Sabura Khanam

KSU ID: 001057990

1. **Train Data Set after applying one-hot encoder and imputing with the median values.**

```
[22] df = pd.get_dummies(df, columns=['School'])
     df
```

	English	Science	Math	School_A	School_B	School_C
0	90.0	50.0	70	1	0	0
1	55.0	75.0	80	1	0	0
2	80.0	85.0	90	0	1	0
3	45.0	90.0	100	0	0	1
4	95.0	45.0	75	0	0	1
5	80.0	70.0	60	0	1	0
6	80.0	80.0	80	1	0	0

Next steps: Generate code with df | View recommended plots

2. **In addition, you have a trained linear regression model with the intercept and coefficients: $\beta 0$=14,βSch_A=−1,βSch_B=−10,βSch_C=10,βEng=0.2, and βSci=0.8. Apply the same data preprocessing techniques as in Q1 and manually calculate the predictions using the regression model. You can either type or scan manually written results. Note that imputation should use the median from the training set.**

Processed Data for question 2

```
df_test = pd.get_dummies(df_test, columns=['School'])
df_test
```

	English	Science	Math	School_A	School_B	School_C
0	85.0	60.0	80	0	0	1
1	75.0	75.0	82	1	0	0
2	75.0	90.0	90	0	1	0
3	60.0	70.0	80	1	0	0
4	80.0	60.0	90	0	0	1

Next steps: Generate code with df_test View recommended plots

<u>Prediction of the Data Set</u>:

```
a = -1*0 + -10*0 + 10*1 + 0.2*85 + 0.8*60 + 14

b = -1*1 + -10*0 + 10*0 + 0.2*75 + 0.8*75 + 14

c = -1*0 + -10*1 + 10*0 + 0.2*75 + 0.8*90 + 14

d = -1*1 + -10*0 + 10*0 + 0.2*60 + 0.8*70 + 14

e = -1*0 + -10*0 + 10*1 + 0.2*80 + 0.8*60 + 14

print(a)
print(b)
print(c)
print(d)
print(e)
```

```
89.0
88.0
91.0
81.0
88.0
```

3. Evaluate the model provided in Q2 by computing the following metrics on the test dataset: Mean Squared Error (MSE), Mean Absolute Error (MAE), and $R2$ score.

Dataset :-

Actual Math (y)	Predicted Math ($\hat{y}$)
80	89
82	88
90	91
80	81
90	88

$$MAE = \frac{1}{n} \sum_{i=0}^{n-1} |y - \hat{y}|$$

$$= \frac{1}{5} \left(|80-89| + |82-88| + |90-91| + |80-81| + |90-80| \right)$$

$$= \frac{1}{5} (9+6+1+1+2)$$

$$= \frac{19}{5} = \boxed{3.8}$$

$$MSE = \frac{1}{5} \left((80-89)^2 + (82-88)^2 + (90-91)^2 + (80-81)^2 + (90-88)^2 \right)$$

$$= \frac{1}{5} \times 123$$

$$= \frac{123}{5}$$

$$= \boxed{24.6}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y - \hat{y})^2}{\sum_{i=1}^{n} (y - \bar{y})^2}$$

$$\bar{y} = \frac{80+82+90+80+90}{5} = \frac{422}{5} = 84.4$$

$$(y-\hat{y})^2 = (80-89)^2 + (82-88)^2 + (90-91)^2 + (80-81)^2 + (90-88)^2$$

$$= 81 + 36 + 1 + 1 + 4$$

$$= 123$$

$$(y-\bar{y})^2 = (80-84.4)^2 + (82-84.4)^2 + (90-84.4)^2 + (80-84.4)^2 + (90-84.4)^2$$

$$= 17.6 + 9.6 + 30.6 + 17.6 + 30.6$$

$$= 105.2$$

$$\therefore R^2 = 1 - \frac{123}{105.2}$$

$$= 1 - 1.168$$

$$= -0.168$$

4. **Suppose you have another model trained with a different dataset, which consists of the following intercept and coefficients: $\beta 0$=14, βSch_A=−2, βSch_B=−11, βSch_C=12, βEng=0.2, and βSci=0.8. Compute the predictions with the test dataset and its MSE and MAE. Compare them with the results in Q3 to explain which metric penalizes the model generating more extreme errors.**

```
m = 0*-2 + 0*-11 + 1*12 + 85*0.2 + 60*0.8 + 14
n = 1*-2 + 0*-11 + 0*12 + 75*0.2 + 75*0.8 + 14
o = 0*-2 + 1*-11 + 0*12 + 75*0.2 + 90*0.8 + 14
p = 1*-2 + 0*-11 + 0*12 + 60*0.2 + 70*0.8 + 14
q = 0*-2 + 0*-11 + 1*12 + 80*0.2 + 60*0.8 + 14

print(m)
print(n)
print(o)
print(p)
print(q)
```

```
91.0
87.0
90.0
80.0
90.0
```

Test Data set

Actual (y)	Predict ($\hat{y}$)
80	91
82	87
90	90
80	80
90	90

$$MAE = \frac{1}{5}\left(|80-91| + |82-87| + |90-90| + |80-80| + |90-90|\right)$$

$$= \frac{1}{5}\left(11+5+0+0+0\right)$$

$$= \frac{1}{5} \times 16$$

$$= \boxed{3.2}$$

$$MSE = \frac{1}{5}\left((80-91)^2 + (82-87)^2 + (90-90)^2 + (80-80)^2 + (90-90)^2\right)$$

$$= \frac{1}{5}\left(121+25+0+0+0\right)$$

$$= \frac{1}{5} \times 146$$

$$= \boxed{29.2}$$

From dataset 3, the higher MSE is penalized more for generating extreme errors compared to dataset from 4. Therefore, MSE is the metric that penalizes the model generating more extreme errors.

$$MAE = \frac{1}{n} \sum_{i=0}^{n=1} |y - \hat{y}|$$

$$= \frac{1}{5} \times [80-91] + [82-87] + [90-90] + [80-80] + [90-90]$$

$$= \frac{1}{5} \times (-11) + (-5) + (0) + (0) + (0)$$

$$= \frac{1}{5} \times -55$$

$$= -11$$

$$MSE = \frac{1}{n} \sum_{i=0}^{n-1} (y - \hat{y})^2$$

$$= \frac{1}{\cancel{9}5} \times (80-91)^2 + (82-87)^2 + (90-90) + (80-80) + (90-90)$$

$$= \frac{1}{5} \times (-11)^2 + (-5)^2 + 0 + 0 + 0$$

$$= \frac{1}{5} \times (121 + 25)$$

$$= \frac{146}{5}$$

$$= 29.2$$