Kaggle Master-1-Final Test

Total points 100/100

Email address *

sevdanurgenc@gmail.com

Name-Last Name: *

Sevdanur GENÇ

✓ Q1- Assume you have a dataset named museum_data and its index is Date column. When you run "museum_data.head()" statement you get the following://Image in Cell D4Which code below achieves the following requirement?"In October 2018, how many more visitors did Avila Adobe receive than the Firehouse Museum?"Note: All dates are in the format of: YYYY-MM-01 (Day is always 01 for every year and month) * 4/4

Date	Avila Adobe	Firehouse Museum	Chinese American Museum	America Tropical Interpretive Center
2014-01-01	24778	4486	1581	6602
2014-02-01	18976	4172	1785	5029
2014-03-01	25231	7082	3229	8129
2014-04-01	26989	6755	2129	2924
2014-05-01	36883	10858	3676	10894

- (●) museum_data[museum_data.index.isin(['2018-10-01'])]['Avila Adobe'].sum() - museum_data[museum_data.index.isin(['2018-10-01'])]['Firehouse Museum'].sum() ✓
- ( ) museum_data[museum_data.index.isin(['2018-10-01'])]['Firehouse Museum'].sum() - museum_data[museum_data.index.isin(['2018-10-01'])]['Avila Adobe'].sum()
- ( ) museum_data[museum_data.index.isin(['2018-10-01'])]['Firehouse Museum'].sum() -

museum_data[museum_data.index.isin(['2018-10-01'])]['Chinese American Museum'].sum()

museum_data[museum_data.index.isin(['2019-10-01'])]['Firehouse Museum'].sum() - museum_data[museum_data.index.isin(['2020-10-01'])]['Avila Adobe'].sum()

✓ Q2- What is the aim of the below code pieces? * 4/4

```
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 data modelling
- For interpreting the data description
- (●) For summarizing model quality ✓

✓ Q3- Which one is false about overfitting and underfitting? * 4/4

- 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. ✓
- Insufficient training (less epoch less batch size), causes underfitting.
- In overfitting accuracy will be very good at train data but will be very bad at unseen data.

✓ Q4- What do the highlighted code pieces mean? * 4/4

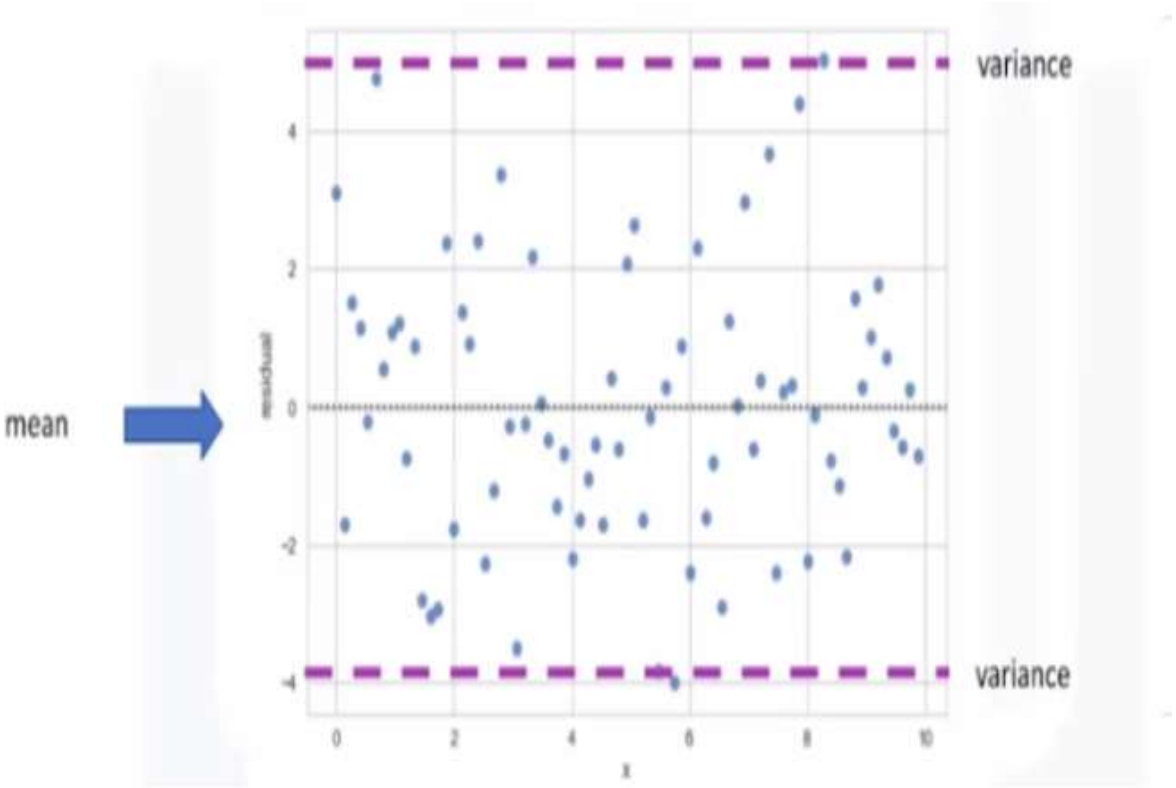
```
X_train_plus = X_train.copy()
X_valid_plus = X_valid.copy()
for col in cols_with_missing:
    X_train_plus[col + '_was_missing'] = X_train_plus[col].isnull()
    X_valid_plus[col + '_was_missing'] = X_valid_plus[col].isnull()
my_imputer = SimpleImputer()
imputed_X_train_plus = pd.DataFrame(my_imputer.fit_transform(X_train_plus))
imputed_X_valid_plus = pd.DataFrame(my_imputer.transform(X_valid_plus))
imputed_X_train_plus.columns = X_train_plus.columns
imputed_X_valid_plus.columns = X_valid_plus.columns
```

- ○ To make copy to avoid changing original data
- ○ For imputation
- ◉ To put removed column names back ✓
- ○ To make new columns indicating what will be imputed

✓ Q5- Which of the following is not an approach that we can use to prepare our categorical data? * 4/4

- ○ One-Hot Encoding
- ◉ Normalization ✓
- ○ Label Encoding
- ○ Drop categorical values

✓ Q6- According to the residual plot below, which statement is false? * 4/4



- Data points spread around randomly.
- The plot is created using seaborn.
- It suggest that the linear model would be appropriate.
- (●) Data points distributed on a curvature. ✓

✓ Q7- Which of the below statement is false? * 4/4

- sns.heatmap - Used to find color-coded patterns in tables of numbers
- sns.distplot - Show the distribution of a single numerical variable

sns.swarmplot - Useful for comparing quantities corresponding to different groups

sns.lmplot - Useful for drawing multiple regression lines, if the scatter plot contains multiple, color-coded groups.

✓ Q8- Which of the following statements are true about “max_depth” hyperparameter in Random Forest? * 4/4

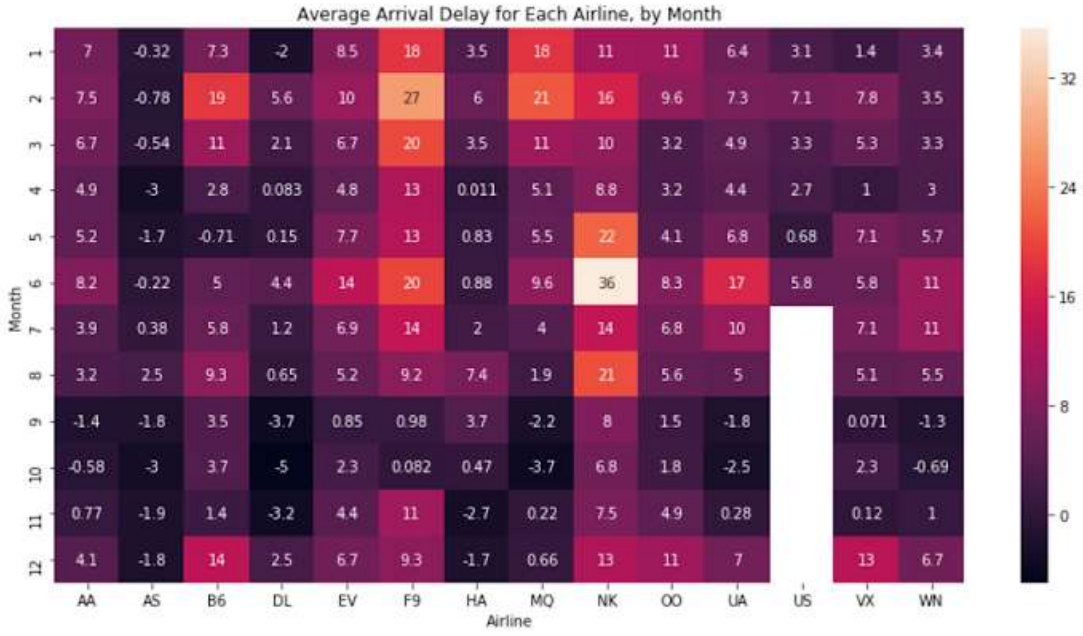
```
I-      Lower is better parameter in case of same validation accuracy
II-     Higher is better parameter in case of same validation accuracy
III-    Increase the value of max_depth may overfit the data
IV-     Increase the value of max_depth may underfit the data
```

- II, III
- I, III ✓
- I, IV
- II, IV

✓ 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? * 4/4

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

✓ Q10- Which one is a true statement about the below visual? * 4/4



- AS Airline has the most delayed flights.
- This shows a bar chart.
- The light boxes are the desired relations if we want to define the least delayed flights.
- (●) The months 9-11 are the best schedules in that year. ✓

✓ Q11- Which plot type does Scatterplot fall into? * 4/4

- Categorical
- Distribution
- Regression
- Relational ✓

Relational ✓

✓ Q12-You will build a model to predict housing prices. The model will be deployed on an ongoing basis, to predict the price of a new house when a description is added to a website. Here are four features that could be used as predictors. Which of the features is most likely to be a source of leakage? * 4/4

- ○ Whether the house has a basement
- ○ Latitude and longitude of the house
- ◉ Average sales price of homes in the same neighborhood ✓
- ○ Size of the house (in square meters)

✓ Q13- How is the Gradient Boosting cycle proceed? Please choose the correct order from the mixed statements below. * 4/4

I- We add the new model to ensemble.
II- We use the current ensemble to generate predictions for each observation in the dataset.
III- We use the loss function to fit a new model that will be added to the ensemble.

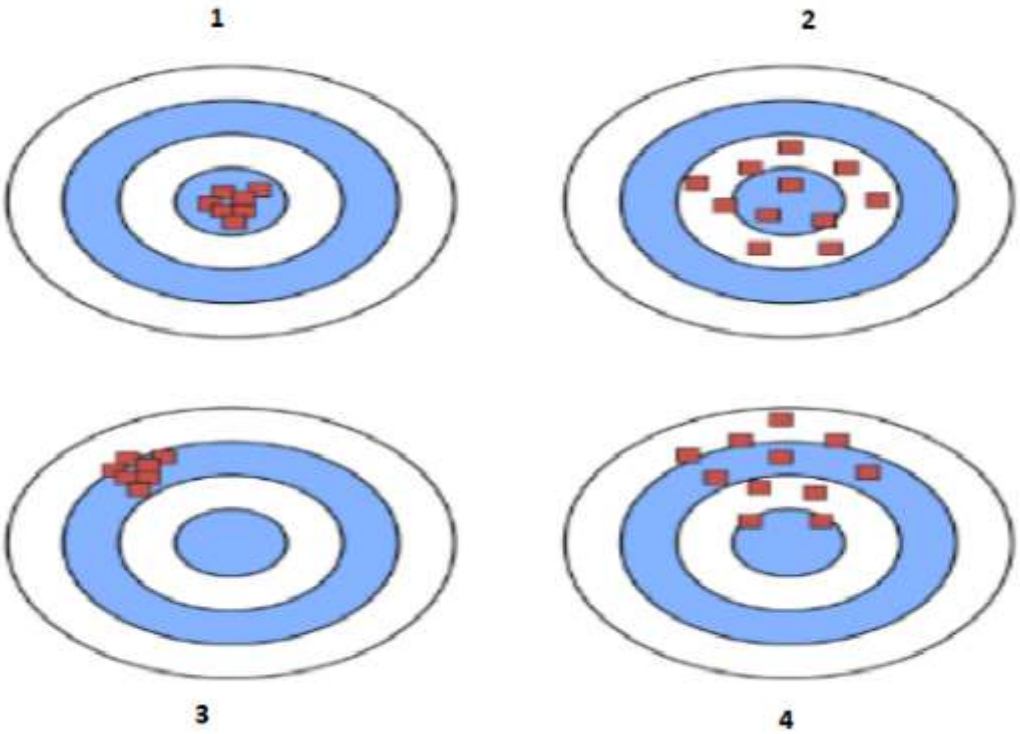
- ○ II-I-III
- ◉ II-III-I ✓
- ○ I-II-III
- ○ I-III-II

✓ Q14- 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. * 4/4

```
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 ,but 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 so the train_X differs from each other.
- ( ) They generate different same number and the train_X is equal to each other.

✓ Q15- According to the shooting clusters scheme above, for each figure which statements are true? Notice that, shooting targets are the centers. * 4/4

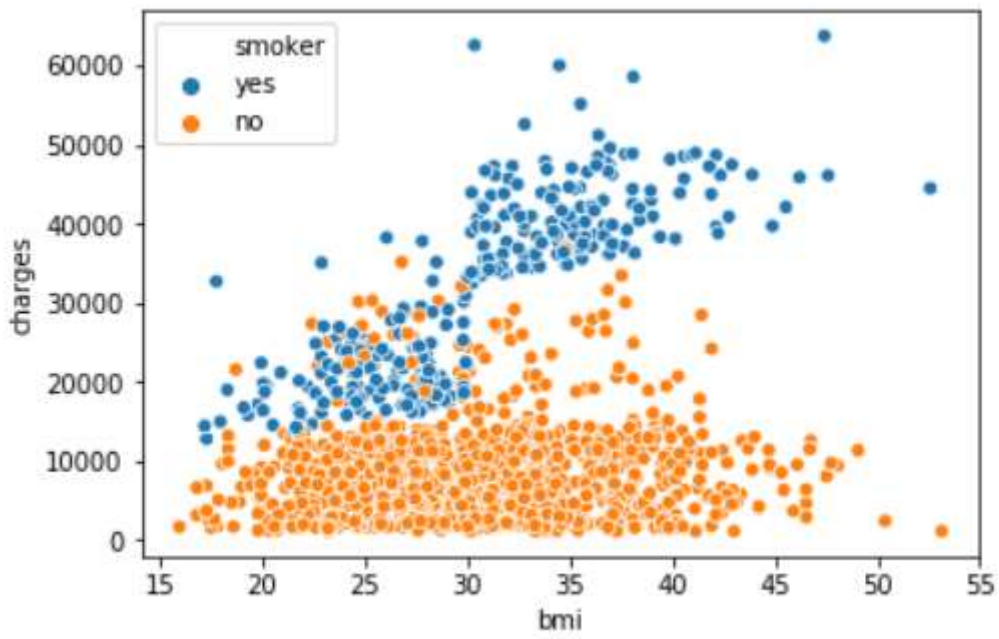


- 1:High Bias- Low Variance 2: High Bias-High Variance 3:Low Bias-Low Variance 4:Low Bias-High Variance
- 1:High Bias- High Variance 2:High Bias-Low Variance 3:Low Bias-High Variance 4:Low Bias-Low Variance
- 1:Low Bias- High Variance 2:Low Bias-Low Variance 3:High Bias-High Variance 4: High Bias-Low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

✓ Q16- Which of the statements below is incorrect for ensemble learning and its techniques? * 4/4

- Its techniques use a combination of learning algorithms to optimize better predictive performance.
- It makes the model more robust.
- Typically, It reduces overfitting in the data.
- Typically, it reduces underfitting in the data. ✓

✓ Q17- Which one is the correct option for the below visualization? * 4/4



- sns.lmplot(x="bmi", y="charges", hue="smoker", data=insurance_data)
- sns.scatterplot(x=insurance_data['bmi'], y=insurance_data['charges'])
- sns.scatterplot(x=insurance_data['bmi'], y=insurance_data['charges'], ✓

hue=insurance_data['smoker'])

sns.regplot(x=insurance_data['bmi'], y=insurance_data['charges'])

✓ Q18- Which of the below is/are nominal variable(s)? * 4/4

I - Gender

II - Genotype

III - Religious preference

IV- IQ

V - Income earned in a week.

I, II

I, II, III ✓

II, III, IV

I, III, IV

✓ Q19- What is the function of parameter ci? * 4/4

Defining the plot type

To make subplot

To highlight the classes of data points

Defining to show confidence interval ✓

✓ Q20- Imagine that you have a company and you would like to create a plot which shows the sales based on date. Which one of the following plot function is less likely suitable for this job? * 4/4

- ○ sns.lineplot
- ◉ sns.heatmap ✓
- ○ sns.barplot
- ○ sns.scatterplot

✓ Q21- Which of the following statement is inconsistent with pipelines? * 4/4

- ○ You won't need to manually keep track of your training and validation data at each step with a pipeline.
- ○ With a pipeline, we can use the cross-validation technique easily.
- ○ With pipelines, there is less probability to forget a preprocessing step.
- ◉ It's hard to productionize a model with pipelines. ✓

✓ Q22- Which of the following statements are true about the intended use of cross-validation? * 4/4

I - To reduce randomness while measuring model performance.

II - To get a better measure of model performance.

III - To increase model's training performance.

IV - To increase MAE (mean absolute error) or MSE (mean squared error).

- (●) I, II ✓
- ( ) II, III
- ( ) II, IV
- ( ) I, IV

✓ Q23- 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 * 4/4

```
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
- ◉ I, II, III ✓
- ○ II, III, IV

✓ Q24- Which of the following statements are true about LabelEncoder and OneHotEncoder? * 4/4

I-They help us to deal with categorical values.

II-Label Encoding assigns each value to a different integer whether it is unique or not.

III-One Hot Encoding creates new column for every possible value in the original data.

IV-For large number of categorical variable count value (such as 15 different values) it is not good to use One Hot Encoder generally.

- ○ I, II, III
- ○ II, III, IV
- ◉ I, III, IV ✓
- ○ All of them

✓ Q25- Which statement is not true with the following code? museum_data = pd.read_csv(museum_filepath,index_col="Date",parse_dates=True) * 4/4

- ○ If we did not use parse_dates, the type of the Date column would not change to datetime. (Assuming that the original type of the column is not datetime)
- ◉ Above code means that creating a new Date column that is not in csv and this column is defined as the index of the museum_data. ✓
- ○ When parse_dates = True, the type of Date column in museum_data becomes datetime
- ○ The index value of the museum_data is the Date column in the csv file.