Assignment 1

In this assignment you will be implementing your own decision trees and random forest for regression problem from scratch.

- Build a **random forest** that consists of multiple **decision trees (for regression)** from the given training data set (Housing Price dataset). Then, apply it on the test set and submit your code to generate predictions.
    - You need to build the random forest and decision trees **from scratch**. (I.e., it is not allowed to use existing machine learning libraries or packages such as sklearn.)
    - You may use any programming language/environment of your choice, but you are required to submit the complete source code to produce the output
        - If you use anything other than jupyter notebook, submit an executable and run that from the main function of the jupyter notebook so that the prediction generation is automated. We can provide assistance with this.
    - The output (a single file with the predictions for each test instance) **must be generated automatically using the approach implemented by you**. Submitting predictions/code from any other source (Internet, another student, etc.) is considered cheating and will result in immediate disqualification (i.e., dismissal from the course).
    - Deadline for the assignment is **02.02.2022 by 23:59**.
    - **Can be done individually or in a group of two.**
    - You are supposed to upload your coding + a short report that also presents the output as well. You will upload in in Canvas
    - The day later, February 3rd at 12:15 you are asked to present in Lab sessions to explain the code you have implemented.
- Dataset:
    - The dataset is for housing price prediction. Training and Testing data set are given
        - The goal is predicting the price of a house given its attributes
        - So, it is a regression problem
        - Therefore, your random forest should be able to predict a value (housing price) rather than a class
        - Use appropriate splitting criterion and error function

        - The dataset has a lot of missing values denoted as NaN you may replace them with appropriate categorical value like None or mean/mode appropriately.
        - At the leaf node you may use the average value as the predicted value or if there are many instances you may use additional regression models like linear regression.
    - The performance is evaluated using RMSE Root Mean Square Error