During this quarter, I explored active learning strategies and sampling methods with a focus on greedy sampling techniques.

Topics Covered:

• Active Learning Fundamentals: Using unlabeled data and iterative labeling to refine model performance until improvements in RMSE plateau.

• Greedy Sampling Methods:

- GSx: Selecting samples based solely on input feature distances. This method computes the shortest distance from each unlabeled sample to the existing labeled set and then selects the sample with the largest distance. It is independent of the regression model, which makes it computationally efficient.

- GSy: Choosing samples by comparing predicted and actual values. In this approach, a regression model predicts values for the unlabeled samples, and the method selects the sample where the difference between the predicted and true value is greatest. This process requires updating the regression model at each iteration, which adds computational complexity.

- IGS: Integrating both GSx and GSy through distance matrix operations. IGS combines the strengths of both methods by computing separate distance matrices for the input features (GSx) and the prediction errors (GSy), and then integrating these matrices.

Special Project Summary:

I implemented a custom variant of these methods by incorporating average distance and standardization. Instead of multiplying distances as done in traditional iGS, I first applied a standardization process using z-scores to normalize both the input feature distances and the differences between predicted and actual outputs. This normalization ensures that both components contribute equally to the decision-making process when selecting the most informative samples. After standardizing, I computed the average distance from each point rather than relying solely on the minimum distance, thereby capturing a broader perspective of the data distribution.

This approach aimed to leverage the benefits of both input space variability and output prediction errors, resulting in a more balanced and informative sample selection strategy. Experimental results demonstrated that the standardized iGS and iGS with average distance variants outperformed the traditional GSx approach as well as the original iGS method. Notably, these variants provided a better trade-off between computational cost and predictive performance, as evidenced by improved error metrics and faster convergence rates in our active learning experiments.

Reflections and Future Work:

This project deepened my understanding of active learning and highlighted the importance of combining theoretical insights with practical algorithm development. Future efforts will explore alternative regression models such as random forests and xgboost, and further test these methods on diverse datasets.