KNN Regression

Statistical Prediction Modeling ***DATA 1204***

Assignment No1

Due Date : , 31st Jan 2022

Jumana Obeid

100832624

DURHAM COLLEGE
SUCCESS MATTERS

1. Create a PowerPoint (PPT) presentation that includes the following:

a. Rational Statement (summary of the problem or problems to be addressed by the PPT) – 2%

We will use KNN (Nearest Neighbor Algorithm) to find the best prediction for any point of Data that we might encounter in the future based on training an available dataset for Energy Use Cooling , this dataset is available in the format of csv file.

The KNN algorithm predicts the output of a value based on the nearest neighbors’ output values, In other words, we can take one neighbor output value as a reference to predict the value of any point , the output of this point will be the exact same value of the output of this neighbor point.

Or we can take two or more nearest points and the value will be the average(mean) of the output of these points. We are using regression and not classification because we are predicting continuous values and not discreate.

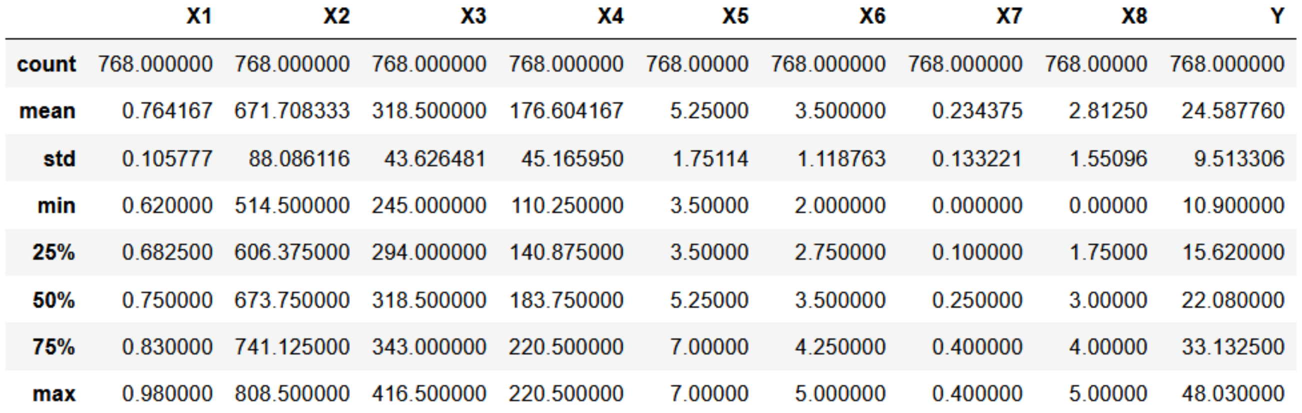
The features (independent variables (X as mentioned above).
The outcome(dependent variable(Y as mentioned above).

Independent Variables:
X1 - Relative Compactness
X2 - Surface Area
X3 - Wall Area
X4 - Roof Area
X5 - Overall Height
X6 - Orientation
X7 - Glazing Area
X8 - Glazing Area
Distribution Dependent Variable:
Y- Cooling Load

The steps are as follows:

1. Prepare your data(dataset).
2. Split the dataset to a training data and validation(test)data.
3. Create Ml model(KNN).
4. Plot a learning Curve for our KNN-(check bias & variance).
5. Review the ML model predictive performance(see accuracy).
6. Use Grid Search to know the best K, (number of Neighbor's)

Insights of the data



	X1	X2	X3	X4	X5	X6	X7	X8	Y
count	768.000000	768.000000	768.000000	768.000000	768.00000	768.000000	768.000000	768.00000	768.000000
mean	0.764167	671.708333	318.500000	176.604167	5.25000	3.500000	0.234375	2.81250	24.587760
std	0.105777	88.086116	43.626481	45.165950	1.75114	1.118763	0.133221	1.55096	9.513306
min	0.620000	514.500000	245.000000	110.250000	3.50000	2.000000	0.000000	0.00000	10.900000
25%	0.682500	606.375000	294.000000	140.875000	3.50000	2.750000	0.100000	1.75000	15.620000
50%	0.750000	673.750000	318.500000	183.750000	5.25000	3.500000	0.250000	3.00000	22.080000
75%	0.830000	741.125000	343.000000	220.500000	7.00000	4.250000	0.400000	4.00000	33.132500
max	0.980000	808.500000	416.500000	220.500000	7.00000	5.000000	0.400000	5.00000	48.030000

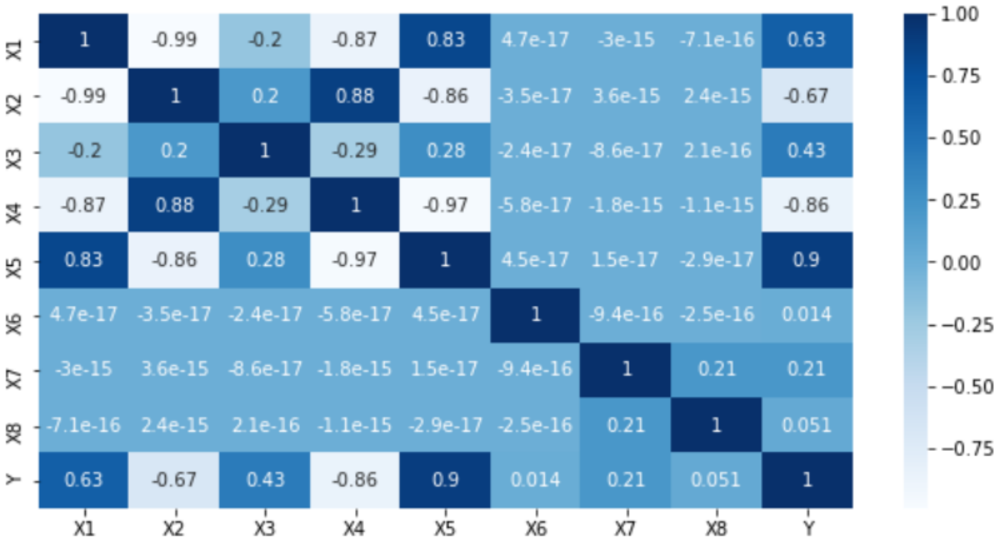
- 75% of the cooling load /output is 33 or below.
- 50% of the overall height is 5.2.This factor has strong correlation with the output as we are going to see.

c. Present the Correlation Heatmap and explain two (2) insights – 2%

As you we see in the heat map(dark areas & very light areas), there is strong correlation more than **8%** between: –

- **Positive correlation:**
- ('X4', 'X2'), ('X5', 'X1'), ('Y', 'X5')
- (RoofArea,SurfaceArea),(OverallHight,Relative Compactness)
- (CollingLoad,OverallHight)

- **Negative correlation:**
- [('X2', 'X1'), ('X4', 'X1'), ('X5', 'X2'),
- ('X5', 'X4'), ('Y', 'X4'),]
- (SurfaceArea,RelativeCompactness)
- (RoofArea,RelativeCompactness)
- (OverallHight,SurfaceArea)
- (OverallHight,RoofArea)
- (CoolingLoad, RoofArea)



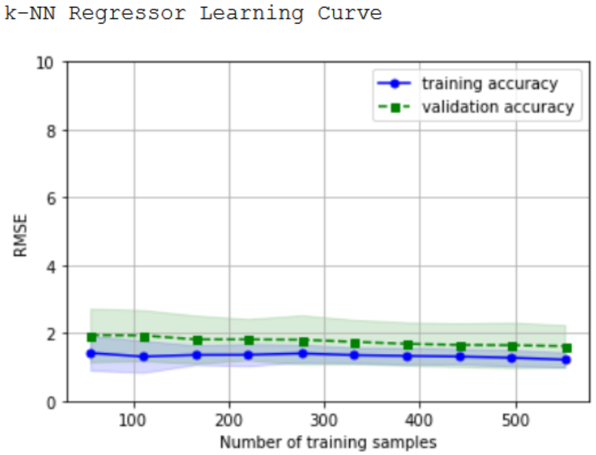
	X1	X2	X3	X4	X5	X6	X7	X8	Y
X1	1	-0.99	-0.2	-0.87	0.83	4.7e-17	-3e-15	-7.1e-16	0.63
X2	-0.99	1	0.2	0.88	-0.86	-3.5e-17	3.6e-15	2.4e-15	-0.67
X3	-0.2	0.2	1	-0.29	0.28	-2.4e-17	-8.6e-17	2.1e-16	0.43
X4	-0.87	0.88	-0.29	1	-0.97	-5.8e-17	-1.8e-15	-1.1e-15	-0.86
X5	0.83	-0.86	0.28	-0.97	1	4.5e-17	1.5e-17	-2.9e-17	0.9
X6	4.7e-17	-3.5e-17	-2.4e-17	-5.8e-17	4.5e-17	1	-9.4e-16	-2.5e-16	0.014
X7	-3e-15	3.6e-15	-8.6e-17	-1.8e-15	1.5e-17	-9.4e-16	1	0.21	0.21
X8	-7.1e-16	2.4e-15	2.1e-16	-1.1e-15	-2.9e-17	-2.5e-16	0.21	1	0.051
Y	0.63	-0.67	0.43	-0.86	0.9	0.014	0.21	0.051	1

d. Present the Learning Curve for the k-NN standard model and explain two (2) insights – 2%

From the learning curve we conclude that the module has high bias, the plot has both low training and cross validation accuracy. This indicates that the module underfits the training data and thus the case of high bias. The validation accuracy is far from the desired accuracy.

To solve this issue, we can either:

- Add more features
- Decrease the degree of regularization .



e. Present and explain two (2) insights from the evaluation metrics (i.e. Adj. R2 , MAE, RMSE) for the Optimized k -NN Regression model – 2%

- Adj_R2: .9 states that 90% of the variance of the dependent variable being studied is explained by the variance of the independent variable.

- RMSE is very close to 10% of the output mean. Which was 2.8.We can say that it is a pretty good module.

f. State and explain two (2) recommendations for Mr. John Hughes for next steps. – 2%

- We recommend that Mr.Johns takes more features into consideration in the coming module to increase the accuracy of the module.(As mentioned above).
- Also, try other algorithms like SVM for example.