# Experiment 1: Feature Selection by using Heart Dataset

In [2]:

```
import pandas as pd
df = pd.read_csv('heart.csv')
df.head()
```

Out[2]:

| | age | sex | cp | trestbps | chol | fbs | restecg | thalach | exang | oldpeak | slope |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 63 | 1 | 3 | 145 | 233 | 1 | 0 | 150 | 0 | 2.3 | 0 |
| **1** | 37 | 1 | 2 | 130 | 250 | 0 | 1 | 187 | 0 | 3.5 | 0 |
| **2** | 41 | 0 | 1 | 130 | 204 | 0 | 0 | 172 | 0 | 1.4 | 2 |
| **3** | 56 | 1 | 1 | 120 | 236 | 0 | 1 | 178 | 0 | 0.8 | 2 |
| **4** | 57 | 0 | 0 | 120 | 354 | 0 | 1 | 163 | 1 | 0.6 | 2 |

In [3]:

```
### Univariate selection
x=df.iloc[:,:-1]
y=df['target']
```

In [4]:

```
x.head()
```

Out[4]:

| | age | sex | cp | trestbps | chol | fbs | restecg | thalach | exang | oldpeak | slope |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 63 | 1 | 3 | 145 | 233 | 1 | 0 | 150 | 0 | 2.3 | 0 |
| **1** | 37 | 1 | 2 | 130 | 250 | 0 | 1 | 187 | 0 | 3.5 | 0 |
| **2** | 41 | 0 | 1 | 130 | 204 | 0 | 0 | 172 | 0 | 1.4 | 2 |
| **3** | 56 | 1 | 1 | 120 | 236 | 0 | 1 | 178 | 0 | 0.8 | 2 |
| **4** | 57 | 0 | 0 | 120 | 354 | 0 | 1 | 163 | 1 | 0.6 | 2 |

In [5]:

```
y.head()
```

Out[5]:

```
0    1
1    1
2    1
3    1
4    1
Name: target, dtype: int64
```

In [6]:

```
from sklearn.feature_selection import SelectKBest
from sklearn.feature_selection import chi2
```

SelectKBest selects the top features based on their scores using a statistical test, such as chi squared test or ANOVA F-test. The score measures the dependency between each feature and the target variable,and the K features with the highest scores with the highest scores are selected.

In [7]:

```
## Apply SelectKBest Algorithm
ordered_rank_features = SelectKBest(score_func=chi2,k=8)
ordered_rank_features
```

Out[7]:

```
▾                    SelectKBest
SelectKBest(k=8, score_func=<function chi2 at 0x00000285A3428A40>)
```

In [8]:

```
ordered_feature = ordered_rank_features.fit(x,y)
ordered_feature
```

Out[8]:

```
▾                    SelectKBest
SelectKBest(k=8, score_func=<function chi2 at 0x00000285A3428A40>)
```

In [9]:

```
dfscore=pd.DataFrame(ordered_feature.scores_,columns=['Score'])
dfcolumns=pd.DataFrame(x.columns)
#dfcolumns
```

In [10]:

```
features_rank = pd.concat([dfcolumns,dfscore],axis=1)
```

In [11]:

```
features_rank.columns=['Features','Score']
features_rank
```

Out[11]:

| | Features | Score |
|---|---|---|
| **0** | age | 23.286624 |
| **1** | sex | 7.576835 |
| **2** | cp | 62.598098 |
| **3** | trestbps | 14.823925 |
| **4** | chol | 23.936394 |
| **5** | fbs | 0.202934 |
| **6** | restecg | 2.978271 |
| **7** | thalach | 188.320472 |
| **8** | exang | 38.914377 |
| **9** | oldpeak | 72.644253 |
| **10** | slope | 9.804095 |
| **11** | ca | 66.440765 |
| **12** | thal | 5.791853 |

In [12]:

```
features_rank.nlargest(10,'Score')
```

Out[12]:

| | Features | Score |
|---|---|---|
| **7** | thalach | 188.320472 |
| **9** | oldpeak | 72.644253 |
| **11** | ca | 66.440765 |
| **2** | cp | 62.598098 |
| **8** | exang | 38.914377 |
| **4** | chol | 23.936394 |
| **0** | age | 23.286624 |
| **3** | trestbps | 14.823925 |
| **10** | slope | 9.804095 |
| **1** | sex | 7.576835 |

# Correlation

In [13]:

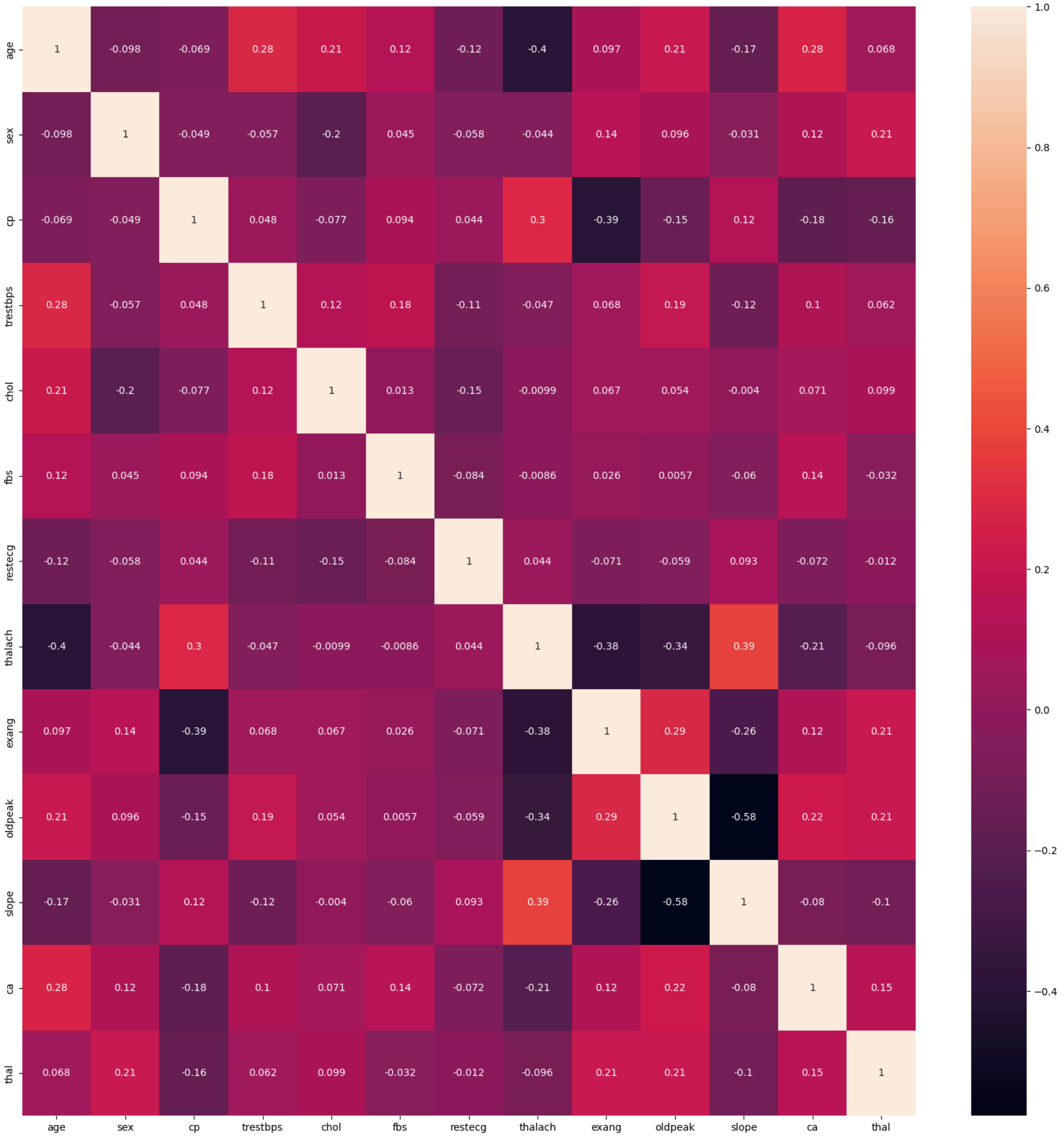
```
import matplotlib .pyplot as plt
import seaborn as sns
corr = df.iloc[:,:-1].corr()
```

In [14]:

```
top_features = corr.index
plt.figure(figsize=(20,20))
sns.heatmap(df[top_features].corr(),annot=True)
```

Out[14]:

```
<Axes: >
```


age
1
-0.098
-0.069
0.28
0.21
0.12
-0.12
-0.4
0.097
0.21
-0.17
0.28
0.068
sex
-0.098
1
-0.049
-0.057
-0.2
0.045
-0.058
-0.044
0.14
0.096
-0.031
0.12
0.21
cp
-0.069
-0.049
1
0.048
-0.077
0.094
0.044
0.3
-0.39
-0.15
0.12
-0.18
-0.16
trestbps
0.28
-0.057
0.048
1
0.12
0.18
-0.11
-0.047
0.068
0.19
-0.12
0.1
0.062
chol
0.21
-0.2
-0.077
0.12
1
0.013
-0.15
-0.0099
0.067
0.054
-0.004
0.071
0.099
fbs
0.12
0.045
0.094
0.18
0.013
1
-0.084
-0.0086
0.026
0.0057
-0.06
0.14
-0.032
restecg
-0.12
-0.058
0.044
-0.11
-0.15
-0.084
1
0.044
-0.071
-0.059
0.093
-0.072
-0.012
thalach
-0.4
-0.044
0.3
-0.047
-0.0099
-0.0086
0.044
1
-0.38
-0.34
0.39
-0.21
-0.096
exang
0.097
0.14
-0.39
0.068
0.067
0.026
-0.071
-0.38
1
0.29
-0.26
0.12
0.21
oldpeak
0.21
0.096
-0.15
0.19
0.054
0.0057
-0.059
-0.34
0.29
1
-0.58
0.22
0.21
slope
-0.17
-0.031
0.12
-0.12
-0.004
-0.06
0.093
0.39
-0.26
-0.58
1
-0.08
-0.1
ca
0.28
0.12
-0.18
0.1
0.071
0.14
-0.072
-0.21
0.12
0.22
-0.08
1
0.15
thal
0.068
0.21
-0.16
0.062
0.099
-0.032
-0.012
-0.096
0.21
0.21
-0.1
0.15
1
age
sex
cp
trestbps
chol
fbs
restecg
thalach
exang
oldpeak
slope
ca
thal
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4