**Solution Sheet**

1. Which model have you used for Covid Cases prediction? Explain your model.

Our model is *CatBoostRegressor* with '*Population [2011]*', '*Toilets Avl*', '*# of hospitals*', '*Median*

*Age*', '*Sex Ratio*', '*Covid Cases\_\_min*' (a State-level Aggregate Feature), '*Water Purity*', and,

'*H Index*' as important features.

We've selected MAPE [Mean Absolute Percentage Error] as the Evaluation Metrics as Absolute Difference (MAE) is not a good measure when Target Column ('*Covid Cases*') has a high Variance.

We've ignored '*Population 2001*' as it is very correlated to '*Population 2011*', but old and completely missing in the Test set. '*Female Population*' seems to be wrong considering '*Population 2011*' and '*Sex Ratio*', so it has been ignored.

We've imputed missing values by the Mean of the value in the respective State. We've created aggregate Features (mean, min, max, sum, nunique, var) of Numerical Columns, and experimented in Model Building.

We used PyCaret for quick Modelling in the beginning.

1. Which model have you used for Foreign Visitors Time series prediction? Explain your model.  
   '*Foreign Visitors*' has a linear relation with the previous month (in the form aX + b + c); where, *X* is the previous month's '*Foreign Visitors*', and *a*, and b are constants, and *c* is a small random variable. That makes it an extremely short Time-Series [with 5 Data Points only], and without a strong pattern/trend. We computed *Percentage Change* for the upcoming month as the Average of Median and Mean of Percentage Change of the Previous Consecutive Months.  
   With that, we have '*Foreign Visitors*' for the next month, and assuming all other columns remain constant, we use the same *CatBoostRegressor* model trained from Problem-1.