Diamonds prices prediction

Dataset:

-Link in Kaggle: Diamonds (kaggle.com)

-Dataset descebtion:

It contains different sizes,prices,types of diamonds

-dataset original shape(53940,11)

Wrangling & Preprocessing

-this data doesn’t contain null or duplicated values but some columns have value ‘0’

So we`ve drop it .

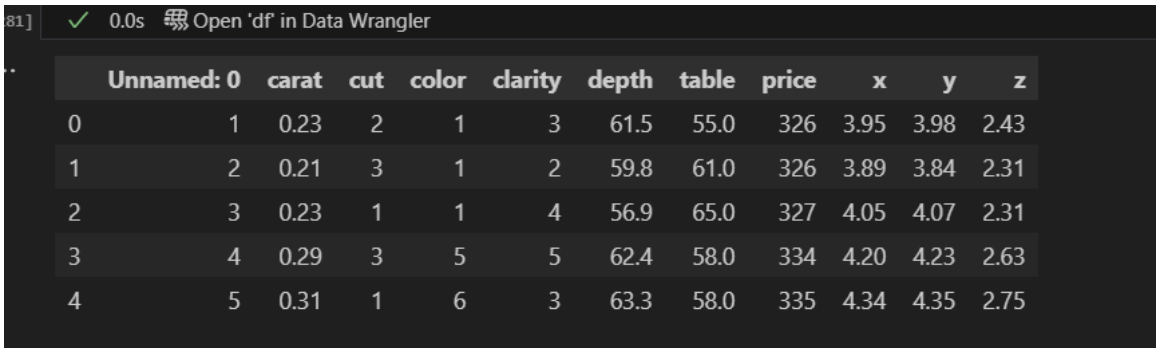
Encoding:

-using LablEncoder

. before encoding:

	Unnamed: 0	carat	cut	color	clarity	depth	table	price	x	y	z
0	1	0.23	Ideal	E	SI2	61.5	55.0	326	3.95	3.98	2.43
1	2	0.21	Premium	E	SI1	59.8	61.0	326	3.89	3.84	2.31
2	3	0.23	Good	E	VS1	56.9	65.0	327	4.05	4.07	2.31
3	4	0.29	Premium	I	VS2	62.4	58.0	334	4.20	4.23	2.63
4	5	0.31	Good	J	SI2	63.3	58.0	335	4.34	4.35	2.75

. after encoding:



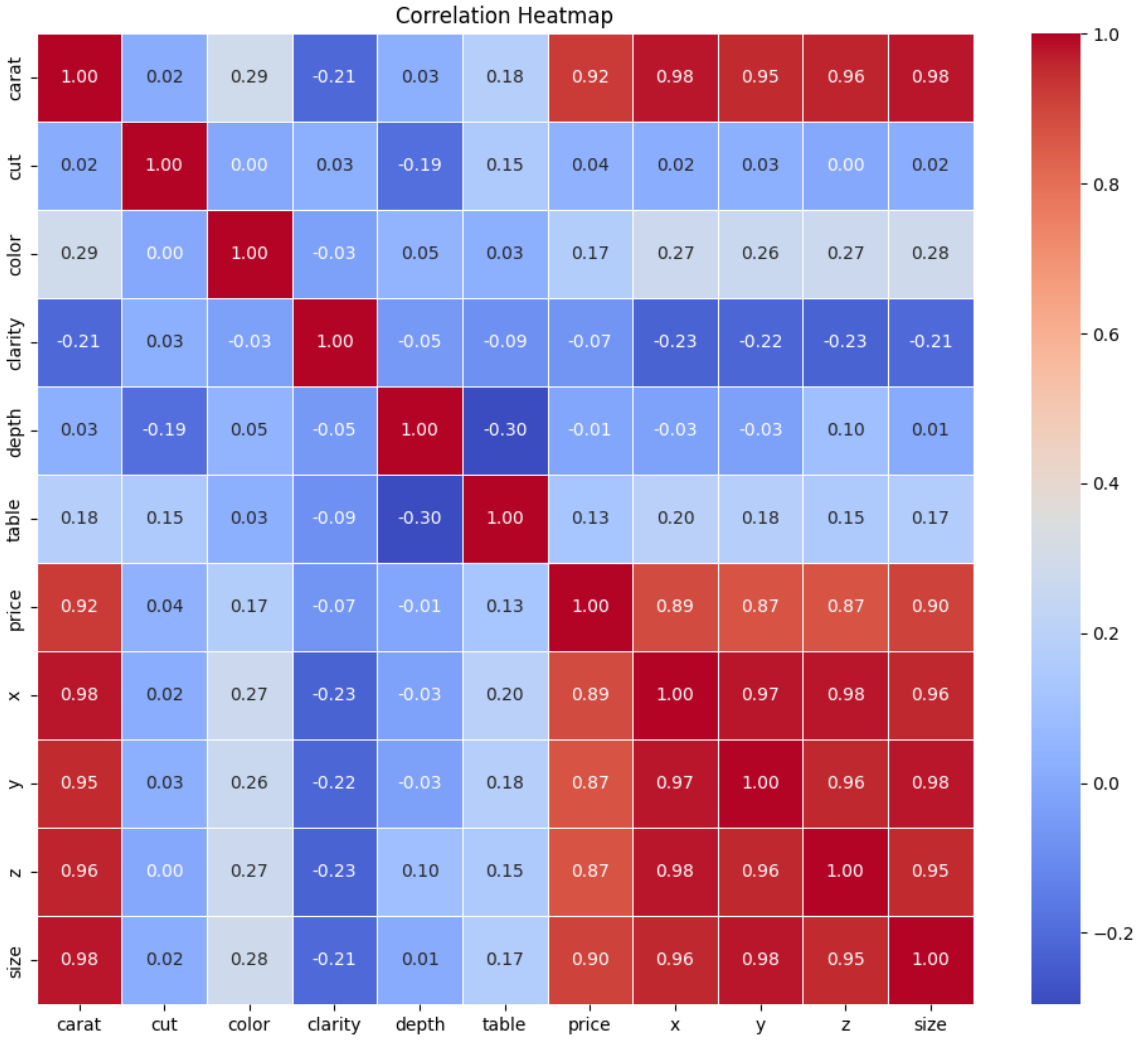
	Unnamed: 0	carat	cut	color	clarity	depth	table	price	x	y	z
0	1	0.23	2	1	3	61.5	55.0	326	3.95	3.98	2.43
1	2	0.21	3	1	2	59.8	61.0	326	3.89	3.84	2.31
2	3	0.23	1	1	4	56.9	65.0	327	4.05	4.07	2.31
3	4	0.29	3	5	5	62.4	58.0	334	4.20	4.23	2.63
4	5	0.31	1	6	3	63.3	58.0	335	4.34	4.35	2.75

-After that dropping column [‘Unnamed 0‘] as we don’t what this represents

---

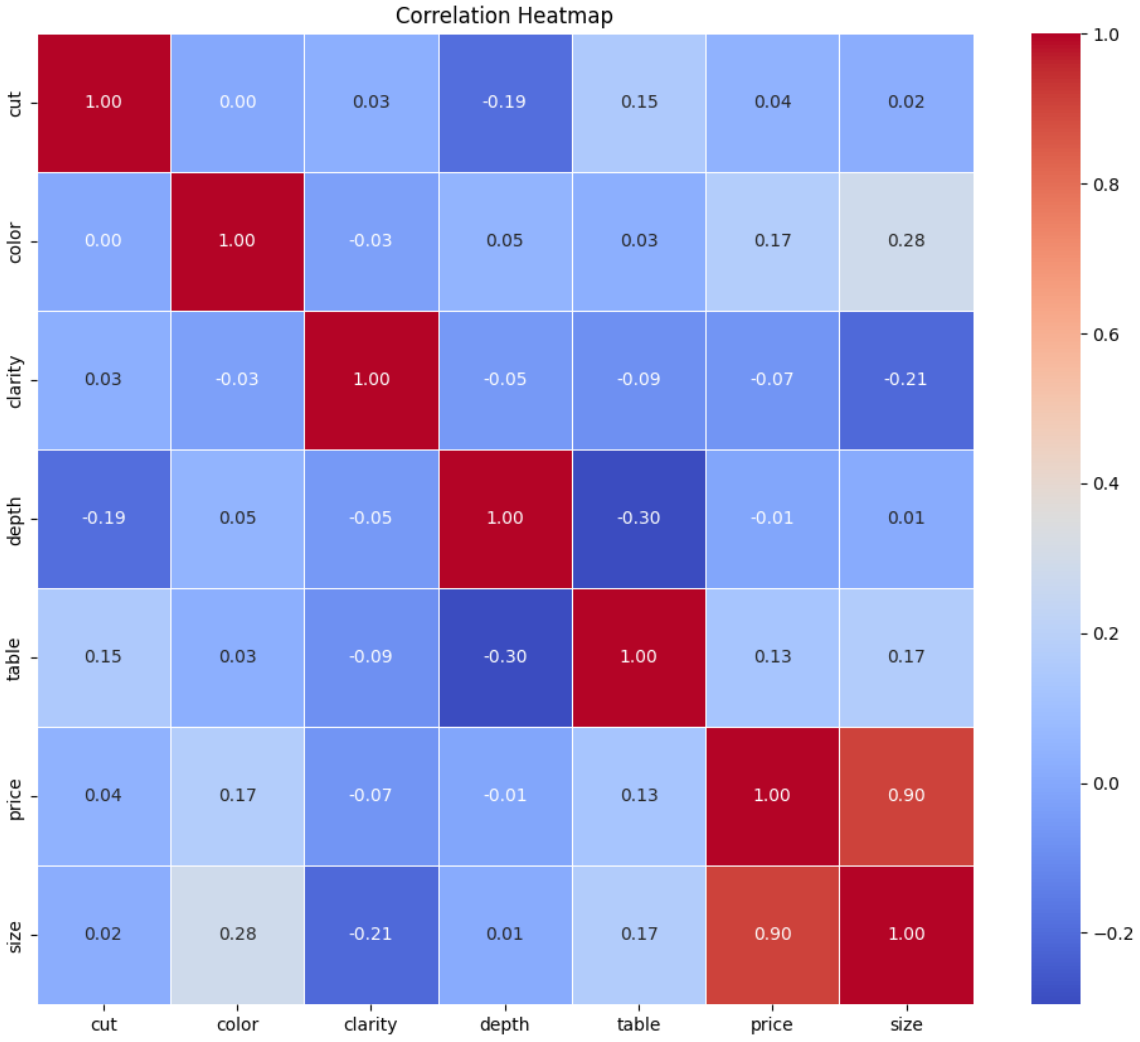
- creating new feature called size by :multiplying x,y,z columns.

Checking correlation:



-since x,y,z are highly correlated we`ve dropped them.

After dropping them:



Normalizing:

-Using StandardScaler.

Splitting data:

Train and test with:test_size=0.2 and random_state=42.

Applying KFold with:

K=5 , shuffle=True,random_state=42.

X_train_fold.shape(34509,6)

Y_train_fold.shape(34509,)

X_test_fold.shape(8627,6)

Y_test_fold.shape(8627,)

Modeling:

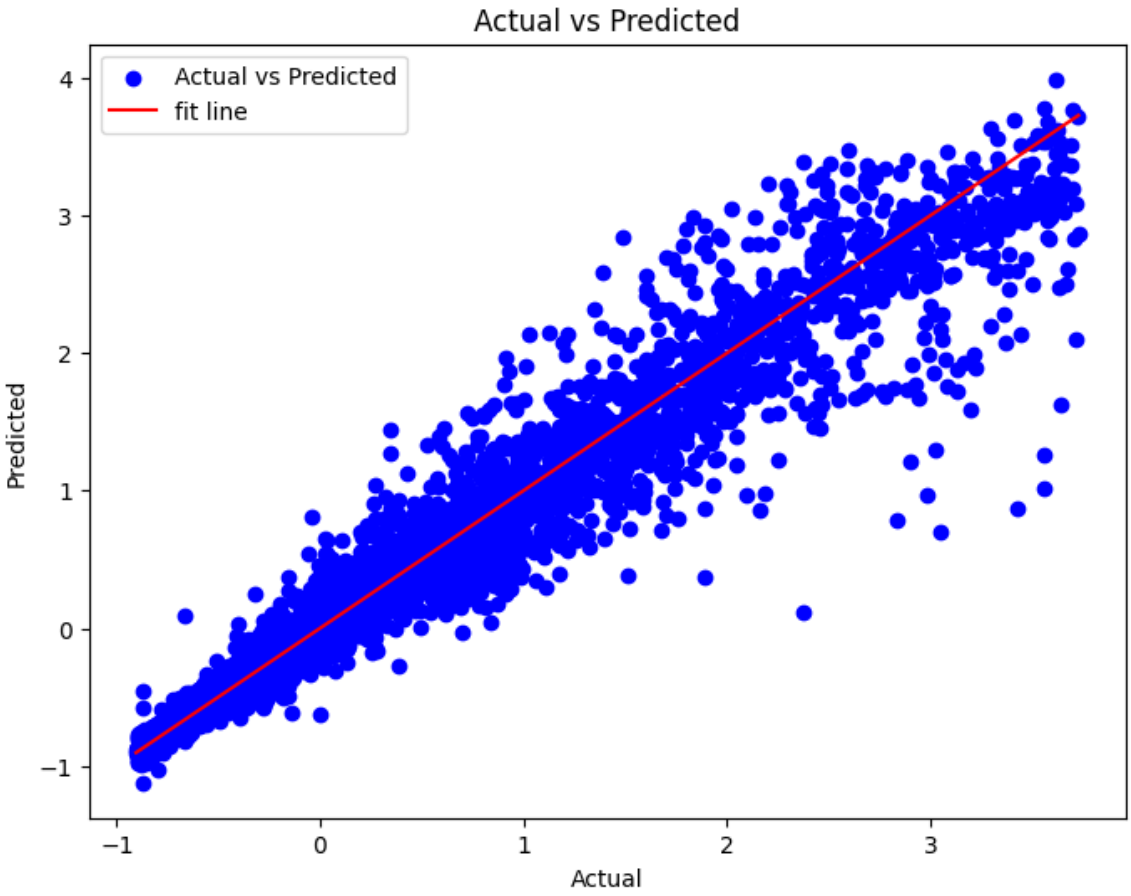
Using SVR with parameters:

[c=10 , kernel='rbf' , gamma='scale']

Evaluation

Applying some metrics:

```
Mean Absolute Error: 0.11067256867197495
...
Mean Squared Error: 0.04255431906748594
Root Mean Squared Error: 0.20628698230253392
Training accuracy: 0.9627855366935698
Test accuracy: 0.9577351451527133
```


Actual vs Predicted
Actual vs Predicted
fit line
Predicted
Actual
4
3
2
1
0
−1
−1
0
1
2
3