ACTU PS5841 Data Science in Finance & Insurance - Autumn 2022 (Y. Wang)
Assignment - 4
Assigned 9/22/22, Due 10/1/22

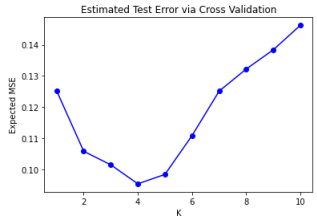
Part A -

Write a python script and use a **6-fold cross-validation** to identify the optimal number of neighbors (K) for the KNN regressor, with respect to data in the file `data_1d_reg.csv`, and then fit a KNN regression with the data.

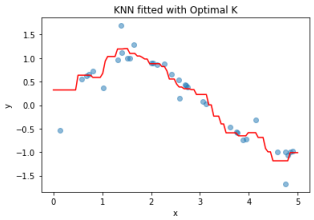
[a] For the optimal K, please code from "scratch" using `KFold()` and **NOT** some wrapper function such as `cross_val_score()`. For convenience of grading, please use the same `random_state` and enable `shuffle`:

```
KFold(n_splits = 6, random_state = 1, shuffle = True)
```

Please enable your script to produce a plot similar to the following.



[b] For the KNN regression, please enable your script to produce a plot similar to the following.

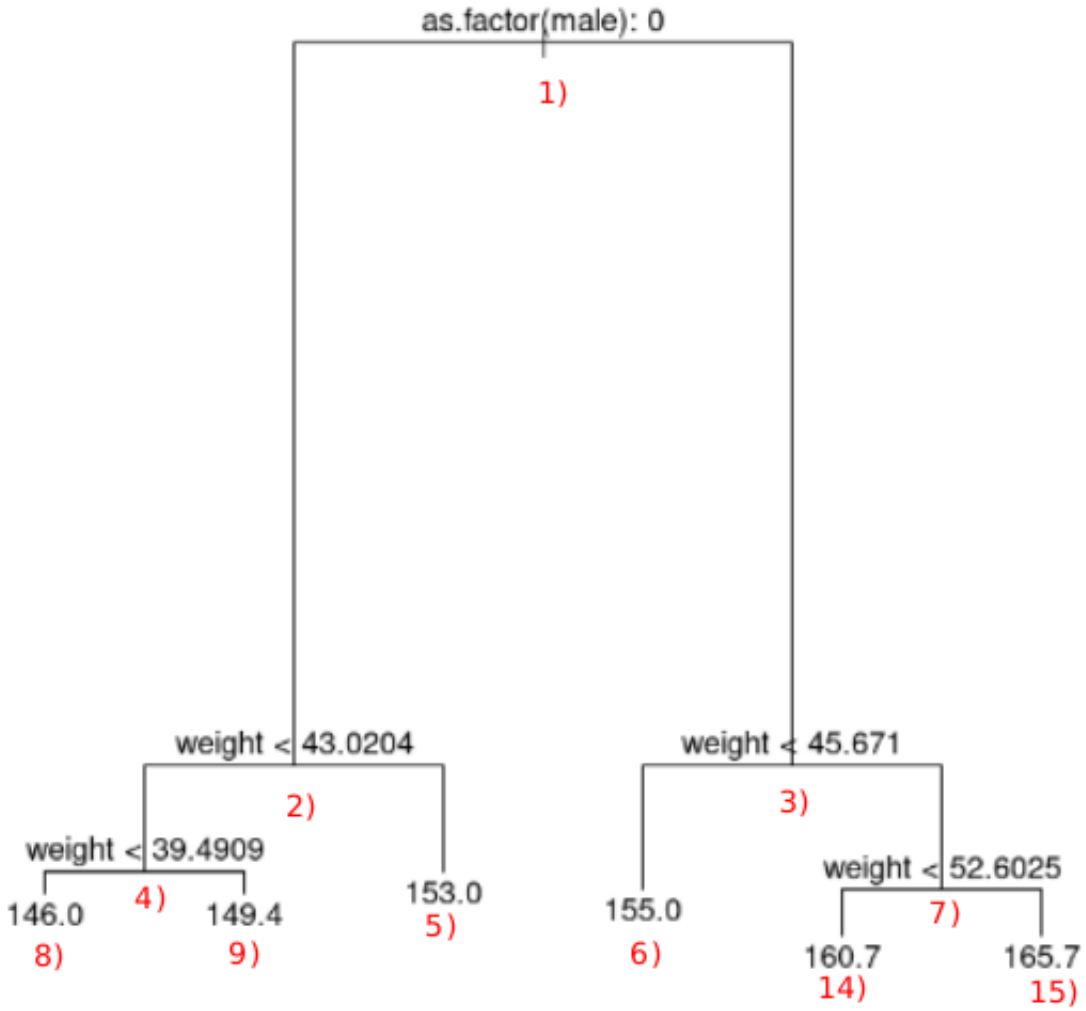


Please submit your work as hw4A.ipnb and hw4A.html to `Canvas`.

Part B -

The Decision Tree on page 2 is generated in R by the model, `tree(height ~ weight + as.factor(male))`, fitted with the Howell dataset on those aged 18-years or older (as discussed in class).
Write a python script to show that, based on the recursive binary splitting scheme, splitting node 3) by weight is more optimal than splitting it by sex.
For each split variable (weight or sex), please let your code to report
(1) the number of samples,
(2) the sum squared error, and
(3) the predicted height
for each of the resulting two nodes (i.e. node 6) and node 7)) as indicated.)

Please submit your work as hw4B.ipnb and hw4B.html to `Canvas`.


as.factor(male): 0
1)
weight < 43.0204
2)
weight < 45.671
3)
weight < 39.4909
4)
153.0
5)
155.0
6)
weight < 52.6025
7)
146.0
8)
149.4
9)
160.7
14)
165.7
15)