Solution Sheet

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

Random Forest Regressor was used for the probability prediction.
Random forest algorithm can be used for both classifications and regression task as it provides higher accuracy and won't allow overfitting trees in the model.
It has the power to handle a large data set with higher dimensionality.

Upon performing an exploratory data analysis on both the data sets, it was observed that the Infect_Prob column in the train data did not follow a normal distribution curve with a noticeable peak at nearly 50% probability and no fruitful correlations were seen in the heatmap.

All the categorical features: Region, Gender, Designation, Married, Occupation, Public Transport, Comorbidity, Pulmonary Score and Cardiological Pressure were encoded using Label Encoder
The train dataset had quite many missing values compared to test dataset, which were filled with the most frequent value of the particular column as the categories seemed to have low correlations with the Infect_Prob.
Dropping the Name column, the datasets were left with all numeric values, which were now ready for model fitting.

Using train-test split, the train data was split and fit in the RF Regressor to find an acceptable RMSE value.
Thus, similar steps were taken to predict the Infect_Prob for the test dataset and stored into a .csv file.

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