گزارش تمرین دوم پیاده سازی داده کاوی

محبوبه شاکری 9531301

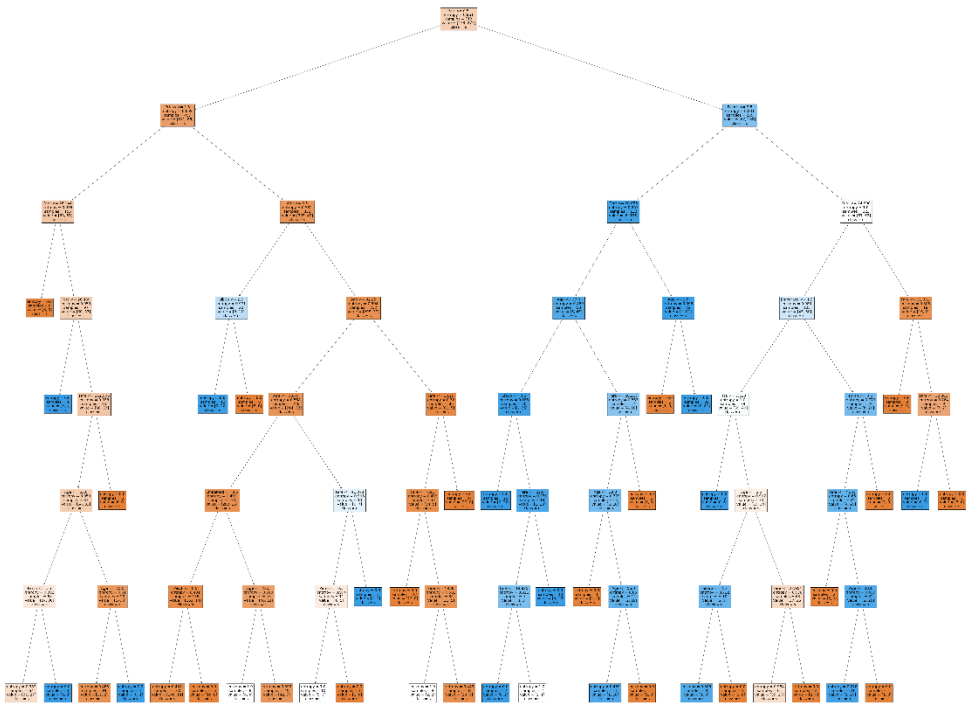
سوال 6)درخت تصمیم

در این بخش در ابتدا داده های غیر عددی مانند جنسیت به داده های عددی تبدیل شده اند و مقادیر null برای ویژگی fare با میانه و ویژگی سن با میانگین داده ها پر شده اند.

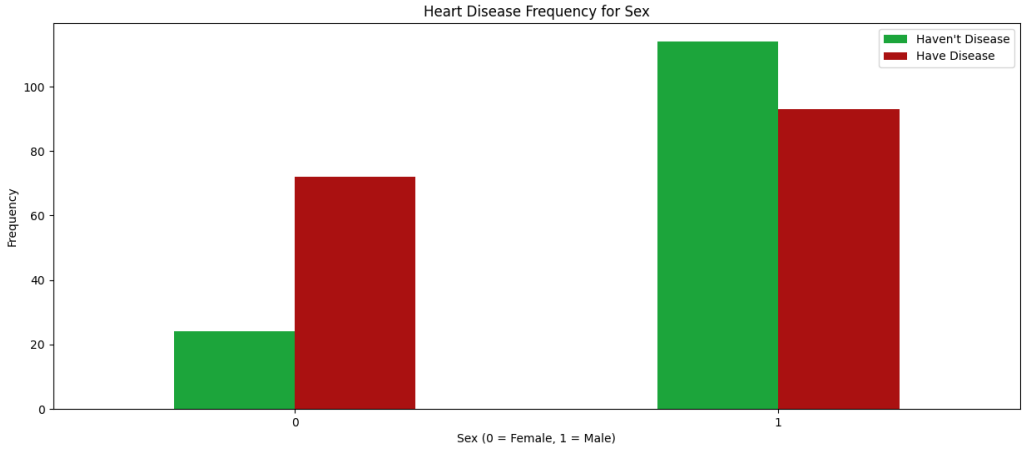
در ادامه بعد از تسیم دیتا به دو بخش تست و ترین و ترین درخت های تصمیم با 7 و5و3 ویژگی انتخابی و همچنین تابع های تقسیم مختلف 6 حالت، accuracy آن ها برای داده های تست محاسبه میشود.که نتیجه به دست امده برای تمام حالات بالای 80 درصد بوده است.

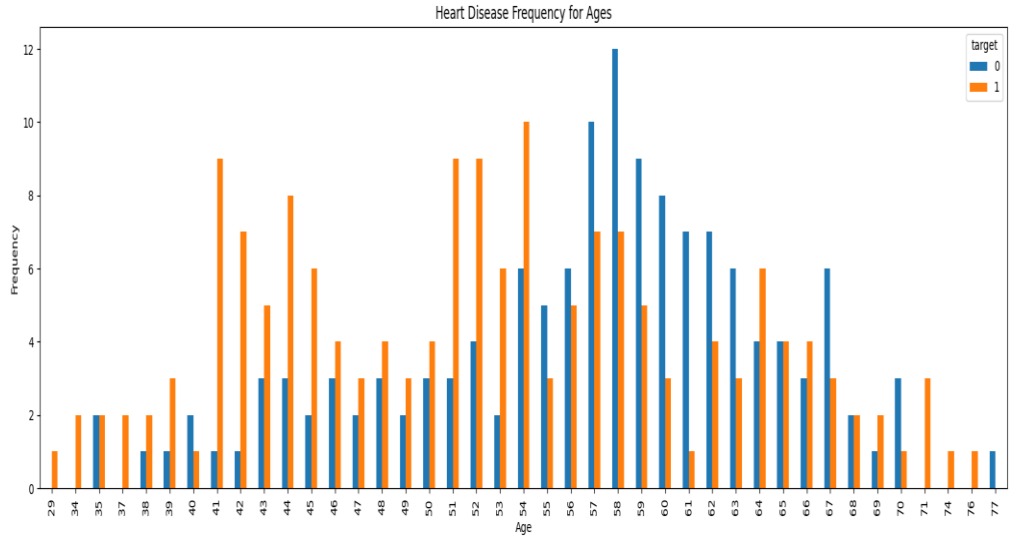
```
(env) F:\term9\dataminig\HW>dt.py
Accuracy for 7features with gini: 0.8100558659217877
Accuracy for 7features with entropy: 0.8324022346368715
Accuracy for 5features with gini: 0.8044692737430168
Accuracy for 5features with entropy: 0.8100558659217877
Accuracy for 3features with gini: 0.8044692737430168
Accuracy for 3features with entropy: 0.8044692737430168
```

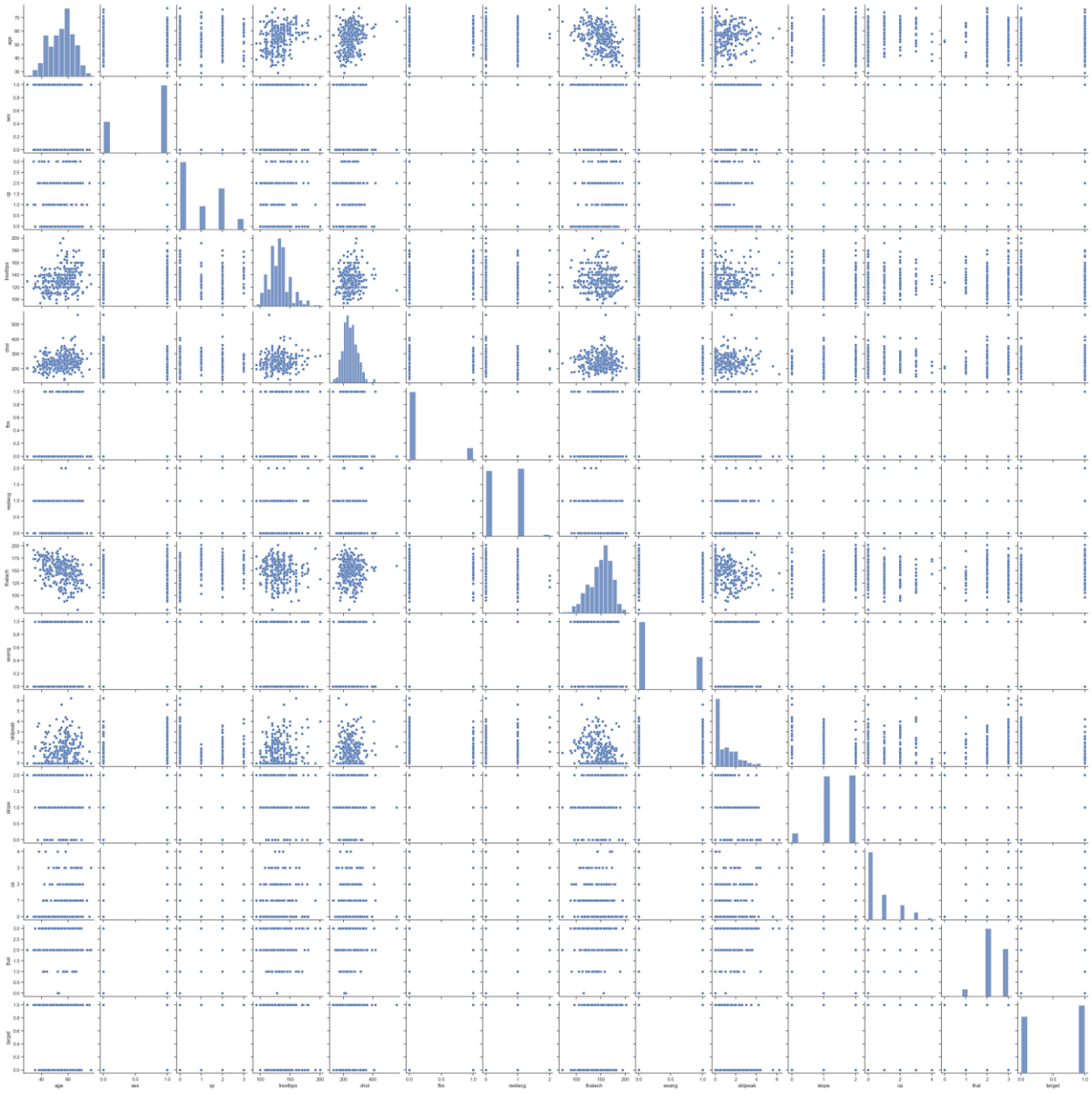
و درخت تصمیم با بیشترین accuracy به صورت گرافیکی نمایش داده شده است.



سوال 7) در بخش اول با بررسی نمودار فیچر های مختلف میتوان به این نتیجه رسید که همانطور که در نمودار بیماری قلبی بر اساس جنسیت، به طوری کلی درصد مراجعه مردان به پزشک و همچنین بیماری قلبی آن ها نسبت به زنان بیشتر بوده است. اما در بین مردان بیش از نیمی از مراجغه کننده گان بیماری قلبی داشته اند. ولی در زنان بین کسانی که به پزشک مراجعه کرده اند درصد کمی بیماری قلبی داشته اند و می توان نتیجه گرفت که جنسیت از اهمیت بالایی برخوردار است. هم چنین از نمودار بیماری قلبی بر اساس سن، روند مشخصی بین این دو فاکتور وجود ندارد و در تمام سنین مراجعه کنندگان داری بیماری قلبی بوده اند.

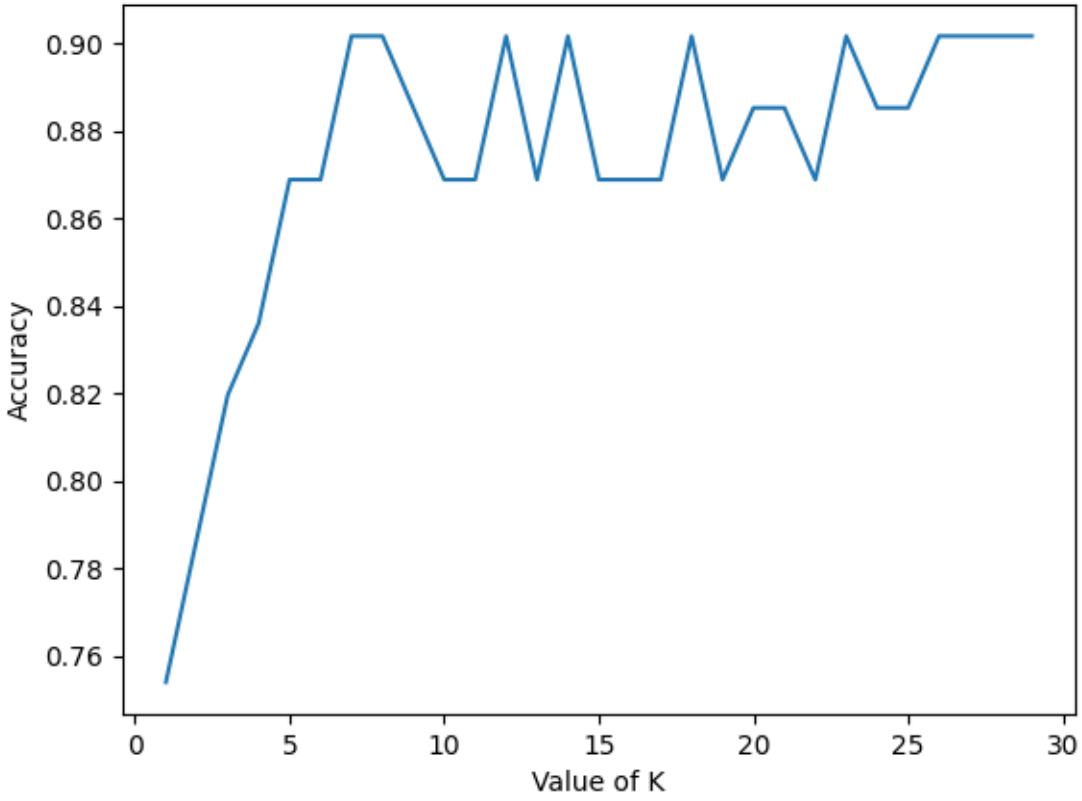

Heart Disease Frequency for Sex
Haven't Disease
Have Disease
Frequency
0
20
40
60
80
100
0
1
Sex (0 = Female, 1 = Male)


Heart Disease Frequency for Ages
target
0
1
Frequency
0
2
4
6
8
10
12
29
34
35
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
74
76
77
Age



در ادامه چون اسکیل داده ها در ویژگی های مختلف با هم متفاوت بود و در روش knn اسکیل داده ها مهم است، داده های تست و ترین که به نسبت 1 به 4 هستند را نرمال کرده ایم و همچنین از تمام ویژگی های موجود در دیتا بیس استفاده شده است.

سپس برای یافتن k مناسب برای داده ها نمودار accuracy نسبت به k در بازه 1 تا 30 را رسم کرده ایم. که با توجه به نمودار k=7 انتخاب شده است.



همچنین برای روش naive base با توجه به ماهیت این روش استفاده از داده های اسکیل شده و نشده تغیری در خروجی نمی گذارد. Accuracy به دست امده در هر دو روش برای داده های تست بالای 80 درصد است.

```
Accuracy KNN: 0.9016393442622951
Accuracy NB: 0.8688524590163934
```

- بخش های مربوط به نمودار ها در کد کامنت شده است.