Multiple Disease Classifier

Team Members:

Name	Repositories
Waqas Kureshy	Project Repo
Neeharika Yeluri	neeharikayeluri
Jasmine Wang	jwang11398

Dataset:

A. **Heart Failure Prediction: Dataset**
B. **Diabetes Dataset: Dataset**
C. **Breast Cancer: Dataset**

Problem description:

Cardiovascular diseases, diabetes, and breast cancer are top 3 causes of death globally, taking millions of lives each year, which accounts for 50% of all deaths worldwide. We will be making a web application through which the user can add in their personal details and our trained models would be able to predict if the patient has one of the diseases mentioned below.

Heart Failure

Heart Failure Prediction contains 11 features that can be used to predict a possible heart disease. People with cardiovascular disease or who are at high cardiovascular risk need early detection and management where a machine learning model can be of great help.

Diabetes

The dataset is originally from the National Institute of Diabetes and Digestive and Kidney Diseases. The objective is to predict based on diagnostic measurements whether a patient has diabetes. It has the following features:

Number of Instances: 768
Number of Attributes: 8 plus class
For Each Attribute: (all numeric-valued)

Breast Cancer

Health is an essential aspect of everyone's life. Breast cancer is found in the body of male or females when the cells in the breast begin to grow out of control. These cells usually form a tumor and can be felt as a lump or could be seen on an x-ray. Cancer can be distinguished as benign, or either can be malignant (cancer). Dataset contains 31 features.

Potential Methods

It can be a supervised logistic regression to predict what kind of disease when the user input some symptoms.

OR it can be an unsupervised learning clustering to find hidden correlation among some features.

This notebook contains the initial analysis of the Heart Health data

In [5]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
```

In [2]:
```
data_info = pd.read_csv('heart.csv')
```

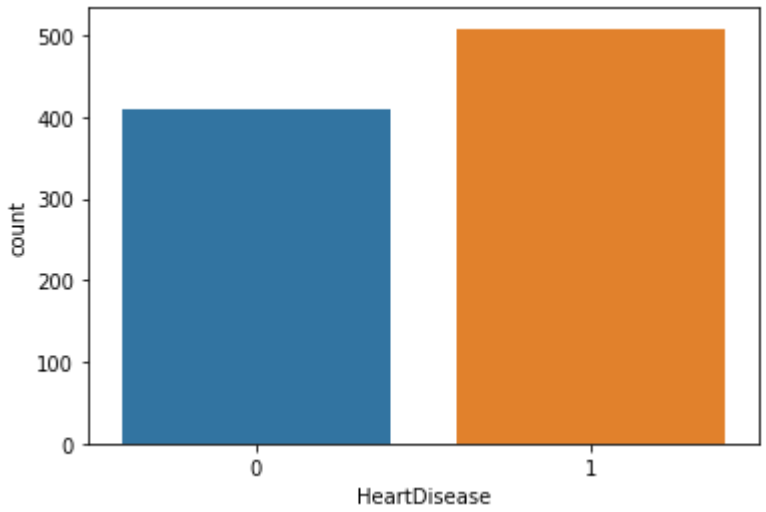
In [4]:
```
data_info.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 918 entries, 0 to 917
Data columns (total 12 columns):
 #   Column          Non-Null Count  Dtype  
---  ------          --------------  -----  
 0   Age             918 non-null    int64  
 1   Sex             918 non-null    object 
 2   ChestPainType   918 non-null    object 
 3   RestingBP       918 non-null    int64  
 4   Cholesterol     918 non-null    int64  
 5   FastingBS       918 non-null    int64  
 6   RestingECG      918 non-null    object 
 7   MaxHR           918 non-null    int64  
 8   ExerciseAngina  918 non-null    object 
 9   Oldpeak         918 non-null    float64
 10  ST_Slope        918 non-null    object 
 11  HeartDisease    918 non-null    int64  
dtypes: float64(1), int64(6), object(5)
memory usage: 86.2+ KB
```

In [6]:
```
sns.countplot('HeartDisease',data=data_info)
```

```
/usr/local/lib/python3.7/dist-packages/seaborn/_decorators.py:43: FutureWarnin
g: Pass the following variable as a keyword arg: x. From version 0.12, the only
valid positional argument will be `data`, and passing other arguments without a
n explicit keyword will result in an error or misinterpretation.
  FutureWarning
```

Out[6]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x7f01bf0b22d0>
```

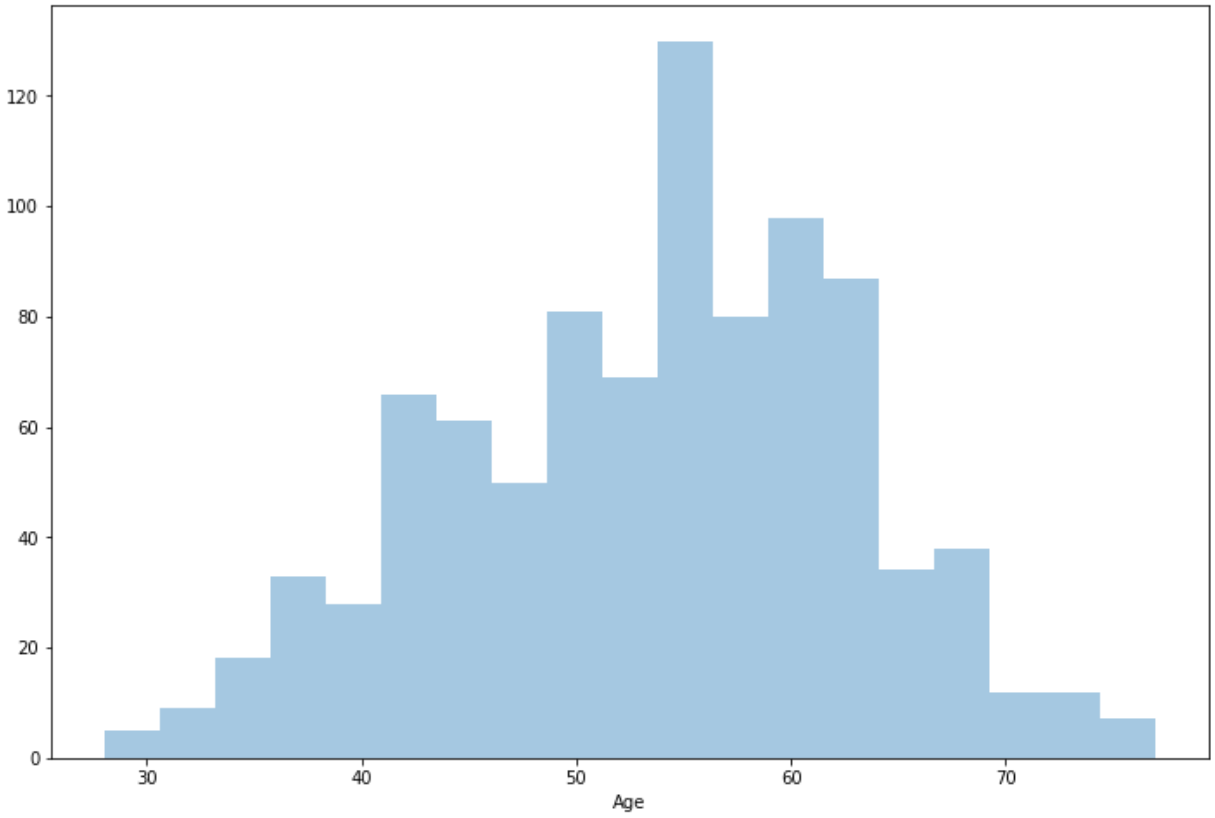


A view of the distribution of data according to Age

In [7]:
```
plt.figure(figsize=(12,8))
sns.distplot(data_info['Age'],kde=False)
```

```
/usr/local/lib/python3.7/dist-packages/seaborn/distributions.py:2619: FutureWar
ning: `distplot` is a deprecated function and will be removed in a future versi
on. Please adapt your code to use either `displot` (a figure-level function wit
h similar flexibility) or `histplot` (an axes-level function for histograms).
  warnings.warn(msg, FutureWarning)
```

Out[7]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x7f01be828c10>
```



Correlation of data

In [8]:
```
data_info.corr()
```

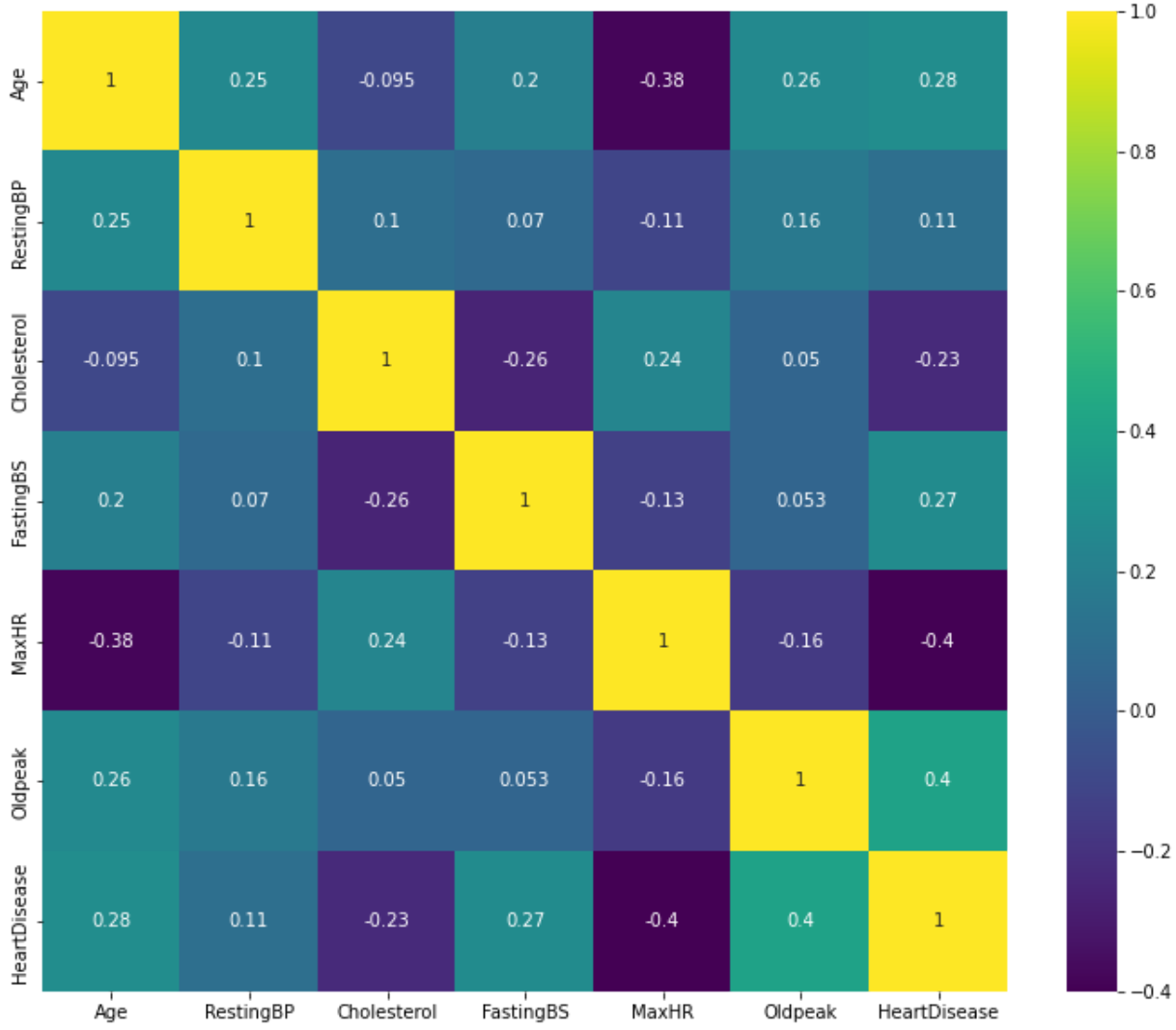
Out[8]:

	Age	RestingBP	Cholesterol	FastingBS	MaxHR	Oldpeak	HeartDisease
Age	1.000000	0.254399	-0.095282	0.198039	-0.382045	0.258612	0.282039
RestingBP	0.254399	1.000000	0.100893	0.070193	-0.112135	0.164803	0.107589
Cholesterol	-0.095282	0.100893	1.000000	-0.260974	0.235792	0.050148	-0.232741
FastingBS	0.198039	0.070193	-0.260974	1.000000	-0.131438	0.052698	0.267291
MaxHR	-0.382045	-0.112135	0.235792	-0.131438	1.000000	-0.160691	-0.400421
Oldpeak	0.258612	0.164803	0.050148	0.052698	-0.160691	1.000000	0.403951
HeartDisease	0.282039	0.107589	-0.232741	0.267291	-0.400421	0.403951	1.000000

Initial analysis suggests that Heart Disease is somewhat correlated with Age

In [9]:
```
plt.figure(figsize=(12,10))
sns.heatmap(data_info.corr(),annot=True,cmap='viridis')
```

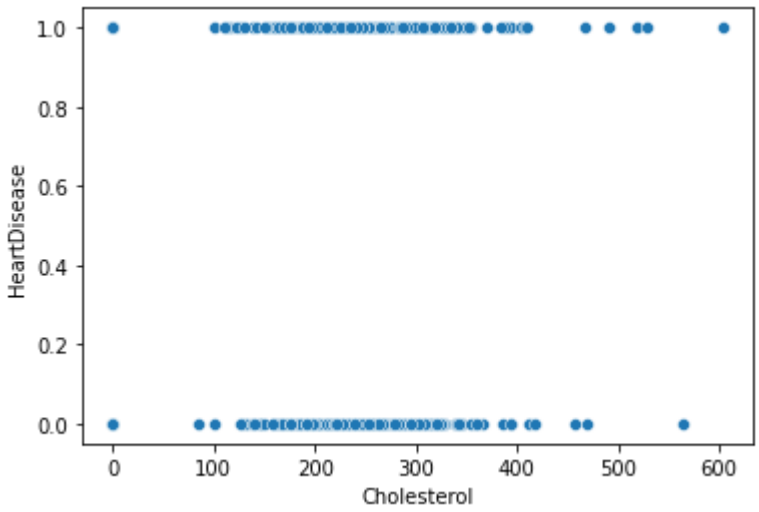
Out[9]: <matplotlib.axes._subplots.AxesSubplot at 0x7f01be830c10>



In [17]:
```
sns.scatterplot('Cholesterol','HeartDisease',data=data_info)
```

```
/usr/local/lib/python3.7/dist-packages/seaborn/_decorators.py:43: FutureWarnin
g: Pass the following variables as keyword args: x, y. From version 0.12, the o
nly valid positional argument will be `data`, and passing other arguments witho
ut an explicit keyword will result in an error or misinterpretation.
  FutureWarning
```

Out[17]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x7f01bb66e0d0>
```



Plot of Age with respect to Heart disease

In [20]:
```
plt.figure(figsize=(12,10))
sns.countplot(data=data_info,x='Age',hue='HeartDisease')
```

Out[20]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x7f01b928a390>
```