

**CE 5319 Machine Learning for Civil Engineers**  
**Exercise Set 8**  
**Feature Reduction and Distance Metric in KNN Regression**

**Exercises**

1. Using the implementation of `KNeighborsRegressor` on the `Solids_in_Rivers` database, complete the following tasks:<sup>1</sup>
  - (a) Use the `permutation_importance` function from `sklearn.inspection` to estimate the importance of each feature in your dataset.
  - (b) Identify the top 5 or 6 most important features based on the permutation importance scores.
  - (c) Retrain the `KNeighborsRegressor` model using only the reduced feature set. Keep the same number of neighbors as in your full model.
  - (d) Compare the performance of the reduced model to the full model using at least the following metrics:
    - Coefficient of determination ( $R^2$  score)
    - Root mean square error (RMSE)
  - (e) For a fixed input vector (you may use the one from the earlier example), show the estimate computed using both models (full and reduced). Are they close? Provide a short explanation.
  - (f) Repeat the modeling using two different distance powers:  $p = 1$  (Manhattan) and  $p = 2$  (Euclidean). Comment on how the distance metric affects:
    - Feature importance rankings
    - Estimated values
    - Model performance

---

<sup>1</sup>This exercise is to develop intuition around (a) which features matter most in KNN regression and (b) how the choice of distance metric can influence model behavior.