Machine Learning Competition

Author: *Ayana Ghosh*

[**Overall suggestion**: In each team divide yourself into two groups. One group may focus on answering Q. (1-3) and other (4-6) but make sure you discuss and work together.] **(Total Points: 100)**

1. Fill in the Table below with information from the Dataset_challenge.csv (a-d) and DatasetII_challenge.csv (e) files.
(**Points: 10**)

(a)	Full Dataset Size	
(b)	Training Set Size (90%)	
(c)	Test Set Size (10%)	
(d)	Number of Descriptors listed	
(e)	Validation Set Size (DatasetII_challenge.csv)	

2. Name the five most important descriptors pertinent to predict the endpoint. (**Points: 20**)
(Hint: Median Analysis)

3. Can you identify any quantitative trend among these 'important descriptors' and the endpoint? (**Points: 10**)

4. Fill in the Table below with information from the RF_Model.R code.
(**Points: 10**)

Name and values of Hyperparameters	

5. Propose a solution to optimize the hyperparameters. (**Points: 10**)

6. Run the RF_Model.R code 10 times with given hyperparameters and list down the root mean square error (RMSE), mean absolute error (MAE) of the model as applied to the training and test. (up to three digits after decimal) (**Points: 10**)

Entry	RMSE_Train	MAE_Train	RMSE_Test	MAE_Test
(a)				
(b)				
(c)				
(d)				
(e)				
(f)				
(g)				
(h)				
(i)				
(j)				

5. Compare the entries (a)-(j). (**Points: 30**)

I. Are the errors, if compared between any two entries or more different ?

II. If so, why ?

III. What does this mean?

IV. Is this an inconsistency of the model, the dataset itself or algorithm ?

V. Propose a solution. (How will you make a more ‘universal’ model?)