# Solution Sheet

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

In the given task, the target was to predict the infection probability, that is, the probability of a person getting infected with Coronavirus using various features such as region, health conditions, sex, etc. The task boils down to a regression problem to predict a float number, that is, the infection probability. The dataset contains missing values for different columns(NaNs), some columns are categorical and some continuous. Therefore, during preprocessing, the NaNs were filled differently for both data types. Also, some columns had high correlation with each other, so one of them was dropped. Finally, features were also scaled using StandardScaler. I used many models but the best cross-validation rmse was of Random Forest Regressor.

Random forest is a Supervised Learning algorithm which uses ensemble learning methods for classification and regression. There are two types of ensemble learning techniques viz. Bagging and Boosting techniques. Random forest is a bagging technique. The trees in random forests are run in parallel. There is no interaction between these trees while building the trees.It operates by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees.

Finally, I experimented with boosting techniques such as CatBoost, but the end rmse was almost the same, if not worse so I removed it.