

# Data Science

A stylized illustration in a flat, modern style. In the center, a man in a blue shirt and blue pants stands with his back to the viewer, looking at a large circular dashboard with various charts and data points. To his right, a woman in a blue shirt and red pants stands, pointing at a bar chart on a screen. The background is filled with various data-related icons: a cloud with a list, a pie chart, a line graph with data points labeled 20, 30, 45, 50, and 65, a magnifying glass, and a computer monitor displaying a dashboard. The overall color palette is muted, with blues, reds, and greys.

Lab Session 2

# Have Google Colab up and running



# Lab Assignment

- Use a supervised learning approach to build a predictive model where the output is “MEDV - Median value of owner-occupied homes in \$1000's”
- Dataset: [Boston Dataset](#)



# Dataset

- CRIM per capita crime rate by town
- ZN proportion of residential land zoned for lots over 25,000 sq.ft.
- INDUS proportion of non-retail business acres per town
- CHAS Charles River dummy variable (= 1 if tract bounds river; 0 otherwise)
- NOX nitric oxides concentration (parts per 10 million)
- RM average number of rooms per dwelling
- AGE proportion of owner-occupied units built prior to 1940
- DIS weighted distances to five Boston employment centres
- RAD index of accessibility to radial highways
- TAX full-value property-tax rate per \$10,000
- PTRATIO pupil-teacher ratio by town
- LSTAT % lower status of the population
- MEDV Median value of owner-occupied homes in \$1000's

# Lab assignment

---

**Step 0 – Visualization and EDA**

**Step 1 – Data Processing and choosing the supervised learning approach**

**Step 2 – Training & Test Dataset**

**Step 3 – Train the Model**

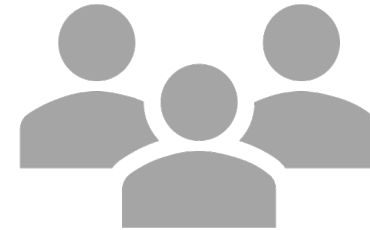
**Step 4 – Performance Evaluation**

# Step 0 – Visualising the Dataset & EDA

---



Finding the solution – 20 min



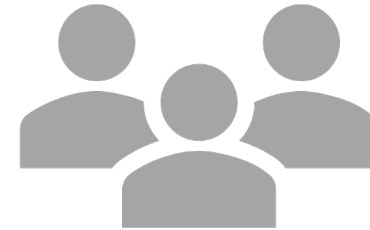
Team presentation – 10 min

# Step I – Data Processing & Model Choice

---



Finding the solution – 20 min



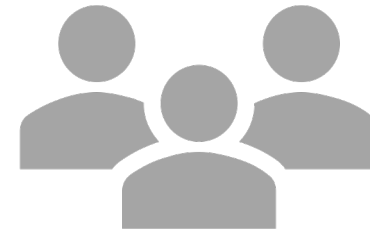
Team presentation – 10 min

# Step 2 – Training and test data set

---



Finding the solution – 10 min



Team presentation – 10 min

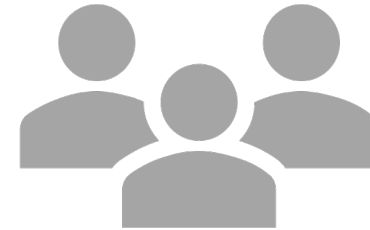


# Step 3 – Train the model

---



Finding the solution – 20 min



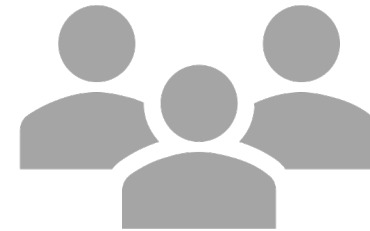
Team presentation – 10 min

# Step 4 – Performance Evaluation

---



Finding the solution – 10 min



Team presentation – 10 min



What is next?

# Compare Results

---

- Choose another supervised learning method and perform the training.
- Discuss results and compare with the original method you chose.
- What are the pros and cons?
- Which one would you choose overall?
- Briefly discuss your analysis and overall results.