Under the supervision of:

Samsung Innovation Campus

HEART DISEASE PREDICTION

Supervised By:

Instructor: Dr. Doaa Mohamed

Facilitator: Eng. Shaimaa

AGENDA

1

Problem Statement

Address the problem we need to find a solution for

2

Key Findings

Demonstrate findings discovered upon analysis

3

Preprocessing

Describe our process in making our data suitable for conducting analysis

4

Models & Evaluation

Models tested and comparison

1

PROBLEM STATEMENT & DESCRIBTION

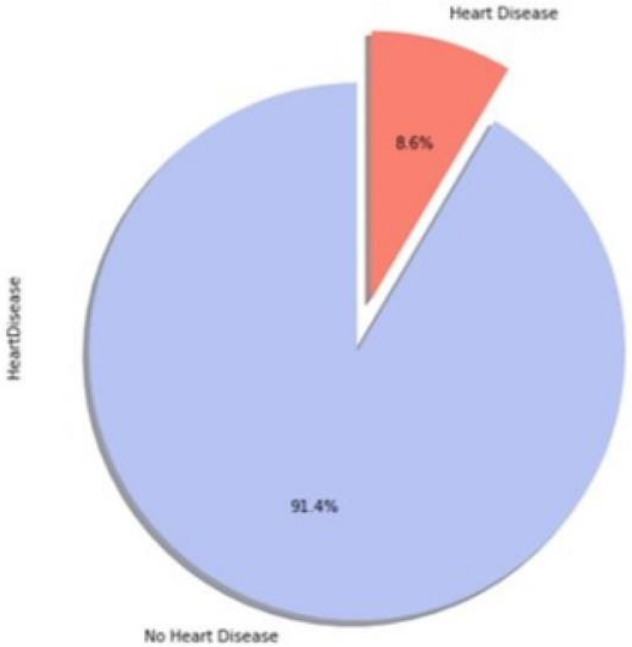
What's our problem and what do we need to achieve?

PROBLEM STATEMENT:

Cardiovascular diseases are one of the main causes of global death, taking an estimated 17.9 million lives each year, which represents about 31% of global deaths. Nowadays more people tend to depend on convenience food due to our packed schedules. Meanwhile, it has been reported that heart diseases are in constant increase the past few years and many are unaware of this.

AN OVERVIEW

- **Our Data Comes from** 2020 annual CDC survey data of 400k adults related to their health status, which conducts annual telephone surveys to gather data on the health status of U.S. residents.
- **The set consists** of 319795 rows and 18 columns.
- **The data set is** highly imbalanced and doesn't contain any missing values as we a large ratio for people without HD than others and this indicates imbalanced data.
- **Our Target is** to predict if the patient has Heart Disease or not.
- **About** 27373 of Adult have Heart disease and 292422 doesn't have Heart Disease.



DATA DESCRIPTION

1 **Heart Disease:**

Have you ever had a heart attack?

(Yes / No)

2 **BMI:** Body Mass Index

3 **Smoking :**
Have you ever smoked? (Yes / No)

4 **Alcohol Drinking :**
Have you ever drank alcohol (Yes / No)

5 **Stroke :**
Have you ever had a stroke? (Yes / No)

6 **Physical Health :**
How many days during the past 30 days was your physical health not good? (0-30 days)

7 **Mental Health :**
How many days during the past 30 days was mental health not good? (0-30 days)

8 **DiffWalking:**
Difficulty walking or climbing stairs
(Yes / No)

9 **Sex :**
Male or Female

DATA DESCRIPTION

10 **Age Category :**
Thirteen-level age category

11 **Physical Activity :**
Doing physical activity or exercise during the past 30 days other than their regular job (Yes / No)

12 **Race :**
Ethnicity

13 **Diabetic :**
Ever told you had diabetes? (Yes / No)

14 **GenHealth :**
General Health

15 **SleepTime:**
Hours of sleeping in 24-hour period

16 **Asthma :**
Ever told you had asthma? (Yes / No)

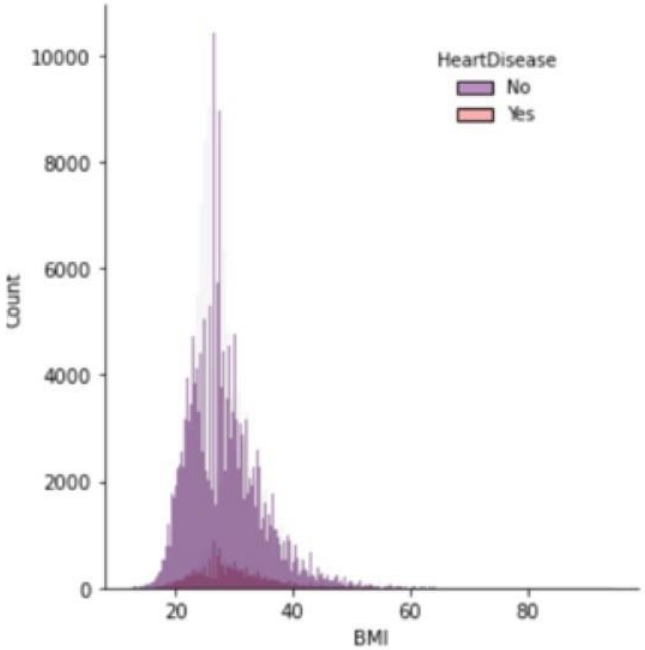
17 **Kidney Disease :**
Ever told you had kidney disease? (Yes / No)

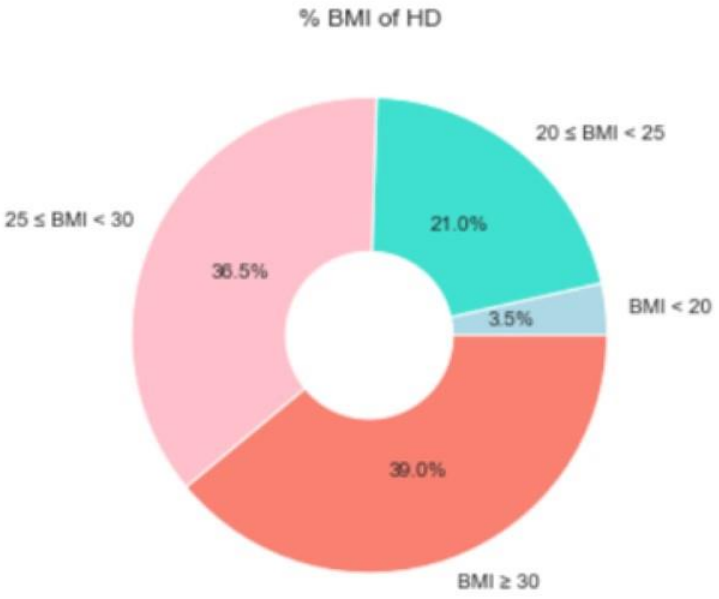
18 **Skin Cancer :**
Ever told you had skin cancer? (Yes / No)

Key Findings

A PICTURE ALWAYS REINFORCES THE CONCEPT

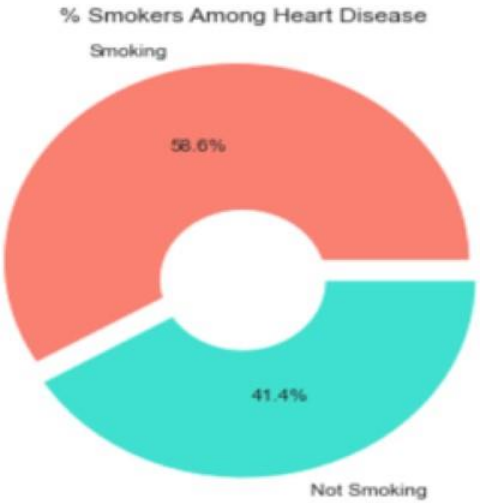
Heart Disease and BMI

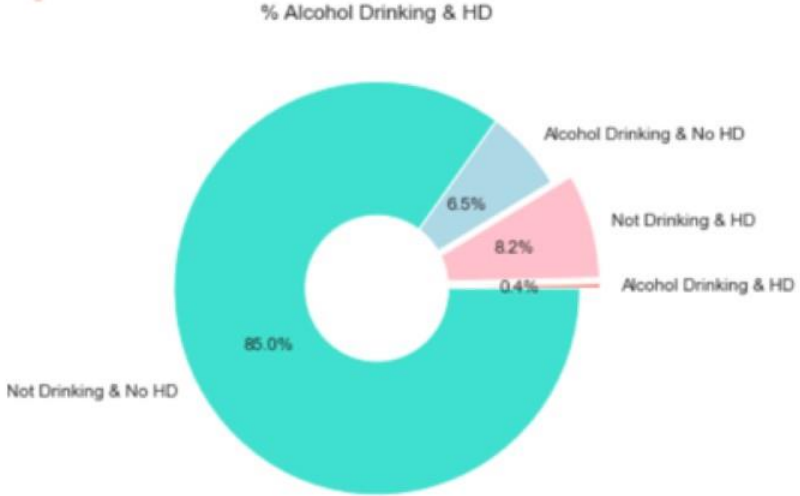




As the body shape index increases, the risk of getting the disease increases which is also stated by British Heart Foundation.(https://www.bhf.org.uk/informationsupport/riskfactors/obesity#:~:text=How%20does%20obesity%20increase%20the%20risk%20of%20heart%20and%20circulatory%20diseases%3F&text=Excess%20weight%20can%20lead%20to,lead%20to%20a%20heart%20attack.)

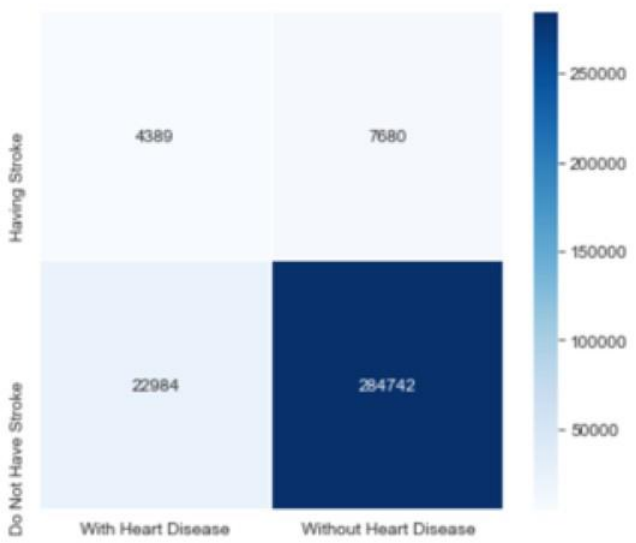
Smoking and Alcohol Drinking

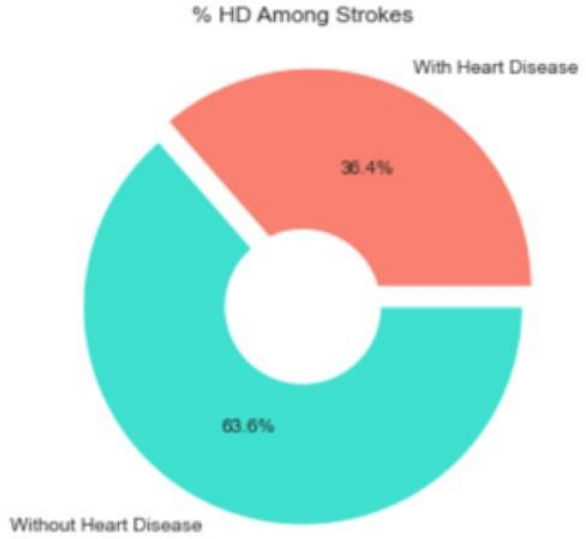




People who smokes are most vulnerable to toxic effects on health and cardiovascular systems which is proved by FDA. That's besides BHF that stated the dangerous effect of drinking alcohol on patients. (https://www.fda.gov/tobacco-products/health-effects-tobacco-use/how-smoking-affects-heart-health)(https://www.bhf.org.uk/informationsupport/heart-matters-magazine/medical/effects-of-alcohol-on-your-heart)

Strokes and Heart disease





Here are the people to have HD without strokes are small, and that is because the data isn't balanced.

According to CDC, people with HD are also vulnerable to having strokes because it's an important risk factor that affects cardiovascular diseases due to unhealthy lifestyles and physical inactivity.
(https://www.cdc.gov/chronicdisease/resources/publications/factsheets/heart-disease-stroke.htm#:~:text=Making%20blood%20sticky%20and%20more,and%20narrowing%20of%20blood%20vessels)

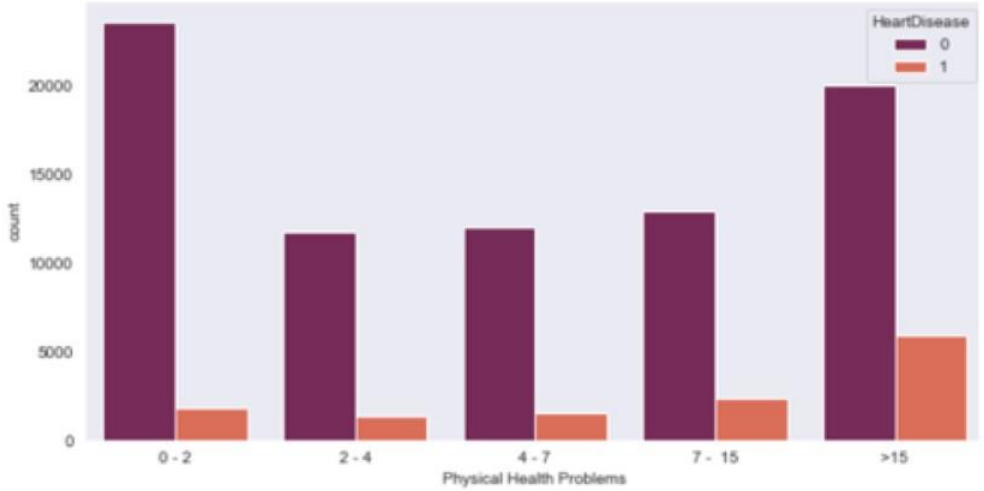
Physical Health Problems:

Symptoms of Heart disease include:

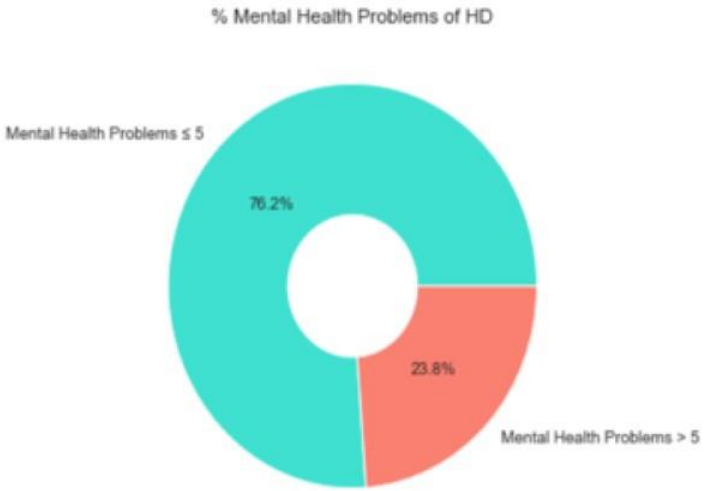
•Chest pain, chest tightness, and Angina

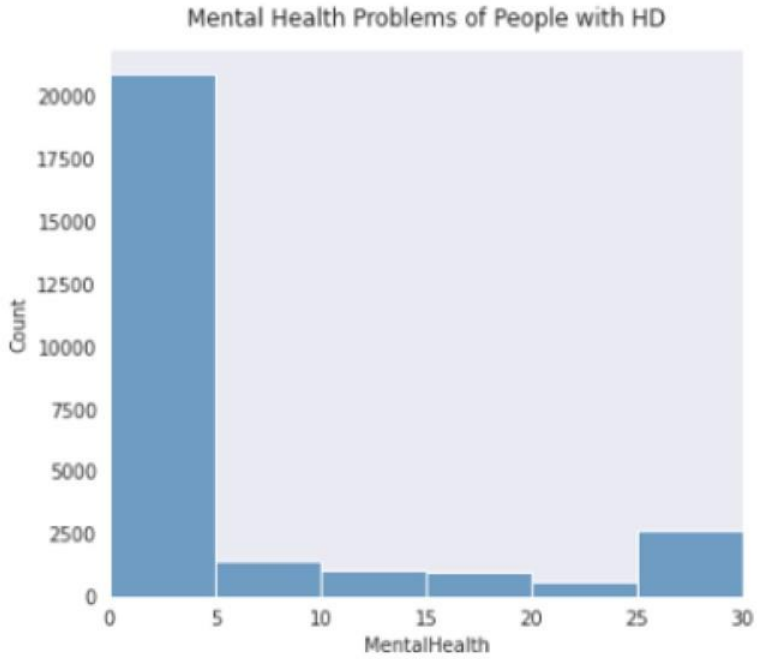
•Shortness of breath

•Pain and coldness in the legs or arms.



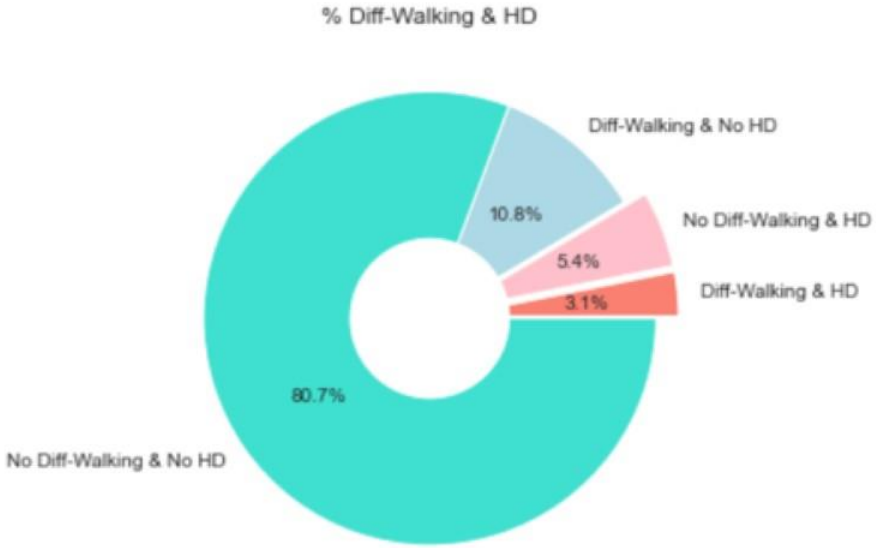
Mental Health Problems and HD:

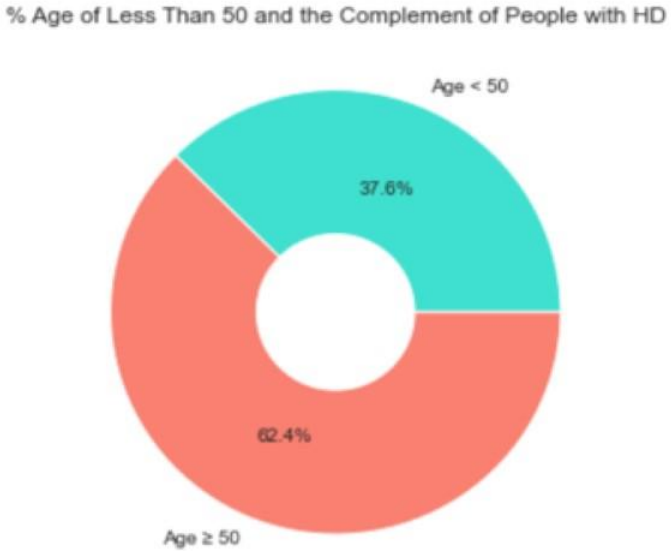




According to CDC, there is a strong relation between mental health problems and heart disease.
(https://www.cdc.gov/heartdisease/mentalhealth.htm)

Difficulty walking and Age Category

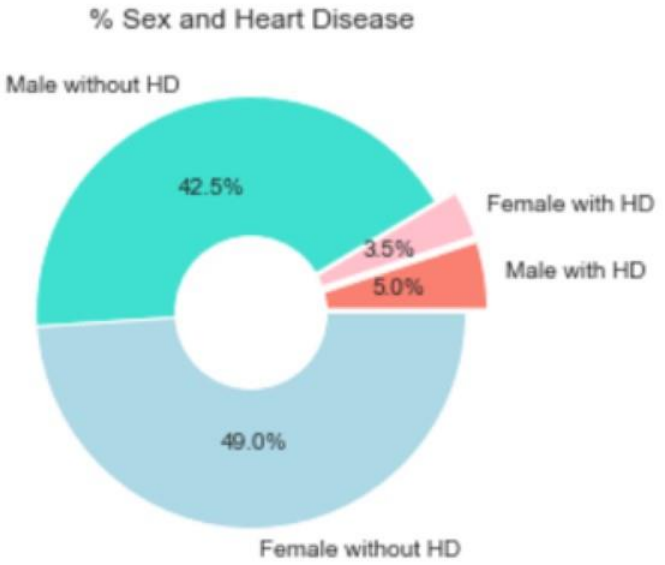




It is clear that old people are exposed to such a disease easily more than young people and sometimes they are even according to CDC (https://www.cdc.gov/heartdisease/any_age.htm)

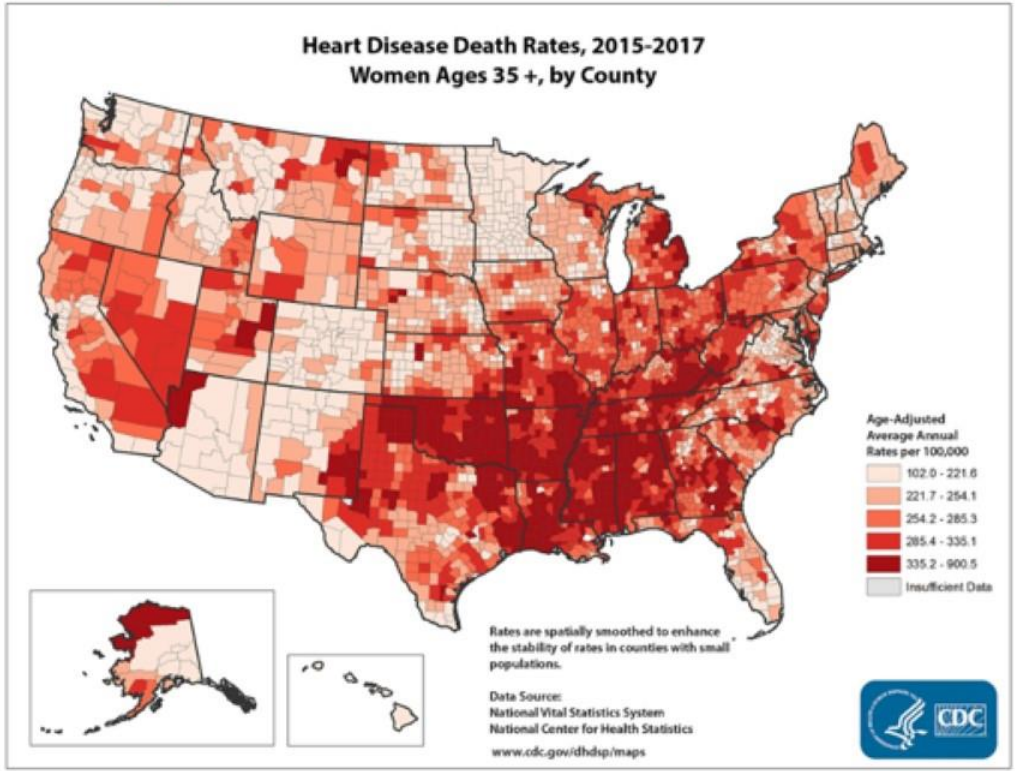
Gender and Heart Disease:

The CDC stated that despite the increases in awareness over the past decades, only about half 56% of women recognize that heart disease is their number 1 killer as HD develops 7 to 10 years later in women than men. In addition, women were more likely than men to be older and have a more complicated medical history at the time of their heart attacks.



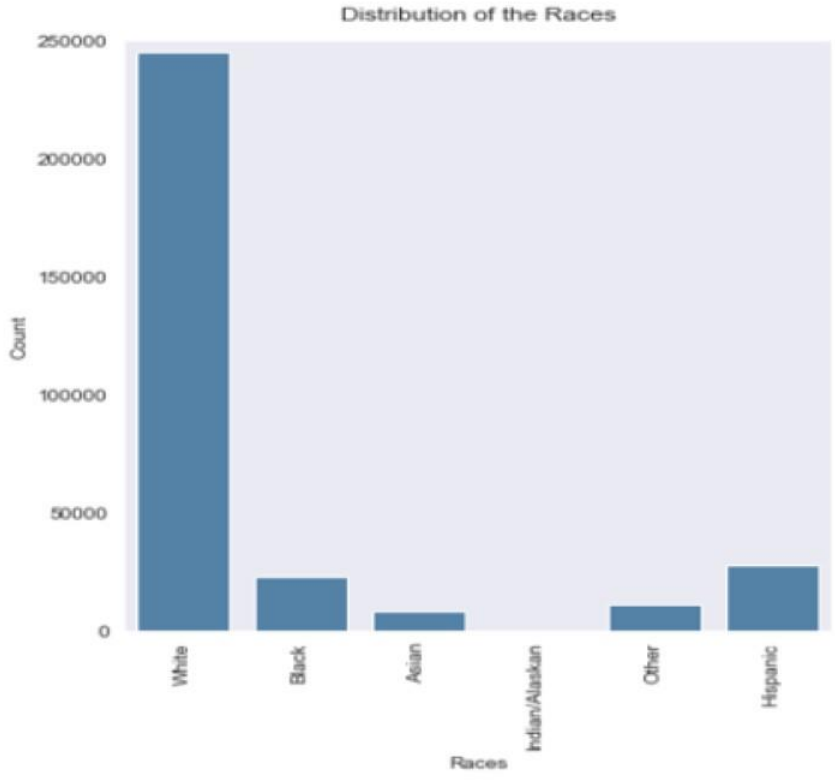
Gender and Heart Disease:

According to CDC, This map shows death rates from heart disease in women in the United States. The darker red indicates a higher death rate. (https://www.cdc.gov/heartdisease/women.htm)



Race and Heart Disease:

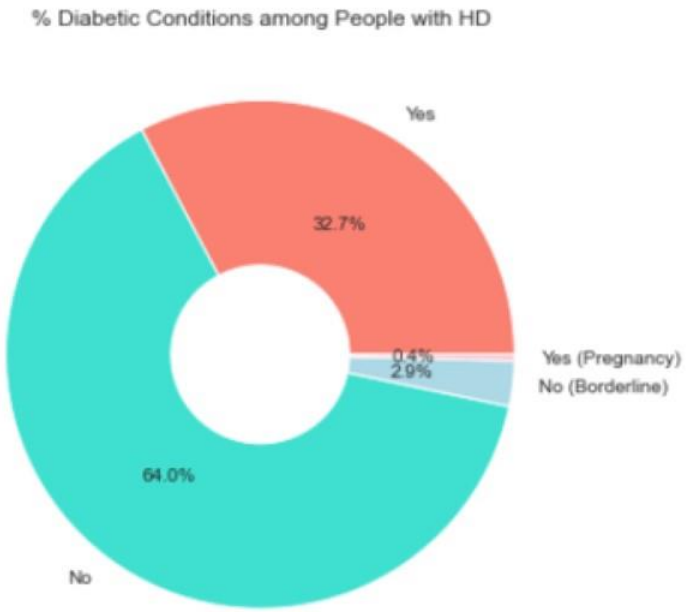
As the Cleveland Clinic website, black people are more vulnerable to heart disease than white people according to some diagnoses, due to the health disparities of race and ethnicity. (Heart Disease Risk: How Race and Ethnicity Play a Role (clevelandclinic.org))



Diabetes and Heart Disease

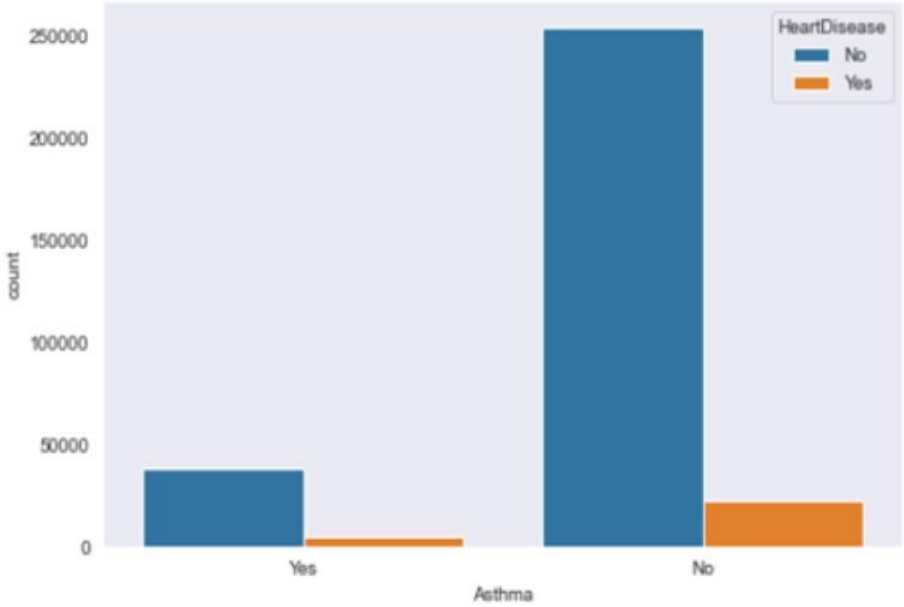
According to CDC, The risk of death from heart disease for adults with diabetes is higher than for adults who do not have diabetes as Diabetes causes sugar to build up in the blood.

(https://www.cdc.gov/diabetes/library/features/diabetes-and-heart.html#:~:text=Over%20time%2C%20high%20blood%20sugar,and%20can%20damage%20artery%20walls)



Asthma Heart Disease:

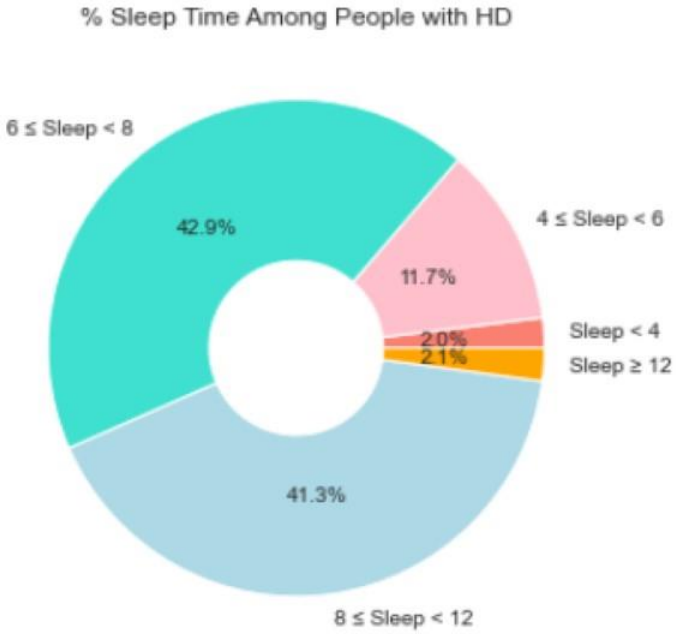
Previous studies have associated asthma with an increased risk of cardiovascular conditions, including Heart attack and cardiovascular disease.



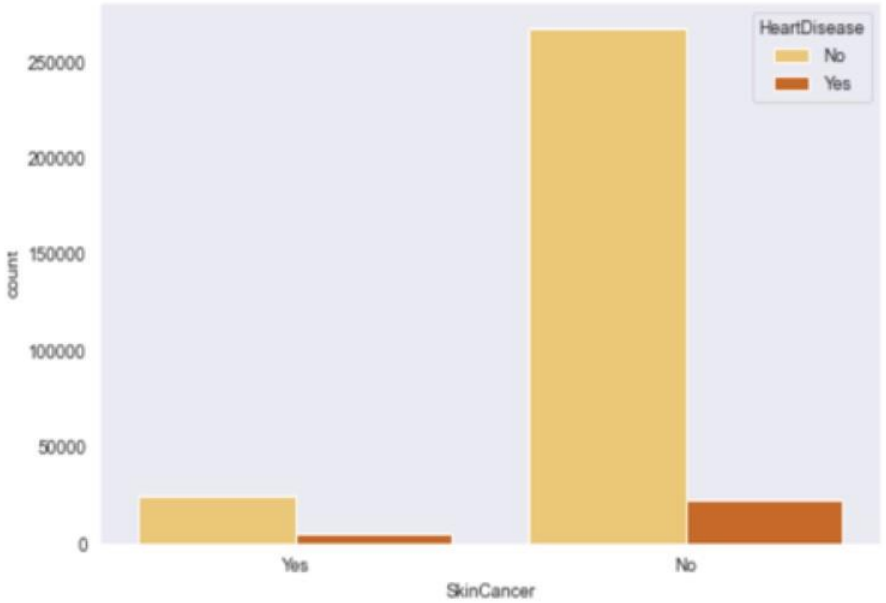
Sleep time and Heart Disease:

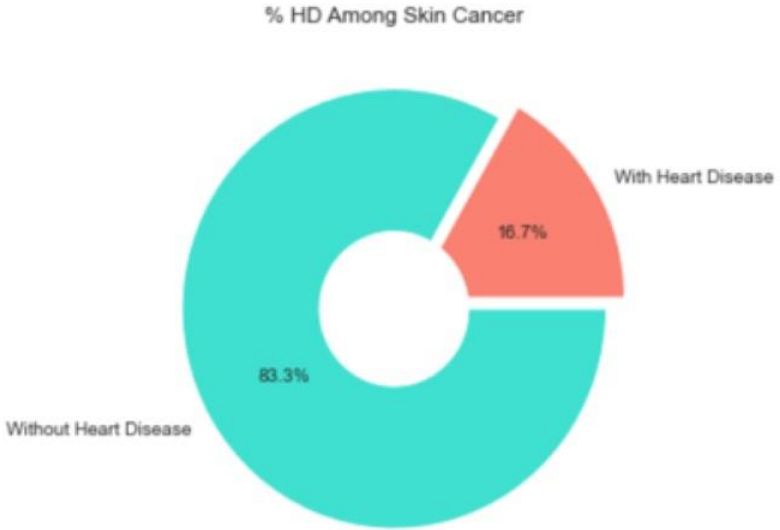
Going to sleep between 10:00 and 11:00 pm is associated with a lower risk of developing heart and circulatory disease compared to earlier or later bedtimes. According to CDC, Adults who sleep less than 7 hours each night are more likely have health problems, including heart attack, asthma, and depression.

(How Does Sleep Affect Your Heart Health? | cdc.gov)(Too Much Sleep May Bring Heart Disease, Death Risk (webmd.com))



Skin Cancer and Heart Disease





It is clear from the graphs above that there is a very weak relationship between skin cancer and heart disease.

3

Data Preprocessing

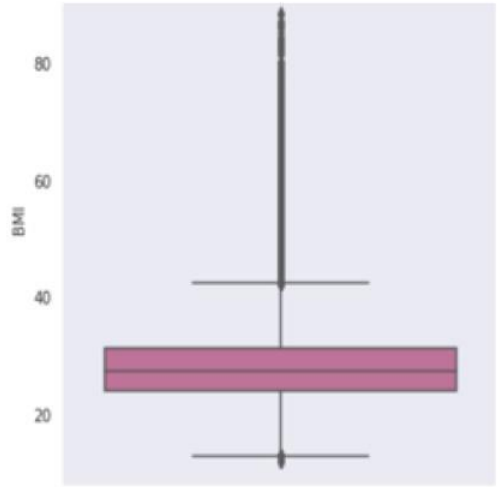
What's should be done for data to fit into the models?

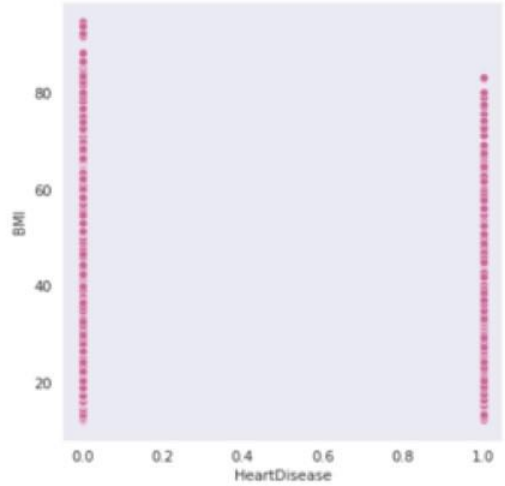
Preprocessing stages:

- Checking null values.
- Encoding categorical data. Remove Unnecessary features.
- Handling outliers

	HeartDisease	BMI	Smoking	AlcoholDrinking	Stroke	PhysicalHealth	MentalHealth
0	No	16.60	Yes	No	No	3.0	30.0
1	No	20.34	No	No	Yes	0.0	0.0

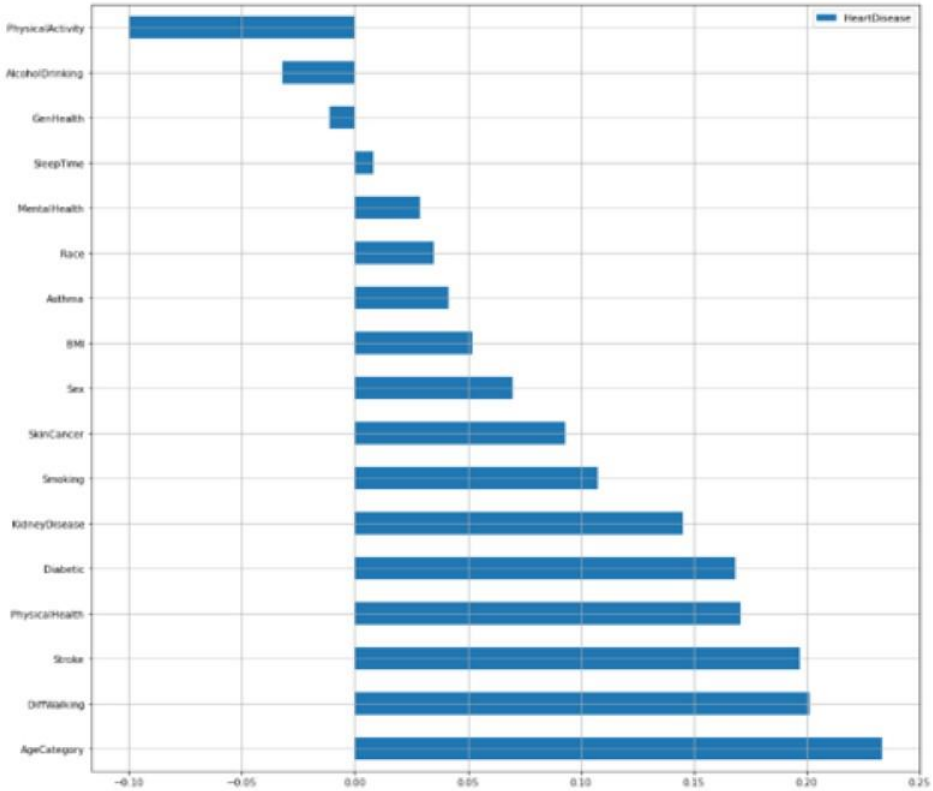
	HeartDisease	BMI	Smoking	Stroke	PhysicalHealth	MentalHealth
0	0	16.60	1	0	3.0	30.0
1	0	20.34	0	1	0.0	0.0





Preprocessing stages:

- We found a lot of weak relationships in our dataset like: SkinCancer & AgeCategory with a ratio of 0.26 and the stronge relations like one between DiffWalking & PhysicalHealth with a ratio of 0.43 and between KindneyDisease & HeartDisease with a ratio of 0.15.
- Imbalanced data Handling
  - Near Miss Sampling
  - Random UnderSampling
  - Random OverSampling
  - Smote Sampling



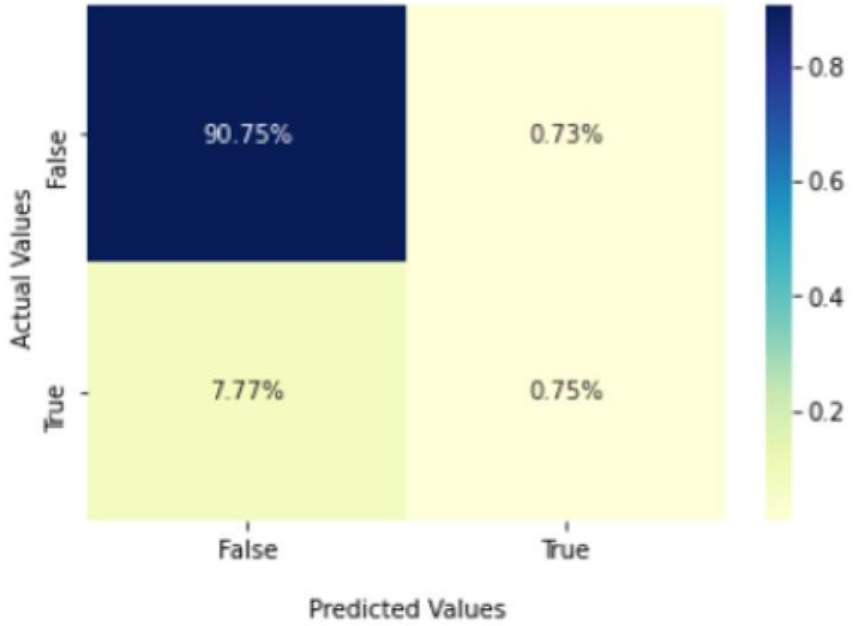
Modeling and Prediction

Logistic Regression Classifier:

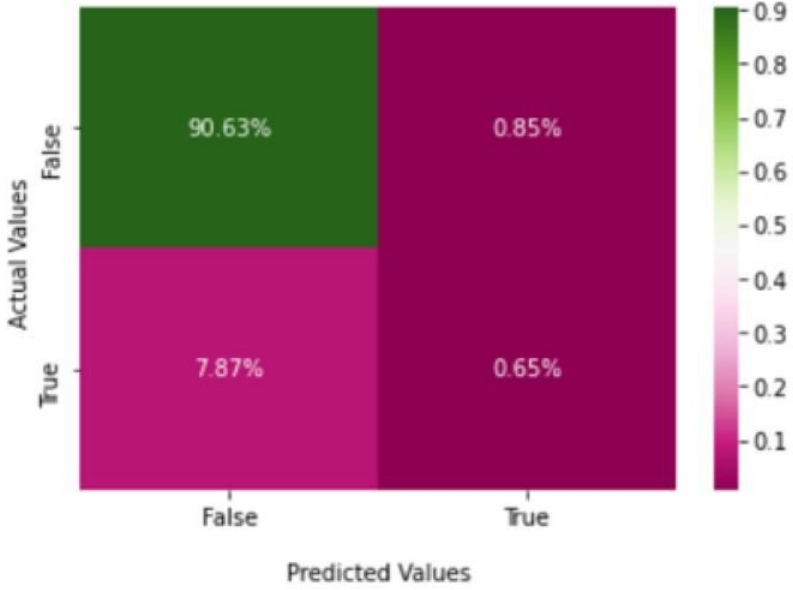
Without using Sampling:

- Accuracy: 91.5%
- Precision: 92%
- Recall: 99%

This model guess the TN more than TP



KNieghbours Classifier:



This model also doesn't need Sampling as it is not influenced in any way by the size of the class:

- Accuracy: 91.28%
- Precision: 92%
- Recall: 99%

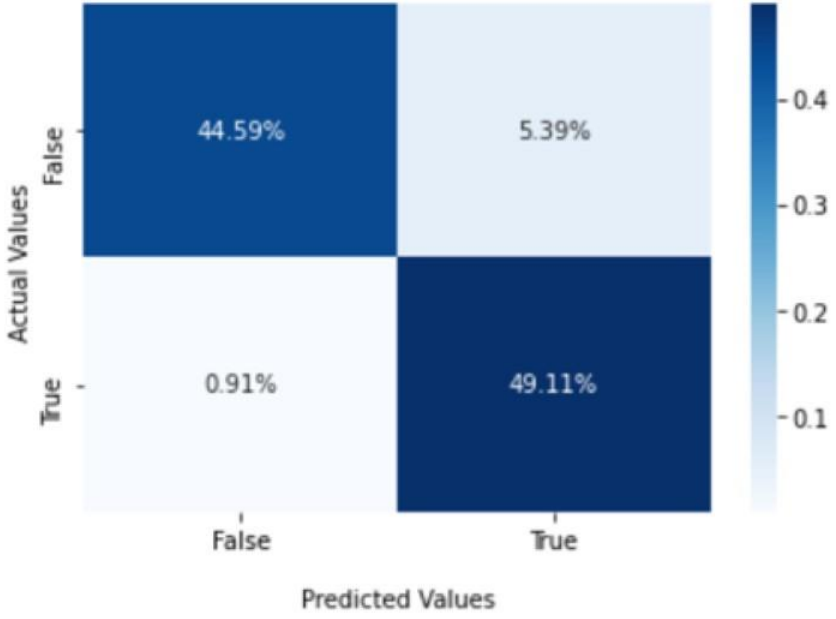
It is a good indicator for TN not for TP

Random Forest Classifier:

Using Random OverSampling:

- Accuracy: 93.72%
- Precision: 98%
- Recall: 89%

This model Guesses the TP more than TN



XGBoost Classifier:

This model deosn't need sampling cause it is known with its great ability to handle overfitting and imbalanced data and This model Guesses the TN way more than TP

- Accuracy: 91.97%
- Precision: 92%
- Recall: 99%

The Best Model for our data is:

Random Forest Classifier:

It produces the best accuracy, which is about 94% under the condition of oversampling, besides that, it predicts the true positives and true negatives in nearly equal ranges unlike other models that only guess the true negatives with high ranges.\

The best-used sampling way is Random over sampling method

Conclusion:

Decrease Heart Disease through:

1. Healthy lifestyle and Diet.

2. More physical activities and have regular excersize.

3. Stop smoking and Alcohol drinking.

4. Keep your blood pressure under control.

SAMSUNG

THANKS!

Together for Tomorrow!

Enabling People

Education for Future Generations