**Exam revision**

1- Use **Boston** dataset and consider **medv** and **chas** as the attributes of interest.

1. Use the appropriate probability models to quantify the uncertainty of two variables

1. Estimate the parameters of your proposed models using the dataset.

1. Predict the future values of variables.

2- Use the dataset **‘Boston’** and consider **chas** as the output variable and rm, age , dis, rad, as the input variables

1. Split the dataset into 80% as the train-set and 20% as the test-set. (use set.seed(1004))

1. Apply SVM algorithm to train the classifier using train-set with kernel ‘poly’. **(5 Marks)**
2. Predict the test-set using the trained model of classifier. **(2.5 Marks)**
3. Provide the confusion matrix and obtain the accuracy. **(5 Marks)**

Redo parts b-d to apply either Naïve Bayes or Decision Tree. Which model does provide the higher accuracy?

3- Use the dataset **‘Boston’** and consider **medv** as the output variable and rm, age , dis, rad, as the input variables,

1. Based on the output variable, which type of GLM is proposed for this analysis.

1. Specify the significant independent variables on the response variable at the level of 𝛼 = 0.05.
2. Using the output, find the optimal predictive model for the response variable.

4- Use **medv** as your time series variable:

1. Validate the assumptions using graphical visualization.

1. Fit the optimized ARIMA model for ***medv*** and provide the coefficient estimates for the fitted model.
2. What is the estimated order for AR and MA?

1. Forecast h=10 step ahead prediction of ***medv*** on the plot of the original time series.

5- Use **Boston** dataset,

1. Use LDA to classify the dataset into few classes so that at least 85% of information of dataset is explained through new classification. (**Hint**: consider the **rad** as the output variable, and tax ptratio black lstat medv as input variables. ) How many LDs do you choose? Explain the reason.
2. Apply PCA to input variables, and identify the important principle components involving at least 90% of dataset variation. Explain your decision strategy? Plot principle components versus their variance
3. Use K-means clustering analysis to input variables and identify the most important classes. How many classes do you select? Why?

1. Split the dataset into two sets of variables so that **X**=( zn, indus, chas, nox) and **Y**=( dis, rad, tax). Apply canonical correlation analysis to find the cross-correlation between **X** and **Y**. What is the correlation between ***indus*** and ***tax***?