Abigail Condliffe

Professor Koehler

Data Bootcamp

13 December, 2024

Data Bootcamp Final - Write Up Version

# Predicting High Blood Pressure Patients Among US Adults

## Introduction to Problem & Data

### Problem Statement:

High Blood Pressure is a chronic condition, referred to as "the silent killer" because it is often undetected, unless tested (Columbia Doctors). High Blood Pressure can develop over time, especially from an inactive lifestyle: diabetes and obesity increase one's risk of High Blood Pressure (CDC). Nearly half of the US adult population suffers from High Blood Pressure, also known as hypertension (HHS). This diagnosis is particularly of high concern because it greatly elevates one's risk of a heart attack or stroke. In extreme causes, it can lead to heart failure.

It is important to predict individuals who are at higher risk of hypertension because of these related serious health risks. This is why in this project I chose to create a predictive model - because of this disease's elusive nature it is even more important to better anticipate High Blood Pressure individuals. An effective predictive model could allow both individuals and health care providers to anticipate a High Blood Pressure diagnosis through various characteristics, even if the individual has not directly sought out blood pressure testing. Even businesses/pharmaceutical companies might be able to better market treatment and remedies to individuals with greater likelihoods of the diagnosis. I will use characteristics such as age, BMI, and gender, to predict a High Blood Pressure diagnosis. I also plan to expand the independent variables to include more demographic characteristics, such as US region of residence and race, which could relate to a positive diagnosis in terms of one's lifestyle.

https://www.cdc.gov/high-blood-pressure/about/index.html

https://www.ncbi.nlm.nih.gov/books/NBK279239/

https://millionhearts.hhs.gov/data-reports/hypertension-prevalence.html

https://www.columbiadoctors.org/news/why-high-blood-pressure-known-silent-killer#:~:text=Medical%20professionals%20call%20high%20blood,determine%20if%20someone%20has%20it.

https://pmc.ncbi.nlm.nih.gov/articles/PMC2643192/

## Dataset Description:

The data is sourced from AHRQ's (the Agency for Healthcare Research and Quality) Medical Expenditure Panel Survey (or MEPS). It is a collection of data/surveys from individuals, families, health care providers, and employers. The data is collected from all across the country and my particular dataset, from 2022, has 20,432 data entries.

A challenge I anticipate is which features to focus on, as the surveys include 1,419 characteristic columns. Thus, I have to use some background knowledge on High Blood Pressure causes to narrow my scope of factors down.

I must also adjust certain variable responses, as some include data unnecessary for the scope of my project (e.g. "Don't Know," "Refused, "Inapplicable").

https://meps.ahrq.gov/mepsweb/data_stats/download_data_files_codebook.jsp?PUFId=H243

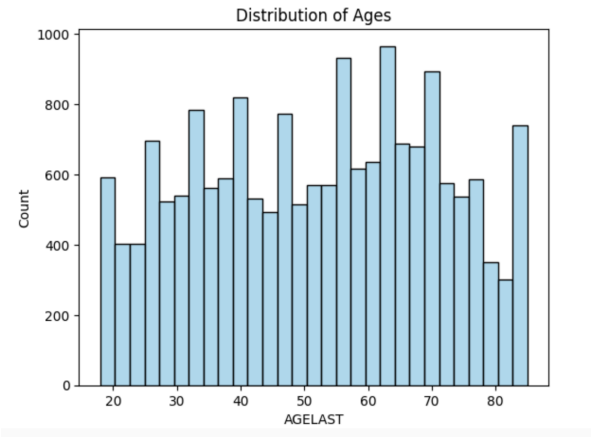
## Data Pre-Processing & Preliminary Examination:

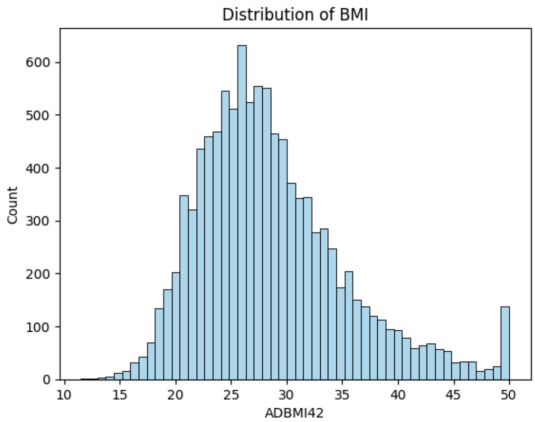
In this stage of pre-processing, I have also filtered the data frame to just include U.S. Adults, of 18 years or older.

## Exploratory Data Analysis:

## Descriptive Statistics:

In this dataset, there are 6,788 individuals (>17) who responded "yes" to being diagnosed with High Blood Pressure. These positive, or "yes", responses equate to roughly 38% of those who responded to the blood pressure survey. There is a count of 11,082 who responded "no," which covers the remaining 62% of the relevant responders.

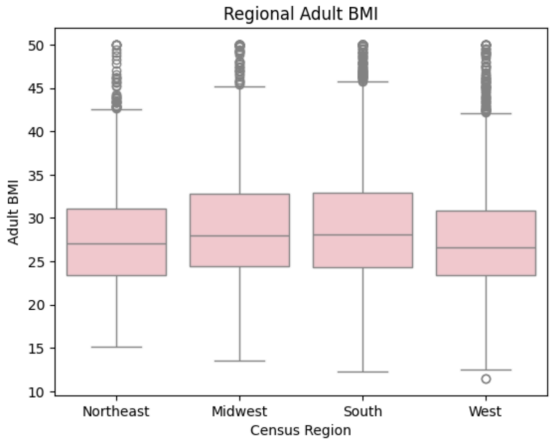


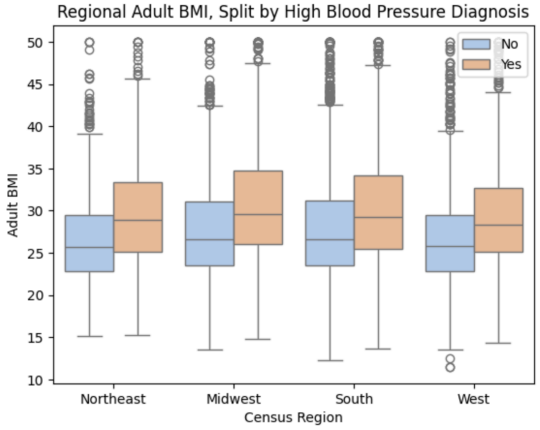
The descriptive statistics show that the individuals considered in this project will be, on average, older as the distribution of ages is slightly skewed left. The oldest age in the data is 85. As for BMIs, the distribution skews to the right. The lowest BMI in the data is 10 while the highest is 50. Some important background on BMIs is that "Normal Weight" individuals score between 18.5 and 24.9 (NIH). Considering that this distribution skews right, this indicates that many in our dataset are Overweight or Obese.

https://www.nhlbi.nih.gov/health/educational/lose_wt/BMI/bmicalc.htm

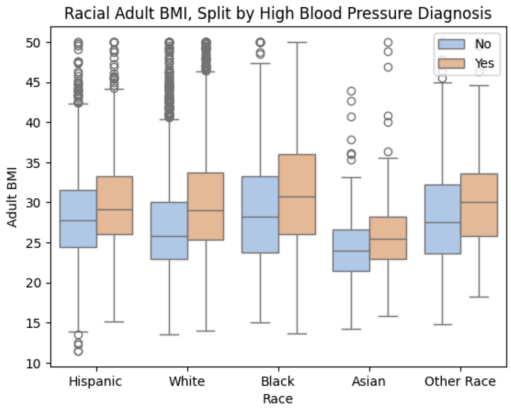
## Initial Visualizations




This box plot above demonstrates that individuals located in the Midwest or South may be more likely to suffer from a higher BMI score (higher median and upper quartile range). This could be for a plethora of reasons, such as lifestyle or access to healthcare. These regions should be further considered, as a high BMI is related to High Blood Pressure.

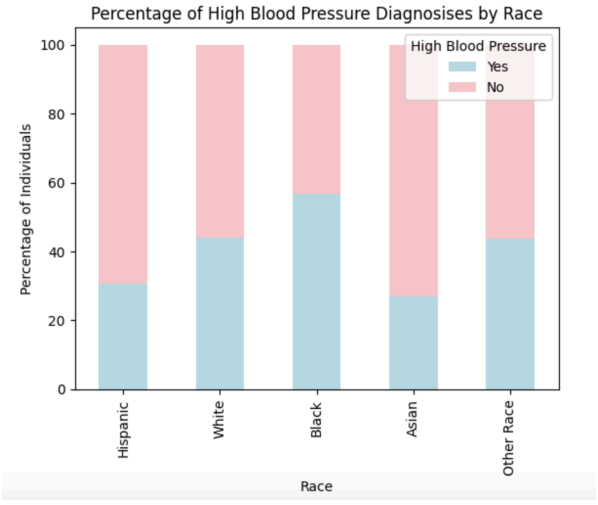



This boxplot above supports my earlier argument, that those diagnosed with High Blood Pressure tend to have higher BMIs -- in this visualization, the orange boxplots, indicating a High Blood Pressure diagnosis, tend to have a higher range and median in terms of BMI.

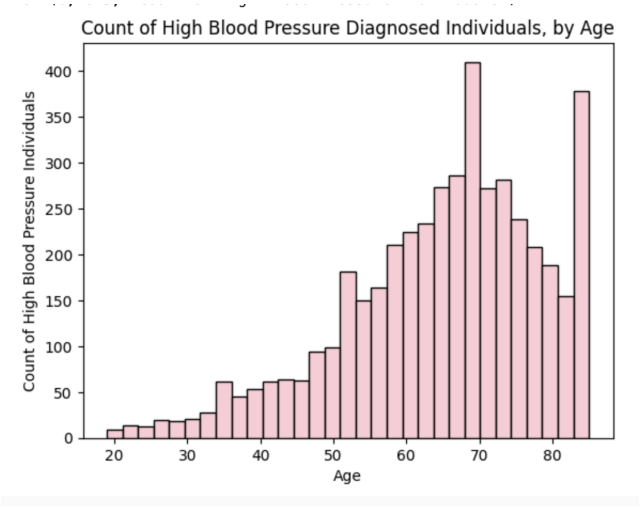



In the box plot above, across all races the trend of higher BMI correlating with a higher blood pressure diagnosis is maintained. Furthermore, this boxplot visualization shows that Black individuals tend to have higher BMIs than the other races (they have a higher median and upper quartile range).

| HIBPDX | 1 | 2 |
|---|---|---|
| RACETHX | | |
| 1 | 30.778515 | 69.221485 |
| 2 | 43.938040 | 56.061960 |
| 3 | 56.657019 | 43.342981 |
| 4 | 26.978417 | 73.021583 |
| 5 | 43.727599 | 56.272401 |

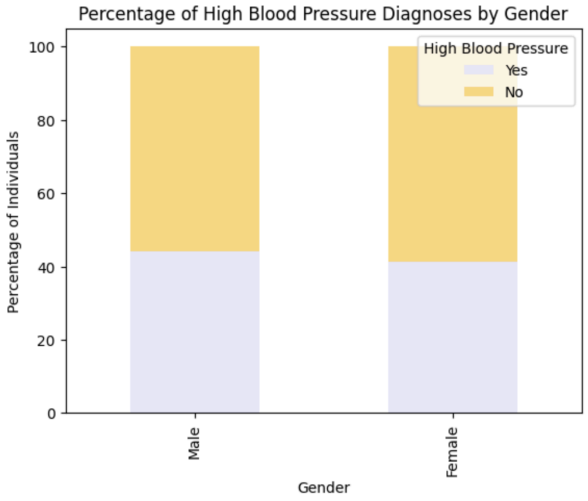



As seen in the bar plot above, individuals identifying as Black have the highest prevalence of blood pressure. This is a further elaboration on my earlier visualization, where Black people, on average, have the highest BMI than other race groups in the dataset.




When reviewing the bar chart above, there is a clear positive relationship between age and High Blood Pressure diagnoses (in terms of count).

| HIBPDX | 1 | 2 |
|---|---|---|
| SEX | | |
| 1 | 44.246862 | 55.753138 |
| 2 | 41.322314 | 58.677686 |




While these two bar splits look comparable, the table above indicates that Male individuals have a greater percentage with High Blood Pressure.

## Modeling & Interpretations

In my predictive models, I will be prioritizing specificity, calculated as $\frac{TP}{(TP+FN)}$, where I will be aiming to minimize False Negatives and maximize True Positives. Therefore, specificity

scores that are higher are the goal. I am aiming to avoid False Negatives as they mean leaving those with High Blood Pressure undiagnosed, and therefore, untreated. I will also be measuring accuracy through '.score,' as it tells me what percentage of the predictions accurately capture the True Positives and the True Negatives. Another note for my models is that I will be using a Train-Test Split, where I will be splitting the data into 75% training and the remaining 25% as the testing set.
(Specificity, as defined in Chapter 7, *Decision Analytic Thinking I: What is a Good Model?*)

## Baseline Model

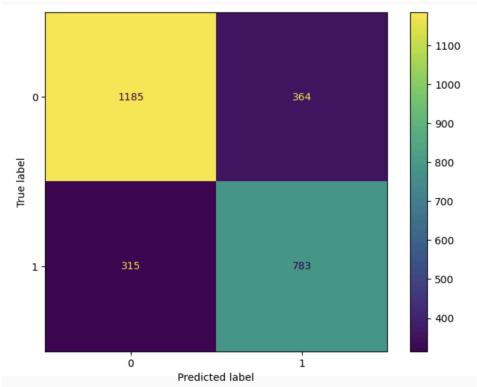
To see how much my different regression models improve upon a baseline, I will calculate the average High Blood Pressure score. Essentially I am using a raw probability as my baseline.

This baseline shows that I should expect 57.4% of individuals in the data to *not* be diagnosed with High Blood Pressure. That being said, this baseline also shows that the baseline is only correct 57.4% of the time. So any models that result in a higher percentage accuracy than this have improved upon the baseline.

## Multiple Regression Model

When choosing between linear and logistic regression as my model, I decided on logistic regression. Logistic regression is better suited for the binary nature of my problem (whether the individuals do or do not have High Blood Pressure).

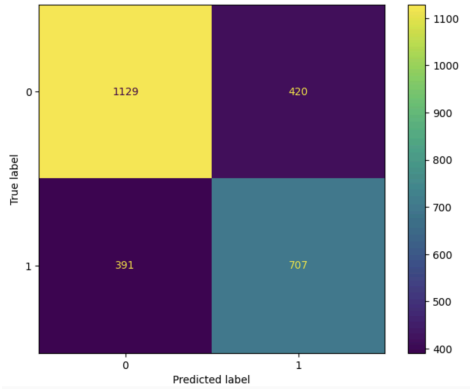
This logistic regression has a much improved score on the baseline, with 73% accuracy for the training data and 74% accuracy with the test.




Upon reviewing the Confusion Matrix, and corresponding specificity score, it shows that logistic regression is relatively good at predicting High Blood Pressure patients (71.3% specificity score). More specifically, in this case there are only 315 False Negatives. But to know if this is actually the best model, I will compare across other models which also fit a Confusion Matrix display to see if False Negatives can be further minimized.

## K-Nearest Neighbors Model

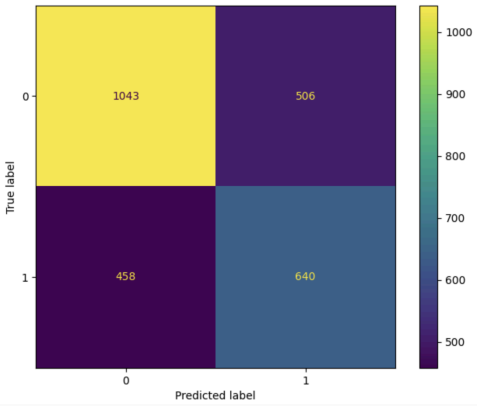
This K-Nearest Neighbors Model did improve on the baseline model, with 79.2% accuracy on the training data and 69.4% on the testing. However, the KNN testing data's accuracy did not improve upon the logistic regression model's score.




Here, using K-Nearest Neighbors, it seems the model has a Specificity Score of 64.4%. It seems that Logistic Regression is more specific with a score of 71.3%. So far, Logistic Regression seems to be superior in terms of the testing data.

## Decision Tree Model

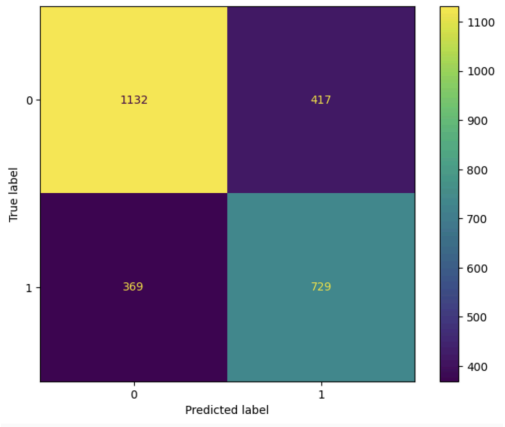
The testing data had an accuracy score of 63.4%, which is better than the baseline but worse than both the Logistic Regression and K-Nearest Neighbors models.




The Decision Tree Model has a specificity score of 58.3%, using the Confusion Matrix results from above. This is worse than both the Logistic Regression and K-Nearest Neighbors' specificity scores of 71.3% and 64.4%, accordingly.

## Random Forest Model

The Random Forest Model does improve upon the Decision Tree Model and K-Nearest Neighbors' scores, in terms of testing data (the Random Forest Model saw an accuracy of 71.6% while the Decision Tree Model saw 63.4% and KNN saw 69.4%). That being said, the Logistic Regression Model still remains the most accurate with the highest score of 74.3%.




The Random Forest model sees a specificity score of 66.4%. While this does improve upon the Decision Tree Model and K-Nearest Neighbor models' specificity scores, it remains lower than the Logistic Regression Model.

Considering that my Logistic Regression Model seems to fare the best, in terms of accuracy and specificity, I want to determine which features this model sees as 'most important.'

| | Feature | Coefficient | Absolute Value Coefficient |
|---|---|---|---|
| 4 | onehotencoder__RACETHX_3 | 0.829953 | 0.829953 |
| 6 | onehotencoder__RACETHX_5 | 0.324441 | 0.324441 |
| 7 | onehotencoder__SEX_2 | -0.270744 | 0.270744 |
| 1 | onehotencoder__REGION22_3 | 0.208415 | 0.208415 |
| 5 | onehotencoder__RACETHX_4 | 0.115717 | 0.115717 |
| 9 | remainder__ADBMI42 | 0.091730 | 0.091730 |
| 8 | remainder__AGELAST | 0.073971 | 0.073971 |
| 2 | onehotencoder__REGION22_4 | -0.068492 | 0.068492 |
| 0 | onehotencoder__REGION22_2 | 0.062100 | 0.062100 |
| 3 | onehotencoder__RACETHX_2 | 0.011954 | 0.011954 |

According to the coefficients, Race is a very important feature, with identifying as "Black," "Other Race," or "Asian" being in the top 5 features ranked. Being male also seems to play a significant role in terms of features, as well as residing in the South. Finally, while it did

not make the top 5, BMI is another important factor in influencing one's likelihood to have High Blood Pressure.

## Next Steps & Discussion

### Summary of Findings

All of my chosen models – Logistic Regression, K-Nearest Neighbors, Decision Tree, Random Forest – improved upon my Baseline Accuracy score. In terms of the best model, I found that the Logistic Regression Model was the best at predictions: it resulted in the highest accuracy score as well as the highest score of specificity, by maximizing True Positives and, through this, minimizing False Negatives.

Key Findings:

1. Logistic Regression Model: Proving that Logistic Regression is the best model, indicates that the data is quite linearly separable. This makes sense seeing that most of the features I considered are binary, with the exception of age and BMI.
2. Influence of BMI: In my early visualizations, it did become apparent that across all races and regions, those diagnosed with High Blood Pressure tended to have higher BMI medians and upper quartile ranges (shown in the box plots). This relationship was further confirmed when I did an assessment of my Logistic Regression's feature importance – BMI was ranked #6 in terms of feature importance.
3. Influence of Racial Identity: Three of the top five important features came from the RACETHX feature. For my own research, seeing how racial identity in the surveys most strongly affected High Blood Pressure diagnoses was an unexpected result. I anticipated age or BMI would rank higher, as these seem more closely related to health issues. So a key finding in this result is the influence of lifestyle or perhaps health care access, related to one's race.

In summary, I found that the Logistic Regression Model was most effective at capturing and predicting the linearly separable nature of High Blood Pressure diagnoses. Features that most

strongly affected the predictions in this model were one's race, sex, and region of residence. These particular features also appear to relate to one's lifestyle — for example, one's race could affect cultural food consumed or perspective on the healthcare system. In terms of region, cuisines vary across the U.S., as well as exercise habits. These top features were closely followed by BMI, which I had anticipated to be a close relationship early on in my research (shown in the visualization portion of this project). In terms of takeaways for health care providers or individuals, this Logistic Regression model and its conclusions indicate there should be a close monitoring of those likely to have High Blood Pressure issues: older, Black, higher BMI, and residing in the South.

## Next Steps/Improvements

If I had more time or further available data, I think it would be interesting to look at the following to increase the depth of my High Blood Pressure analysis:

- High Blood Pressure rates over time:
    - I would love to explore how rates of hypertension in the U.S. have changed over time – have they increased as nutrition and food ingredient qualities have declined? Or has health care technology improved, thus lowering the rate?
- High Blood Pressure costs:
    - I would also be curious to explore more financial and/or insurance data to understand the financial burden of the disease. If it is left undetected and therefore untreated, how does the High Blood Pressure treatment change in terms of cost and intensity?
- Lifestyle by Region, Race, and Gender:
    - After understanding that these are the three most 'important' features used in the Logistic Regression Model (the features with the greatest coefficients), I would want to explore if my assumptions about lifestyle are reflected in the data. For example, I assume that the reason those in the South are generally more prone to the High Blood Pressure diagnosis is due to the fact that cuisine in the South tends

to be heavier. If there was data to measure such lifestyle factors to better attribute the cause, that would further support my arguments.

- Assessing Feature Importance:
    - One thing I considered in my data is adjusting the BMI and age features to be dummy variables, as they differ from the other features in their continuous nature. In practice, BMI could have been 0 - not obese, and 1 - obese; age as 0 - 50 or younger, 1 - 50 or older. That being said, I thought these cutoffs were too blunt and I wanted to keep the data more granular. But that could be a 'next step' to try in my data if I wanted to continue different types of assessments. This adjustment might also affect these features' rank of importance; by keeping these features as continuous, their wider range might dilute their importance in the model.

By exploring these 'next steps,' I would better be able to highlight the gravity of High Blood Pressure diagnoses. For example, if the rates are increasing over time it is even more important to identify and test individuals at greater risk of Higher Blood Pressure. Or if costs incurred are heavy burdens then perhaps these individuals need more financial support. In terms of understanding feature importance, this could better attribute the direct causes for increasing rates of High Blood Pressure, and thus, provide actionable feedback to individuals trying to prevent higher blood pressure rates.