Title

Predicting Hypertension with Add Health Dataset using Machine Learning Models

Author

Fan, Zihan

Publication Date

2024

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA
Los Angeles

Predicting Hypertension with Add Health Dataset using Machine Learning Models

A thesis submitted in partial satisfaction
of the requirements for the degree
Master of Applied Statistics and Data Science

by

Zihan Fan

2024

ABSTRACT OF THE THESIS

Predicting Hypertension with Add Health Dataset
using Machine Learning Models

by

Zihan Fan
Master of Applied Statistics and Data Science
University of California, Los Angeles, 2024
Professor Yingnian Wu, Chair

High blood pressure is a prevalent health concern worldwide, and identifying the factors that contribute to its development is crucial for prevention and management strategies. This study aimed to investigate the influence of sex, hereditary factors, habitats, and BMI on the risk of high blood pressure using machine learning techniques. Several models, including Logistic Regression, Decision Trees, Random Forests, XGBoost, Support Vector Machines, and Neural Networks are employed on the public-use sample from the Add Health dataset.

The thesis of Zihan Fan is approved.

Maria Cha

Frederic Paik Schoenberg

Guang Cheng

Yingnian Wu, Committee Chair

University of California, Los Angeles

2024

To my family ...

who always support me

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

Introduction

I was inspired to work on this specific topic because of my personal experience. During a recent annual medical examination, my 25-year-old friend received a warning from her doctor. The doctor cautioned that if she did not make significant lifestyle changes, she would likely develop diabetes and high blood pressure in the near future. This news was alarming, given her young age.

Around the same time, I had a routine blood test. As the nurse drew my blood, I noticed a distinct yellow substance in the sample tube, which the nurse identified as fat. This immediately sparked concern about my own health, as I feared my triglyceride levels would be dangerously high due to my dietary habits. Although my test results were within the normal range, the experience left a lasting impact on me.

Motivated by these events, I conducted research and discovered a worrying trend [HAB20]. A significant number of young people were found to have visible fat in their blood samples, and many chronic diseases typically associated with older age groups were increasingly affecting younger individuals.

These personal experiences and the concerning patterns I uncovered in my research ignited a passion within me to investigate the factors that contribute to the development of chronic diseases, particularly high blood pressure, among younger populations. By understanding the complex interplay of variables such as sex, heredity, habitats, and BMI, I hope to contribute to the development of more effective prevention strategies and targeted interventions. This thesis represents my dedication to addressing this critical public health issue and promoting healthier lifestyles for generations to come.

To further investigate the factors contributing to high blood pressure, I conducted a little research and identified several key risk factors, including family history, obesity, lack of exercise, and excessive salt intake. These findings align with the growing body of evidence suggesting that lifestyle factors play a crucial role in the development of hypertension.

I wanted to explore these risk factors using real-world data and advanced analytical techniques. To achieve this, I decided to work with the Add Health dataset, a comprehensive longitudinal study that provides a wealth of information on the health and well-being of adolescents to adults in the United States.

By leveraging this dataset and applying various machine learning models, such as Logistic Regression, Decision Trees, Random Forests, XGBoost, Support Vector Machines, and Neural Networks, I aim to gain a more nuanced understanding of how factors like sex, heredity, habitats, and BMI influence an individual's likelihood of developing high blood pressure. These powerful models will allow me to uncover complex patterns and relationships within the data that might not be apparent through traditional statistical methods.

Through this research, I hope to contribute to the development of more accurate risk assessment tools and targeted prevention strategies. By identifying the most important predictors of high blood pressure risk, healthcare professionals and policymakers can design more effective interventions and educational campaigns to promote healthier lifestyles and reduce the burden of this chronic condition on individuals and society as a whole.

In the following sections of this thesis, I will describe the methodology used in my analysis, present the results of my machine learning models, and discuss the implications of my findings for public health practice and future research.

CHAPTER 2

Data

2.1 Data Introduction

The National Longitudinal Study of Adolescent to Adult Health, also know as Add Health is a nationally representative, longitudinal study of adolescents in the United States.[HU09] The study first started in 1994-1995 with a sample of more than 20,000 adolescents in grades 7-12 and has followed the participants into adulthood. Recently, in 2016-2018, the most recent wave of data, wave 5 was collected. Add Health employs a multistage, stratified, school-based, cluster sampling design, which ensures that the sample is representative of U.S. adolescents with respect to region, urbanicity, school size, school type, and ethnicity. The study collects data on a wide range of topics, including physical and mental health, social and economic well-being, education, employment, relationships, and health-related behaviors.

For the purpose of this thesis, I will be using data from multiple waves of the Add Health study, focusing on variables related to high blood pressure risk factors, such as sex, heredity, habitats, BMI, and lifestyle behaviors.

For this study, I will be using data from Waves I, IV, and V of the Add Health dataset. Wave I, conducted between 1994 and 1995, included a 45-minute in-school questionnaire completed by 90,118 students from 145 middle, junior, and high schools. A subset of 20,745 adolescents was then selected for an in-home interview. Family history data was also collected during Wave I through a family questionnaire.

Wave IV, conducted in 2008 with 15,701 original Wave I respondents, focused on the develop-

mental and health trajectories from adolescence into young adulthood. The survey questions were expanded to cover a wide range of topics, including educational transitions, economic status, sleep patterns, illnesses and medications, physical activities, emotional content and quality of relationships, childhood maltreatment, substance addiction, and work-family balance. For this study, I will use data from Wave IV that contains information about respondents' lifestyles and their height/weight measurements.

Wave V, the most recent wave of data collection, took place during 2016-2018. The primary objective was to collect social, environmental, behavioral, and biological data to track the emergence of chronic diseases as the cohort progressed through their fourth decade of life. The Wave V design included a mixed-mode survey, with an in-home interview administered to a sub-sample of respondents to analyze mode effects. Repeat measurements of anthropometric, cardiovascular, metabolic, and inflammatory indicators were collected to assess the change in and/or onset of chronic diseases, including obesity, hypertension, diabetes, and dyslipidemia. New biomarkers of chronic kidney disease were also introduced. For this study, I will use data from Wave V to determine whether respondents have developed high blood pressure.

One thing to notice is that the data is divided into public-use data and restricted-use data. In this paper, the public-use data will be studied. This give me less observation to work with which is around 3000. By leveraging data from these three waves of the Add Health study, I aim to investigate the complex relationships between lifestyle factors, family history, and the development of high blood pressure over time. The longitudinal nature of the dataset allows for a unique opportunity to examine how risk factors in adolescence and young adulthood may influence the onset of chronic diseases later in life.

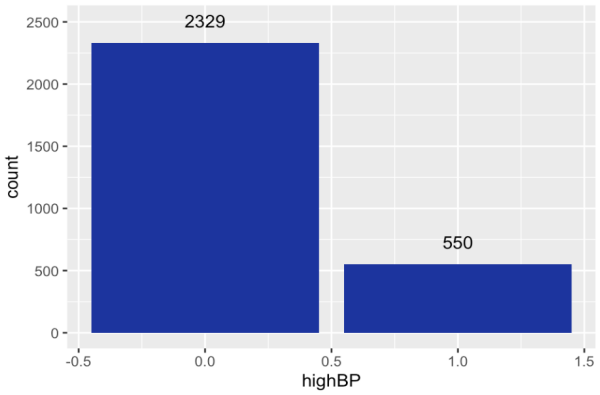

Figure 2.1: Frequency of High Blood Pressure before resampling

2.2 Data Preparation

While exploring the dataset, there are several things need to be considered. Because different models have different assumptions and requirements, data is cleaned in order to improve the model performance. The problems including the conversion of categorical variables, the imbalance in the target variable.

The dataset contains a categorical variable called "inspection", which represents the last time the interviewee has had any health inspection, the possible value of this variable include: '(1) (1) Within the past 3 months', '(2) (2) 4 to 6 months ago', '(3) (3) 7 to 9 months ago', '(4) (4) 10 to 12 months ago', '(5) (5) Longer than 1 year ago but less than 2 years ago', '(6) (6) 2 years ago or longer', '(7) (7) Never'.

Since some machine learning algorithms cannot directly handle categorical variables, I applied one-hot encoding to convert the "inspection" variable into multiple binary variables. One-hot encoding works by creating a new set of binary variables for each unique category in the original variable. A value of 1 indicates the presence of that category and 0 otherwise. This transformation allows the models to effectively incorporate the information from the categorical variable.

	0	1
Frequency	1864	1864

Table 2.1: Class distribution of High Blood Pressure after SMOTE

Upon examining the target variable, the next issue with the data is discovered to be class imbalance. The imbalance is very obvious from figure 2.1. Intuitively, it is easy to understand this imbalance, because naturally there would be way more people who does not have hypertension than those who have it. But this class imbalance can lead to biased models that primarily predict the majority class, resulting in high overall accuracy but poor performance in identifying the minority class .

There are several different ways often used to address this issue. The one used in this study is the Synthetic Minority Over-sampling Technique (SMOTE)[CBH02] to resample the training data. SMOTE is an oversampling technique that creates synthetic examples of the minority class by interpolating between existing minority instances. By generating additional synthetic samples, SMOTE helps to balance the class distribution in the training data, allowing the models to learn more effectively from both classes.

After applying SMOTE, a significant improvement happened in the models' performance, particularly in terms of the ROC AUC metric. The ROC AUC increased from around 0.35 to over 0.5, indicating that the models were better able to discriminate between the two classes after resampling.

The other way to resolve this problem is to undersample the majority class, but the big draw back is that some important feature could be losed during the process. The model results also agress with the drawback of the model because the model performance is not improved with this method, so SMOTE is used in the training data of all the models used in this study.

CHAPTER 3

Models

3.1 Logistic Regression

Binary logistic regression is a statistical method used to model the relationship between a binary dependent variable and independent variables. It is a popular choice for binary classification problems due to its simplicity and interpretability. The probability of an instance belonging to a particular class is modeled based on a linear combination of the input features. The probabilistic outputs make it easy to interpret and set decision thresholds. Logistic regression is suitable for this problem as it can effectively model the relationship between lifestyle and demographic factors and the likelihood of developing high blood pressure. However, it assumes linearity in the data, which may not always be the case, and can be sensitive to outliers and multicollinearity.

In this study, the dependent variable is hypertension (1 = yes, 0 = no). The independent variables include lifestyle factors, such as diet, physical activity, as well as demographic characteristics like sex, age, and family history of hypertension.

The logistic regression model estimates the probability of the dependent variable taking the value of 1 (presence of hypertension) based on the values of the independent variables. The general form of the logistic regression equation is:

$$P(Y = 1 \mid X) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_n X_n)}}$$

In this equation, p is the probability of the dependent variable being 1, $\beta_0$is the intercept term, $\beta_1, \beta_2, ..., \beta_k$are the regression coefficients for the independent variables $X_1, X_2, ..., X_k$

The regression coefficients represent the change in the log odds of the dependent variable for a one-unit change in the corresponding independent variable, holding all other variables constant. A positive coefficient indicates that an increase in the independent variable is associated with an increased likelihood of the dependent variable being 1, while a negative coefficient suggests the opposite.

Table 3.1 shows the coefficients of logistic regression. According to the table, family history plays a significant role in determining an individual's likelihood of developing high blood pressure. The model coefficients reveal that having a mother with diabetes (diabetes mom) increases the risk of high blood pressure, as indicated by the positive coefficient of 0.269159. Similarly, having a father with diabetes (diabetes dad) also contributes to an increased risk, albeit to a lesser extent, with a coefficient of 0.105080. These findings suggest that genetic factors and shared environmental influences within families can have an impact on an individual's blood pressure.

One interesting fact from this table is that, the model coefficient for obesity in the mother (obesity mom) is negative (-0.351274), which is unexpected and contradicts the general understanding that obesity is a risk factor for high blood pressure. This result should be interpreted with caution and may require further investigation. On the other hand, having a father with obesity (obesity dad) is associated with an increased risk of high blood pressure, shown by the positive coefficient of 0.297618. This could reveal the complex interplay between genetic predisposition and environmental factors in the development of high blood pressure.

The model also takes into account of an individual's last health inspection. The inspection variable is one-hot encoded, meaning that it is represented by multiple binary variables indicating different time intervals since the last inspection. The coefficients for these variables provide insights into how people treat health check-ups affect the risk of high blood pressure. For example, having had an inspection within the past 3 months has a positive coefficient of 0.041420, suggesting a slightly increased risk compared to the reference category. As

Feature	Coefficient
SEX	-0.665395
diabete mom	0.269159
diabete dad	0.105080
obesity mom	-0.351274
obesity dad	0.297618
bmi	0.084645
fast food	0.050631
sweetdrink	0.010986
exercise	-0.033571
inspection(1) (1) Within the past 3 months	0.041420
inspection(2) (2) 4 to 6 months ago	-0.312523
inspection(3) (3) 7 to 9 months ago	-0.437495
inspection(4) (4) 10 to 12 months ago	-0.348100
inspection(5) (5) Longer than 1 year ago but ...	-0.307282
inspection(6) (6) 2 years ago or longer	-0.587113
inspection(7) (7) Never	-0.336818

Table 3.1: Logistic Regression Coefficients

the time since the last inspection increases, the coefficients become more negative, with the most negative coefficient (-0.587113) associated with inspections that occurred 2 years ago or longer (inspection(6)). This trend implies that regular health check-ups and timely interventions may play a role in managing and preventing high blood pressure.

It is important to note that the model intercept is -2.63599118, which represents the baseline risk of high blood pressure when all other variables are zero. This negative intercept suggests that, in the absence of other risk factors, the probability of having high blood pressure is relatively low.

In conclusion, the logistic regression model provides valuable insights into the factors influencing the risk of high blood pressure. Family history, particularly diabetes in parents, emerges as a significant predictor, highlighting the role of genetic predisposition. The timing of health inspections also appears to be associated with the likelihood of developing high blood pressure, emphasizing the importance of regular check-ups and early intervention. However, the unexpected finding regarding obesity in mothers warrants further investigation to better understand its implications. These results contribute to our understanding of the complex interplay between various factors in the development of high blood pressure and can inform strategies for prevention and management.

3.2 Decision tree and Random Forest

3.2.1 Decision Tree

Decision tree is a machine learning algorithm that recursively partitions the data into smaller subsets using the explanatory variables. It is also used a lot for binary classification problems due to the ability to handle both numerical and categorical data. One big advantage that increase the popularity of decision tree is its interpretability. This model can capture complex, non-linear relationships. Decision trees are suitable for this problem as they can effectively model the interactions between various lifestyle and demographic factors. However, they are prone to overfitting, especially when the trees have more levels. Small changes in the data can lead to different tree structures. Additionally, their performance might not be as good for high-dimensional data.

At each internal node, the tree selects a single explanatory variable and creates a binary split based on a simple rule. For numeric variables, the split takes the form of $x < k$ vs $x \geq k$, where k is a threshold value. For categorical variables, the split groups the categories into two subsets, such as $x \in \{A, B\}$ vs $x \in \{C, D\}$.

The binary partitioning process continues until a stopping criterion is met, such as reaching a maximum depth or a minimum number of instances per leaf. The final subsets, represented by the leaf nodes, contain observations that share a common predicted value. In other words, a decision tree can only make a limited number of unique predictions, equal to the number of leaf nodes.

In this study, a classification tree will be employed to predict hypertension, which is a binary categorical variable. The tree will evaluate all possible splits at each node and select the one that minimizes an impurity measure. Gini impurity index is used in this model.

The Gini impurity index measures the probability of misclassifying a randomly chosen instance if it were randomly labeled according to the distribution of classes in the subset.

	Precision	Recall	F1-Score	Support
0.0	0.81	0.78	0.79	465
1.0	0.21	0.24	0.22	111
accuracy			0.67	576

Table 3.2: Classification Report for Decision Tree

The Gini impurity[BFO84] for a node t is calculated as:

$$\text{Gini}(t) = 1 - \Sigma_i \, (p_i)^2$$

where p_i is the proportion of instances belonging to class i in the node. A Gini impurity of 0 indicates a pure node, where all instances belong to the same class. Conversely, a Gini impurity of 0.5 indicates an equal mix of classes, representing the maximum impurity.
At each split, the decision tree algorithm selects the feature and threshold that minimize the weighted average of the Gini impurity of the child nodes. This greedy approach aims to create subsets that are as homogeneous as possible compared to the target variable.

3.2.2 Random Forest

Random forests are an ensemble learning method that builds upon the concepts of decision trees to improve their performance and overcome some of their limitations, such as overfitting. They are suitable for this problem as they can capture the intricate interactions between various risk factors and provide insights into the most important predictors of high blood pressure. However, they are less interpretable than individual decision trees and have increased computational complexity compared to single decision trees. The key idea behind random forests is to introduce randomness into the tree-building process, creating a diverse set of trees that collectively make more accurate predictions.

In a random forest, a number of decision trees are constructed using different subsets of the training data and different subsets of the explanatory variables. This randomization helps to reduce overfitting and increases the model's generalization ability.

The random forest algorithm works as follows:

Bootstrap resampling: Create multiple bootstrap samples from the original training data. Each bootstrap sample is generated by randomly selecting instances with replacement

Random feature subset selection: At each node of a tree, instead of considering all available explanatory variables for splitting, only a random subset of variables is considered. This step further increases the diversity among the trees in the forest.

Tree construction: Build a decision tree for each bootstrap sample using the randomly selected feature subsets. Each tree is grown to its maximum depth without pruning, allowing them to capture complex interactions among the variables.

Ensemble prediction: To predict for a new instance, the random forest aggregates the predictions from all the individual trees. For classification tasks, the final prediction is determined by majority voting.

By combining the predictions of multiple diverse trees, random forests aim to achieve better performance than any single tree. The randomization introduced in the tree-building process helps to decorrelate the trees, reducing the risk of overfitting and improving the model's

	Precision	Recall	F1-Score	Support
0.0	0.83	0.87	0.85	465
1.0	0.32	0.25	0.28	111
accuracy			0.75	576

Table 3.3: Classification Report for Random Forest

ability to generalize to unseen data.

One of the advantages of random forests is their ability to provide a measure of variable importance. By calculating the average decrease in impurity across all trees when a specific variable is used for splitting, random forests can identify the most influential predictors of high blood pressure. This information can be valuable for understanding the underlying risk factors and guiding future research and interventions.

The classification reports for the Decision Tree and Random Forest models in table 3.2 and table 3.3 indicate notable differences in their performance. The definition of metrics used here will be discussed in the Result section of this paper.

In terms of precision, the Random Forest model achieves a slightly higher precision for class 0.0, with a score of 0.83 compared to the Decision Tree's 0.81. For result 1.0, which means the people that do have high blood pressure, the improvement is more significant, with the Random Forest achieving a precision of 0.32, whereas the Decision Tree only reaches 0.21. This indicates that the Random Forest model is more accurate in predicting the minority class.

Regarding recall, the Random Forest model also outperforms the Decision Tree. For the negative class, the Random Forest has a recall of 0.87, which is higher than the Decision Tree's recall of 0.78. For positive group, although the recall improvement is modest, the Random Forest still performs slightly better with a recall of 0.25 compared to the Decision

Tree's 0.24. This suggests that the Random Forest model is more effective at identifying true positives, particularly for the majority class.

When considering the F1-score, which balances precision and recall, the Random Forest model again shows superior performance. For negative group, the F1-score of the Random Forest is 0.85, while the Decision Tree achieves 0.79. For positive group, the Random Forest's F1-score is 0.28, significantly higher than the Decision Tree's 0.22. This indicates that the Random Forest model maintains a better balance between precision and recall, especially for the minority class.

Overall accuracy is another key metric where the Random Forest model demonstrates its superiority. The Random Forest model achieves an accuracy of 0.75, compared to the Decision Tree's 0.67. This overall higher accuracy reflects the Random Forest's ability to generalize better across both classes, making fewer errors in classification.

In conclusion, the Random Forest model has consistently better performance compared to the Decision Tree model. This is shown from in its higher precision, recall, and F1-scores for both classes, with particularly good improvements for the minority class. These improvements suggest that the Random Forest model is more robust and reliable, making it a more suitable choice for this classification task.

3.3 XGBoost

Extreme Gradient Boosting, also known as XGBoost, works by combining multiple decision trees and minimize the loss function to create a strong predictive model. XGBoost is an implementation of the gradient boosting framework. They are suitable for this problem as they can capture the intricate interactions between various risk factors and provide insights into the most important predictors of high blood pressure. However, they are less interpretable than individual decision trees and have increased computational complexity compared to single decision trees.

The objective function of XGboost is:[CG16]

$$\mathcal{L}^{(t)} = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{(t-1)} + f_t\left(\mathbf{x}_i\right)\right) + \Omega\left(f_t\right)$$

Where l represents the loss function, for this binary classification task, log likelihood of the bernoulli distribution is used as the loss function.

The term f_t refers to the t^{th} tree, and $y^{(t-1)}$ represents the prediction made by the model for the i^{th} instance at the previous iteration $(t-1)$. Ω is the regularization term which is used to penalize the complexity of the model to prevent overfitting. The other method used in this model to reduce overfitting is a max_depth parameter. This parameter limit the maximum depth of the trees to achieve that purpose.

The feature importance plot for the XGBoost model provides valuable insights into the relative importance of various features in predicting the target variable.

The most important feature in the XGBoost model is "fast_food", with an importance score of 172.0. This suggests that an individual's fast food consumption habits play an important role in the model's predictions. The high importance of this feature indicates that it has a significant impact on the target variable and that the model heavily relies on this information for making accurate predictions.

The second most important feature is "sweet_drink", with an importance score of 116.0. This

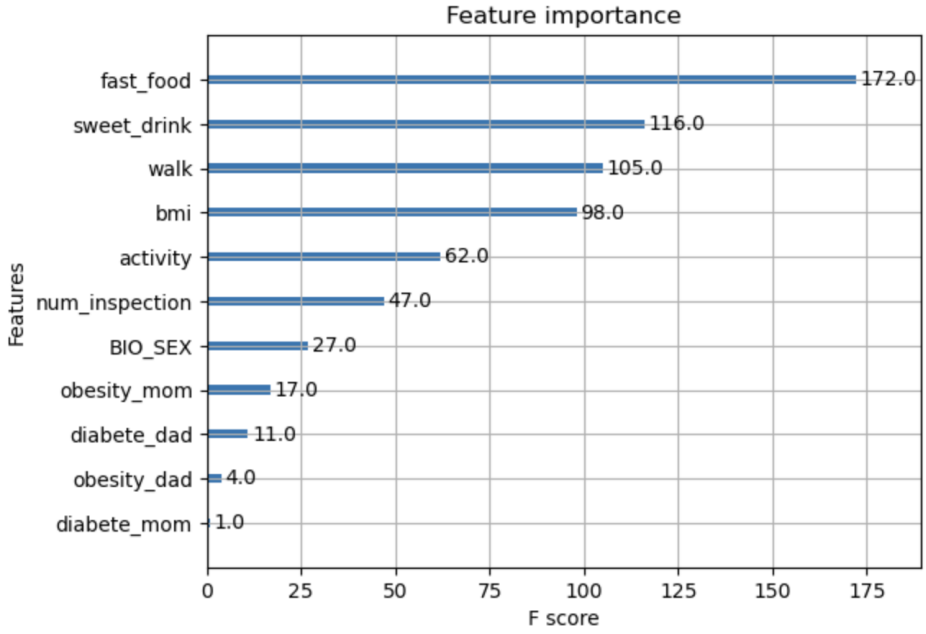

Figure 3.1: Importance rank with XGBoost

implies that an individual's consumption of sweet drinks is also a key factor in the XGBoost model's decision-making process. The model considers this feature to be highly informative in predicting the target variable.

"walk", with an importance score of 105.0, is the third most important feature. This suggests that an individual's walking habits or physical activity levels are significant predictors in the XGBoost model. The model likely captures patterns related to the relationship between walking and the target variable.

"bmi" (Body Mass Index) is the fourth most important feature, with a score of 98.0. This indicates that an individual's BMI plays a notable role in the model's predictions. The XGBoost model considers BMI to be a relevant factor in determining the target variable.

"activity", with an importance score of 62.0, and "num_inspection", with a score of 47.0, are the fifth and sixth most important features, respectively. These features relate to an individual's overall physical activity levels and the number of health inspections they have

undergone. The model assigns moderate importance to these features, suggesting that they contribute to the predictions but to a lesser extent compared to the top four features.

The remaining features, such as "BIO_SEX" (biological sex), parental obesity ("obesity_mom" and "obesity_dad"), and parental diabetes ("diabete_dad" and "diabete_mom"), have relatively lower importance scores. This indicates that these features have a limited impact on the XGBoost model's predictions compared to the lifestyle and health-related features mentioned above.

In summary, the feature importance plot for the XGBoost model highlights the significant role of lifestyle factors in predicting the target variable. Fast food consumption, sweet drink consumption, walking habits, and BMI are the most influential features in the model's decision-making process. Physical activity levels and the number of health inspections also contribute to the predictions, but to a lesser extent. Demographic factors, such as biological sex and parental health conditions, have relatively lower importance in the XGBoost model. These findings provide valuable insights into the key drivers of the model's predictions and can guide further analysis and interpretation of the results.

3.4 Support Vector Machines(SVM)

Support Vector Machines is a powerful machine learning algorithm used for classification and regression tasks. SVMs can effectively handle non-linear relationships. They are suitable for this problem as they can capture complex interactions between risk factors and are robust to outliers and noise in the data. However, they are computationally expensive for large datasets, sensitive to the choice of kernel function and hyperparameters, and less interpretable than other models.

The main idea behind SVM is to find the optimal hyperplane, which could be considered as a straight line in two dimention, that separates the different classes with maximum margin. The margin refers to the distance between the hyperplane and the closest data points from each class, called support vectors. These support vectors define the hyperplane's position.

The primal problem is a constrained optimization problem. It maximizes the margin of hyperplane and minimize the classification error. The equation used by primal problem is to[sks]

$$\begin{aligned}&\text{minimize}(1/2)w^\top w + C\sum_{i=1}{}^{n}\xi_i\\&\text{subject to } y_i\left(w^\top\Phi\left(x_i\right)+b\right)\geq 1-\xi_i, \xi_i\geq 0, i=1,\ldots,n\end{aligned}$$

Here, w is the weight vector, C is the regularization parameter, ξ_i are the slack variables, y_i are the class labels, $\Phi\left(x_i\right)$ is the feature mapping function, and b is the bias term. The objective is to minimize the norm of the weight vector and maximizes the margin while penalizing misclassifications through the slack variables. The primal problem is then simplified into dual problem, with the introduce of lagrange multipliers as follows:

$$\begin{aligned}\min_{\alpha,\alpha^*}\tfrac{1}{2}\left(\alpha-\alpha^*\right)^T Q\left(\alpha-\alpha^*\right)+\varepsilon e^T\left(\alpha+\alpha^*\right)-y^T\left(\alpha-\alpha^*\right)\\\text{subject to } e^T\left(\alpha-\alpha^*\right)=0\\0\leq\alpha_i,\alpha_i^*\leq C, i=1,\ldots,n\end{aligned}$$

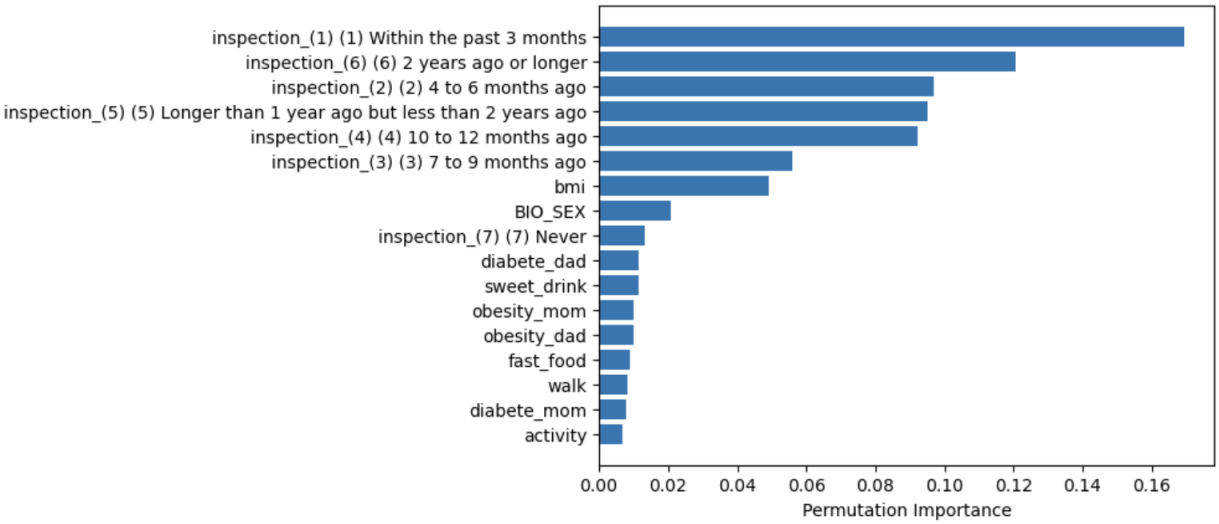

Figure 3.2: Importance rank with SVM

The prediction function that is used to classify the output is:

$$\sum_{i \in SV} (\alpha_i - \alpha_i^*) K(x_i, x) + b$$

Where α_i and α_i^* are the Lagrange multipliers for the support vectors from training process. $K(x_i, x)$represents the kernel function, and the kernel function i used is 'rbf' defined as:

$$K(\mathbf{x}_1, \mathbf{x}_2) = \exp\left(-\gamma \cdot \|\mathbf{x}_1 - \mathbf{x}_2\|^2\right)$$

The permutation importance plot for the Support Vector Machine model provides insights into the relative importance of various features in predicting the target variable. The one-hot encoded "inspection" variable, which represents the timing of an individual's last health inspection, plays a dominant role in the SVM model's predictions.

The top three most important features are all related to the "inspection" variable. "inspection_(1) (1) Within the past 3 months" has the highest permutation importance, indicating that individuals who have had a health inspection within the past 3 months significantly

contribute to the model's predictions. "inspection_(6) (6) 2 years ago or longer" and "inspection_(2) (2) 4 to 6 months ago" also have high importance scores, suggesting that inspections conducted 2 years ago or longer and those conducted 4 to 6 months ago are influential factors in the model's decision-making process.

Other inspection-related features, such as "inspection_(5) (5) Longer than 1 year ago but less than 2 years ago", "inspection_(4) (4) 10 to 12 months ago", and "inspection_(3) (3) 7 to 9 months ago", also have notable permutation importance scores. This reinforces the idea that the timing of health inspections at various intervals plays a crucial role in the SVM model's predictions.

However, unlike most of the other models used in this study "bmi" (Body Mass Index) has a lower permutation importance. While still an important feature, its relative influence on the SVM model's predictions is less important than the inspection-related features.

The "BIO_SEX" (biological sex) feature and the "inspection_(7) (7) Never" feature have similar permutation importance scores, suggesting that they contribute to the model's predictions to a certain extent, but their impact is relatively lower compared to the other inspection-related features.

Lifestyle factors and family history, such as "sweet_drink" consumption, parental obesity ("obesity_mom" and "obesity_dad"), and "fast_food" consumption, have lower permutation importance scores in the SVM model. This indicates that these features have a limited influence on the model's predictions compared to the inspection-related features and BMI.

Parental diabetes ("diabete_dad" and "diabete_mom"), "walk"ing habits, and physical "activity" have the lowest permutation importance scores among the given features, suggesting that they have minimal impact on the SVM model's predictions.

In summary, the permutation importance plot for the SVM model highlights the dominant role of the one-hot encoded "inspection" variable in predicting the target variable. The timing of health inspections, particularly recent inspections within the past 3 months, inspections conducted 2 years ago or longer, and those conducted 4 to 6 months ago, are the

most influential factors in the model's predictions. BMI, biological sex, and the absence of health inspections ("inspection_(7) (7) Never") also contribute to the model's decision-making process, but to a lesser extent compared to the inspection-related features. Lifestyle factors and parental health conditions have relatively lower importance in the SVM model. These findings provide valuable insights into the factors driving the SVM model's predictions and can guide further analysis and interpretation of the results.

3.5 Neural Networks

Neural networks are a class of machine learning algorithms inspired by the human brain. They are suitable for this problem as they can effectively model the intricate relationships between lifestyle, demographic, and other risk factors, and the development of high blood pressure. However, one big problem with neural networks is that they get overfitted easily, especially with small datasets. Neural networks are also less interpretable than other models. In this study, a neural network trained with backpropagation and ReLU activation will be employed. The structure of the neural network is consisted three parts: an input layer, one or more hidden layers, and an output layer. The input layer will have a number of neurons equal to the number of input features. The hidden layers will contain a varying number of neurons, which will be determined through training. The output layer will have a single neuron with a ReLU activation function, producing a probability estimate for the presence of high blood pressure.

The neurons in each layer are connected to the neurons in the subsequent layer by weighted edges. The weights represent the strength of the connections and are learned during the training process. Each neuron receives input from the previous layer, computes a weighted sum of the inputs, and applies an activation function to introduce non-linearity.

ReLU(x) m̄ax(0, x)

Cross entropy loss function is:[Bis06]

$$L = -\frac{1}{N}\left[\sum_{j=1}^{N}\left[t_j \log\left(p_j\right) + \left(1 - t_j\right)\log\left(1 - p_j\right)\right]\right]$$

It returns 0 for negative inputs and the input value itself for positive inputs, introducing a non-linearity that helps the network learn complex patterns.

The training process will be performed using backpropagation, a supervised learning algorithm that iteratively adjusts the weights of the network to minimize a loss function. The steps involved in backpropagation are: Forward propagation: The input features are passed through the network, and the output is computed based on the current weights.

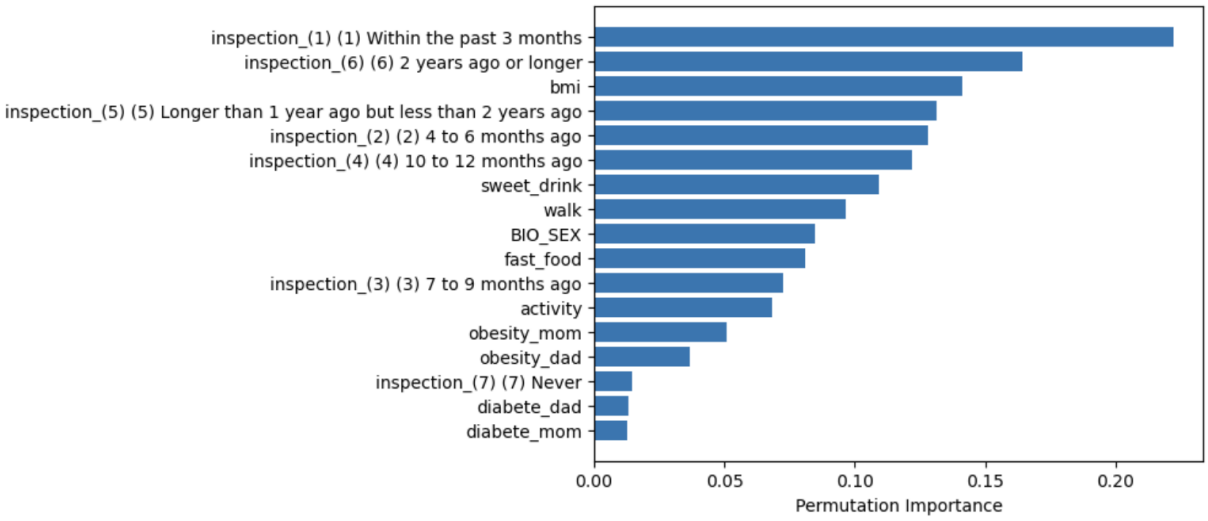

Figure 3.3: Importance rank with Neural Network

Loss computation: The predicted output is compared to the true labels using a loss function, such as binary cross-entropy for binary classification tasks.

Backpropagation: The gradients of the loss function with respect to the weights are computed using the chain rule, starting from the output layer and propagating backwards through the network. The ReLU activation function has a simple derivative:

ReLU'(x) = 1 for $x > 0$

ReLU'(x) = 0 for $x \leq 0$

This property makes the computation of gradients efficient during backpropagation.

Weight update: The weights are updated using an optimization algorithm.

Iteration: Steps 1-4 are repeated for a fixed number of epochs or until a stopping criterion is met.

The permutation importance plot provides insights into the relative importance of various features in the neural network model's predictions.

The most striking observation from the plot is the dominant importance of the "inspection"

variable. The two most important features, "inspection_(1) (1) Within the past 3 months" and "inspection_(6) (6) 2 years ago or longer", suggest that the timing of the last health inspection plays a crucial role in the model's predictions. Individuals who have had an inspection within the past 3 months or those whose last inspection was 2 years ago or longer seem to have a significant impact on the model's output. This finding highlights the potential predictive value of the recency of health inspections in relation to the target variable.

The importance of the "inspection" variable is further emphasized by the presence of other inspection-related features, such as "inspection_(5) (5) Longer than 1 year ago but less than 2 years ago", "inspection_(2) (2) 4 to 6 months ago", and "inspection_(4) (4) 10 to 12 months ago", which have moderate permutation importance scores. These features collectively suggest that the model is capturing patterns related to the timing of health inspections at various intervals, and this information contributes to the model's predictive performance.

Interestingly, the "inspection_(7) (7) Never" feature has a very low importance score, indicating that individuals who have never had a health inspection do not significantly influence the model's predictions. This observation could suggest that the absence of health inspection data may not be as informative as the presence of inspection data at different time points.

Apart from the "inspection" variable, the plot also highlights the importance of other features. "bmi" (Body Mass Index) emerges as the third most important feature, suggesting that an individual's BMI is a significant factor in the model's predictions. This finding aligns with the well-established relationship between BMI and various health outcomes.

Lifestyle factors, such as "sweet_drink" consumption, "walk"ing habits, "fast_food" consumption, and physical "activity", have relatively lower permutation importance compared to the inspection and BMI features. This suggests that while these lifestyle factors contribute to the model's predictions, their impact may be less pronounced than the timing of health inspections and BMI.

Demographic variables, including "BIO_SEX" (biological sex), and family history, including parental obesity ("obesity_mom" and "obesity_dad"), and parental diabetes ("diabete_dad"

and "diabete_mom"), have the lowest permutation importance scores among the given features. This indicates that these demographic factors and family history have a limited influence on the model's predictions compared to the other features in the dataset.

In summary, the permutation importance plot emphasizes the significant role of the one-hot encoded "inspection" variable in the neural network model's predictions. The timing of health inspections, particularly recent inspections within the past 3 months and those conducted 2 years ago or longer, appears to be the most influential factor. BMI is also a key feature, while lifestyle factors and demographic variables have relatively lower importance. Valuable insights are given to the factors driving the model's predictions. These findings can guide further analysis and interpretation of the results.

CHAPTER 4

Results

In this section, I present the results of the machine learning models applied to predict the presence or absence of high blood pressure. The models evaluated include Logistic Regression, Random Forests, XGBoost, Support Vector Machines (SVM), and Neural Networks. The performance of each model was assessed using the following metrics: Precision, Recall, F1-Score, and ROC AUC.

Before discussing the results, it is essential to define the evaluation metrics used: Accuracy measures the proportion of correct predictions (both true positives and true negatives) among all instances.

$$\text{Accuracy} = \frac{\text{True Positive} + \text{False Positive}}{\text{True Positive} + \text{True Negative} + \text{False Positive} + \text{False Negative}}$$

Precision measures the proportion of true positive predictions among all positive predictions.

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}}$$

Recall, also known as sensitivity, measures the proportion of true positive predictions among all actual positive instances.

$$\text{Recall} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}}$$

Model	Accuracy	ROC_AUC
Logistic Regression	0.800	0.695
Random Forests	0.753	0.634
XGBoost	0.731	0.549
SVM	0.783	0.389
Neural Networks	0.736	0.500

Table 4.1: Model Performance Comparison

The F1-score is the harmonic mean of precision and recall, providing a balanced measure of a model's performance.

$$\text{F1-Score} = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}$$

The Receiver Operating Characteristic (ROC) curve plots the true positive rate (recall) against the false positive rate at various classification thresholds. The Area Under the ROC Curve (AUC) is a summary metric that measures the model's ability to discriminate between classes. An AUC of 1 represents a perfect classifier, while an AUC of less than 0.5 indicates a random classifier.

According to table 4.1, logistic regression has the highest accuracy and highest ROC-AUC; XGBoost has the lowest accuracy and SVM has the lowest ROC-AUC. Although I had already oversample the training data, the test data still has a great imbalance in target variable. Consequently, while accuracy remains a fundamental metric for evaluating model performance, its significance is somewhat diminished in this context due to the substantial class imbalance present in the test data. This underscores the critical importance of utilizing alternative performance metrics, such as ROC-AUC (Receiver Operating Characteristic - Area Under the Curve), which are less susceptible to the effects of imbalanced datasets.

In light of this, logistic regression emerges as the most robust performer, exhibiting the highest accuracy and ROC-AUC among the models evaluated. Conversely, XGBoost demonstrates comparatively lower accuracy, while SVM exhibits the lowest ROC-AUC performance. These findings underscore the nuanced interplay between model selection, performance metrics, and the intricacies of dataset characteristics, providing valuable insights for future research and practical application.

CHAPTER 5

Conclusion

In this study, several machine learning models is used, including Logistic Regression, Decision Trees, Random Forests, XGBoost, Support Vector Machines, and Neural Networks, to investigate the influence of various factors such as sex, hereditary, habitats, and BMI on the risk of developing high blood pressure using the Add Health dataset. While the models provided valuable insights into the associations between these factors and high blood pressure, the relatively low ROC-AUC scores indicate that the models have difficulty accurately predicting an individual's likelihood of having high blood pressure based solely on the selected variables.

However, it is essential to acknowledge that the limited predictive power of the models does not necessarily imply that the chosen factors do not contribute to the development of high blood pressure. Several reasons could account for the models' suboptimal performance. Firstly, the study utilized the public-use sample from the Add Health dataset, which contains a significantly reduced number of observations compared to the full dataset. This limitation in sample size may have hindered the models' ability to capture the complex relationships between the predictors and the outcome variable.

Secondly, the etiology of high blood pressure is multifaceted and is very hard to be fully explained by a limited set of factors. Real-life situations involve almost infinite interrelated variablesmay influence an individual's risk of developing high blood pressure. These factors includes both the ones captured within the dataset and the ones present in the real world. A lot of these factors, not accounted for in the current study, could be lifestyle choices, environmental exposures, stress levels, and other medical conditions.

Despite these limitations, the study provides a foundation for understanding the complex nature of high blood pressure and highlights the importance of considering multiple factors when assessing an individual's risk. The findings underscore the need for further research using more comprehensive datasets and incorporating a broader range of variables to develop more accurate predictive models.

In conclusion, while the machine learning models in this study demonstrated limited predictive power, they have shed light on the intricate relationships between sex, hereditary, habitats, BMI, and high blood pressure. Future research should focus on integrating a wider array of variables and exploring the potential of advanced machine learning techniques to enhance the predictive capabilities of the models, ultimately facilitating early detection, prevention, and personalized management strategies for high blood pressure.

REFERENCES

[BFO84] Leo Breiman, Jerome Friedman, Richard Olshen, and Charles Stone. *Classification and Regression Trees*. Wadsworth and Brooks, 1984.

[Bis06] Christopher M. Bishop. "Pattern Recognition and Machine Learning." In *Information Science and Statistics*, pp. 209–210. Springer-Verlag, Berlin, Heidelberg, 2006. Cross-Entropy.

[CBH02] Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. "SMOTE: Synthetic Minority Over-sampling Technique." *Journal of Artificial Intelligence Research*, **16**:321–357, 2002.

[CG16] Tianqi Chen and Carlos Guestrin. "XGBoost:A Scalable Tree Boosting System." 2016.

[HAB20] Thomas C. Hinton, Zoe H. Adams, Richard P. Baker, Katrina A. Hope, Julian F.R. Paton, Emma C. Hart, and Angus K. Nightingale. "Investigation and treatment of high blood pressure in young people." *Hypertension*, **75**:16–22, Nov 2020. Published online: 18 Nov 2019.

[HU09] Kathleen Mullan Harris and J. Richard Udry. "The National Longitudinal Study of Adolescent to Adult Health (Add Health), 1994-2008.", 2009. [Accessed: 07 June 2024].

[sks] "Support Vector Machines." [Accessed: 07 June 2024].