VIETNAM NATIONAL UNIVERSITY – HO CHI MINH CITY
INTERNATIONAL UNIVERSITY
SCHOOL OF COMPUTER SCIENCE AND ENGINEERING

MACHINE LEARNING-BASED PREDICTION OF CARDIOVASCULAR DISEASE RISK USING LIFESTYLE FACTORS

By
LE NGOC UYEN PHUONG
ITDSIU20079

A thesis submitted to the School of Computer Science and Engineering
in partial fulfillment of the requirements for the degree of
Bachelor of Data Science

Ho Chi Minh City, Vietnam
March, 2024

MACHINE LEARNING-BASED PREDICTION OF CARDIOVASCULAR DISEASE RISK USING LIFESTYLE FACTORS

APPROVED BY ADVISOR

_____
Assoc. Prof. Nguyen Thi Thuy Loan

APPROVED BY THESIS COMMITTEE

_____
Assoc. Prof. Nguyen Van Sinh

_____
Assoc. Prof. Huynh Kha Tu

_____
Assoc. Prof. Nguyen Thi Thuy Loan

_____
Dr. Nguyen Trung Ky

THESIS COMMITTEE
(Whichever applies)

ACKNOWLEDGMENTS

I am truly grateful for the valuable support, guidance, and encouragement extended to me by all those who have helped complete my thesis.

To begin with, I wish to express my heartfelt thanks to my thesis supervisor, Dr. Nguyen Thi Thuy Loan, for her great support, expertise, and patience. Her insightful feedback and advice have greatly guided and motivated my thesis's direction, quality, and product until now. I am grateful for her continuous willingness to share her knowledge and encourage me to continue my thesis progress.

Secondly, I am also thankful to the School of Computer Science and Engineering for providing me with a good learning environment and fundamental knowledge in Data Science to complete this thesis.

Thirdly, I am indebted to someone who voluntarily spent time sharing his feedback and support on this research. His contributions have improved the findings of my thesis.

Finally, I want to extend my special thanks to my family and close friends, who have been important parts of my time at the university. Their support and encouragement have helped me overcome obstacles during this journey and kept my motivation high.

Ho Chi Minh City, January 12th, 2024

Le Ngoc Uyen Phuong

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

Word	Explanation
CatBoost	Categorical Boosting
CDC	Centers for Disease Control and Prevention
COVID-19	Coronavirus disease of 2019
CSV	Comma Separated Values
CVD	Cardiovascular Disease
DEMO	Demographics
ID	Identifier
IDE	Integrated Development Environment
KNN	k-Nearest Neighbors
KNNImputer	k-Nearest Neighbors Imputations
MEC	Mobile Examination Center
MNAR	Missing Not at Random
NaN	Not a Number (missing value)
NCHS	National Center for Health Statistics
NHANES	National Health and Nutrition Examination Survey
NHES	National Health Examination Survey
ROC	Receiver Operating Characteristic curve
SEQN	Sequence Number
SMOTE	The Synthetic Minority Oversampling Technique
SVM	Support Vector Machine
WHO	World Health Organization
XGBoost	eXtreme Gradient Boosting

ABSTRACT

In recent times, healthcare has become one of the most significant global concerns, especially following the COVID-19 pandemic. Chronic diseases, such as cardiovascular diseases (CVDs), remain a leading cause of mortality worldwide, particularly in the United States. CVDs are usually appropriately recognized based on indicators measured in the hospital, but due to people's busy schedules, many are unwilling to spend time and effort for check-ups in the hospital. Lifestyle habits are assumed to play a role in the development of CVDs, making it crucial to modify these behaviors to predict and prevent these diseases.

Previous studies have used machine learning to predict disease risk based on various factors. However, these studies may not fully capture the dynamic relationship between lifestyle data and the presence of cardiovascular diseases. To address this, this thesis aims to build on previous findings and improve the accuracy of predicting CVDs by analyzing data from the NHANES dataset, which combines information on lifestyle habits and health indicators, using machine learning techniques such as CatBoost, Gradient Boosting, XGBoost, Linear Regression, and Support Vector Machines. The ultimate objective is to identify the relationship between lifestyle habits and CVDs, and thus reduce the number of potential patients.

Keywords: Cardiovascular Disease, Lifestyle Behaviors, Habits, Disease Prediction, Machine Learning, Health Informatics

CHAPTER 1

INTRODUCTION

1.1. Background

Healthcare has recently emerged as one of the most important global concerns, particularly in the aftermath of the COVID-19 epidemic. People are prioritizing their health more than they did in the past (Cordina et al., 2022) [1]. They are becoming more aware of their health risks, seeking new and creative ways to connect with their doctors to improve their health, and shifting their perspectives on data privacy.

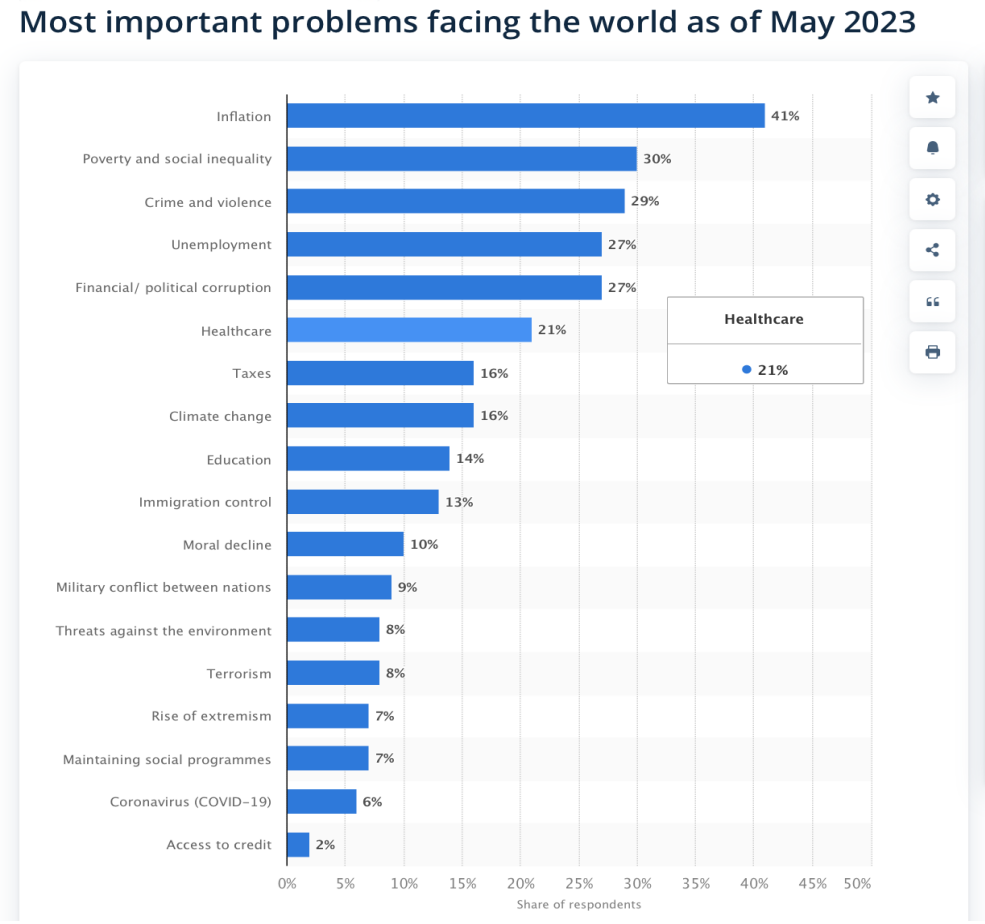


Figure 1.1. The most important problems facing the world as of May 2023 [2]

Because of the reluctance to walk outside during the COVID-19 outbreak, an increasing number of people who using technology for health tracking are willing to provide their information to get better results and measurements. As a result, new healthcare systems have been developed that track and analyze their patients' health data to suit their demands and discover disease risk factors. This approach is particularly beneficial to individuals with chronic diseases, as they often require ongoing medical attention and care from doctors or medical professionals (Betts et al., 2020) [3].

Cardiovascular diseases (CVDs), a type of chronic disease, are the leading cause of death with an estimated 17.9 million cases globally each year. CVDs contributed to over 30% of the 17 million premature deaths in 2019 [4]. As an urgent worldwide issue, people need to know about the causes and consequences of CVDs to protect their own and their family members' health.

So far, cardiovascular disease is usually detected based on indicators measured and tested in the hospital, or until the symptoms become serious enough to require hospital admissions and testing. However, as people's lives get increasingly hectic, the signs of the condition are more readily overlooked, and individuals are even reluctant to visit the hospital for check-ups due to the cost and time involved. This is a significant contributor to the high rates of cardiovascular disease-related deaths in many countries today.

Despite advances in science and medicine, the prevalence of cardiovascular disease continues to rise at an alarming rate. In recent years, specialists have pointed out that the majority of serious diseases, including CVDs, are associated with the patient's lifestyle and dietary habits. Furthermore, these disorders can be treated or prevented by lifestyle modifications and risk reduction strategies [5].

Having a risk factor does not indicate that the heart disease will develop, but the more of their appearance, the higher our chances of suffering from those types of diseases.

Therefore, we need to take action. We need to find the relationships between lifestyle habits and this type of disease to predict and reduce the number of potential patients.

1.2. Problem Statement

Cardiovascular diseases (CVDs) remain a leading global health concern, necessitating innovative approaches for early risk prediction and intervention. The current challenge lies in the inadequacy of traditional risk assessment methods, which often overlook the intricate influence of lifestyle factors. Lifestyle choices, such as diet, exercise, and sleep patterns, play a substantial role in CVD development, yet integrating these into predictive models is a complex task. Existing risk prediction models may lack granularity and fail to capture the dynamic interactions of diverse lifestyle elements. Therefore, a machine learning-based solution is essential to leverage the power of large-scale data, considering the multifaceted nature of individual lifestyles, to enhance the accuracy and personalized nature of cardiovascular risk assessments. Addressing this gap is crucial for advancing preventive healthcare strategies and minimizing the societal burden of cardiovascular diseases.

1.3. Scope and Objectives

The overall aim of this thesis is to develop appropriate machine learning models that can predict an individual's risk of cardiovascular disease based on his or her lifestyle habits data. In terms of the scope of this thesis, the objectives will be specified as follows:

1. Collect lifestyle habits and health statistics data, including physical activity levels, dietary patterns, smoking habits, and Cholesterol levels, from individuals with a known cardiovascular disease history, to create a comprehensive dataset.
2. Identify key features within the data that lead to cardiovascular disease risks among the patients.

3. Utilize machine learning techniques for data preprocessing processes, including missing value imputation.
4. Build and compare the performance of several machine learning models including CatBoost, Logistic Regression, Gradient Boosting, XGBoost, and SVM, to identify the most effective model for predicting cardiovascular disease risks based on lifestyle habits and health statistics data.
5. Develop user-friendly machine learning models that can provide accurate predictions of cardiovascular disease risks, so that people can predict their disease risks by completing some survey questions on their health status and lifestyle habits.
6. Determine the impact of individual lifestyle behaviors on the predictions generated by the machine learning models, which will provide insight into the potential for individualized therapies for reducing cardiovascular disease risk.

1.4. Assumptions and Solutions

Assumptions:

In developing a machine learning-based prediction model for cardiovascular disease (CVD) risk using lifestyle factors, several assumptions can be made. Firstly, it is assumed that the selected lifestyle factors, such as diet and physical activity, significantly contribute to the overall risk of cardiovascular diseases. This assumption is based on existing medical literature and epidemiological studies that highlight the impact of lifestyle choices on heart health. Additionally, it is assumed that the dataset used for training the machine learning model is representative of diverse populations, encompassing various demographics and geographic regions. This assumption is crucial for the model's generalizability and applicability across different communities. Furthermore, it is assumed that the features

selected for the model adequately capture the multifaceted nature of lifestyle and its relationship with cardiovascular health, considering both individual and synergistic effects.

Solutions:

To address the assumptions mentioned, a robust approach is required in the development of the machine learning-based prediction model. Firstly, an extensive literature review and expert consultations should be conducted to refine the selection of lifestyle factors and medical indicators, ensuring that the chosen variables align with the latest medical knowledge and research findings. To mitigate potential biases, the dataset used for training should be carefully curated to represent diverse populations and should include a sufficient number of samples from underrepresented groups. Employing techniques such as stratified sampling can help ensure a balanced representation. Moreover, feature engineering methods should be applied to extract relevant information from lifestyle factors, considering potential interactions and nonlinear relationships. Regular updates to the model should be scheduled to accommodate evolving scientific understanding and changing lifestyle patterns. Lastly, the model's transparency and interpretability should be prioritized to foster trust among healthcare professionals and individuals, facilitating better adoption and understanding of the predictive results in a clinical setting.

1.5. Structure of the Thesis

This Thesis is comprised of 6 chapters: Introduction, Literature Review, Methodology, Implementation & Results, Discussion & Evaluation, and Conclusion & Future work.

Chapter 1 - Introduction describes the thesis topic and its application in the healthcare industry, the motivation of the thesis, general scopes and objectives, some necessary assumptions with potential solutions for the development, as well as an overview of the subsequent chapters in this Thesis.

Chapter 2 - Literature Review identifies essential findings and research gaps after summarizing the fundamental ideas, theories, and procedures used in previous studies related to the research topic. It briefly describes how machine learning is applied to this kind of data for prediction purposes.

Chapter 3 - Methodology describes the overall research design, including the approach (e.g., qualitative, quantitative, mixed methods), data overview, data collection methods, and data processing techniques, such as statistical analysis, machine learning models, and the tools that are used to implement this thesis project.

Chapter 4 - Implementation and Results gives a detailed description of how the proposed algorithms would be implemented, followed by an analysis and comparison of the preliminary findings of each algorithm.

Chapter 5 - Discussion and Evaluation compares the results of this paper with existing literature and studies in the field, then highlights similarities, differences, or unexpected findings. This chapter also discusses any patterns or trends observed in the data and evaluates the effectiveness of applied machine-learning models.

Chapter 6 - Conclusion and Future work provides a comprehensive summary of the primary findings and the achievement of the research objectives. After acknowledging the study's limitations, it suggests potential areas for improvement or further research.

CHAPTER 2

LITERATURE REVIEW

2.1. Cardiovascular Diseases

Cardiovascular disease (CVD) [6], a non-communicable disease, is a general term that describes the disorders affecting the heart and blood vessels. There are various types of CVD diseases, such as:

- Coronary heart disease (heart attack): a condition that affects the blood vessels that supply the heart muscle due to a buildup of plaque in your coronary arteries.
- Cerebrovascular disease (stroke): a type of disease that affects the blood vessels that supply the brain.
- Congestive heart failure: a long-term condition that happens when your heart can't pump blood well enough to give your body a normal supply.
- Rheumatic heart disease: a condition that damages the heart muscle and heart valves due to rheumatic fever, which is caused by streptococcal bacteria.
- Peripheral arterial disease: a type of disease that affects the blood vessels that supply the arms and legs.
- Congenital heart disease: a type of birth defect that affects the normal function of the heart due to malformations in its structure from birth.
- Deep vein thrombosis and pulmonary embolism: a condition that involves blood clots in the leg veins that can dislodge and travel to the heart and lungs.

The two most prevalent forms of cardiovascular disease, heart attacks, and strokes, are typically triggered by blockages that restrict blood flow to the human heart or brain. The primary cause of these blockages is the fatty deposits along the inner walls of the blood vessels that provide oxygen and nutrients to these vital organs.

The symptoms of diseases related to heart and blood vessels are usually not clear or do not exist. A heart attack, stroke, or any of the CVD disease types may be the first symptom of underlying diseases. Therefore, people should be aware of some earlier signs of a potential heart attack or a stroke, such as:

Table 2.1. Heart attack and stroke symptoms [6]

Heart attack	**Stroke**
• Pains or discomforts in the center of the chest. • Pains or discomforts in the arms, the left shoulder, elbows, jaw, or back. • Difficulty in breathing or shortness of breath. • Nausea or vomiting. • Light-headedness or faintness. • A cold sweat.	• Sudden loss of strength in the face, arm, or leg, typically on one side of the body. • Numbness of the face, arm, or leg, especially on one side of the body. • Confusion or difficulty in speaking or understanding speech. • Difficulty seeing with one or both eyes. • Severe headache without a known cause. • Fainting or unconsciousness.

Cardiovascular diseases (CVDs) remain a major cause of death worldwide each year. In 2019, the World Health Organization (WHO) reported that CVD caused approximately 17.9 million people to die, with heart attacks and strokes accounting for 85% of these deaths. Over 75% of these deaths occur in low- and middle-income countries, and CVDs are responsible for over 30% of premature deaths (those under the age of 70) in 2019 [6]. Because of this, people need to know about the causes and effects of CVDs to protect themselves and others' health in their families.

Most of the causes that lead to cardiovascular diseases are from the patient's lifestyle habits, which can then be called behavioral risk factors. Lifestyle habits concentrate on the activities we do, the food we eat, the types of drinks we consume, or other related habits from the beginning until the end of the day. Therefore, behavioral risk factors such as

lifestyle behaviors such as diet, physical activity, smoking, and stress levels can significantly impact the development and progression of CVD.

Other forms of risk factors include:

- Non-modifiable risk factors, which are the factors that we are unable to change, such as age, ethnic background, genes, and gender.
- Physiological variables include high blood pressure (hypertension), high cholesterol, and excessive blood sugar or glucose levels.

The relationships between lifestyle habits and this type of disease should be early identified to forecast and reduce the number of potential patients.

2.2. Disease Prediction

In the field of healthcare, the ability to predict human diseases plays a critical role in healthcare, enabling doctors to provide more effective treatment to patients. Medical professionals base their diagnosis on an analysis of the patient's symptoms and medical history to evaluate the likelihood of a particular condition or disease and recommend appropriate treatment options. By keeping track of the patient's health history, doctors can provide personalized care that is customized to the patient's individual health needs, leading to improved treatment outcomes [7]. In today's world, advancements in science and medicine have made it possible to identify and treat many diseases and illnesses that were once considered incurable. However, it is still essential for individuals to be proactive about their health and take steps to early maintain good health and prevent serious illnesses.

Disease prediction is a classification problem that involves predicting whether an individual has a disease or not or classifying diseases into different categories. Therefore, supervised machine learning techniques have demonstrated remarkable potential in predicting and preventing diseases by analyzing health data. The accuracy of detecting diseases based on symptoms has been assessed using various algorithms, including Support

Vector Machine (SVM), K-nearest neighbors (KNN), and Logistic Regression. Extensive research in the field of medical science has led to the successful implementation of machine learning models for predicting various common diseases. These models have been found to be particularly effective in classifying patients with cardiovascular disease (CVD) among individuals with diabetes [8]. By analyzing large volumes of patient data, these algorithms can identify patterns and make predictions with high accuracy, leading to earlier diagnosis and more effective treatment plans.

The research titled "A data-driven approach to predicting diabetes and cardiovascular disease with machine learning", which has been authored by Dinh et al., presents a comprehensive evaluation of the performance of various supervised machine learning models in classifying individuals with diabetes and cardiovascular disease in the United States. The study utilizes the National Health and Nutrition Examination Survey (NHANES) dataset from 2007 to 2014, which consists of survey data and lab indicators. The objective is to identify data-driven techniques that employ supervised machine learning models to detect individuals at risk while determining the essential data factors that contribute to such disorders. From the findings of this and previous research on CVDs, the key contributors to the diseases can be listed in Figure 2.1, including age, systolic blood pressure, self-reported weight, occurrence of chest pain, and diastolic blood pressure.

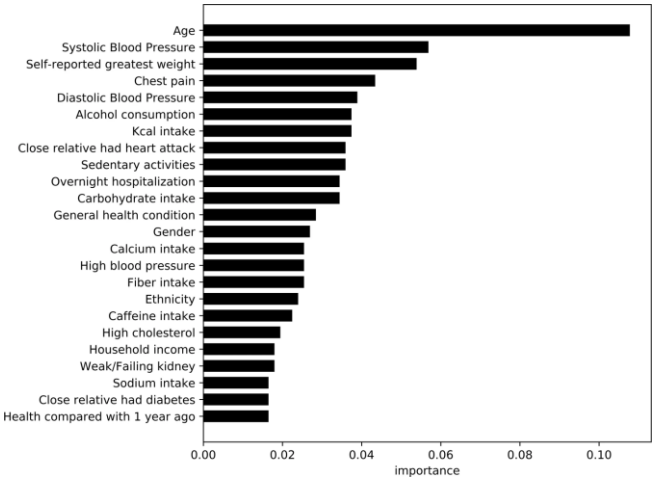


Figure 2.1. Feature importance for cardiovascular disease classifier without lab results

CHAPTER 3

METHODOLOGY

3.1. Workflow

In this part, the workflow of data analysis steps of the thesis will be shown.

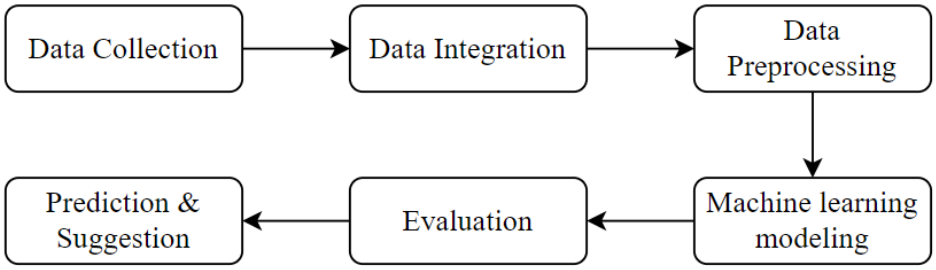


Figure 3.1. Data Analysis Workflow

1. **Data collection**: Collect relevant data from various data files, ensuring it includes the necessary features and information for machine learning model prediction, particularly the age, body measurements, and lifestyle habits of people with and without a history of cardiovascular disease.
2. **Data integration**: Combine and merge the data obtained from the previous step to generate a single dataset that provides a full view of the information, since the datasets are segregated into various files for each type of information.
3. **Data preprocessing**: Clean and transform the integrated data to handle missing values, ambiguous values, imbalanced labels, and label assignments to ensure uniformity and make it appropriate for machine learning algorithms.
4. **Machine learning modeling:** Divide the preprocessed dataset into the train, test, and validation sets to develop models on the train data set and evaluate their performance on unseen data.

5. **Result evaluation:** Assess the model's performance metrics, such as accuracy, to see how well it predicts outcomes relative to the ground truth.
6. **Prediction & Suggestion:** Utilize the trained models to make predictions on new data, delivering insights and suggestions based on the learned patterns.

3.2. Data Collection

3.2.1. About NHANES

The Centers for Disease Control and Prevention (CDC) [9-10] is the national leading science-based and data-driven service organization dedicated to protecting public health in the United States. For over 70 years, they have developed a wealth of information and resources that individuals, businesses, and communities require to promote good health, prevent disease, manage injuries and disabilities, and prepare for potential health threats that may arise in the future.

The National Health and Nutrition Examination Survey (NHANES) [11] survey program, administered by the National Center for Health Statistics (NCHS) [12], a division of the CDC, is a crucial tool that assesses the health and nutritional status of individuals in the United States. The data collected by NHANES is used to generate vital health statistics that inform doctors, patients, researchers, and policymakers about the prevalence of major diseases and risk factors, as well as the association between nutrition status and overall health problems in the United States [13]. Moreover, NHANES findings are instrumental in the development of public health policies, programs, and services, as well as raising awareness about national health concerns, such as "What percent of adults in the United States have high blood pressure, high cholesterol, or diabetes?" [14].

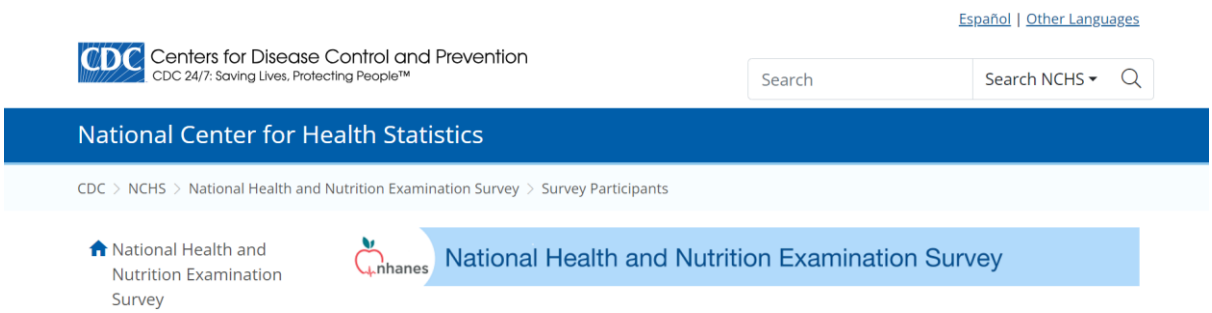


Figure 3.2. NHANES website

The NHANES program sets itself apart by utilizing both interview data and health indicators gathered from physical examinations, which means that the data can not only be collected from medical tests but also be gathered from surveys or questionnaire answers. The NHANES program also excels at collecting and examining the risk factors associated with certain diseases or conditions. Factors such as smoking, alcohol consumption, drug use, physical fitness and activity, weight, and dietary intake are all carefully studied.

Following are some of the diseases, medical conditions, and health indicators that NHANES studied from the participants' data [11]:

- Cardiovascular disease
- Diabetes
- Eye diseases
- Hearing loss
- Infectious diseases
- Kidney disease
- Nutrition
- Obesity
- Oral health
- Physical fitness and physical functioning
- Reproductive history and sexual behavior
- Respiratory disease (asthma, chronic bronchitis, emphysema)
- Vision

Beginning in the early 1960s, the NHANES program has been conducted as a sequence of surveys, each of which targets specific population groups or health-related topics [15].

- NHES I, II, and III (1959–1970)
- NHANES I, II, and III (1971 - 1994)

- Continuous NHANES (1999–present)

Starting in 1999, the survey evolved into a continuous program that adapts to changing needs by focusing on a variety of health and nutrition measurements. Since then, the survey has examined a nationally representative sample of about 5,000 participants of all ages in 15 different counties across the United States each year. Each participant plays a crucial role in the study, representing approximately 65,000 others in the country who share similar characteristics. In this thesis, data from the Continuous NHANES (2011-2020) are gathered as they are close to the patient's health condition today.

NHANES data are accurate and reliable since they were carefully gathered and analyzed by the organization's professionals. The participants are interviewed at home and then complete the survey's health assessment at a Mobile Examination Center (MEC), which provides an ideal environment for collecting high-quality data. The data at the MEC can be efficiently transmitted into the central databases with the aid of digital scales and stadiometers. The examination component comprises examinations of the patient's physical, mental, and dental health, as well as laboratory tests performed by highly qualified healthcare professionals, including doctors, dentists, physicians, as well as medical and health technicians. Each of them has received component-specific training and assessments to ensure that they understood the standard NHANES protocol for every examination they conducted [11]. In addition, all data collected in NHANES surveys are kept strictly confidential and are protected by public laws.

3.2.2. Data details

Since 1999, data has been generated in two-year cycles, which is known as the current NHANES or Continuous NHANES. This approach involves dividing each cycle into five distinct categories labeled by collection method: Demographics, Examination, Laboratory, Questionnaire, and Dietary [16].

- The **Demographics** file contains data at the individual, family, and household levels on a variety of issues, including age, weight, pregnancy status, household and family size, income, individual education level, marital status, military service status, country of birth, citizenship, and length of time spent living in the United States.
- **Examination** files contain data gathered during physical examinations, including audiometry, blood pressure, body measures, muscular strength, oral health, vision exam, etc. Additionally, depending on the cycle, specialty exams like dental and eye exams may also be performed.
- **The Laboratory** category focuses on obtaining biological samples from participants, such as blood, urine, hair, air, tuberculosis skin test, HIV, heavy metals, plasma glucose, total cholesterol, triglycerides, etc. The blood chemistry, cholesterol levels, hormone levels, food levels, and markers of diseases or ailments are among the health-related metrics that can be obtained from these samples when they are tested in a laboratory setting.
- **Questionnaire** files contain data collected from interviews conducted at homes and Mobile Examination Centers (MEC). These interviews go across a wide range of health-related topics, including medical history, health behaviors (such as smoking habit, alcohol use, and physical activity), medication usage, mental health, access to healthcare, and other relevant factors.
- **Dietary** files involve collecting data on participants' dietary intake, including details on the kinds and amounts of food and beverages consumed. It includes 24-hour dietary recall interviews carried out by professional interviewers to capture detailed information about participants' food consumption.

The picture below depicts the data structure used for each survey cycle of the Continuous NHANES, which includes the five primary components indicated above as well as some examples of the data files that comprise each primary component.

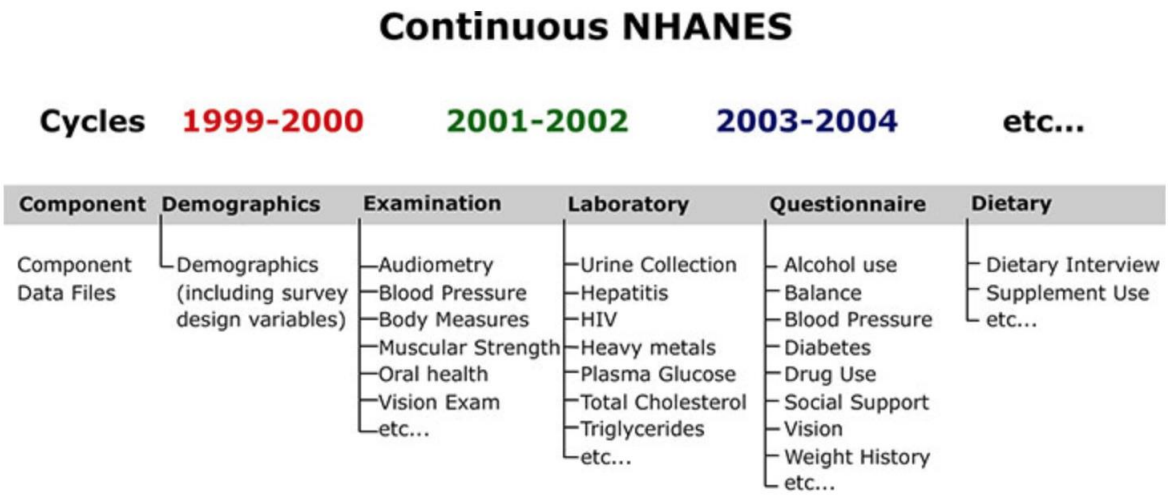


Figure 3.3. Continuous NHANES Cycles and Primary components [16]

The information regarding a person's daily dietary intake, physical activity levels, usage of alcohol use, blood pressure, smoking habits, etc., which is primarily accessible through Dietary, Examination, and Questionnaire data files, can be employed in this thesis to describe the individual's lifestyle habits.

3.2.3. Advantages & Limitations

On the one hand, the most significant advantage of the NHANES data is that it provides a variety of participant information, from lifestyle behaviors, and dietary status, to medical issues and disease history, all of which are valuable data for this thesis research and are difficult to be found in other data sources. These data have been then utilized to achieve various interesting objectives, one of which is to investigate the relationship between diet, nutrition, and health. These are some of the objectives of NHANES data:

- Determine the prevalence and risk factors of major diseases.
- Assess nutritional status and its association with disease prevention.
- Help epidemiological studies and health sciences research.

- Find out the link between high cholesterol levels & heart disease (1960s).

Furthermore, the NHANES provides data analysts with a variety of tutorials and recommendations [17-18] for analysis methodologies, as well as guidelines on how to merge datasets of each two-year survey data, how to deal with missing values, how to construct the appropriate weight in making estimates, and how to properly calculate the variance for subgroups of interest.

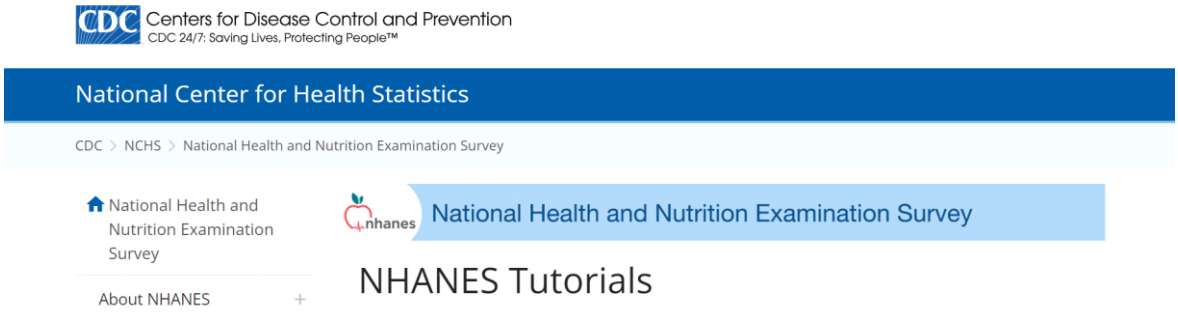


Figure 3.4. NHANES Tutorial Resources

On the other hand, this NHANES dataset has numerous drawbacks. Since there is so much information, or features, for each participant, there are so many columns in each category data file that it is difficult to manage all of them, and they must be chosen for further research. For example, the demographics file from cycle 2013-2014 contains 47 columns, while the others contain more than 100. Besides, the quantity of data points, or rows, in each data file, is not the same, and some data files contain many null values that must be addressed before training prediction models.

3.2.4. Data files downloads and imports

NHANES data is widely accessible to the public and can be found on the Survey Cycle's Questionnaires, Datasets, and Related Documentation page on the official NHANES website. Not only are the historical and current data files released publicly on the website, but detailed descriptions and instructions to use in data analytics tasks are also included.

NHANES data are saved in SAS transport (.XPT) files. SAS transport files can be extracted using a variety of software packages, including SUDAAN, SPSS, Stata, and R. In this thesis project, Python can also be used to import and read all these SAS transport files downloaded from the data sources to generate the analytic dataset for the project.

Python Pandas [19] can read two SAS file formats, which are SAS xports (.XPT) and SAS data files (.sas7bdat), using pandas.read_sas(), which is as simple as reading Comma Separated Values (.csv) files in Python.

3.3. Data Preprocessing

Due to the complexity of the NHANES dataset, which contains multiple features and inherent problems, an extensive preprocessing process is required to transform it into a clean dataset suitable for modeling steps. Figure 3.5 provides a detailed overview of the necessary steps involved in this stage, which are critical to ensure the accuracy and reliability of the modeling results.

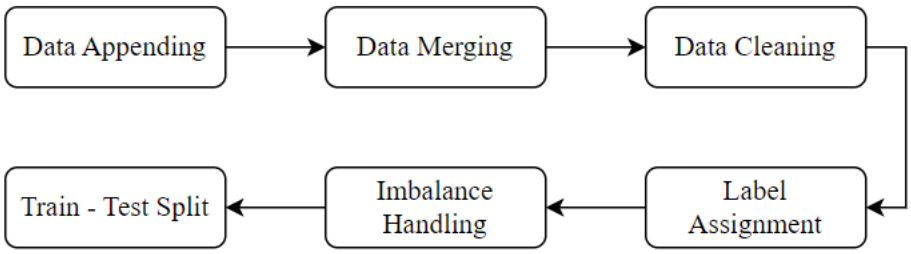


Figure 3.5. Data Preprocessing Process

3.3.1. Dataset appending

The dataset used in this paper will typically include data from 2011 to 2020, corresponding to five survey cycles: 2011-2012, 2013-2014, 2015-2016, 2017-2018, and 2017-2020. The data acquired after each survey cycle is recorded in different files, so a data appending procedure should be implemented to combine the years of data to create a

complete dataset from 2011 to 2020. The data files' contents should be reviewed before appending because the shape and variable names may change from cycle to cycle, and recorded or derived variables may be added in various cycles. For example, some features in this cycle will not exist in the next cycle, and vice versa.

The SEQN is an important column that should be included in appended NHANES datasets. SEQN, or Sequence Number, is a short eight-character unique identifier (ID) assigned for each observation (participant) in NHANES to allow the SAS transport files to be compatible with the SAS Statistical Package and other relevant tools.

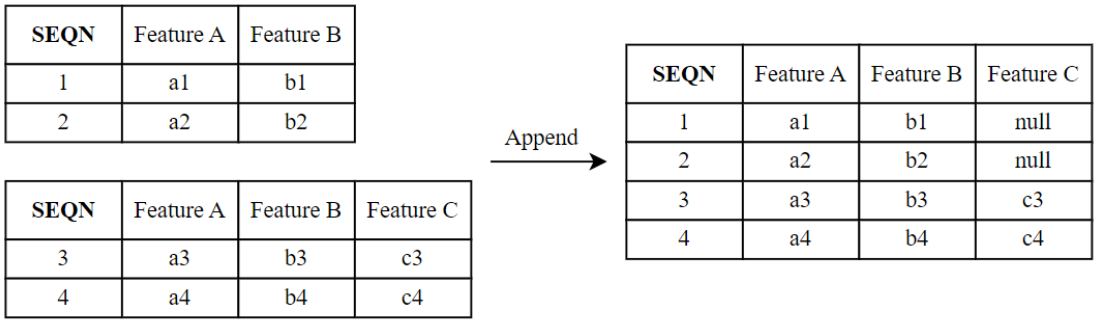


Figure 3.6. Dataset appending example

3.3.2. Dataset merging

Data files for each data cycle are grouped by the method of collection, which may be classified into one of five categories: Demographics, Examination, Laboratory, Questionnaire, and Dietary. The analysis dataset needs variables from more than one component. As a result, the selected data files must be combined to include variables from several components.

The process of consolidating individual component data files into one comprehensive dataset is called merging. The initial step involves organizing each data file by a distinct identifier. In the case of NHANES data, the exclusive sequence number (SEQN) serves as the unique identifier for each sample participant. Most NHANES data files have only one entry for each sample person who took part in that specific component. However, due to

various restrictions such as age, gender, etc., not all sample participants took part in every component. For instance, the Demographic Variables and Sample Weights (DEMO) file has a single record for each sample person, while the Body Measures (BMX) file has one record for each person who took the MEC test. Since SEQN is the key variable that identifies each participant in these files, it is necessary to use it as the merging variable for the dataset.

All participants must provide their information in the Demographics component so that all participants in a survey cycle may be found in that cycle's Demographic Variables and Sample Weights (DEMO) file.

In this thesis, most of the important features are in the Questionnaire component datasets, so the final dataset for analytics will be merged on the sequence number (SEQN) of participants in the Questionnaire file.

While SEQN is the unique identifier for the majority of NHANES data files, it may not be unique in others since each sample individual may have numerous records in those files. Following are some examples of data files using this multi-record structure [16].

- Prescription Medications (RXQ_RX),
- Dietary Interview - Total Nutrient Intakes:
- First Day (DR1TOT) and
- Second Day (DR2TOT)
- Dietary Interview - Individual Foods:
- First Day (DR1IFF) and
- Second Day (DR2IFF)
- Dietary Supplement Use - Total Dietary Supplements:
- First Day of 24-hour recall (DS1TOT)
- Second Day of 24-Hour Recall (DS2TOT)
- 30-Day (DSQTOT)

- Dietary Supplement Use - Individual Dietary Supplements:
- First Day of 24-Hour Recall (DS1IDS)
- Second Day of 24-Hour Recall (DS2IDS)
- 30-Day (DSQIDS)

As a result, SEQN must be carefully considered before merging files. For example, as instructed by NHANES, the data files with multi-record structure would need to be converted from the detailed drug-level file into a person-level file (with a single record for every person) before being merged with NHANES Demographics and other data files, utilizing SEQN as the unique identifier. Moreover, the record counts should be checked after each merging step to ensure that everything is in order.

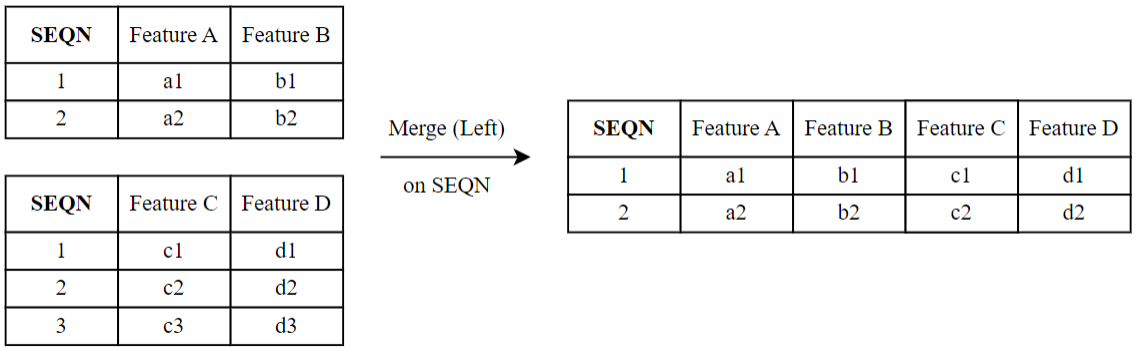


Figure 3.7. Dataset merging example

3.3.3. Data cleaning

Each data file only contains the records of survey participants who were qualified for inclusion in that component. People who rejected or did not have enough time to take the survey would have their responses recorded as missing values. Missing data refers to values or data that are not recorded (or do not exist) for some variable(s) in a dataset [20], and the absence of it can cause. Missing values in NHANES are assigned as in the Table below.

Table 3.1. Missing values code in NHANES data [16]

NHANES codes	Description	Action
. (period)	missing numeric value	None
(blank)	missing character value	None
7 or 77 or 777 or 7777 or 77777	"refused" response	Code as missing (period or blank)
9 or 99 or 999 or 9999 or 99999	"don't know" response	Code as missing (period or blank)

According to Table, the value of "refused" and "don't know" responses in various analytic dataset features must be transformed to missing values, otherwise they may result in inaccurate conclusions. The values of the two answer types are typically found together in the same feature (column) of a dataset, hence those features just need to be identified and the assigned values of both responses can be easily replaced with NaN values.

BPQ020 - Ever told you had high blood pressure

Variable Name: BPQ020

SAS Label: Ever told you had high blood pressure

English Text: {Have you/Has SP} ever been told by a doctor or other health professional that {you/s/he} had hypertension, also called high blood pressure?

Target: Both males and females 16 YEARS - 150 YEARS

Code or Value	Value Description	Count	Cumulative	Skip to Item
1	Yes	2174	2174	
2	No	4285	6459	BPQ056
7	Refused	0	6459	BPQ056
9	Don't know	5	6464	BPQ056
.	Missing	0	6464	

Figure 3.8. Missing values in a feature [21]

Missing values might affect analytic findings. Thus, their extent in the datasets must be assessed to decide whether the data can be used without extra re-weighting for item non-response. If less than 10% of the primary outcome variable data is missing for a component, it can be safely ignored, and analysis can continue as normal. However, if more than 10% of

the data is missing, it's essential to assess respondents and non-respondents to determine if imputation of missing values or adjusted weights are necessary [16]. This helps maintain data integrity and reliable analysis.

If the codebook and frequency counts indicate that an item has multiple "true" missing values, such as a period for a missing numeric value or a blank for a missing character value, it's important to review the documentation of the data file. This will help determine why that item was not evaluated for those respondents. It's possible that some items or features have a limited target gender or age range compared to other items in the data file, or they may be part of a skip pattern, which means that certain respondents are not eligible to participate in or respond to these variables.

It is essential to note that NHANES warns against dropping any records from the analysis dataset, even for individuals with missing values for a variable of interest [16]. Therefore, it's crucial to review the missing values and consider imputing or adjusting weights as needed.

The NHANES dataset utilized for this thesis still has several features, with more than 90% of the data missing values. Those features will be removed because they cannot give any information with such a small dataset, and the large number of missing values may cause bias in the prediction models. After that, the missing values imputation method may be applied to the remaining columns.

In data science, missing values can be imputed by various methods. They can be replaced by an arbitrary value, such as the column's mean, mode, or median. Because the features in this dataset are associated with one another in varying degrees, the SimpleImputer and KNNImputer imputation methods were used to fill in missing values depending on their relationship or correlation.

scikit-learn offers SimpleImputer [20] as an approach to fill in the missing values using mean or median. It creates a simple model that takes into consideration a single feature and the missing value in that feature can be filled in using the mean or median of that feature.

k-Nearest Neighbors Imputations (KNNImputer) by scikit-learn [22] is an imputation method that implements the k-Nearest Neighbors (KNN) approach to fill out or predict missing values in a dataset based on the correlation between the imputed column and the others. With "k" being the number of nearest neighbors with the row that requires imputing, this technique identifies the k neighboring points using a Euclidean distance metric. Then it takes the average or mode of their complete values to impute missing values.

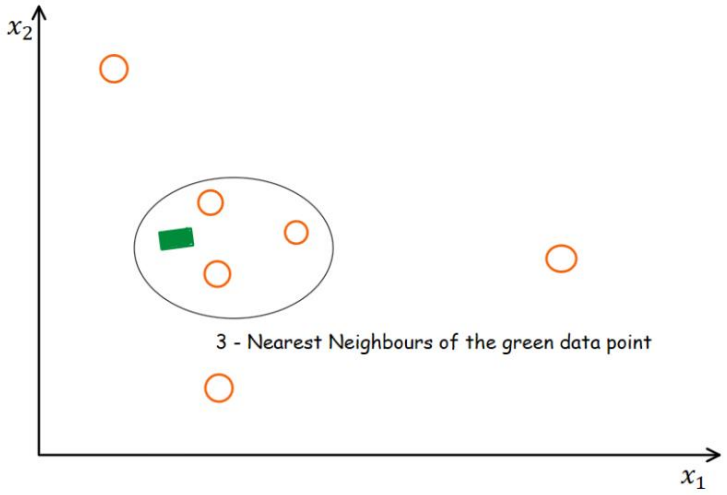


Figure 3.9. KNN algorithm approach (k=3) [23]

In the case of missing coordinates, the Euclidean distance is calculated by scaling up the weight of the non-missing coordinates and excluding the missing values [23]. The formula can be as follows.

$$d(x,y) = \sqrt{weight * \sum_{i=1}^{n}(x_i - y_i)^2} \quad (1)$$

where,

$$weight = \frac{Total\ number\ of\ coordinates}{Number\ of\ present\ coordinates} \quad (2)$$

When determining the nearest neighbors for the row that has to be imputed, KNNImputer considers all the features supplied to it. As a result, it is especially effective when dealing with Missing Not at Random (MNAR) values, and it can lead to more accurate and trustworthy machine-learning algorithms. The following is an example of how KNNImputer can fill in missing values.

```
from sklearn.impute import KNNImputer
impute_knn = KNNImputer(n_neighbors = 2)
df_test = impute_knn.fit_transform(df2)
```

Age	Cholesterol	Blood pressure	Hypertension
22.0	168.0	82.0	2.0
44.0	196.0	56.0	2.0
80.0	138.0	64.0	1.0
50.0	187.0	74.0	1.0
75.0	148.0	80.0	1.0
42.0	240.0	88.0	1.0
41.0	NaN	70.0	2.0

	Age	Cholesterol	Blood pressure	Hypertension
0	22.0	168.0	82.0	2.0
1	44.0	196.0	56.0	2.0
2	80.0	138.0	64.0	1.0
3	50.0	187.0	74.0	1.0
4	75.0	148.0	80.0	1.0
5	42.0	240.0	88.0	1.0
6	41.0	191.5	70.0	2.0

Figure 3.10. Example of KNNImputer

This is a sample of the thesis' analytics dataset. Let the last row be the one with a missing value in the "Cholesterol" level that has to be filled in. In the above example, the n_neighbors parameter is set to 2, so KNNImputer will identify the two most similar rows based on how near the "Blood Pressure" and "Hypertension" values are to this row. The distance between rows with indices 6 and 1 is calculated using the Euclidean distance formula described above.

$$d(6,1) = \sqrt{\frac{4}{4} * [(41-44)^2 + (70-56)^2 + (2-2)^2} = 14.3178 \qquad (3)$$

Table 3.2. Euclidean distance between the row with missing value (index = 6) and others

d	**0**	**1**	**2**	**3**	**4**	**5**	**6**
6	22.4722	14.3178	39.4715	9.8995	35.4542	18.0555	0

The rows with indexes 1 and 3 were found to have the closest values for the two features, with the smallest distance values of 9.8995 and 14.3178 respectively in Table 3.2. As a result, the imputed value is calculated as the average of the "Cholesterol" values from these two rows.

$$Cholesterol_6 = \frac{(Cholesterol_1 + Cholesterol_3)}{2} = \frac{(196+187)}{2} = 191.5 \qquad (4)$$

3.3.4. Label assignment

One of the significant drawbacks of the dataset used for this thesis is that it needs a label feature. The problem of cardiovascular disease prediction requires a label for supervised learning models in the classification problem. As a result, the label feature must be developed or assigned based on the information provided by each participant.

Based on the primary symptoms of cardiovascular disease [24] and the success of the previous study work (Dinh et al., 2019), the label feature can be assigned as follows: If the participant answers "Yes" (value = 1) to at least one of these four questions on cardiovascular symptoms or conditions, then the subject will be classified as having the disease (label=1). Otherwise, if the responses are "No" to all four questions, the participant will be categorized as not having the disease (label=0).

- Variable “MCQ160B”: Has a doctor or any other health professional ever diagnosed you with congestive heart failure?
- Variable “MCQ160C”: Has a doctor or other health professional ever told you that you had coronary (kor-o-nare-ee) heart disease?
- Variable “MCQ160E”: Has a doctor or other health professional ever told you that you had a heart attack (also called a myocardial infarction (my-o-car-dee-al in-fark-shun))?
- Variable “MCQ160F”: Has a doctor or other health professional ever told you that you had a stroke?

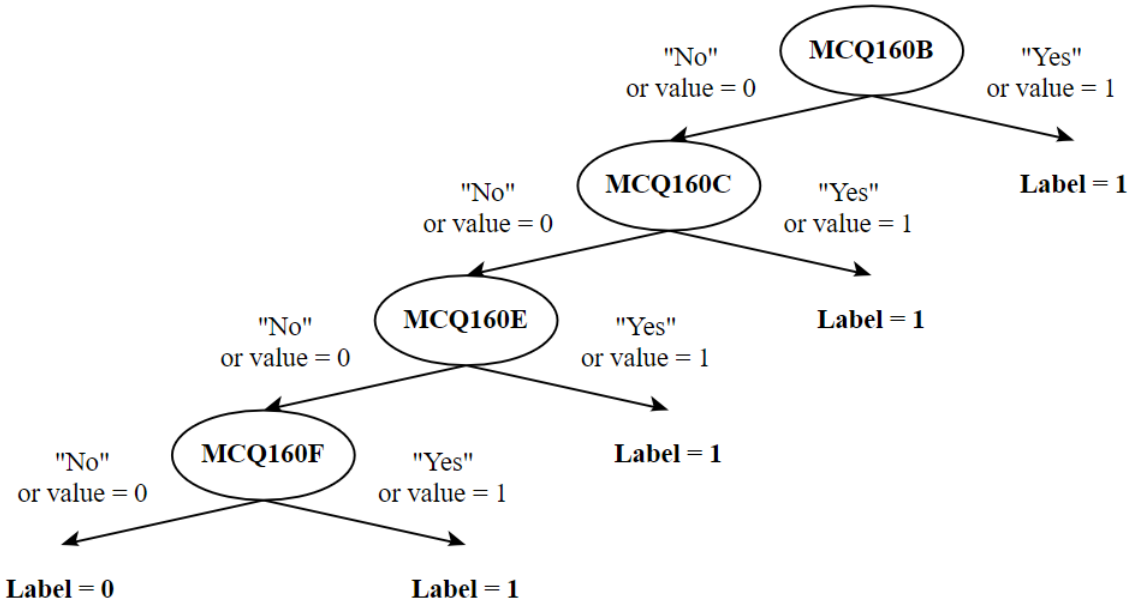


Figure 3.11. Label Assignment process

3.3.5. Imbalanced data handling

The observations in the dataset are not evenly distributed, or imbalanced, among the class labels, resulting in a significant number of observations for one label and fewer for the other. This can lead to bias in our classifier's predictions. To address this issue, we can use resampling techniques. These techniques involve either reducing the samples from the majority class (under-sampling) or adding more instances from the minority class (over-sampling) to balance out the dataset.

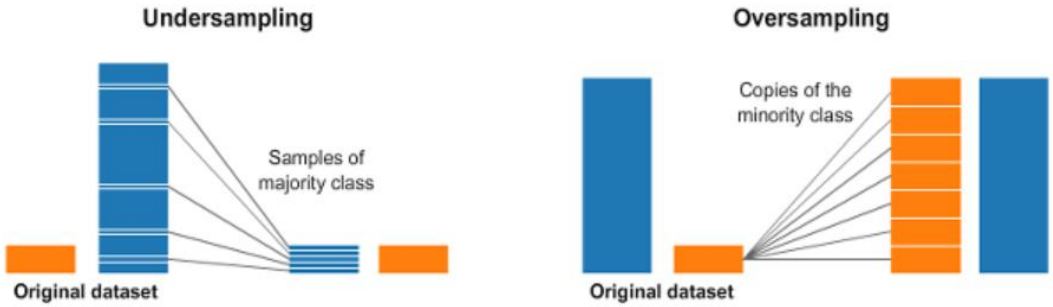


Figure 3.12. Resampling Techniques to Solve Class Imbalance [25]

Undersampling involves removing some observations of the majority class until both labels are balanced. However, a drawback to this technique is that it may remove some

valuable information. On the other hand, oversampling is used when there is insufficient data. It balances the dataset by increasing the size of rare samples using repetition, bootstrapping, or SMOTE. The Synthetic Minority Oversampling Technique (SMOTE) addresses imbalanced datasets by creating new instances for the minority class using k-nearest neighbors [26]. This method can lead to biased machine learning models due to the use of artificially generated data. Therefore, to address the issue of imbalanced labeling and prioritize the inclusion of real data in the modeling stages, this thesis exclusively utilized undersampling. It is important to note that fake data was not included in the test sets.

Figure 3.13 shows imbalanced labels in the dataset. There are 28,286 observations labeled as "0", while only 3,333 observations are labeled as "1", which means having cardiovascular diseases.

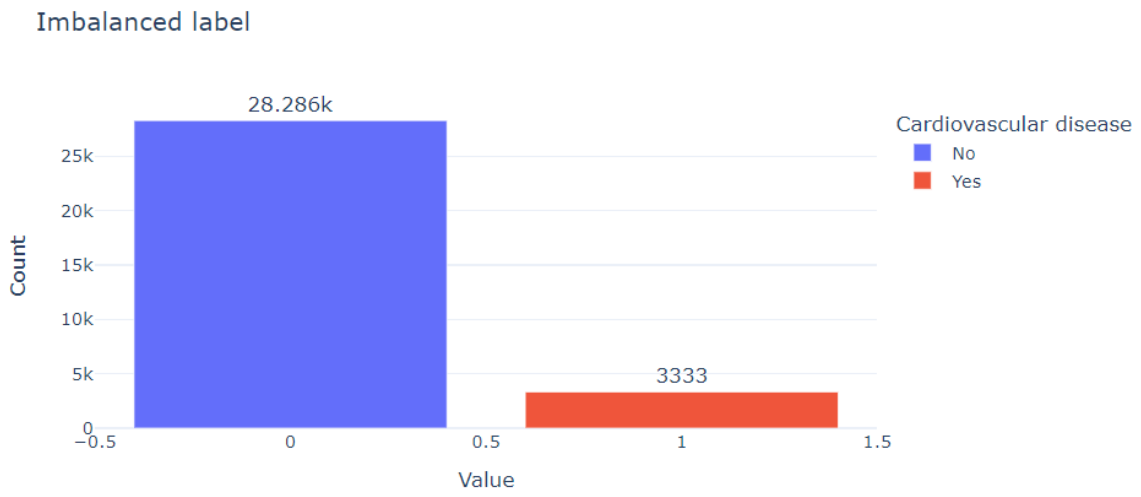


Figure 3.13. Imbalanced label

To achieve a more balanced dataset, the majority label with a value of "0" was under-sampled by randomly selecting 6,000 data points out of 28,286. Meanwhile, the minority label with a value of "1" remained unchanged at a total of 3,333 data points. As a result, the labels became much more balanced, with the difference between them not being so significant as before. It is worth noting that this approach resulted in the loss of 22,286 data points, which could contain valuable insights. However, after trying to keep the number of

the majority class as 3,333, 4,000, or 5,000, the way of undersampling 6,000 data points of label “0” gave better results as it lost fewer data. This technique can ensure that all data are real. As a result, the remaining data, which represents about 30% of the original dataset, can still provide a comprehensive overview of the entire dataset because of the distinct patterns of the data points.

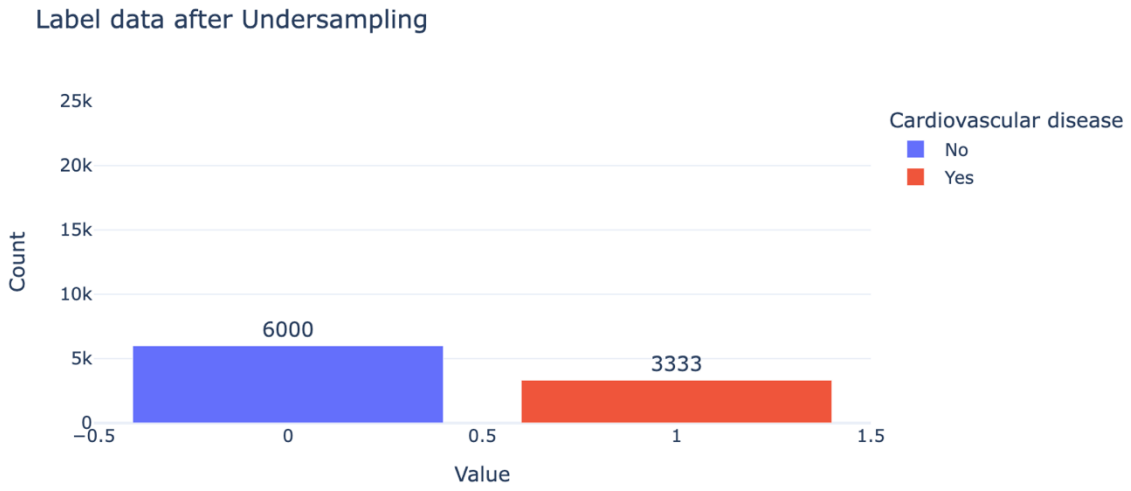


Figure 3.14. Imbalanced labels after Undersampling

3.4. Machine Learning Models

Disease risk prediction is a classification problem that has a discrete label variable categorized into “Yes” or “No”, which expresses the risk of developing the disease for an individual [27]. In this type of problem, various supervised learning models are utilized to define the association between the features and the disease presence to identify cardiovascular disease risk factors, as well as to predict the probability of potential disease patients based on their data (e.g., age, weight, lifestyle, eating habits, etc.).

3.4.1. Gradient Boosting

The machine learning boosting algorithm [28] is a method for building an ensemble, which is an iteratively built combination of simple individual models that work together to generate a more powerful new model. It begins with creating a model (e.g., a tree or linear

regression) on the training dataset, followed by the creation of a second model to correct the error in the first one. The primary idea behind this approach is to create models sequentially, with succeeding models attempting to reduce the errors of the prior model and minimizing the loss function using gradient descent.

Gradient Boosting [28] is a machine learning boosting approach that iteratively combines the predictions of many weak learners, usually decision trees. Each subsequent tree improves the performance of the previous tree, which leads to better results by minimizing the loss function using gradient descent and improving the model's accuracy. This is why the name Gradient Boosting arises. Gradient Boosting is notable for its prediction speed and accuracy, particularly on big and complicated datasets. Because this thesis is a classification problem with a discrete target feature, Gradient Boosting Classifiers should be used.

The Gradient Boosting Tree consists of N trees. Tree 1 is trained using the feature matrix X and labels y. The predicted labels $\widehat{y_1}$ are used to determine the training set residual errors r_1.

$$r_1 = y_1 - \widehat{y_1} \tag{5}$$

The second tree (Tree 2) is then trained with the feature matrix X and the residual errors r_1 of Tree 1 as labels. The predicted outcomes $\widehat{r_1}$ are then utilized to calculate the residual errors r_2. The procedure is repeated until all the N trees in the ensemble are trained. Upon completion of training all trees, the ultimate prediction can be made using the formula provided below.

$$y(pred) = y_1 + (\eta * r_1) + (\eta * r_2) + \ldots + (\eta * r_N) \tag{6}$$

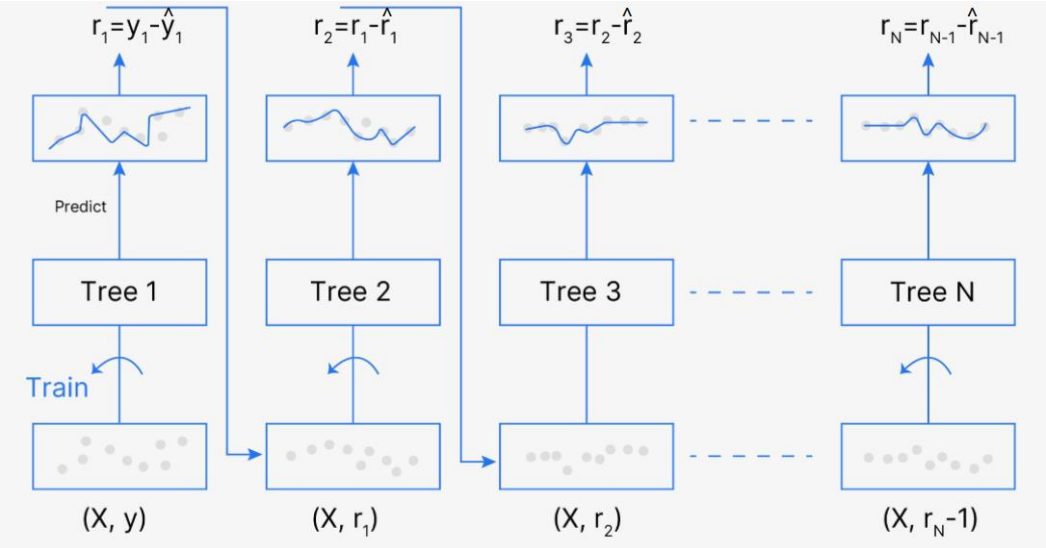


Figure 3.15. Gradient Boosting Tree [28]

3.4.2. CatBoost

CatBoost, or Categorical Boosting [29], is a supervised machine learning method that implements Gradient Boosting on Decision Tree. It is an open-source algorithm that is developed by the researchers and engineers of Yandex, a Russian technology company that builds intelligent products and services powered by machine learning. As for its name, CatBoost has two main features: it works with categorical data (the “Cat”) and it uses Gradient Boosting (the “Boost”). Like other Boosting algorithms, CatBoost combines numerous decision trees in the background, known as an ensemble of trees, to predict a classification label. This algorithm can add trees that have been trained to correct the mistakes made by previous trees while minimizing a differentiable loss function. This repeated strategy gradually improves the model's prediction performance.

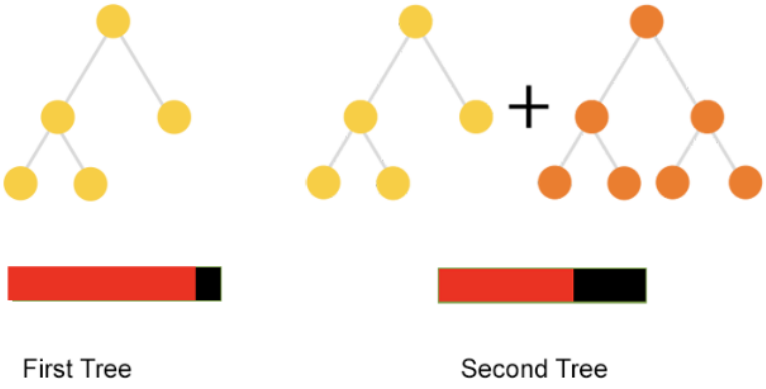


Figure 3.16. CatBoost's first and second trees [30]

Most of the existing machine learning models or decision tree-based approaches require categorical values in the training dataset to be converted to numerical values. This may take some time and effort as the encoding procedure involves many steps and it extends the dataset with various new columns. CatBoost may overcome this limitation by dealing with both category and numerical data without the requirement for preprocessing. The algorithm includes pre-processing data, which encodes categorical variables using Ordered encoding. The ordered encoding uses the target statistics from all rows preceding a data point to produce a value to replace the categorical feature. In addition, CatBoost leverages the Symmetric Weighted Quantile Sketch (SWQS) technique which effectively handles missing values in each dataset to reduce overfitting and improve the overall performance [30].

The building of a baseline model is required to obtain an understanding of the complexity of the problem and dataset, as well as to provide a performance baseline for developing more complicated models. The baseline model is usually a simple build without any heavy data cleaning and preprocessing. Therefore, CatBoost is used as a baseline model for this thesis project.

3.4.3. XGBoost

XGBoost, or eXtreme Gradient Boosting [31], is an efficient distributed Gradient Boosting method that uses parallel tree boosting (also known as GBDT, GBM) to train

models on big datasets in a reasonable time. It inherits the characteristics of the Gradient Boosting algorithm, which creates a predictive model by iteratively combining the predictions of multiple individual models, typically decision trees, to minimize a predefined loss function during training using a gradient descent optimization technique. Known for its computational efficiency, feature importance analysis, and handling of missing values, XGBoost is widely used for tasks such as regression, classification, and ranking. XGBoost, which is known for its computational speed, feature significance analysis, and management of missing information, is commonly used for tasks including regression and classification.

3.4.4. Logistic Regression

Logistic Regression [32] is a supervised machine learning algorithm that is commonly used for Binary Classification problems, such as predicting whether an individual will develop a particular disease. In the context of lifestyle habits and disease prediction, logistic regression can be used to model the association between specific daily habits and the risk of developing certain diseases.

Logistic Regression uses a sigmoid function to map prediction outcomes and their probabilities. The sigmoid function is an S-shaped curve that transforms any real number to a range between 0 and 1. Therefore, this approach can be used to predict the probability that a person will have cardiovascular disease shortly based on their habits data.

$$\sigma(x) = \frac{1}{1+e^{-x}} \tag{7}$$

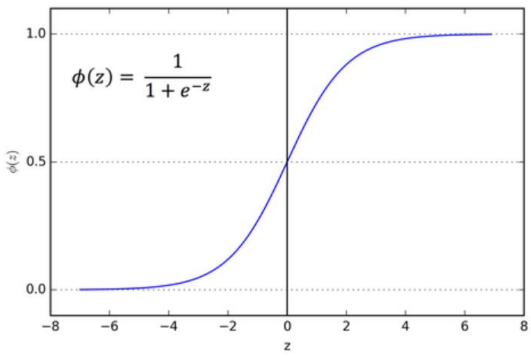


Figure 3.17. Sigmoid function [32]

Moreover, if the output of the sigmoid function exceeds a predetermined threshold on the graph, the model predicts that the instance belongs to that class. If the estimated probability falls below the predefined threshold, the model concludes that the instance does not belong to the class. For instance, if the output of the sigmoid function is above 0.5, the output is considered as 1 (having the disease). On the other hand, if the output is less than 0.5, the output is classified as 0 (not having the disease).

To evaluate model performance, the most used loss function is the mean squared error. However, because Logistic Regression returns a probability value between 0 and 1, the cross-entropy loss function should be utilized instead.

Figure 3.18 depicts the summary operation of the Logistic Regression model. A linear equation (z) is fed into a sigmoid function (σ) to predict the result (ŷ).

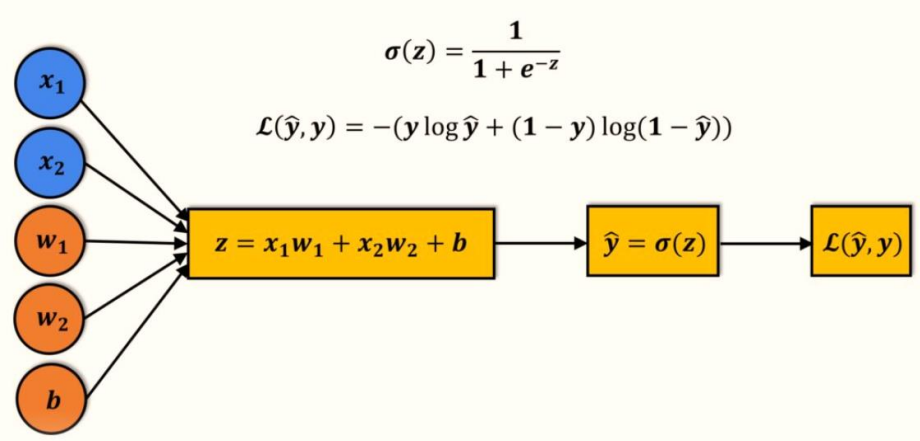


Figure 3.18. The working of the Logistic regression model [32]

3.4.5. Support Vector Machine (SVM)

Support Vector Machine (SVM) [33] is a supervised learning approach for classification and regression problems that determines the best decision boundary, known as a hyperplane, to separate an n-dimensional space into classes. It is mostly used for classification problems, such as disease prediction.

Support vectors are the extreme data points or vectors that are closest to the hyperplane. SVM will find the maximum distance from the support vectors and the hyperplane, known as the maximum margin, and place the best hyperplane at that position. Consider Figure 3.19 as an example of classifying 2 different categories.

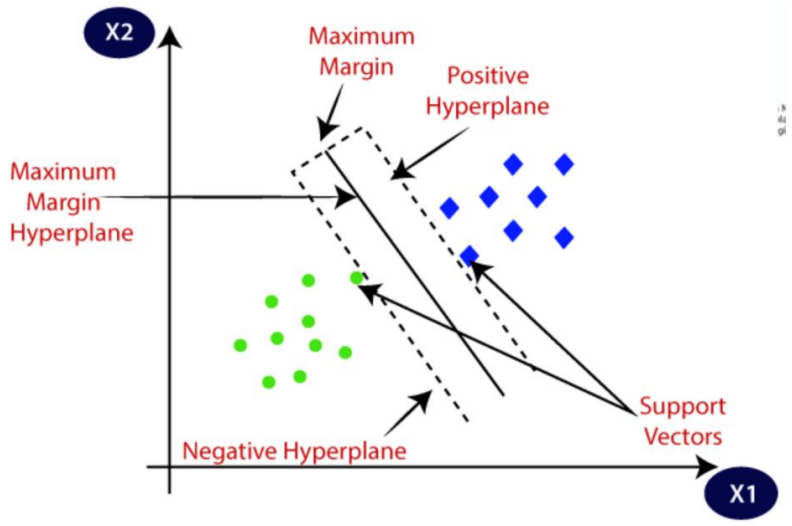


Figure 3.19. Support Vector Machine [33]

In this thesis, a Support Vector Machine will be utilized to categorize individuals based on their likelihood of getting cardiovascular disease. For example, an SVM model may be trained on a dataset comprising information on food and lifestyle behaviors, as well as other risk variables such as blood pressure, cholesterol levels, and smoking status, to predict whether a person is at high or low risk of developing cardiovascular disease.

3.4.6. Random Forest

The Random Forest model [34] is a valuable supervised learning algorithm that that utilizes ensemble learning, combining predictions from multiple models to produce superior results. It works by building a group of decision trees, similar to a forest, that make predictions based on a series of yes or no questions about the data's features. This algorithm can handle complex datasets with both continuous and categorical variables, making it a versatile tool for a variety of predictive tasks in machine learning. During the training phase,

Random Forest randomly selects observations and builds numerous decision trees, with the final classification prediction being the class that receives the most votes from the individual trees in the forest. This randomness introduces variability among the individual trees, mitigating the risk of overfitting and improving the overall prediction performance.

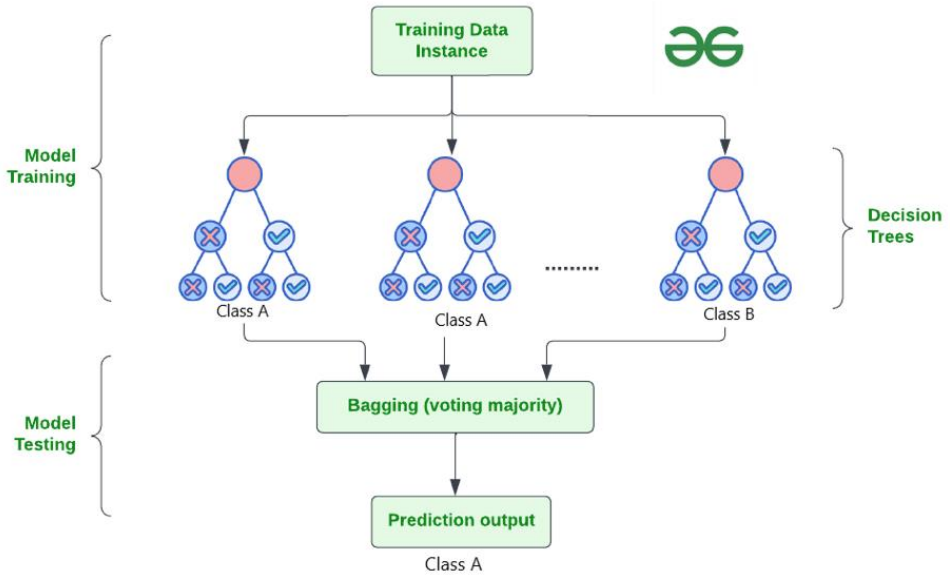


Figure 3.20. Random Forest [34]

3.5. Development Tools

For this machine learning thesis project, Python [35] is an excellent programming language that offers numerous advantages. One of the significant benefits of using Python is its extensive libraries, including NumPy, Pandas, and Scikit-Learn. These libraries simplify data manipulation, analysis, and preprocessing, which makes handling and managing diverse data sets more efficient. Python's code readability and simplicity also make it easier for researchers to develop their projects rapidly. The use of powerful visualization libraries like Matplotlib, Seaborn, or Plotly allows for insightful data exploration and presentation.

Figure 3.21. Python

The coding for this thesis was first developed on a Colab notebook, which is a Jupyter notebook hosted by Colab. Google Colab [36], also known as "Colaboratory," is a cloud-based Jupyter notebook environment offered by Google Research that allows users to write and execute Python code directly in their web browser without any setup required. The platform provides access to GPUs and TPUs at no cost and enables easy sharing of Colab notebooks, which combine executable code and rich text into a single document. However, as the memory for the free version of Colab, it can take time to load notebooks with a vast amount of code and information. Therefore, the coding files of this thesis is now using Visual Studio Code, or VS Code. As it is a fast source code editor, it saves a lot of time in editing and running the code with various data files and models.

A simple web page is built for the prediction objective of this thesis using Streamlit. Streamlit [37] is an open-source Python framework that allows users to easily develop web applications for Machine Learning and Data Science projects. The web app created by Steamlit has an interactive, clear, and dynamic appearance that is not only user-friendly, but also easy for updating tasks.

CHAPTER 4

IMPLEMENTATION AND RESULTS

4.1. Data files importing

All the data files extracted from the NHANES datasets are SAS transport (.XPT) files, so Python's Pandas library is used to import and read the data files. Each imported data file can become a dataframe for further processing steps.

```
import pandas as pd
# Body Measures
bmx_1112= pd.read_sas('BMX_G.XPT')
bmx_1314= pd.read_sas('BMX_H.XPT')
bmx_1516= pd.read_sas('BMX_I.XPT')
bmx_1718= pd.read_sas('BMX_J.XPT')
bmx_1720= pd.read_sas('P_BMX.XPT')
```

Figure 4.1. Data files importing

4.2. Data Preprocessing

The data files from five survey cycles from 2011 to 2020 were appended using Pandas pd.concat, then the appended datasets of each component were merged into a dataset based on their common column of Sequence Number "SEQN".

```
# Body Measures
bmx_1112= pd.read_sas('BMX_G.XPT')
bmx_1314= pd.read_sas('BMX_H.XPT')
bmx_1516= pd.read_sas('BMX_I.XPT')
bmx_1718= pd.read_sas('BMX_J.XPT')
bmx_1720= pd.read_sas('P_BMX.XPT')
bmx = pd.concat([bmx_1112, bmx_1314, bmx_1516, bmx_1718, bmx_1720], ignore_index=True)

exam = pd.merge(bp,bmx,how='right',on='SEQN')
```

Figure 4.2. Data appending and merging

After appending and merging all the data files, we have 4 primary datasets for 4 different categories: Demographics, Examination, Laboratory, and Questionnaire. The following is a full description of each component dataset, including the data files utilized,

the shape of each data file, the shape of the merged datasets, and whether or not the datasets have duplicate values.

Table 4.1. Demographics data files

Component	Cycle	Data files	# Data points	# Features	Merged shape		Duplicate
Demographics	11-12	DEMO_G.XPT	9,756	48	54716 x 56	**54716 x 56**	0
	13-14	DEMO_H.XPT	10,175	47			
	15-16	DEMO_I.XPT	9,971	47			
	17-18	DEMO_J.XPT	9,254	46			
	17-20	P_DEMO.XPT	15,560	29			

Table 4.2. Examination data files

Component	Cycle	Data files	# Data points	# Features	Merged shape		Duplicate
Blood Pressure	11-12	BPX_G.XPT	9,338	27	49055 x 39	**51699 x 66**	0
	13-14	BPX_H.XPT	9,813	23			
	15-16	BPX_I.XPT	9,544	21			
	17-18	BPX_J.XPT, BPXO_J.XPT	8,704	33			
	17-20	P_BPXO	11,656	12			
Body Measures	11-12	BMX_G.XPT	9,338	26	51699 x 28		0
	13-14	BMX_H.XPT	9,813	26			
	15-16	BMX_I.XPT	9,544	26			
	17-18	BMX_J.XPT	8,704	21			
	17-20	P_BMX.XPT	14,300	22			

Table 4.3. Laboratory data files

Component	Cycle	Data files	# Data points	# Features	Merged shape		Duplicate
Fasting Questionnaire	11-12	FASTQX_G.XPT	8,956	19	49681 x 19	**43766 x 73**	0
	13-14	FASTQX_H.XPT	9,422	19			
	15-16	FASTQX_I.XPT	9,165	19			
	17-18	FASTQX_J.XPT	8,366	19			

	17-20	P_FASTQX.XPT	13,772	19			
Standard Biochemistry Profile	11-12	BIOPRO_G.XPT	6,549	38	37082 x 41		0
	13-14	BIOPRO_H.XPT	6,979	38			
	15-16	BIOPRO_I.XPT	6,744	38			
	17-18	BIOPRO_J.XPT	6,401	41			
	17-20	P_BIOPRO.XPT	10,409	41			
Cholesterol - Total	11-12	TCHOL_G.XPT	7,821	3	43766 x 3		0
	13-14	TCHOL_H.XPT	8,291	3			
	15-16	TCHOL_I.XPT	8,021	3			
	17-18	TCHOL_J.XPT	7,435	3			
	17-20	P_TCHOL.XPT	12,198	3			
Cholesterol - HDL	11-12	HDL_G.XPT	7,821	3	43766 x 3		0
	13-14	HDL_H.XPT	8,291	3			
	15-16	HDL_I.XPT	8,021	3			
	17-18	HDL_I.XPT	7,435	3			
	17-20	P_HDL.XPT	12,198	3			
Cholesterol - LDL & Triglycerides	11-12	TRIGLY_G.XPT	3,239	6	17885 x 11		0
	13-14	TRIGLY_H.XPT	3,329	6			
	15-16	TRIGLY_I.XPT	3,191	6			
	17-18	TRIGLY_J.XPT	3,036	10			
	17-20	P_TRIGLY.XPT	5,090	10			

Table 4.4. Questionnaire data files

Component	Cycle	Data files	# Data points	# Features	Merged shape		Duplicate
Alcohol use	11-12	ALQ_G.XPT	5,615	10	31772 x 18	**52592 x 182**	0
	13-14	ALQ_H.XPT	5,924	10			
	15-16	ALQ_I.XPT	5,735	10			
	17-18	ALQ_J.XPT	5,533	10			
	17-20	P_ALQ.XPT	8,965	10			
Blood Pressure & Cholesterol	11-12	BPQ_G.XPT	6,175	15	35322 x 15		0
	13-14	BPQ_H.XPT	6,464	14			
	15-16	BPQ_I.XPT	6,327	11			

	17-18	BPQ_J.XPT	6,161	11			
	17-20	P_BPQ.XPT	10,195	11			
Cardio -vascular Health	11-12	CDQ_G.XPT	3,603	17	21499 x 17		0
	13-14	CDQ_H.XPT	3,815	17			
	15-16	CDQ_I.XPT	3,766	17			
	17-18	CDQ_J.XPT	3,882	17			
	17-20	P_CDQ.XPT	6,433	17			
Medical Conditions	11-12	MCQ_G.XPT	9,364	92	52592 x 135		0
	13-14	MCQ_H.XPT	9,770	95			
	15-16	MCQ_I.XPT	9,575	90			
	17-18	MCQ_J.XPT	8,897	76			
	17-20	P_MCQ.XPT	14,986	63			

All four appended component datasets above were then integrated into the final merged dataset using the SEQN column, allowing each participant to have one row for all the attributes associated with that individual.

```
df = pd.merge(pd.merge(pd.merge(demo, ques, how='right', on='SEQN'),
                       exam, how='left', on='SEQN'),
              lab, how='left', on='SEQN')
```

Figure 4.3. Final Dataset merging

The shape of the final merged dataset is 52,592 rows by 374 columns. Table 4.1 reveals there may be 54,716 participants in the final dataset, but because some of them did not have their cardiovascular disease history documented in the Questionnaire dataset, the final dataset was only merged to have the information for 52,592 participants in the Questionnaire file. Figure 4.4 shows a general sample of the final merged dataset. It is also obvious that all the variable names in the NHANES data files are expressed in some type of code because the original variable names are so long that they need to be abbreviated into a combination of starting letters. The explanation of each feature name used in this dataset is attached in the Appendix section.

	SEQN	SDDSRVYR	RIDSTATR	RIAGENDR	RIDAGEYR	RIDAGEMN	RIDRETH1	RIDRETH3	RIDEXMON	RIDEXAGY	...	PHAGUMMN	PHQ050	PHAANTHR	PHAANTMN	PHQ060	PHASUPHR	PHASUPMN	PHAFSTHR	PHAFSTMN	PHDSESN
0	62161.0	7.0	2.0	1.0	22.0	NaN	3.0	3.0	2.0	NaN	...	NaN	2.0	NaN	NaN	2.0	NaN	NaN	14.0	37.0	5.397605e-79
1	62162.0	7.0	2.0	2.0	3.0	NaN	1.0	1.0	1.0	3.0	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2	62163.0	7.0	2.0	1.0	14.0	NaN	5.0	6.0	2.0	14.0	...	NaN	2.0	NaN	NaN	2.0	NaN	NaN	17.0	55.0	1.000000e+00
3	62164.0	7.0	2.0	2.0	44.0	NaN	3.0	3.0	1.0	NaN	...	NaN	2.0	NaN	NaN	2.0	NaN	NaN	11.0	6.0	5.397605e-79
4	62165.0	7.0	2.0	2.0	14.0	NaN	4.0	4.0	2.0	14.0	...	NaN	2.0	NaN	NaN	2.0	NaN	NaN	12.0	11.0	5.397605e-79
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
52587	124818.0	66.0	2.0	1.0	40.0	NaN	4.0	4.0	1.0	NaN	...	NaN	2.0	NaN	NaN	2.0	NaN	NaN	6.0	4.0	1.000000e+00
52588	124819.0	66.0	2.0	1.0	2.0	NaN	4.0	4.0	2.0	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
52589	124820.0	66.0	2.0	2.0	7.0	NaN	3.0	3.0	2.0	NaN	...	NaN	2.0	NaN	NaN	2.0	NaN	NaN	4.0	42.0	1.000000e+00
52590	124821.0	66.0	2.0	1.0	63.0	NaN	4.0	4.0	1.0	NaN	...	NaN	2.0	NaN	NaN	2.0	NaN	NaN	10.0	8.0	5.397605e-79
52591	124822.0	66.0	2.0	1.0	74.0	NaN	2.0	2.0	2.0	NaN	...	NaN	2.0	NaN	NaN	2.0	NaN	NaN	12.0	52.0	5.397605e-79

52592 rows × 374 columns

Figure 4.4. Merged dataset overview

The label feature of cardiovascular disease presence was developed using four Questionnaire variables representing the disease types or symptoms, as described in Chapter 3. Figure 4.5 below shows the description of one of the four variables. The feature contains four distinct values, with “1” and “2” indicating the presence of each participant’s disease in the past, and “7” and “9” representing missing values. The final row of “Missing” indicates the number of participants who are ineligible for this question due to the target age restriction of 20 years old or older.

MCQ160b - Ever told had congestive heart failure

Variable Name: MCQ160b

SAS Label: Ever told had congestive heart failure

English Text: Has a doctor or other health professional ever told {you/SP} that {you/s/he} . . .had congestive heart failure?

Target: Both males and females 20 YEARS - 150 YEARS

Code or Value	Value Description	Count	Cumulative	Skip to Item
1	Yes	201	201	
2	No	5351	5552	
7	Refused	0	5552	
9	Don't know	17	5569	
.	Missing	3328	8897	

Figure 4.5. Description of a feature to create label

To avoid inaccurate insights, the four characteristics used to give the label should only have two unique values of "Yes" or "No", hence their "7" and "9" values were transformed to NaN. In NHANES Questionnaire data files, the value of "No" is represented as "2" instead

of "0" in common sense, thus they were converted to "0" values using the Python code provided below.

```
# Change all rows with values 7 and 9 in label columns to NaN and change 2 to 0

for col in ['MCQ160B', 'MCQ160C', 'MCQ160E', 'MCQ160F']:
  df[col] = np.where(~df[col].isin([1, 2]), np.nan, np.where(df[col] == 1, 1, 0))
```

Figure 4.6. Data cleaning for features used to create the label

Following the transformation procedures, it was clear that the missing values would appear together in the four features, implying that if one of the four columns has a missing value, the same happens for the others. Therefore, all rows with missing values in these four columns were then removed from the dataset.

```
# Drop all rows with 4 label columns are null

df = df[(df['MCQ160B'].notna()) & (df['MCQ160C'].notna()) & (df['MCQ160E'].notna()) & (df['MCQ160F'].notna())]
```

Figure 4.7. Deletion of rows containing missing values

Other Questionnaire features have the characteristics of the four critical elements listed above, thus any confusing values should be handled. There are several types of missing values, as indicated in Table 3.1. The numbers have the form of one to five figures, thus the attributes associated with these values must be identified to treat each sort of missing value effectively.

```
# Change and drop some ambiguous values

to_change_cols = ['DMDEDUC2', 'ALQ101', 'ALQ120U', 'ALQ151', 'ALQ111', 'BPQ020', 'BPQ030', 'BPQ040A', 'BPQ050A', 'BPQ057', 'BPQ056',
                  'BPQ059', 'BPQ080', 'BPQ060', 'BPQ070', 'BPQ090D', 'BPQ100D', 'CDQ001', 'CDQ002', 'CDQ008', 'CDQ010', 'DMDCITZN',
                  'DMDEDUC2', 'DMDHREDU', 'DMDHSEDU', 'DMDHREDZ', 'MCQ010', 'MCQ035', 'MCQ053', 'MCQ070', 'MCQ080', 'MCQ082', 'MCQ086',
                  'MCQ092', 'MCQ140', 'MCQ160A', 'MCQ195', 'MCQ160N', 'MCQ160G', 'MCQ160M', 'MCQ160K', 'OSQ230', 'MCQ160L', 'MCQ220',
                  'MCQ300A', 'MCQ300B', 'MCQ300C', 'MCQ370A', 'MCQ370B', 'MCQ370C', 'MCQ370D', 'AGQ030', 'MCQ160O', 'MCQ203', 'MCQ520',
                  'MCQ550', 'MCQ560', 'MCQ366A', 'MCQ366B', 'MCQ366C', 'MCQ366D', 'MCQ371A', 'MCQ371B', 'MCQ371C', 'MCQ371D', 'MCQ160P',
                  'MCQ160B', 'MCQ160C', 'MCQ160E', 'MCQ160F', 'MCQ365A', 'MCQ365B', 'MCQ365C', 'MCQ365D', 'MCQ084', 'MCQ170M', 'MCQ540',
                  'ALQ110', 'MCQ160D', 'DMQMILIZ']

for col in to_change_cols:
  df[col] = np.where(df[col].isin([7, 9]), np.nan, df[col])

for col in ['ALQ120Q', 'ALQ130', 'ALQ141Q', 'BPD035', 'ALQ170']:
  df[col] = np.where(df[col].isin([777, 999]), np.nan, df[col])

for col in ['MCQ025']:
  df[col] = np.where(df[col].isin([77777, 99999]), np.nan, df[col])

for col in ['DMDYRSUS', 'DMDMARTL', 'DMDHRBR4', 'DMDHRMAR', 'DMDHRMAZ', 'DMDMARTZ', 'DMDBORN4']:
  df[col] = np.where(df[col].isin([77, 99]), np.nan, df[col])

for col in ['BPAARM', 'BPXML1']:
  df[col] = np.where(df[col] == 888, np.nan, df[col])

for col in ['BPXPTY', 'BPAEN1', 'BPAEN2', 'BPAEN3']:
  df[col] = np.where(df[col] == 8, np.nan, df[col])
```

Figure 4.8. Change of ambiguous values into missing values

After having ambiguous values transformed into the appropriate category of missing values, the dataset contained several features with more than 90% missing values. These features may be unable to add to the thesis's ideas and outcomes due to a lack of knowledge. Therefore, to prevent biased findings, they were removed from the dataset. The values "2" in certain other Questionnaire features were also changed to "0" to ensure data integrity.

```
# Drop all columns with more than 90% null

to_drop = []

for col in df.columns:
  if df[col].isna().sum() / len(df) > 0.9:
    to_drop.append(col)

df = df.drop(columns = to_drop)
```

Figure 4.9. Data cleaning: Deletion of columns with more than 90% of values missing

After removing the selected rows and columns with missing values, the remaining dataset has 243 features for 31,619 data points of participant information. However, the dataset was still not clean, implying that there were numerous missing values in the dataset. As a result, following the NHANES organization's caution, missing values were imputed using scikit-learn's SimpleImputer and KNNImputer.

SimpleImputer method fills in the missing values using the feature's mean or median. The results using the datasets with missing values imputed by mean, median, and KNNImputer will be compared in the next chapter.

```
# Impute missing values using mean

from sklearn.impute import SimpleImputer

imputer = SimpleImputer(strategy = 'mean')
df_imputed = imputer.fit_transform(df)

df = pd.DataFrame(data = df_imputed, columns = df.columns)
```

```
# Impute missing values using median

from sklearn.impute import SimpleImputer

imputer = SimpleImputer(strategy = 'median')
df_imputed = imputer.fit_transform(df)

df = pd.DataFrame(data = df_imputed, columns = df.columns)
```

Figure 4.10. Missing values imputation using SimpleImputer

For this dataset, the parameter n_neighbors was set to 3, which means that KNNImputer would discover the three closest data points to the one that needs to be imputed, and the average value of those three "neighbors" could be used to fill in the missing values.

```
# Impute missing values

from sklearn.impute import KNNImputer

impute_knn = KNNImputer(n_neighbors = 3)
df_impute = impute_knn.fit_transform(df)

df = pd.DataFrame(data = df_impute, columns = df.columns)
```

Figure 4.11. Missing values imputation using KNNImputer

Section 3.3.4 assigned the label indicating the presence or absence of cardiovascular disease in the dataset, but it was found to be imbalanced. The "No" values far exceeded the "Yes" values, making the data imbalanced. To address this, 6,000 data points out of a total of 28,286 were randomly selected from the "No" group, while the "Yes" group was left unchanged with 3,333 data points. Section 3.3.5 provides further information on this balancing approach, which avoided the creation of fake data through oversampling, and kept

an appropriate number of data points using undersampling that could mostly represent the whole dataset patterns.

```
df = pd.concat([df[df['Label'] == 1], df[df['Label'] == 0].sample(n = 6000, random_state = 1204)]).sample(frac = 1, random_state = 1204)
```
✓ 0.4s

```
df.value_counts('Label')
```
✓ 0.0s

```
Label
0    6000
1    3333
Name: count, dtype: int64
```

Figure 4.12. Label Imbalance handling

The resulting dataset after preprocessing was divided into train, validation, and test sets for machine learning modeling. Because the dataset comprises 9,333 data points, the train set has data points ranging from 0 to 5,000, the validation set from 5,000 to 5,500, and the remainder is in the test set.

```
df_train = df.iloc[:5000]
df_val = df.iloc[5000:5500]
df_test = df.iloc[5500:]
```

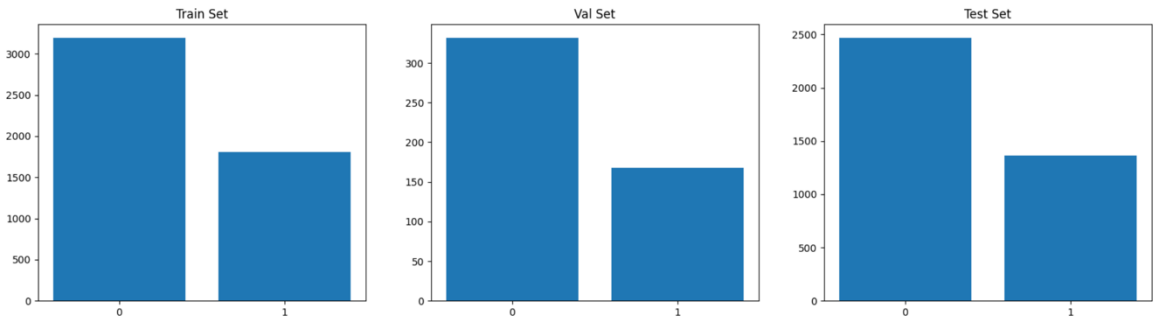
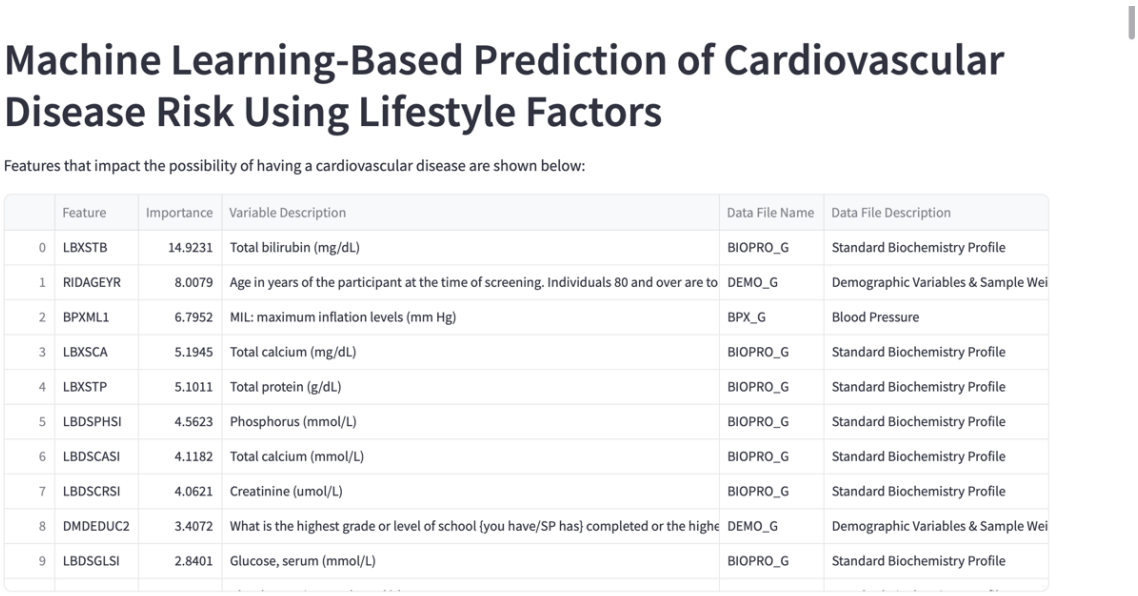


Figure 4.13. Train-test dataset split

A simple web page is built for the prediction objective of this thesis using Streamlit. Below is the brief overview of the web page, which shows the topic name and the list of key features have an significant impact on the disease.



	Feature	Importance	Variable Description	Data File Name	Data File Description
0	LBXSTB	14.9231	Total bilirubin (mg/dL)	BIOPRO_G	Standard Biochemistry Profile
1	RIDAGEYR	8.0079	Age in years of the participant at the time of screening. Individuals 80 and over are to	DEMO_G	Demographic Variables & Sample Wei
2	BPXML1	6.7952	MIL: maximum inflation levels (mm Hg)	BPX_G	Blood Pressure
3	LBXSCA	5.1945	Total calcium (mg/dL)	BIOPRO_G	Standard Biochemistry Profile
4	LBXSTP	5.1011	Total protein (g/dL)	BIOPRO_G	Standard Biochemistry Profile
5	LBDSPHSI	4.5623	Phosphorus (mmol/L)	BIOPRO_G	Standard Biochemistry Profile
6	LBDSCASI	4.1182	Total calcium (mmol/L)	BIOPRO_G	Standard Biochemistry Profile
7	LBDSCRSI	4.0621	Creatinine (umol/L)	BIOPRO_G	Standard Biochemistry Profile
8	DMDEDUC2	3.4072	What is the highest grade or level of school {you have/SP has} completed or the highe	DEMO_G	Demographic Variables & Sample Wei
9	LBDSGLSI	2.8401	Glucose, serum (mmol/L)	BIOPRO_G	Standard Biochemistry Profile

Figure 4.14. Demo Website (1)

The website aims to predict the possibility that a people can have cardiovascular diseases in the future based on the answers they give for the questions of key features in the list above. The key features list consists of various data types, including binary, integer, and float values. Therefore, the next part of the web app allows people to answer some questions about their information and health status data by choosing from select boxes and filling in numerical values.

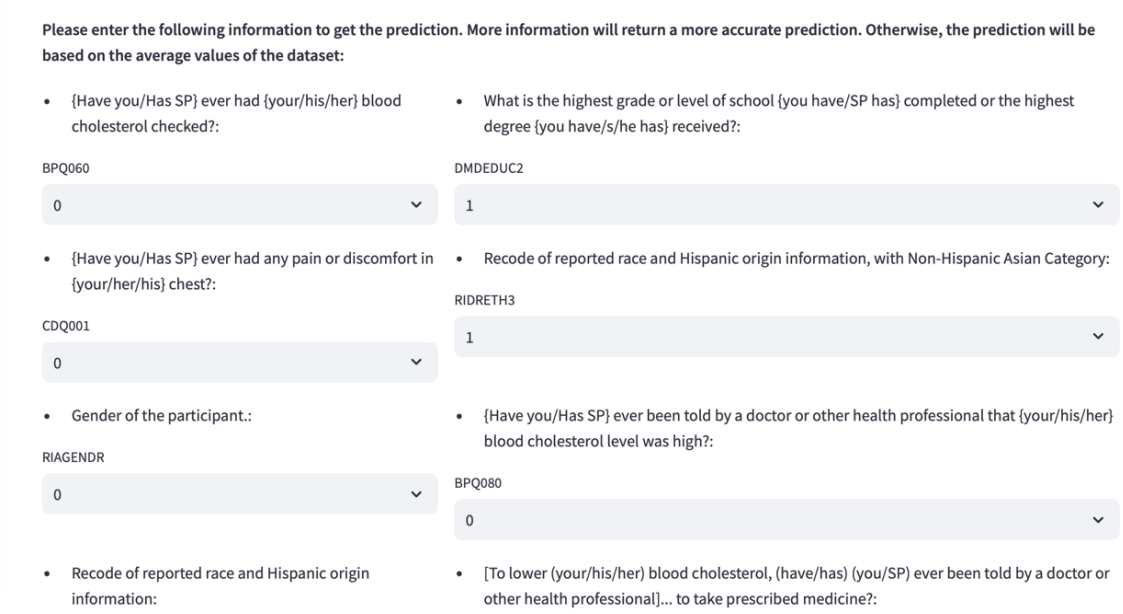


Figure 4.15. Demo Website (2)

The data inputted by the user is recorded to a dataframe, then the data is applied to the model for prediction. This web app use CatBoost model as it has the best performance with the highest accuracy for the data, and it can reduce errors by its boosting algorithm. For the questions that the user does not answer, the mean value of that feature will be filled in as default for avoiding missing values. At the end of the web application, there is a "Predict" button to show the prediction results from the obtained data above. The result can be high or low possibility of having cardiovascular diseases based on the user's data, along with a short suggestion for the user to have a check-up in the hospitals to keep track of their health better.

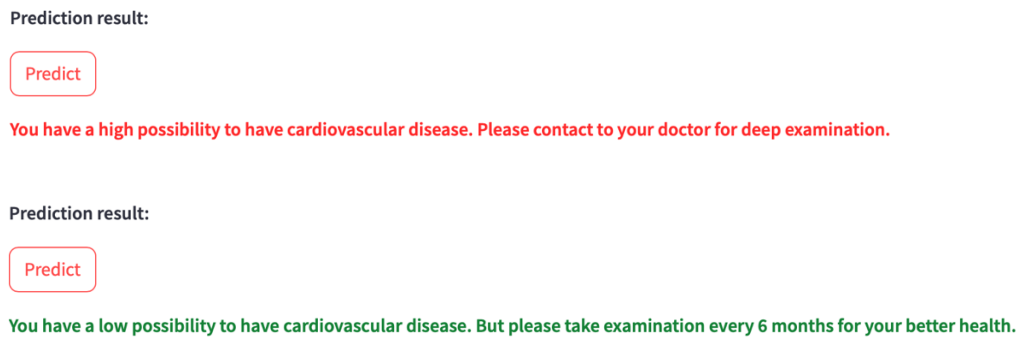


Figure 4.16. Prediction Results

CHAPTER 5

DISCUSSION AND EVALUATION

After conducting a thorough evaluation of various machine learning models, it has become apparent that each algorithm exhibits unique performance metrics. The CatBoost algorithm achieved the highest accuracy, displaying its robust predictive capabilities, and XGBoost closely follows with an impressive accuracy rate, demonstrating its ability to handle complex datasets. Gradient Boosting and Random Forest also performed well, indicating their suitability for predictive tasks. However, Logistic Regression and Support Vector Machines showed comparatively lower predictive accuracy, which suggest potential limitations in handling complex datasets. Although the predictions were highly accurate, the top three models required significant time for both training and testing due to the complexity of multiple data points and features. Additionally, their boosting algorithms worked to optimize the results of each previous sub-model, further contributing to the required time investment. The varying performance across these models emphasizes the importance of selecting an algorithm aligned with the specific characteristics of the data for optimal predictive outcomes.

At the beginning, KNNImputer was used as the main approach to impute missing values as its ability to take feature associations into consideration. However, because it needs to find the number of k most related data points and take their average value to fill in the missing place, it took far a lot of time for processing, especially around 53 to 90 minutes. As a result, other approaches are applied in comparison with KNNImputer. By using mean or median, the missing values are imputed only about 1 minute, and those cleaned datasets can even provide the models with better performance, as described in Table 5.1.

```
from sklearn.model_selection import cross_val_score

def model_accuracy_cv(model, X_test, y_test):
    return cross_val_score(model, X_test, y_test, cv = 10, scoring = 'accuracy').mean()
```

Figure 5.1. Cross Validation for Model Accuracy Comparison

In order to effectively compare the performance of the machine learning models utilized in this thesis project, cross-validation techniques were employed. Specifically, the k-fold cross-validation [38] is a technique utilized to evaluate predictive models. This technique involves dividing the dataset into k subsets or folds, training and evaluating the model k times. The performance metrics obtained from each fold are then averaged to estimate the model's generalization performance. This method facilitates model assessment and results in a more reliable measure of a model's effectiveness. The accuracy of each model was calculated using the cross_val_score function of sklearn, with k-folds set to 10. The average accuracy of these 10 folds and the accuracy for normal train – test split method above are presented in Tables 5.1 and 5.2, with different numbers of features applied.

Table 5.1. Models Evaluation (Accuracy, 239 features)

Model	**Train - test split**			**Cross validation (10 folds)**		
	Mean	**Median**	**KNNImputer**	**Mean**	**Median**	**KNNImputer**
CatBoost	0.8257	0.8281	0.8216	0.8534	0.8566	0.8404
XGBoost	0.8276	0.8182	0.8153	0.8444	0.8451	0.8345
Gradient Boosting	0.8229	0.8197	0.8182	0.8275	0.8235	0.8250
Logistic Regression	0.6961	0.6940	0.6908	0.7068	0.7117	0.7004
Support Vector Machines (SVM)	0.6444	0.6444	0.6447	0.6429	0.6429	0.6429
Random Forest	0.8257	0.8304	0.8161	0.8459	0.8476	0.8278

The dataset used to train the models has 239 features, which is such a lot of features for a person to input. Therefore, a list of most important features is created from the CatBoost model to show which features have association with the label of disease appearance. The

performance of models using 30 most important features has a slight decrease in accuracy because of the little loss in information, but it is still good for use.

Table 5.2. Models Evaluation (Accuracy, 30 most important features)

Model	**Train - test split**			**Cross validation (10 folds)**		
	Mean	**Median**	**KNNImputer**	**Mean**	**Median**	**KNNImputer**
CatBoost	0.8242	0.8296	0.8111	0.8422	0.8438	0.8285
XGBoost	0.8192	0.8140	0.8041	0.8321	0.8254	0.8187
Gradient Boosting	0.8153	0.8179	0.8109	0.8175	0.8210	0.8162
Logistic Regression	0.7975	0.7910	0.7801	0.7963	0.7995	0.7721
Support Vector Machines (SVM)	0.7514	0.7480	0.7472	0.7629	0.7601	0.7581
Random Forest	0.8299	0.8322	0.8150	0.8507	0.8480	0.8338

CHAPTER 6

CONCLUSION AND FUTURE WORK

6.1. Conclusion

In conclusion, this thesis has developed machine learning models for the prediction of cardiovascular diseases utilizing lifestyle habits data from the NHANES dataset. This dataset, which is complicated to analyze, has had most of its issues resolved during the preprocessing stage before being employed in machine learning models. The thesis transformed this dataset from a collection of discrete data files with numerous missing values, a lack of labels, and imbalanced labels into a comprehensive clean dataset for analytics and modeling. The application of six distinct models—CatBoost, Gradient Boosting, XGBoost, Linear Regression, Support Vector Machines and Random Forest—has provided valuable insights into the complex relationship between lifestyle choices and cardiovascular health. The findings emphasize the promising capabilities of machine learning in accurately predicting cardiovascular illnesses, with certain models displaying significant performance. Alongside analyzing a variety of machine learning models for the predictive purpose of this thesis, a simple web application has been developed to assess an individual's likelihood of developing cardiovascular diseases based on their responses to questions about their lifestyle and health status statistics on the web app.

It's worth noting that despite some notable achievements, there are certain limitations to this thesis that should be taken into consideration. The dietary habits data could not be utilized by prediction models due to its multi-record structure, and the Dietary component dataset was not included in the final analytics dataset. Moreover, the current method of selecting data points for training and testing models isn't optimal, as it results in the loss of information, although it can solve the issue of fake data in the dataset used for model

creation. Additionally, while some models have shown promising accuracy, others still have relatively low accuracy, suggesting that there is room for improvement in achieving more precise disease predictions.

This thesis has laid the foundation for future research in predictive modeling for cardiovascular diseases based on lifestyle habits data. Such research can contribute to advancing our understanding of the relationship between lifestyle habits and cardiovascular diseases, and aid in the development of more effective prevention and treatment strategies.

6.2. Future work

The thesis serves as a foundation for future research, offering avenues for improvement and expansion. It is essential to explore advanced feature engineering and model optimization techniques to strengthen the accuracy and robustness of predictive models. This entails addressing challenges such as selecting the optimal test set and handling imbalanced labels, identifying the most impactful lifestyle habits that influence the prediction of cardiovascular diseases, and exploring hyperparameter tuning and optimization approaches to achieve better efficiency and wider applicability across diverse datasets.

After data analysis has been conducted to answer crucial research questions, it is highly advisable to create a more refined and user-friendly website that allows individuals to take part in a survey concerning their daily lifestyle habits, dietary and overall health status. The website should be an improved version of the existing simple web application that can provide the user with prediction results in the form of a percentage probability of developing cardiovascular disease, followed by personalized recommendations for behavior modification based on that probability. This will enable users to take proactive steps towards maintaining a healthy lifestyle and reducing their risk of cardiovascular disease.

REFERENCES

[1] Cordina, J., Levin, E., & Stein, G. (2022, March 25). Consumer Health Insights: How respondents are adapting to the "new normal". *McKinsey 2022 Consumer Health Insights COVID-19 Wave 1 Survey*. https://www.mckinsey.com/industries/healthcare/our-insights/covid-19-consumer-healthcare-insights-what-2021-may-hold

[2] *Most important issues globally 2023*. (n.d.). Statista. https://www.statista.com/statistics/946266/most-worrying-topics-worldwide/

[3] Betts, D., Korenda, L., & Giuliani, S. (2020, August 13). *2020 Health care consumer survey: consumer health trends*. Deloitte. https://www2.deloitte.com/xe/en/insights/industry/health-care/consumer-health-trends.html

[4] World Health Organization (WHO). (n.d.). *Cardiovascular diseases*. World Health Organization (WHO). https://www.who.int/health-topics/cardiovascular-diseases#tab=tab_1

[5] *Preventing Heart Disease | The Nutrition Source | Harvard T.H. Chan School of Public Health*. (2022, August). Harvard T.H. Chan School of Public Health. https://www.hsph.harvard.edu/nutritionsource/disease-prevention/cardiovascular-disease/preventing-cvd/

[6] World Health Organization (WHO). (2021, June 11). *Cardiovascular diseases (CVDs)*. World Health Organization (WHO). https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)

[7] Gaurav, K., Kumar, A., Singh, P., Kumari, A., Kasar, M., & Suryawanshi, T. (2023). Human Disease Prediction using Machine Learning Techniques and Real-life Parameters. *International Journal of Engineering*, 36(6), 1092-1098. doi: 10.5829/ije.2023.36.06c.07

[8] Dinh, A., Miertschin, S., Young, A., & Mohanty, S. D. (2019, November 06). A data-driven approach to predicting diabetes and cardiovascular disease with machine learning. *BMC Medical Informatics and Decision Making 19, 211*. https://doi.org/10.1186/s12911-019-0918-5

[9] Centers for Disease Control and Prevention. (n.d.). *About CDC | About*. Centers for Disease Control and Prevention (CDC). https://www.cdc.gov/about/

[10] USAGov. (n.d.). *Centers for Disease Control and Prevention (CDC)*. USAGov. https://www.usa.gov/agencies/centers-for-disease-control-and-prevention

[11] National Center for Health Statistics. (2023, May 31). *NHANES - About the National Health and Nutrition Examination Survey*. CDC. https://www.cdc.gov/nchs/nhanes/about_nhanes.htm

[12] Centers for Disease Control and Prevention. (2022, July 15). *About NCHS - Homepage*. CDC. https://www.cdc.gov/nchs/about/index.htm

[13] National Center for Health Statistics. (2015, November 6). *NHANES - Data Accomplishments*. CDC. https://www.cdc.gov/nchs/nhanes/dataaccomp.htm#print

[14] National Center for Health Statistics. (2021, July 22). *NHANES - Participants - About*. CDC. https://www.cdc.gov/nchs/nhanes/participant/participant-about.htm

[15] National Center for Health Statistics. (n.d.). *NHANES Questionnaires, Datasets, and Related Documentation*. CDC. https://wwwn.cdc.gov/nchs/nhanes/

[16] National Center for Health Statistics. (n.d.). *Datasets and Documentation*. CDC. https://wwwn.cdc.gov/nchs/nhanes/tutorials/Datasets.aspx

[17] National Center for Health Statistics. (n.d.). *NHANES Tutorials*. CDC. https://wwwn.cdc.gov/nchs/nhanes/tutorials/default.aspx

[18] National Center for Health Statistics. (2015). *NHANES General Information about NHANES Documentation Files*. CDC. https://wwwn.cdc.gov/nchs/nhanes/continuousnhanes/doccontents.aspx?BeginYear=2015

[19] *pandas.read_sas — pandas 2.1.4 documentation*. (n.d.). Pandas. https://pandas.pydata.org/docs/reference/api/pandas.read_sas.html

[20] Tamboli, N. (2023, July 14). *Effective Strategies for Handling Missing Values in Data Analysis (Updated 2023)*. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2021/10/handling-missing-value/

[21] NHANES. (n.d.). *BPQ_H*. CDC. https://wwwn.cdc.gov/Nchs/Nhanes/2013-2014/BPQ_H.htm#BPQ020

[22] scikit-learn developers. (n.d.). *sklearn.impute.KNNImputer — scikit-learn 1.3.2 documentation*. Scikit-learn. https://scikit-learn.org/stable/modules/generated/sklearn.impute.KNNImputer.html

[23] Roy, K. (2020, July 20). *KNNImputer | Way To Impute Missing Values*. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2020/07/knnimputer-a-robust-way-to-impute-missing-values-using-scikit-learn/

[24] Centers for Disease Control and Prevention. (2017, April 6). *Indicator Definitions - Cardiovascular Disease | CDI | DPH*. CDC. https://www.cdc.gov/cdi/definitions/cardiovascular-disease.html

[25] Class *Imbalance in ML: 10 Best Ways to Solve it Using Python*. (2024, January 17). Analytics Vidhya. https://www.analyticsvidhya.com/blog/2020/07/10-techniques-to-deal-with-class-imbalance-in-machine-learning/

[26] Mazumder, S. (2023, September 27). *What is Imbalanced Data | Techniques to Handle Imbalanced Data*. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2021/06/5-techniques-to-handle-imbalanced-data-for-a-classification-problem/

[27] Uddin, S., Khan, A., Hossain, M. et al. (2019, December 21). Comparing different supervised machine learning algorithms for disease prediction. *BMC Medical Informatics and Decision Making 19, 281*. https://doi.org/10.1186/s12911-019-1004-8

[28] Saini, A. (2024, January 10). *Gradient Boosting: A Step-by-Step Guide*. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2021/09/gradient-boosting-algorithm-a-complete-guide-for-beginners/

[29] Yandex. (n.d.). *CatBoost - open-source gradient boosting library*. https://catboost.ai/

[30] Shaikh, B. (2023, July 27). *CatBoost: A Solution for Building Model with Categorical Data.* Analytics Vidhya. https://www.analyticsvidhya.com/blog/2023/07/catboost-building-model-with-categorical-data/

[31] *What Is XGBoost and How Does It Improve Machine Learning?* (2018, September 6). Analytics Vidhya. https://www.analyticsvidhya.com/blog/2018/09/an-end-to-end-guide-to-understand-the-math-behind-xgboost/

[32] Bonthu, H. (2021, July 11). *An Introduction to Logistic Regression.* Analytics Vidhya. https://www.analyticsvidhya.com/blog/2021/07/an-introduction-to-logistic-regression/

[33] Saini, A. (2024, January 23). *Guide on Support Vector Machine (SVM) Algorithm*. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2021/10/support-vector-machinessvm-a-complete-guide-for-beginners/

[34] Jain, S. (2024, February 22). Random Forest Algorithm in Machine Learning. GeeksforGeeks. https://www.geeksforgeeks.org/random-forest-algorithm-in-machine-learning/

[35] Python Software Foundation. (n.d.). Python. Welcome to Python.org. https://www.python.org/

[36] Google Research. (n.d.). Google Colab. Google Research. https://research.google.com/colaboratory/faq.html

[37] Streamlit Inc. (n.d.). *Streamlit*. Streamlit • A faster way to build and share data apps. https://streamlit.io

[38] Pandian, S. (2023, November 17). K-Fold Cross Validation Technique and its Essentials. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2022/02/k-fold-cross-validation-technique-and-its-essentials/

APPENDIX A: Codebook for NHANES Data

#	Variable Name	Variable Description	Data File Name	Data File Description
1	AGQ030	During the past 12 months, {have you/has SP} had an episode of hay fever?	RDQ_G	Respiratory Health
2	ALQ101	The next questions are about drinking alcoholic beverages. Included are liquor (such as whiskey or gin), beer, wine, wine coolers, and any other type of alcoholic beverage.In any one year, {have you/has SP} had at least 12 drinks of any type of alcoholic beverage? By a drink, I mean a 12 oz. beer, a 5 oz. glass of wine, or one and half ounces of liquor.	ALQ_G	Alcohol Use
3	ALQ110	In {your/SP's} entire life, {have you/has he/ has she} had at least 12 drinks of any type of alcoholic beverage?	ALQ_G	Alcohol Use
4	ALQ111	The next questions are about drinking alcoholic beverages. Included are liquor (such as whiskey or gin), beer, wine, wine coolers, and any other type of alcoholic beverage. In {your/SP's} entire life, {have you/has he/has she} had at least 1 drink of any kind of alcohol, not counting small tastes or sips? By a drink, I mean a 12 oz. beer, a 5 oz. glass of wine, or one and a half ounces of liquor.	ALQ_J	Alcohol Use
5	ALQ121	During the past 12 months, about how often did {you/SP} drink any type of alcoholic beverage? PROBE: How many days per week, per month, or per year did {you/SP} drink?	ALQ_J	Alcohol Use
6	ALQ130	In the past 12 months, on those days that {you/SP} drank alcoholic beverages, on the average, how many drinks did {you/he/she} have?	ALQ_G	Alcohol Use
7	ALQ141Q	In the past 12 months, on how many days did {you/SP} have {DISPLAY NUMBER} or more drinks of any alcoholic beverage? PROBE: How many days per week, per month, or per year did {you/SP} have {DISPLAY NUMBER} or more drinks in a single day?	ALQ_G	Alcohol Use
8	ALQ142	During the past 12 months, about how often did {you/SP} have {DISPLAY NUMBER} or more drinks of any alcoholic beverage? PROBE: How many days per week, per month, or per year did {you/SP} have {DISPLAY NUMBER} or more drinks in a single day?	ALQ_J	Alcohol Use
9	ALQ151	Was there ever a time or times in {your/SP's} life when {you/he/she} drank {DISPLAY NUMBER} or more drinks of any kind of alcoholic beverage almost every day?	ALQ_G	Alcohol Use
10	ALQ170	Considering all types of alcoholic beverages, during the past 30 days, how many times did you have {5/4} or more drinks on an occasion?	ALQ_J	Alcohol Use

11	ALQ270	During the past 12 months, about how often did {you/SP} have {DISPLAY NUMBER} or more drinks in a period of two hours or less?	ALQ_J	Alcohol Use
12	ALQ280	During the past 12 months, about how often did {you/SP} have 8 or more drinks in a single day?	ALQ_J	Alcohol Use
13	BMXBMI	Body Mass Index (kg/m**2)	BMX_G	Body Measures
14	BMXWT	Weight (kg)	BMX_G	Body Measures
15	BPAOMNTS	Difference in minutes between blood pressure obtained by a physician with a mercury sphygmomanometer (legacy) and blood pressure obtained by a health technician with an oscillometric device.	BPXO_J	Blood Pressure - Oscillometric Measurements
16	BPD035	How old {were you/was SP} when {you were/he/she was} first told that {you/he/she} had hypertension or high blood pressure?	BPQ_G	Blood Pressure & Cholesterol
17	BPQ020	{Have you/Has SP} ever been told by a doctor or other health professional that {you/s/he} had hypertension, also called high blood pressure?	BPQ_G	Blood Pressure & Cholesterol
18	BPQ030	{Were you/Was SP} told on 2 or more different visits that {you/s/he} had hypertension, also called high blood pressure?	BPQ_G	Blood Pressure & Cholesterol
19	BPQ040A	Because of {your/SP's} (high blood pressure/hypertension), {have you/has s/he} ever been told to . . . take prescribed medicine?	BPQ_G	Blood Pressure & Cholesterol
20	BPQ050A	HELP AVAILABLE (Are you/Is SP) now taking prescribed medicine	BPQ_G	Blood Pressure & Cholesterol
21	BPQ056	{Did you/Did SP} take {your/his/her} blood pressure at home during the last 12 months?	BPQ_G	Blood Pressure & Cholesterol
22	BPQ057	{Have you/Has SP} ever been told by a doctor or other health professional that {you have/s/he has} high normal blood pressure or borderline hypertension?	BPQ_G	Blood Pressure & Cholesterol
23	BPQ059	Did a doctor or other health professional tell {you/SP} to take {your/his/her} blood pressure at home?	BPQ_G	Blood Pressure & Cholesterol
24	BPQ060	{Have you/Has SP} ever had {your/his/her} blood cholesterol checked?	BPQ_G	Blood Pressure & Cholesterol
25	BPQ070	About how long has it been since {you/SP} last had {your/his/her} blood cholesterol checked? Has it been...	BPQ_G	Blood Pressure & Cholesterol
26	BPQ080	{Have you/Has SP} ever been told by a doctor or other health professional that {your/his/her} blood cholesterol level was high?	BPQ_G	Blood Pressure & Cholesterol
27	BPQ090D	[To lower (your/his/her) blood cholesterol, (have/has) (you/SP) ever been told by a doctor or other health professional]... to take prescribed medicine?	BPQ_G	Blood Pressure & Cholesterol
28	BPQ100D	(Are you/Is SP) now following this advice to take prescribed medicine?	BPQ_G	Blood Pressure & Cholesterol
29	BPQ150A	Have you had any of the following in the past 30 minutes?: Food	BPX_G	Blood Pressure
30	BPQ150B	Have you had any of the following in the past 30 minutes?: Alcohol	BPX_G	Blood Pressure
31	BPQ150C	Have you had any of the following in the past 30 minutes?: Coffee	BPX_G	Blood Pressure

32	BPQ150D	Have you had any of the following in the past 30 minutes?: Cigarettes	BPX_G	Blood Pressure
33	BPXDI1	Diastolic: Blood pressure (first reading) mm Hg	BPX_G	Blood Pressure
34	BPXDI2	Diastolic: Blood pressure (second reading) mm Hg	BPX_G	Blood Pressure
35	BPXDI3	Diastolic: Blood pressure (third reading) mm Hg	BPX_G	Blood Pressure
36	BPXML1	MIL: maximum inflation levels (mm Hg)	BPX_G	Blood Pressure
37	BPXODI1	Diastolic - 1st oscillometric reading	BPXO_J	Blood Pressure - Oscillometric Measurements
38	BPXODI2	Diastolic - 2nd oscillometric reading	BPXO_J	Blood Pressure - Oscillometric Measurements
39	BPXODI3	Diastolic - 3rd oscillometric reading	BPXO_J	Blood Pressure - Oscillometric Measurements
40	BPXOPLS1	Pulse - 1st oscillometric reading	BPXO_J	Blood Pressure - Oscillometric Measurements
41	BPXOPLS2	Pulse - 2nd oscillometric reading	BPXO_J	Blood Pressure - Oscillometric Measurements
42	BPXOPLS3	Pulse - 3rd oscillometric reading	BPXO_J	Blood Pressure - Oscillometric Measurements
43	BPXOSY1	Systolic - 1st oscillometric reading	BPXO_J	Blood Pressure - Oscillometric Measurements
44	BPXOSY2	Systolic - 2nd oscillometric reading	BPXO_J	Blood Pressure - Oscillometric Measurements
45	BPXOSY3	Systolic - 3rd oscillometric reading	BPXO_J	Blood Pressure - Oscillometric Measurements
46	BPXPLS	60 sec. pulse (30 sec. pulse * 2)	BPX_G	Blood Pressure
47	BPXPTY	Pulse type	BPX_G	Blood Pressure
48	BPXPULS	Pulse regular or irregular?	BPX_G	Blood Pressure
49	BPXSY1	Systolic: Blood pressure (first reading) mm Hg	BPX_G	Blood Pressure
50	BPXSY2	Systolic: Blood pressure (second reading) mm Hg	BPX_G	Blood Pressure
51	BPXSY3	Systolic: Blood pressure (third reading) mm Hg	BPX_G	Blood Pressure
52	CDQ001	{Have you/Has SP} ever had any pain or discomfort in {your/her/his} chest?	CDQ_G	Cardiovascular Health
53	CDQ002	{Do you/Does she/Does he} get it when {you/she/he} walk uphill or hurry?	CDQ_G	Cardiovascular Health
54	CDQ008	Have {you/she/he} ever had a severe pain across the front of {your/her/his} chest lasting for half an hour or more?	CDQ_G	Cardiovascular Health
55	CDQ010	{Have you/Has SP} had shortness of breath either when hurrying on the level or walking up a slight hill?	CDQ_G	Cardiovascular Health

56	DMDBORN4	In what country {were you/was SP} born?	DEMO_G	Demographic Variables & Sample Weights
57	DMDEDUC2	What is the highest grade or level of school {you have/SP has} completed or the highest degree {you have/s/he has} received?	DEMO_G	Demographic Variables & Sample Weights
58	DMDMARTL	Marital status	DEMO_G	Demographic Variables & Sample Weights
59	DMDMARTZ	Marital status	P_DEMO	Demographic Variables and Sample Weights
60	DMDYRSUS	Length of time the participant has been in the US.	DEMO_G	Demographic Variables & Sample Weights
61	INDFMIN2	Total family income (reported as a range value in dollars)	DEMO_G	Demographic Variables & Sample Weights
62	INDFMPIR	A ratio of family income to poverty guidelines.	DEMO_G	Demographic Variables & Sample Weights
63	LBDHDD	Direct HDL-Cholesterol (mg/dL)	HDL_G	Cholesterol - HDL
64	LBDHDDSI	Direct HDL-Cholesterol (mmol/L)	HDL_G	Cholesterol - HDL
65	LBDLDL	LDL-cholesterol (mg/dL)	TRIGLY_G	Cholesterol - LDL & Triglycerides
66	LBDLDLM	LDL-Cholesterol, Martin-Hopkins equation (mg/dL). LBDLDLM = (LBXTC-(LBDHDD + LBXTR/Adjustable Factor), round to 0 decimal places) for LBXTR less than 400 mg/dL, and missing for LBXTR greater than 400 mg/dL. LBDHDD from public release file HDL_J	TRIGLY_J	Cholesterol - Low-Density Lipoproteins (LDL) & Triglycerides
67	LBDLDLSI	LDL-cholesterol (mmol/L)	TRIGLY_G	Cholesterol - LDL & Triglycerides
68	LBDLDMSI	LDL-Cholesterol, Martin-Hopkins equation (mmol/L)	TRIGLY_J	Cholesterol - Low-Density Lipoproteins (LDL) & Triglycerides
69	LBDLDNSI	LDL-Cholesterol, NIH equation 2 (mmol/L)	TRIGLY_J	Cholesterol - Low-Density Lipoproteins (LDL) & Triglycerides

70	LBDSBUSI	Blood urea nitrogen (mmol/L)	BIOPRO_G	Standard Biochemistry Profile
71	LBDSCASI	Total calcium (mmol/L)	BIOPRO_G	Standard Biochemistry Profile
72	LBDSCHSI	Cholesterol (mmol/L)	BIOPRO_G	Standard Biochemistry Profile
73	LBDSCRSI	Creatinine (umol/L)	BIOPRO_G	Standard Biochemistry Profile
74	LBDSGLSI	Glucose, serum (mmol/L)	BIOPRO_G	Standard Biochemistry Profile
75	LBDSIRSI	Iron, refigerated (umol/L)	BIOPRO_G	Standard Biochemistry Profile
76	LBDSPHSI	Phosphorus (mmol/L)	BIOPRO_G	Standard Biochemistry Profile
77	LBDSTBSI	Total bilirubin (umol/L)	BIOPRO_G	Standard Biochemistry Profile
78	LBDSTPSI	Total protein (g/L)	BIOPRO_G	Standard Biochemistry Profile
79	LBDSTRSI	Triglycerides (mmol/L)	BIOPRO_G	Standard Biochemistry Profile
80	LBDSUASI	Uric acid (umol/L)	BIOPRO_G	Standard Biochemistry Profile
81	LBDTCSI	Total Cholesterol(mmol/L)	TCHOL_G	Cholesterol - Total
82	LBDTRSI	Triglyceride (mmol/L)	TRIGLY_G	Cholesterol - LDL & Triglycerides
83	LBXSBU	Blood urea nitrogen (mg/dL)	BIOPRO_G	Standard Biochemistry Profile
84	LBXSCA	Total calcium (mg/dL)	BIOPRO_G	Standard Biochemistry Profile
85	LBXSCH	Cholesterol (mg/dL)	BIOPRO_G	Standard Biochemistry Profile
86	LBXSCLSI	Chloride (mmol/L)	BIOPRO_G	Standard Biochemistry Profile
87	LBXSGL	Glucose, serum (mg/dL)	BIOPRO_G	Standard Biochemistry Profile
88	LBXSIR	Iron, refigerated (ug/dL)	BIOPRO_G	Standard Biochemistry Profile
89	LBXSNASI	Sodium (mmol/L)	BIOPRO_G	Standard Biochemistry Profile

90	LBXSTB	Total bilirubin (mg/dL)	BIOPRO_G	Standard Biochemistry Profile
91	LBXSTP	Total protein (g/dL)	BIOPRO_G	Standard Biochemistry Profile
92	LBXSTR	Triglycerides (mg/dL)	BIOPRO_G	Standard Biochemistry Profile
93	LBXTC	Total Cholesterol(mg/dL)	TCHOL_G	Cholesterol - Total
94	LBXTR	Triglyceride (mg/dL)	TRIGLY_G	Cholesterol - LDL & Triglycerides
95	MCQ080	Has a doctor or other health professional ever told {you/SP} that {you were/s/he/SP was} overweight?	MCQ_G	Medical Conditions
96	MCQ086	{Are you/is SP} on a gluten-free diet?	MCQ_G	Medical Conditions
97	MCQ160A	Has a doctor or other health professional ever told {you/SP} that {you/s/he} . . .had arthritis (ar-thry-tis)?	MCQ_G	Medical Conditions
98	MCQ160D	Has a doctor or other health professional ever told {you/SP} that {you/s/he} . . .had angina (an-gi-na), also called angina pectoris?	MCQ_G	Medical Conditions
99	MCQ160K	Has a doctor or other health professional ever told {you/SP} that {you/s/he} . . .had chronic bronchitis?	MCQ_G	Medical Conditions
100	MCQ160M	Has a doctor or other health professional ever told {you/SP} that {you/s/he} . . .had another thyroid (thigh-roid) problem?	MCQ_G	Medical Conditions
101	MCQ160N	Has a doctor or other health professional ever told {you/SP} that {you/s/he} . . .had gout?	MCQ_G	Medical Conditions
102	MCQ160O	Has a doctor or other health professional ever told {you/SP} that {you/s/he} . . .had COPD?	MCQ_H	Medical Conditions
103	MCQ170M	{Do you/Does SP} still . . . have another thyroid problem?	MCQ_G	Medical Conditions
104	MCQ180A	How old {were you/was SP} when {you were/s/he was} first told {you/s/he} . . . had arthritis?	MCQ_G	Medical Conditions
105	MCQ195	Which type of arthritis was it?	MCQ_G	Medical Conditions
106	MCQ300A	Including living and deceased, were any of {SP's/your} close biological that is, blood relatives including father, mother, sisters or brothers, ever told by a health professional that they had a heart attack or angina (an-gi-na) before the age of 50?	MCQ_G	Medical Conditions
107	MCQ300C	Including living and deceased, were any of {SP's/your} close biological that is, blood relatives including father, mother, sisters or brothers, ever told by a health professional that they had diabetes?	MCQ_G	Medical Conditions
108	MCQ365A	To lower {your/SP's} risk for certain diseases, during the past 12 months {have you/has s/he} ever been told by a doctor or health professional to: control {your/his/her} weight or lose weight?	MCQ_G	Medical Conditions

109	MCQ365B	To lower {your/SP's} risk for certain diseases, during the past 12 months {have you/has s/he} ever been told by a doctor or health professional to: increase {your/his/her} physical activity or exercise?	MCQ_G	Medical Conditions
110	MCQ365C	To lower {your/SP's} risk for certain diseases, during the past 12 months {have you/has s/he} ever been told by a doctor or health professional to: reduce the amount of sodium or salt in {your/his/her} diet?	MCQ_G	Medical Conditions
111	MCQ365D	To lower {your/SP's} risk for certain diseases, during the past 12 months {have you/has s/he} ever been told by a doctor or health professional to: reduce the amount of fat or calories in {your/his/her} diet?	MCQ_G	Medical Conditions
112	MCQ366A	During the past 12 months {have you/has s/he} ever been told by a doctor or health professional to: control {your/his/her} weight or lose weight?	MCQ_J	Medical Conditions
113	MCQ366B	During the past 12 months {have you/has s/he} ever been told by a doctor or health professional to:increase {your/his/her} physical activity or exercise?	MCQ_J	Medical Conditions
114	MCQ366C	During the past 12 months {have you/has s/he} ever been told by a doctor or health professional to:watch or reduce the amount of sodium or salt in {your/his/her} diet?	MCQ_J	Medical Conditions
115	MCQ366D	During the past 12 months {have you/has s/he} ever been told by a doctor or health professional to: watch or reduce the amount of fat or calories in {your/his/her} diet?	MCQ_J	Medical Conditions
116	MCQ370A	To lower {your/his/her} risk for certain diseases, {are you/is s/he} now doing any of the following: controlling {your/his/her} weight or losing weight?	MCQ_G	Medical Conditions
117	MCQ370B	To lower {your/his/her} risk for certain diseases, {are you/is s/he} now doing any of the following: increasing {your/his/her} physical activity or exercise?	MCQ_G	Medical Conditions
118	MCQ370C	To lower {your/his/her} risk for certain diseases, {are you/is s/he} now doing any of the following: reducing the amount of sodium or salt in {your/his/her} diet?	MCQ_G	Medical Conditions
119	MCQ370D	To lower {your/his/her} risk for certain diseases, {are you/is s/he} now doing any of the following: reducing the amount of fat or calories in {your/his/her} diet?	MCQ_G	Medical Conditions
120	MCQ371A	{Are you/Is s/he} now doing any of the following:controlling {your/his/her} weight or losing weight?	MCQ_J	Medical Conditions
121	MCQ371B	{Are you/Is s/he} now doing any of the following: increasing {your/his/her} physical activity or exercise?	MCQ_J	Medical Conditions
122	MCQ371C	{Are you/Is s/he} now doing any of the following: watching or reducing the amount of sodium or salt in {your/his/her} diet?	MCQ_J	Medical Conditions
123	MCQ371D	{Are you/Is s/he} now doing any of the following: watching or reducing the amount of fat or calories in {your/his/her} diet?	MCQ_J	Medical Conditions

124	MCQ560	Have {you/s/he} ever had gallbladder surgery?	MCQ_J	Medical Conditions
125	PEASCST1	Blood Pressure Status	BPX_G	Blood Pressure
126	PEASCTM1	Blood Pressure Time in Seconds	BPX_G	Blood Pressure
127	PHQ020	Coffee or tea with cream or sugar? [Include milk or non-dairy creamers.]	FASTQX_G	Fasting Questionnaire
128	PHQ030	Alcohol, such as beer, wine, or liquor?	FASTQX_G	Fasting Questionnaire
129	RIAGENDR	Gender of the participant.	DEMO_G	Demographic Variables & Sample Weights
130	RIDAGEYR	Age in years of the participant at the time of screening. Individuals 80 and over are topcoded at 80 years of age.	DEMO_G	Demographic Variables & Sample Weights
131	RIDEXPRG	Pregnancy status for females between 20 and 44 years of age at the time of MEC exam.	DEMO_G	Demographic Variables & Sample Weights
132	RIDRETH1	Recode of reported race and Hispanic origin information	DEMO_G	Demographic Variables & Sample Weights
133	RIDRETH3	Recode of reported race and Hispanic origin information, with Non-Hispanic Asian Category	DEMO_G	Demographic Variables & Sample Weights
134	SDDSRVYR	Data release cycle	DEMO_G	Demographic Variables & Sample Weights
135	SEQN	Respondent sequence number.	DEMO_G	Demographic Variables & Sample Weights
136	WTINTPRP	Full sample interview weight	P_DEMO	Demographic Variables and Sample Weights
137	WTMECPRP	Full sample MEC exam weight	P_DEMO	Demographic Variables and Sample Weights
138	WTSAF2YR	Fasting Subsample 2 Year MEC Weight	GLU_G	Plasma Fasting Glucose & Insulin
139	WTSAFPRP	Fasting Subsample Weight	P_TRIGLY	Cholesterol - Low-Density Lipoproteins (LDL) & Triglycerides