Kaggle Master-Week 1 Quiz

Toplam puan 100/100

Name-Last Name: *

Sevdanur GENÇ

E-mail address: *

sevdanurgenc@gmail.com

Q1- After training our decision tree model, we saw that the model is overfitted on the training data and it has bad performance on the test data. Which hyper-parameter could help us to get rid of this problem? Note: You can use sklearn.tree.DecisionTreeClassifier documentation.https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html#sklearn.tree.DecisionTreeClassifier * 10/10

- ○ criterion
- ◉ max_depth
- ○ random_state
- ○ splitter

Q2- Which of the below can be said definitely according to the results table taken from the data.describe() method? I. 75% of the values in the Rooms column are greater than 2. II. There are some houses with a land size of 0. III. There are missing values in the BuildingArea column. IV. There is no house with 9 rooms in the data set *

10/10

```
In [2]: import pandas as pd

        data = pd.read_csv("/home/fatih/Desktop/melb_data.csv")

        data.describe()
```

Out[2]:

	Rooms	Price	Distance	Postcode	Bedroom2	Bathroom	Car	Landsize	BuildingArea	YearBuilt
count	13580.000000	1.358000e+04	13580.000000	13580.000000	13580.000000	13580.000000	13518.000000	13580.000000	7130.000000	8205.000000
mean	2.937997	1.075684e+06	10.137776	3105.301915	2.914728	1.534242	1.610075	558.416127	151.967650	1964.684217
std	0.955748	6.393107e+05	5.868725	90.676964	0.965921	0.691712	0.962634	3990.669241	541.014538	37.273762
min	1.000000	8.500000e+04	0.000000	3000.000000	0.000000	0.000000	0.000000	0.000000	0.000000	1196.000000
25%	2.000000	6.500000e+05	6.100000	3044.000000	2.000000	1.000000	1.000000	177.000000	93.000000	1940.000000
50%	3.000000	9.030000e+05	9.200000	3084.000000	3.000000	1.000000	2.000000	440.000000	126.000000	1970.000000
75%	3.000000	1.330000e+06	13.000000	3148.000000	3.000000	2.000000	2.000000	651.000000	174.000000	1999.000000
max	10.000000	9.000000e+06	48.100000	3977.000000	20.000000	8.000000	10.000000	433014.000000	44515.000000	2018.000000

- ○ I, II
- ○ II, III
- ○ II, III, IV
- ◉ I, II, III

Q3- Which one is false about overfitting and underfitting? *

10/10

- ○ Insufficient training (less epoch less batch size), causes underfitting.
- ○ Training on too much epoch and batch size causes overfitting.
- ◉ Splitting dataset as train and test datasets will always be enough to prevent overfitting, no need for validation datasets.
- ○ In overfitting accuracy will be very good at train data but will be very bad at unseen data.

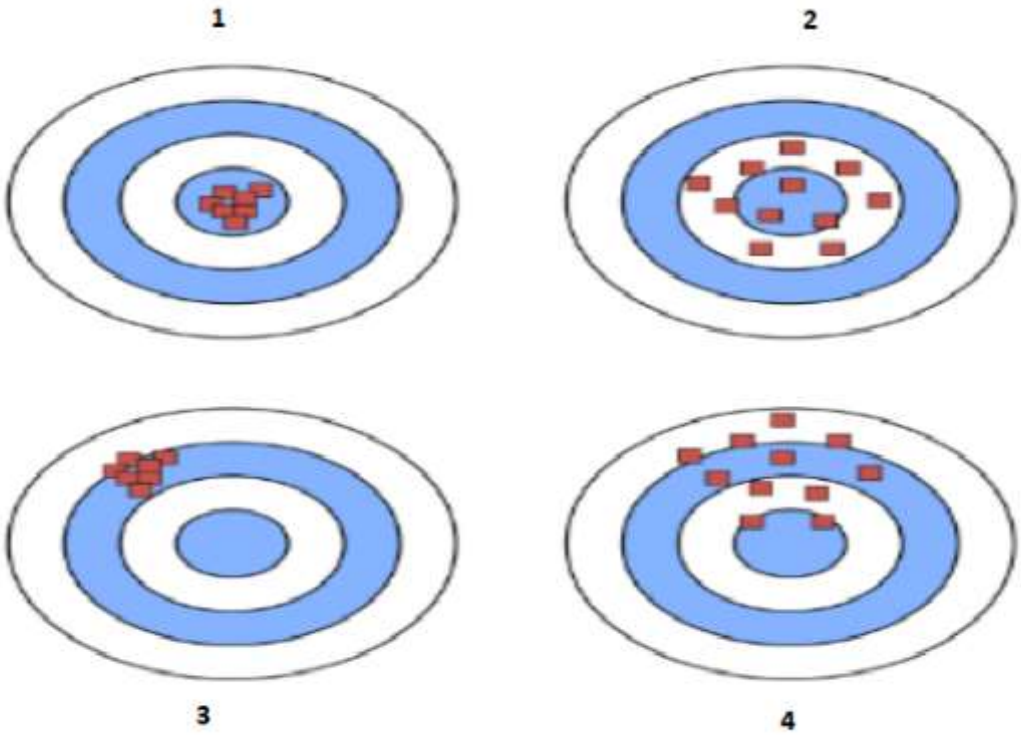
Q4- Which of the following is false regarding pandas and scikit-learn methods? *

10/10

- DataFrame.head(x) shows x samples in the DataFrame from the beginning.
- DataFrame.describe() shows summary of the data.
- model.predict() determines how accurate the model's predictions are.
- DataFrame.dropna(axis=0) drops missing values.

Q5- According to the shooting clusters scheme above, for each figure which statements are true? Notice that, shooting targets are the centers. *

10/10



- (●) 1:Low Bias- Low Variance 2:Low Bias-High Variance 3:High Bias-Low Variance 4: High Bias-High Variance
- ( ) 1:Low Bias- High Variance 2:Low Bias-Low Variance 3:High Bias-High Variance 4: High Bias-Low Variance
- ( ) 1:High Bias- Low Variance 2: High Bias-High Variance 3:Low Bias-Low Variance 4:Low Bias-High Variance
- ( ) 1:High Bias- High Variance 2:High Bias-Low Variance 3:Low Bias-High Variance 4:Low Bias-Low Variance

Q6- According to the random forests algorithm, which of the below statements are true? * 10/10

I - It is an algorithm that aims to increase the classification value by producing multiple decision trees.

II - It was created by combining Bagging and Random Subspace methods.

III - While creating the tree, it is made performance evaluation with 2/3 of the data set.

- ○ I, III
- ○ II, III
- ◉ I, II
- ○ I, II, III

Q7- What do you think about train_X when line 1 and line 2 are executed separately? The rest of the code is exactly the same. * 10/10

```
Line 1. train_X, val_X, train_y, val_y = train_test_split(X, y, random_state = 2,shuffle=False)
Line 2. train_X, val_X, train_y, val_y = train_test_split(X, y, random_state = 1,shuffle=False)
```

- ○ They generate different random number so the train_X differs from each other.
- ○ They generate different same number and the train_X is equal to each other.
- ○ They generate different random number so the train_X is equal to each other.
- ◉ They generate different random number ,but the train_X is equal to each other.

Q8- Trees have their length and we call that the depth of the tree. RandomForestRegressor, in scikit-learn library, has a maximum leaf (max_depth) parameter which is None as default which means nodes are expanded until all leaves are pure. What can be said if we change the number of maximum leaf nodes of a random forest? *

10/10

- ○ Length of a tree does not affect any of the results.
- ◉ Model may overfit for large depth values.
- ○ The longer tree is the better tree.
- ○ Short trees more precise than long trees.

Q9- Let assume, we have a data set called home_data with 3 features names; LotArea, YearBuilt, PoolArea. How do you define non-missing values for the feature LotArea? *

10/10

- ○ non_missings = home_data["LotArea"].mean()
- ○ non_missings = home_data.count()
- ◉ non_missings = home_data["LotArea"].count()
- ○ non_missings = home_data.mean()

Q10- What is the aim of the below code pieces? *

10/10

```
from sklearn.metrics import mean_absolute_error

predicted_home_prices = melbourne_model.predict(X)
mean_absolute_error(y, predicted_home_prices)
```

- ( ) For splitting the data as test and train
- ( ) For interpreting the data description
- (•) For summarizing model quality
- ( ) For data modelling