Regression Task

Load the Data

Using the pandas read_excel function we imported the Real Estate dataset then converted it to a numpy array. We also removed the header and stored it separately so we could actually do things with the data.

Meet the Data

There are 7 features.
The names of the features were:

- X1 transaction date
- X2 house age
- X3 distance to the nearest MRT station
- X4 number of convenience stores
- X5 latitude
- X6 longitude
- Y house price of unit area

The target column was “Y house price of unit area”.
There were 414 samples.
The first 5 rows of the data were:

- 2012.9166667, 32, 84.87882, 10, 24.98298, 121.54024, 37.9
- 2012.9166667, 19.5, 306.5947, 9, 24.98034, 121.53951, 42.2
- 2013.5833333, 13.3, 561.9845, 5, 24.98746, 121.54391, 47.3
- 2013.5, 13.3, 561.9845, 5, 24.98746, 121.54391, 54.8
- 2012.8333333, 5, 390.5684, 5, 24.97937, 121.54245, 43.1

The correlation matrix for the data was as follows:

	X1	X2	X3	X4	X5	X6	X7
X1	1	.018	.061	.010	.035	-.041	.088
X2	.018	1	.026	.050	.054	-.049	-.211
X3	.061	.026	1	-.603	-.591	-.806	-.674
X4	.010	.050	-.603	1	.444	.449	.571
X5	.035	.054	-.591	.444	1	.413	.546
X6	-.041	-.049	-.806	.449	.413	1	.523

X7	.088	-.211	-.674	.571	.546	.523	1

Model Fitting and Parameter tuning

Lasso and Ridge both preferred alpha = 0 but this is some error within our code which makes it complicated, they start off with 100% accuracy at a0 and move around 20% accuracy afterwards.

Metrics

When trying to make a confusion matrix to collect the data needed for R2 and RMSE I was unable to generate the matrix, I got stuck with a “continuous value not supported” and could not find a solution.

Comparison of Different Models

With the dysfunctional code for the model it is not possible to make comparisons in the weighting of the coefficients because the models are performing so badly that changing the weights of the parameters seems to have 0 effect.

Classification Task

Load the Data

Using basic python loading techniques, we loaded the banknote authentication database into a normal python array. As there was no header in the original dataset, there is no need to remove the first line.

Meet the Data

There are four features in the data set: variance, skewness, curtosis, entropy. Target is just called class without much else description. There are 1372 samples in the dataset, with the first five rows of data being:

3.6216	8.6661	-2.8073	-0.44699	0
4.5459	8.1674	-2.4586	-1.4621	0
3.866	-2.6383	1.9242	0.10645	0
3.4566	9.5228	-4.0112	-3.5944	0

0.32924	-4.4552	4.5718	-0.9888	0

Correlation Matrix:

	variance	skewness	curtosis	entropy	class
variance	1	0.264	-0.381	0.277	-0.725
skewness	0.264	1	-0.787	-0.526	-0.445
curtosis	-0.381	-0.787	1	0.319	0.156
entropy	0.277	-0.526	0.319	1	-0.023
class	-0.725	-0.445	0.156	-0.023	1

Model Fitting

Accuracy: 0.9854545454545455

Metrics

Class of 0

- Precision: 0.9872
- Recall: 0.9872
- Accuracy: 0.9855
- Sensitivity: 0.9832

Class of 1

- Precision: 0.9832
- Recall: 0.9832
- Accuracy: 0.9855
- Sensitivity: 0.9832

Metrics (ROC curve)

Our ROC curve was lacking in points as a consequence of some issue giving our model unrealistically high accuracy on our dataset.