

OPTIMIZATION FOR DATA SCIENCE



Project - Heart Diagnosis

Submitted by

Kouadio Yao innocent junior

Supervisor

Professor: Loualitene FATAH

EPITA : École Pour l'Informatique et les Techniques Avancées - 19/02/2020

Table of Contents

Introduction	3
Dataset	4
Dataset One	4
Variables of the first dataset.....	4
Dataset Two	5
Variable for the second dataset.....	5
Exploratory Data Analysis	7
Integer variables	7
Binary Variables (Integer Variables)	9
Dataset Two.....	10
Cleaning Data - Dealing with Missing Data	11
Questions raised from the dataset	12
Deep Analysis the Dataset	12
Which Factors have the correlation to each other's?	18
Prediction Model 1 - The Algorithm to checking Heart Disease	22
Different Logistic Regression Tested for Our model.....	23
The extension of the project	24
The external dataset analysis	24
Prediction Model 2 - The Algorithm to checking Heart Disease	27
Conclusion	27
Acknowledgements	28
Appendix.....	29



Introduction

Heart disease is one of the most dangerous human diseases of all times. With the evolution in technology and medicine, we now can have more chances to improve the human's life, especially doctors recently can have a lot of information about the patient from medical devices to support the diagnosis of health condition. Hence, the measurement data from medical devices are the real gold for doctors to have the right decision on each of every person. Since the importance of understanding the data in heart disease and interest in data analysis and prediction model. Our team would like to investigate heart disease through the project Heart Diagnosis with Data Optimization.

Dataset

Dataset One

Our team has chosen a Dataset of Heart Disease from the website Kaggle.com. This Dataset is based on the real case study from the Hungarian Institute of Cardiology, University of Hospital Zurich, University of Hospital Basel and V.A. Medical Center, Long Beach and Cleveland Clinic Foundation USA. The Dataset has 14 columns and 304 rows with 13 types of integer variables (including 4 binary variables) and 1 float variables. Each of the variables describes the factor (index) that is related to heart disease (target). We are going to explain further the variables in the next parts.

Variables of the first dataset

Features

- Age: All the patient has declared his or her age in the data. The minimum age is 29 and the maximum age is 77
- Sex: 0 means the female gender, 1 means the male gender
- CP: CP measures that different types of chest pain (4 values). There are 4 types of chest pain: Value 0 - typical angina, Value 1 - atypical angina, Value 2 - non anginal type, Value 3 - asymptomatic
- Trestbps: This means "resting blood pressure".
- Chol: This data shows the index of serum cholesterol in mg/dl
- FSB: fasting blood sugar > 120 mg/dl (1 = true; 0 = false)
- Restecg: resting electrocardiographic results (values 0,1,2). There are 3 types of Restecg: Value 0 - normal, Value 1 - having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV), Value 2 - showing probable or definite left ventricular hypertrophy by Estes criteria
- Thalach: shows the maximum heart rate achieved
- Exang: exercise induced angina (1 = yes, 0 = no)

- Oldpeak: Oldpeak means ST depression induced by exercise relative to rest
- Slope: the slope of the peak exercise ST segment. Value 1 - upsloping, Value 2 - flat, Value 3 - downsloping
- CA: number of major vessels (0-4) colored by fluoroscopy
- Thal: the meaning of the data is “ 3 = normal; 6 = fixed defect; 7 = reversible defect”
- Target: 0 means don't have heart disease, 1 means that he or she is having heart disease. This is a binary classification problem.

Description of Dataset One:

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
count	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000
mean	54.366337	0.683168	0.966997	131.623762	246.264026	0.148515	0.528053	149.646865	0.326733	1.039604	1.399340	0.729373	2.313531	0.544554
std	9.082101	0.466011	1.032052	17.538143	51.830751	0.356198	0.525860	22.905161	0.469794	1.161075	0.616226	1.022606	0.612277	0.498835
min	29.000000	0.000000	0.000000	94.000000	126.000000	0.000000	0.000000	71.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	47.500000	0.000000	0.000000	120.000000	211.000000	0.000000	0.000000	133.500000	0.000000	0.000000	1.000000	0.000000	2.000000	0.000000
50%	55.000000	1.000000	1.000000	130.000000	240.000000	0.000000	1.000000	153.000000	0.000000	0.800000	1.000000	0.000000	2.000000	1.000000
75%	61.000000	1.000000	2.000000	140.000000	274.500000	0.000000	1.000000	166.000000	1.000000	1.600000	2.000000	1.000000	3.000000	1.000000
max	77.000000	1.000000	3.000000	200.000000	564.000000	1.000000	2.000000	202.000000	1.000000	6.200000	2.000000	4.000000	3.000000	1.000000

Figure 1 – Dataset 1 Summary table

Dataset Two

The second dataset has been selected after analysis of the first dataset, A question has arisen in us to know why man is a potential subject of heart issue over women , due to that doubt we have download the second dataset from Kaggle which give us some information about human activities that might be a factor of a heart disease . This data set is real data collected from patients after examination.

Variable for the second dataset

Features

- Age | Objective Feature | age | int (days)`
- Height | Objective Feature | height | int (cm) |

- Weight | Objective Feature | weight | float (kg) |
- Gender | Objective Feature | gender | categorical code |
- Systolic blood pressure | Examination Feature | ap_hi | int |
- Diastolic blood pressure | Examination Feature | ap_lo | int |
- Cholesterol | Examination Feature | cholesterol | 1: normal, 2: above normal, 3: well above normal |
- Glucose | Examination Feature | gluc | 1: normal, 2: above normal, 3: well above normal |
- Smoke: Smoking | Subjective Feature | binary |
- Alco: Alcohol intake | Subjective Feature | binary |
- Active: Physical activity | Subjective Feature | binary |
- Cardio: Presence or absence of cardiovascular disease | Target Variable | cardio | binary |

Description of Dataset Two

	age	gender	height	weight	ap_hi	ap_lo	cholesterol	gluc	smoke	alco	active	cardio
count	70000.000000	70000.000000	70000.000000	70000.000000	70000.000000	70000.000000	70000.000000	70000.000000	70000.000000	70000.000000	70000.000000	70000.000000
mean	53.583786	1.349571	164.359229	74.205690	128.817286	96.630414	1.366871	1.226457	0.088129	0.053771	0.803729	0.499700
std	6.860581	0.476838	8.210126	14.395757	154.011419	188.472530	0.680250	0.572270	0.283484	0.225568	0.397179	0.500003
min	29.000000	1.000000	55.000000	10.000000	-150.000000	-70.000000	1.000000	1.000000	0.000000	0.000000	0.000000	0.000000
25%	49.000000	1.000000	159.000000	65.000000	120.000000	80.000000	1.000000	1.000000	0.000000	0.000000	1.000000	0.000000
50%	54.000000	1.000000	165.000000	72.000000	120.000000	80.000000	1.000000	1.000000	0.000000	0.000000	1.000000	0.000000
75%	59.000000	2.000000	170.000000	82.000000	140.000000	90.000000	2.000000	1.000000	0.000000	0.000000	1.000000	1.000000
max	65.000000	2.000000	250.000000	200.000000	16020.000000	11000.000000	3.000000	3.000000	1.000000	1.000000	1.000000	1.000000

Figure 2 – Dataset Two Summary table

Exploratory Data Analysis

Dataset One

Integer variables

We will discover the insights from the distribution of the variables:

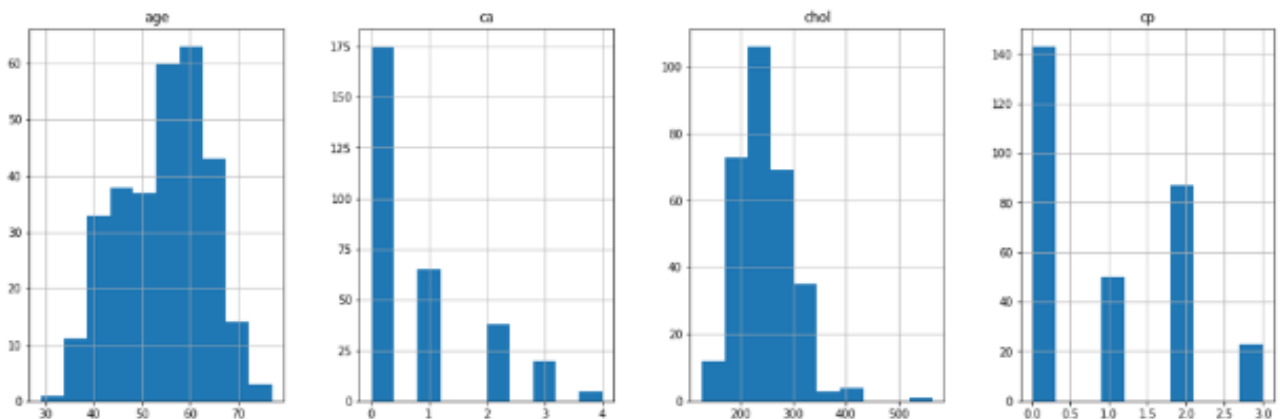


Figure 3 - From left to right: Age, Ca, Chol, Cp

We can observe the frequency of age , chest pain , cholesterol and number of major vessels .

- We observe the age range of people in the data set slice from]30,70[years old ,also from graph we have more subject people in the range of 50 to 60 years old who have been diagnostic from the hospital , we can make a conclusion that people slicing between 50 to 60+ years old check their heart health more frequently compare to others .
- Number of major vessels 0 3 colored by fluoroscopy range from (0 to 4), if the number is greater than zero then is a potential disease otherwise if it is 0 implies no potential

disease, from how dataset we state that we have more people with no potential heart issue .

- A cholesterol level higher than 200 for an adult is considered high ,we can observe some value reaching 300 which is a potential risk patient
- Any chest pain value greater than 0 , is view as a potential risk of heart disease.

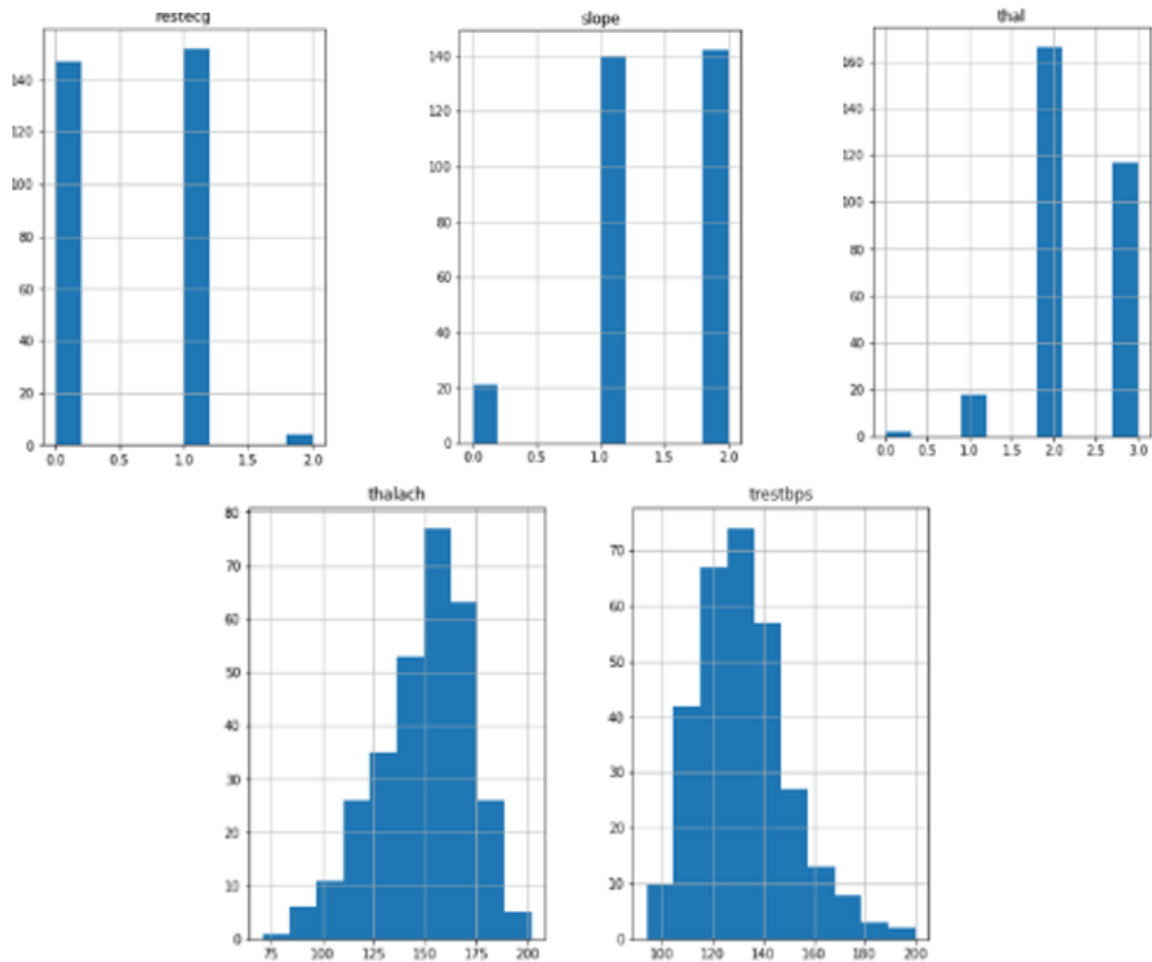


Figure 4: From left to right and from top to bottom, histogram of Restecg, Slope, Thai, Thalach, Trestbps

As in the Figure, we can see the highest value for Thalach skews lies between 125 to 175, meanwhile the Trestbps skews lies between 110 to 150 respectively.. For Thai, the figure also shows the highest type which is 2 and 3, at the same time. In terms of Chol factor, the data is skewed into 150 and 300 which shows the result of the heart beat over a group of patients .

Binary Variables (Integer Variables)

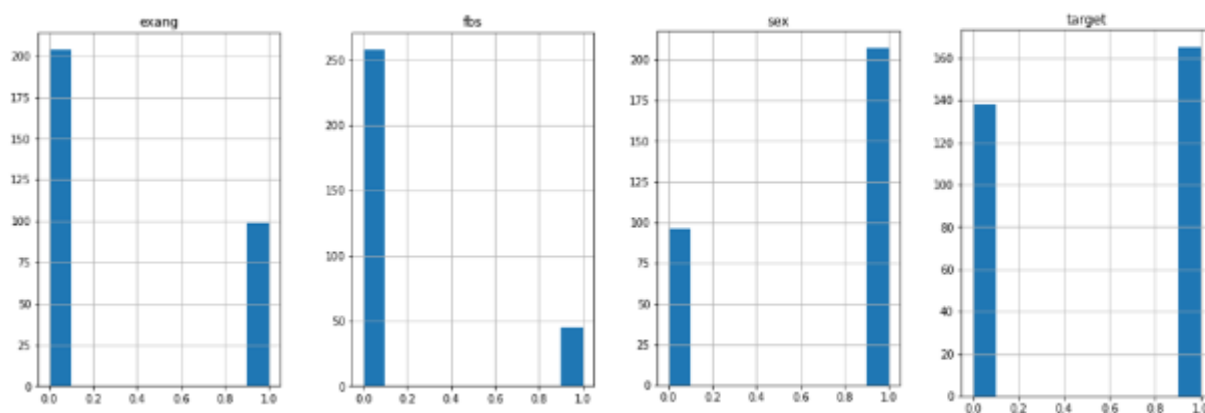


Figure 5: Histogram of Exang, Fbs, Sec and Target

In terms of gender, we can see the number of men is double that of women. On the other side, on Target side we observe the number of heart diseases of people are quite similar versus people with non heart disease. The figure also shows that Majority of Exang and Fbs are 0.

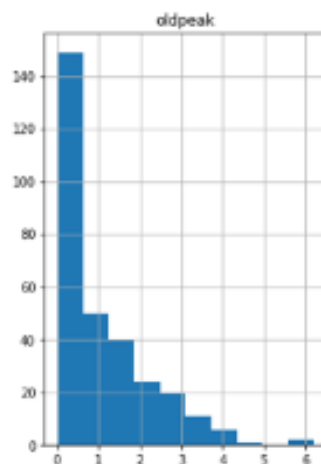


Figure 6: Float Variables - Oldpeak

From the Oldpeak data, we can see that the mean score is around 1.04 points, and the majority of indexes are below 1.

Dataset Two

Since, some of the factors are duplicate with the Dataset One, we chose the most relevant data factors from Dataset Two to analyze.

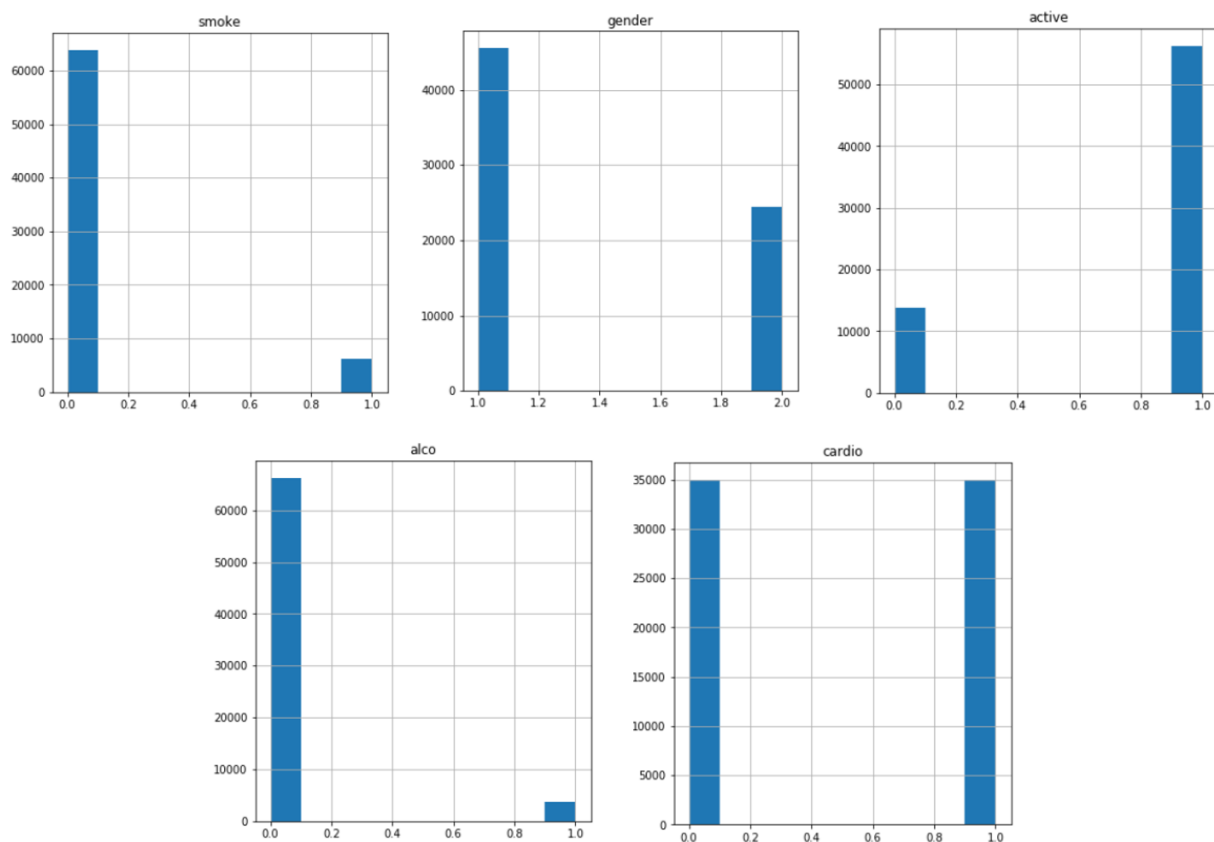


Figure 7 - From left to right, top to bottom - Smoke, Gender, Active, Alco and Cardio

As we have mentioned above, this extra dataset is to find the root cause and activities that lead to heart disease, in this case the classification data is Cardio (0 means normal, 1 means heart disease). There are 4 types of activities: Smoke, Alcohol, Activity (Physical activity). For Smoke

and Alcohol, we can see that 90% of patients did not smoke or drink alcohol. In terms of activity, the majority of patients had physical activity.

There is also a small difference vs Dataset One; 1 means woman and 2 means man.

Cleaning Data - Dealing with Missing Data

The dataset has no missing data; hence the data is ready to analyze. For the session of dealing with missing data, but we apply the data standardization.

<pre> <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 trestbps 303 non-null int64 4 chol 303 non-null int64 5 fbs 303 non-null int64 6 restecg 303 non-null int64 7 thalach 303 non-null int64 8 exang 303 non-null int64 9 oldpeak 303 non-null float64 10 slope 303 non-null int64 11 ca 303 non-null int64 12 thal 303 non-null int64 13 target 303 non-null int64 dtypes: float64(1), int64(13) memory usage: 33.3 KB </pre>	<pre> <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 </pre>
--	---

Figure 8: Data type and Non-Null Count of Dataset One and Two

From Data Standardization, we change the age from number of days to years to ensure the harmonizing between two Dataset.

```
1 import numpy as np
2 import pandas as pd
3 import seaborn as sns
4 import matplotlib.pyplot as plt
5 from sklearn import linear_model
6
7
8
9
10
11
12 activity_set = pd.read_csv('CardioActivity.csv', sep=';')
13
14
15
16
17
18 act = (np.array(activity_set['age']) / 360).astype(int)
19 activity_set['age'] = act
20
21
22
23
24
25 sns.scatterplot(activity_set['smoke'], activity_set['active'], hue=activity_set['cardio'])
26
27
28 #a = activity_set.groupby(['gender', 'smoke'])
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65
66
67
68
69
70
71
72
73
74
75
76
77
78
79
80
81
82
83
84
85
86
87
88
89
90
91
92
93
94
95
96
97
98
99
100
101
102
103
104
105
106
107
108
109
110
111
112
113
114
115
116
117
118
119
120
121
122
123
124
125
126
127
128
129
130
131
132
133
134
135
136
137
138
139
140
141
142
143
144
145
146
147
148
149
150
151
152
153
154
155
156
157
158
159
160
161
162
163
164
165
166
167
168
169
170
171
172
173
174
175
176
177
178
179
180
181
182
183
184
185
186
187
188
189
190
191
192
193
194
195
196
197
198
199
200
201
202
203
204
205
206
207
208
209
210
211
212
213
214
215
216
217
218
219
220
221
222
223
224
225
226
227
228
229
230
231
232
233
234
235
236
237
238
239
240
241
242
243
244
245
246
247
248
249
250
251
252
253
254
255
256
257
258
259
260
261
262
263
264
265
266
267
268
269
270
271
272
273
274
275
276
277
278
279
280
281
282
283
284
285
286
287
288
289
290
291
292
293
294
295
296
297
298
299
300
301
302
303
304
305
306
307
308
309
310
311
312
313
314
315
316
317
318
319
320
321
322
323
324
325
326
327
328
329
330
331
332
333
334
335
336
337
338
339
340
341
342
343
344
345
346
347
348
349
350
351
352
353
354
355
356
357
358
359
360
361
362
363
364
365
366
367
368
369
370
371
372
373
374
375
376
377
378
379
380
381
382
383
384
385
386
387
388
389
390
391
392
393
394
395
396
397
398
399
400
401
402
403
404
405
406
407
408
409
410
411
412
413
414
415
416
417
418
419
420
421
422
423
424
425
426
427
428
429
430
431
432
433
434
435
436
437
438
439
440
441
442
443
444
445
446
447
448
449
450
451
452
453
454
455
456
457
458
459
460
461
462
463
464
465
466
467
468
469
470
471
472
473
474
475
476
477
478
479
480
481
482
483
484
485
486
487
488
489
490
491
492
493
494
495
496
497
498
499
500
501
502
503
504
505
506
507
508
509
510
511
512
513
514
515
516
517
518
519
520
521
522
523
524
525
526
527
528
529
530
531
532
533
534
535
536
537
538
539
540
541
542
543
544
545
546
547
548
549
550
551
552
553
554
555
556
557
558
559
560
561
562
563
564
565
566
567
568
569
570
571
572
573
574
575
576
577
578
579
580
581
582
583
584
585
586
587
588
589
590
591
592
593
594
595
596
597
598
599
600
601
602
603
604
605
606
607
608
609
610
611
612
613
614
615
616
617
618
619
620
621
622
623
624
625
626
627
628
629
630
631
632
633
634
635
636
637
638
639
640
641
642
643
644
645
646
647
648
649
650
651
652
653
654
655
656
657
658
659
660
661
662
663
664
665
666
667
668
669
670
671
672
673
674
675
676
677
678
679
680
681
682
683
684
685
686
687
688
689
690
691
692
693
694
695
696
697
698
699
700
701
702
703
704
705
706
707
708
709
710
711
712
713
714
715
716
717
718
719
720
721
722
723
724
725
726
727
728
729
730
731
732
733
734
735
736
737
738
739
740
741
742
743
744
745
746
747
748
749
750
751
752
753
754
755
756
757
758
759
760
761
762
763
764
765
766
767
768
769
770
771
772
773
774
775
776
777
778
779
780
781
782
783
784
785
786
787
788
789
790
791
792
793
794
795
796
797
798
799
800
801
802
803
804
805
806
807
808
809
810
811
812
813
814
815
816
817
818
819
820
821
822
823
824
825
826
827
828
829
830
831
832
833
834
835
836
837
838
839
840
841
842
843
844
845
846
847
848
849
850
851
852
853
854
855
856
857
858
859
860
861
862
863
864
865
866
867
868
869
870
871
872
873
874
875
876
877
878
879
880
881
882
883
884
885
886
887
888
889
890
891
892
893
894
895
896
897
898
899
900
901
902
903
904
905
906
907
908
909
910
911
912
913
914
915
916
917
918
919
920
921
922
923
924
925
926
927
928
929
930
931
932
933
934
935
936
937
938
939
940
941
942
943
944
945
946
947
948
949
950
951
952
953
954
955
956
957
958
959
960
961
962
963
964
965
966
967
968
969
970
971
972
973
974
975
976
977
978
979
980
981
982
983
984
985
986
987
988
989
990
991
992
993
994
995
996
997
998
999
1000
1001
1002
1003
1004
1005
1006
1007
1008
1009
1010
1011
1012
1013
1014
1015
1016
1017
1018
1019
1020
1021
1022
1023
1024
1025
1026
1027
1028
1029
1030
1031
1032
1033
1034
1035
1036
1037
1038
1039
1040
1041
1042
1043
1044
1045
1046
1047
1048
1049
1050
1051
1052
1053
1054
1055
1056
1057
1058
1059
1060
1061
1062
1063
1064
1065
1066
1067
1068
1069
1070
1071
1072
1073
1074
1075
1076
1077
1078
1079
1080
1081
1082
1083
1084
1085
1086
1087
1088
1089
1090
1091
1092
1093
1094
1095
1096
1097
1098
1099
1100
1101
1102
1103
1104
1105
1106
1107
1108
1109
1110
1111
1112
1113
1114
1115
1116
1117
1118
1119
1120
1121
1122
1123
1124
1125
1126
1127
1128
1129
1130
1131
1132
1133
1134
1135
1136
1137
1138
1139
1140
1141
1142
1143
1144
1145
1146
1147
1148
1149
1150
1151
1152
1153
1154
1155
1156
1157
1158
1159
1160
1161
1162
1163
1164
1165
1166
1167
1168
1169
1170
1171
1172
1173
1174
1175
1176
1177
1178
1179
1180
1181
1182
1183
1184
1185
1186
1187
1188
1189
1190
1191
1192
1193
1194
1195
1196
1197
1198
1199
1200
1201
1202
1203
1204
1205
1206
1207
1208
1209
1210
1211
1212
1213
1214
1215
1216
1217
1218
1219
1220
1221
1222
1223
1224
1225
1226
1227
1228
1229
1230
1231
1232
1233
1234
1235
1236
1237
1238
1239
1240
1241
1242
1243
1244
1245
1246
1247
1248
1249
1250
1251
1252
1253
1254
1255
1256
1257
1258
1259
1260
1261
1262
1263
1264
1265
1266
1267
1268
1269
1270
1271
1272
1273
1274
1275
1276
1277
1278
1279
1280
1281
1282
1283
1284
1285
1286
1287
1288
1289
1290
1291
1292
1293
1294
1295
1296
1297
1298
1299
1300
1301
1302
1303
1304
1305
1306
1307
1308
1309
1310
1311
1312
1313
1314
1315
1316
1317
1318
1319
1320
1321
1322
1323
1324
1325
1326
1327
1328
1329
1330
1331
1332
1333
1334
1335
1336
1337
1338
1339
1340
1341
1342
1343
1344
1345
1346
1347
1348
1349
1350
1351
1352
1353
1354
1355
1356
1357
1358
1359
1360
1361
1362
1363
1364
1365
1366
1367
1368
1369
1370
1371
1372
1373
1374
1375
1376
1377
1378
1379
1380
1381
1382
1383
1384
1385
1386
1387
1388
1389
1390
1391
1392
1393
1394
1395
1396
1397
1398
1399
1400
1401
1402
1403
1404
1405
1406
1407
1408
1409
1410
1411
1412
1413
1414
1415
1416
1417
1418
1419
1420
1421
1422
1423
1424
1425
1426
1427
1428
1429
1430
1431
1432
1433
1434
1435
1436
1437
1438
1439
1440
1441
1442
1443
1444
1445
1446
1447
1448
1449
1450
1451
1452
1453
1454
1455
1456
1457
1458
1459
1460
1461
1462
1463
1464
1465
1466
1467
1468
1469
1470
1471
1472
1473
1474
1475
1476
1477
1478
1479
1480
1481
1482
1483
1484
1485
1486
1487
1488
1489
1490
1491
1492
1493
1494
1495
1496
1497
1498
1499
1500
1501
1502
1503
1504
1505
1506
1507
1508
1509
1510
1511
1512
1513
1514
1515
1516
1517
1518
1519
1520
1521
1522
1523
1524
1525
1526
1527
1528
1529
1530
1531
1532
1533
1534
1535
1536
1537
1538
1539
1540
1541
1542
1543
1544
1545
1546
1547
1548
1549
1550
1551
1552
1553
1554
1555
1556
1557
1558
1559
1560
1561
1562
1563
1564
1565
1566
1567
1568
1569
1570
1571
1572
1573
1574
1575
1576
1577
1578
1579
1580
1581
1582
1583
1584
1585
1586
1587
1588
1589
1590
1591
1592
1593
1594
1595
1596
1597
1598
1599
1600
1601
1602
1603
1604
1605
1606
1607
1608
1609
1610
1611
1612
1613
1614
1615
1616
1617
1618
1619
1620
1621
1622
1623
1624
1625
1626
1627
1628
1629
1630
1631
1632
1633
1634
1635
1636
1637
1638
1639
1640
1641
1642
1643
1644
1645
1646
1647
1648
1649
1650
1651
1652
1653
1654
1655
1656
1657
1658
1659
1660
1661
1662
1663
1664
1665
1666
1667
1668
1669
1670
1671
1672
1673
1674
1675
1676
1677
1678
1679
1680
1681
1682
1683
1684
1685
1686
1687
1688
1689
1690
1691
1692
1693
1694
1695
1696
1697
1698
1699
1700
1701
1702
1703
1704
1705
1706
1707
1708
1709
1710
1711
1712
1713
1714
1715
1716
1717
1718
1719
1720
1721
1722
1723
1724
1725
1726
1727
1728
1729
1730
1731
1732
1733
1734
1735
1736
1737
1738
1739
1740
1741
1742
1743
1744
1745
1746
1747
1748
1749
1750
1751
1752
1753
1754
1755
1756
1757
1758
1759
1760
1761
1762
1763
1764
1765
1766
1767
1768
1769
1770
1771
1772
1773
1774
1775
1776
1777
1778
1779
1780
1781
1782
1783
1784
1785
1786
1787
1788
1789
1790
1791
1792
1793
1794
1795
1796
1797
1798
1799
1800
1801
1802
1803
1804
1805
1806
1807
1808
1809
1810
1811
1812
1813
1814
1815
1816
1817
1818
1819
1820
1821
1822
1823
1824
1825
1826
1827
1828
1829
1830
1831
1832
1833
1834
1835
1836
1837
1838
1839
1840
1841
1842
1843
1844
1845
1846
1847
1848
1849
1850
1851
1852
1853
1854
1855
1856
1857
1858
1859
1860
1861
1862
1863
1864
1865
1866
1867
1868
1869
1870
1871
1872
1873
1874
1875
1876
1877
1878
1879
1880
1881
1882
1883
1884
1885
1886
1887
1888
1889
1890
1891
1892
1893
1894
1895
1896
1897
1898
1899
1900
1901
1902
1903
1904
1905
1906
1907
1908
1909
1910
1911
1912
1913
1914
1915
1916
1917
1918
1919
1920
1921
1922
1923
1924
1925
1926
1927
1928
1929
1930
1931
1932
1933
1934
1935
1936
1937
1938
1939
1940
1941
1942
1943
1944
1945
1946
1947
1948
1949
1950
1951
1952
1953
1954
1955
1956
1957
1958
1959
1960
1961
1962
1963
1964
1965
1966
1967
1968
1969
1970
1971
1972
1973
1974
1975
1976
1977
1978
1979
1980
1981
1982
1983
1984
1985
1986
1987
1988
1989
1990
1991
1992
1993
1994
1995
1996
1997
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
2021
2022
2023
2024
2025
2026
2027
2028
2029
2030
2031
2032
2033
2034
2035
2036
2037
2038
2039
2040
2041
2042
2043
2044
2045
2046
2047
2048
2049
2050
2051
2052
2053
2054
2055
2056
2057
2058
2059
2060
2061
2062
2063
2064
2065
2066
2067
2068
2069
2070
2071
2072
2073
2074
2075
2076
2077
2078
2079
2080
2081
2082
2083
2084
2085
2086
2087
2088
2089
2090
2091
2092
2093
2094
2095
2096
2097
2098
2099
2100
2101
2102
2103
2104
2105
2106
2107
2108
2109
2110
2111
2112
2113
2114
2115
2116
2117
2118
2119
2120
2121
2122
2123
2124
2125
2126
2127
2128
2129
2130
2131
2132
2133
2134
2135
2136
2137
2138
2139
2140
2141
2142
2143
2144
2145
2146
2147
2148
2149
2150
2151
2152
2153
2154
2155
2156
2157
2158
2159
2160
2161
2162
2163
2164
2165
2166
2167
2168
2169
2170
2171
2172
2173
2174
2175
2176
2177
2178
2179
2180
2181
2182
2183
2184
2185
2186
2187
2188
2189
2190
2191
2192
2193
2194
2195
2196
2197
2198
2199
2200
2201
2202
2203
2204
2205
2206
2207
2208
2209
2210
2211
2212
2213
2214
2215
2216
2217
2218
2219
2220
2221
2222
2223
2224
2225
2226
2227
2228
2229
2230
2231
2232
2233
2234
2235
2236
2237
2238
2239
2240
2241
2242
2243
2244
2245
2246
2247
2248
2249
2250
2251
2252
2253
2254
2255
2256
2257
2258
2259
2260
2261
2262
2263
2264
2265
2266
2267
2268
2269
2270
2271
2272
2273
2274
2275
2276
2277
2278
2279
2280
2281
2282
2283
2284
2285
2286
2287
2288
2289
2290
2291
2292
2293
2294
2295
2296
2297
2298
2299
2300
2301
2302
2303
2304
2305
2306
2307
2308
2309
2310
2311
2312
2313
2314
2315
2316
2317
2318
2319
2320
2321
2322
2323
2324
2325
2326
2327
2328
2329
2330
2331
2332
2333
2334
2335
2336
2337
2338
2339
2340
2341
2342
2343
2344
2345
2346
2347
2348
2349
2350
2351
2352
2353
2354
2355
2356
2357
2358
2359
2360
2361
2362
2363
2364
2365
2366
2367
2368
2369
2370
2371
2372
2373
2374
2375
2376
2377
2378
2379
2380
2381
2382
2383
2384
2385
2386
2387
2388
2389
2390
2391
2392
2393
2394
2395
2396
2397
2398
2399
2400
2401
2402
2403
2404
2405
2406
2407
2408
2409
2410
2411
2412
2413
2414
2415
2416
2417
2418
2419
2420
2421
2422
2423
2424
2425
2426
2427
2428
2429
2430
2431
2432
2433
2434
2435
2436
2437
2438
2439
2440
2441
2442
2443
2444
2445
2446
2447
2448
2449
2450
2451
2452
2453
2454
2455
2456
2457
2458
2459
2460
2461
2462
2463
2464
2465
2466
2467
2468
2469
2470
2471
2472
2473
2474
2475
2476
2477
2478
2479
2480
2481
2482
2483
2484
2485
2486
2487
2488
2489
2490
2491
2492
2493
2494
2495
2496
2497
2498
2499
2500
2501
2502
2503
2504
2505
2506
2507
2508
2509
2510
2511
2512
2513
2514
2515
2516
2517
2518
2519
2520
2521
2522
2523
2524
2525
2526
2527
2528
2529
2530
2531
2532
2533
2534
2535
2536
2537
2538
2539
2540
2541
2542
2543
2544
2545
2546
2547
2548
2549
2550
2551
2552
2553
2554
2555
2556
2557
2558
2559
2560
2561
2562
2563
2564
2565
2566
2567
2568
2569
2570
2571
2572
2573
2574
2575
2576
2577
2578
2579
2580
2581
2582
2583
2584
2585
2586
2587
2588
2589
2590
2591
2592
2593
2594
2595
2596
2597
2598
2599
2600
2601
2602
2603
2604
2605
2606
2607
2608
2609
2610
2611
2612
2613
2614
2615
2616
2617
2618
2619
2620
2621
2622
2623
2624
2625
2626
2627
2628
2629
2630
2631
2632
2633
2634
2635
2636
2637
2638
2639
2640
2641
2642
2643
2644
2645
2646
2647
2648
2649
2650
2651
2652
2653
2654
2655
2656
2657
2658
2659
2660
2661
2662
2663
2664
2665
2666
2667
2668
2669
2670
2671
2672
267
```

1. Which gender has the most heart disease?

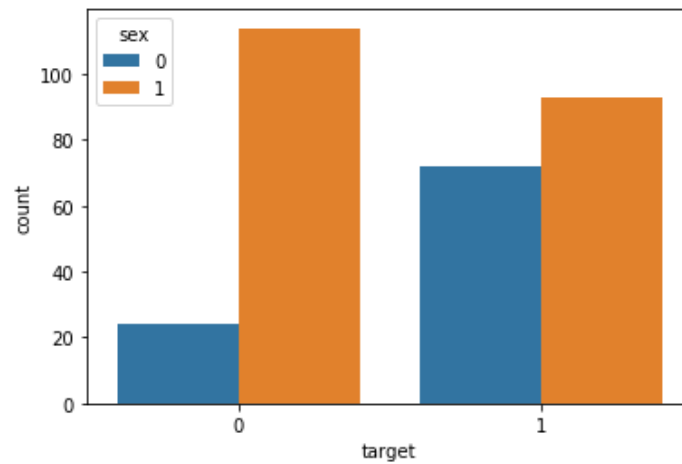


Figure 9 – Data of sex vs Target

(0 is female, non-disease; 1 is male, disease)

We observe that men are potential subjects of heart disease compared to women. To understand the root cause, we will deep dive in the extension part (Dataset Two).

2. What is the age range affecting the most by heart disease?

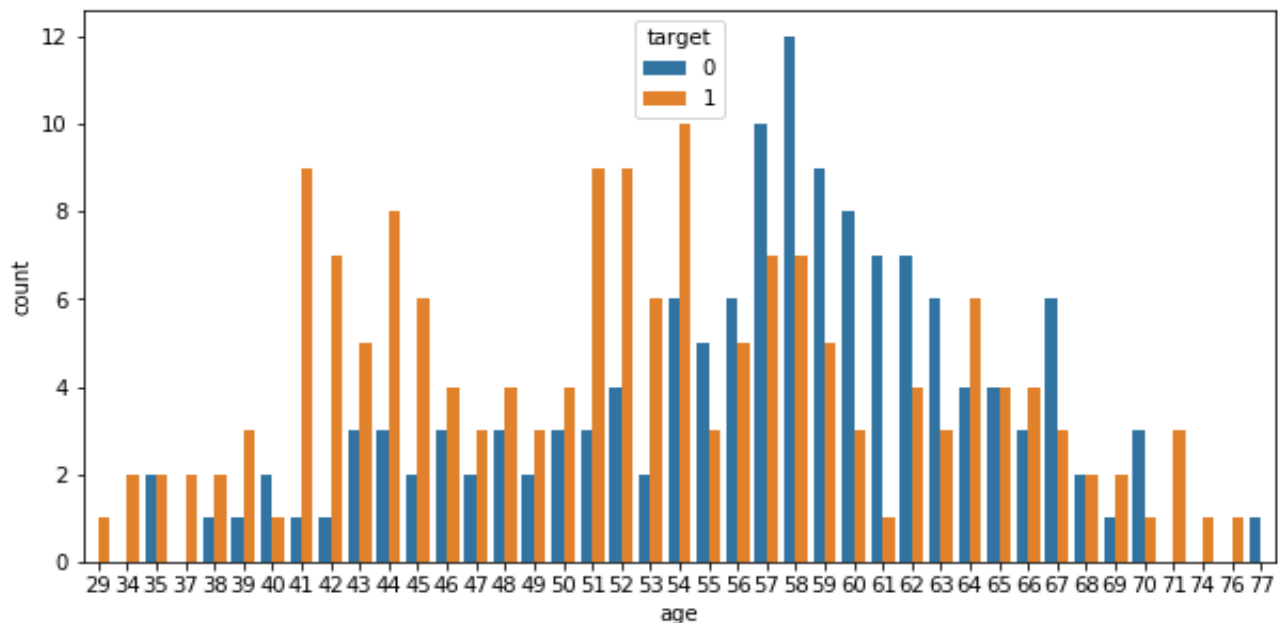


Figure 10 – Data of Age vs Target

('Target ' 0 non-disease; 1 is disease)

The figure showed that all potential heart issues are discovered on people whose age range lies between 41 to 54.

3. Which are the factors that lead to heart disease?

Relation between old peak and heart disease

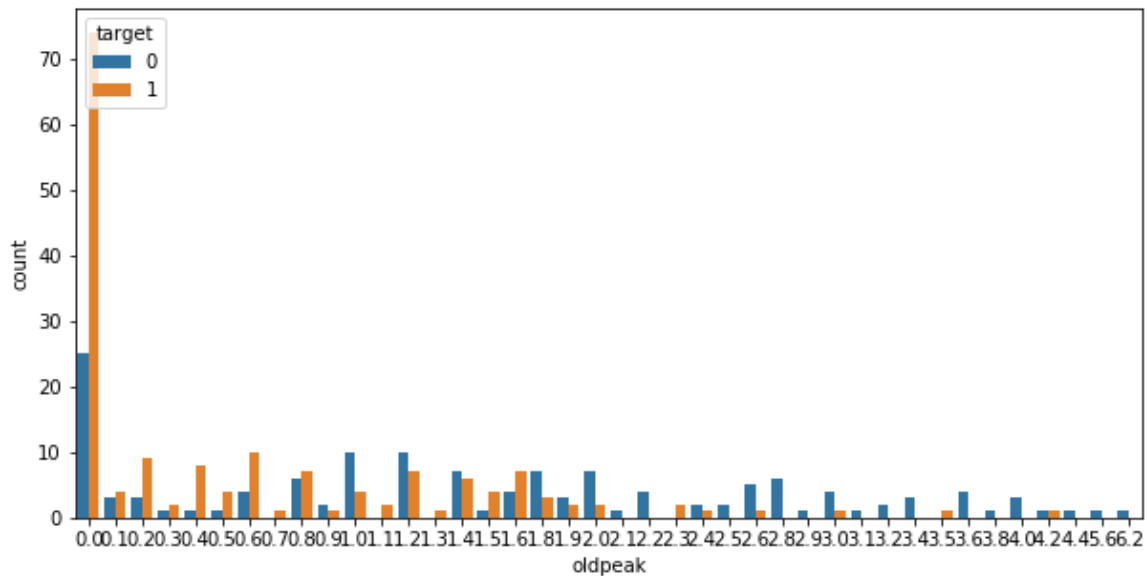


Figure 11 – Data of Oldpeak vs Target

(Target: 0 non-disease; 1 is disease)

The graph shows that the risk of having a heart disease , A subject with an oldpeak less the 0.12 is potential subject to a heart disease , the most subject to a heart issue at 100% is when the oldpeak is equal to zero .

4. Is there a relation between chest pain and heart disease?

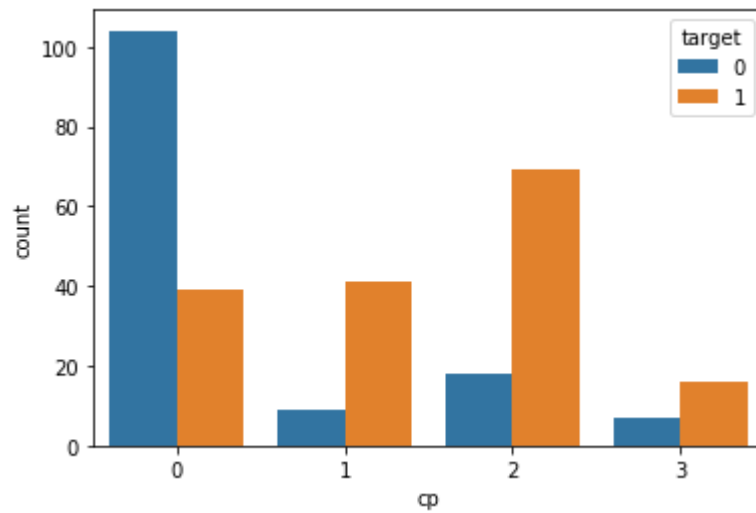


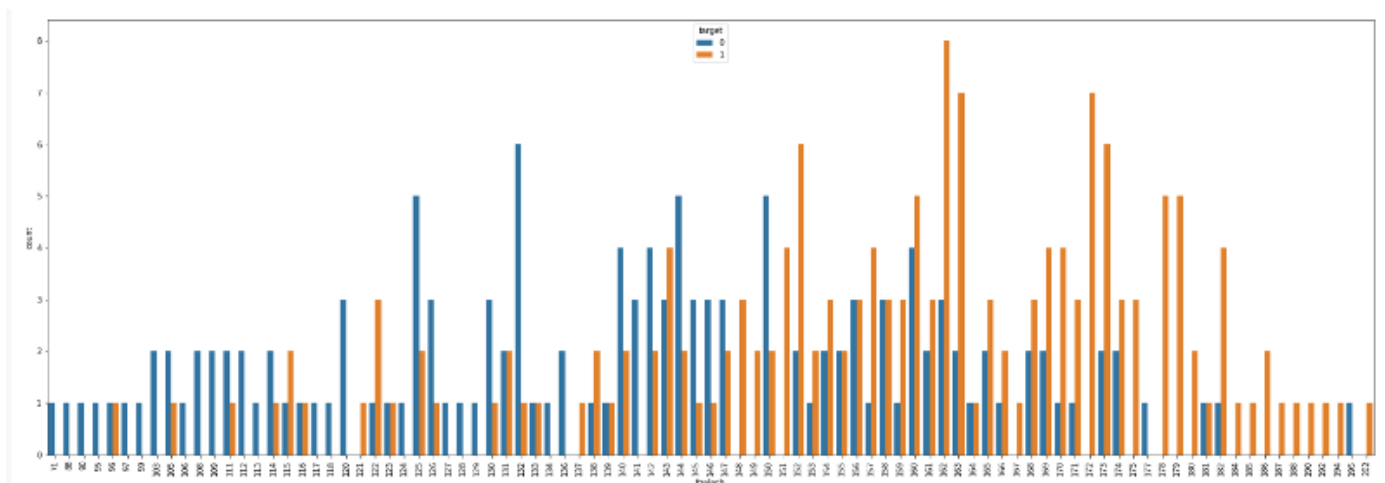
Figure 12 – Data of Cp vs Target

(Target: 0 non-disease; 1 is disease)

We come across a potential relation between chest pain and heart issue, we can state that the fact of having a chest pain level greater than 0 , is a factor of heart dysfonctionnement

6. Relation between chest pain and heart disease?

Figure 13 – Data of Thalach vs Target (Target: 0 non-disease; 1 is disease)



We observe that if the maximum heart rate achieved by a patient lies from 140 to 186 , then there is a potential chance that that patient has a heart issue .

7. Relation between heart issue and resting blood pressure?

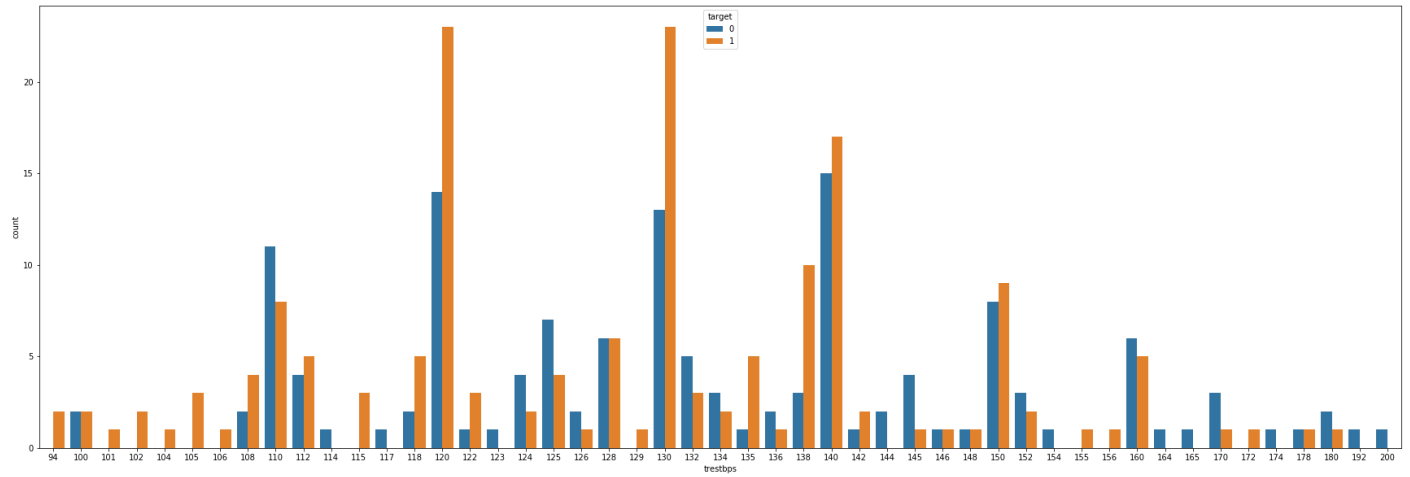


Figure 14 – Data of Trestbps vs Target

(Target: 0 non-disease; 1 is disease)

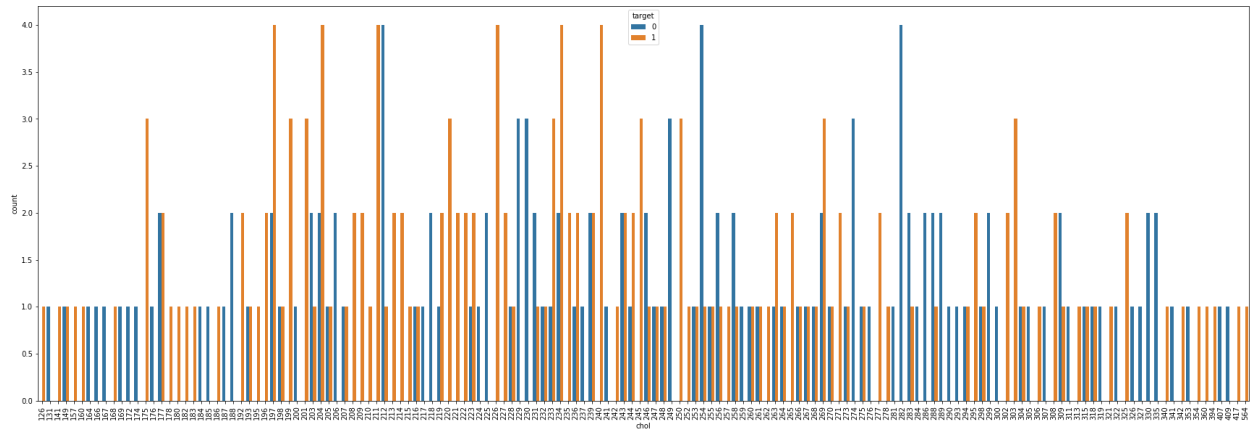


Figure 15 – Data of Chol vs Target

(Target: 0 non-disease; 1 is disease)

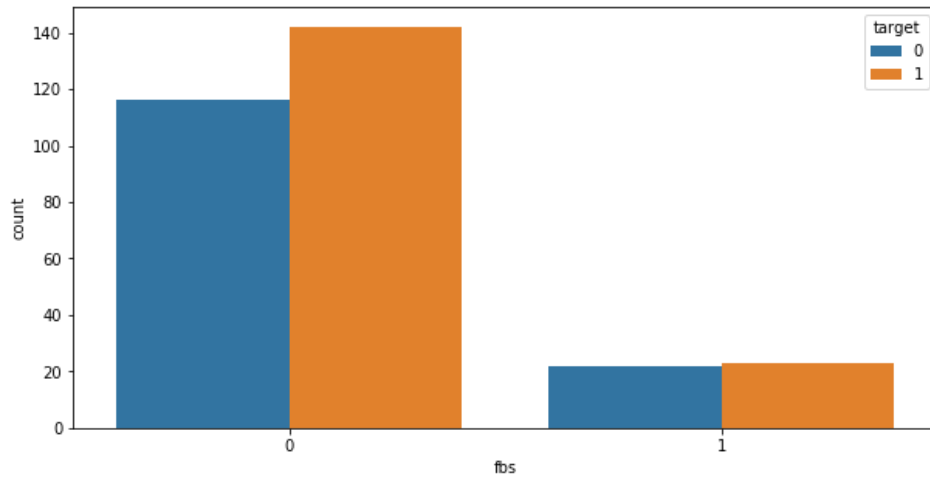


Figure 16 – Data of Fbs vs Target
(Target: 0 non-disease; 1 is disease)

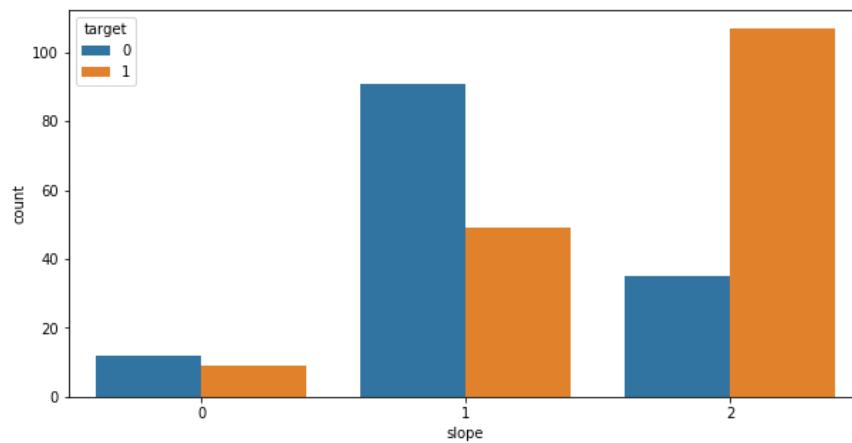


Figure 17 – Data of Slope vs Target
(Target: 0 non-disease; 1 is disease)

From the Figure, we can see that in case the Slope is 1, the chance of heart disease is 35%; in case the Slope is 2, the chance of heart disease is 80%.

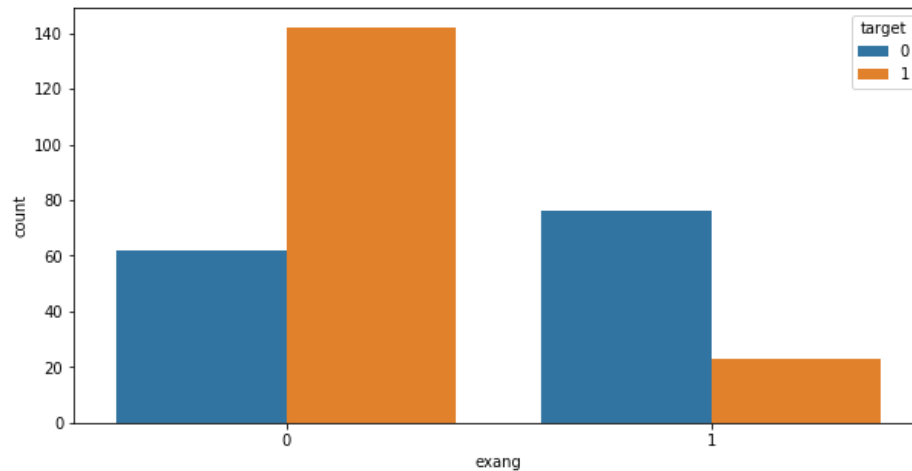


Figure 18 – Data of Exang vs Target
(Target: 0 non-disease; 1 is disease)

From the Figure, we can see that in case the Exang is 0, the chance of heart disease happened more than 80%.

Conclusion

From the multiple figures above, we can see that Oldpeak (at 0.0 to 0.3), CP (type 0, 1, 2), Thalach (from 140 to 186), Trestbps (110 to 150), Slope (1, 2), Exang (0) and Sex (Male) have the tendency to lead heart disease.

Which Factors have the correlation to each other's?

Correlation between variables



Figure 19 – Correlation Matrix of Dataset

To validate the learning above, we now run the correlation matrix chart to see the relation between the factors. From the figure, we can clearly see these are strong correlation between:

- ✓ Cp and Exang
- ✓ Slope and Oldpeak
- ✓ Target and Ca
- ✓ Target and Thai
- ✓ Target and Exang
- ✓ Target and Oldpeak
- ✓ Thalach and age

Plot of Correlation between variables and Target

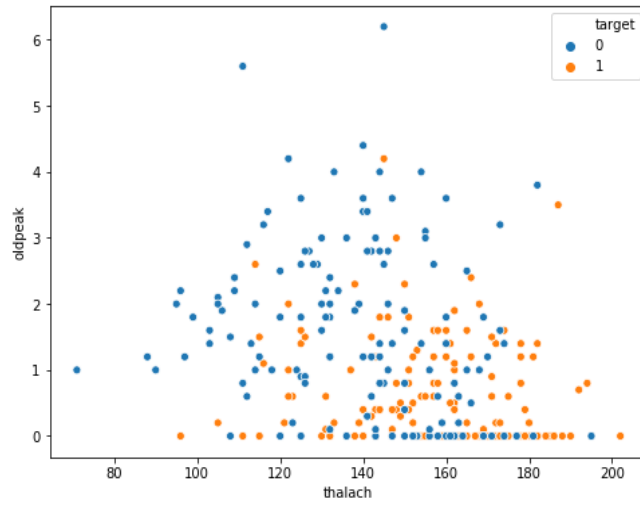


Figure 20 – Scatterplot (Thalach, Oldpeak and Target)

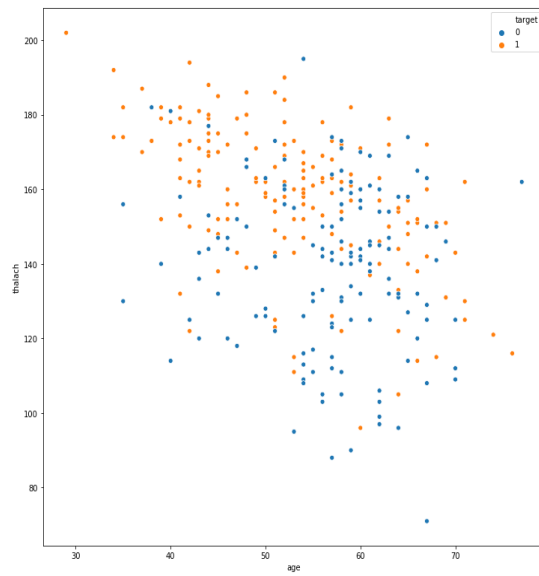


Figure 21 – Scatterplot (Thalach, Age and Target)

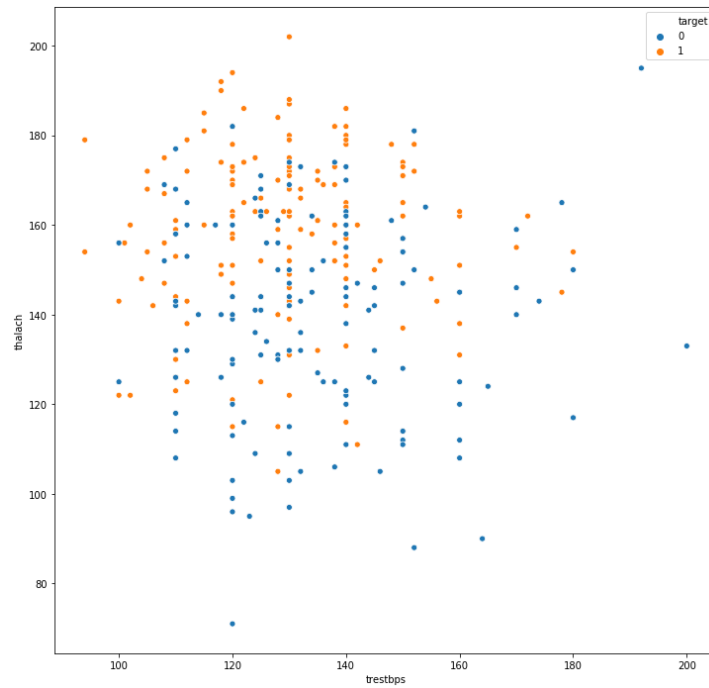


Figure 22 – Scatterplot (Thalach, Trestbps and Target)

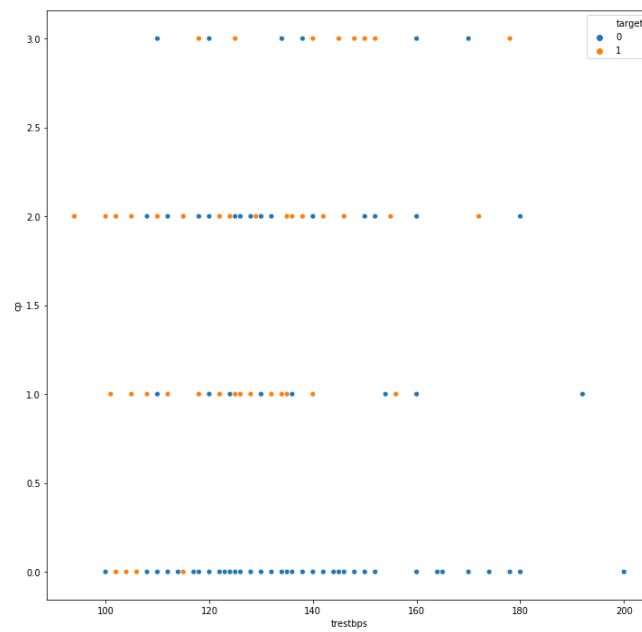


Figure 23 – Scatterplot (Cp, Trestbps and Target)

Prediction Model 1 - The Algorithm to checking Heart Disease

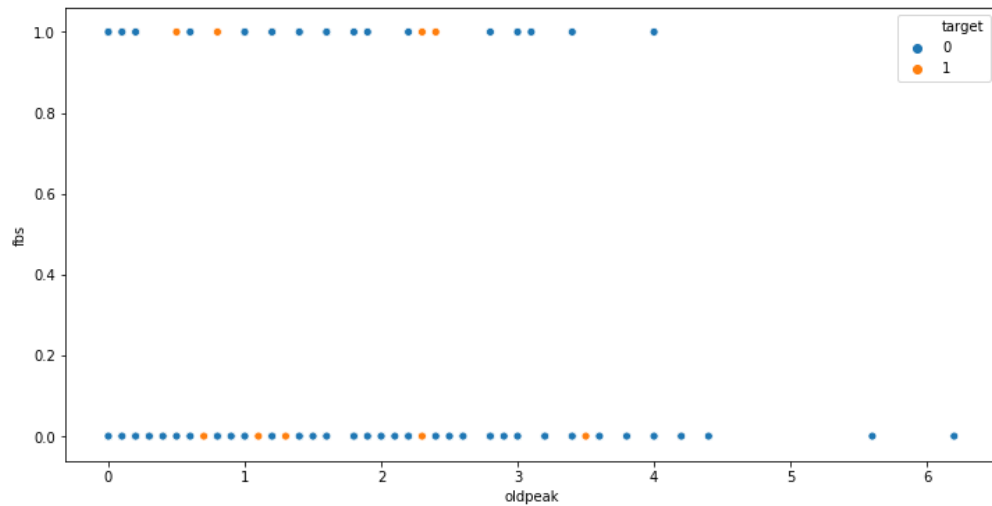


Figure 24

As we can see from the above fig , Our model looks like a Logistic Regression where we have to classify people having heart disease and people not having a heart issue base on 'sex', 'age', 'oldpeak', 'fbs' as input data and predict whether they might be a subject of a heart disease (Dataset one) .

Why Logistic Regression Classification?

A logistic Regression is a type of regression based on the mathematics equation $1/1+e^{-x}$, this regression as a basic shape of S which fits perfectly our model.

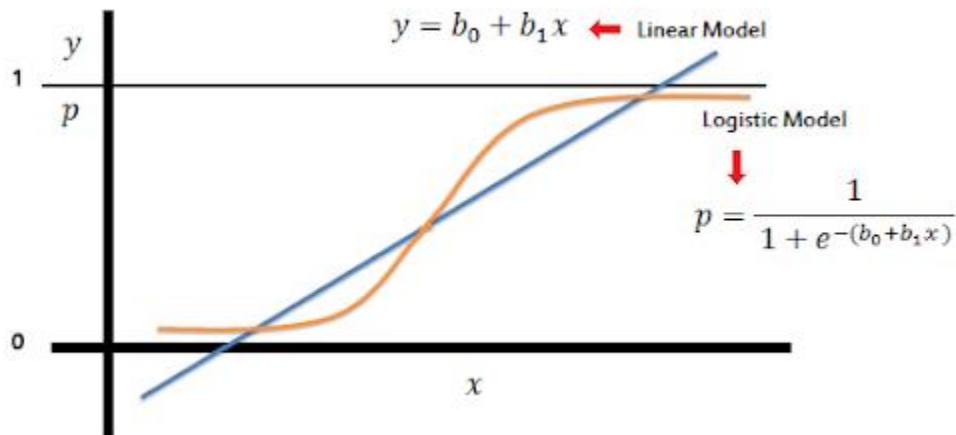


Figure 25 – Logistic Model

Our model can be viewed as multiple regression model , where many features are independent variable and only one dependent variable feature , ($y = a + bx + cx^2 + \dots +$).

Different Logistic Regression Tested for Our model

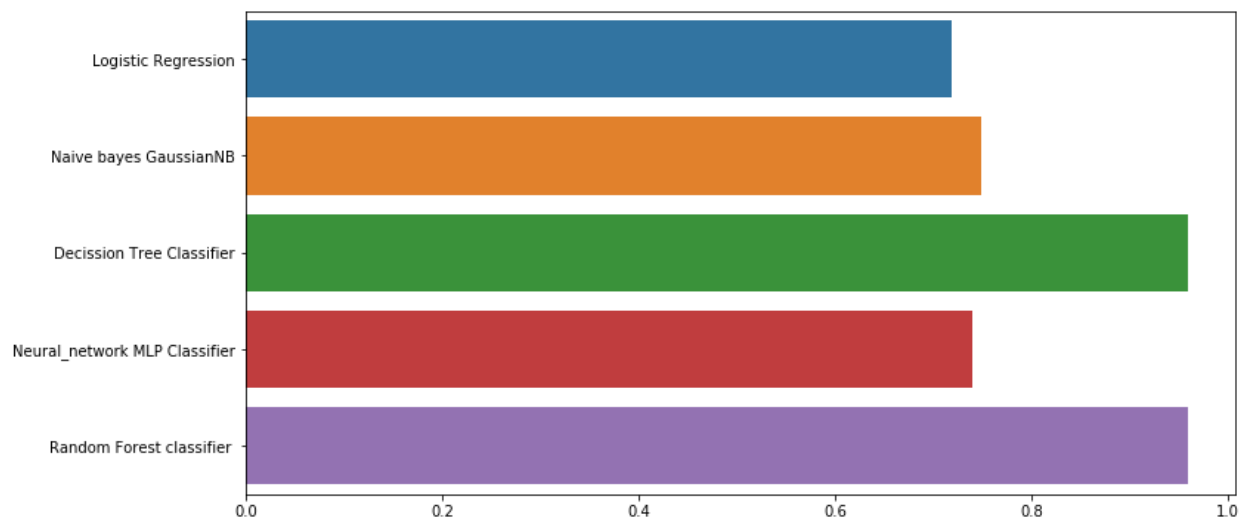


Figure 26 - Plot of different Logistic classifier result

The most accurate algorithm for our model is either Decision Tree classifier or random forest classifier which give us a prediction almost close to 80% percent . from our testing set data .

The extension of the project

As mentioned above, to further understand the root causes that impact the heart disease we have investigated in external data set of cardio_train.csv from the website Kaggle.com to see other related factors to heart disease.

The external dataset analysis

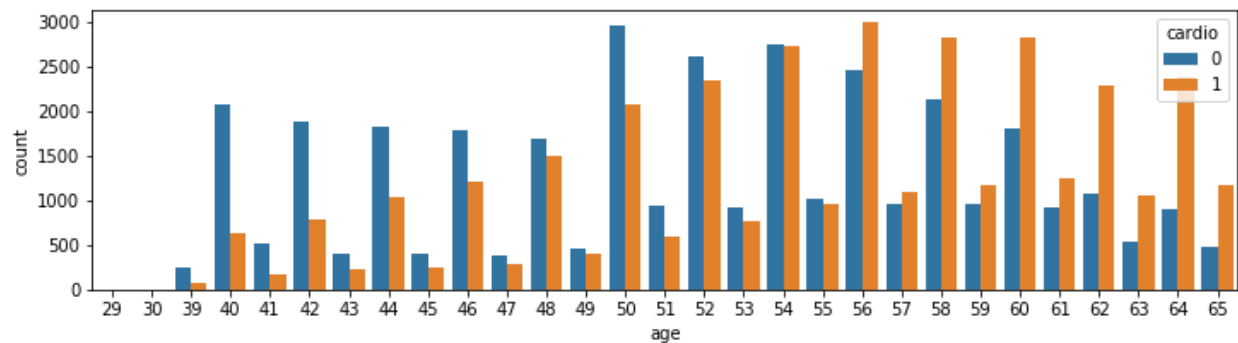


Figure 27 – Age vs Cardio

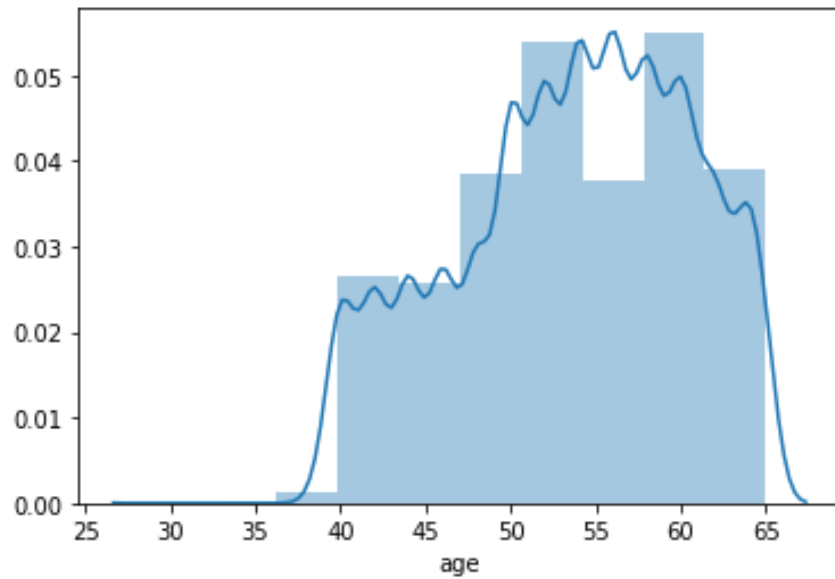


Figure 28 – Distribution of Age

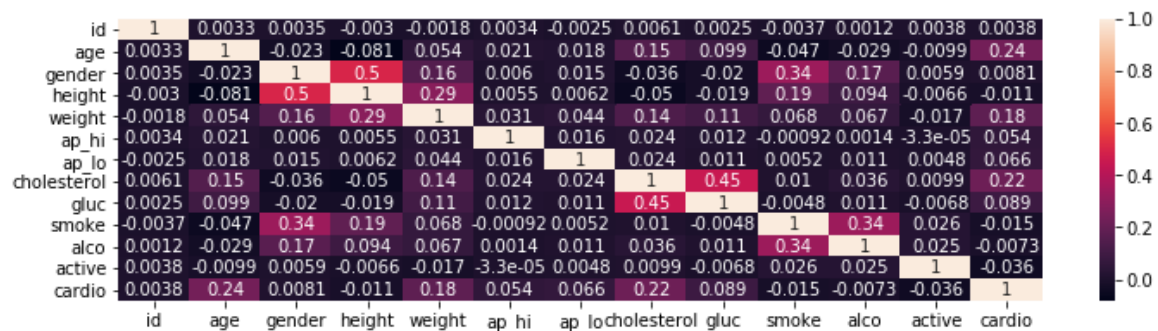


Figure 29 – Correlation Matrix

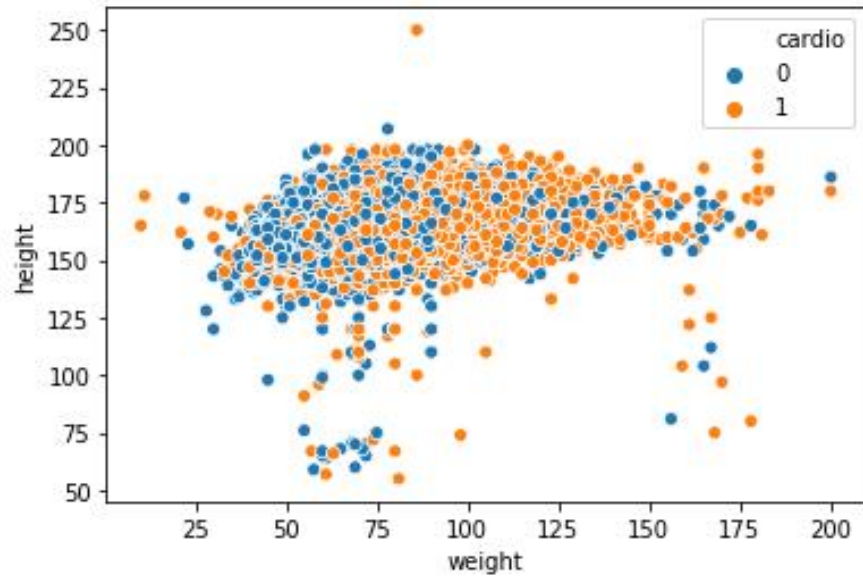


Figure 30 – Correlation Height, Weight and Cardio

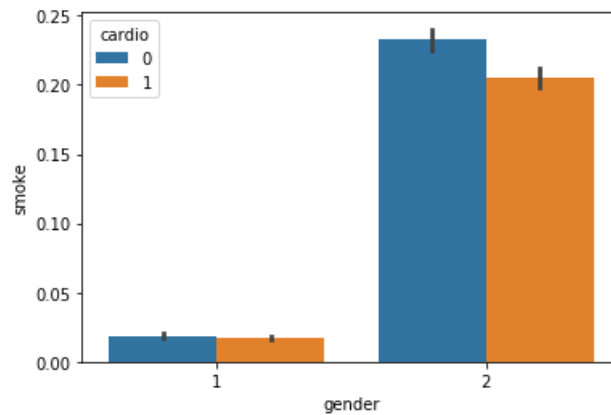
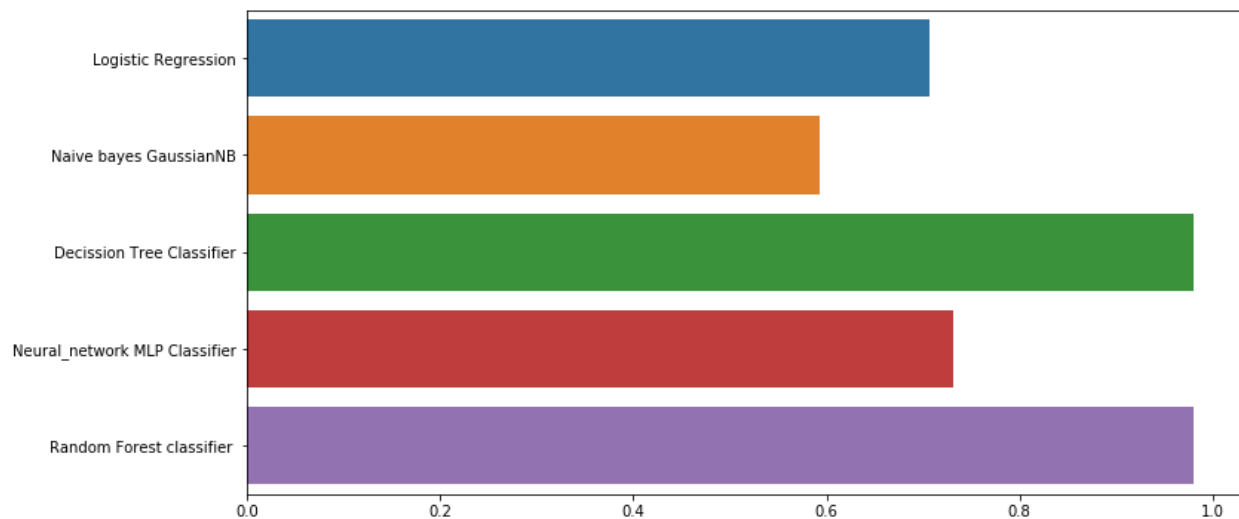


Figure 31 – Correlation Smoke, Gender and Cardio

From the above study (Dataset One), we can see that male had the tendency to have heart disease more than females. In this figure of Dataset Two, smoke shows that men smoke much more women. Hence, smoking can be a factor that relates to the heart disease gender differentiation.

Prediction Model 2 - The Algorithm to checking Heart Disease

How second data sets also classify , people base their activity to predict whether they can be a subject for a heart disease , those factors are based on 'ap_lo', 'ap_hi', 'smoke', 'alco', 'active', 'age', 'gender', 'cholesterol', 'gluc', 'height', 'weight'.



After testing different algorithm Decision Tree and Random Forest fit our model about 70 percent accuracy

Conclusion

Through the project of Heart Diagnosis, we have applied the Logistic Regression Classification with 5 different types of Algorithm including: Logistic Regression, Naïve bayes GaussianNB, Decision Tree Classifier, Neural_network NLP Classifier, Random Forest classifier. From that we can have the prediction model to help diagnose heart disease with the accuracy 80% for Dataset One and 70% for the Dataset Two. The learning from the project helps us to practice from the Data Analysis, Deep dive understanding, Making the hypothesis and validating the hypothesis by Optimizing data and Applying Logistic Regression Classification.

Acknowledgements

Creators of Dataset:

1. *Hungarian Institute of Cardiology. Budapest: Andras Janosi, M.D.*
2. *University Hospital, Zurich, Switzerland: William Steinbrunn, M.D.*
3. *University Hospital, Basel, Switzerland: Matthias Pfisterer, M.D.*
4. *V.A. Medical Center, Long Beach and Cleveland Clinic Foundation: Robert Detrano, M.D., Ph.D.*

Appendix

Pairplot

(sns.pairplot(mydata))

