DSAI 3202 – Parallel and distributed computing

Lab – 3 Part 1: Data Parallel Model

# Objectives:

* Build a parallel program to enhance the training of a machine learning model.
* Learn when to use thread and processes.

# Tasks

## Download and read the data.

1. Download the file “housing\_prices\_data.zip”
2. Copy the file top the remote (*host*) machine:
   1. Open the terminal (*Do not connect the host yet.*)
   2. In the terminal, go to the folder where the file is downloaded.
   3. Copy it to the host machine using scp:

scp housing\_prices\_data.zip student@10.102.10.XX:/home/student/

*Do not forget to correct XX to your address.*

1. Repeat the same process for the notebook file “ModelingWithRandomForests.ipynb”.
2. Connect to the host machine and copy the folder to your repository:
   1. Use ssh to connect to your host machine.
   2. Once connected, copy the downloaded files from the home student to your repository. You can see the list of files using ls.

(base) student@dsai3203-template:~$ ls

anaconda3 examples.desktop ModelingWithRandomForests.ipynb

Anaconda3-2023.09-0-Linux-x86\_64.sh housing\_prices\_data.zip parallel-and-distributed-computing-dsai3202

*This is my machine, yours might look slightly different.*

* 1. Move these files into your repository (The name of my repository is *parallel-and-distributed-computing-dsai3202,* yours might be different).

(base) student@dsai3203-template:~$ mv ModelingWithRandomForests.ipynb parallel-and-distributed-computing-dsai3202/

(base) student@dsai3203-template:~$ mv housing\_prices\_data.zip parallel-and-distributed-computing-dsai3202/

* 1. Move into your repository using cd.
  2. Move the notebook to the notebooks folder using mv.
  3. Create a data folder using mkdir.

(base) student@dsai3203-template:~/parallel-and-distributed-computing-dsai3202$ mkdir data

*You can check that it was created using ls.*

1. Move the “housing\_prices\_data.zip” to the data folder using mv.
2. Go into the data folder using *cd*.
3. Unzip the “housing\_prices\_data.zip”, using *unzip*.

(base) student@dsai3203-template:~/parallel-and-distributed-computing-dsai3202/data$ unzip housing\_prices\_data.zip

1. Delete the zip after extraction, using *del*.

(base) student@dsai3203-template:~/parallel-and-distributed-computing-dsai3202/data$ rm housing\_prices\_data.zip

(base) student@dsai3203-template:~/parallel-and-distributed-computing-dsai3202/data$ ls

housing\_prices\_data

## Test the machine learning program.

1. In Visual Studio code, connect to your host machine—if you have not done so already—and in your repository open the Jupyter file you have just downloaded.
2. Read the file and test the first machine learning program.

## Sequential search for the best parameters

1. Run the program for seeking the best parameters.
2. Time its execution.

## Parallelize with Threading and processes:

1. Modify your program to use the threading module to parallelize the seeking of the best parameters.
2. Modify your program to use the multiprocessing module to parallelize the seeking of the best parameters.

## Questions:

1. How does the execution time change when moving from sequential to threaded to multiprocessing implementations?
2. Compute the performance metrics for the threaded and multiprocessing executions.