

E.4. Random Forest: Random Forest models can also be used for predicting hypertension. As an ensemble machine learning method, Random Forest uses a combination of multiple Decision Trees to make predictions in a process called bagging. Decision Trees are trained on randomly selected subsets of the original data and features while recording the prediction from each tree. The model then averages all the predictions or finds the mode in order to make the final prediction. We used the Random Forest model because of the randomness it adds to our data and features, which can help address overfitting in smaller datasets. To maximize the model's predictive accuracy, we generated optimized values for hyperparameters such as maximum depth (maximum number of levels in each tree), minimum sample leaf size (minimum number of data points in a leaf node), and minimum sample split (minimum number of data points before the split), before and after resampling.

F. Empirical Analysis

After selecting three different types of models to predict hypertension we used three metrics to assess the quality of model predictions. Accuracy in this context is the number of individuals correctly identified as having or not having hypertension divided by the number of all individuals. Sensitivity (also called recall for binary classifications) is the ratio of the number of correctly identified hypertensive participants to the number of hypertensive participants. Sensitivity will indicate the rate at which the model classifies those with hypertension, with values closer to 1 indicating that most individuals with hypertension are classified as such. Specificity (also called the True Negative Rate) is the ratio of the model correctly identifying those with regular blood pressure to all individuals with regular blood pressure. The specificity will indicate the capability of the model to correctly identify healthy individuals.

One problem with our dataset that challenged our models was a class imbalance: approximately 16% of the individuals in this study had hypertension. To address this issue, we chose two resampling techniques and explored a clustering approach. One resampling technique is to under-sample the dataset by only keeping a percentage of non-hypertension data by randomly removing a proportion of people who do not have hypertension from the dataset. As a result, the two classes have a similar amount of data. The other resampling technique that we chose is to over-sample, which is to create some duplicates of the hypertension data. One potential drawback to resampling is that we could introduce sampling bias based on how the data are split. To preserve general trends in the individuals without hypertension, we clustered the non-hypertension data points, with the number of clusters equalling the number of individuals with hypertension. Then, we replaced the non-hypertension individuals in the original dataset with the centroids of these clusters for classification. There are many applicable clustering methods to balance the two classes; however, in this project, we applied K-Means Clustering on the dataset to obtain the centroids. We applied the under sampling method to each of the models and tested the over sampling as well as clustering method using the Logistic Regression models.

F.1. Logistic Regression: When we applied Logistic Regression to the original dataset, the class imbalance resulted in the majority of individuals classified as not having hypertension. Even though the prediction accuracy was high, it was not a suitable model for hypertension classification because the cost of giving false negative results is much higher than giving false positive results. We therefore tuned the threshold of the Logistic Regression model to achieve a higher sensitivity score. The accuracy, sensitivity, and specificity scores we obtained using different thresholds are shown in Figure A3. We applied the model we obtained using 8-fold cross-validation, and with a threshold of 0.2, the model was able to classify hypertension with high sensitivity and specificity.

In addition to tuning the threshold, we tried balancing the two classes with resampling and clustering methods to minimize the impact of having an imbalanced dataset. After balancing the dataset, the model no longer predicted most individuals as non-hypertensive. The results of applying the logistic model to the dataset balanced with a resampling method and the dataset balanced with a clustering method were similar, as is shown in Table 1. While the accuracy decreased compared to the results obtained using an imbalanced dataset without tuning the threshold, the sensitivity of the model increased. Since both of these methods have a random element, the classification results and significant variables can be slightly different each run. But according to the p-values (as shown in Figure A4), age, BMI, and gender are most significantly associated with hypertension. This is consistent with our hypothesis that inherited factors are more important for predicting hypertension than environmental factors. Other variables such as caffeine intake, the number of medications a person is on, and whether a person is on cardiac medications are also related to hypertension (the corresponding variable names are shown in Table A1). Despite the randomness, the Logistic Regression model can predict hypertension with approximately 75% accuracy and high sensitivity and specificity.

Table 1: Classification Results of Logistic Regression Models

	Accuracy	Sensitivity	Specificity
Using the imbalanced dataset, threshold = 0.5	85.91%	21.79%	96.93%
Using the imbalanced dataset, threshold = 0.2	77.00%	74.29%	77.47%
Using dataset balanced with the under sampling method	76.13%	75.87%	76.35%
Using dataset balanced with the clustering method	73.28%	79.17%	67.75%

F.2. K-Nearest Neighbors: The class imbalance was also a challenge for our KNN models. Figure 6 shows the extensive results of finding the optimal K from cross-validation. Using the 'elbow method', we decided that k = 5 was the optimal value of k for our models. We optimized the KNN parameters and applied an under sampling technique with the goal of increasing accuracy and sensitivity. The results of our models are summarized in Table 2. The accuracy score of the testing data when using k = 5 is 86.08%; however, the sensitivity of 17.36% is low. A low sensitivity means the model fails to classify more than 80% of the individuals with hypertension as having hypertension. In addition to the number of neighbors, we tuned two more parameters, weight and power. Weight refers to the amount of data that is given to the neighbor values. Power refers to the power applied to the Minkowski Distance (e.g., a power of two is commonly known as the Euclidean Distance). The overall accuracy increased slightly from 86.08% to 87.96%. The sensitivity worsened, decreasing from 17.36% to 15.85%. Similar to what was observed in the Logistic Regression models, the original dataset with a class imbalance had a high accuracy but a low sensitivity. Adjusting the ratio of samples with the under-sampling method applied to the original dataset improved the overall accuracy, the sensitivity, the specificity, and the applicability of the model. The accuracy after under sampling the original dataset is 92.34%, indicating that the resampling helped the model in its predictions. The sensitivity increased from 17.36% to 87.46%, illustrating the increase in true positive samples shown in the confusion matrix (Figure A5).

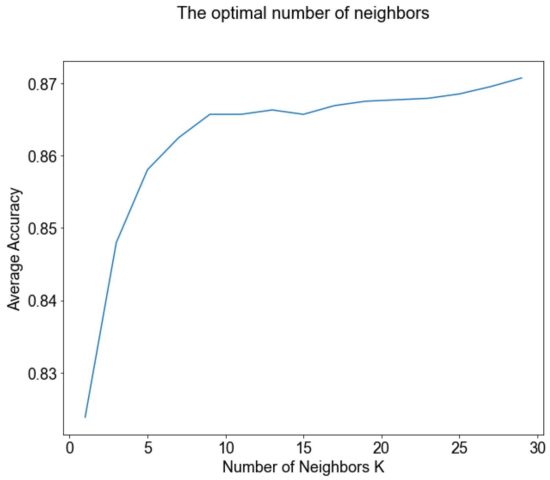

Figure 6: The average accuracy across 8 folds was plotted against the number of neighbors selected for each KNN model. The slope has a sharp increase before plateauing around k = 5. Then, the average accuracy increases gradually.

Table 2: Classification Results of K-Nearest Neighbors Models

	Accuracy	Sensitivity	Specificity
Using imbalance dataset with the optimal k = 5	86.08%	17.36%	95.86%
Using imbalance dataset with the best parameters-k, power, and weight	87.96%	15.85%	98.23%
Using under-sampling dataset	92.34%	87.46%	97.40%

Unlike the Logistic Regression and Random Forest models, KNN cannot identify important features because the model is not designed to do so. However, the results shown in Table 2 illustrate that KNN is an appropriate choice to predict hypertension since the overall accuracy of the best performing model is 92.88% with a sensitivity of 86.74% after under-sampling the original dataset. These were the best results of all models.

F.3. Random Forest: For Random Forest, resampling negatively impacted the model's overall performance except for its sensitivity which increased by a high margin, as shown in Table 3. The results obtained came after tuning the hyperparameters for both the imbalanced and balanced data by separately performing GridSearch cross-validation on each dataset and obtaining optimal parameters for maximum depth, minimum samples leaf' and minimum samples split. The slight drop in accuracy (from 86.27% to 86.13%) after resampling was mainly because of the reduction of data points after under sampling that constrained the model to not get enough random subsets of the original data. While the model's specificity, or ability to predict

healthy individuals, was able to maintain a high score (above 99% before and after resampling), its sensitivity score was still relatively low after resampling. This means the model was constrained by the number of positive cases that it had to learn from even after resampling.

Table 3: Classification Results of Random Forest Models

	Accuracy	Sensitivity	Specificity
Using imbalanced dataset with no resampling	86.27%	2.72%	99.67%
Using balanced dataset with random under sampling	86.13%	44.22%	99.45%

As part of our analysis with Random Forest, we also sought to assess which features were important to our results and which ones were simply adding noise. Doing this enabled us to compare (on the basis of our hypothesis) whether hereditary traits are more important than environmental factors and by how much. As shown in Figure 7, two out of top three most important factors for predicting hypertension using the Random Forest model are hereditary (age and BMI), implying hereditary factors are important in predicting hypertension. Comparing these results with the ones obtained by Logistic Regression, age and BMI still seem to greatly influence the prediction of hypertension.

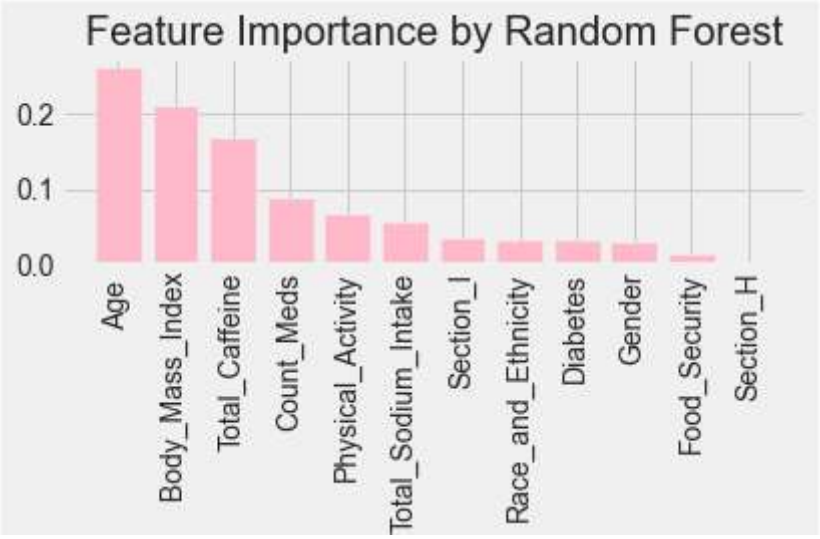

Figure 7: This is a plot of feature importance generated by running the tuned Random Forest model on the undersampled dataset. Larger values correspond to more important features.

To evaluate the fit of our model, we created a learning curve by fitting the Random Forest using 8-fold cross-validation. For each sample, we obtained the training score and the validation score as shown in Figure A6. We observed a small difference between the training and the validation curves that decreased significantly as the model trained on a larger fraction of the

training data. The existing thin margin between these two curves annotates a good fit of the model with little variance across samples.

G. Conclusion

In this project, we built classification models to predict whether a person has hypertension and explored heritable and environmental risk factors that are significantly associated with hypertension. Using the NHANES 2015-2016 dataset, the challenges we faced were the difficulty in identifying important risk factors and the problems of having an imbalanced dataset. It was hard to select pertinent predictors for our models because there were initially hundreds of variables in the datasets with some variables interacting with one another. To address the number of potential predictor variables for hypertension, we applied a series of Forward Stepwise Regression models to extract the overall top 10 variables as predictors to build our classification models for hypertension. To minimize the impact of having an imbalanced dataset, we applied resampling and clustering methods to balance the two classes to avoid obtaining imbalanced classification results. In addition, we wanted to emphasize the models' ability to predict hypertension correctly, so other than prediction accuracy, we calculated sensitivity and specificity metrics to evaluate the models.

We implemented Logistic Regression, K-Nearest Neighbors, and Random Forest models to predict hypertension. With the balanced dataset, Logistic Regression achieved more than 75% accuracy, sensitivity, and specificity score and showed the three variables that are most significantly associated with hypertension are age, BMI, and gender. This model explicitly identified statistically significant predictor variables, which was used to assess if inherited characteristics were more important than environmental factors. The Random Forest models were able to predict hypertension with approximately 86% accuracy and the learning curve indicated that overfitting was minimal. While the specificity score of the best-performing Random Forest model was nearly 100% with a resampled dataset, its sensitivity score was only 44.22%, suggesting its struggles in correctly identifying those with hypertension. In the best-performing Random Forest model, age and BMI were the two most important predictor variables in predicting hypertension. The best-performing K-Nearest Neighbors model had the best sensitivity out of the three classification model types and predicted hypertension most accurately, with 92.9% accuracy achieved on the under resampled dataset.

The results of our models show that heritable factors are more significant predictors than environmental factors in predicting hypertension. The high accuracy, specificity, and sensitivity also indicate that the models we chose are competent in predicting hypertension. Despite the satisfactory results, more work remains to be done. For instance, we could make simpler models that can still maintain comparable accuracies, sensitivities, and specificities as we found. We could also look into other features that may be relevant to hypertension or apply feature transformations since most selected predictors are not statistically significant using linear methods.

H. References

"About Sodium." *Centers for Disease Control and Prevention*, Centers for Disease Control and Prevention, 23 Aug. 2022, www.cdc.gov/salt/food.htm.

Beilin, Lawrence J. and Puddey, Ian B. "Alcohol and Hypertension." *Hypertension,* Vol. 47, No. 6, 2006, pp 1035-1038, https://doi.org/10.1161/01.HYP.0000218586.21932.3c.

Benetos, Athanase et al. "Hypertension Management in Older and Frail Older Patients." *Circulation Research,* Vol. 124 No. 7, 2019, pp 1045-1060, https://doi.org/10.1161/CIRCRESAHA.118.313236.

Ehlers, Cindy L, and Ian R Gizer. "Evidence for a genetic component for substance dependence in Native Americans." *The American Journal of Psychiatry,* Vol. 170, No. 2, 2013, pp 154-64, doi:10.1176/appi.ajp.2012.12010113.

Gu, Anna et al. "Racial and Ethnic Differences in Antihypertensive Medication Use and Blood Pressure Control Among US Adults With Hypertension." *Circulation: Cardiovascular Quality and Outcomes,* Vol. 10, No. 1, e003166, 2017, https://doi.org/10.1161/CIRCOUTCOMES.116.003166.

Husain, Kazim et al. "Alcohol-induced hypertension: Mechanism and prevention." *World journal of Cardiology,* Vol. 6, No. 5, 2014, pp 245-52, doi:10.4330/wjc.v6.i5.245.

Imamura et al. "Association Between Na, K, and Lipid Intake in Each Meal and Blood Pressure". Frontiers in Nutrition., 2022, doi.org/10.3389/fnut.2022.853118.

Johnston JG, Pollock DM. Circadian regulation of renal function. *Free Radic Biol Med.* 2018, 119:93–107. doi: 10.1016/j.freeradbiomed.2018.01.018.

Kachuee, Mohammad, et al. "Cost-sensitive diagnosis and learning leveraging public health data." arXiv preprint, 2019, arXiv:1902.07102.

Le Brigant, Alice, et al. "Classifying histograms of medical data using information geometry of beta distributions." *IFAC-PapersOnLine* Vol. 54 No. 9, 2021, pp 514-520, https://doi.org/10.1016/j.ifacol.2021.06.110.

Li, Yiming, et al. "SNPs Filtered by Allele Frequency Improve the Prediction of Hypertension Subtypes." *IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 2021, pp. 2796-2802, doi: 10.1109/BIBM52615.2021.9669758.

Perez V and Chang ET. "Sodium-to-potassium ratio and blood pressure, hypertension, related Factors". *Adv Nutr*. Vol. 5, 2014, pp. 712–41, doi: 10.3945/an.114.006783.

Russell, Marcia et al. "A longitudinal study of stress, alcohol, and blood pressure in community-based samples of blacks and non-blacks." *Alcohol Res Health,* Vol 23, No. 4, 1999, pp 299-306, PMCID: PMC6760380.

Sabin, Janice A. "Perspective: Tackling Implicit Bias in Health Care." *New England Journal of Medicine,* Vol. 387, 2022, pp 105-107, DOI: 10.1056/NEJMp2201180.

Santana, Nathália Miguel Teixeira et al. “Consumption of alcohol and blood pressure: Results of the ELSA-Brasil study.” *PloS one,* Vol. 13, No. 1, 2018, e0190239, doi:10.1371/journal.pone.0190239.

Weaver, Connie M., “Potassium and Health.” *Adv Nutr.,* Vol. 4, No. 3, 2013, pp 368S-377S, doi: 10.3945/an.112.003533.

Williams, D R, and T D Rucker. “Understanding and addressing racial disparities in health care.” *Health Care Financing Review,* Vol. 21, No. 4, 2000, pp 75-90, PMCID: PMC4194634.

Wu, Chen-Yi et al. “High Blood Pressure and All-Cause and Cardiovascular Disease Mortalities in Community-Dwelling Older Adults.” *Medicine,* Vol. 94, No. 47, 2015, e2160, doi:10.1097/MD.0000000000002160.

Yatabe MS, Iwahori T, Watanabe A, Takano K, Sanada H, Watanabe T, et al. “Urinary sodium-to-potassium ratio tracks the changes in salt intake during an experimental feeding study using standardized low-salt and high-salt meals among healthy Japanese volunteers.” *Nutrients*, Vol. 9, No. 9, 2017, p 951, doi: 10.3390/nu9090951.

I. Appendix

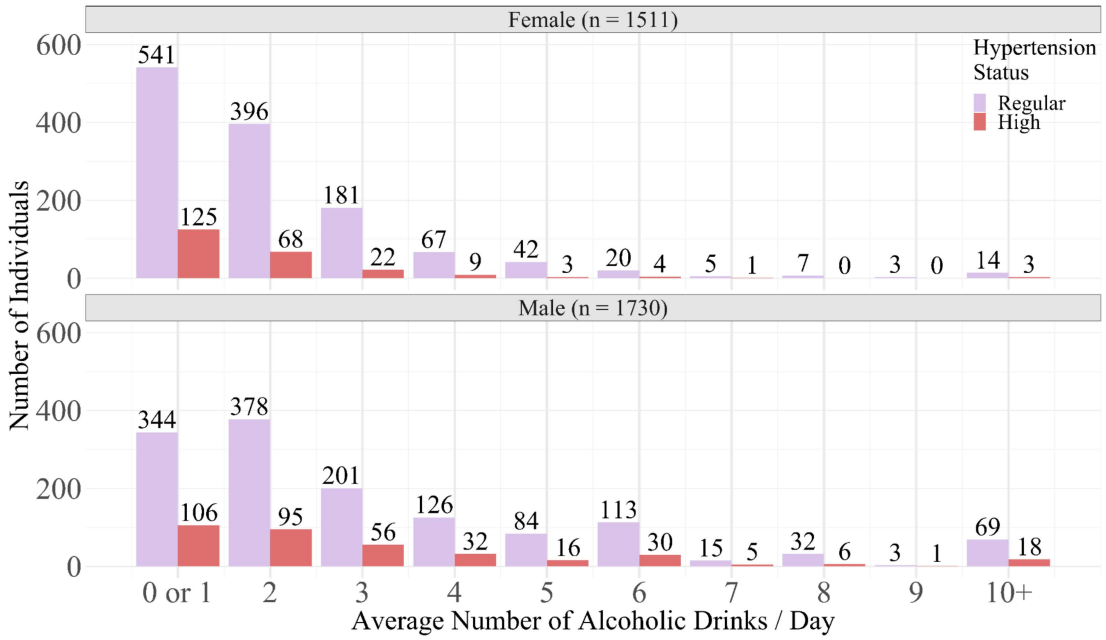

Figure A1: This faceted histogram by gender displays the raw counts of individuals with and without hypertension per drinking level. A larger fraction of women drink 2 drinks or less per day (74.78%) compared to men (53.35%).

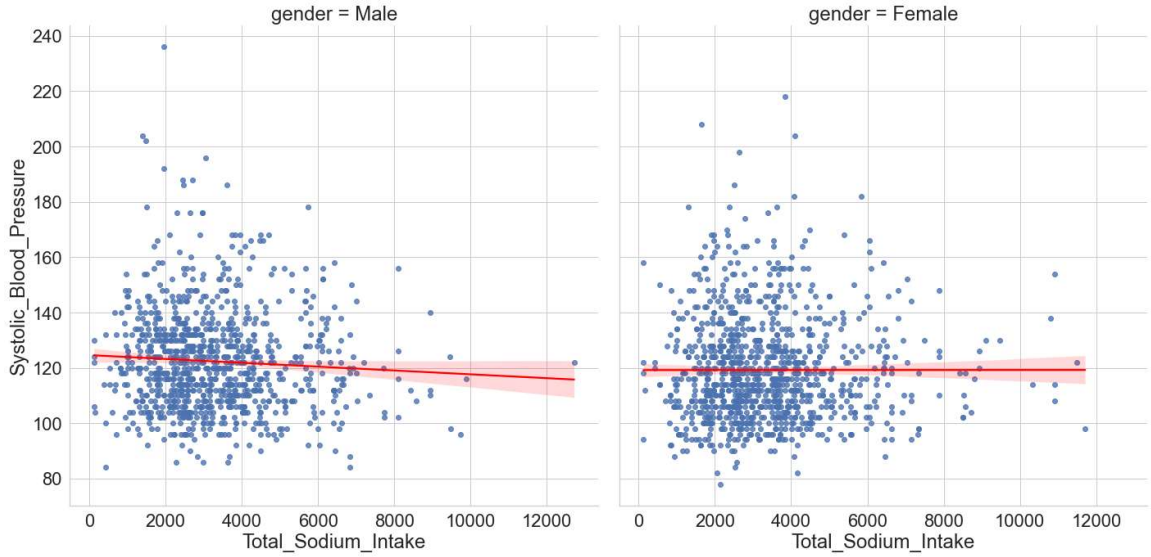

Figure A2: On the basis of our literature review, we decided to investigate the potential relationship between sodium intake and systolic blood pressure. The average sodium intake in mg per day, in our sample, is 3280 for men and 3250 for women. The regression lines are fairly horizontal, implying that there is a low correlation between sodium intake and systolic blood pressure. Contrary to what the literature suggested, it is unlikely that sodium intake as a predictor variable can be used to correctly predict hypertension.

Table A1: Documentation of Predictor Variables

Descriptive Name	NHANES Code Name	NHANES Dataset
Age in Years	ridageyr	DEMO_I[Demo]
BMI	bmxbmi	BMX_I[Exam]
Gender	riagendr	DEMO_I[Demo]
Number of Hours of Computer Usage in the Past 30 Days	paq715	PAQ_I[Q]
Daily Total Caffeine Intake	total_caffeine *	DR1IFF_I[Diet]
Food Stamp Expectancy	fsd855	FSQ_I[Q]
Number of Prescribed Medications	count_meds*	RXQ_RX_I[Q]
If Prescribed Heart Medication	section_I*	RXQ_RX_I[Q]
If Prescribed Eye Medication	section_H*	RXQ_RX_I[Q]
Diagnosed with Diabetes by a Doctor	diq010	DIQ_I[Q]

This table is a descriptive list of the most important predictor variables pulled from the Forward Selection process. * indicates that we performed data cleaning to obtain these variables. The total caffeine intake was calculated by summing the entries per individual, using the original dr1Icaff variable. These data were taken over a 24-hour period to determine the typical caffeine intake of a particular individual. The number of medications was determined in a similar manner, but this time summing all types of medication per individual. The section I and H refer to the ICD-10-CM codes of prescription medications provided by NHANES. The different superscripts in the NHANES Dataset refer to where the data were downloaded from: Demo means from Demographics Data, Diet means from Dietary Data, Exam means from Examination Data, and Q means from Questionnaire Data.

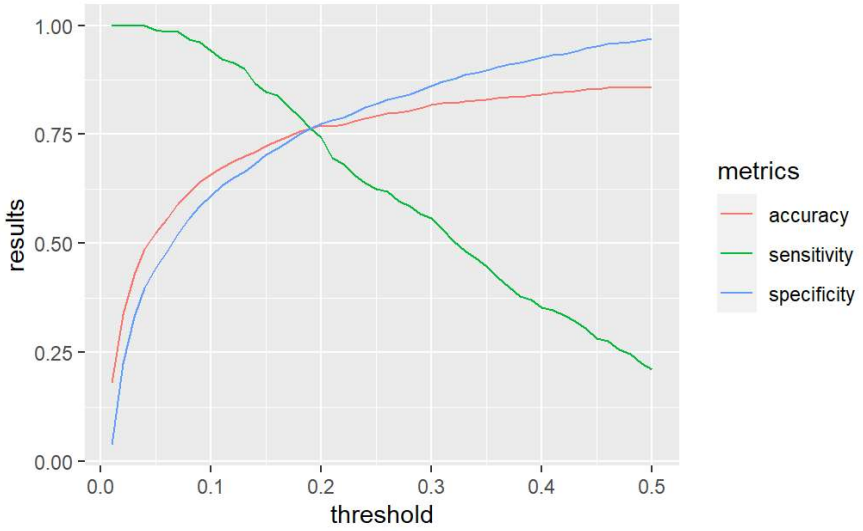

Figure A3: This is a line plot showing the accuracy, sensitivity, and specificity score using different thresholds for the Logistic Regression model.

```
Call:
glm(formula = hypertension ~ ridageyr + bmxbmi + riagendr + count_meds,
    family = "binomial", data = traindata)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.3211  -0.6712  -0.2318   0.7818   2.2991

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -4.112180   0.390392 -10.533  < 2e-16 ***
ridageyr     0.073861   0.004170  17.711  < 2e-16 ***
bmxbmi       0.038467   0.009384   4.099 4.15e-05 ***
riagendr    -0.360800   0.133557  -2.701   0.0069 **
count_meds  -0.061765   0.027352  -2.258   0.0239 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2002.3  on 1444  degrees of freedom
Residual deviance: 1391.2  on 1440  degrees of freedom
AIC: 1401.2

Number of Fisher Scoring iterations: 5
```

Figure A4: This is the table of coefficients obtained by applying the Logistic Regression model on the under-sampled balanced dataset. Only significant predictors were kept in this model.

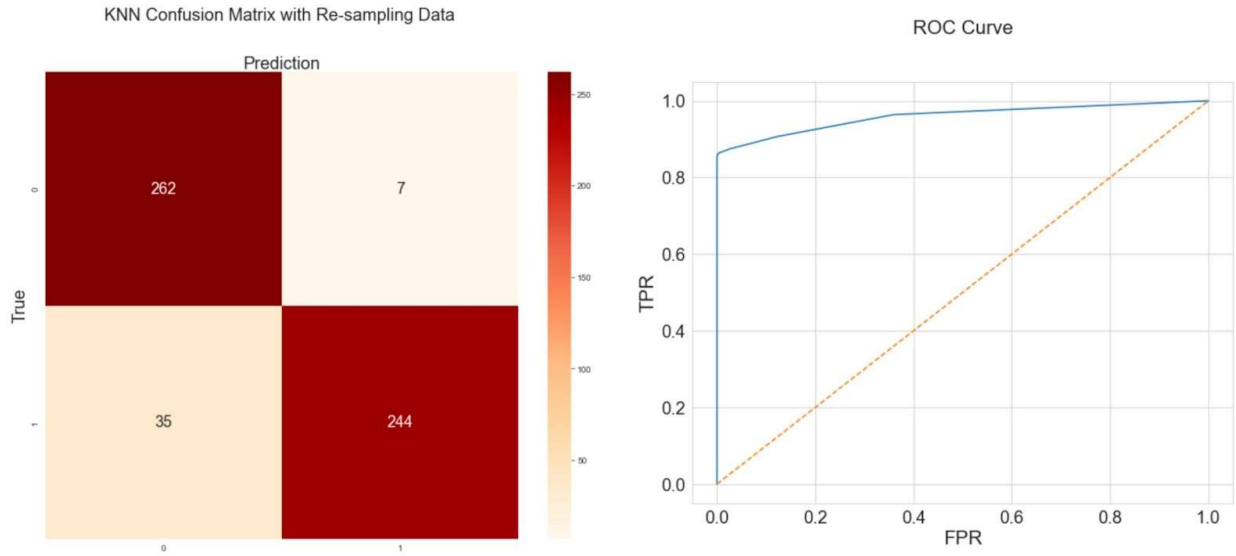

Figure A5: The confusion matrix is the result of the KNN model with k = 5 applied to the undersampled data. The ROC curve represents the model performance. The blue line is the ratio of TPR and FPR. It is a steep slope which means the model fits the data well.

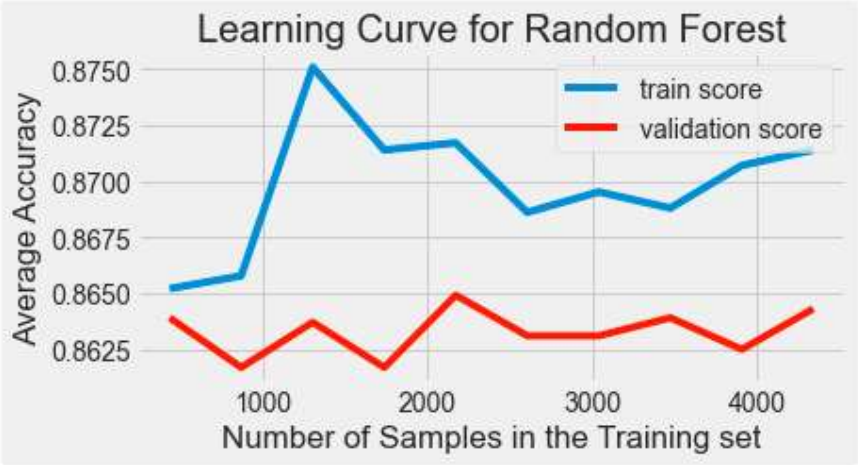

Figure A6: This Learning Curve for the Random Forest model using the under sampled data illustrates that overfitting is minimal since the average accuracy across 8-folds for training and validation are very close.

All code used for our analysis can be found on our GitHub Repository.