



THE UNIVERSITY OF NAIROBI  
FACULTY OF SCIENCE AND TECHNOLOGY  
SCHOOL OF MATHEMATICS

STA 420: PROJECT IN STATISTICS

**ASSESSMENT OF RISK FACTORS FOR PRE-HYPERTENSION AND  
HYPERTENSION  
A MULTINOMIAL LOGIT MODEL  
BY:**

NYAMBOK ALLAN JOE: I63/2405/2020

MUTUKU BRIAN UHURU: I63/4593/2020

MUSAU VERONICA NZILANI: I63/4576/2020

GICHOHI BENSON KARANJA: I63/4612/2020

GATHURI VANESSA WAMBUI: I63/139914/2020

**SUPERVISOR**

DR. ANN WANJIRU WANG'OMBE

This research project is submitted to the University of Nairobi in partial fulfillment of the requirements for the award of a Bachelor degree in Statistics.

© 2024

# DECLARATION

We solemnly declare that this project is our own original work and has not been submitted to any university either in parts or in its entirety for any purpose of examination. All the sources of information and materials used in this work has been dully acknowledged. We also acknowledge that this project is being submitted for assessment and that it may be made available for use and reference by other students.

**Registration Number    Name**

**Signature**

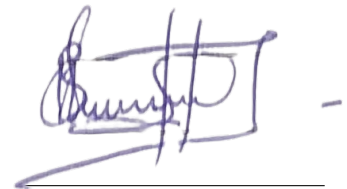
I63/139914/2020

GATHURI VANESSA WAMBUI



I63/4612/2020

GICHOHI BENSON KARANJA



I63/2405/2020

NYAMBOK ALLAN JOE



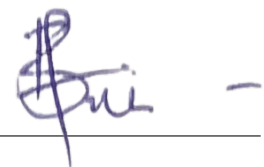
I63/4576/2020

MUSAU VERONICA NZILANI



I63/4593/2020

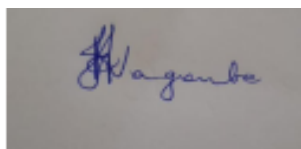
MUTUKU BRIAN UHURU



This research project has been submitted for examination with my approval as the University project supervisor.

DR. ANN WANG'OMBE

SIGNATURE:

A rectangular box containing a handwritten signature in blue ink. The signature appears to be 'A. Wang'ombe' written in a cursive style.

LECTURER, SCHOOL OF MATHEMATICS.  
UNIVERSITY OF NAIROBI.

# Acknowledgement

We want to thank a number of people for their help and valuable advice for the successful completion of our project. First, we send our special thanks to Dr. Ann Wanjiru Wang'ombe for her advice, encouragement, and constructive remarks while conducting this research. Without her knowledge and hard work, this project would not have been possible, as would the scholarly credibility of this work. We would also like to thank the faculty and staff in the School of Mathematics at the University of Nairobi for providing us with the necessary tools to finish this research. This is specifically dedicated to our teachers and tutors who have guided us in order to be on the right track, inculcating discipline and knowledge that we would apply in our educational endeavors. We also extend our appreciation to all those who contributed to this study, as their efforts contributed to the completion of this research. It has been in great measure due to the willingness to invest the time and relinquish such knowledge in the study. At last, we wanted to express individual thanks to family and friends who always supported us and encouraged us during the most challenging period of this work. We also accept that they are the ones who encouraged most of the efforts and successes. We are grateful a hundredfold to be able to have your valuable possessions, and we thank you for your support of such a worthy cause.

# List of Abbreviations and Acronyms

- AHP - Allied health professional
- BMI - Body Mass Index
- BP - Blood pressure
- CI - Confidence interval
- CVD - Cardiovascular Disease
- FBG - Fasting blood glucose
- HDL-C - High-density lipoprotein cholesterol
- HT - Hypertension
- JNC-7 - The Seventh Report of the Joint National Committee on Prevention, Detection, Evaluation, and Treatment of High Blood Pressure
- LDL-C - Low-density lipoprotein-cholesterol
- preHT - Prehypertension
- SSA - Sub-saharan Africa
- WHO - World Health Organization

# ABSTRACT

This study employs multinomial logistic regression to model the determinants of individual status as normal, prehypertensive, or hypertensive among the Kenyan population. The analysis aims to identify and quantify the impact of various predictors on blood pressure status, thereby providing insights for targeted public health interventions. Data were collected from a representative sample encircling diverse demographic and socioeconomic backgrounds. The predictors examined include age, gender, obesity, smoking status, heart rate, and cholesterol levels. The multinomial logistic regression model exhibited an overall accuracy of 48%, with precision and recall also at 48%.

The analysis identified several significant predictors of blood pressure status. Age being a significant predictor, older individuals were more likely to fall into prehypertensive or hypertensive categories. Gender also showed significant associations, with females exhibiting higher odds of prehypertension and hypertension compared to males. Obesity was another critical factor, basically increasing the likelihood of elevated blood pressure. Smoking status was found to be significantly associated with both prehypertension and hypertension, underscoring the detrimental effects of tobacco use on cardiovascular health. Additionally, elevated heart rate and higher cholesterol levels were significant predictors, further highlighting their roles in influencing blood pressure status. These findings suggest that proper management of prehypertension and hypertension in Kenya requires a comprehensive approach addressing these key risk factors. Public health strategies should prioritize lifestyle interventions aimed at reducing obesity, smoking cessation programs, and routine monitoring of heart rate and cholesterol levels. Moreover, age and gender-specific health education and prevention programs could enhance the effectiveness of interventions. In conclusion, this study provides valuable insights into the factors influencing blood pressure status in Kenya and gives emphasis on the importance of a flexible approach in managing and preventing prehypertension and hypertension. In this situation, the use of multinomial logistic regression provides a strong foundation for understanding the complex interactions between many predictors and guides the creation of focused, evidence-based public health strategies.

# Contents

<b>1</b>	<b>INTRODUCTION</b>	<b>8</b>
1.1	Background of the Study . . . . .	8
1.2	Significance of the Study . . . . .	8
1.3	Problem Statement . . . . .	9
1.4	Objectives of The Study . . . . .	9
1.4.1	General Objective . . . . .	9
1.4.2	Specific Objectives . . . . .	9
<b>2</b>	<b>LITERATURE REVIEW</b>	<b>10</b>
2.1	Introduction . . . . .	10
2.2	Review . . . . .	10
<b>3</b>	<b>METHODOLOGY</b>	<b>12</b>
3.1	Introduction . . . . .	12
3.1.1	Research Design . . . . .	12
3.1.2	Data Source . . . . .	12
3.1.3	Study Population . . . . .	12
3.2	Model Specification . . . . .	13
3.2.1	Logistic Regression Model . . . . .	13
3.2.2	Multinomial logistic regression model . . . . .	13
3.2.3	Interpretation of model coefficients . . . . .	14
3.2.4	Model Assumptions . . . . .	14
3.3	Variable Selection . . . . .	15
3.4	Data Splitting . . . . .	15
3.5	Data Cleaning . . . . .	16
3.6	Model Inference . . . . .	16
3.6.1	Assessing the significance of the predictors . . . . .	16
3.6.2	Confusion Matrix . . . . .	16
<b>4</b>	<b>RESULTS</b>	<b>18</b>
4.1	Dataset description . . . . .	18
4.2	Sample Description . . . . .	18
4.3	Explanatory Data Analysis . . . . .	18
4.3.1	Correlation Analysis . . . . .	19
4.3.2	Testing for Multicollinearity . . . . .	20
4.3.3	Data Pre-processing . . . . .	20
4.4	Uni-variate Exploration . . . . .	21
4.5	Analysis of Personal Attributes . . . . .	22
4.6	Model Summary . . . . .	27
4.6.1	Interpretation of Model Coefficients . . . . .	27
4.7	Model Evaluation . . . . .	31
<b>5</b>	<b>RECOMMENDATIONS AND CONCLUSIONS</b>	<b>32</b>
5.1	Conclusions . . . . .	32
5.2	Limitations of the study . . . . .	32
5.3	Recommendations for future works . . . . .	33

# Chapter 1

## INTRODUCTION

### 1.1 Background of the Study

Hypertension is a major public health problem worldwide, with the World Health Organization (WHO) estimating that 46% of individuals over 25 years of age in sub-Saharan Africa have hypertension. Kenya is no exception with a high prevalence of hypertension and associated cardiovascular risk factors. The burden of hypertension in Kenya is of particular concern because of the country's rapid urbanization and epidemiological transformation, which have contributed to increased risk factors such as changes in dietary patterns and sedentary lifestyles.

Studies have consistently shown that hypertension is a major modifiable risk factor for cardiovascular diseases (CVD) globally. In low- and middle-income settings, including sub-Saharan Africa, hypertension prevalence has been increasing rapidly over the past several decades. The WHO estimates that 46% of individuals over 25 years in SSA have hypertension, with rising rates due to demographic transitions that have led to sedentary lifestyles, smoking, harmful alcohol use, and consumption of processed foods.

In Kenya, hypertension prevalence estimates are high, ranging from 12.6–36.9%, with higher rates in urban areas. Older age, higher body mass index (BMI), alcohol consumption, cigarette smoking, and higher socioeconomic status have been associated with hypertension in previous studies in Kenya. However, diagnosis and treatment of hypertension are often delayed due to its asymptomatic nature, leading to an increased risk of complications and mortality.

The prevalence of prehypertension, a condition characterized by blood pressure levels that are higher than normal but not yet at the level of hypertension, is also a significant concern. Prehypertension is a known risk factor for the development of hypertension and CVD. In Kenya, the burden of prehypertension among adults is substantial, with a prevalence of 54.5%.

This study aims to investigate the prevalence of prehypertension and hypertension in adults in Kenya using a multinomial logit model. The model will analyze the relationships between various demographic and lifestyle factors and the likelihood that an individual will have normal blood pressure, prehypertension or hypertension. The findings of this study will provide valuable insights into the epidemiology of hypertension and prehypertension in Kenya and inform the development of targeted interventions to reduce the burden of these conditions and associated CVD.

### 1.2 Significance of the Study

The findings from the study will be of great importance to the health sector and the citizens at large. Furthermore, other researchers and analysts will use the findings of the study for future comparatives. In addition, the study gaps which emanated from the study will present a strong foundation for future studies. Researchers may also use the findings from the study to find the viability of the present-day managing of hypertensive and prehypertensive cases. The analysts too may get a deeper relationship between variables for the betterment of the health sector in managing hypertensive and prehypertensive cases. Health Research Institutions will also use the findings of the study to detect various loopholes in managing of prehypertensive and hypertensive cases.



## 1.3 Problem Statement

Hypertension is the primary factor contributing to cardiovascular disease. It is accountable for more than 50% of the worldwide morbidity resulting from stroke and heart disease combined. Annually, a staggering 9.4 million people succumb to hypertension-related complications, resulting in their demise. In Kenya, stroke cases attributed to it constitute 64% of the total, according to the World Health Organization (2019). Within Africa, hypertension impacts about 46% of the adult population, with the average blood pressure in the region being notably higher than the global norm (WHO, 2017). Although East Africa currently has lower rates of hypertension compared to the rest of the continent, there are indications that an epidemic may be emerging, particularly due to the rise in urbanization in the region (Stuart Francesc, 2017). In Kenya, 20–30% of the adults were stated to have hypertension in 2013 (Department of Non Communicable Diseases, 2014). According to Yaya et al. (2021), the high blood pressure prevalence among people from rural and urban populations in Kenya was 23.7%. The expenses associated with hypertension have consistently risen over time, and these figures will continue to escalate if adequate management of hypertension in the population is not implemented. Wierzejska (2020). The cost of healthcare specifically related to hypertension in the United States of America has reached a staggering \$131 billion. Kirkland et al. (2018). This underscores the necessity for action to treat and contain hypertension in the community. Developing countries have the largest burden of hypertension, which is due to the high levels of risk factors. The increase in risk factors for hypertension has been mainly attributable to the growth in urbanization that has been noticed in the region (Vijver, 2014). There are few studies and research projects that have been done on the prevalence and contributing factors of hypertension among adult patients attending outpatient clinics in Africa, while in Kenya, no study has been done on the same. A study done on adult patients in Ethiopia found that the prevalence among the adult patients was 27.3%, with alcohol intake, obesity, and abdominal obesity exhibiting five relationships with hypertension. Belachew (2018). However, data concerning the prevalence as well as risk factors of elevated blood pressure in adult patients seeking healthcare in Kenya, specifically Nairobi, is not available. The study aims to investigate the prevalence and contributing variables to high BP among adult patients seeking medical care in Nairobi, Kenya.

## 1.4 Objectives of The Study

### 1.4.1 General Objective

The aim of the research project was to analyze and predict hypertensive status among individuals.

### 1.4.2 Specific Objectives

- To identify risk factors that are significantly associated with hypertensive status of an individual and use these factors to build a logistic regression model for predictions.
- To evaluate the performance of the multinomial logistic regression in predicting hypertensive status of an individual.

## Chapter 2

# LITERATURE REVIEW

### 2.1 Introduction

In this segment we choose to focus on studies that have been done by other scholars with regards to hypertension and prehypertension. By so doing we recognize the strength of these studies, identify weaknesses as well as gaps that may exist in these studies.

### 2.2 Review

Cihangir Erem et al (2008) conducted a study on Prevalence of prehypertension and hypertension and associated risk factors among Turkish adults: Trabzon Hypertension Study. In this cross-sectional survey, a sample of households was systematically selected from the central province of Trabzon and its nine towns. A total of 4809 adult subjects (2601 women and 2208 men) were included in the study. Demographic and socioeconomic factors, family history of selected medical conditions, and lifestyle factors were obtained for all participants. Systolic blood pressure (BP) and diastolic BP levels were measured for all subjects. The persons included in the questionnaire were invited to the local medical centers for blood examination between 08:00-10:00 following 12 hours of fasting. The levels of serum glucose (FBG), total cholesterol (Total-C), high density cholesterol (HDL-C), low density cholesterol (LDL-C) and triglycerides were measured with autoanalyzer. Definition and classification of HT was performed according to guidelines from the US JNC-7 report. Prevalence, awareness, treatment and control of HT were assessed. The prevalences of HT and preHT were 44.0(46.1% in women and 41.6% in men) and 14.5% (12.6% in women and 16.8% in men), respectively. Overall, only 41% of the hypertensive individuals had been previously diagnosed. Furthermore, 54.5% of the hypertensive subjects were being treated with antihypertensive drugs (AHD), but only 24.3% of treated subjects had their BP adequately controlled. Among all hypertensive subjects (known and newly diagnosed), only 5.43% had their BP under control. The prevalence of HT increased with age, being highest in the 60- to 69-year-old age group (84.4%) but lower again in the 70 age group. Interestingly, the prevalence was 16.9% in the 20-to 29-year-old age group. HT was associated positively with marital status, parity, cessation of cigarette smoking, and negatively with level of education, alcohol consumption, current cigarette use, and physical activity. Multinomial logistic regression analysis revealed that HT were significantly associated with age, male gender, BMI, low education level, nonsmoking, positive family history of selected medical conditions, occupation, and parity. The Trabzon Hypertension Study data indicated that HT is very common and is an important health problem in the adult population of Trabzon. Patients who are unaware of their status and treated uncontrolled hypertensives are at high risk of early cardiovascular morbidity and mortality. To control preHT and HT, effective public health education and urgent precautions are needed. The precautions include serious health education, a well-balanced diet and increasing physical activity.

Wesly Jeune et al (2018) conducted a study: Multinomial Logistic Regression and Random Forest Classifiers in Digital Mapping of Soil Classes in Western Haiti and further used a confusion matrix to evaluate the accuracy of the model. The confusion matrix for the classification showed an overall accuracy of 55%.

Diego Chambergó-Michilot et al (2021) conducted a study on Socioeconomic determinants of hypertension and prehypertension in Peru. Their main aim was to assess the association between socioeconomic determinants, hypertension and prehypertension using a nationally representative survey of Peruvians. They performed a cross-sectional analysis of the Peruvian Demographic and Health Survey (2018), which is a two-staged regional-level representative survey. They used data from 33,336 people aged 15 and older. The dependent variable was blood pressure classification (normal, prehypertension and hypertension) following the Seventh Report of the Joint National Committee (JNC-7) on hypertension management. Independent variables were socioeconomic: age, sex, marital status, wealth index, health insurance, education, region and area of residence. Due to the nature of the dependent variable (more than two categories), they opted to use the multinomial regression model, adjusting the effect of the multistage sample using the svy command. They tested interactions with the adjusted Wald test. The prevalence of prehypertension and hypertension was 33.68% and 19.77%, respectively. Awareness was higher in urban than in rural areas (9.61% vs. 8.31%,  $p = 0.008$ ). Factors associated with a higher prevalence ratio of both prehypertension and hypertension were age (ratios rose with each age group), male sex (prehypertension aRPR 5.15, 95%CI 4.63–5.73; hypertension aRPR 3.85, 95% CI 3.37–4.40) and abdominal obesity (prehypertension aRPR 2.11, 95%CI 1.92–2.31; hypertension aRPR 3.04, 95% CI 2.69–3.43). Factors with a lower prevalence ratio of both diseases were secondary education (prehypertension aRPR 0.76, 95%CI 0.60–0.95; hypertension aRPR 0.75, 95% CI 0.58–0.97), higher education (prehypertension aRPR 0.78, 95%CI 0.61–0.99; hypertension aRPR 0.62, 95% CI 0.46–0.82), being married/cohabiting (prehypertension aRPR 0.87, 95%CI 0.79–0.95; hypertension aRPR 0.77, 95% CI 0.68–0.87), richest wealth index (only prehypertension aRPR 0.76, 95% CI 0.63–0.92) and living in cities different to Lima (rest of the Coastline, Highlands and Jungle). Having health insurance (only hypertension aRPR 1.26, 95%CI 1.03–1.53) and current drinking (only prehypertension aRPR 1.15, 95%CI 1.01–1.32) became significant factors in rural areas. Their conclusion evidenced socioeconomic disparities among people with hypertension and prehypertension. Better health policies on reducing the burden of risk factors are needed, besides, policy decision makers should focus on hypertension preventive strategies in Peru.

Dil Bahadur Rahut et al (2023) did research on the prevalence of prehypertension and hypertension among the adults in South Asia. The study used the nationally representative data collected through a recent round of DHS in India (2019–2021), Nepal (2016), and Bangladesh (2017–2018). The data collection applied a multi-stage sampling process. The enumeration areas and the primary sampling units were selected in the first step. In the second stage, sample households were selected from the primary sampling units. As a result, 637,396 respondents in India, 8,924 in Nepal, and 8,613 in Bangladesh, aged 18–49 years, were selected for blood pressure measurement. Individuals on medication for hypertension were excluded from the analysis, and other individuals with the missing outcome and dependent variables were also excluded from the study. Details about the data collection process are available on the DHS website. The ethical review of the data collection was done by the DHS and the government of the country involved. Descriptive and econometric analyses were done separately for each country, and accordingly, tables reporting. Since the dependent variable is three discrete and mutually exclusive, multinomial logistic regression was used to understand the factors associated with prehypertension and hypertension among adults aged 18–49 years. This study had empirically analyzed the determinants of prehypertension and hypertension together by looking at socio-demographic characteristics in a single framework utilizing the recent extensive sample data from Bangladesh, India, and Nepal. The study showed that geographical locations (division in Bangladesh, states in India, and province in Nepal), gender differences, level of education only in the case of India, wealth, and overweight and obesity are important determinants of the prevalence of hypertension and prehypertension. Most importantly, the patterns for prehypertension follow a similar trend in these countries, such that prehypertension is highly prevalent in the young and economically active age group. In this way, understanding the prevalence of prehypertension in South Asia and its alarming rate mandates preventive measures. The prevalence of hypertension and prehypertension and the disease burdens in the productive mid-life period shall adversely affect workforce productivity and economic development in these regions. In conclusion the above studies have shown that social-demographic factors such as age, level of education, gender, marital status, BMI, smoking status are significantly associated with prehypertension and hypertension. This offered a compelling argument in favor of including them as research variables in this study.

## Chapter 3

# METHODOLOGY

### 3.1 Introduction

The purpose of this study is to find out how common adult hypertension is in Nairobi, Kenya, as well as what variables contribute to its development. The objective of this study is to give a thorough understanding of the determinants of hypertension in an urban Kenyan population by looking at a variety of demographic, behavioral, and clinical factors. The results will enhance the body of knowledge already available on hypertension in sub-Saharan Africa and aid in the creation of focused interventions aimed at lowering the prevalence of hypertension and enhancing cardiovascular health outcomes in Kenya.

This chapter covers the many techniques that were employed to guarantee precise responses to the research proposal's questions and goals. The research design, sampling strategy, target population, data collection techniques, validity and reliability, data analysis techniques, and ethical considerations are all covered in this chapter.

#### 3.1.1 Research Design

An observational cohort study was used to carry out this investigation. This research design had the advantage that it was inexpensive and was relatively easy to conduct. Descriptive and analytic data analysis methods. Descriptive analytical method was used to create a summary of the data and point to observed patterns. Data analytics was used to find trends in the information gathered and draw conclusions from the data sets.

#### 3.1.2 Data Source

An online data repository served as the secondary source from which the study's data were gathered. A wealth of health-related data, such as blood pressure readings, lifestyle factors, medical history, and demographics, were accessible through this repository. We made sure to use secondary data in a way that was both economical and quick in order to collect a significant amount of pertinent data for our investigation. The repository is renowned for its dependability, thoroughness, and observance of moral principles when gathering and handling data.

#### 3.1.3 Study Population

The study's target population consisted of persons who were at risk of developing hypertension and were between the ages of 30 and 70. Pre-existing conditions, family history, lifestyle choices, and prior medical records showing high blood pressure were taken into consideration for determining this risk. The sample was selected from the resident population of a US state, guaranteeing a cohort that was defined geographically. This selection criteria improved the study's relevance and application to similar groups by enabling a targeted inquiry into the incidence and risk factors of hypertension within a particular age range and demographic.

## 3.2 Model Specification

Statistics model-building strategies seek the best-fitting and most logical model to describe the relationship between an outcome variable and a set of predictors (independent) variables. These independent variables are referred to as covariates. A typical technique that is frequently used is linear regression model with a continuous outcome variable. For this study however, the outcome variable was the hypertensive status which is not a continuous variable. The logistic regression model was therefore used.

### 3.2.1 Logistic Regression Model

A multinomial logistic regression model is a simple extension of the binary logistic regression model that allows for more than two categories of the dependent variables. It is a model that is used to predict the probabilities of the different possible outcomes of a categorically distributed dependent variables. Since the response variable is a non-numeric a dummy variable is created which is of the form; The response

Table 3.1: Hypertensive Status

Hypertensive Status	
0	If a person is normal
1	If a person is hypertensive
2	If a person is pre-hypertensive

variable is therefore a random variable governed by the multinomial distribution.

### 3.2.2 Multinomial logistic regression model

A multinomial logistic regression model uses a predictor linear function that constructs a score from a set of regression coefficients that are linearly combined with the explanatory variables of a given observation using a dot product score. According to Hosmer and Lemeshow (2004), code the outcome variable, Y as 0, 1, or 2. In the three outcomes, category model to logit functions are needed. We use  $Y = 0$  as the referent, or baseline, outcome and to form logit functions comparing each other category to it. To develop the model, assume there are p covariates and a constant term, denoted by vector x, of length p+1, where  $x_0 = 1$ . The two logit functions are denoted as:

$$g_1(x) = \ln \left[ \frac{\Pr(Y = 1|x)}{\Pr(Y = 0|x)} \right]$$

$$= \beta_{10} + \beta_{11}x_1 + \beta_{12}x_2 + \dots\beta_{16}x_6$$

$$= x' \beta_1$$

$$g_2(x) = \ln \left[ \frac{\Pr(Y = 2|x)}{\Pr(Y = 0|x)} \right]$$

$$= \beta_{20} + \beta_{21}x_1 + \beta_{22}x_2 + \dots\beta_{26}x_6$$

$$= x' \beta_2$$

It follows that the conditional probabilities of each outcome category given the covariate vectors are

$$\Pr(Y = 0|x) = \frac{1}{1 + e^{g_1(x)} + e^{g_2(x)}}$$

$$\Pr(Y = 1|x) = \frac{1}{1 + e^{g_1(x)} + e^{g_2(x)}}$$

and

$$\Pr(Y = 2|x) = \frac{e^{g_2(x)}}{1 + e^{g_1(x)} + e^{g_2(x)}}$$

A general expression for the conditional probability in the three-category model is

$$\pi_j(x) = \Pr(Y = j|x) = \frac{e^{g_j(x)}}{\sum_{k=0}^2 e^{g_k(x)}}$$

where the vector  $\beta_0 = 0$  and  $g_0(x) = 0$

### 3.2.3 Interpretation of model coefficients

In order to interpret the coefficients of our logistic regression model we need to understand the relationship between the predictor variable and the probability of prehypertensive and hypertensive. The interpretation of categorical and continuous predictors shows the magnitude and direction of the effect for example if the coefficient for “Gender” predictor is positive, it indicates that belonging to that gender, increases the log-odds of the outcome when compared to the reference group. The continuous coefficient indicates the change in log-odds or odd ratio for a one-unit increase in the predictor. A positive coefficient indicates an increase in the log-odds ratio as the predictor increases by one unit while a negative coefficient suggests a decrease.

### 3.2.4 Model Assumptions

- I. Appropriate Result Type:** The multinomial logit model assumes that the outcome variable is categorical with more than two unordered categories. Categories of the outcome variable must be mutually exclusive and exhaustive. It can be, for example, a choice between several political parties or a preference for different types of food.
- II. Independence of Irrelevant Alternatives (IIA):** This assumption states that the chances of choosing one category over another are independent of the presence or absence of other categories. In other words, the relative probability of choosing between options A and B should not be affected by the presence of option C.
- III. Absence of perfect multicollinearity:** Similar to binary logistic regression, the multinomial logit model assumes the absence of perfect multicollinearity between the independent variables. Perfect multicollinearity occurs when one independent variable is a perfect linear combination of the others, making it impossible to estimate unique coefficients.
- IV. Independence of observations:** The observations used to estimate the multinomial logit model should be independent of each other. This means that the choice made by one individual should not affect the choice of other individuals in the sample.
- V. Sufficient sample size:** An adequate sample size is required to ensure reliable estimates of model parameters. A small sample size can lead to unstable parameter estimates and unreliable predictions.
- VI. Proportional odds:** Also known as the parallel regression assumption, this assumption states that the effect of the independent variables on the log odds of each category versus the reference category is constant across categories. In other words, the relationship between the independent variables and the outcome is consistent across categories.

### 3.3 Variable Selection

Our variables are defined in as the table below:

Table 3.2: Explanation of Response Variable

Response Variable Y	Explanation
Hypstatus	<p>Categorizes individuals based on their hypertension status.</p> <p><b>Normal (not hypertensive):</b> Individuals who do not exhibit hypertension.</p> <p><b>Pre-hypertensive:</b> Individuals who have blood pressure levels higher than normal but not high enough to be classified as hypertensive.</p> <p><b>Hypertensive:</b> Individuals diagnosed with hypertension, indicating consistently high blood pressure.</p>

Table 3.3: Explanation of Predictors

$X_i$	Predictor	Explanation	Coding
$X_1$	Sex	Categorizes individuals based on their gender.	0=Male 1=Female
$X_2$	Heart rate	Measures the number of heartbeats per minute.	Continuous variable
$X_3$	Obesity	Indicates whether an individual is obese or not.	Obese Not Obese
$X_4$	Smoking Status	Categorizes individuals based on their smoking habit.	Smoker Non smoker
$X_5$	Cholesterol	Measures the level of cholesterol in the blood.	Continuous variable
$X_6$	Age	Represents the age of the individual.	Continuous variable

### 3.4 Data Splitting

We employed the practice of splitting our data-set into two distinct subsets: a training set and a testing set. This method ensures that we can train the model on one portion of the data and evaluate its performance on an unseen portion. Specifically, we used 75% of the data for model training and the remaining 25% for testing.

## 3.5 Data Cleaning

The effectiveness of the final model hinges significantly on the quality of the underlying dataset. Some primary issues that affect data quality are misinterpretation, inadequate capture and missing values. These significantly impact the reliability and accuracy of the final model.

A Multinomial logistic regression is sensitive to data quality and it requires complete datasets without missing values to ensure robust model estimation. However it is flexible in that missing values can be incorporated as a predictor, this provides insights into their relationship with the other predictors and the potential influence it has on the outcome.

Missing values carry informative signals and are not randomly distributed. Individuals with certain characteristics could be more likely to have missing data, this indicates an underlying patterns with other variables.

## 3.6 Model Inference

### 3.6.1 Assessing the significance of the predictors

There are two major approaches used to assess the significance of predictors. The two methods are discussed below:

1. Testing the statistical hypothesis that  $H_0 : \beta_0 = 0$  against the alternative hypothesis that  $H_1 : \beta_0 \neq 0$ . The Z-test statistic is used, given by  $Z = \frac{\beta_1}{s.e(\beta_1)}$ . Reject  $H_0$  if  $|Z| > Z_{\alpha/2}$ , and in this case, we conclude that  $X$  is a significant predictor.
2. The second approach is to test the hypothesis that the odds ratio  $e^{\beta_i} = 1$  against the alternative hypothesis that  $e^{\beta_i} \neq 1$ . The hypothesis is evaluated using a  $100(1 - \alpha)$  confidence interval  $e^{\beta_1} - Z_{\alpha/2} \cdot s.e(\beta_1); e^{\beta_1} + Z_{\alpha/2} \cdot s.e(\beta_1)$ . The predictor is statistically significant if the value 1 is not included in the interval, i.e., the lower confidence limit is less than 1 and the upper confidence limit is greater than 1.

### 3.6.2 Confusion Matrix

The confusion matrix was utilized in our study to evaluate the effectiveness of our logistic regression model. Here is a quick summary of the main ideas behind the approach. The confusion matrix is a useful instrument for evaluating a logistic regression model's effectiveness. It gives a detailed explanation of how the model's predictions and the actual results differ. We can gain insight about the model's accuracy, precision, recall, and overall efficacy in predicting the target variable by looking at the numbers in the confusion matrix.

The components of the confusion matrix in the context of multinomial logistic regression are:

1. True Positives (TP): These are the cases where the model correctly predicts the positive class for a given category.
2. True Negatives (TN): These are the cases where the model correctly predicts the negative class for a given category. In multinomial logistic regression, each category has its own set of negative classes.
3. False Positives (FP): These are the cases where the model incorrectly predicts the positive class for a given category.
4. False Negatives (FN): These are the cases where the model incorrectly predicts the negative class for a given category.

We can as well derive numerous evaluation metrics from the confusion matrix; Accuracy is determined as  $\frac{TP+TN}{TP+TN+FP+FN}$  and represents the total correctness of the forecasts. It represents the percentage of correctly classified instances. Precision, also known as positive predictive value, is the ratio of genuine positives to predicted positives computed as  $\frac{TP}{TP+FP}$ . Precision is concerned with the accuracy of positive



predictions, reflecting the model's capacity to reduce false positives. The fraction of true positives among actual positive events, computed as  $\frac{TP}{TP+FN}$ , is known as recall. Recall is a metric that assesses a model's ability to correctly identify positive instances and is significant in situations where minimizing false negatives is critical. The F1 score is the harmonic mean of precision and recall and gives a balanced measure of the model's performance, considering both precision and recall.

## Chapter 4

# RESULTS

### 4.1 Dataset description

The research used a hypertension dataset. The dataset contains anonymous demographic information about the individuals such as age and gender. The primary goal of the project is to investigate the prevalence of prehypertension and hypertension. This dataset was chosen for its ability to pinpoint key indicators linked to hypertension and aid the healthcare sector to properly assess the condition. We intend to construct a predictive model using this data to investigate how different variables relate to hypertension. The following is a brief description of the variables contained in the data. Each individual is assigned a unique ID. 'SEX' denotes the gender, with 1 for male and 0 for female. 'AGE' represents individuals' ages in years. 'CHOLESTEROL' shows individuals' cholesterol levels in mg/dL. 'HEART' indicates individuals' heart rate in beats per minute. 'HYPSTATUS' displays whether an individual is normal, prehypertensive or hypertensive. 'SMOKING' reveals if an individual is a smoker or non-smoker. 'OBESE' indicates if an individual is obese or non-obese.

### 4.2 Sample Description

The sample size consisted of 4,187 individuals whom were of various hypertensive status. Of the individuals; 1021 were normal constituting to 24.39%. 1494 were of prehypertensive status constituting to 35.68% and 1672 individuals of hypertensive status translating to about 39.93%. The table below gives a description of the sample.

Hypertensive Status	Count	Percentage
0 - Normal	1021	24.39%%
1 - Hypertensive	1672	39.33%%
2 - Pre-hypertensive	1494	35.68%%

Table 4.1: A 3x3 Table

### 4.3 Explanatory Data Analysis

Exploratory data analysis was performed on the data set so as to get a deeper insight of the data set. This entailed handling of missing values, cleaning the data as well as performing uni-variate exploration to be able to understand the data set. This analysis was aided and conducted using python and R programming languages.

### 4.3.1 Correlation Analysis

A correlation heatmap was plotted to determine the variables that were strongly related with each other.

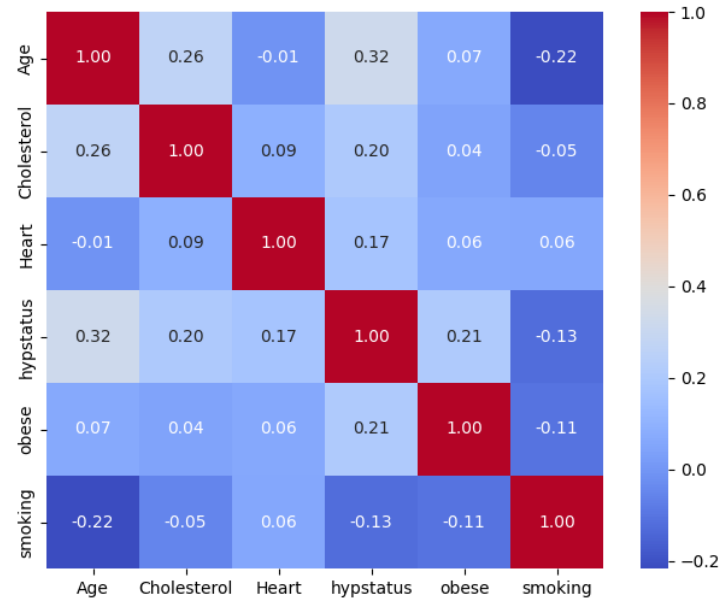


Figure 4.1: A sample graph

Variables are considered to be strongly correlated with a value of 0.7 or greater from the heatmap. However, from the above heatmap there is no correlation of close to 0.7. The highest correlation is 0.32 between hypstatus and age and the lowest is between heart and age that is 0.01. From the above values we can conclude that the variables are weakly correlated.

### 4.3.2 Testing for Multicollinearity

A heatmap plot with the variables age, heart rate and cholesterol levels were used to test for multicollinearity and the results are as shown below:

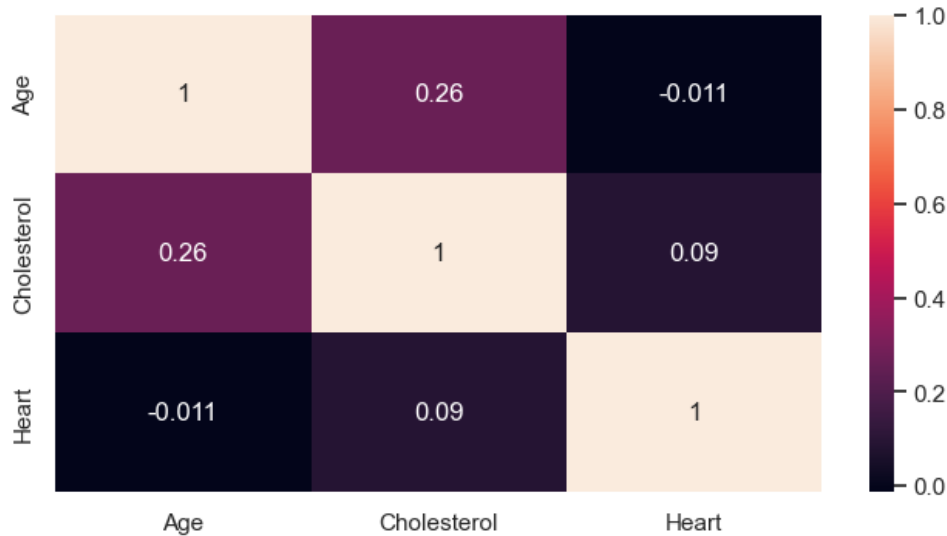


Figure 4.2: A sample graph

From the above heatmap it is sufficient to note that there is absence of multicollinearity between the variables.

### 4.3.3 Data Pre-processing

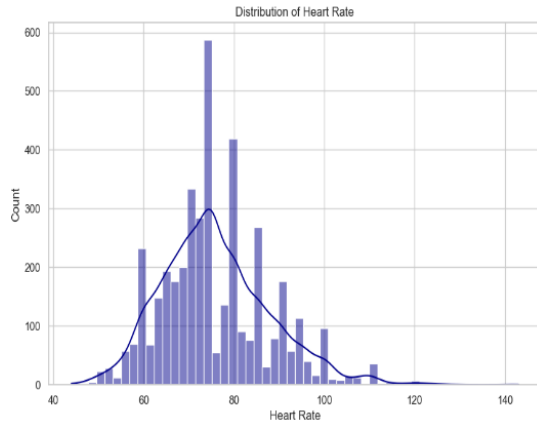
This entailed naming and defining all available variables in the data, determining the data types of values within instances of that data, and ensuring compliance with implied data type requirements for variable values in all circumstances chosen to be part of the study. Data exploration was done by calculating averages and counts for relevant variables, maximum values and minimums, medians and modes, in order to gain a meaningful understanding of the data and the variables contained within it, as well as to observe the balance/imbalance portrayed by the data in relation to the target variable under consideration. Values in the various fields of the data were also examined to see if there were any obvious patterns or remarkable features, such as variables that appeared to be categorical variables (and those that were anticipated to be categorical variables by their nature) were transformed into factors for easier analysis. There were no case of null values in the data.

#### Data Cleaning

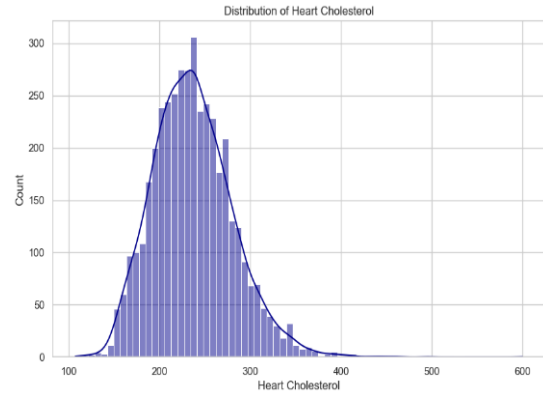
None of the 7 variables had a missing value. Quality of data affects the performance statistical machine learning models. Logistic regression is therefore not an exception so data should be tidy! For our data however, there were no cases of missing values-all the data points were present. We only re-coded some variable types into an appropriate format that was suitable to be understood and implemented by our programming/data analysis tool (Python)

## 4.4 Uni-variate Exploration

By graphing the continuous variable heart rate and cholesterol level we can observe a bell-shaped curve implying the data is normally distributed.



(a) Caption for Image 1



(b) Caption for Image 2

Figure 4.3: Comparison of two images

## 4.5 Analysis of Personal Attributes

### Gender

A bar graph was used to do comparative analysis of two genders and their relationship to hypstatus and the results are as shown below:

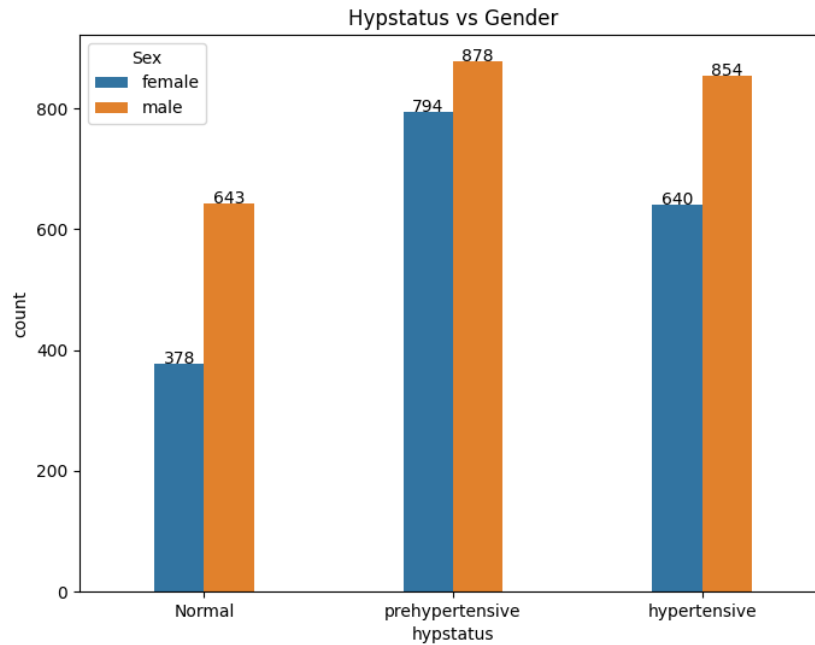


Figure 4.4: A sample graph

Of the 4,187 individuals, 2,375(56.72%) were males while the remaining 1,812(43.28%) were females. Of the 2,375 males 643(27.07%) were normal, 878(36.97%) were prehypertensive and finally the remaining 854(35.96%) were hypertensive. For the females 378(20.86%) were normal, 794(43.82%) were prehypertensive and 640(35.32%) were hypertensive.

### Age

A bar graph was used to do comparative analysis on the ages of the individuals and the results are shown below:

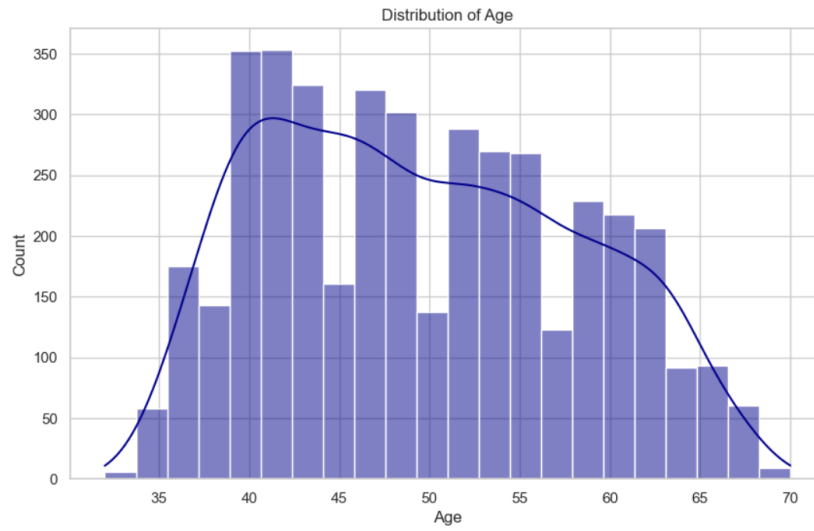


Figure 4.5: A sample graph

From the above, age is normally distributed with mean 49.56 and median 49.00

## Heart Rate

We used a bar graph to do comparative analysis between different heart beat rates. The results are as presented below:

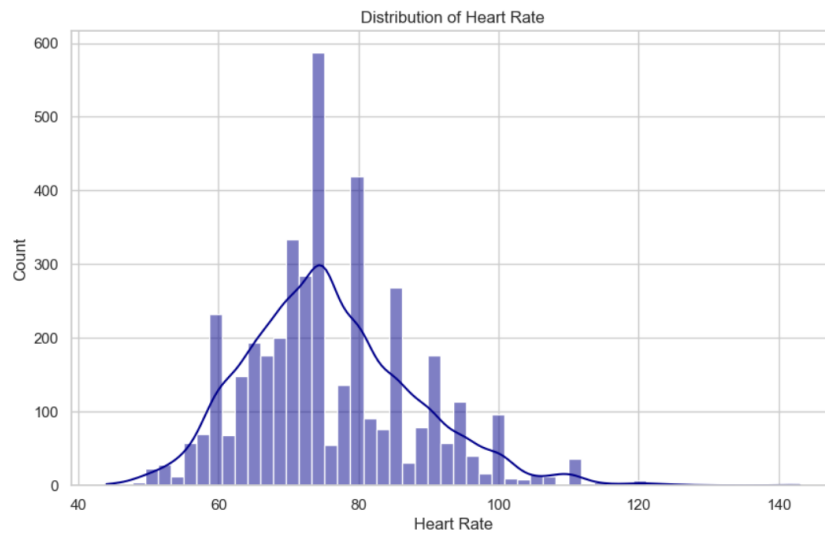


Figure 4.6: A sample graph

From the above, heart rate is noted to be normally distributed and consequently has a mean of 75.86 and median of 75.00



### Cholesterol Level

We used a bar graph to comparative analysis on the cholesterol levels. The results are as interpreted below:

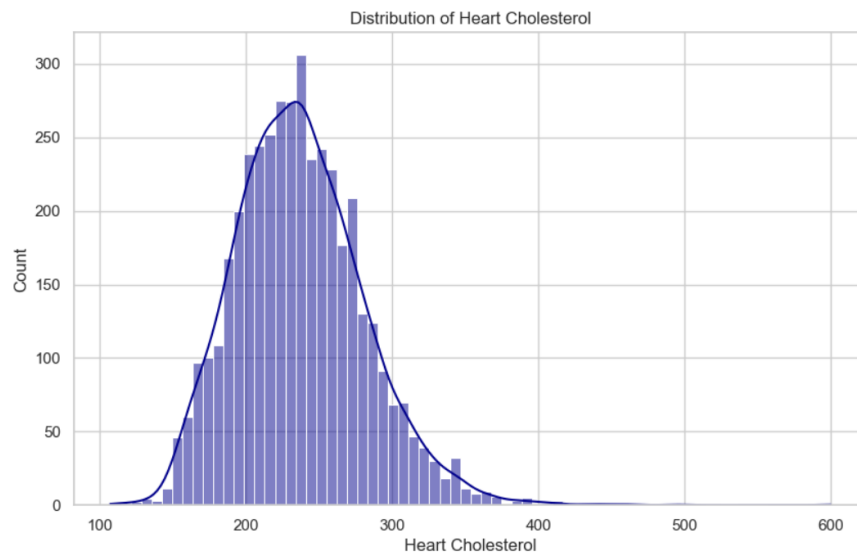


Figure 4.7: A sample graph

Cholesterol Level is normally distributed with a mean of 236.7 and has a median of 234.0

### Smoking Status

We a bar graph to do comparative analysis on the smoking status of an individual and the results are as shown below:

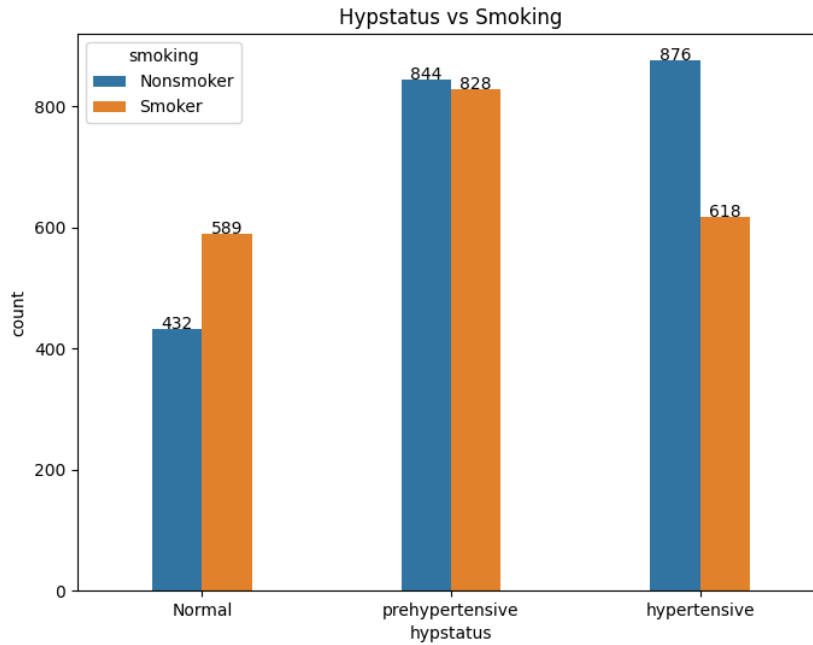


Figure 4.8: A sample graph

Of the 4, 187 individuals 2,035(48.60%) were smokers while the remaining 2,152(51.40%) were non-smokers. Of the smokers 589(28.94%) were normal, 828(40.69%) were prehypertensive and finally 618(30.37%) were hypertensive. Of the non-smokers 432(20.07%) were normal, 844(39.22%) were prehypertensive and finally 876(40.71%) were hypertensive.

### Obesity Status

We a bar graph to do comparative analysis on the obesity status of an individual and the results are as shown below:

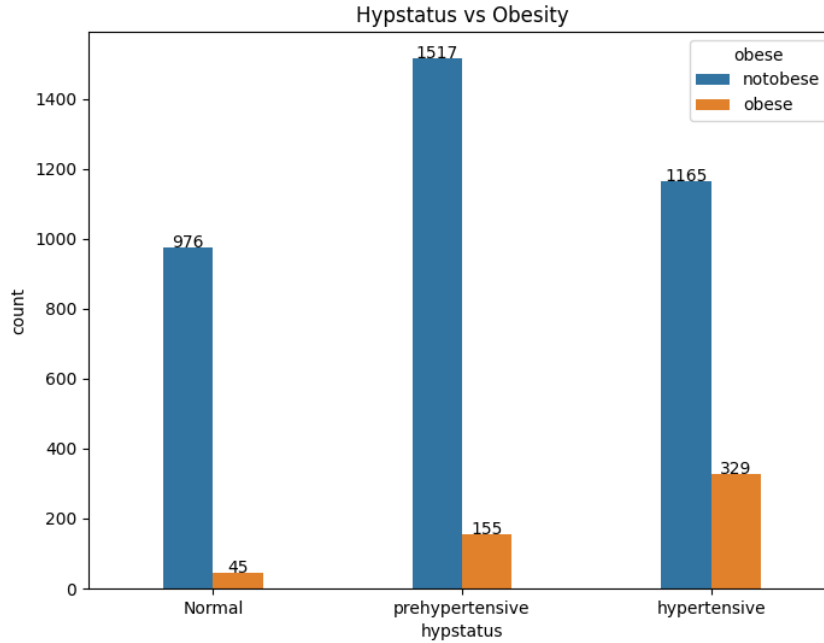


Figure 4.9: A sample graph

Of the 4187 individuals 529(12.63%) were obese while the remaining 3,658(87.37%) were non-obese. Of the obese individuals 45(8.51%) were normal 155(29.30%) were prehypertensive and finally 329(62.19%) were hypertensive. Of the non-obese individuals 976(26.68%) were normal, 1517(41.47%) were prehypertensive and finally 1165(31.85%) were hypertensive.

## 4.6 Model Summary

All our independent variables: age, cholesterol levels, heart rate, smoking status, obesity status and gender were all included since they had a p-value less than 0.05. This implied that they were significant in predicting the hypertensive status among the individuals under study. To build the multinomial logistic regression we split our data in the ratio 7.5:2.5 where by 75% of the data was used in training and the remaining 25% was used for evaluating the model. The performance of the model was assessed using a confusion matrix. From the confusion we obtained three metrics which we used to evaluate our model; accuracy, recall and precision. The model had an accuracy, recall and precision of 48%, 48% and 48% respectively. It can therefore be concluded that the model performed averagely in carrying out predictions.

### 4.6.1 Interpretation of Model Coefficients

The response variable was the hypertensive status of an individual with normal coded as 0, hypertensive as 1 and prehypertensive as 2. That meant that the base group was normal individuals. The classical interpretations of the coefficients of the model are made as follows. The following table shows the summary output of the model together with the coefficients, probability value and 95% confidence interval.

MNLogit Regression Results						
=====						
Dep. Variable:	hypstatus		No. Observations:		3140	
Model:	MNLogit		Df Residuals:		3126	
Method:	MLE		Df Model:		12	
Date:	Sat, 18 May 2024		Pseudo R-squ.:		0.1074	
Time:	13:00:02		Log-Likelihood:		-3022.7	
converged:	True		LL-Null:		-3386.3	
Covariance Type:	nonrobust		LLR p-value:		6.314e-148	
=====						
hypstatus=1	coef	std err	z	P> z	[0.025	0.975]
-----						
const	-10.4590	0.568	-18.411	0.000	-11.572	-9.346
Sex	0.5393	0.110	4.887	0.000	0.323	0.756
Age	0.0987	0.007	14.179	0.000	0.085	0.112
Cholesterol	0.0093	0.001	7.226	0.000	0.007	0.012
Heart	0.0471	0.005	10.155	0.000	0.038	0.056
smoking	-0.4623	0.110	-4.210	0.000	-0.677	-0.247
obese	1.7066	0.200	8.541	0.000	1.315	2.098
-----						
hypstatus=2	coef	std err	z	P> z	[0.025	0.975]
-----						
const	-4.1863	0.485	-8.634	0.000	-5.137	-3.236
Sex	0.5383	0.099	5.448	0.000	0.345	0.732
Age	0.0351	0.006	5.510	0.000	0.023	0.048
Cholesterol	0.0060	0.001	5.025	0.000	0.004	0.008
Heart	0.0210	0.004	4.889	0.000	0.013	0.029
smoking	-0.3415	0.098	-3.469	0.001	-0.534	-0.149
obese	0.6204	0.205	3.021	0.003	0.218	1.023
=====						

Figure 4.10: A sample graph

The model coefficients were exponentiated and converted to odd ratio for easier interpretation and the results are represented in the table below.

## Odds Ratios:

	0	1
const	0.000029	0.015203
Sex	1.714776	1.713118
Age	1.103763	1.035772
Cholesterol	1.009377	1.006019
Heart	1.048220	1.021220
smoking	0.629856	0.710723
obese	5.510182	1.859741

Figure 4.11: Odd ratios

## Hypertensive (compared to normal)

### 1. Sex

The sex variable had 2 categories, male and female with males being used as the base group. The coefficient for sex was 0.5393 with a corresponding odds ratio of 1.714776 implying that being female increases the odds of one being hypertensive by 71.48%.

### 2. Age

The coefficient for age was 0.0987 with a corresponding odds ratio of 1.103763 implying that an increase in age increases the odds of being hypertensive by 10.38%.

### 3. Cholesterol Level

The coefficient for cholesterol level was 0.0093 with a corresponding odds ratio of 1.009377 suggesting that a unit increase in cholesterol level increases the odds of being hypertensive by 0.94%.

### 4. Heart rate

The coefficient for heart rate was 0.0471 with a corresponding odds ratio of 1.048220 implying that an increase in heart rate increases the odds of being hypertensive by 4.82%.

### 5. Smoking

The coefficient of smoking was  $-0.4623$  with a corresponding odds ratio of 0.629856 implying that being a smoker decreases the odds of being hypertensive by 37.01%.

### 6. Obese

The coefficient of being obese was 1.7066 with a corresponding odds ratio of 5.510182 implying that being obese increases the odds of being hypertensive by 451.02%.

## Prehypertensive (compared to normal)

### 1. **Sex**

The sex variable had 2 categories, male and female with males being used as the base group. The coefficient for sex was 0.5383 with a corresponding odds ratio of 1.713118 implying that being female increases the odds of prehypertension by 71.31%.

### 2. **Age**

The coefficient for age was 0.0351 with a corresponding odds ratio of 1.035772 implying that an increase in age increases the odds of being prehypertensive by 3.58%.

### 3. **Cholesterol Levels**

The coefficient for cholesterol was 0.0060 with a corresponding odds ratio of 1.006019 implying that a unit increase in cholesterol levels increases the odds of prehypertension by 0.60%.

### 4. **Heart rate**

The coefficient for heart rate was 0.0210 with a corresponding odds ratio of 1.021220 implying that an increase in heart rate increases the odds of being prehypertensive by 2.12%.

### 5. **Smoking**

The coefficient for smoking was  $-0.3415$  with a corresponding odds ratio of 0.710723 implying that being a smoker decreases the odds of being prehypertensive by 28.93%.

### 6. **Obesity**

The coefficient for obesity was 0.6204 with a corresponding odds ratio of 1.859741 implying that being obese increases the odds of being prehypertensive by 85.97%.

## 4.7 Model Evaluation

### Confusion Matrix

A confusion matrix was used to evaluate the model performance as shown below:

Table 4.2: Confusion Matrix

	Normal	Hypertensive	Prehypertensive
True Positive	88	196	240
False Positive	66	161	178
True Negative	167	178	296
False Negative	726	512	333

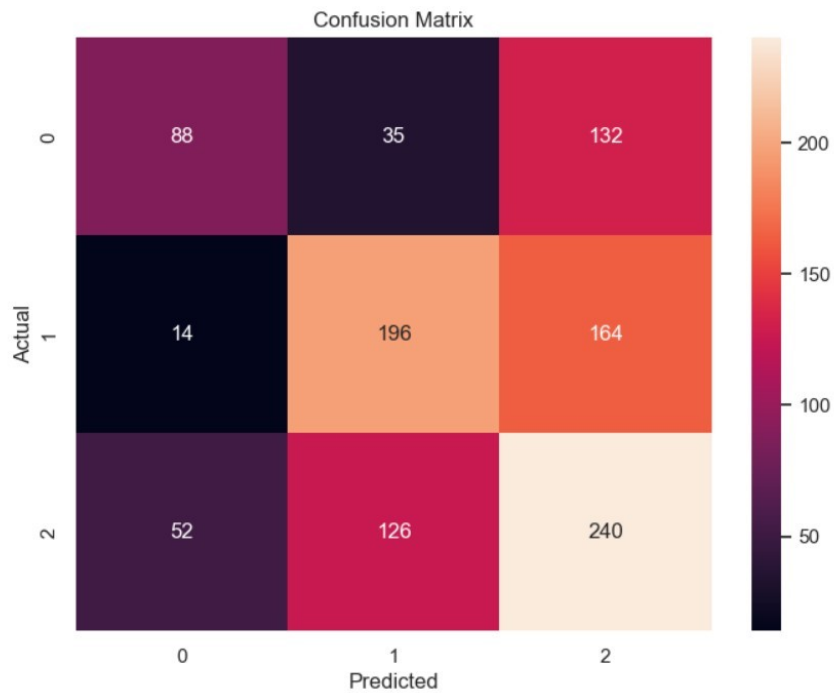


Figure 4.12: A sample graph

The classification report matrices such as accuracy, precision, recall and f1 score that were used in evaluating the model are represented in the table below:

Table 4.3: Classification Report

	Precision	Recall	F1 Score	Support
0	0.44	0.51	0.48	255
1	0.53	0.53	0.53	374
2	0.46	0.42	0.44	418
<b>Accuracy</b>			0.48	1047
<b>Macro Avg</b>	0.48	0.49	0.48	1047
<b>Weighted Avg</b>	0.48	0.48	0.48	1047

The explanation for the above is given in section 5.1 under summary of the logistic regression model.

## Chapter 5

# RECOMMENDATIONS AND CONCLUSIONS

### 5.1 Conclusions

In order to analyze and predict hypertensive status among individuals, this research has illustrated an assessment of various risk factors. This study was focused on achieving the aforementioned objectives with the overall purpose of building a hypertensive status prediction model. The goals were as follows:

- (i) To identify risk factors significantly associated with hypertensive status. Our research established that personal characteristics like age, gender, heart rate, cholesterol level, smoking status, and obesity status have a significant influence on an individual's hypertensive status. This was supported by both the literature reviewed and the findings of our investigation.
- (ii) To evaluate the performance of the multinomial logistic regression in predicting hypertensive status. An examination of how these risk factors influence hypertensive conditions suggests that variables like age and heart rate are critical predictors of hypertension. It is important to note that while these factors are weakly correlated, their collective impact is essential in the model's predictive capability.
- (iii) Build a model that can predict hypertensive status. In our work, multinomial logistic regression analysis was utilized to construct a prediction model. The model accurately classified individuals into normal, prehypertensive, and hypertensive categories. This suggests that multinomial logistic regression is a good fit for similar contexts and, depending on the problem area, other machine learning approaches are also worthwhile considerations for enhancing predictive accuracy.

In summary, this study underscores the importance of thorough data analysis and the application of logistic regression in predicting health conditions like hypertension. The findings suggest that such models can be robustly applied to similar datasets, contributing valuable insights for healthcare providers and policymakers. Future research could further refine these models and explore additional variables to improve performance on hypertensive status prediction.

### 5.2 Limitations of the study

Our study had a few drawbacks which are briefly explained below. Our data which we used in the study came from a single institution. The results could have been different if we used data from a different institution. Furthermore, interpretation of model coefficients and odds ratio was more intricate which led to difficulty in explaining how changes in age, gender, obesity, heart rate, smoking status and cholesterol levels impact hypertensive status simultaneously. We clarified the interpretation process essentially for accurate understanding.



### 5.3 Recommendations for future works

The study focused on a few aspects related to hypertension therefore we recommend that future studies look into other aspects such as genetics, life choices like diet and physical activity, stress level and underlying health conditions such as kidney disease or diabetes. These factors could provide variable insights in understanding and management of hypertension. Also, the study can be modelled using various techniques. Some techniques are more accurate than others and a technique's accuracy may vary based on the data set employed. As a result, we recommend future studies to incorporate various modelling techniques in the study. New technologies such as artificial intelligence are continuously improving to increase performance of the models. Therefore, we encourage future scholars to employ these emerging technologies in the study of hypertension.

## References

- Chambergo-Michilot, D., Rebatta-Acuña, A., Delgado-Flores, C. J., & Toro-Huamanchumo, C. J. (2021). Socioeconomic determinants of hypertension and prehypertension in Peru: Evidence from the Peruvian Demographic and Health Survey. *PLoS ONE*, 16(1), e0245730.
- Rahut, D. B., Mishra, R., Sonobe, T., & Timilsina, R. R. (2023). Prevalence of prehypertension and hypertension among the adults in South Asia: A multinomial logit model. *Frontiers in Public Health*, 10, 1006457.
- Karadeniz Technical University, Faculty of Medicine, Department of Internal Medicine, Division of Endocrinology and Metabolism, The Trabzon Endocrinological Studies Group, Trabzon, Turkey; Department of Biochemistry, Trabzon Endocrinological Studies Group, Trabzon, Turkey; Department of Public Health, The Trabzon Endocrinological Studies Group, Trabzon, Turkey. Address correspondence to Cihangir Erem, E-mail: [cihangirerem@hotmail.com](mailto:cihangirerem@hotmail.com)/[cihangirerem@netscape.net](mailto:cihangirerem@netscape.net)
- Hosmer and Lemeshow type Goodness-of-Fit Statistics for the Cox Proportional Hazards Model
- Belachew, A., Tewabe, T., Miskir, Y., Melese, E., Wubet, E., Alemu, S., Tesfa, G. (2018). Prevalence and associated factors of hypertension among adult patients in Felege-Hiwot Comprehensive Referral Hospitals, northwest, Ethiopia: a cross-sectional study. *BMC research notes*, 11(1), 1-6.
- Stuart, A., Francesc, X. (2017). Hypertension is the silent killer disease spreading across an Africa that is not ready.
- Kirkland, E. B., Heincelman, M., Bishu, K. G., Schumann, S. O., Schreiner, A., Axon, R. N., Mauldin, P. D., Moran, W. P. (2018).
- Vijver, S., Akinyi, H., Oti, S., Olajide, A., Agyemang, C., Aboderin, I., Kyobutungi, C. (2013). Status report on hypertension in Africa—consultative review for the 6th Session of the African Union Conference of Ministers of Health on NCD's. *The Pan African medical journal*, 16(38), 1-8
- Wierzejska, E., Giernaś, B., Lipiak, A., Karasiewicz, M., Cofta, M., Staszewski, R. (2020). A global perspective on the costs of hypertension: a systematic review. *Archives of medical science: AMS*, 16(5), 1078–1091.
- Yaya, S., El-Khatib, Z., Ahinkorah, B. O., Budu, E., Bishwajit, G. (2021). Prevalence and Socioeconomic Factors of Diabetes and High Blood Pressure among Women in Kenya: A Cross-Sectional Study. *Journal of epidemiology and global health*, 11(4), 397–404.
- Department of non-communicable diseases. (2014). Non Communicable diseases.
- World Health organization (2014). Blood Pressure. Global Health Observatory (GHO).