HEART DISEASE DATASET

This section contains the first dataset and a detailed description of the dataset with some preparation steps

Dataset collection

The used dataset was published in the repository of UCI and collected at (the University of California, Irvine C.A) Center for Machine Learning and Intelligent Systems.

Dataset Description

The dataset is clinical contains historical medical records, habits, and demographic information for 920 cases with 14 features for each case, the four datasets were combined into a single dataset (i.e., combined dataset) for better model performance. The combined dataset consists of 920 instances with 14 attributes. The characteristics of four heart disease datasets and the combined dataset are shown in Table 1, and Table 2 shows the 14 attributes/features as they exist in the dataset alongside the description of each attribute.

Table 1: combined dataset

Id	Datasets	No of Instances	No of Attribute	Missing Value
1	Cleveland	303	14	No
2	Hungarian	294	14	Yes
3	V.A.	200	14	Yes
4	Switzerland	123	14	Yes
5	Combined Dataset	920	14	Missing Values replaced with Mode Value

Table 2: Description of attributes in the dataset

Id	Attribute	Description
1	Age	Age (continues data) 29 to 77
2	Sex	Sex: 1: male, 0: female
3	CP	Chest pain type: 1: Typical Angina 2: atypical angina 3: non-anginal pain 4: asymptomatic
4	Trestbps	Diastolic blood pressure (mm Hg)
5	Chol	Cholesterol in mg/dl
6	FBS	Fasting blood sugar>120 mg/dl, 1: true, 0: false
7	Restecg	Resting ECG: 0: Normal, 1: ST-T abnormal, 2: Left V. Hypertrophy
8	Thali	Maxi's mum's heart rate achieved
9	Exang	Exercise-induced angina (1=yes, 0=no)
10	old peak	ST depression induced by exercise relative to rest
11	Slope	The slope of the peak exercise ST Segment: 1: upsloping, 2: flat, 3: downsloping
12	Ca	Number of major vessels colored by fluoroscopy (0-3)
13	Thal	Defect type: 3: Normal, 6: fixed defect, 7: reversible defect
14	Target	Heart disease (0-4): 0=Healthy,1= low,2= medium, 3=high, 4=serious.

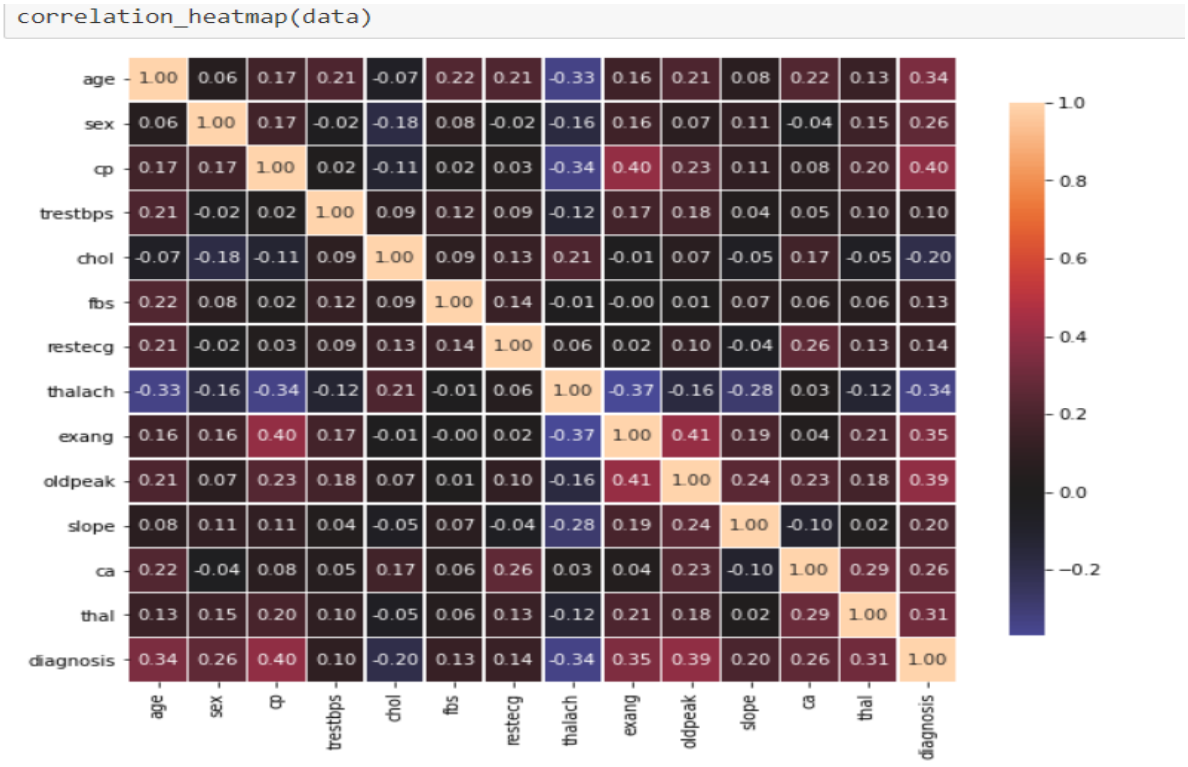


Figure 1 Correlation of input and target output.

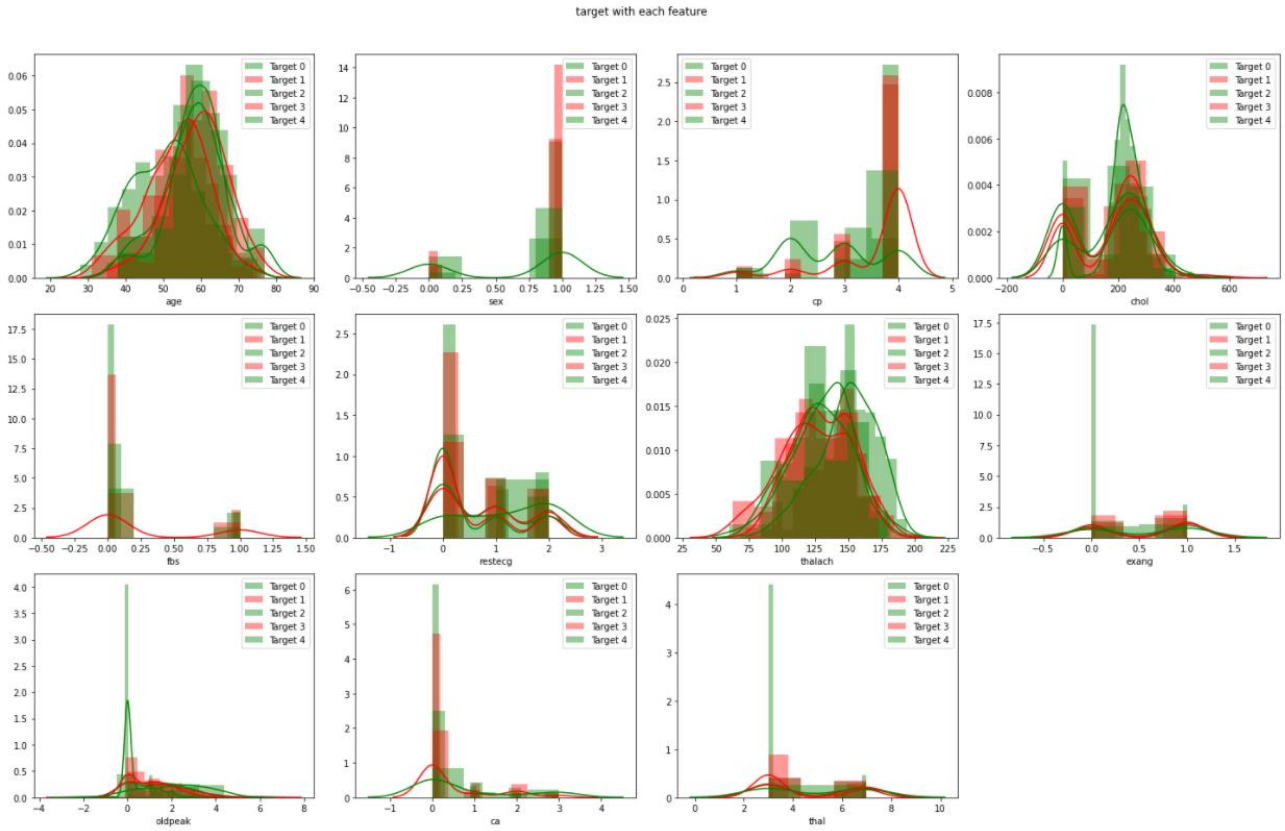


figure 2 Frequency Distribution of Features for All Participants

Dataset preparation

Table 2 also shows a lot of missing values can be found in features 4 and 11, for the leakage of information features, features 4 and 11 are removed and the number of features then decreased to 12 features. To handle the missing values the mean equation is used as shown in (Equation 4-1).

$$\bar{x} = \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right) = \frac{x_1 + x_2 + \cdots + x_n}{n} \qquad \textit{(Equation 4-1)}$$

Exploratory Data Analysis

The first step was to ascertain the distribution of totally different attributes and this was best envisioned by histograms

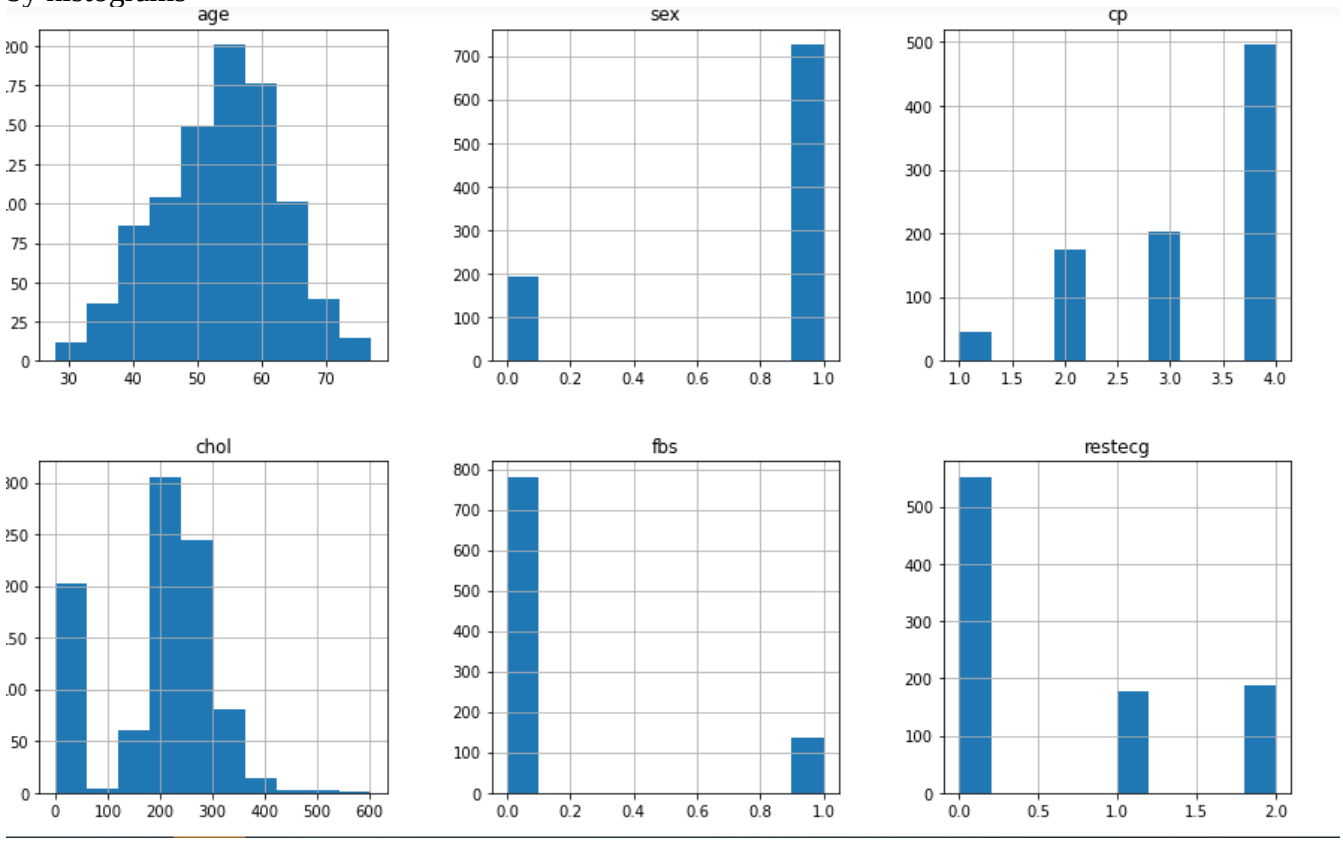


figure 3 data description

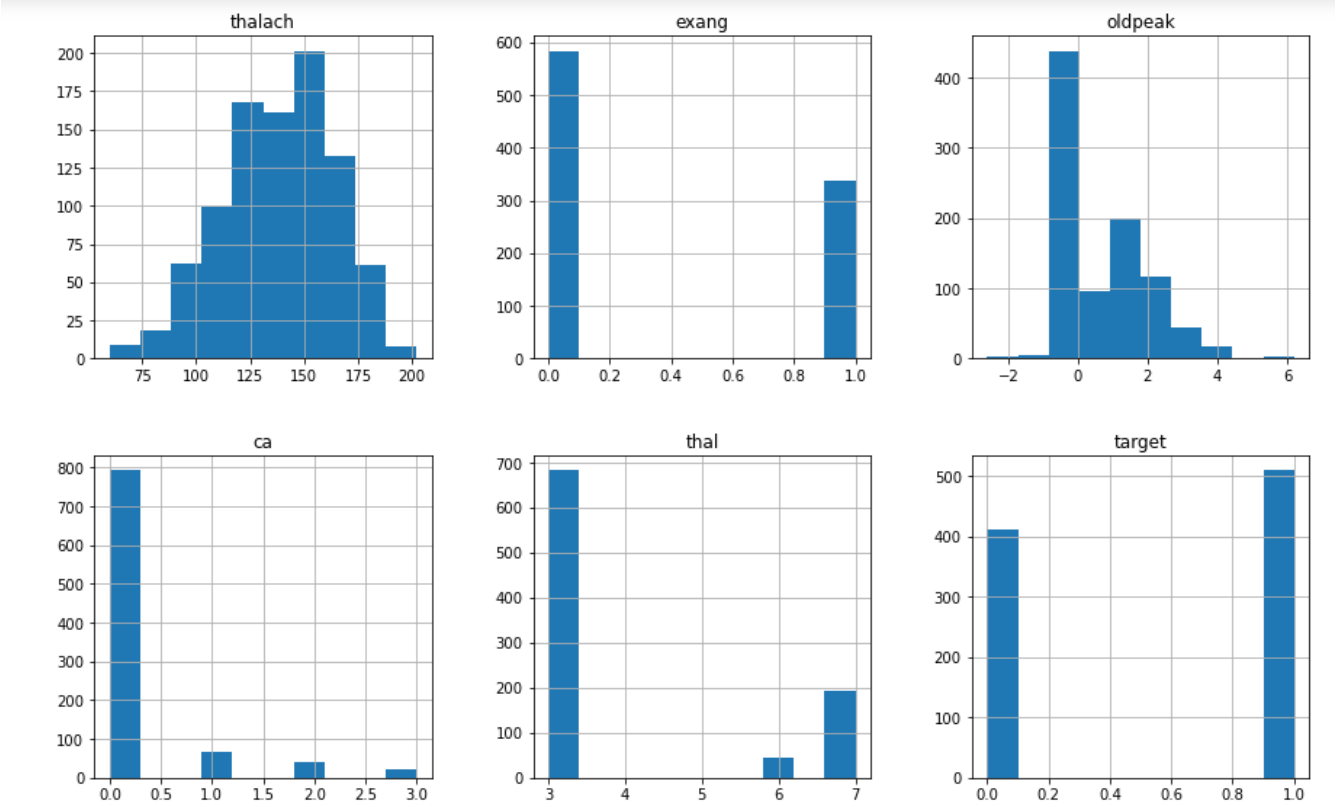
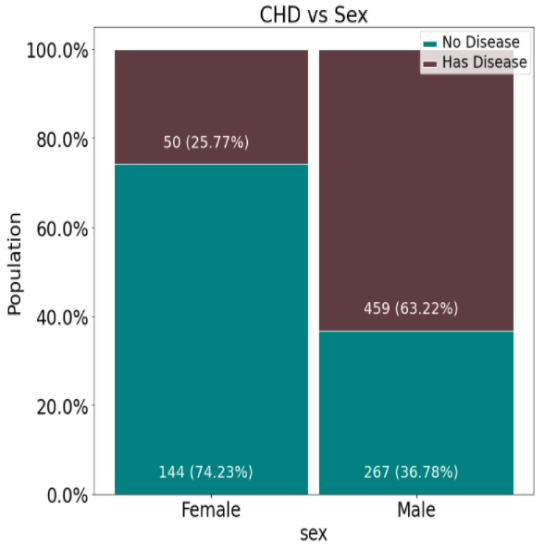


figure 4. 1 data description

To gain a lot of insight into the info I checked the proportions of positive and negative cases in every class.



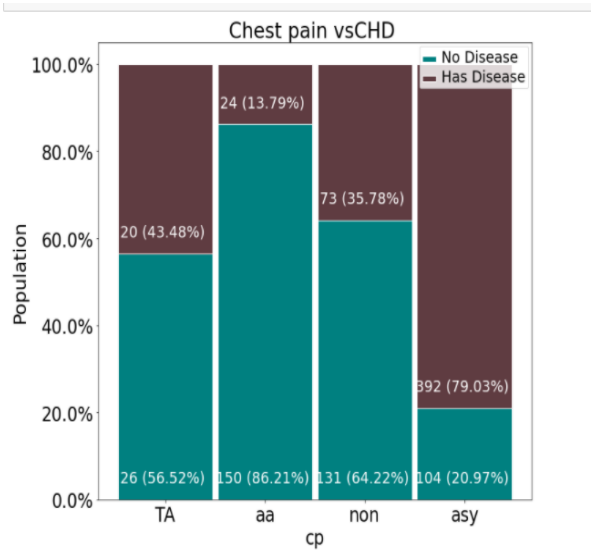
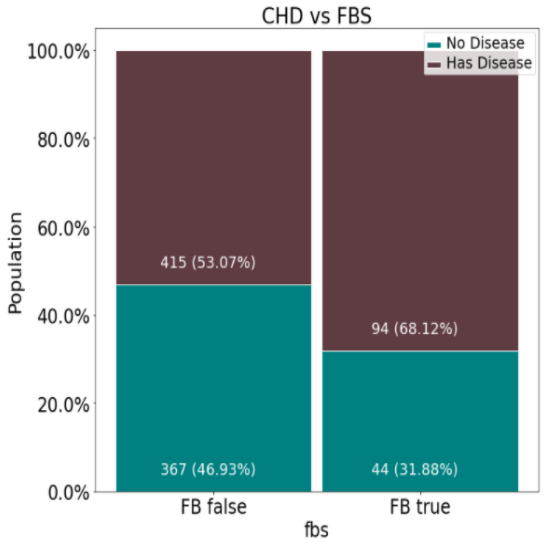


figure 4. 2 Chest pain VS CHD

figure 4. 3 CHD VS SEX



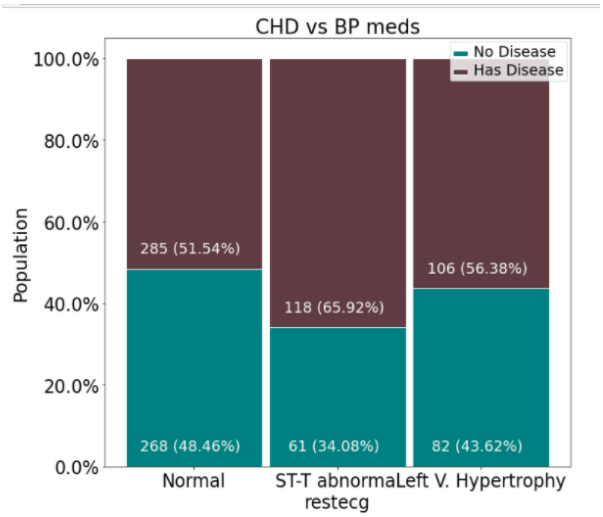
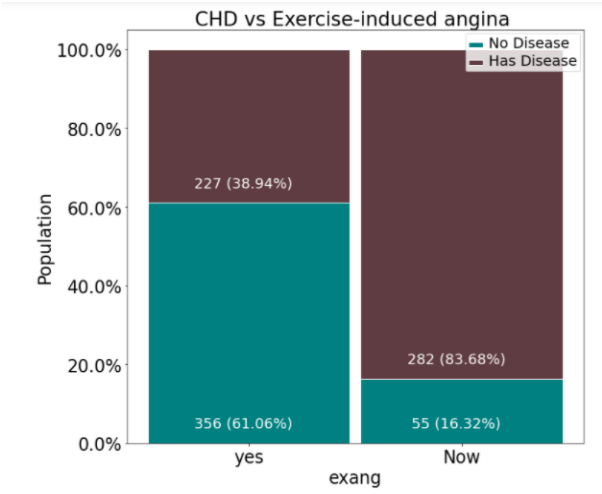


figure 4. 4 CHD VS BP MEDS

figure 4. 5 CHD VS FBS



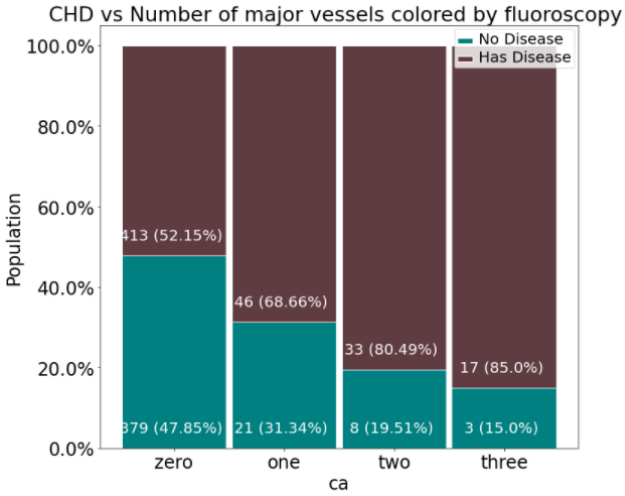


figure 4. 6 CHD VS number of major vessels colored by fluoroscopy

figure 4. 7 CHD VS EXERCISE- induced angina

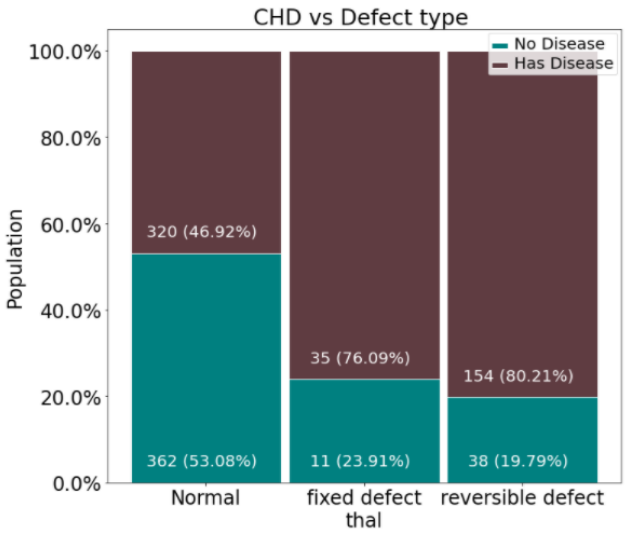


figure 4. 8 CHD VS Defect Type

Due to the unbalanced nature of the info set it had been troublesome to create conclusions that supported what's determined however these area units the conclusions that would be drawn:

Risk factors

- Slightly more males are suffering from CHD than females.
- The percentage of people who have CHD is increasing when having Chest pain type asymptomatic
- A larger percentage of the people who have CHD have Fasting blood sugar
- The percentage of people who have CHD is almost equal between Resting ECG
- A larger percentage of the people who have CHD has not Exercise-induced angina
- And the people who have a number of major vessels colored by fluoroscopy 2 or 3 has A larger percentage of CHD
- And the people who have Defect type fixed defect, or reversible defect have A larger percentage of CHD

Another interesting trend I checked for was the distribution of the ages of the people who had CHD and the number of the sick generally increased with age with the peak being at 59 years old.

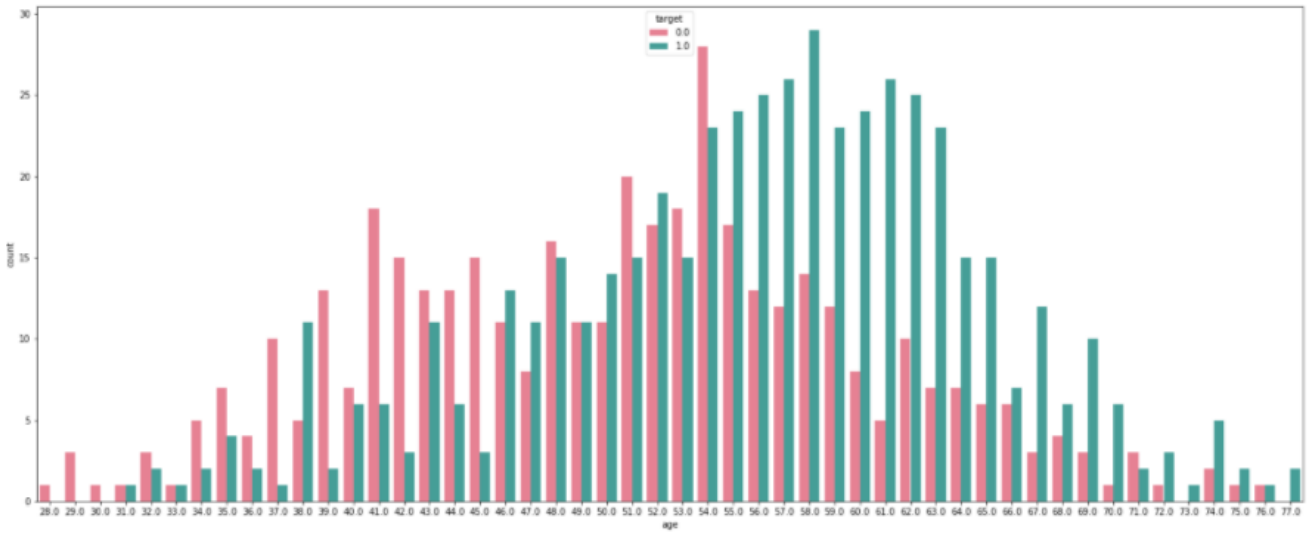


figure 4. 9 CHD VS AGE