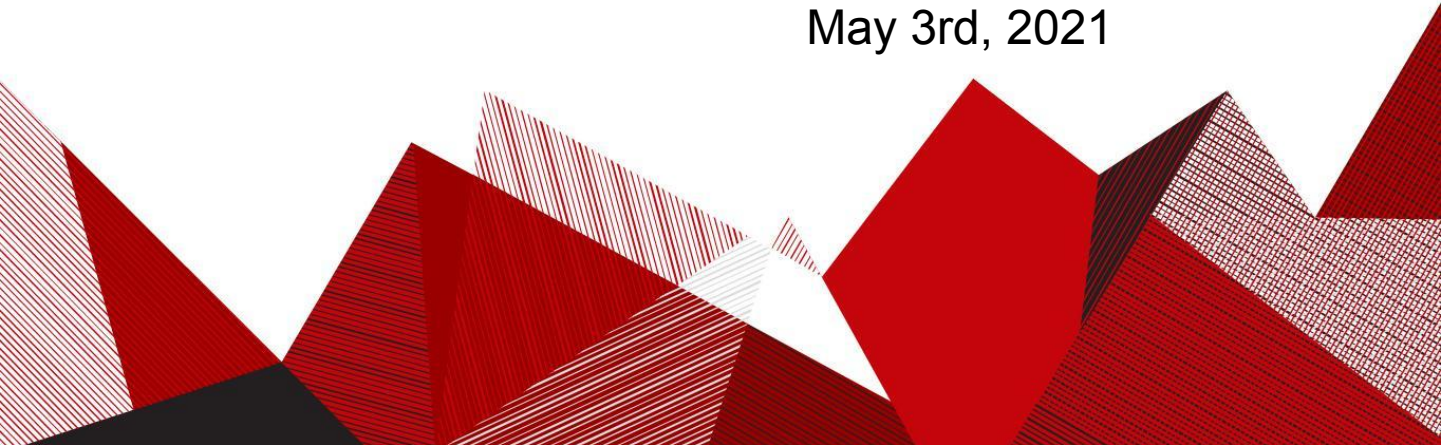
Machine Learning Domain and Error Analysis

Skunkworks Informatics (MSE 299)

Final Report

May 3rd, 2021

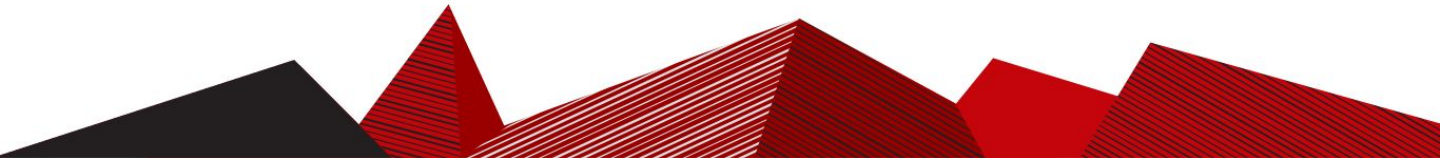


Gautam Agarwal

Highlights

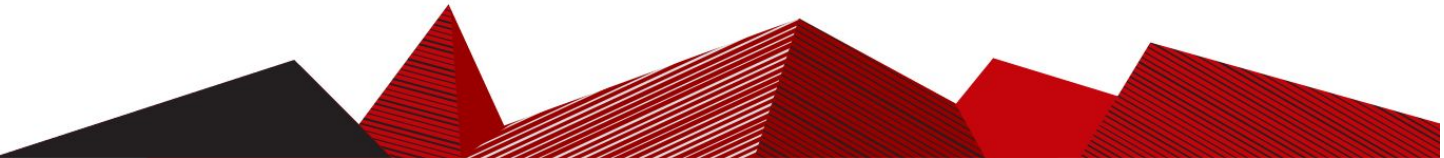
- Worked with Domain and Error Analysis Group
- Mentored by Professor Dane Morgan
- Adapted MastML, lolopy
- Created code repository:

  https://github.com/GAInTheHouse/domain-error



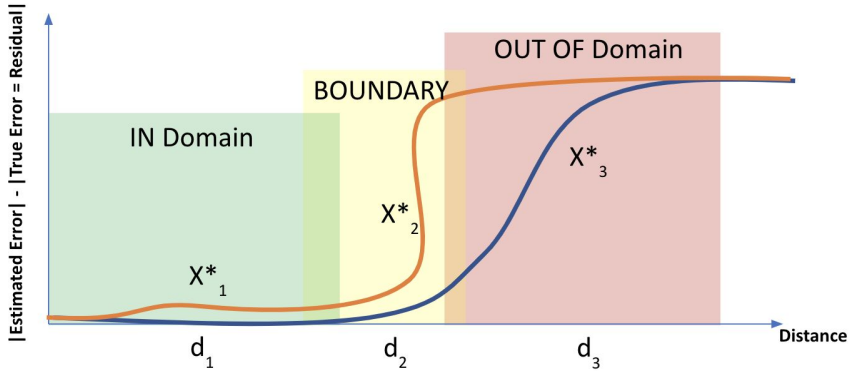
Work Errors

Week 2: 12 hours
Week 3: 11 hours
Week 4: 16 hours
Week 5: 10 hours
Week 6: 11 hours
Week 7: 9 hours
Week 8: 9 hours
Week 9: 11 hours
Week 10: 9 hours
Week 11: 13 hours
Week 12: 8 hours
Week 13: 10 hours
Week 14: 14 hours



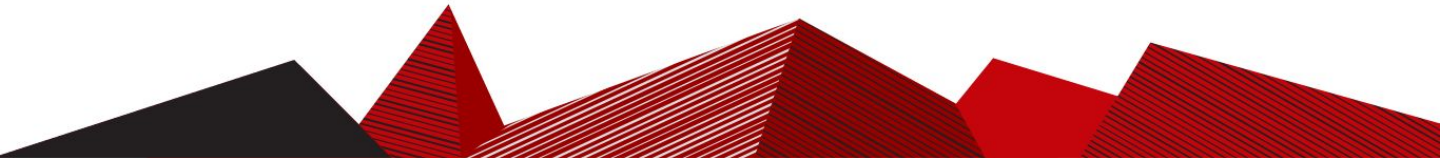
Problem

- Identifying Domain of Machine Learning Data
- Observing error in error metric



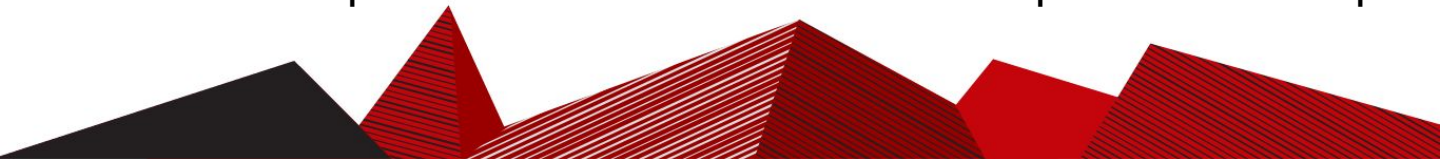
Datasets Used

- Friedman Data
- Diffusion Data
- Superconductor Data



Data Split Techniques Observed

- General Techniques
  - Clustering the data
- Dataset-specific Techniques
  - Friedman’s Data: Based on first column
  - Diffusion Data: Based on element
  - Superconductor Data: Based on cuprates / non-cuprates

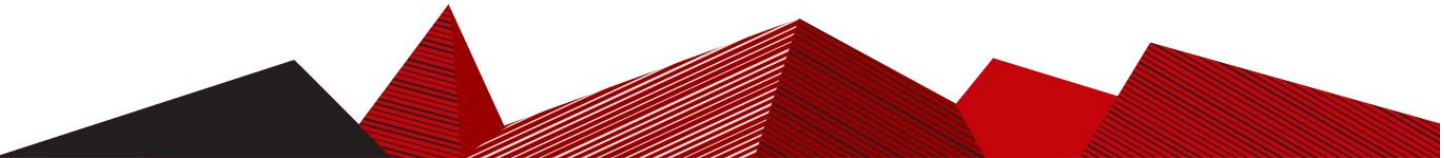


Error/ Error Analysis Metrics Used

- Error Metric
  - Residual Error: Predicted - Actual
  - Model Error: σ(Predicted)
  - Error Bar Length = 2* 1.96 * σ

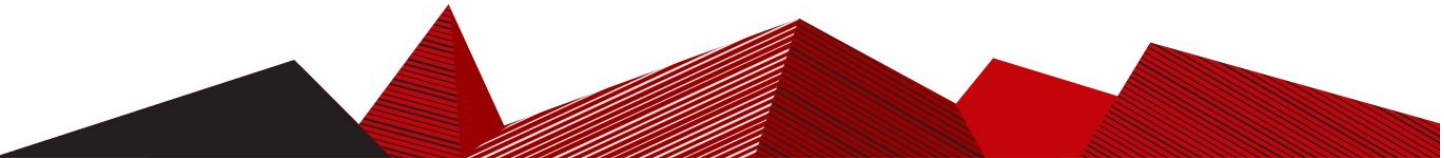
- Error in Error
  - Residual Error - Model Error
  - Residual Error / Model Error

- Error Analysis Metrics also varied in their Modulus Operands



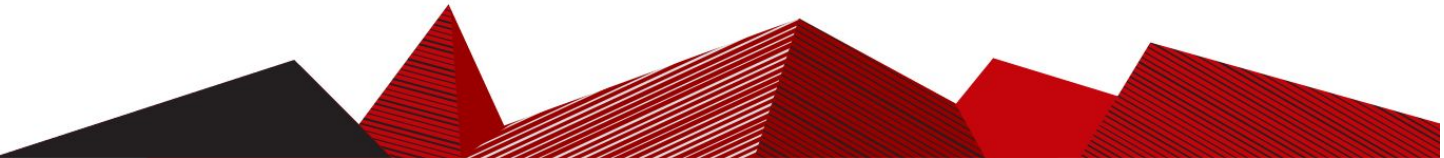
Distance Metrics Used

- Euclidean Distance
  - Average distance between a test point and each training point
  - Distance between a test point and nearest training point
- Mahalanobis Distance



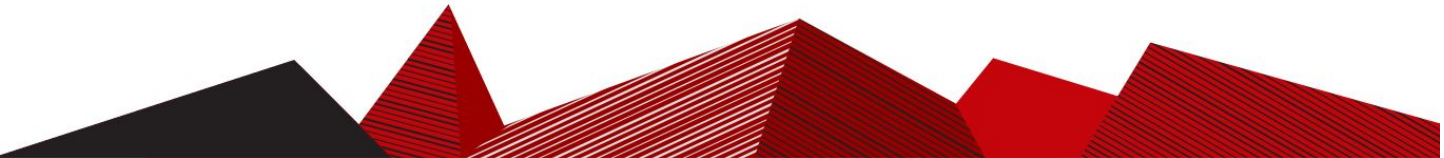
Pre-Processing

- Split the dataset into training and testing set.
- Identify out-of domain points in test data set



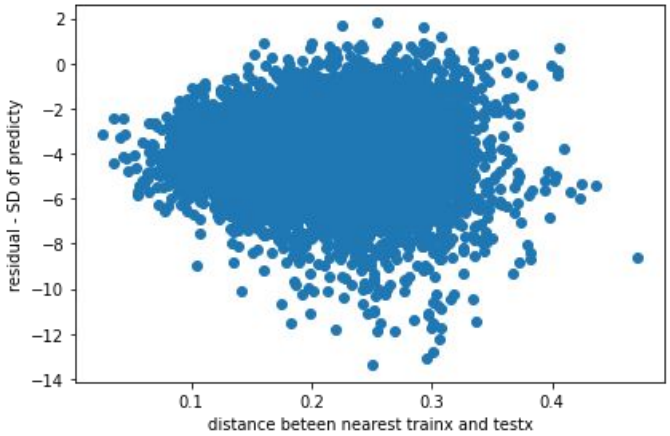
Procedure 1

- Plot Error in the error against distance in the test data
- Expectation: Error in error becomes ambiguous after some distance
- Goal:
  - Find a good error analysis metric
  - Briefly analyze behavior to different data metrics

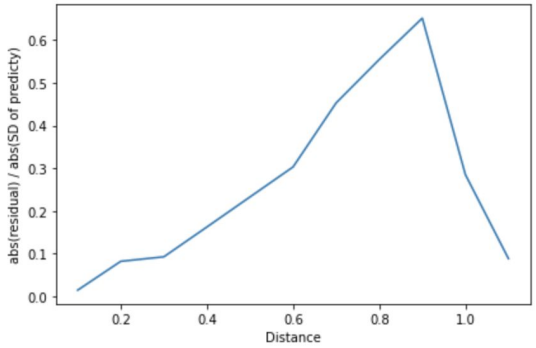


Observations

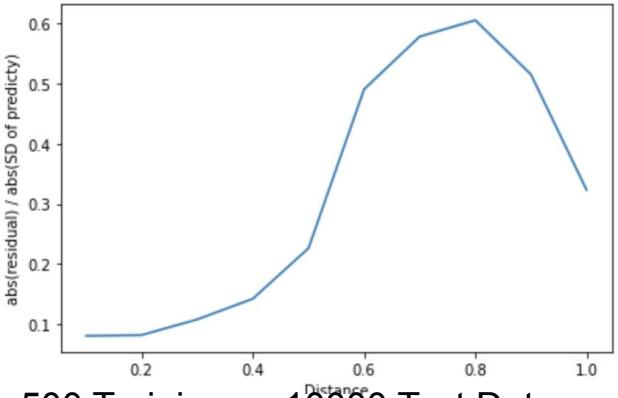
- Data: Friedman’s data
- Distance: Nearest Euclidean Distance between a training point and each test point
- Clustering type: Random
- Model: Random Forest
- Error in Error: Residual - SD(predicted)
- Each Point: Test Data Point



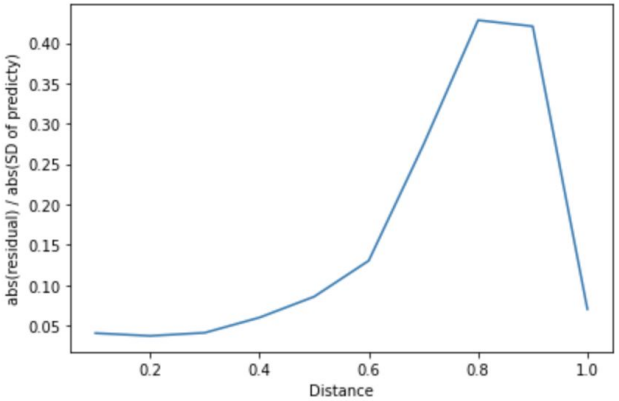
- Data: Friedman's Data (Split: Random)
- Distance: Euclidean distance of test data from average of training data point
- Model: Random Forest
- Error in Error: |Residual Error / Predicted Error|, where Predicted Error is the standard deviation of predicted y on test data x
- Each Line: Highest value of error in error for every bin of distance



500 Training vs 1000 Test Data

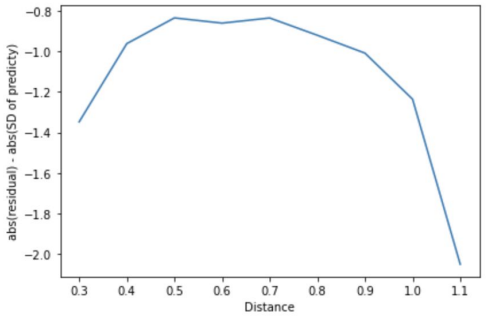


500 Training vs 10000 Test Data

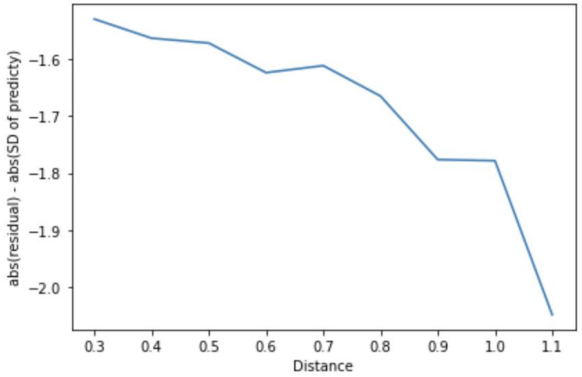


10000 Training & Testing Data

- Data: Friedman’s Data (Split: Based on Clusters)
- Distance: Euclidean distance of test data from average of training data point
- Model: Random Forest
- Error in Error: |Residual Error| -|Predicted Error|, where Predicted Error is the standard deviation of predicted y on test data x
- Each Line: Highest and average values of error in error for every bin of distance

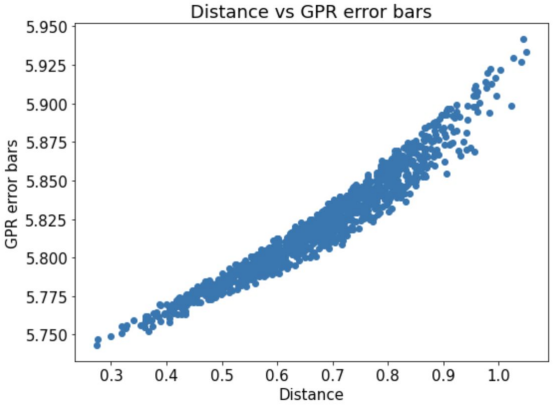


Highest Error in Error

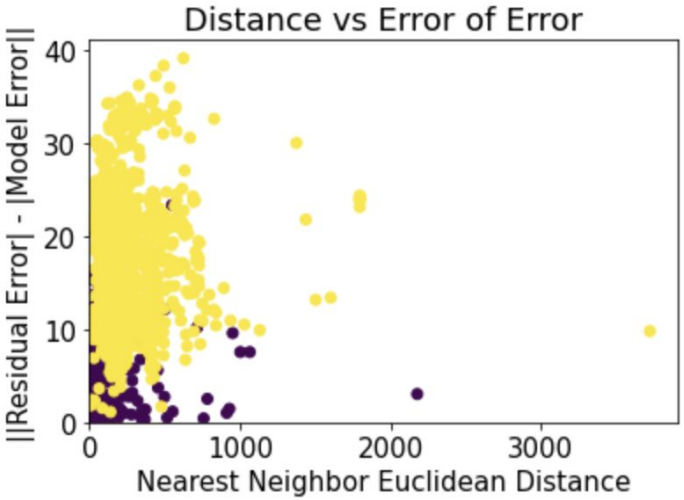


Average Error in Error

- Data: Friedman's Data (Split: Based on Clusters)
- Distance: Euclidean distance of test data point from the nearest training data
- Error Bar: 2*1.96*σ
- Each Point: Testing Data
- Model: GPR

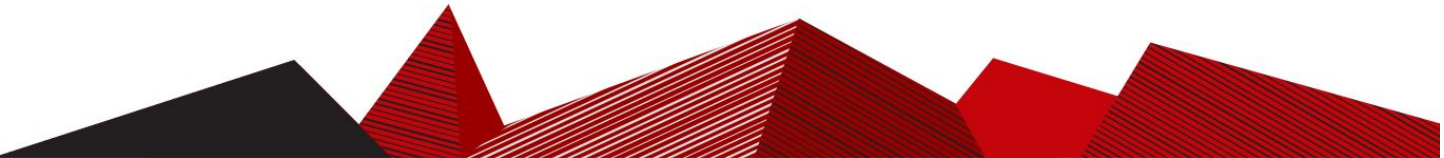


- Data: Superconductor Data (Split: Based on Domain)
- Distance: Euclidean distance of test data point from the nearest training data
- Error in Error: | |Residual Error| -|Predicted Error| |, where Predicted Error is the standard deviation of predicted y on test data x
- Each Point: Testing Data
- Model: Random Forest



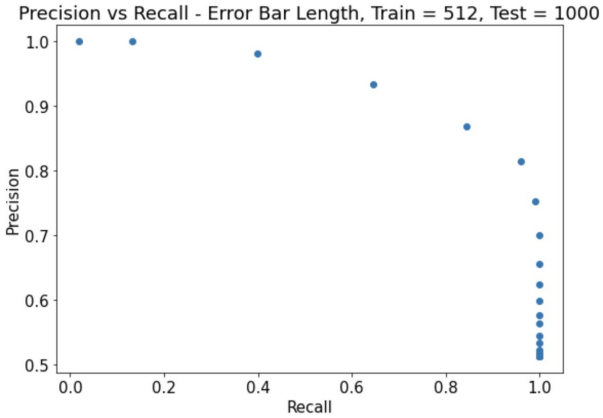
Procedure 2

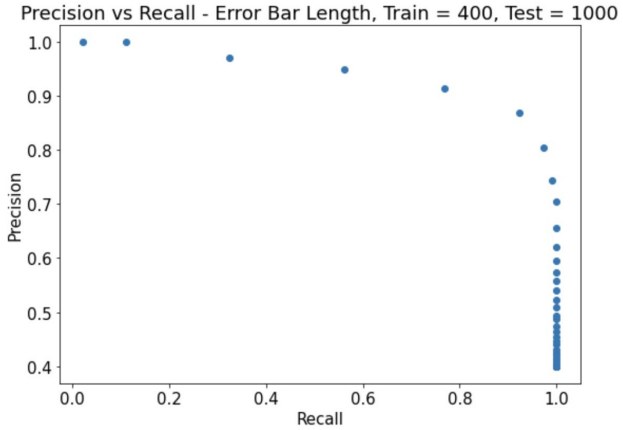
- Plot Precision/ Recall based on:
  - Domain, &
  - Either Distance, Error Bar Length or Error in Error
- Goal:
  - Briefly analyze behavior to different distance metrics
  - Find a distance metric giving consistent results across different datasets



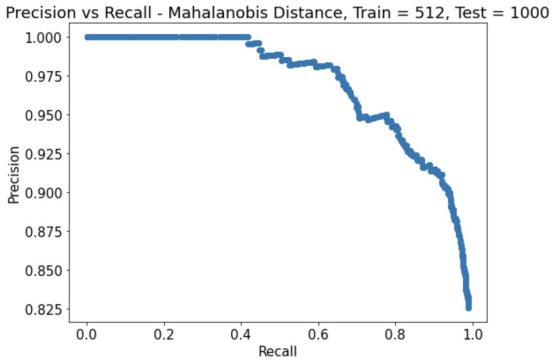
Observations 2

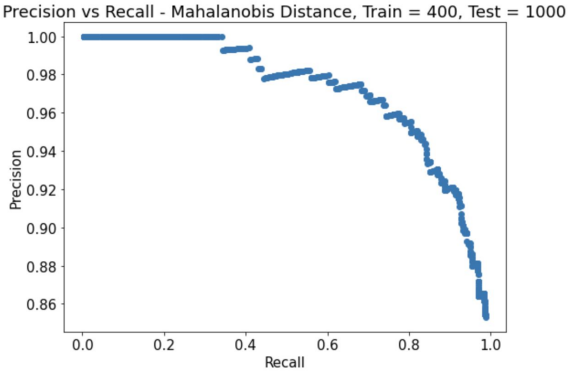
- Data: Friedman's Data (Split: Based on Clusters)
- Threshold Criteria: Error Bar Length (2*1.96*σ)
- Each Point: Testing Data



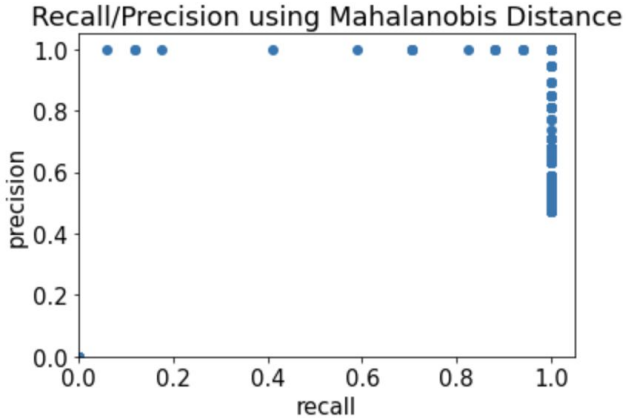


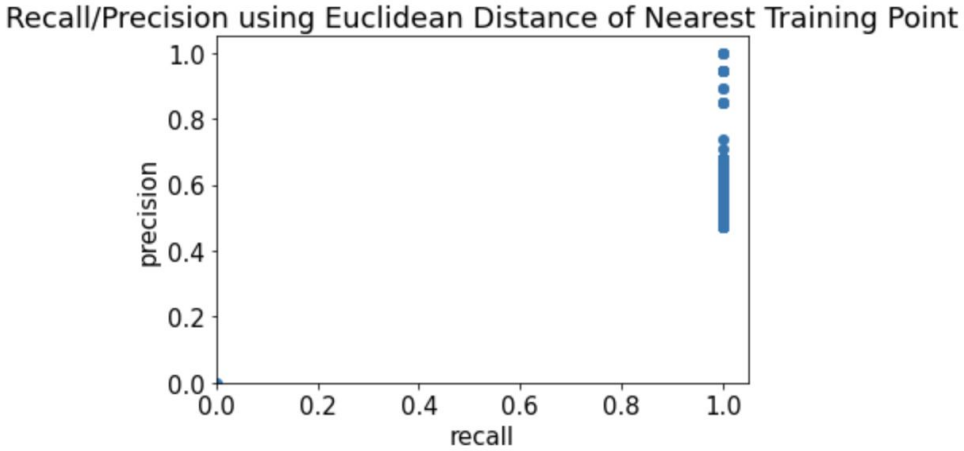
- Data: Friedman’s Data (Split: Based on Clusters)
- Threshold Criteria: Mahalanobis Distance
- Each Point: Testing Data



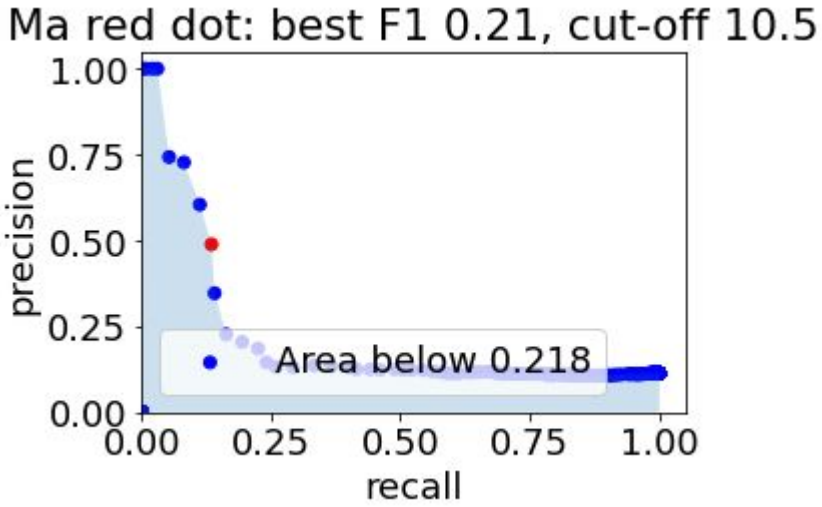


- Data: Diffusion Data (Split: Based on elements)
- Threshold Criteria: Mahalanobis Distance & Euclidean Distance
- Each Point: Testing Data





- Data: Superconductor Data (Split: Based on elements)
- Threshold Criteria: Mahalanobis Distance
- Each Point: Testing Data

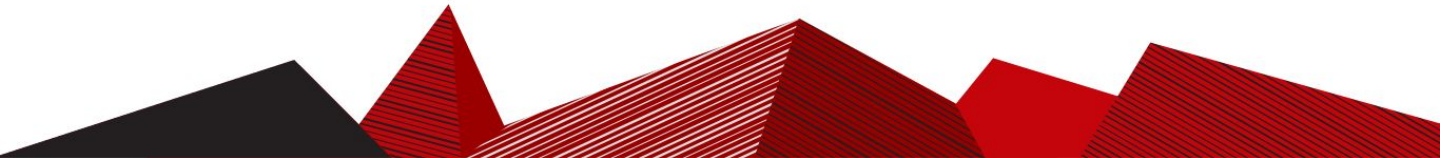


Other Key Takeaways

- Proposal Making for Hilldale Fellowship
    - Result: Not-selected
    - Takeaway: Proposal Writing, and explanation to general audience

Summer Goals

Apply standard scaling, normalization before using Mahalanobis



Thank You !!!