DEPI Graduation Project

HealthCare Predictive analysis

Team Members:

- Amina Zeyad Atta Ibrahim
- Amr Ashour Mohamed Ali
- Elham Mokhtar Ahmed Mohammed
- Enas Samir Abbas Mohamed
- Esraa Mohamed Sayed Mohamed

Training Company: CLS

Group Number: CLS__CAI2_AIS4_G4

Instructor: Mahmoud El-Sayed

HealthCare Predictive analysis

Contents

List of Figures

1. Project Overview

The Healthcare Predictive Analytics specially **Cardiovascular Disease (CVD)** project focuses on developing a classification model to classify the patient if he/she suffers from cardiovascular disease or not by providing data-driven insights. The model will be designed to help healthcare professionals with tasks such as patient risk detection, making informed decisions based on predictive analytics. The project will utilize a classification model.

2. Milestone1

2.1 Data Collection:

Our Dataset was collected from Kaggle
Link:https://www.kaggle.com/datasets/scientificstephen/medical-examination-dataset-analysis

Link of Dataset on our drive:
https://drive.google.com/file/d/1uD5d16AkU_fdwTs3Fq6xqlGCN09A7o_7/view?usp=drive_link

2.2 Data Exploration:

Healthcare cardiovascular disease dataset consists of 70000 records of people and 13 factors taken into consideration in the dataset. The below figure shows a sample of the dataset

id	age	gender	height	weight	ap_hi	ap_lo	cholesterol	gluc	smoke	alco	active	cardio
0	18393	2	168	62	110	80	1	1	0	0	1	0
1	20228	1	156	85	140	90	3	1	0	0	1	1
2	18857	1	165	64	130	70	3	1	0	0	0	1
3	17623	2	169	82	150	100	1	1	0	0	1	1
4	17474	1	156	56	100	60	1	1	0	0	0	0
8	21914	1	151	67	120	80	2	2	0	0	0	0
9	22113	1	157	93	130	80	3	1	0	0	1	0
12	22584	2	178	95	130	90	3	3	0	0	1	1
13	17668	1	158	71	110	70	1	1	0	0	1	0
14	19834	1	164	68	110	60	1	1	0	0	0	0
15	22530	1	169	80	120	80	1	1	0	0	1	0
16	18815	2	173	60	120	80	1	1	0	0	1	0
18	14791	2	165	60	120	80	1	1	0	0	0	0
21	19809	1	158	78	110	70	1	1	0	0	1	0

Figure 1 Original Dataset sample

Description of each column (Key Features):

- id: this is just a number used to identify the patient.
- age: contains the age of each patient in days.
- gender: The column identify the sex of each patient (1: Female,2: Male).
- height: contains the height of each patient in meters (m).
- weight: contains the weight of each patient in kilograms (kg).
- ap_hi: Systolic of the patient in mmHg.
- ap_lo: Diastolic of the patient in mmHg.
- cholesterol: categorize the cholesterol level of each patient (1:Low, 2:Medium, 3:High).
- gluc: categorize the glucose level of each patient (1:Low, 2:Medium, 3:High).

- Smoke: Categorize the patient if he/she is a smoker or not.(1: Smoker, 0: Not a smoker)
- alco: Categorize the patient if he/she drinks alcohol or not. (1: drinker, 0: not a drinker)
- active: Categorize if the patient practices any sport or not (1: practice any activity, 0: not practice any activity).
- cardio: this is the target column in which classifies the patient if he/ she suffers from cardiovascular disease or not.

Data Summary:

1- Size: (13*70000) Thousands of individual records.

2- Type: Mixed numeric and categorical data.

3- Challenges: Includes outliers and categorical data requiring cleaning and preprocessing.

2.2.1 Check Data Statistics; First 5 records, Last 5 records, NULLs , Duplicates in Dataset and check if there is any categorical column needs encoding:

First 5 rows of the dataset:

	id	age	gender	height	weight	ap_hi	ap_lo	cholesterol	gluc	smoke	\
0	0	18393	2	168	62.0	110	80	1	1	0	
1	1	20228	1	156	85.0	140	90	3	1	0	
2	2	18857	1	165	64.0	130	70	3	1	0	
3	3	17623	2	169	82.0	150	100	1	1	0	
4	4	17474	1	156	56.0	100	60	1	1	0	

	alco	active	cardio
0	0	1	0
1	0	1	1
2	0	0	1
3	0	1	1
4	0	0	0

Figure 2 First 5 records in the dataset

Last 5 rows of the dataset:

	id	age	gender	height	weight	ap_hi	ap_lo	cholesterol	gluc	\
69995	99993	19240	2	168	76.0	120	80	1	1	
69996	99995	22601	1	158	126.0	140	90	2	2	
69997	99996	19066	2	183	105.0	180	90	3	1	
69998	99998	22431	1	163	72.0	135	80	1	2	
69999	99999	20540	1	170	72.0	120	80	2	1	

	smoke	alco	active	cardio
69995	1	0	1	0
69996	0	0	1	1
69997	0	1	0	1
69998	0	0	0	1
69999	0	0	1	0

Figure 3 Last 5 records in the dataset

```
Dataset Info:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 70000 entries, 0 to 69999
Data columns (total 13 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   id           70000 non-null  int64
 1   age          70000 non-null  int64
 2   gender       70000 non-null  int64
 3   height       70000 non-null  int64
 4   weight       70000 non-null  float64
 5   ap_hi        70000 non-null  int64
 6   ap_lo        70000 non-null  int64
 7   cholesterol  70000 non-null  int64
 8   gluc         70000 non-null  int64
 9   smoke        70000 non-null  int64
 10  alco         70000 non-null  int64
 11  active       70000 non-null  int64
 12  cardio       70000 non-null  int64
dtypes: float64(1), int64(12)
memory usage: 6.9 MB
```

Figure 4 Checks on Data(columns type,NULLS)

```
Shape of Healthcare dataset ---> (70000, 13)
Check Duplication in the dataset ---> 0
Check Nulls in the dataset --->
 id             0
age             0
gender          0
height          0
weight          0
ap_hi           0
ap_lo           0
cholesterol     0
gluc            0
smoke           0
alco            0
active          0
cardio          0
dtype: int64
```

Figure 5 Checks on Duplicates and Nulls

From the above figure it is found that:

- No Duplicates.
- No NULLs.
- No need for encoding.

2.2.2 Data Distribution and Handling:

	id	age	gender	height	weight	ap_hi	ap_lo	cholesterol	gluc	smoke
count	70000.000000	70000.000000	70000.000000	70000.000000	70000.000000	70000.000000	70000.000000	70000.000000	70000.000000	70000.000000
mean	49972.419900	19468.865814	1.349843	164.359229	74.205690	128.817286	96.630414	1.366871	1.226457	0.088129
std	28851.302323	2467.251667	0.477253	8.210126	14.395757	154.011419	188.472530	0.680250	0.572270	0.283484
min	0.000000	10798.000000	1.000000	55.000000	10.000000	-150.000000	-70.000000	1.000000	1.000000	0.000000
25%	25006.750000	17664.000000	1.000000	159.000000	65.000000	120.000000	80.000000	1.000000	1.000000	0.000000
50%	50001.500000	19703.000000	1.000000	165.000000	72.000000	120.000000	80.000000	1.000000	1.000000	0.000000
75%	74889.250000	21327.000000	2.000000	170.000000	82.000000	140.000000	90.000000	2.000000	1.000000	0.000000
max	99999.000000	23713.000000	3.000000	250.000000	200.000000	16020.000000	11000.000000	3.000000	3.000000	1.000000

Figure 6 Data Description Part1

alco	active	cardio
70000.000000	70000.000000	70000.000000
0.053771	0.803729	0.499700
0.225568	0.397179	0.500003
0.000000	0.000000	0.000000
0.000000	1.000000	0.000000
0.000000	1.000000	0.000000
0.000000	1.000000	1.000000
1.000000	1.000000	1.000000

Figure 7 Data Description part2

From the above figures, it is found that the age has very large numbers as it is calculated in days.

So we have converted days into years then we have shown a sample of the data after conversion as follow

	id	age	gender	height	weight	ap_hi	ap_lo	cholesterol	gluc	smoke	alco	active	cardio
0	0	50	2	168	62.0	110	80	1	1	0	0	1	0
1	1	55	1	156	85.0	140	90	3	1	0	0	1	1
2	2	51	1	165	64.0	130	70	3	1	0	0	0	1
3	3	48	2	169	82.0	150	100	1	1	0	0	1	1
4	4	47	1	156	56.0	100	60	1	1	0	0	0	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...
69995	99993	52	2	168	76.0	120	80	1	1	1	0	1	0
69996	99995	61	1	158	126.0	140	90	2	2	0	0	1	1
69997	99996	52	2	183	105.0	180	90	3	1	0	1	0	1
69998	99998	61	1	163	72.0	135	80	1	2	0	0	0	1
69999	99999	56	1	170	72.0	120	80	2	1	0	0	1	0

70000 rows × 13 columns

Figure 8 Sample of Data after age conversion

```
[ ] Healthcare_Cleaned.describe()
```

	id	age	gender	height	weight	ap_hi	ap_lo	cholesterol	gluc
count	70000.000000	70000.000000	70000.000000	70000.000000	70000.000000	70000.000000	70000.000000	70000.000000	70000.000000
mean	49972.419900	52.840671	1.349843	164.359229	74.205690	128.817286	96.630414	1.366871	1.226457
std	28851.302323	6.766774	0.477253	8.210126	14.395757	154.011419	188.472530	0.680250	0.572270
min	0.000000	29.000000	1.000000	55.000000	10.000000	-150.000000	-70.000000	1.000000	1.000000
25%	25006.750000	48.000000	1.000000	159.000000	65.000000	120.000000	80.000000	1.000000	1.000000
50%	50001.500000	53.000000	1.000000	165.000000	72.000000	120.000000	80.000000	1.000000	1.000000
75%	74889.250000	58.000000	2.000000	170.000000	82.000000	140.000000	90.000000	2.000000	1.000000
max	99999.000000	64.000000	3.000000	250.000000	200.000000	16020.000000	11000.000000	3.000000	3.000000

Figure 9 Data Description after age conversion. Changes are in age

Then we have drown distribution of each key feature as follow:

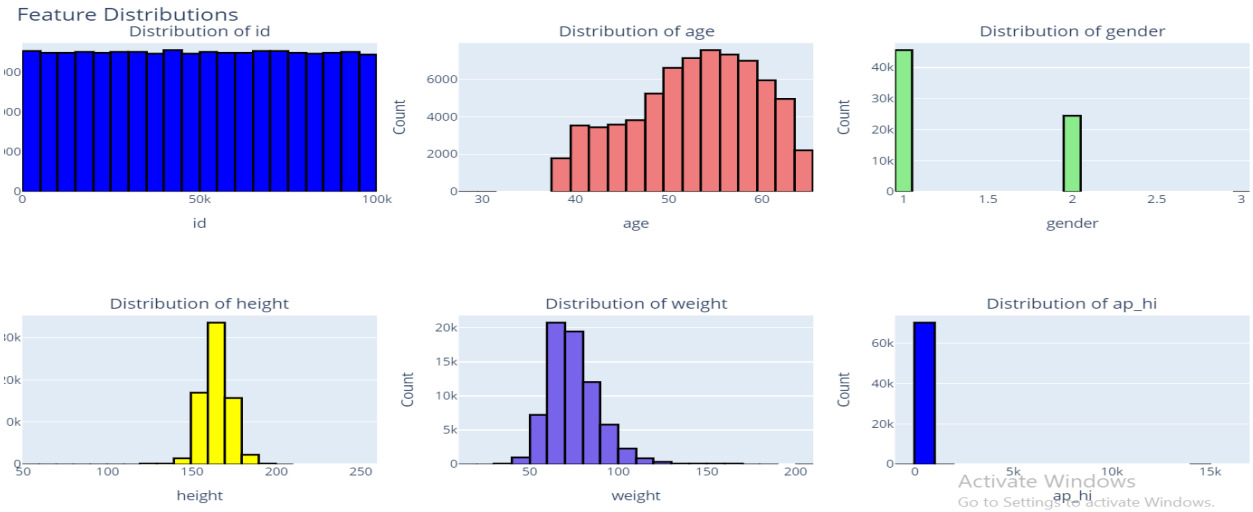

Figure 10 Features Distribution part1

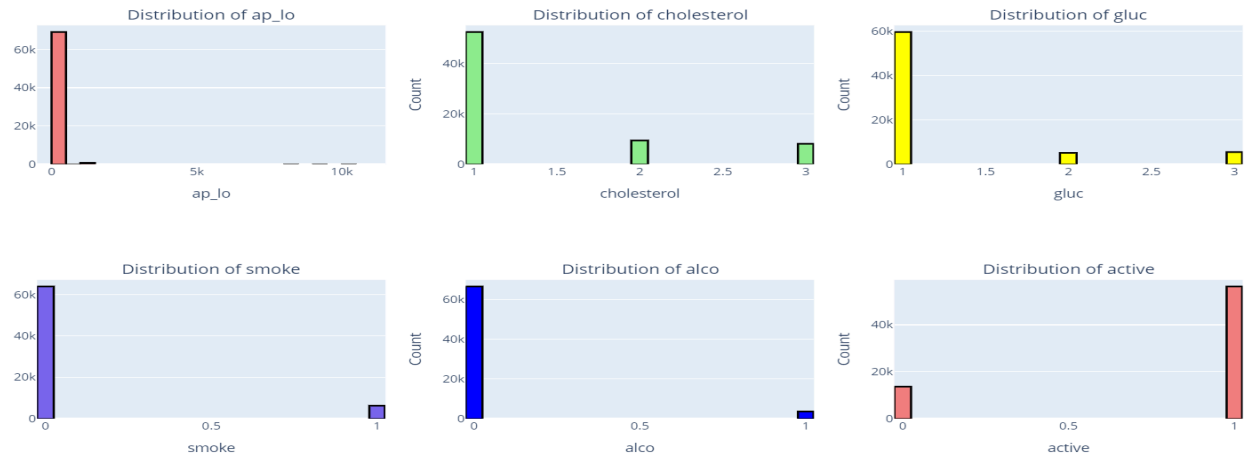

Figure 11 Features Distribution Part2

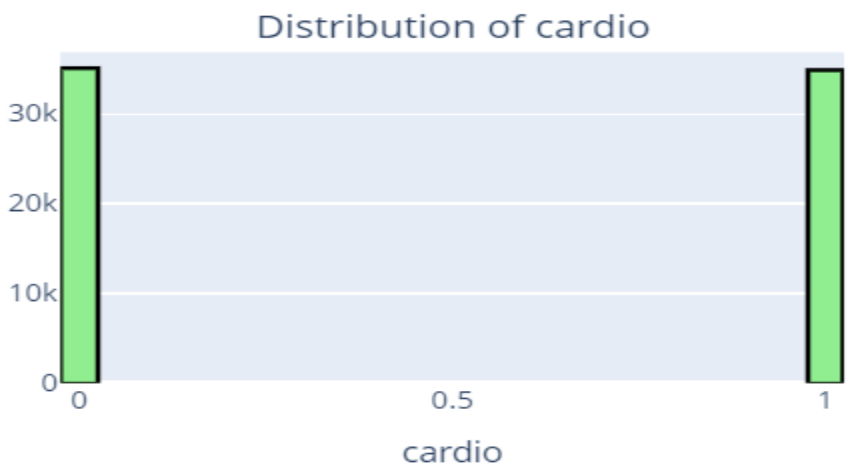

Figure 12 Features Distribution Part3

From the above figure we found that:

- Age distribution is left skewed data so it needs normalization to convert the left skewed to Gaussian distribution.
- Gender contains 3 categories which is not logic as 0 for Males, 1 for Females and 2 for what?! so it depends on the number of samples of this category. the number of samples is 11 records as shown below, it is recommended to eliminate it.

```
gender
1    45522
2    24467
3       11
Name: count, dtype: int64
gender
1    45522
2    24467
Name: count, dtype: int64
```

Figure 13 Number of records in which gender=3

- Height and weight are almost Gaussian distribution.
- ap_hi and ap_lo seem to have outliers as the maximum values from data description are 16020 and 11000 respectively which are not logic values and number of samples at these values are not large so we can handle them by elimination as follow.
- In features the most dominant samples are the normal samples of people but cardiovascular column is balanced which means that there are outliers in the dataset.

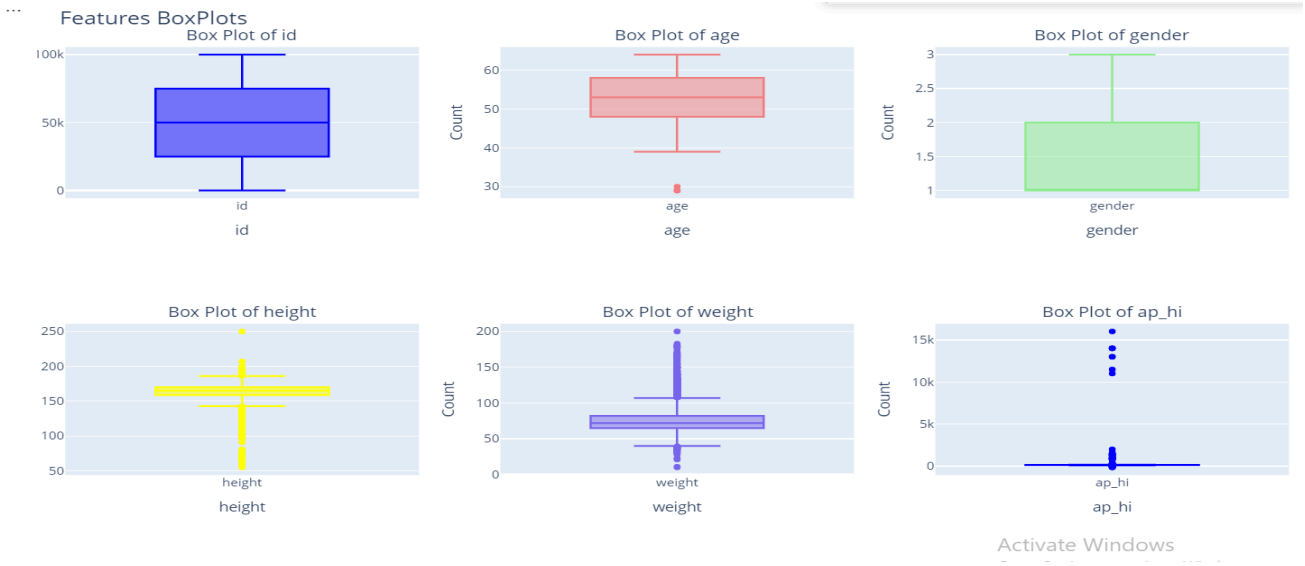

Figure 14 Features Box plots part1

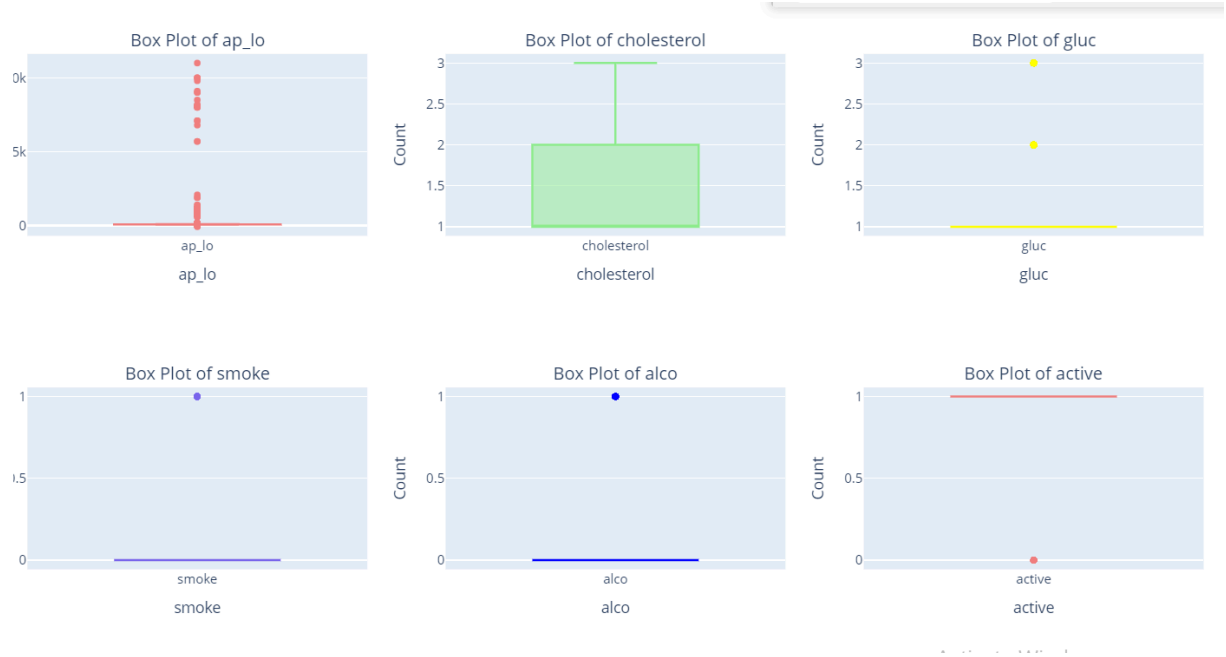

Figure 15 Features Box Plots part2

From the above figures we found that:

- The minimum and maximum values of age are 29 and 64 respectively which are normal values so we decided to keep it as it is.

- In height column minimum and maximum values are 55 and 250 cm respectively. But some of these values are not logic and not correlated with age. Thresholds we have chosen are 100 cm for the lower threshold and 200 cm for the upper threshold. Based on these thresholds outlier samples count was 31 sample so we decided to eliminate them.
- In weight column minimum and maximum values are 10 and 200 kg respectively. Value 10 does not match with the minimum value of age 29 so our thresholds for weight were 45 kg as a lower threshold and 190 kg as an upper threshold. We found that 304 samples were out of this range (outlier) so we decided to eliminate them.
- For ap_hi and ap_lo thresholds chosen are dependent on medical domain as follow:

Condition	Possible Lowest Systolic (`ap_hi`)	Description
Healthy (Normal Range)	90 mmHg	Generally considered the safe lower limit for normal individuals.
Hypotension	60 - 89 mmHg	May cause dizziness, fainting, risk of shock if too low.
Critical Hypotension	Below 60 mmHg	Associated with severe conditions like shock, organ failure, or trauma.

Figure 16 Possible Lowest Systolic (ap_hi)

From the above figure we found that the possible lowest systolic value is 60 mmHg below that the human will not be alive. But if there is a noise in the device used to measure the pressure this may affect the measurement to we have chosen the lower threshold to be 50 mmHg.

Category	Systolic (ap_hi)	Description
Normal	90 - 119 mmHg	Ideal blood pressure.
Elevated	120 - 129 mmHg	Increased risk if not managed.
Hypertension Stage 1	130 - 139 mmHg	Requires lifestyle changes or medication.
Hypertension Stage 2	140 - 179 mmHg	High risk of cardiovascular disease; treatment needed.
Hypertensive Crisis	180 mmHg and above	Immediate medical attention required.

Figure 17 Possible Highest Systolic(ap_hi)

From the above figure we found that the maximum value for ap_hi is 180 mmHg but during search we found that the maximum value ap_hi can be taken and the human is a live is around 250 mmHg.

So thresholds chosen for ap_hi from 50 mmHg to 250 mmHg.

Based on the above thresholds it is found that 224 samples are considered as outliers so we decided to eliminate these samples.

- For ap_lo from search we found that:

Condition	Possible Diastolic (ap_lo) Range	Description
Normal	60 - 79 mmHg	Healthy blood pressure.
Elevated	80 - 89 mmHg	Potential risk of hypertension.
Hypertension Stage 1	90 - 99 mmHg	Requires monitoring and treatment.
Hypertension Stage 2	100 - 119 mmHg	High risk; medical treatment often required.
Hypertensive Crisis	120 - 150 mmHg	Emergency; very rare to exceed this range in valid data.
Hypotension (Low BP)	40 - 59 mmHg	May cause dizziness and fainting; emergency if too low.
Critical Low (Possible Error)	Below 40 mmHg	Unlikely to be physiologically valid, potential data error.

Figure 18 Thresholds for ap_lo

From The above figure we found that the minimum value for ap_lo is 40 mmHg and the upper threshold 150 mmHg.

Based on the previous thresholds we found that 1013 samples as outlier we decided not to eliminate these samples as they represent 1.4%.

The logic used to impute the outliers is as follow:

- By logic, people whose ap_lo is lower than 40 mmHg and higher than 150 mmHg definitely suffer from cardiovascular disease which means cardio flag=1. This will give us subset of data. Number of samples in this subset is 837 records. we decide to calculate the mean of ap_lo of these people and impute the outlier values with this mean.

	id	age	gender	height	weight	ap_hi	ap_lo	cholesterol	gluc	smoke
count	33912.000000	33912.000000	33912.000000	33912.000000	33912.000000	33912.000000	33912.000000	33912.000000	33912.000000	33912.000000
mean	50080.200047	54.462373	1.353149	164.323602	76.795081	133.829323	84.620370	1.516749	1.278043	0.083510
std	28825.283297	6.353931	0.478202	8.057732	14.745338	17.342840	9.630336	0.776686	0.625256	0.276656
min	1.000000	39.000000	1.000000	100.000000	45.000000	70.000000	45.000000	1.000000	1.000000	0.000000
25%	25257.500000	50.000000	1.000000	159.000000	66.000000	120.000000	80.000000	1.000000	1.000000	0.000000
50%	50134.500000	55.000000	1.000000	165.000000	75.000000	130.000000	80.000000	1.000000	1.000000	0.000000
75%	74996.500000	60.000000	2.000000	170.000000	85.000000	140.000000	90.000000	2.000000	1.000000	0.000000
max	99998.000000	64.000000	3.000000	198.000000	183.000000	240.000000	140.000000	3.000000	3.000000	1.000000

Figure 19 Mean of ap_hi of subset (ap_hi < 40mmHg & ap_hi > 150)

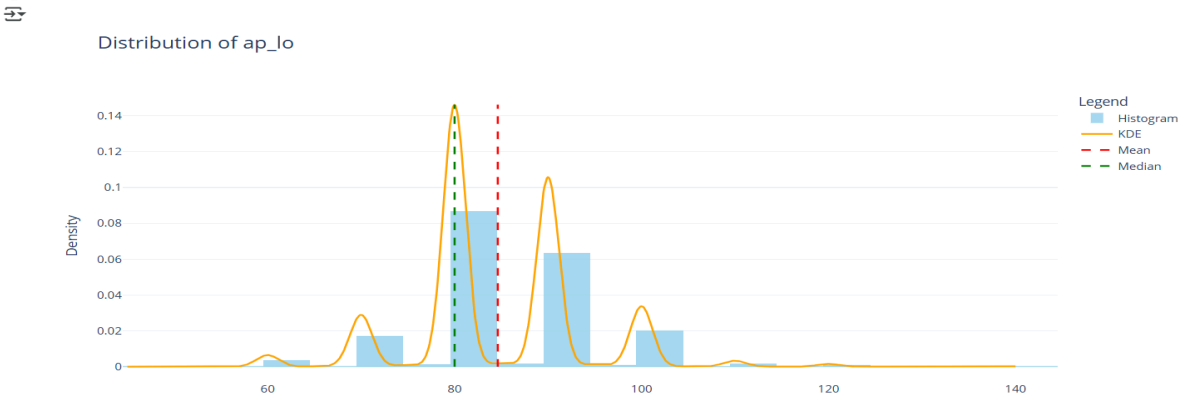

Figure 20 Distribution of ap_lo filtered on (<40 & >150 & cardio=1)

From the above figure we found that mean=84 and median=80 so we substituted by mean.

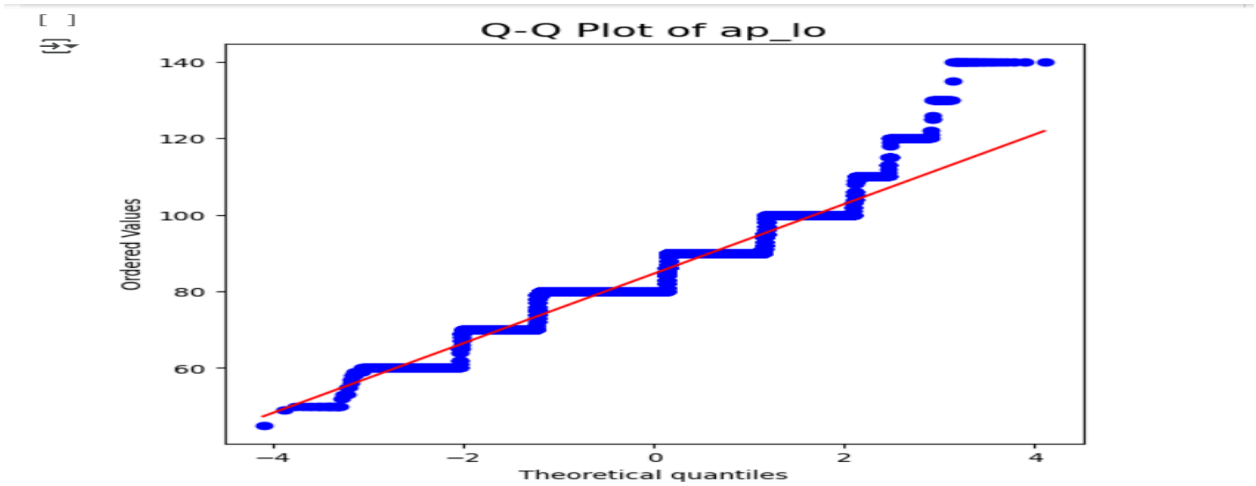

Figure 21 Q-Q Plot of ap_lo after removing outliers

This is a Q-Q plot of ap_lo after substitution which means that the data distribution almost became normal distribution.

```
[ ]  Healthcare_Cleaned
```

	id	age	gender	height	weight	ap_hi	ap_lo	cholesterol	gluc	smoke	alco	active	cardio
0	0	50	2	168	62.0	110	80	1	1	0	0	1	0
1	1	55	1	156	85.0	140	90	3	1	0	0	1	1
2	2	51	1	165	64.0	130	70	3	1	0	0	0	1
3	3	48	2	169	82.0	150	100	1	1	0	0	1	1
4	4	47	1	156	56.0	100	60	1	1	0	0	0	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...
69995	99993	52	2	168	76.0	120	80	1	1	1	0	1	0
69996	99995	61	1	158	126.0	140	90	2	2	0	0	1	1
69997	99996	52	2	183	105.0	180	90	3	1	0	1	0	1
69998	99998	61	1	163	72.0	135	80	1	2	0	0	0	1
69999	99999	56	1	170	72.0	120	80	2	1	0	0	1	0

69441 rows × 13 columns

Figure 22 Sample of Dataset after cleaning

```
[ ]  Healthcare_Cleaned.describe()
```

	id	age	gender	height	weight	ap_hi	ap_lo	cholesterol	gluc
count	69416.000000	69416.000000	69416.000000	69416.000000	69416.000000	69416.000000	69416.000000	69416.000000	69416.000000
mean	49957.639492	52.847903	1.351086	164.442780	74.349992	127.060303	81.424124	1.367711	1.227008
std	28854.763496	6.761097	0.477645	7.953601	14.227458	17.068139	9.477888	0.681076	0.572872
min	0.000000	29.000000	1.000000	100.000000	45.000000	60.000000	40.000000	1.000000	1.000000
25%	24982.750000	48.000000	1.000000	159.000000	65.000000	120.000000	80.000000	1.000000	1.000000
50%	49980.500000	53.000000	1.000000	165.000000	72.000000	120.000000	80.000000	1.000000	1.000000
75%	74882.000000	58.000000	2.000000	170.000000	82.000000	140.000000	90.000000	2.000000	1.000000
max	99999.000000	64.000000	3.000000	200.000000	183.000000	240.000000	150.000000	3.000000	3.000000

Figure 23 Data Description after cleaning

There is a bug found during data investigation:

It is found that there are some records in which ap_lo > ap_hi which does not make sense.

2.2.2.1 Features Distribution and Box plots after data handling

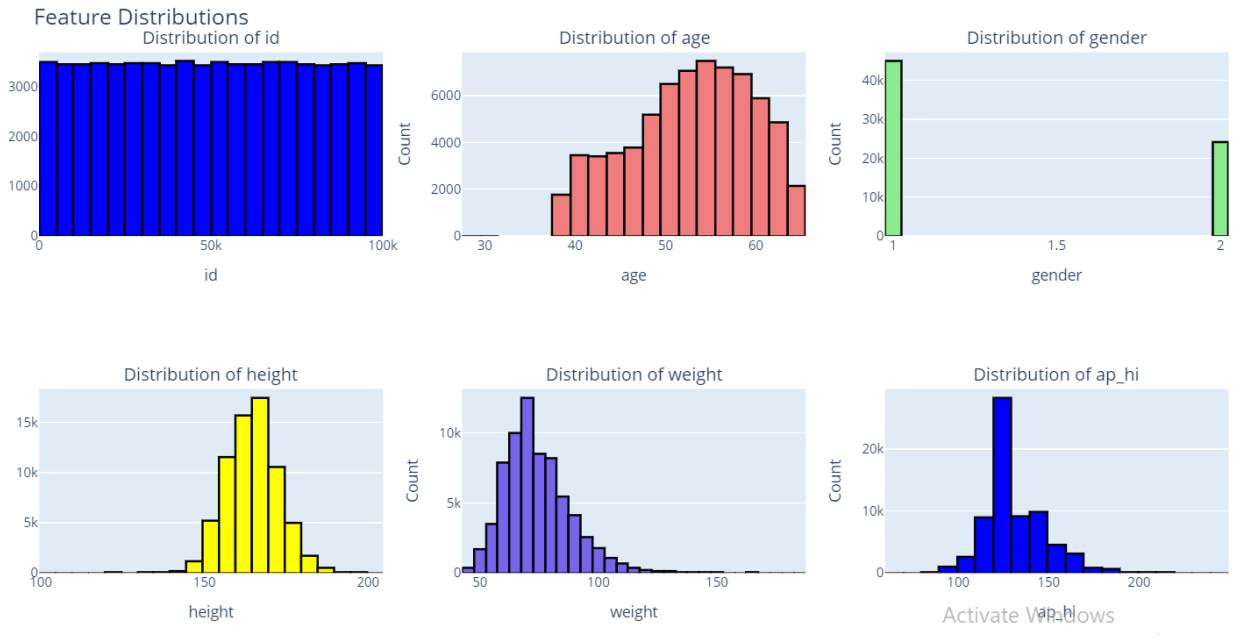

Figure 24 Features Distribution Part1

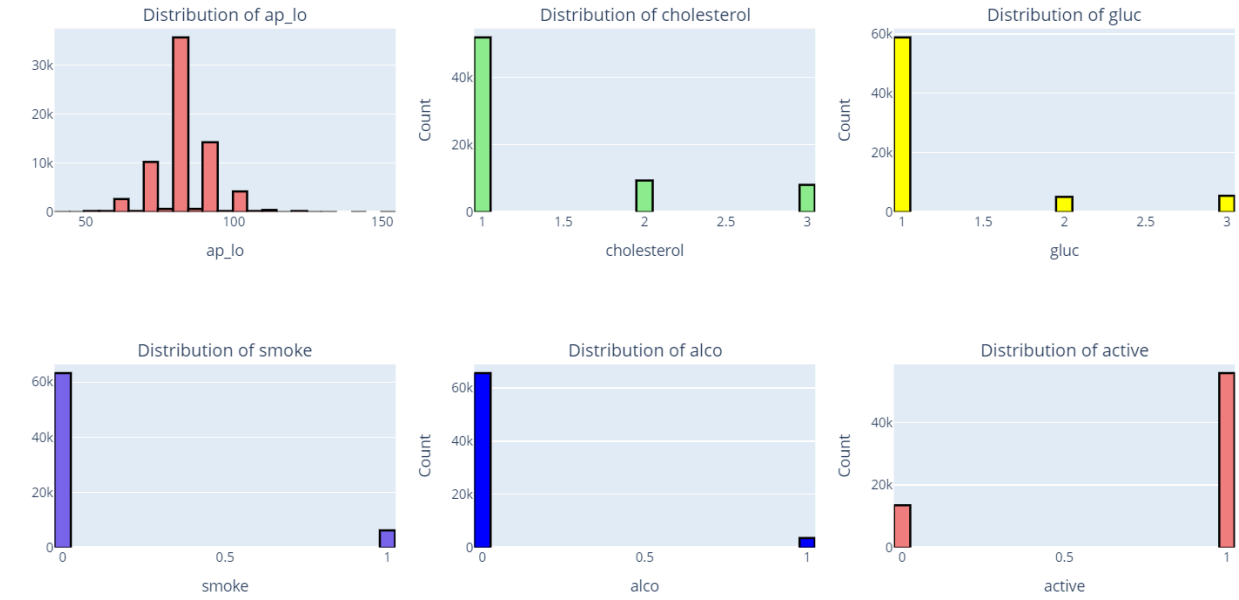

Figure 25 Features Distribution Part2

From the above figures it is found that:

- *Gender feature now contains only two categories.*
- *Distribution of ap_hi & ap_lo are enhanced*