**Solution Sheet**

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

I imported several libraries for the project

**Numpy** : To work with arrays

**pandas**: To work with csv files and dataframes

**train\_test\_split**: To split the dataset into training and testing data

**StandardScaler**: To scale all the features, so that the Machine Learning model better adapts to the dataset

I used  read\_csv() to read the dataset and save it to the data variable

**Data processing**:

To work with categorical variables, we should break each Categorical column into dummy columns

To get this done, we use the get\_dummies  method from pandas. Next, we need to scale the dataset for which we will use the StandardScaler. The  fit\_transform() method of the scaler scales the data and we update the columns.

I split dataset into 33% test data and 67% train data

I used RandomForest Classifier

This classifier takes the concept of decision trees to the next level. It creates a forest of trees where each tree is formed by a random selection of features from the total features

We make class predictions (predict) as well as predicted probabilities (predict\_proba) to calculate the ROC AUC/Accuracy. Once we have the testing predictions, we can calculate the ROC AUC

Then I calculated the probabilities for the values in Test\_dataset by training the model by randomForest

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