**Solution Sheet**

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

In the task given to us under the Flipr Hackathon Hiring Program 4.0 , we have to predict the probability of a person being affected with the virus COVID-19 given the various features describing the health and the environment of the person. The given problem is essentially a regression based problem. The given dataset consists of various categorical and continuous variables of different scales, therefore, preprocessing steps like missing value treatment, multicollinearity checking,feature scaling etc. were performed. Finally , I used the **Catboost Regressor** model from the Catboost library in python to predict the probability of infection of a person on 20th March. CatBoost is based on gradient boosted decision trees. During training, a set of decision trees is built consecutively. Each successive tree is built with reduced loss compared to the previous trees.

1. Which model have you used for Diuresis Time series prediction? Explain your model.

In task 2, we had to predict infection probability using the diuresis value of the person on 27th March using the values for past 7 days. Essentially this is a time series based problem .I used Cubic Spline Interpolation for time series from the python Scipy library to predict the value of Diuresis on 27th March. Cubic Spline Interpolation fits the given points by a piecewise polynomial function, which is a composite function formed by low-degree polynomials.

Now, since the people\_ID in the test set don’t match with the ids given in the diuresis sheet in training data, according to my understanding of task , I have replaced the diuresis value of 27th march in the training data and then have trained the task 1 model again and predicted the new Infect\_prob on test data.