**DSA5208 Project 1: Kernel Ridge Regression**

**Objective:** Implement and evaluate kernel ridge regression (KRR) with different kernels for predicting house price in California

**Dataset:** California Housing Dataset

**Data Pre-processing:**

* Standardized input features
  + Ensured input features is scaled to a range of [0,1] for faster convergence, using MinMaxScalar
  + Input features includes all columns except the target feature (`medianHouseValue`)
* Split data into training and testing sets (70:30 ratio)
  + Used random state 2 to split 70:30 between training and testing sets

**Method:**

* The Gaussian Kernel was used
* MPI implementation was written using mpi4py library
* Hyperparameter tuning of s and lambda, the smoothness and regularization parameters respectively, was conducted

**Steps:**

Initialisation of MPI

* MPI Communicator: `comm = MPI.COMM\_WORLD`
* Rank: `rank = comm.Get\_rank()` # identifies the process ID
* Size: `size = comm.Get\_size()` # number of processes

Data Initialisation and Distribution

* Broadcast all training and test data to all the processes using comm.bcast

Distributed computation of kernel matrix

Conjugate gradient method

Prediction and test using root mean squared error (RMSE)

Hyperparameter tuning

* Coarse grain tuning: utilise a large range of [1e-6, 1e-5, 1e-4, 1e-3, 1e-2, 1e-1] for s and [1e-5, 1e-4, 1e-3, 1e-2, 1e-1, 1, 10, 100] for lambda
* Result: Best s 1.00e-01, Best lambda 1.00e-01, Best RMSE\_test 59770.2649
* Hence, proceeded with fine grain tuning, utilising a range of XXX
* Result: Best s 1.20e-01, Best lambda 1.10e-01, **Best RMSE\_test 56805.2882**