PART-2 CLASSIFICATION ANALYSIS MODEL EQUATION AND DATA VALIDATION

Take a dataset for classification analysis with a minimum of 3 predictors (e.g.: acclerometer data from a smart phone for human activity recognition). Describe the dataset (predictors, response, number of data points). Perform exploratory data analysis to understand the relationships between the variables. Please divide the data for training and testing. What algorithm is used for fitting the model? What accuracy is achieved when the model is fit? Which of the predictors are significant?

How well did the model fit the test data? What is the confusion matrix for this model and what do you infer from it? Please describe in words about the relationship between the significant predictors and the response.

Please share the notebook (the code and the screenshots of the visualizations) to show this work with the assignment.

The dataset taken from https://Kaggle.com. The dataset comprises 768 samples and 8 features, aiming to predict two real valued responses. It can also be used as a multi-class classification problem if the response is rounded to the nearest integer.

Attribute Information:

The dataset contains eight attributes (or features, denoted by X1...X8) and two responses (or outcomes, denoted by y1 and y2). The aim is to use the eight features to predict each of the two responses.

- Predictor Variables:

1. X1 Relative Compactness
2. X2 Surface Area
3. X3 Wall Area
4. X4 Roof Area
5. X5 Overall Heigh
6. tX6 Orientation
7. X7 Glazing Area
8. X8 Glazing Area Distribution

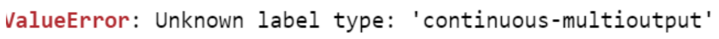
- Response Variables:

1. y1 Heating Load
2. y2 Cooling Load

- No. of data points = 768

- Decision Tree algorithm has been used for model fitting.

- *##While creating confusion matrix, following error was received:*



```
ValueError: Unknown label type: 'continuous-multioutput'
```

- *Model Equations:*

Heating Load(Y1) = 84.0145 – (64.7740*relative compactness) - (0.0626 * surface area) + (0.0361 *Wall area) - (0.0494 *Roof area) + (4.1699 * overall height) - (0.0233 *Orientation) + (19.9327* Glazing Area) + (0.2038* glazing area distribution)

Cooling Load(Y2) = 97.2457– (70.7877*relative compactness) - (0.0661* surface area) + (0.0225*Wall area) - (0.0443*Roof area) + (4.2838* overall height) – (0.1215 *Orientation) + (14.7171* Glazing Area) + (0.0407* glazing area distribution)

- The results indicate that relative compactness, wall area and roof area appear mostly associated with HL andCL.

DATA VALIDATION IN EXCEL

Heating Load(Y1)

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.957
R Square	0.916
Adjusted R Square	0.914
Standard Error	2.934
Observations	768.000

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	8.000	71546.076	8943.260	1187.063	0.000
Residual	760.000	6543.766	8.610		
Total	768.000	78089.842			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	84.015	19.034	4.414	0.000	46.650	121.379	46.650	121.379
X Variable 1	-64.774	10.289	-6.295	0.000	-84.973	-44.575	-84.973	-44.575
X Variable 2	-0.087	0.017	-5.112	0.000	-0.121	-0.054	-0.121	-0.054
X Variable 3	0.061	0.007	9.148	0.000	0.048	0.074	0.048	0.074
X Variable 4	0.000	0.000	65535.000	#NUM!	0.000	0.000	0.000	0.000
X Variable 5	4.170	0.338	12.337	#NUM!	3.506	4.833	3.506	4.833
X Variable 6	-0.023	0.095	-0.246	0.805	-0.209	0.163	-0.209	0.163
X Variable 7	19.933	0.814	24.488	0.000	18.335	21.531	18.335	21.531
X Variable 8	0.204	0.070	2.914	0.004	0.067	0.341	0.067	0.341

Cooling Load(Y2)

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.942
R Square	0.888
Adjusted R Square	0.885
Standard Error	3.201
Observations	768.000

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	8.000	61627.583	7703.448	859.119	0.000
Residual	760.000	7788.205	10.248		
Total	768.000	69415.788			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	97.246	20.765	4.683	0.000	56.483	138.009	56.483	138.009
X Variable 1	-70.788	11.225	-6.306	0.000	-92.824	-48.751	-92.824	-48.751
X Variable 2	-0.088	0.019	-4.737	0.000	-0.125	-0.052	-0.125	-0.052
X Variable 3	0.045	0.007	6.161	0.000	0.030	0.059	0.030	0.059
X Variable 4	0.000	0.000	65535.000	#NUM!	0.000	0.000	0.000	0.000
X Variable 5	4.284	0.369	11.618	#NUM!	3.560	5.008	3.560	5.008
X Variable 6	0.122	0.103	1.176	0.240	-0.081	0.324	-0.081	0.324
X Variable 7	14.717	0.888	16.573	0.000	12.974	16.460	12.974	16.460
X Variable 8	0.041	0.076	0.534	0.594	-0.109	0.190	-0.109	0.190