SI618 Final Project Report: Predicting Hypertension Using NHANES Data

Group members: Yukuan Zhu, Shuo Deng

1. Introduction

This project aims to predict hypertension using datasets obtained from Kaggle's National Health and Nutrition Examination Survey (NHANES). This report outlines the methodology employed to construct a predictive model for hypertension based on demographic.csv, diet.csv, and examinations.csv datasets.

1.1 Data Sources

The data utilized in this study originates from the National Health and Nutrition Examination Survey (NHANES), available on Kaggle.(National Health and Nutrition Examination Survey (kaggle.com))

NHANES is a program conducted by the Centers for Disease Control and Prevention (CDC) in the United States. It is designed to assess the health and nutritional status of adults and children in the U.S., providing valuable insights into various health-related factors. NHANES data is widely used in public health research, epidemiology, and policy-making due to its comprehensive nature and representative sampling of the U.S. population. Recently, thousands of research findings have been published using the NHANES data, which can be accessed via NHANES - Search Results - PubMed (nih.gov).

From the Kaggle repository containing the National Health and Nutrition Examination Survey (NHANES) datasets, six files were available for analysis: demographic.csv, diet.csv, examinations.csv, labs.csv, medications.csv, questionnaire.csv.

For this project, we focused on the analysis of the following three files:

demographic.csv: Contains essential demographic details such as age, gender, race, and socioeconomic status of the survey participants. Data size: 10175 rows, 47 columns

diet.csv: Includes data regarding participants' dietary habits, nutrient intake, and food frequency. Data size: 9813 rows, 168 columns.

examinations.csv: Comprises measurements from physical examinations, including blood pressure, BMI, cholesterol levels, etc. Data size: 9813 rows, 224 columns.

The remaining three documents were not used for the following reasons.

labs.csv, medications.csv: These files contain substantial data directly related to blood pressure, for example antihypertensive drug use, which might influence our predictive model. Hence, to maintain model independence and accuracy, these files were excluded from the analysis.

questionnaire.csv: There exist a considerable number of missing values in this file, which significantly impacted its utility for analysis. Due to poor data quality and limited usability, the questionnaire data file was not considered for this study.

1.2 Motivations

Initially, the project aimed to predict diabetes using machine learning techniques. However, the diagnosis criteria for diabetes exhibit significant variations, lacking a standardized blood glucose threshold for diagnosis. This variation in diagnostic criteria posed a challenge, as it hindered the establishment of a unified benchmark for diabetes diagnosis based on the available datasets.

Reflecting on homework we've been doing this semester, we found that in a prior assignment, blood pressure was categorized into multiple levels. Hypertension also stands as a common and significant health concern among individuals, making it an essential target for predictive modeling.

Hence, the project's motivation shifted towards leveraging the available NHANES datasets—specifically demographic, dietary, and examination data—to construct a classification model for predicting hypertension. By exploring the relationship between various health parameters and the incidence of hypertension, the objective became to develop a robust predictive model capable of identifying individuals at risk of high blood pressure based on diverse health factors.
The decision to focus on hypertension prediction was rooted in its prevalence, significant health implications, and the availability of more consistent diagnostic criteria compared to diabetes. Through this shift in focus, the project aims to contribute to the development of effective preventive healthcare measures and interventions for hypertension.

2. Data Processing

2.1 Blood Pressure Data Preprocessing

In our project, we accessed blood pressure data from the examination.csv file, encompassing measurements for both systolic (high) and diastolic (low) blood pressure.
To derive more reliable blood pressure indicators, we first identified and removed values falling outside the physiological ranges expected for blood pressure readings. Specifically, diastolic values below 30 or above 180, as well as systolic values below 50 or above 250, were considered outliers and subsequently excluded from the dataset.
After that, we proceeded to compute averaged blood pressure values for each individual. This involved calculating the mean of all non-Nan blood pressure measurements for every participant. For example, if a person has blood pressure data that meets the criteria on three out of four occasions, we add up their blood pressure and divide by 3.
As a result of these preprocessing steps, we obtained a more robust and trustworthy blood pressure dataset. This refined dataset will serve as a solid foundation for subsequent modeling and analysis endeavors, ensuring greater accuracy and reliability in our predictive models for hypertension.
After that, considering our goal of establishing a classification model, we aimed to categorize individuals into specific stages of hypertension based on the obtained blood pressure indicators. Adhering to the criteria outlined in a previous assignment, we categorized the blood pressure readings into three stages:
Normal Range: Diastolic pressure below 80 and systolic pressure below 130 indicated normal blood pressure.
Stage 1 Hypertension: Diastolic pressure ranging from 80 to less than 90, coupled with systolic pressure below 140, or diastolic pressure below 90 combined with systolic pressure ranging from 130 to less than 140.
Stage 2 Hypertension: Diastolic pressure equal to or above 90, or systolic pressure equal to or above 140.
This stratification into distinct stages aligns with established medical classifications and will serve as the cornerstone for our classification model.

2.2 Feature data preprocessing

In the preprocessing phase of the feature data, several steps were undertaken to ensure data accuracy and utility. Initially, similar to the blood pressure data treatment, repetitive measurements within the feature data columns were addressed. These columns underwent a transformation where their multiple measurements were dropped and replaced by their mean values. Subsequently, columns with excessive missing values were removed from consideration.

The decision to eliminate such columns stemmed from the necessity to avoid an overwhelming amount of missing values that could significantly impact the efficacy of our predictive models. Following this, the data was categorized into two types: categorical and numerical. Due to a majority of the dataset's object types being floats, directly identifying whether a column was categorical or numerical via the info.Dtype function was unfeasible. Hence, a custom algorithm was devised; columns with a unique count (nunique) exceeding 10 were categorized as numerical, while those below this threshold were classified as categorical.

3. Feature selection and EDA

3.1 Feature selection

During the feature selection phase, we tried to identify and retain columns significantly associated with blood pressure, thereby enhancing the model's predictive capability. The process began with an initial review of all column descriptions across the dataset to eliminate irrelevant columns, such as 'SEQN', which is the participants' ID.

Subsequently, the processed blood pressure data was merged with all three tables—demographic, dietary, and examinations datasets. Upon merging, correlations between all columns and both low and high blood pressure were computed.

After that, features are selected by a criterion: any column demonstrating a correlation coefficient exceeding 0.05 with either low or high blood pressure was considered influential and selected as a potential feature for our predictive models. This criterion was chosen to ensure the inclusion of features demonstrating a reasonable degree of correlation with blood pressure, thereby enhancing the predictive power of the model. Totally, 70 columns are selected as our features.

3.2 EDA and visualization of the selected features

During the exploratory phase of our experiment, we initiated our analysis by conducting histogram plot visualizations for the systolic and diastolic blood pressure—two crucial indicators we aim to predict. This allowed us to assess their distribution characteristics and ascertain if they align with a normal distribution.

Subsequently, employing the describe() function, we obtained an overview of the numerical features within the dataset. This summary statistics method enabled us to identify potential instances of skewness (notably, substantial disparities between median and mean values) or the presence of extreme outliers. For these identified features, we further conducted visualizations to gain deeper insights into their distribution.

Below are some exemplary plots depicting these investigations:

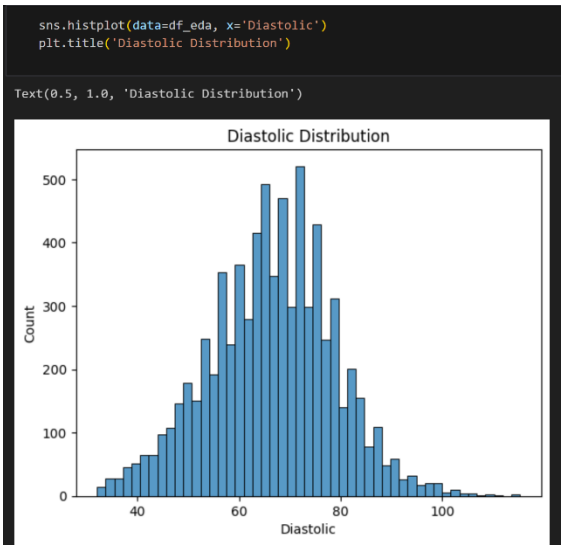

Figure 1. Histplot of Diastolic blood pressure

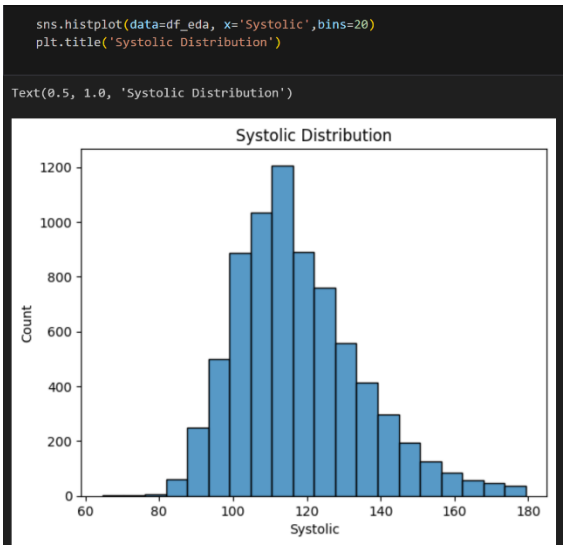

Figure 2. Histplot of Systolic blood pressure

From Figure 1 and Figure 2, it can be found that both the diastolic blood pressure and the systolic blood pressure generally follow a normal distribution, which meets our expectations.

For the features columns, after detecting the skewness on some columns by looking at the overview of them, we do histogram plots on those columns. For columns displaying left-skewed distributions, we utilized a square transformation approach. Conversely, for columns exhibiting right-skewed distributions, a logarithmic transformation was applied. For instance, as is shown in Figure 3, the histogram plot for water intake demonstrates a notable right-skewed distribution. As a remedial measure, a logarithmic transformation was applied to this specific column. Post-transformation, the data for water intake exhibited a distribution that more closely approximated a normal distribution. For other columns displaying similar skewness characteristics, the same approach of logarithmic transformation was applied. (To check details, please check the code for every transformation).

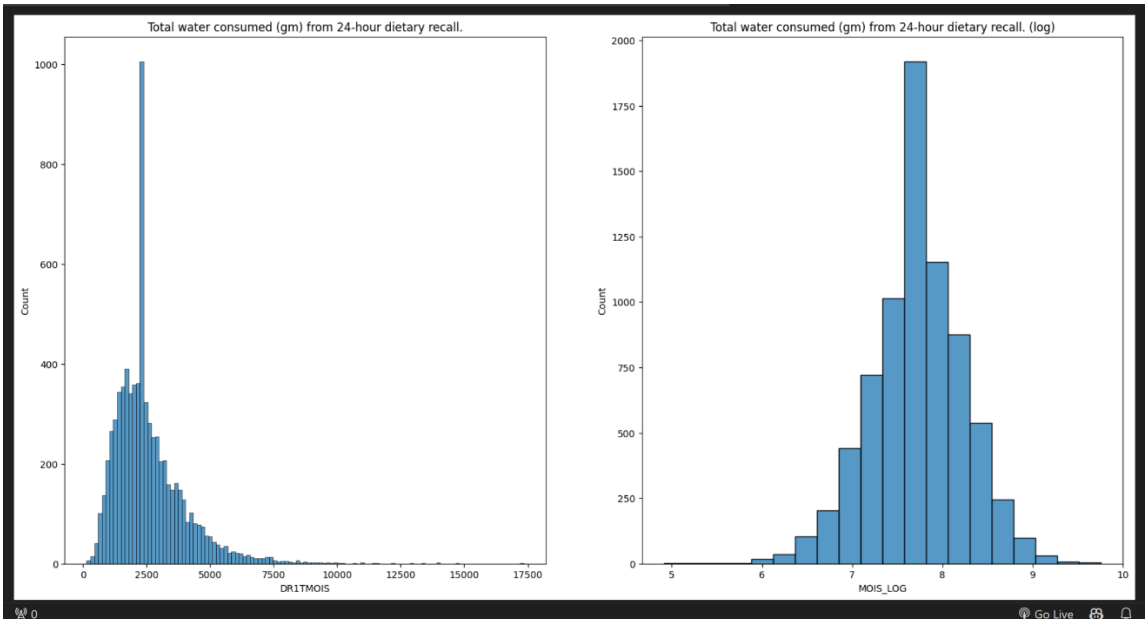

Figure 3. Histplot of water intake

A noteworthy aspect within our dataset pertains to the column representing alcohol intake. As depicted in Figure 4, the data in the 'DR1TALCO' column (which indicates the alcohol intake) exhibits a distinct right-skewed distribution. However, conventional logarithmic transformations were found ineffective in rectifying this skewness.

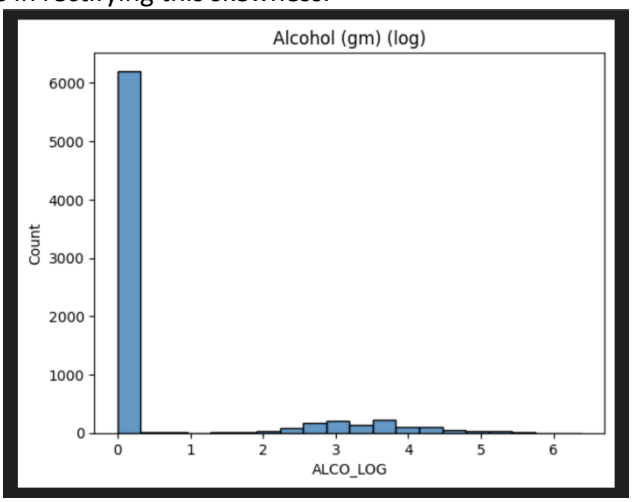

Figure 4. Histplot of alcohol intake after logarithmic transformation

Upon further investigation, illustrated in Figure 5, it was observed that a significant proportion of participants reported zero alcohol intake. This prevalence of zero values led to the ineffectiveness of the logarithmic transformation.

```
df_eda['DRINKER'] = np.where(df_eda['DR1TALCO']>0, 1, 0)
df_eda['DRINKER'].value_counts()
```

✓ 0.0s Python

```
DRINKER
0    6196
1    1221
Name: count, dtype: int64
```

Figure 5. value counts of the DR1TALCO column

To address this issue and ensure improved integration of this column into our predictive models, a decision was made to transform it into a categorical variable. Specifically, individuals who do not consume alcohol were categorized as '0', while those who do were categorized as '1'. This categorical representation allows us to capture the essence of alcohol intake in a manner that aligns more effectively with the modeling process.

3.3 Evaluation of Feature Interrelationships

To evaluate the interrelationships among features, a heatmap (Figure 6) was constructed to visualize the correlations between all features. This heatmap provided an overview of the correlation matrix, aiding in identifying potential instances of high correlation between variables. Upon examination, it was discerned that certain feature pairs exhibited notably high levels of correlation. Such high correlations suggested a potential redundancy between these variables, signaling the need to drop one of the correlated features to enhance model performance and mitigate multicollinearity issues.

As an example, a high correlation was observed between 'WTMEC2YR_log' (which indicates weight at the MEC exam) and 'WTINT2YR_log'(which indicates weight at the home exam). To further investigate this correlation, as is shown in Figure 7, a scatterplot was generated, followed by a linear regression analysis between these two parameters.

With a predefined significance of 0.05, the p-values obtained for both slope and intercept were found to be less than the alpha value. (Figure 8) This observation suggested a potential complete linear relationship between these two features. Consequently, it is advisable to eliminate one of the columns to prevent redundant information in the dataset. In addition to this example, similar checks for excessive correlation were conducted among various other feature pairs. Several columns displaying a too-high correlation with other features were consequently removed. Further details and comprehensive insights regarding this process can be found in the code file.

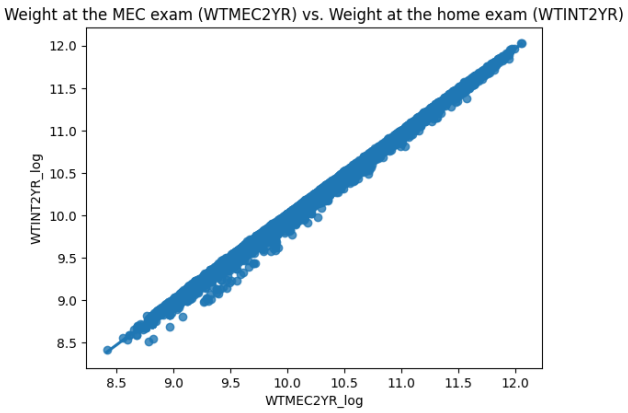

Figure 7. Scatterplot between WTMEC2YR and WTINT2YR

OLS Regression Results

Dep. Variable:	Q('WTMEC2YR_log')	**R-squared:**	1.000
Model:	OLS	**Adj. R-squared:**	1.000
Method:	Least Squares	**F-statistic:**	1.377e+32
Date:	Mon, 04 Dec 2023	**Prob (F-statistic):**	0.00
Time:	11:59:46	**Log-Likelihood:**	2.3319e+05
No. Observations:	7417	**AIC:**	-4.664e+05
Df Residuals:	7415	**BIC:**	-4.664e+05
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	1.141e-14	8.72e-16	13.079	0.000	9.7e-15	1.31e-14
Q('WTMEC2YR_log')	1.0000	8.52e-17	1.17e+16	0.000	1.000	1.000

Omnibus:	1547.785	**Durbin-Watson:**	0.110
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	308.122
Skew:	0.089	**Prob(JB):**	1.24e-67
Kurtosis:	2.018	**Cond. No.**	145.

Figure 8. Summary of the Regression Model

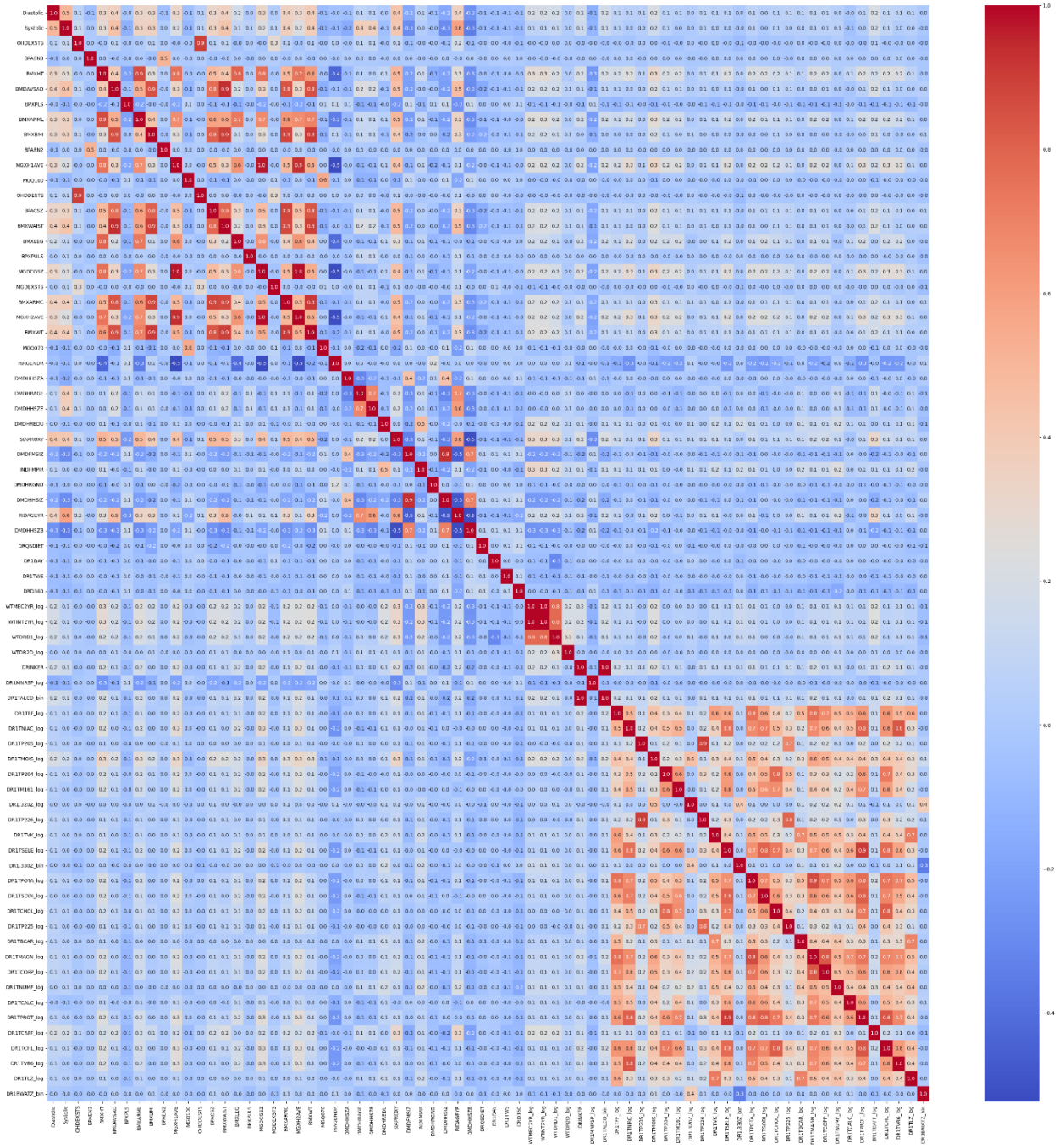

Figure6. Heatmap of all feature columns

3.4 Validation of Feature Relevance to Blood Pressure

While retaining features with correlations above 0.05 with blood pressure during the feature selection process, there remained a need to ascertain the definite correlation of certain features with blood pressure. To address this, we conducted Chi-Squared and ANOVA tests on these specific features to confirm their direct association with blood pressure.

These tests were pivotal in determining whether these features exhibited significant differences across various categories, such as distinct ranges of blood pressure levels. The Chi-square test focused on assessing relationships among categorical variables, while the ANOVA test was used to evaluate differences among continuous variables within different blood pressure categories.

For instance, within our feature set, one column named 'SIAPROXY' denotes whether a proxy respondent was utilized during the Sample Person (SP) interview. Although this column exhibits a correlation with blood pressure exceeding the 0.05 threshold, concerns arose regarding its definitive relationship with blood pressure based on its description. Therefore, we subjected it to a chi-square test to validate its association. As is shown in Figure 9, setting the significance level at 0.05, the chi-square test yielded a p-value significantly lower than the threshold. This outcome indicates a robust correlation between 'SIAPROXY' and blood pressure, affirming the substantive relevance of this feature in our predictive modeling endeavors. Hence, it is imperative to include this feature in our modeling process, given its established association with blood pressure.

Another example for the ANOVA test. To validate the relationship between the user's weight ('BMXWT') and blood pressure, we initiated our analysis by plotting its distribution using kernel density estimation (KDE) and boxplots concerning different blood pressure stages. As is shown in Figure 10 and Figure 11, this exploratory visualization revealed distinct variations in the distribution of 'BMXWT' across various blood pressure categories.

After this visual inspection, an ANOVA test was conducted. Setting the significance level at 0.05, as is shown in Figure 12, the resultant p-value from the ANOVA test was notably lower than 0.05, signifying a robust and statistically significant correlation between 'BMXWT' and blood pressure.

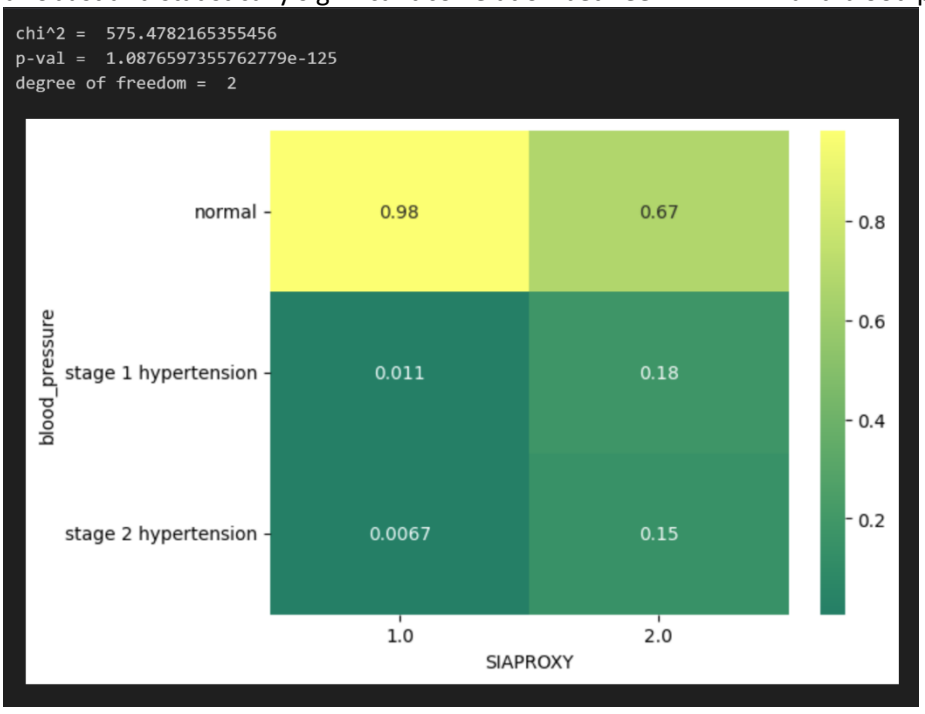

Figure 9. The Chi-Square test result for 'SIAPROXY'

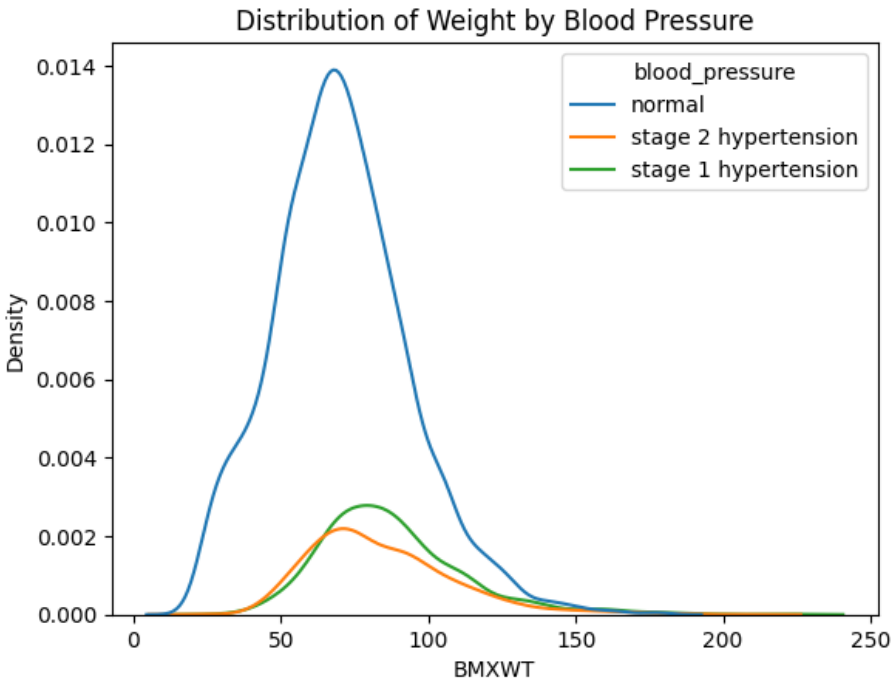

Figure 10. The kde plot of BMXWT

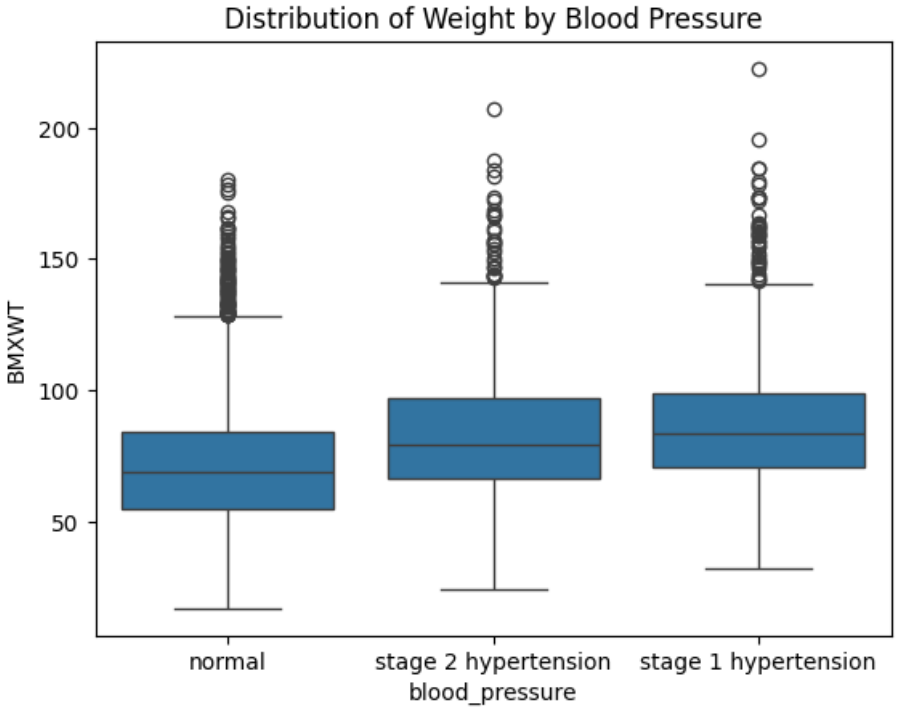

Figure 11. The boxplot of BMXWT

	sum_sq	df	F	PR(>F)
blood_pressure	3.497763e+05	2.0	312.258464	6.276902e-131
Residual	4.152396e+06	7414.0	NaN	NaN

Multiple Comparison of Means - Tukey HSD, FWER=0.05

group1	group2	meandiff	p-adj	lower	upper	reject
normal	stage 1 hypertension	16.9467	0.0	15.0956	18.7978	True
normal	stage 2 hypertension	13.6628	0.0	11.6486	15.677	True
stage 1 hypertension	stage 2 hypertension	-3.2838	0.0064	-5.805	-0.7627	True

Figure 12. The ANOVA test result of BMXWT

Similar assessments were conducted for other columns within our dataset. For comprehensive details regarding these analyses, please refer to the code file.

4. Modeling

4.1 Preprocessing

During the model construction phase, our initial step involved the preprocessing of data features. Categorical features underwent one-hot encoding to transform them into a suitable format for analysis, while numerical features were standardized to ensure consistency in scale across the dataset.

Subsequently, given the extensive number of columns initially considered for predictive modeling, we employed the SelectKBest method using f_classif with a criterion to retain the top 40 features. This selection approach aimed to minimize the reduction in variability while effectively reducing the feature set to 40, optimizing the model's efficiency and computational performance.

4.2 Classifier Selection and Parameter Optimization

During the model-building phase, we evaluated three different classifiers: RandomForest, HistGradientBoosting, and MLP. These classifiers were chosen based on their potential to handle our dataset's characteristics effectively.

Subsequently, to optimize these classifiers' performance concerning our dataset, we employed GridSearchCV. This technique allowed us to conduct an exhaustive search over specified parameter values for each classifier, aiming to identify the optimal hyperparameters tailored to our dataset's nuances. For comprehensive details regarding the best parameters discovered for each classifier, please refer to the code file.

Following the parameter tuning process, we ensembled these three optimized classifiers using a VotingClassifier. This ensemble approach enabled us to combine the predictive strengths of individual classifiers, leveraging their diverse algorithms and learning strategies to generate our final, optimized classifier.

4.3 Model training and Cross Validation

Our classifier yielded a cross-validation score exceeding 0.83, indicating a satisfactory performance. This result affirms our confidence in the model's ability to make accurate predictions across diverse datasets.

Additionally, to further evaluate the classifier's effectiveness, we generated a confusion matrix. This visualization allowed us to comprehensively assess the model's predictive capabilities by illustrating the true positive, true negative, false positive, and false negative predictions across different hypertension stages.

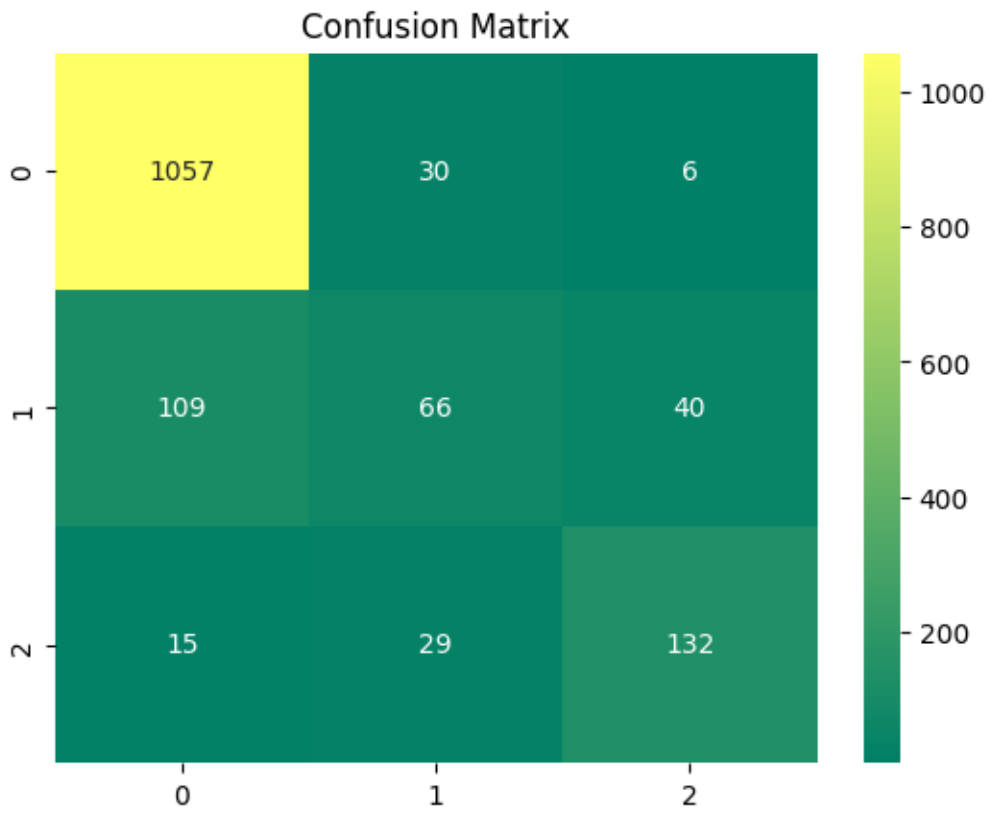

Figure 13. Confusion Matrix

5. Conclusion

In this project, we extensively utilized Kaggle's National Health and Nutrition Examination Survey (NHANES) dataset to acquire a wealth of data. Commencing with meticulous data cleansing, we proceeded with thorough Exploratory Data Analysis (EDA) to identify pertinent features suitable for predicting hypertension stages. Subsequently, many columns underwent preprocessing to enhance their predictive capacity. Ultimately, leveraging machine learning techniques, we developed a model capable of predicting normal, stage 1, and stage 2 hypertension stages, achieving a cross-validation value of 0.83.

6. Distribution Statement

In this project, Yukuan Zhu and Shuo Deng worked together on everything from data collection, data cleaning, EDA and report writing. The workload of both of them is almost 50% of the project. Shuo Deng puts more effort on model parameter optimization, while Yukuan Zhu takes on more report writing tasks.