# CS F320 - Foundations of Data Science

## Assignment 2

**Submission Time & Date: 1659hrs on 6th Dec 2021**
**Max Marks:20**

**General Instructions:**

This assignment is a coding project and is expected to be done in groups. Each group can contain at most three members. Make sure that all members in the group are registered to this course.

This assignment is expected to be done in Python using standard libraries like NumPy, Pandas and Matplotlib. You can use Jupyter Notebook. No other ML library like scikit/sklearn, TensorFlow, Torch etc. should be used.

Refrain from directly copying codes/snippets from other groups or the internet as all codes will be put through a plagiarism check.

All deliverable items (ex. .py files, .ipynb files, reports, images) should be put together in a single .zip file. Rename this file as A2_<id-of-first-member>_<id-of-second-member>_<id-of-third-member> before submission.

Submit the zip file on or before the aforementioned deadline. Please note that this is a hard deadline and no extensions/exemptions will be given. The demos for this assignment will be held on a later date which shall be conveyed to you by the IC.

All group members are expected to be present during the demo.

**Problem statement**

In this assignment, your aim is to use linear regression to predict house prices using the other 13 attributes (bedrooms, bathrooms, sqft_living etc.) of the dataset attached below. In order to achieve the same, you required to

1. Perform a thorough pre-processing of the dataset. The techniques you are expected to use include (but are not limited to) standardization, normalization, detecting outliers and handling missing values.
2. Perform greedy forward feature selection to find the subset of features that provide the optimal regression model. Find the minimum training and testing error of the optimal model obtained.

3. Perform greedy backward feature selection to find the subset of features that provide the optimal regression model. Find the minimum training and testing error of the optimal model obtained.

Try to write a clean, modularized and vectorized code which can solve the above problem. Please refrain from hardcoding any part of your code, unless it is absolutely necessary.

**What needs to be documented in your report:**

Give a brief description of your model, algorithms and how you implemented greedy forward and greedy backward feature selection.

Specify the subset of features that provide the optimal model for

1. Greedy forward feature selection
2. Greedy backward feature selection

Tabulate the minimum training and testing error obtained from

1. Greedy forward feature selection
2. Greedy backward feature selection
3. Linear regression model without any pre-processing and feature selection.

Link to the dataset : https://drive.google.com/file/d/1dXtp4aAJxehLQRVx-lQoSfDClf2cnmGB/view?usp=sharing

***Whom to contact for queries:***

Please contact Mr.Achyuta Krishna V (f20180165@hyderabad.bits-pilani.ac.in) for any queries.