20200389	كريم احمد
20200031	احمد علي علي محمود
20200278	عبدالرحمن احمد سمير
20200332	علي ماهر
20200005	ابراهيم هشام احمد
20200347	عمر صلاح الدين احمد

Dataset:

We used Heart Attack analysis & Prediction Dataset (Heart Attack Analysis & Prediction Dataset (kaggle.com).

It's a dataset for heart attack classification, it contains data about Age , Sex , Exercise induced angina (Yes , No) , Number of major vessels (0-3) , Chest Pain Type (4 options) , Resting blood pressure , Cholestral in mg/dl , Fasting blood sugar (True , False) , Resting electrocardiographic results (3 options) , Maximum heart rate achieved and Target (Yes , No).

Methodology:

1- Data Preprocessing:

First we started by checking the features' data type:

```
(303, 14)
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 303 entries, 0 to 302
Data columns (total 14 columns):
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   age       303 non-null    int64
 1   sex       303 non-null    int64
 2   cp        303 non-null    int64
 3   trtbps    303 non-null    int64
 4   chol      303 non-null    int64
 5   fbs       303 non-null    int64
 6   restecg   303 non-null    int64
 7   thalachh  303 non-null    int64
 8   exng      303 non-null    int64
 9   oldpeak   303 non-null    float64
 10  slp       303 non-null    int64
 11  caa       303 non-null    int64
 12  thall     303 non-null    int64
 13  output    303 non-null    int64
dtypes: float64(1), int64(13)
memory usage: 33.3 KB
```

After that we checked if there are any null values in our data:

```
age         0
sex         0
cp          0
trtbps      0
chol        0
fbs         0
restecg     0
thalachh    0
exng        0
oldpeak     0
slp         0
caa         0
thall       0
output      0
dtype: int64
```

We checked if there are any duplicated values in the data to be cleared:

```
1
Cleared
```

We checked the Range (Min , Max) for each column:

```
Range of col age: Max: 77, Min: 29
**************************************************************************************************

Range of col sex: Max: 1, Min: 0
**************************************************************************************************

Range of col cp: Max: 3, Min: 0
**************************************************************************************************

Range of col trtbps: Max: 200, Min: 94
**************************************************************************************************

Range of col chol: Max: 564, Min: 126
**************************************************************************************************

Range of col fbs: Max: 1, Min: 0
**************************************************************************************************

Range of col restecg: Max: 2, Min: 0
**************************************************************************************************

Range of col thalachh: Max: 202, Min: 71
**************************************************************************************************

Range of col exng: Max: 1, Min: 0
**************************************************************************************************

Range of col oldpeak: Max: 6.2, Min: 0.0
**************************************************************************************************

Range of col slp: Max: 2, Min: 0
**************************************************************************************************

Range of col caa: Max: 4, Min: 0
**************************************************************************************************

Range of col thall: Max: 3, Min: 0
**************************************************************************************************

Range of col output: Max: 1, Min: 0
**************************************************************************************************
```

We also checked columns which have continuous values:

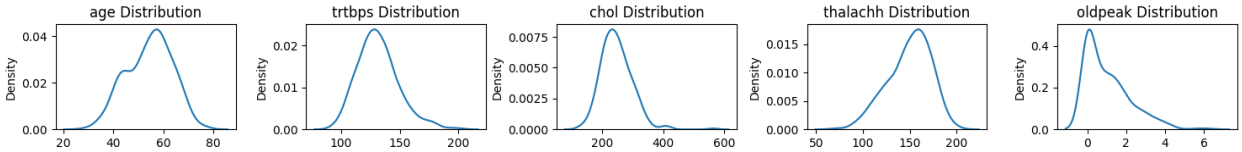
[‘age’ , ‘trtbps’ , ‘chol’ , ‘thalachh’ , ‘oldpeak’]

We checked the outliers of each column:

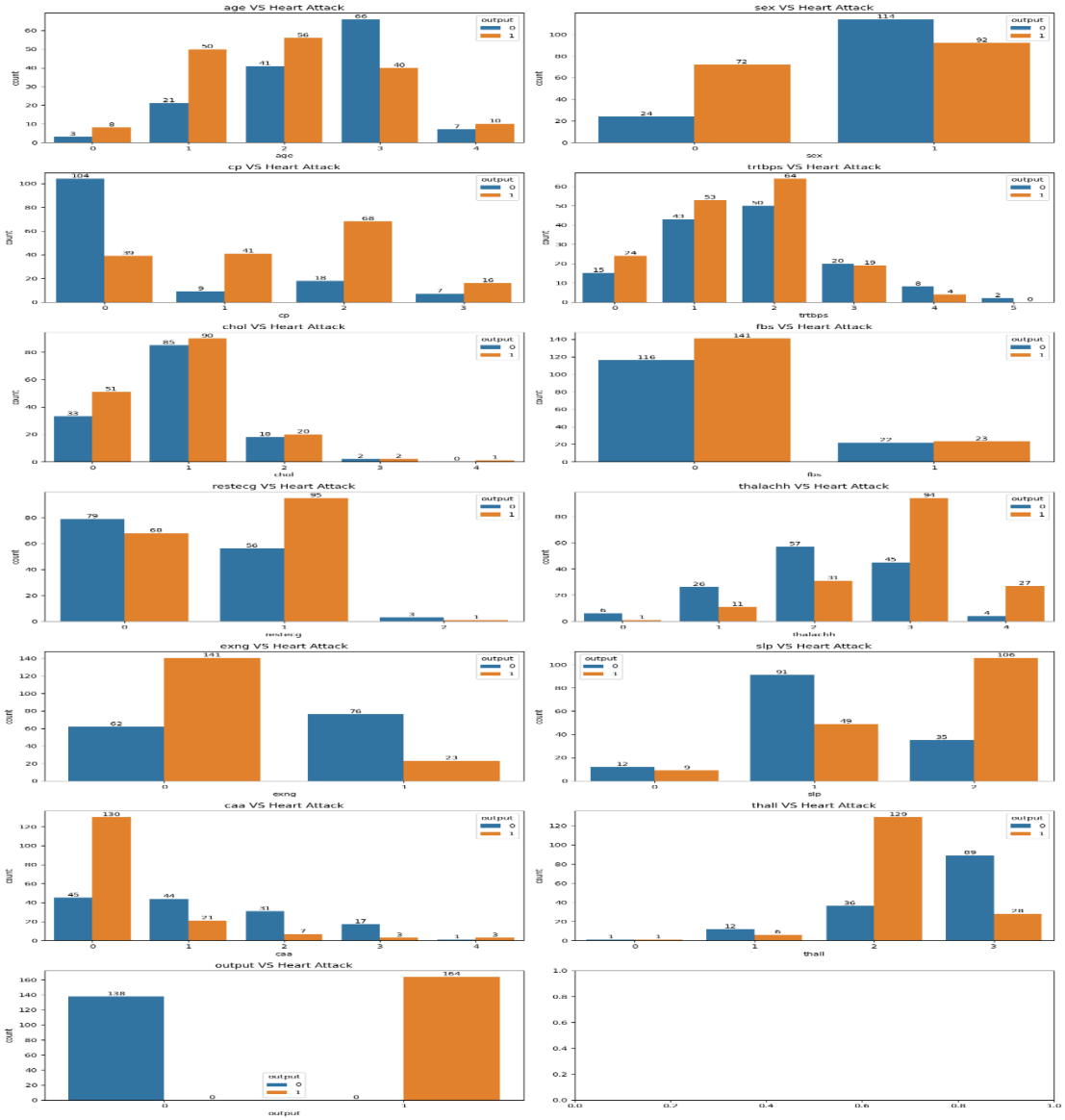
```
Outlier count in age: 0
Variance: 0.9358319948956019
**********************************************************************************************************

Outlier count in sex: 0
Variance: 0.2175529691315923
**********************************************************************************************************

Outlier count in cp: 0
Variance: 1.0651140788981541
**********************************************************************************************************

Outlier count in trtbps: 14
Variance: 1.0647620514400122
**********************************************************************************************************

Outlier count in chol: 5
Variance: 0.4748960418912675
**********************************************************************************************************

Outlier count in fbs: 45
Variance: 0.12722492354403644
**********************************************************************************************************

Outlier count in restecg: 0
Variance: 0.2767045829574707
**********************************************************************************************************

Outlier count in thalachh: 7
Variance: 0.8421816901718332
**********************************************************************************************************

Outlier count in exng: 0
Variance: 0.22108424457107653
**********************************************************************************************************

Outlier count in oldpeak: 5
Variance: 1.3489714197707423
**********************************************************************************************************

Outlier count in slp: 0
Variance: 0.37979362390266447
**********************************************************************************************************

Outlier count in caa: 24
Variance: 1.0135420562803898
**********************************************************************************************************

Outlier count in thall: 2
Variance: 0.3758003124243691
**********************************************************************************************************

Outlier count in output: 0
Variance: 0.24897141977074214
**********************************************************************************************************
```

2- **Data Visualization:**

In this step we wanted to know more about data in an understandable way like visualization for each feature and check it's **distribution (continuous columns)** :



After that we checked the relation between each feature and heart attack and how it affects the output result :



3- **Data Splitting:**

We split the data into 2 sets (training , testing) , with testing size = 0.33

4- **Feature Selection:**

For this process we used the Genetic Algorithm to pick the best set of features to use in training our model.

We used RandomForest Classifier in G.A just for testing our data in each iteration in the algorithm to check the best-fit features.

We made the population size = 10

Number of generations = 15

Mutation rate = 0.01

Here are the results after applying Genetic Algorithm:

```
Generation 1: Fitness: 0.83
Generation 2: Fitness: 0.81
Generation 3: Fitness: 0.83
Generation 4: Fitness: 0.82
Generation 5: Fitness: 0.81
Generation 6: Fitness: 0.81
Generation 7: Fitness: 0.82
Generation 8: Fitness: 0.82
Generation 9: Fitness: 0.82
Generation 10: Fitness: 0.8
Generation 11: Fitness: 0.8
Generation 12: Fitness: 0.84
Generation 13: Fitness: 0.84
Generation 14: Fitness: 0.84
Generation 15: Fitness: 0.81

Best Features: [0, 1, 0, 0, 0, 1, 1, 1, 1, 0, 0, 1, 1]
```

Best set of features selected:

```
['sex', 'fbs', 'restecg', 'thalachh', 'exng', 'caa', 'thall']
```

5- **Feature Reduction:**

For reducing the dimensionality of our features , we used PCA:

We set the number of components = 6

The shape of (X) after applying PCA is : (302 , 6)

6- **Modeling:**

We used KNN to be our Classifier Model:

We set the Neighbors(K) = 5

Then we trained the model using the training set we got from the split we have done earlier.

After that we used the test set for prediction to check the performance of our model.

Classification Report:

	precision	recall	f1-score	support
0	0.72	0.85	0.78	27
1	0.89	0.78	0.83	40
accuracy			0.81	67
macro avg	0.80	0.81	0.80	67
weighted avg	0.82	0.81	0.81	67