

Centre for Development of Advanced Computing, Mumbai

**REPORT ON**

**HEART ATTACK PREDICTION**

PG-DBDA SEP 2022

Submitted by:

Project Team 17

Chandrashekhar Bhavsar Shivangi Singh

Digvijay Pawar Rohan Nipurte

Sumeet Yadav

Mr. Abhishek Raghorte

Project Guide

1. INTRODUCTION

“CDC (Centre for disease control and prevention) is the nation’s leading science-based, data-driven, service organization that protects the public’s health. For more than 70 years. CDC is one of the major operating components of the Department of Health and Human Services.”

CDC works 24/7 to protect America from health, safety and security threats, both foreign and in the U.S. Whether diseases start at home or abroad, are chronic or acute, curable or preventable, human error or deliberate attack, CDC fights disease and supports communities and citizens to do the same.

Heart disease is the leading cause of death for people. Heart dataset collected by CDC to help people stay healthy. The diagnosis of heart disease is usually based on signs, symptoms and physical examination of the patient. There are several factors responsible for heart disease, these factors are need to be analysed for heart attack prediction.

1.1 Dataset:

The Behavioural Risk Factor Surveillance System (BRFSS) is the nation’s premier system of health-related telephone surveys that collect state data about U.S. residents regarding their health-related risk behaviours, chronic health conditions, and use of preventive services. Established in 1984 with 15 states, BRFSS now collects data in all 50 states as well as the District of Columbia and three U.S. territories. BRFSS completes more than 400,000 adult interviews each year, making it the largest continuously conducted health survey system in the world.

By collecting behavioural health risk data at the state and local level, BRFSS has become a powerful tool for targeting and building health promotion activities. As a result, BRFSS users have increasingly demanded more data and asked for more questions on the survey. Currently, there is a wide sponsorship of the BRFSS survey, including most divisions in the CDC National Centre for Chronic Disease Prevention and Health Promotion; other CDC centre’s; and federal agencies, such as the Health Resources and Services Administration, Administration on Aging, Department of Veterans Affairs, and Substance Abuse and Mental Health Services Administration.

This Dataset surveyed by CDC contains several parameters on which people have given their responses.

Dataset consists of 401958 records and 279 features, which is surveyed data of 2020 and 2021.

* Heart.csv

Contains 401958 records and following fields.

* \_STATE : State FIPS Code
* FMONTH : File Month
* IDATE : Interview Date
* IMONTH : Interview Month
* IDAY : Interview Day
* IYEAR : Interview Year
* DISPCODE : Final Disposition
* SEQNO : Annual Sequence Number
* \_PSU : Primary Sampling Unit
* CTELENM1 : is this (phone number) ?
* PVTRESD1 : Is this a private residence?
* COLGHOUS : Do you live in college housing?
* STATERE1 : Do you currently live in \_\_\_\_(state)\_\_\_\_?
* CELPHONE : Is this a cell telephone?
* LADULT1 : Are you 18 years of age or older?
* COLGSEX : Are you male or female?
* NUMADULT : I need to randomly select one adult who lives in your household to be interviewed. Excluding adults living away from home, such as students away at college, how many members of your household, including yourself, are 18 years of age or older?
* LANDSEX : Are you male or female?
* NUMMEN : How many of these adults are men?
* NUMWOMEN : So the number of women in the household is [X]. Is that correct?
* RESPSLCT : The person in your household that I need to speak with is [XXX]. Are you the [XXX] in this household
* SAFETIME : Is this a safe time to talk with you?
* CTELNUM1 : Is this (phone number) ?
* CELLFON5 : Is this a cell phone?
* CADULT1 : Are you 18 years of age or older?
* CELLSEX : Are you male or female?
* PVTRESD3 : Do you live in a private residence? (By private residence, we mean someplace like a house or apartment.)
* CCLGHOUS : Do you live in college housing?
* CSTATE1 : Do you currently live in \_\_\_\_(state)\_\_\_\_?
* LANDLINE : Do you also have a landline telephone in your home that is used to make and receive calls?
* HHADULT : How many members of your household, including yourself, are 18 years of age or older?
* SEXVAR : Sex of Respondent
* GENHLTH : Would you say that in general your health is:
* PHYSHLTH : Now thinking about your physical health, which includes physical illness and injury, for how many days during the past 30 days was your physical health not good?
* MENTHLTH : Now thinking about your mental health, which includes stress, depression, and problems with emotions, for how many days during the past 30 days was your mental health not good?
* POORHLTH : During the past 30 days, for about how many days did poor physical or mental health keep you from doing your usual activities, such as self-care, work, or recreation?
* HLTHPLN1 : Do you have any kind of health care coverage, including health insurance, prepaid plans such as HMOs, or government plans such as Medicare, or Indian Health Service?
* PERSDOC2 : Do you have one person you think of as your personal doctor or health care provider?
* MEDCOST : Was there a time in the past 12 months when you needed to see a doctor but could not because of cost?
* CHECKUP1 : About how long has it been since you last visited a doctor for a routine checkup?
* EXERANY2 : During the past month, other than your regular job, did you participate in any physical activities or exercises such as running, calisthenics, golf, gardening, or walking for exercise?
* SLEPTIM1 : On average, how many hours of sleep do you get in a 24-hour period?
* CVDINFR4 : (Ever told) you had a heart attack, also called a myocardial infarction?
* CVDCRHD4 : (Ever told) (you had) angina or coronary heart disease?
* CVDSTRK3 : (Ever told) (you had) a stroke.
* ASTHMA3 : (Ever told) (you had) asthma?
* ASTHNOW : Do you still have asthma?
* CHCSCNCR : (Ever told) (you had) skin cancer?
* CHCOCNCR : (Ever told) (you had) any other types of cancer?
* CHCCOPD2 : (Ever told) (you had) chronic obstructive pulmonary disease, C.O.P.D., emphysema or chronic bronchitis?
* HAVARTH4 : (Ever told) (you had) some form of arthritis, rheumatoid arthritis, gout, lupus, or fibromyalgia?
* ADDEPEV3 : (Ever told) (you had) a depressive disorder (including depression, major depression, dysthymia, or minor depression)?
* CHCKDNY2 : Not including kidney stones, bladder infection or incontinence, were you ever told you had kidney disease?
* DIABETE4 : (Ever told) (you had) diabetes?
* DIABAGE3 : How old were you when you were told you had diabetes?
* LASTDEN4 : Including all types of dentists, such as orthodontists, oral surgeons, and all other dental specialists, as well as dental hygienists, how long has it been since you last visited a dentist or a dental clinic for any reason?
* RMVTETH4 : Not including teeth lost for injury or orthodontics, how many of your permanent teeth have been removed because of tooth decay or gum disease?
* MARITAL : Are you: (marital status)
* EDUCA : What is the highest grade or year of school you completed?
* RENTHOM1 : Do you own or rent your home?
* NUMHHOL3 : Not including cell phones or numbers used for computers, fax machines or security systems, do you have more than one telephone number in your household?
* NUMPHON3 : How many of these telephone numbers are residential numbers?
* CPDEMO1B : How many cell phones do you have for personal use?
* VETERAN3 : Have you ever served on active duty in the United States Armed Forces, either in the regular military or in a National Guard or military reserve unit?
* EMPLOY1 : Are you currently ?
* CHILDREN : How many children less than 18 years of age live in your household?
* INCOME2 : Is your annual household income from all sources: (If respondent refuses at any income level, code ´Refused.´)
* PREGNANT : To your knowledge, are you now pregnant?
* WEIGHT2 : About how much do you weigh without shoes?
* HEIGHT3 : About how tall are you without shoes?
* DEAF : Are you deaf or do you have serious difficulty hearing?
* BLIND : Are you blind or do you have serious difficulty seeing, even when wearing glasses?
* DECIDE : Because of a physical, mental, or emotional condition, do you have serious difficulty concentrating, remembering, or making decisions?
* DIFFWALK : Do you have serious difficulty walking or climbing stairs?
* DIFFDRES : Do you have difficulty dressing or bathing?
* DIFFALON : Because of a physical, mental, or emotional condition, do you have difficulty doing errands alone such as visiting a doctor´s office or shopping?
* SMOKE100 : Have you smoked at least 100 cigarettes in your entire life?
* SMOKDAY2 : Do you now smoke cigarettes every day, some days, or not at all?
* STOPSMK2 : During the past 12 months, have you stopped smoking for one day or longer because you were trying to quit smoking?
* LASTSMK2 : How long has it been since you last smoked a cigarette, even one or two puffs?
* USENOW3 : Do you currently use chewing tobacco, snuff, or snus every day, some days, or not at all?
* ALCDAY5 : During the past 30 days, how many days per week or per month did you have at least one drink of any alcoholic beverage such as beer, wine, a malt beverage or liquor?
* AVEDRNK3 : One drink is equivalent to a 12-ounce beer, a 5-ounce glass of wine, or a drink with one shot of liquor. During the past 30 days, on the days when you drank, about how many drinks did you drink on the average?
* DRNK3GE5 : Considering all types of alcoholic beverages, how many times during the past 30 days did you have 5 or more drinks for men or 4 or more drinks for women on an occasion?
* MAXDRNKS : During the past 30 days, what is the largest number of drinks you had on any occasion?
* FLUSHOT7 : During the past 12 months, have you had either flu vaccine that was sprayed in your nose or flu shot injected into your arm?
* FLSHTMY3 : During what month and year did you receive your most recent flu vaccine that was sprayed in your nose or flu shot injected into your arm?
* SHINGLE2 : Have you ever had the shingles or zoster vaccine?
* PNEUVAC4 : Have you ever had a pneumonia shot also known as a pneumococcal vaccine?
* FALL12MN : In the past 12 months, how many times have you fallen?
* FALLINJ4 : How many of these falls caused an injury that limited your regular activities for at least a day?
* SEATBELT : How often do you use seat belts when you drive or ride in a car? Would you say
* DRNKDRI2 : During the past 30 days, how many times have you driven when you’ve had perhaps too much to drink?
* HADMAM : Have you ever had a mammogram?
* HOWLONG : How long has it been since you had your last mammogram?
* HADPAP2 : Have you ever had a Pap test?
* LASTPAP2 : How long has it been since you had your last Pap test?
* HPVTEST : An H.P.V. test is sometimes given with the Pap test for cervical cancer screening. Have you ever had an H.P.V. test?
* HPLSTTST : How long has it been since you had your last H.P.V. test?
* HADHYST2 : Have you had a hysterectomy?
* PCPSAAD3 : Has a doctor, nurse, or other health professional ever talked with you about the advantages of the ProstateSpecific Antigen or P.S.A. test
* PCPSADI1 : Has a doctor, nurse, or other health professional ever talked with you about the disadvantages of the PSA test?
* PCPSARE1 : Has a doctor, nurse, or other health professional ever recommended that you have a P.S.A. test?
* PSATEST1 : Have you ever had a P.S.A. test?
* PSATIME : How long has it been since you had your last P.S.A. test?
* PCPSARS1 : What was the MAIN reason you had this P.S.A. test – was it?
* COLNSCPY : Have you ever had a colonoscopy?
* COLNTEST : How long has it been since you had this test?
* SIGMSCPY : Have you ever had a sigmoidoscopy?
* SIGMTEST : How long has it been since you had this test?
* BLDSTOL1 : Have you ever had this test using a home kit?
* LSTBLDS4 : How long has it been since you had this test?
* STOOLDNA : Have you ever had this test?
* SDNATEST : How long has it been since you had this test?
* VIRCOLON : Have you ever had a virtual colonoscopy?
* VCLNTEST : How long has it been since you had this test?
* HIVTST7 : Including fluid testing from your mouth, but not including tests you may have had for blood donation, have you ever been tested for H.I.V?
* HIVTSTD3 : Not including blood donations, in what month and year was your last H.I.V. test?
* HIVRISK5 : I am going to read you a list. When I am done, please tell me if any of the situations apply to you. You do not need to tell me which one. You have injected any drug other than those prescribed for you in the past year. You have been treated for a sexually transmitted disease or STD in the past year. You have given or received money or drugs in exchange for sex in the past year.
* PDIABTST : Have you had a test for high blood sugar or diabetes within the past three years?
* PREDIAB1 : Have you ever been told by a doctor or other health professional that you have pre-diabetes or borderline diabetes?
* INSULIN1 : Are you now taking insulin?
* BLDSUGAR : About how often do you check your blood for glucose or sugar?
* FEETCHK3 : Including times when checked by a family member or friend, about how often do you check your feet for any sores or irritations?
* DOCTDIAB : About how many times in the past 12 months have you seen a doctor, nurse, or other health professional for your diabetes?
* CHKHEMO3 : About how many times in the past 12 months has a doctor, nurse, or other health professional checked you for A-one-C?
* FEETCHK : About how many times in the past 12 months has a health professional checked your feet for any sores or irritations?
* EYEEXAM1 : When was the last time you had an eye exam in which the pupils were dilated, making you temporarily sensitive to bright light?
* DIABEYE : Has a doctor ever told you that diabetes has affected your eyes or that you had retinopathy?
* DIABEDU : Have you ever taken a course or class in how to manage your diabetes yourself?
* TOLDCFS : Have you ever been told by a doctor or other health professional that you had Chronic Fatigue Syndrome (CFS) or (Myalgic Encephalomyelitis) ME?
* HAVECFS : Do you still have Chronic Fatigue Syndrome (CFS) or (Myalgic Encephalomyelitis) ME?
* WORKCFS : Thinking about your CFS or ME, during the past 6 months, how many hours a week on average have you been able to work at a job or business for pay?
* TOLDHEPC : Have you ever been told by a doctor or other health professional that you had Hepatitis C?
* TRETHEPC : Were you treated for Hepatitis C in 2015 or after?
* PRIRHEPC : Were you treated for Hepatitis C prior to 2015?
* HAVEHEPC : Do you still have Hepatitis C?
* HAVEHEPB : Has a doctor, nurse, or other health professional ever told you that you had hepatitis B?
* MEDSHEPB : Are you currently taking medicine to treat hepatitis B?
* HLTHCVR1 : What is the primary source of your health care coverage?
* CIMEMLOS : During the past 12 months, have you experienced confusion or memory loss that is happening more often or is getting worse?
* CDHOUSE : During the past 12 months, as a result of confusion or memory loss, how often have you given up day-to-day household activities or chores you used to do, such as cooking, cleaning, taking medications, driving, or paying bills?
* CDASSIST : As a result of confusion or memory loss, how often do you need assistance with these day-to-day activities?
* CDHELP : When you need help with these day-to-day activities, how often are you able to get the help that you need?
* CDSOCIAL : During the past 12 months, how often has confusion or memory loss interfered with your ability to work, volunteer, or engage in social activities outside the home?
* CDDISCUS : Have you or anyone else discussed your confusion or memory loss with a health care professional?
* CAREGIV1 : During the past 30 days, did you provide regular care or assistance to a friend or family member who has a health problem or disability?
* CRGVREL4 : What is his or her relationship to you?
* CRGVLNG1 : For how long have you provided care for that person?
* CRGVHRS1 : In an average week, how many hours do you provide care or assistance?
* CRGVPRB3 : What is the main health problem, long-term illness, or disability that the person you care for has?
* CRGVALZD : Does the person you care for also have Alzheimer´s disease, dementia or other cognitive impairment disorder?
* CRGVPER1 : In the past 30 days, did you provide care for this person by managing personal care such as giving medications, feeding, dressing, or bathing?
* CRGVHOU1 : In the past 30 days, did you provide care for this person by managing household tasks such as cleaning, managing money, or preparing meals?
* CRGVEXPT : n the next 2 years, do you expect to provide care or assistance to a friend or family member who has a health problem or disability?
* ECIGARET : Have you ever used an e-cigarette or other electronic vaping product, even just one time, in your entire life?
* ECIGNOW : Do you now use e-cigarettes or other electronic vaping products every day, some days, or not at all?
* MARIJAN1 : During the past 30 days, on how many days did you use marijuana or cannabis?
* USEMRJN2 : During the past 30 days, which one of the following ways did you use marijuana the most often? Did you usually
* RSNMRJN1 : When you used marijuana or cannabis during the past 30 days, was it usually
* LCSFIRST : How old were you when you first started to smoke cigarettes regularly
* LCSLAST : How old were you when you last smoked cigarettes regularly?
* LCSNUMCG : On average, when you smoke/smoked regularly, about how many cigarettes do/did you usually smoke each day?
* LCSCTSCN : In the last 12 months, did you have a CT or CAT scan?
* CNCRDIFF : How many different types of cancer have you had?
* CNCRAGE : At what age were you told that you had cancer?
* CNCRTYP1 : What type of cancer was it?
* CSRVTRT3 : Are you currently receiving treatment for cancer?
* CSRVDOC1 : What type of doctor provides the majority of your health care?
* CSRVSUM : Did any doctor, nurse, or other health professional ever give you a written summary of all the cancer treatments that you received?
* CSRVRTRN : Have you ever received instructions from a doctor, nurse, or other health professional about where you should return or who you should see for routine cancer check-ups after completing treatment for cancer?
* CSRVINST : Were these instructions written down or printed on paper for you?
* CSRVINSR : With your most recent diagnosis of cancer, did you have health insurance that paid for all or part of your cancer treatment?
* CSRVDEIN : Were you ever denied health insurance or life insurance coverage because of your cancer?
* CSRVCLIN : Did you participate in a clinical trial as part of your cancer treatment?
* CSRVPAIN : Do you currently have physical pain caused by your cancer or cancer treatment?
* CSRVCTL2 : Would you say your pain is currently under control…?
* PCPSADE1 : Which of the following best describes the decision to have the P.S.A. test done?
* PCDMDEC1 : Who made the decision with you?
* HPVADVC4 : Have you ever had an H.P.V. vaccination?
* HPVADSHT : How many HPV shots did you receive?
* TETANUS1 : Have you received a tetanus shot in the past 10 years?
* IMFVPLA1 : At what kind of place did you get your last flu shot or vaccine?
* BIRTHSEX : What was your sex at birth? Was it male or female?
* SOMALE : Which of the following best represents how you think of yourself?
* SOFEMALE : Which of the following best represents how you think of yourself?
* TRNSGNDR : Do you consider yourself to be transgender?
* ACEDEPRS : Did you live with anyone who was depressed, mentally ill, or suicidal?
* ACEDRINK : Did you live with anyone who was a problem drinker or alcoholic?
* ACEDRUGS : Did you live with anyone who used illegal street drugs or who abused prescription medications?
* ACEPRISN : Did you live with anyone who served time or was sentenced to serve time in a prison, jail, or other correctional facility?
* ACEDIVRC : Were your parents separated or divorced?
* ACEPUNCH : How often did your parents or adults in your home ever slap, hit, kick, punch or beat each other up?
* ACEHURT1 : Not including spanking, (before age 18), how often did a parent or adult in your home ever hit, beat, kick, or physically hurt you in any way?
* ACESWEAR : How often did a parent or adult in your home ever swear at you, insult you, or put you down?
* ACETOUCH : How often did anyone at least 5 years older than you or an adult, ever touch you sexually?
* ACETTHEM : How often did anyone at least 5 years older than you or an adult, try to make you touch them sexually?
* ACEHVSEX : How often did anyone at least 5 years older than you or an adult, force you to have sex?
* RCSGENDR : Is the child a boy or a girl?
* RCSRLTN2 : How are you related to the child?
* CASTHDX2 : Has a doctor, nurse or other health professional EVER said that the child has asthma?
* CASTHNO2 : Does the child still have asthma?
* QSTVER : Questionnaire Version Identifier
* QSTLANG : Language identifier
* \_METSTAT : Metropolitan Status
* \_URBSTAT : Urban/Rural Status
* MSCODE : Metropolitan Status Code
* \_STSTR : Sample Design Stratification Variable
* \_STRWT : Stratum weight
* \_RAWRAKE : Raw weighting factor used in raking
* \_WT2RAKE : Design weight used in raking (Stratum weight (\_STRWT) multiplied by the raw weighting factor (\_RAWRAKE).
* \_IMPRACE : Imputed race/ethnicity value
* \_CHISPNC : Child Hispanic, Latino/a, or Spanish origin calculated variable
* \_CRACE1 : Child multiracial race categorization
* \_CPRACE : Preferred Child Race Categories
* \_CLLCPWT : Final child weight: Land-line and Cell-Phone data
* \_DUALUSE : Dual Phone Use Categories
* \_DUALCOR : Dual phone use correction factor
* \_LLCPWT2 : Truncated design weight used in adult combined land line and cell phone raking
* \_LLCPWT : Final weight assigned to each respondent: Land-line and cell-phone data
* \_RFHLTH : Adults with good or better health
* \_PHYS14D : 3 level not good physical health status: 0 days, 1-13 days, 14-30 days
* \_MENT14D : 3 level not good mental health status: 0 days, 1-13 days, 14-30 days
* \_HCVU651 : Respondents aged 18-64 who have any form of health care coverage
* \_TOTINDA : Adults who reported doing physical activity or exercise during the past 30 days other than their regular job
* \_MICHD : Respondents that have ever reported having coronary heart disease (CHD) or myocardial infarction (MI)
* \_LTASTH1 : Adults who have ever been told they have asthma
* \_CASTHM1 : Adults who have been told they currently have asthma
* \_ASTHMS1 : Computed asthma status
* \_DRDXAR2 : Respondents who have had a doctor diagnose them as having some form of arthritis
* \_EXTETH3 : Adults aged 18+ who have had permanent teeth extracted
* \_ALTETH3 : Adults aged 65+ who have had all their natural teeth extracted
* \_DENVST3 : Adults who have visited a dentist, dental hygenist or dental clinic within the past year
* \_PRACE1 : Preferred race category
* \_MRACE1 : Calculated multiracial race categorization
* \_HISPANC : Hispanic, Latino/a, or Spanish origin calculated variable
* \_RACE : Race/ethnicity categories
* \_RACEG21 : White non-Hispanic race group
* \_RACEGR3 : Five-level race/ethnicity category
* \_RACEPRV : Computed race groups used for internet prevalence tables
* \_SEX : Calculated sex variable
* \_AGEG5YR : Fourteen-level age category
* \_AGE65YR : Two-level age category
* \_AGE80 : Imputed Age value collapsed above 80
* \_AGE\_G : Six-level imputed age category
* HTIN4 : Reported height in inches
* HTM4 : Reported height in meters
* WTKG3 : Reported weight in kilograms
* \_BMI5 : Body Mass Index (BMI)
* \_BMI5CAT : Four-categories of Body Mass Index (BMI)
* \_RFBMI5 : Adults who have a body mass index greater than 25.00 (Overweight or Obese)
* \_CHLDCNT : Number of children in household
* \_EDUCAG : Level of education completed
* \_INCOMG : Income categories
* \_SMOKER3 : Four-level smoker status: Everyday smoker, Someday smoker, Former smoker, Non-smoker
* \_RFSMOK3 : Adults who are current smokers
* DRNKANY5 : Adults who reported having had at least one drink of alcohol in the past 30 days
* DROCDY3\_ : Drink-occasions-per-day
* \_RFBING5 : Binge drinkers
* \_DRNKWK1 : Calculated total number of alcoholic beverages consumed per week
* \_RFDRHV7 : Heavy drinkers
* \_FLSHOT7 : Adults aged 65+ who have had a flu shot within the past year
* \_PNEUMO3 : Adults aged 65+ who have ever had a pneumonia vaccination
* \_RFSEAT2 : Always or Nearly Always Wear Seat Belts Calculated Variable
* \_RFSEAT3 : Always Wear Seat Belts Calculated Variable
* \_DRNKDRV : Drinking and Driving
* \_RFMAM22 : Women respondents aged 40+ who have had a mammogram in the past two years
* \_MAM5023 : Women respondents aged 50-74 who have had a mammogram in the past two years
* \_RFPAP35 : Women respondents aged 21-65 who have had a pap test in the past three years
* \_RFPSA23 : Male respondents aged 40+ who have had a PSA test in the past 2 years
* \_CLNSCPY : Respondents aged 50-75 who have had a colonoscopy within the past ten years
* \_SGMSCPY : Respondents aged 50-75 who have had a sigmoidoscopy within the past five years
* \_SGMS10Y : Respondents aged 50-75 who have had a sigmoidoscopy within the past ten years
* \_RFBLDS4 : Respondents aged 50-75 who have had a blood stool test within the past year
* \_STOLDNA : Respondents aged 50-75 who have had a stool DNA test within the past three years
* \_VIRCOLN : Respondents aged 50-75 who have had a virtual colonoscopy within the past five years
* \_SBONTIM : Respondents aged 50-75 who have had a sigmoidoscopy within the past ten years and a blood stool test in the past year
* \_CRCREC1 : Respondents aged 50-75 who have fully met the USPSTF recommendations
* \_AIDTST4 : Adults who have ever been tested

# PROBLEM STATEMENT

Cardiovascular diseases (CVDs) are the number 1 cause of death globally, taking an estimated 17.9 million lives each year, which accounts for 31% of all deaths worldwide. Four out of 5CVD deaths are due to heart attacks and strokes, and one-third of these deaths occur prematurely in people under 70 years of age. Heart failure is a common event caused by CVDs and this dataset contains 21 features that can be used to predict a possible heart disease.

People with cardiovascular disease or who are at high cardiovascular risk (due to the presence of one or more risk factors such as hypertension, diabetes, hyperlipidaemia or already established disease) need early detection and management wherein a machine learning model can be of great help.

2.1 Objective:

Our aim in this project is to build a Machine Learning model which predicts coronary heart disease chances. With this model, people can take needed precautions and regularly visit doctor.

The correct prediction of heart disease can prevent life threats, and incorrect prediction can prove to be fatal at the same time. In this paper different machine learning algorithms are applied to compare the results and analysis of CDC Dataset.

Wherein we are predicting chances of heart attack on the basis of surveyed data.

* Performing exploratory data analysis of CDC surveyed dataset.
* Identifying the factors that impact coronary heart disease.
* Developing machine learning models to predict Chances of coronary heart disease.

# DATA PREPARATION

3.1 Missing Values:

There are 229 columns that contain missing values. Since most missing values exist because there was no information available at a specific time, fields containing missing values are left as ‘NA’.

3.2 Don’t Know and Refused Values:

There are various columns that contain don’t know and refused response from the people. Replacing those values with mode of that columns.

* 1. Feature Selection:

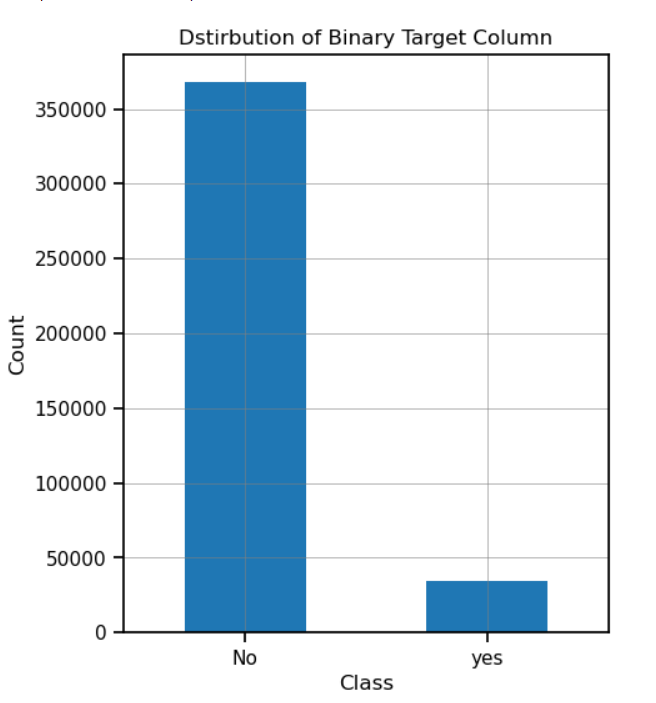
Based on Meta Data, there are 21 columns that are relevant to our model building and prediction. Out of these columns 20 columns are categorical data and 1 column is continuous data.

* 1. Binary Target Columns Analysis:

There is imbalance distribution of data in target Binary column. This bias in the training dataset can influence many machine learning algorithms, leading some to ignore the minority class entirely. This is a problem as it is typically the minority class on which predictions are most important.

One approach to addressing the problem of class imbalance is to randomly resample the training dataset.

Resampling involves creating a new transformed version of the training dataset.



* 1. Chi Square Test:

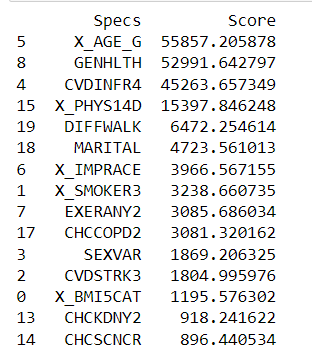
For our Algorithm we require some features, which is best for prediction, so we selected some of the columns on the basis of chi square test.

Pearson's chi-squared test is used to determine whether there is a statistically significant difference between the expected frequencies and the observed frequencies in one or more categories of a contingency table.

Below are the scores of features:



Sort listed top 15 features for our algorithm, these features are:



3.7 Data Splitting into Training, Testing:

The main difference between training data and testing data is that training data is the subset of original data that is used to train the machine learning model, whereas testing data is used to check the accuracy of the model. The training dataset is generally larger in size compared to the testing dataset.

One important aspect of all machine learning models is to determine their accuracy. Now, in order to determine their accuracy, one can train the model using the given dataset and then predict the response values for the same dataset using that model and hence, find the accuracy of the model. A better option is to split our data into two parts: first one for training our machine learning model, and second one for testing our model. Split the dataset into two pieces: a training set and a testing set. Train the model on the training set. Test the model on the testing set, and evaluate how well our model did.

Here we use train test split function to split or data set we have split 20% data as test data and remaining 80% as training data

Advantages of train/test split:

* Model can be trained and tested on different data than the one used for training.
* Response values are known for the test dataset, hence predictions can be evaluated
* Testing accuracy is a better estimate than training accuracy of out-of-sample performance.

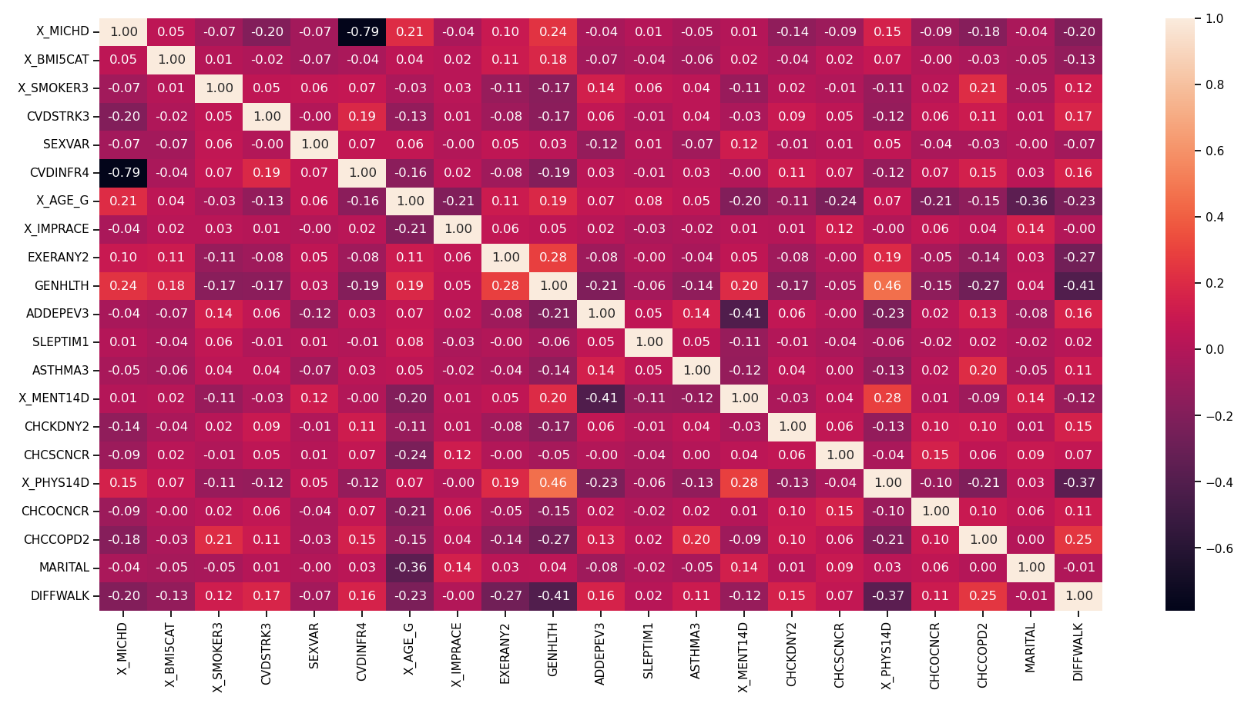
# EXPLORATORY DATA ANALYSIS

* 1. Multivariate Data Analysis:

A correlation heatmap is a heatmap that shows a 2D correlation matrix between two discrete dimensions, using coloured cells to represent data from usually a monochromatic scale. The values of the first dimension appear as the rows of the table while of the second dimension as a column. The colour of the cell is proportional to the number of measurements that match the dimensional value. This makes correlation heatmaps ideal for data analysis since it makes patterns easily readable and highlights the differences and variation in the same data. A correlation heatmap, like a regular heatmap, is assisted by a colorbar making data easily readable and comprehensible.

The heatmap bellow shows that there is a strong correlation between the following variables:

* Coronary Heart Disease and CVDINFR4(myocardial infarction)



4.2 Conclusion On EDA:

Walmart’s stores are classified into three types: A, B, and C. Type A stores often have high sales, big store’s sizes, large number of departments, and large markdown values. Type C stores often have low sales, small store sizes, small number of departments, and small markdown values. Walmart’s sales are often at peak during the week of Thanksgiving and three weeks after Thanksgiving. External factors like temperature, fuel price, consumer price index, and unemployment rate do not have significant impact on Walmart’s sales. Promotional markdown events before holidays seem to increase Walmart’s sales except for Christmas.

# FUTURE IMPROVEMENTS

Data will be made more stationary with different techniques.

More detailed feature engineering and feature selection will be done.

More data can be found to observe holiday effects on sales and different holidays will be added like Easter, Halloween and Come Back to School times.

Markdown effects on model will be improved according to department sales.

Different models can be build for special stores or departments.

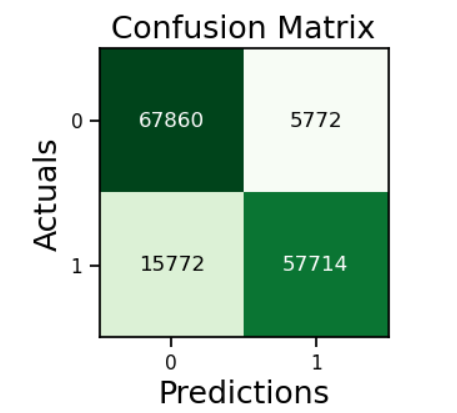
Market basket analysis can be done to find higher demand items of departments.

# MODEL TESTING

6.1 Linear Regression:

Linear regression analysis is used to predict the value of a variable based on the value of another variable. The variable you want to predict is called the dependent variable. The variable you are using to predict the other variable's value is called the independent variable.

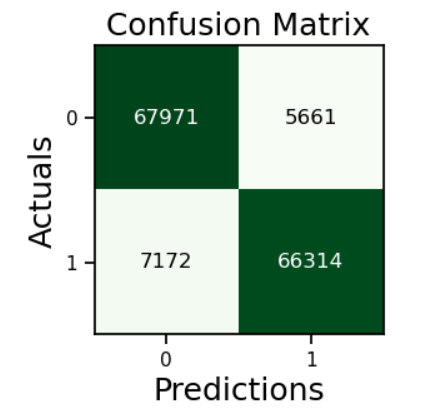
Confusion matrix Plot for Linear Regression



In this linear regression model, we got an accuracy is equal to 85.35%

6.2 Decision Tree:

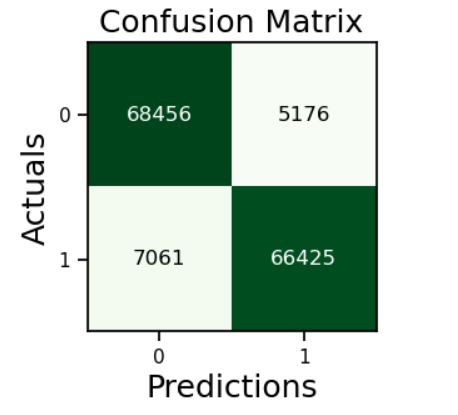
Decision Tree is the most powerful and popular tool for classification and prediction. A Decision tree is a flowchart-like tree structure, where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node (terminal node) holds a class label.



In this model we got an accuracy is equal to 97.27%

6.3 Random Forest:

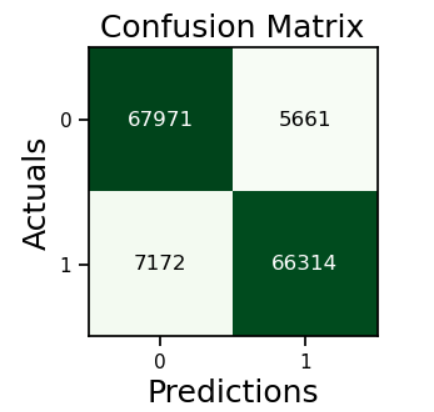
Random Forest Regression is a supervised learning algorithm that uses ensemble learning method for regression. Ensemble learning method is a technique that combines predictions from multiple machine learning algorithms to make a more accurate prediction than a single model.



In this model we got an accuracy is equal to 91.68%

6.4 Naïve Bayes:

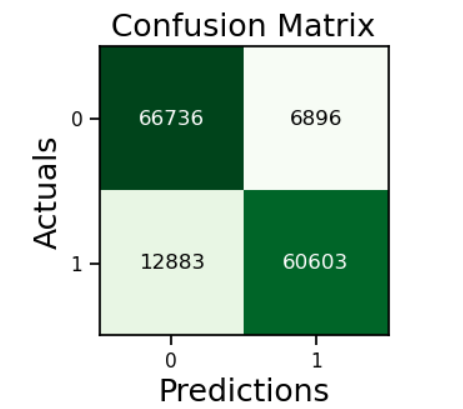
Naive Bayes classifiers are a collection of classification algorithms based on Bayes’ Theorem. It is not a single algorithm but a family of algorithms where all of them share a common principle, i.e. every pair of features being classified is independent of each other.



In this model we got an accuracy is equal to 91.27%

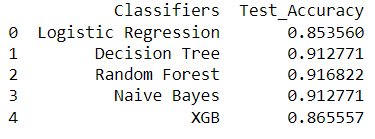
6.5 XGB Classifier:

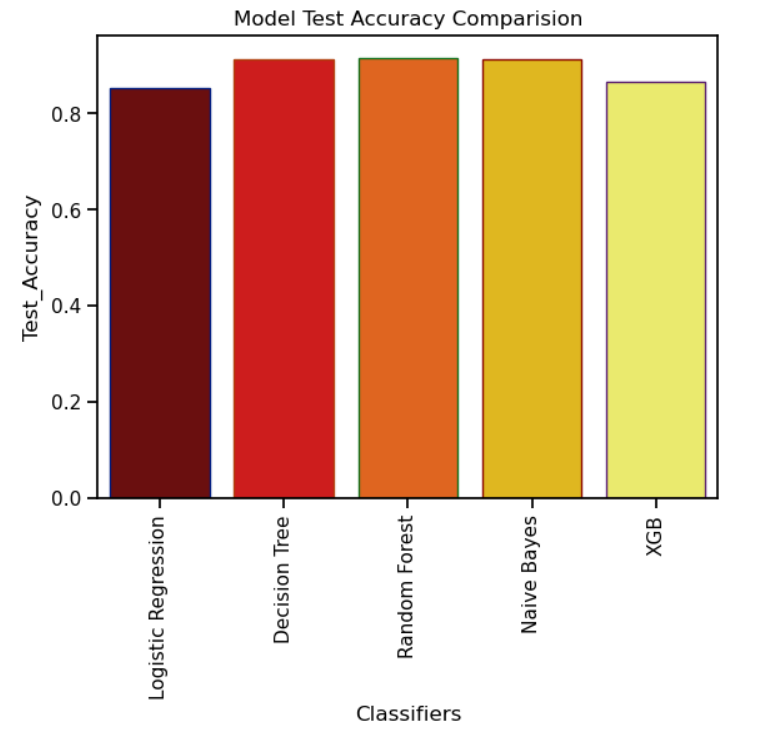
XGBoost is an optimized distributed gradient boosting library designed for efficient and scalable training of machine learning models. It is an ensemble learning method that combines the predictions of multiple weak models to produce a stronger prediction. XGBoost stands for “Extreme Gradient Boosting” and it has become one of the most popular and widely used machine learning algorithms due to its ability to handle large datasets and its ability to achieve state-of-the-art performance in many machine learning tasks such as classification and regression.



In this model we got an accuracy is equal to 86.5%

# COMPARING MODELS





From above figures we can see that random forest regression model has the highest accuracy

Which is 91.68 %

So, we will use random forest regression model for Heart Attack prediction of this datasets other models also perform well as their accuracies are:

Linear regression = 85.35 %

Decision Tree = 91.27 %

Naïve Bayes = 91.27%

XGB regression = 86.55 %

# CONCLUSION

In conclusion, we find that our regression equation is quite accurate (91.68% accuracy) in predicting the heart disease with Random Forest Regressor. People can use it to Predict their Heart Health Status. People need to focus on their Physical Health, General Health, Smoking Habits, Exercises, BMI, Kidney Disease.

# REFERENCES

Dataset link:

https://www.kaggle.com/datasets/divyajeetthakur/walmart-sales-prediction

Models:

1. Linear Regression:

https://scikit-learn.org/stable/modules/linear\_model.html#ordinary-leastsquares

1. Random Forest Regression:

https://scikit-learn.org/stable/modules/ensemble.html#forests-of-randomizedtrees

1. Decision Tree:

<https://scikit-learn.org/stable/modules/tree.html#>

1. Naïve Bayes:

<https://scikit-learn.org/stable/modules/naive_bayes.html#>