Appendix: Detailed Results

Dauda Sheni[1][0000-0002-4634-2538], Anton Basson[1][0000-0001-7998-330X], Jacomine Grobler[2][0000-0002-1868-0759]

1. Introduction

This document serves as an appendix to the paper titled "Evaluating human-in-the-loop machine-learning (HITL-ML) querying strategies for missing value imputation (MVI)" [1]. This appendix includes the results from the model-feature hyperparameter grid search, the summarised training and cross-validation results, as well as the results from the two-sided Mann-Whitney U test. The decision tree-based active learners emerge as the best-performing, demonstrating efficient learning and generalization across all features. The query-by-committee (QBC) algorithm generally outperforms other strategies, improving the performance of weak learners such as the MLP.

2. Grid search results

Hyperparameters are selected following a ten-fold cross-validation grid search for each model-feature combination. Grid searches are conducted on all available data, withholding data used for the holdout sets. The results for the decision tree, k-NN, and MLP algorithms are shown in Tables 10, 11, and 12, respectively.

Table 1: Decision Tree Grid Search Results

Dataset	Feature	Problem Type	Criterion	Max Depth	Min Samples Leaf	Splitter	Metric	Best Score
HVB	Material Weight	Regression	squared_error	10	1	best	R2	0.999998
	Material Part Number	Regression	squared_error	10	1	best	R2	0.985151
	Minimum Capacity	Regression	squared_error	15	1	random	R2	0.999973
	Capacity Throughput	Regression	absolute_error	20	1	best	R2	0.995930
	Voltage	Regression	friedman_mse	7	32	best	R2	0.752215
	Model Code	Classification	gini	10	1	random	Accuracy	0.999773
Red wine-quality	quality	Classification	gini	15	1	best	Accuracy	0.6219

[1] Sheni, D.N., Basson, A.H., Grobler, J.: Evaluating human-in-the-loop machine-learning (HITL-ML) querying strategies for missing value imputation (MVI). (2025). Submitted to Elsevier Knowledge-based Systems 2025.

Abalone	Rings	Regression	squared_error	10	32	best	R2	0.5951

For the k-NN algorithm (Table 11), the cosine distance metric is selected for most features and shows significantly better training performance than other distance metrics in the conducted experiments.

Table 2: k-NN Grid Search Results

Dataset	**Feature**	**Problem Type**	**Metric (Distance)**	**Neighbours**	**Weights**	**Metric**	**Best Score**
HVB	Material Weight	Regression	cosine	3	distance	R2	0.995242
	Material Part Number	Regression	cosine	3	distance	R2	0.920068
	Minimum Capacity	Regression	cosine	3	distance	R2	0.994859
	Capacity Throughput	Regression	cosine	3	distance	R2	0.764981
	Voltage	Regression	cosine	11	uniform	R2	0.705323
	Model Code	Classification	manhattan	3	distance	Accuracy	0.704433
Red wine-quality	quality	Classification	cosine	11	distance	Accuracy	0.6594
Abalone	Rings	Regression	cosine	15	distance	R2	0.5878

The results for the MLP algorithm are shown in Table 12. The adaptive moment estimation (Adam) solver was selected for all MLP-based active learners, as the datasets are large [53]. The maximum iterations were tuned to 2000 to balance model fit and computational effort. L2 regularisation was tested separately, and an 'alpha' value of 0.0001 was selected.

Table 3: MLP Grid Search Results

Dataset	**Feature**	**Problem Type**	**Activation**	**Hidden Layers**	**Learning Rate**	**Metric**	**Best Score**
HVB	Material Weight	Regression	logistic	(130, 50)	0.001	R2	0.905671
	Material Part Number	Regression	relu	(70, 10)	0.001	R2	0.374984
	Minimum Capacity	Regression	logistic	(100, 70)	0.001	R2	0.942476
	Capacity Throughput	Regression	identity	(150, 10)	0.01	R2	0.212695
	Voltage	Regression	identity	(90, 50)	0.01	R2	0.320200
	Model Code	Classification	logistic	(70, 10)	0.001	Accuracy	0.824967
Red wine-quality	quality	Classification	logistic	(50, 130)	0.0001	Accuracy	0.575
Abalone	Rings	Regression	logistic	(80, 30)	0.0001	R2	0.2682

3. Additional classification results

Figure 2 shows the sample performance of the three model types on the *"Quality"* feature from the red wine dataset. The decision tree learners generally outperform other model types, exhibiting robust performance. While model types such as the k-NN improve as more queries are retrieved, the performance does not reach that of the decision tree models that are trained on the initial training data. Although models such as k-NN improved as more queries were added, their performance did not reach that of decision trees trained on the initial dataset. Model performance can be improved by removing noise from the training sets, such as noise created by querying similar data.

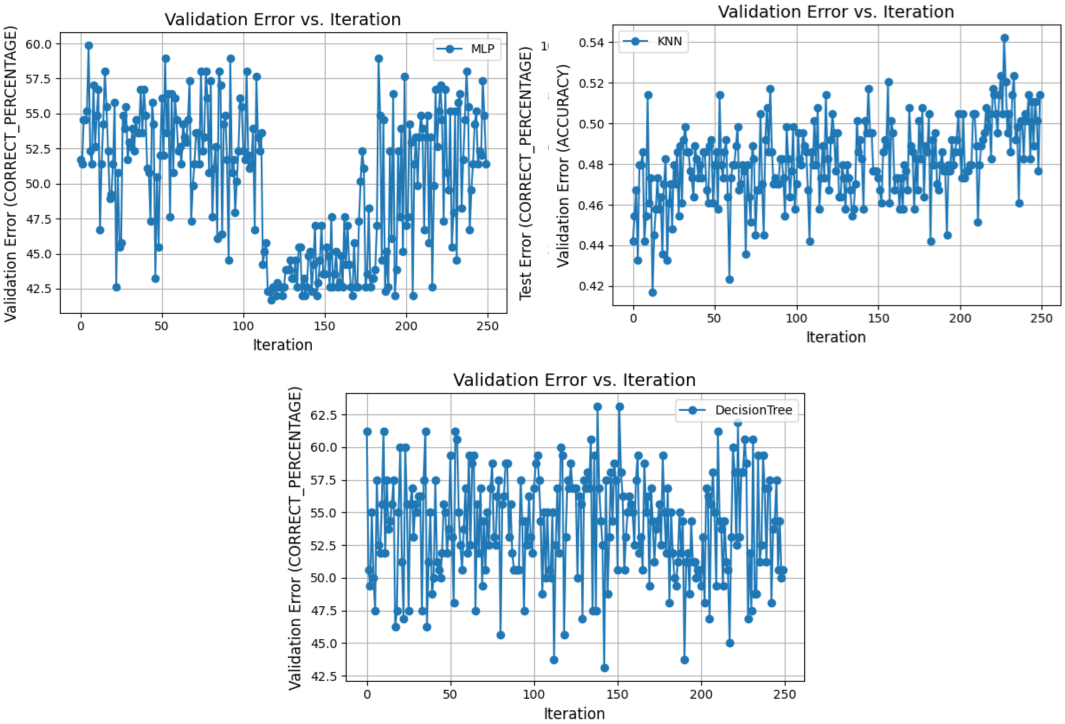

Figure 1: Sample distributions of model performance for the accuracy on the "Quality" feature, grouped by model type, with the MLP algorithm exhibiting catastrophic forgetting (top-left), the KNN algorithm showing improvement in accuracy as the active learning process continues (top-right), and the decision tree algorithm exhibiting robust performance similar to a random distribution (bottom).

3.1 Red wine – quality

The Decision tree (QBC and least confidence) learners are tied for first place based on the statistical significance. The entropy sampling yields the lowest performance across model types on this feature.

Model (Strategy)	**Total Wins**	**Total Losses**	**Total Draws**	**Performance Score**
Decision tree (QBC)	11	0	3	0.89
Decision tree (Least confidence)	11	0	3	0.89
Decision tree (Margin sampling)	10	0	4	0.86
Decision tree (Random)	10	0	4	0.86
Decision tree (Entropy sampling)	10	2	2	0.79
MLP (Random)	6	5	3	0.54
MLP (QBC)	5	5	4	0.50
MLP (Least confidence)	5	5	4	0.50
MLP (Margin sampling)	5	5	4	0.50
MLP (Entropy sampling)	5	6	3	0.46
KNN (QBC)	0	10	4	0.14
KNN (Least confidence)	0	10	4	0.14
KNN (Random)	0	10	4	0.14
KNN (Margin sampling)	0	10	4	0.14
KNN (Entropy sampling)	0	10	4	0.14

Figure 3 shows the accuracy, as a percentage, across all active learners and per model type for the *"Quality"* feature. The QBC strategy consistently has a better performance, with a better distribution and higher maximum values. The entropy sampling strategy shows the opposite, with relatively longer tails and less robust performance.

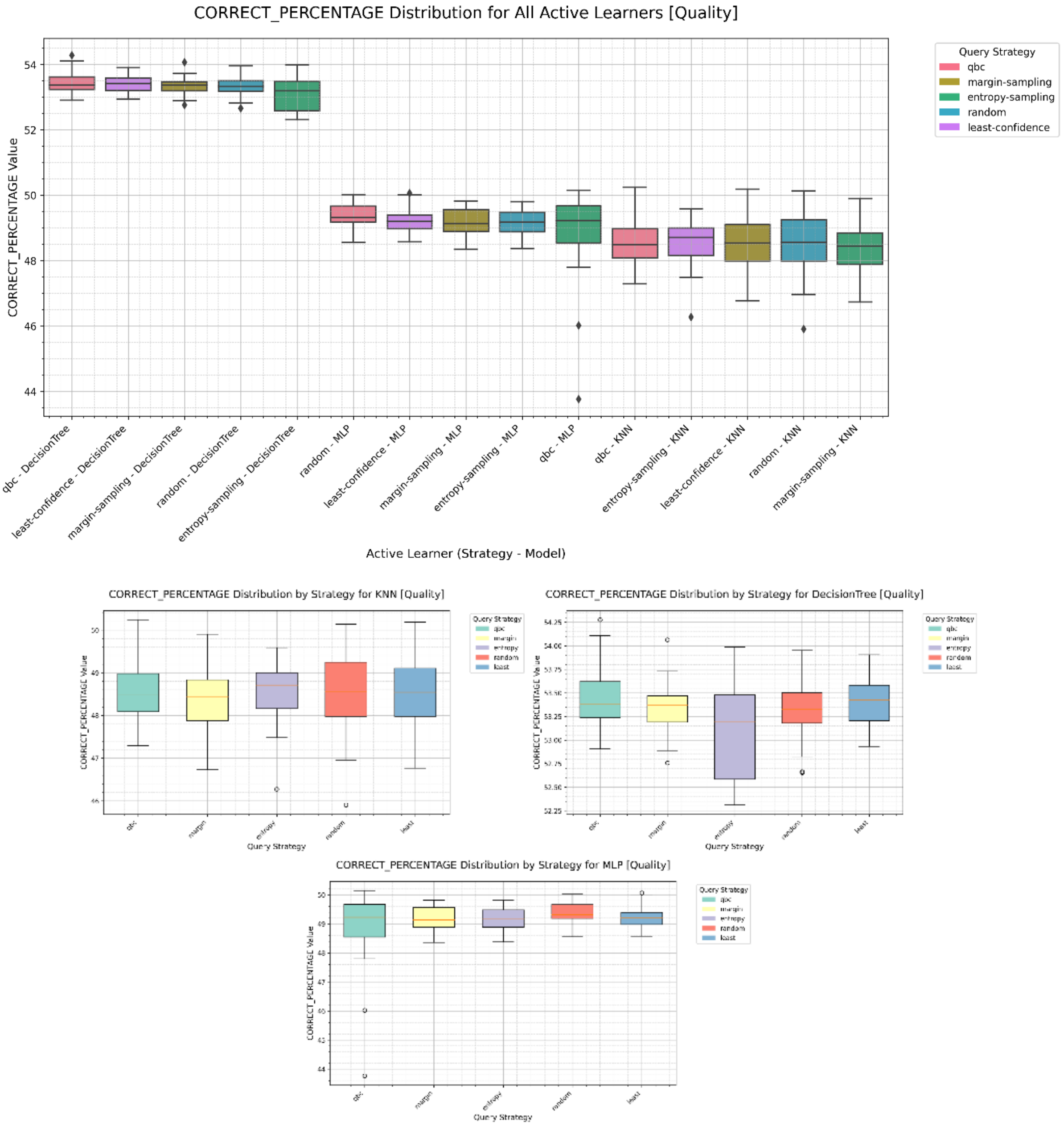

Figure 2: Accuracy results from all active learners (top) and per model type (bottom) on the Quality feature from the Red wine dataset

3.2 Model Code

The decision tree (Random) learner dominates in this dataset, being the only statistically significant result of its model type, with the QBC learners of the k-NN and MLP have slightly better performance in their specific model types, but results are not found to be statistically significant.

Table 4: Mann-Whitney U-test results for the accuracy metric on the Model Code feature from the HVB dataset

Model (Strategy)	Total Wins	Total Losses	Total Draws	Performance Score
Decision tree (Random)	11	0	3	0.89
Decision tree (QBC)	10	0	4	0.86
Decision tree (Margin sampling)	10	0	4	0.86
Decision tree (Least confidence)	10	0	4	0.86
Decision tree (Entropy sampling)	10	1	3	0.82
KNN (QBC)	5	5	4	0.5
KNN (Margin sampling)	5	5	4	0.5
KNN (Least confidence)	5	5	4	0.5
KNN (Random)	5	5	4	0.5
KNN (Entropy sampling)	5	5	4	0.5
MLP (QBC)	0	10	4	0.14
MLP (Margin sampling)	0	10	4	0.14
MLP (Least confidence)	0	10	4	0.14
MLP (Random)	0	10	4	0.14
MLP (Entropy sampling)	0	10	4	0.14

Figure 4 shows the accuracy, as a percentage, across all active learners and per model type for the *"Model Code"* feature. The random sampling seems to show more robust performance in this dataset. The QBC and least confidence performance seem to be less robust in this feature, indicating a relatively even distribution of class instances, attributed to the larger dataset size, as compared to the red wine dataset.

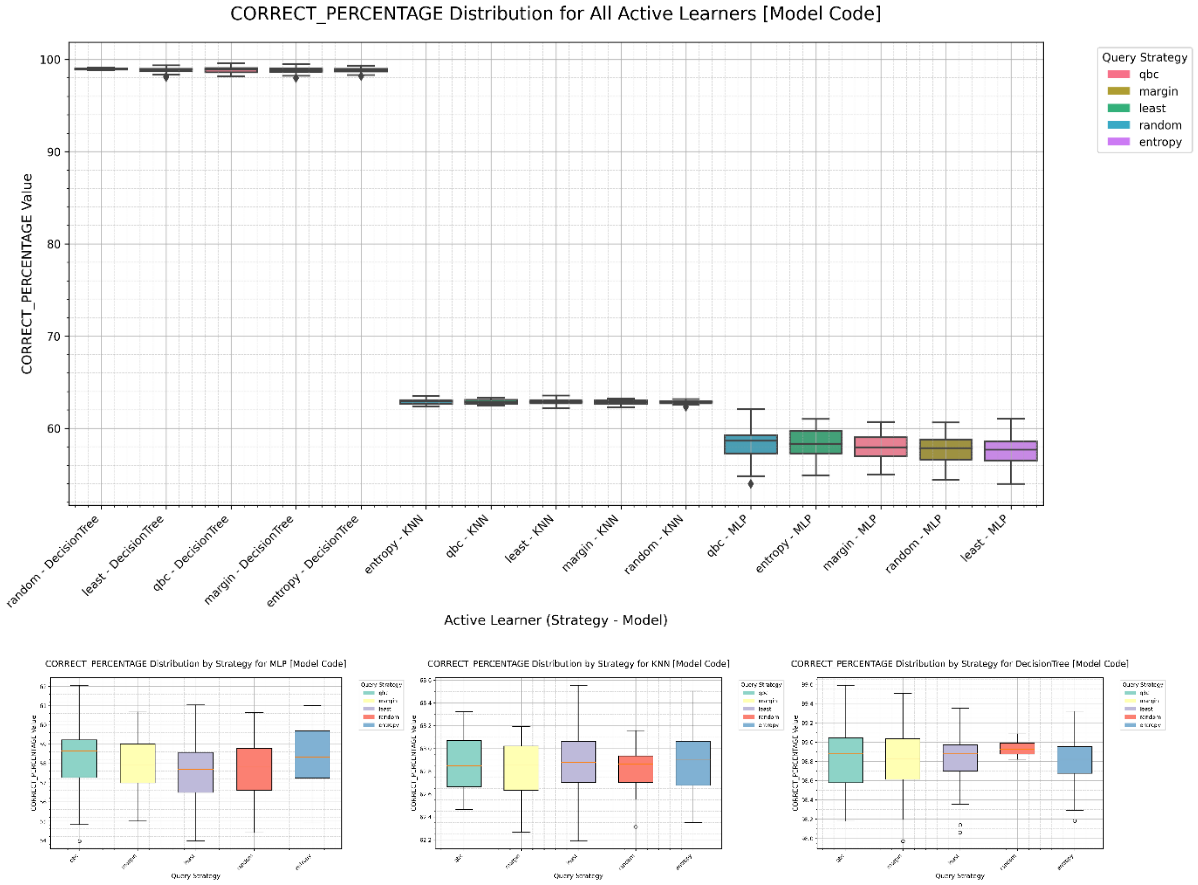

Figure 3: Accuracy results from all active learners (top) and per model type (bottom) on the Model Code feature from the HVB dataset

4. Additional regression results

On the HVB dataset, QBC outperformed random sampling across most models and features. Decision trees achieved the highest R^2 scores (≥ 0.99) and the lowest NRMSEs on *"Minimum Capacity"* and *"Capacity Throughput"* (Tables 5 and 6), with QBC again slightly outperforming random sampling. KNN also performed well, while MLPs showed limited effectiveness despite optimisation. The rankings are consistent across all metrics for each feature. Each features learner rankings, RMSE, MAE, and R^2 distributions and distributions for R^2 for each model type are included under the relevant subheading.

4.1 Abalone – Rings

Decision tree (QBC) achieved the highest performance score, winning all of its pairwise comparisons, while MLPs underperformed across both strategies. Full model rankings based on Mann-Whitney U-tests are shown below. Notable is the KNN (QBC) learner tied in second place with the decision tree (Random) learner. Figures of RMSE, MAE, and R^2 are shown in Figures 5,6, and 7, respectively.

Table 5: Mann-Whitney U-test results rankings on the Rings feature from the abalone dataset

Model (Strategy)	**Total Wins**	**Total Losses**	**Total Draws**	**Performance Score**
Decision tree (QBC)	5	0	0	1
KNN (QBC)	3	1	1	0.7
Decision tree (random)	3	1	1	0.7
KNN (Random)	2	3	0	0.4
MLP (QBC)	0	4	1	0.1
MLP (Random)	0	4	1	0.1

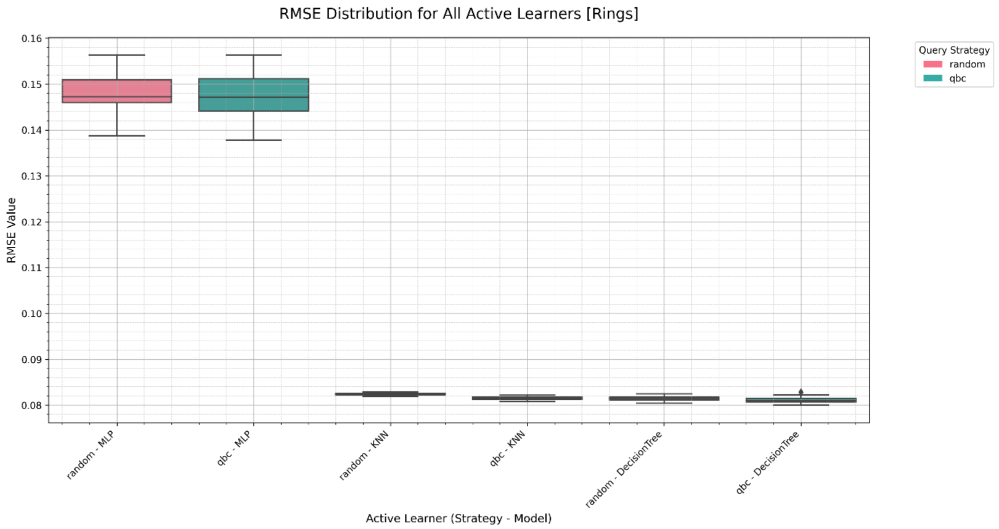

Figure 4: RMSE results for all active learners on the "Rings" feature.

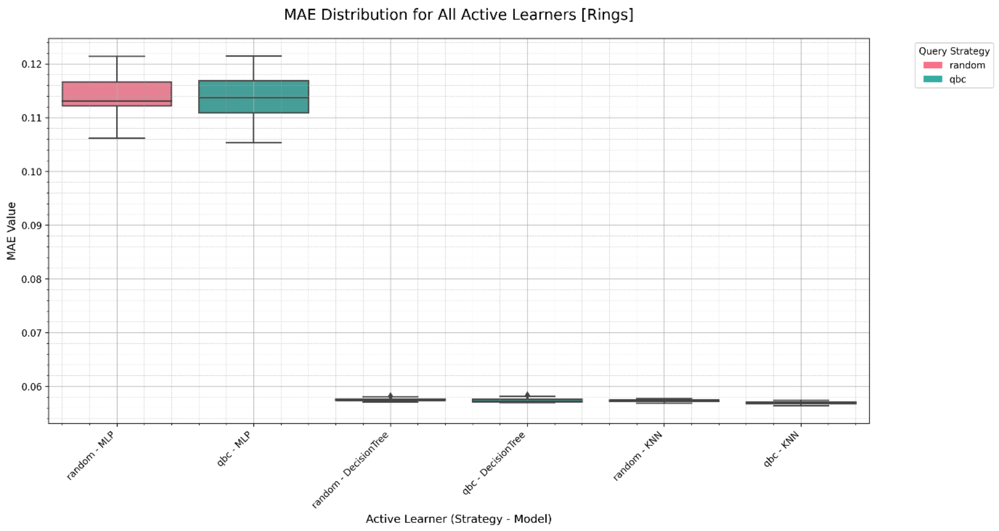

Figure 5: MAE results for all active learners on the "Rings" feature.

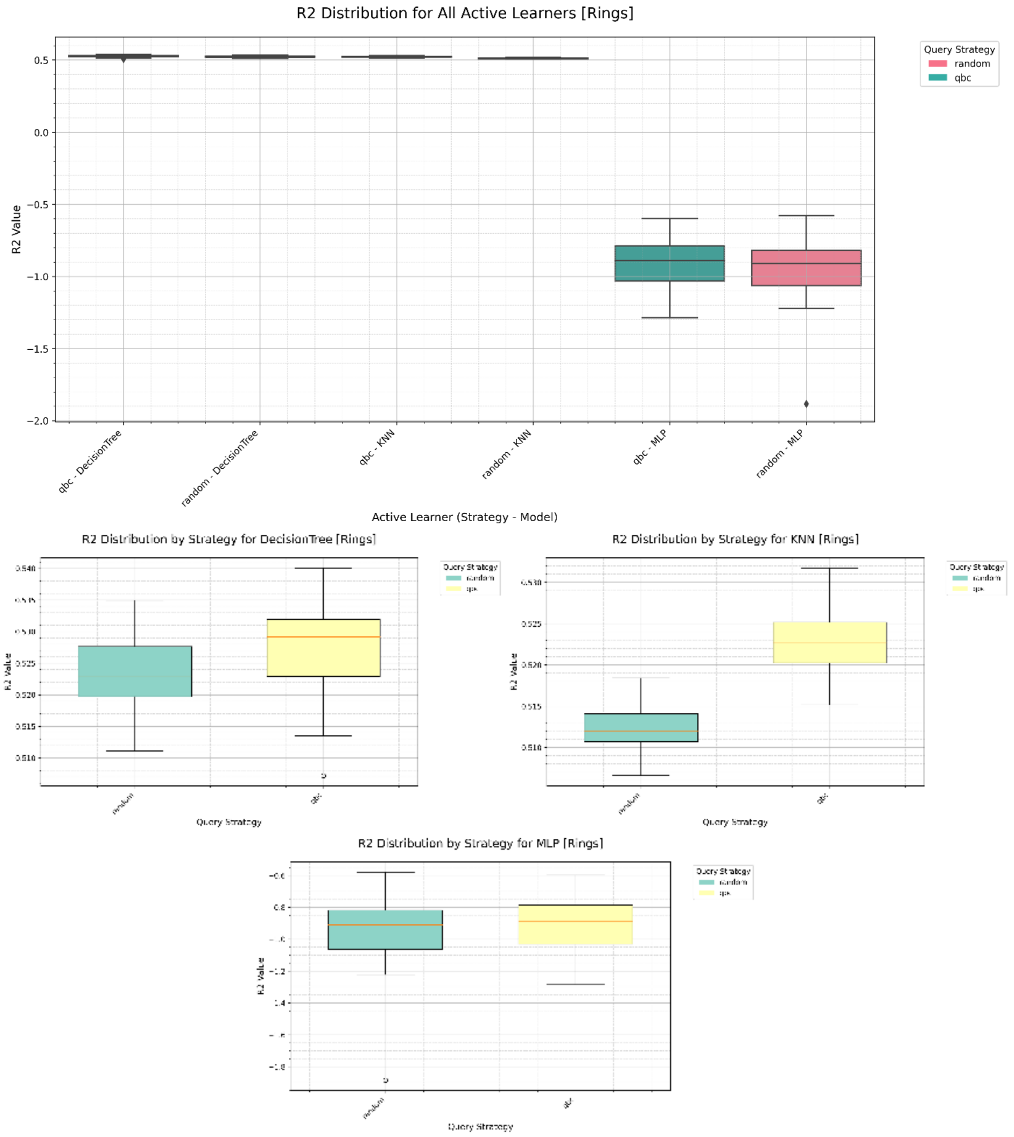

Figure 6: R2 results from all active learners (top) and per model type (bottom) on the Rings feature from the abalone dataset

4.2 Capacity Throughput

Decision tree (QBC) achieved the top performance score (Table 15), outperforming all other learners in its pairwise comparisons. The random strategy version of the decision tree also performed strongly, ranking second. MLPs, particularly under random sampling, consistently underperformed. KNN (QBC) achieved moderate performance but was outperformed by both decision tree variants.

Table 6: Mann-Whitney U-test results rankings on the Capacity Throughput feature from the HVB dataset

Model (Strategy)	**Total Wins**	**Total Losses**	**Total Draws**	**Performance Score**
Decision tree (QBC)	5	0	0	1
Decision tree (Random)	4	1	0	0.8
KNN (QBC)	3	2	0	0.6
KNN (Random)	2	3	0	0.4
MLP (QBC)	1	4	0	0.2
MLP (Random)	0	5	0	0

Figures of RMSE, MAE, and R^2 are shown in Figures 8, 9, and 10, respectively. There is a clear difference between the QBC variants and the random sampling variants, with statistical significance confirmed for each model type.

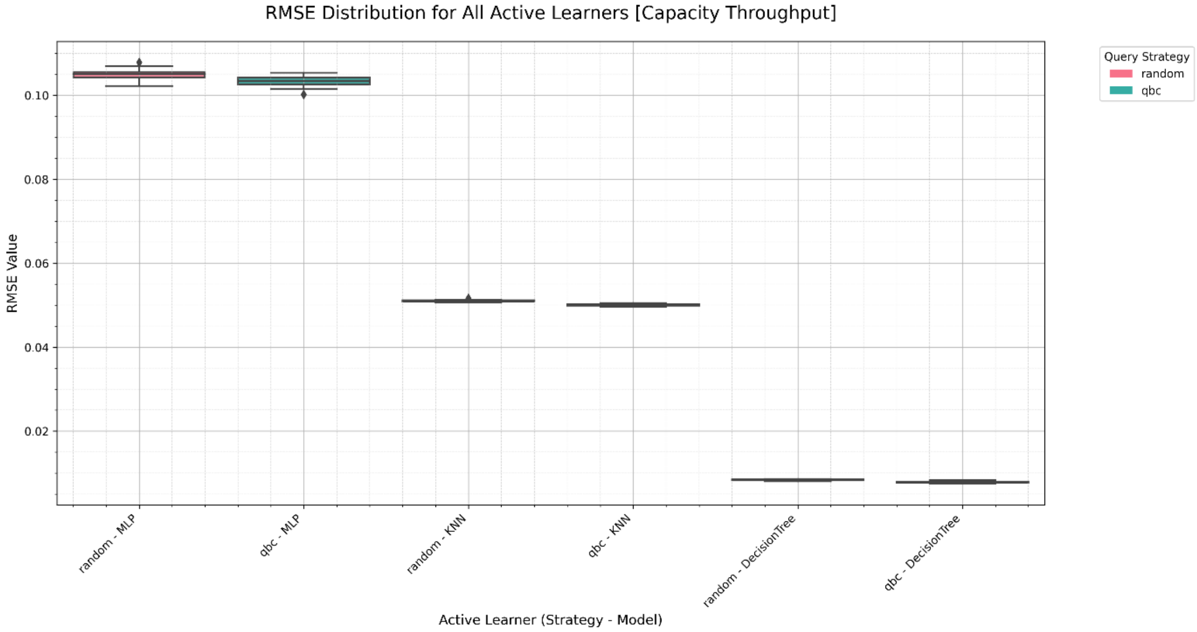

Figure 7: RMSE results for all active learners on the "Capacity Throughput" feature.

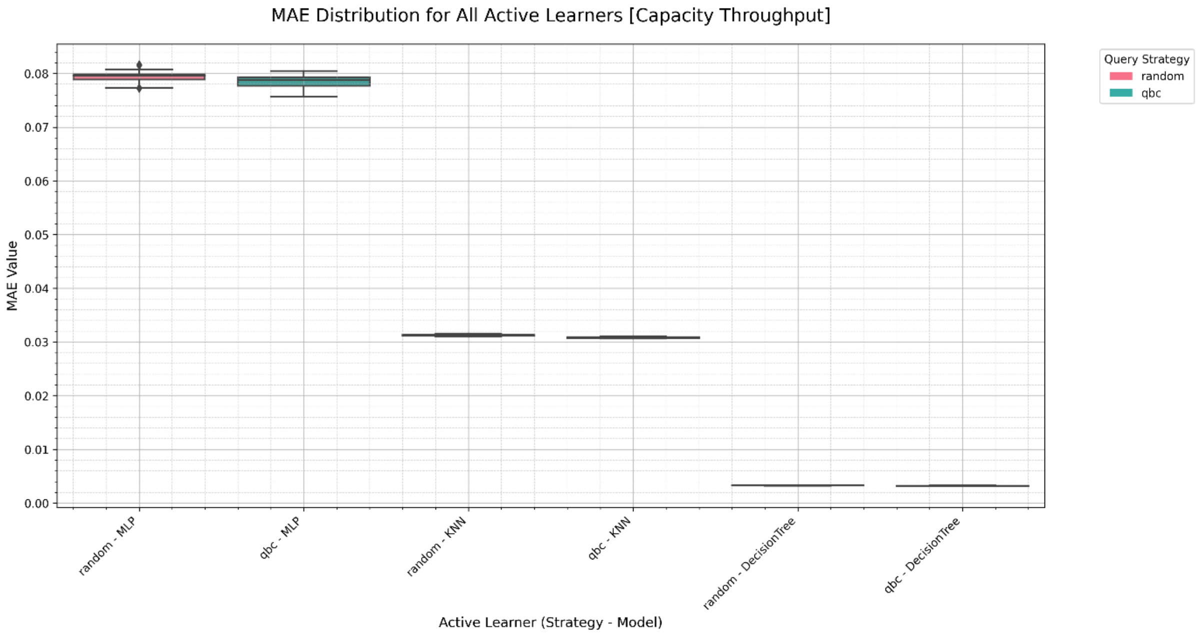

Figure 8: MAE results for all active learners on the “Capacity Throughput” feature

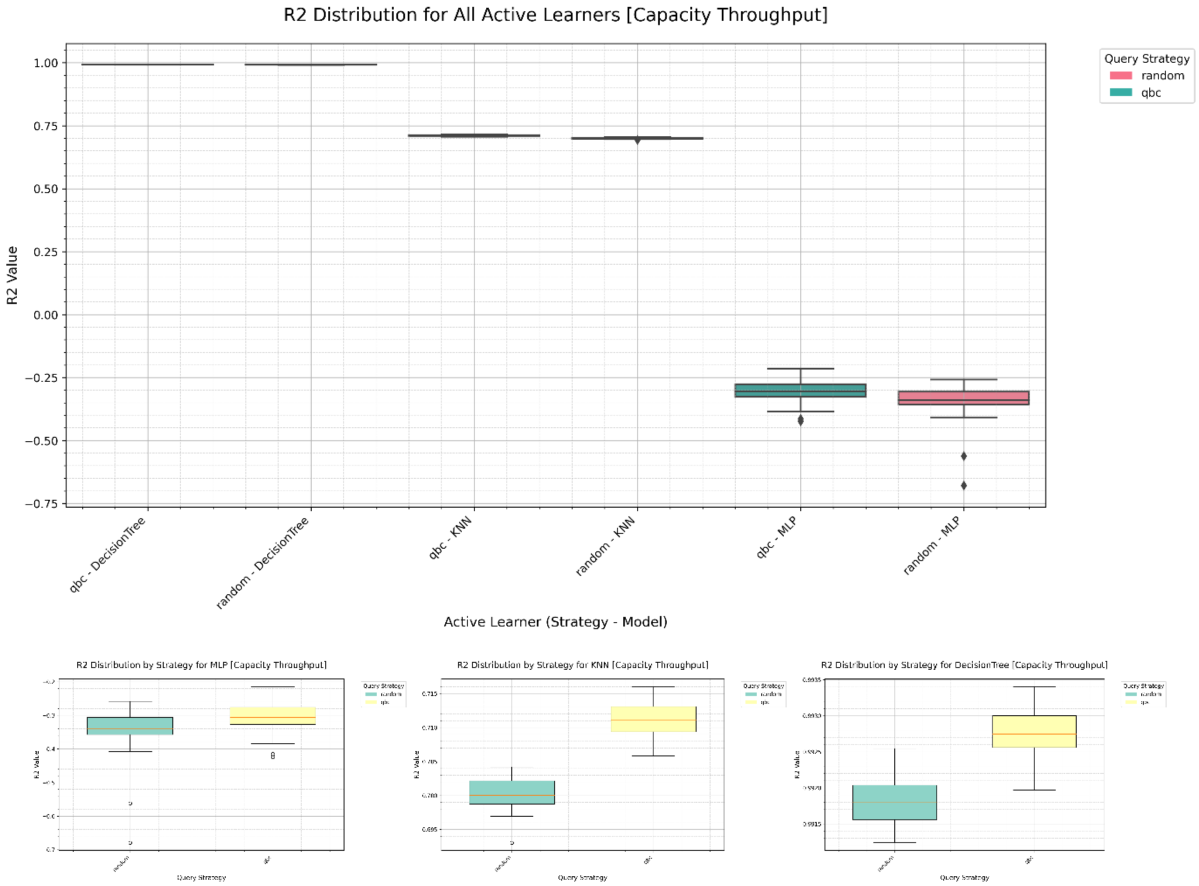

Figure 9: R^2 results from all active learners (top) and per model type (bottom) on the Capacity Throughput feature from the HVB dataset.

4.3 Minimum Capacity

Decision tree (Random) and decision tree (QBC) learners tied for the highest performance score, each winning four pairwise comparisons and drawing once. KNN (QBC) outperformed its random counterpart, while both MLP variants underperformed, with identical scores. These results demonstrate consistency in decision tree robustness across querying strategies.

Table 7: Mann-Whitney U-test ranking results on the Minimum Capacity feature from the HVB dataset

Model (Strategy)	**Total Wins**	**Total Losses**	**Total Draws**	**Performance Score**
Decision tree (Random)	4	0	1	0.9
Decision tree (QBC)	4	0	1	0.9
KNN (QBC)	3	2	0	0.6
KNN (Random)	2	3	0	0.4
MLP (Random)	0	4	1	0.1
MLP (QBC)	0	4	1	0.1

The distributions of MAE, RMSE, and R² are shown in Figures 11, 12, and 13, respectively. Only the querying k-NN (QBC) learner outperforms its random sampling variant. The MLP (Random) learn yields lower error and a better R^2 score, but this is not confirmed to be statistically significant.

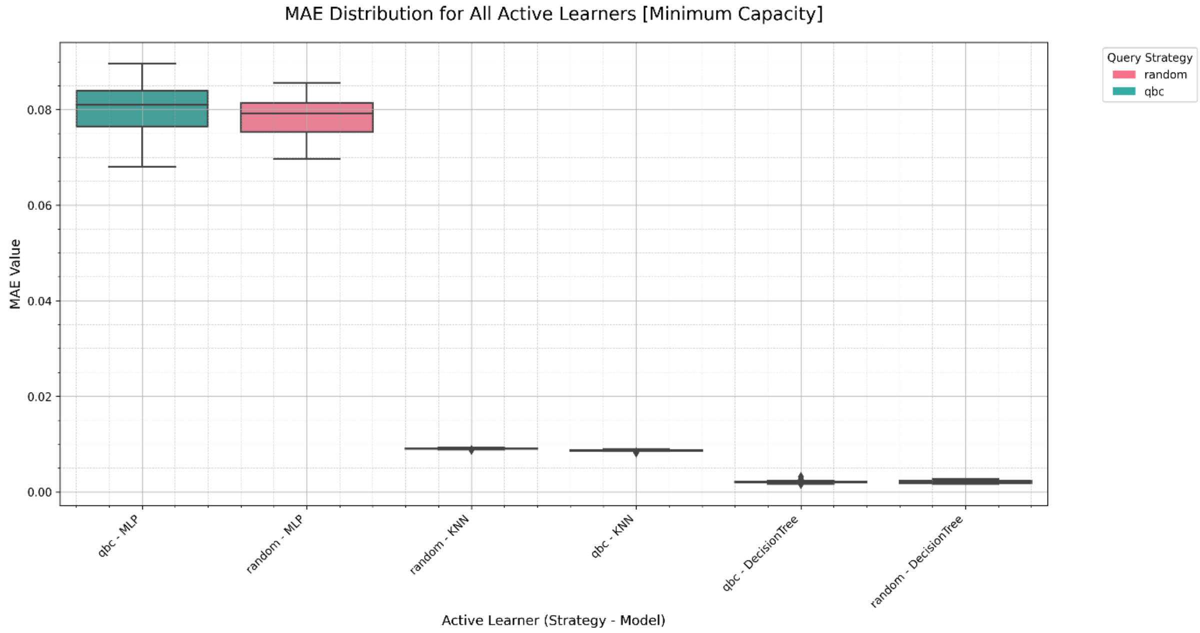

Figure 10: MAE results for all active learners on the "Minimum Capacity" feature.

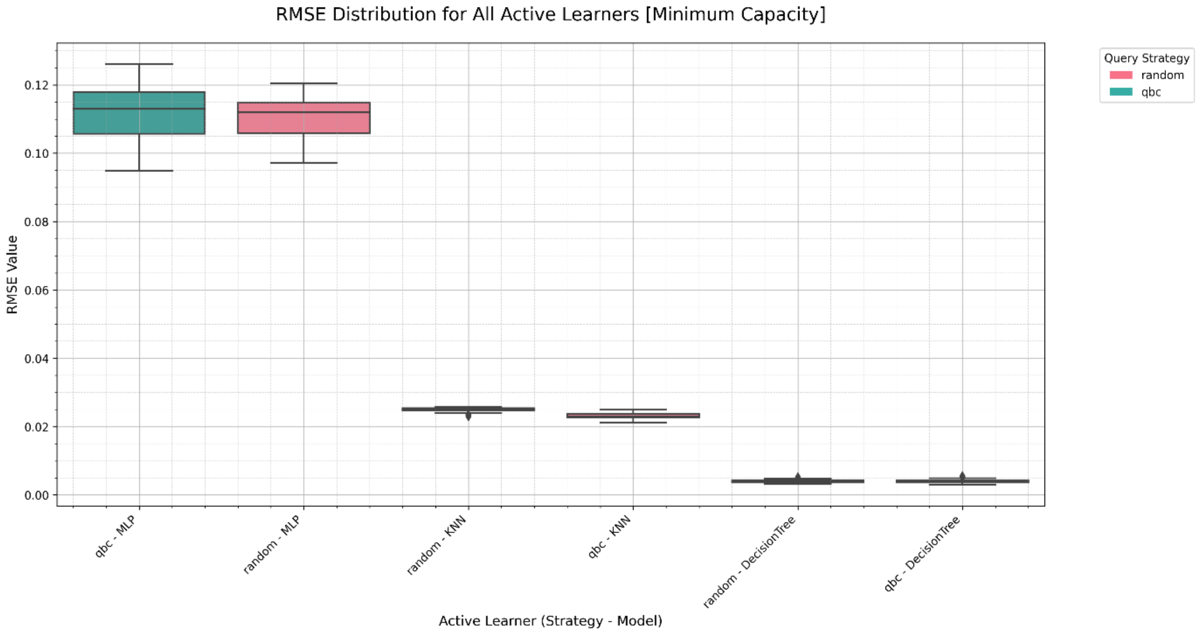

Figure 11: RMSE results for all active learners on the "Minimum Capacity" feature.

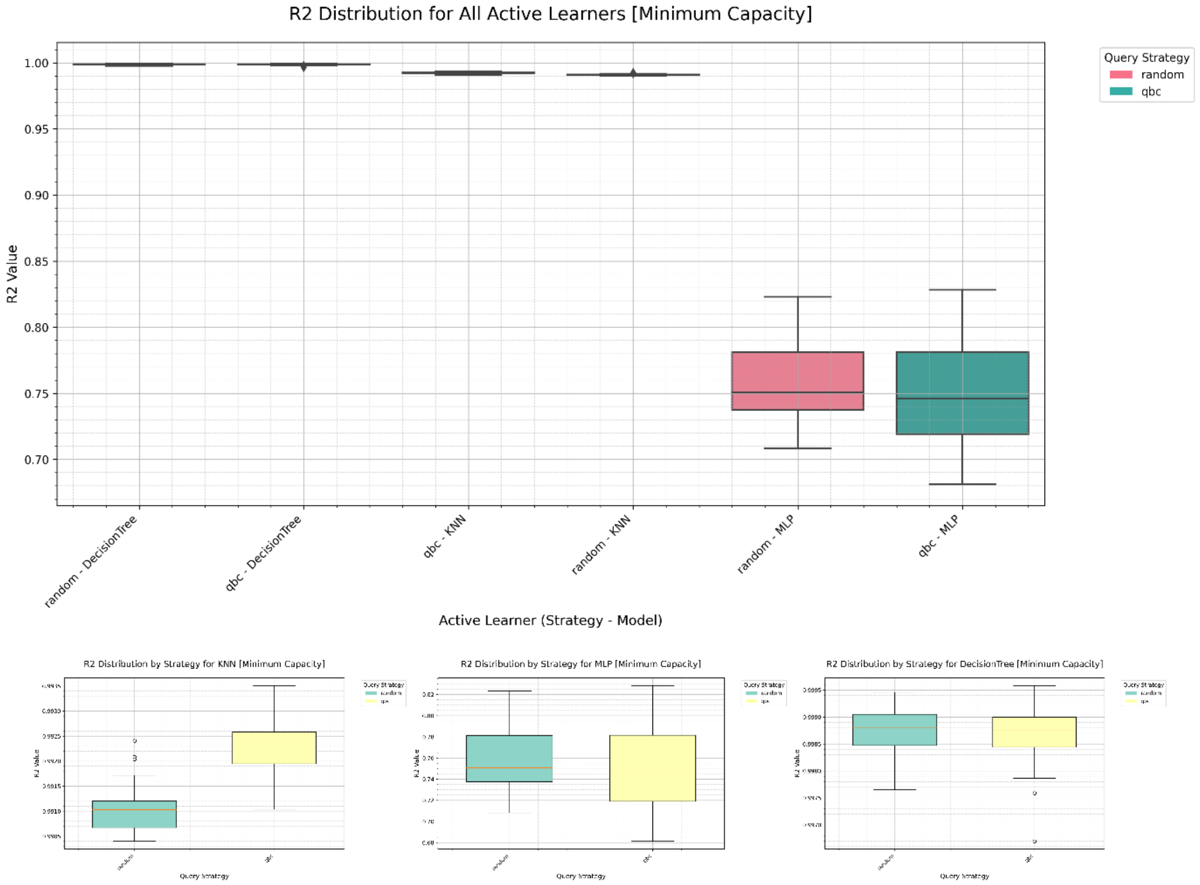

Figure 12: R2 results from all active learners (top) and per model type (bottom) on the Minimum Capacity feature from the HVB dataset.

4.4 Voltage

Decision tree learners, using both QBC and random sampling, tied for the top performance score, each winning four comparisons and drawing one. KNN (QBC) again outperformed KNN (Random), consistent with previous features. The MLP (QBC) learner outperforms its random counterpart. These findings further highlight the reliability of decision trees and the added value of QBC for k-NN and MLP.

Table 8: Mann-Whitney U-test results rankings on the Voltage feature from the HVB dataset

Model (Strategy)	**Total Wins**	**Total Losses**	**Total Draws**	**Performance Score**
Decision tree (QBC)	4	0	1	0.9
Decision tree (Random)	4	0	1	0.9
KNN (QBC)	3	2	0	0.6
KNN (Random)	2	3	0	0.4
MLP (QBC)	1	4	0	0.2
MLP (Random)	0	5	0	0

Figures 14, 15, and 15 show the RMSE, MAE, and R² distributions, respectively.

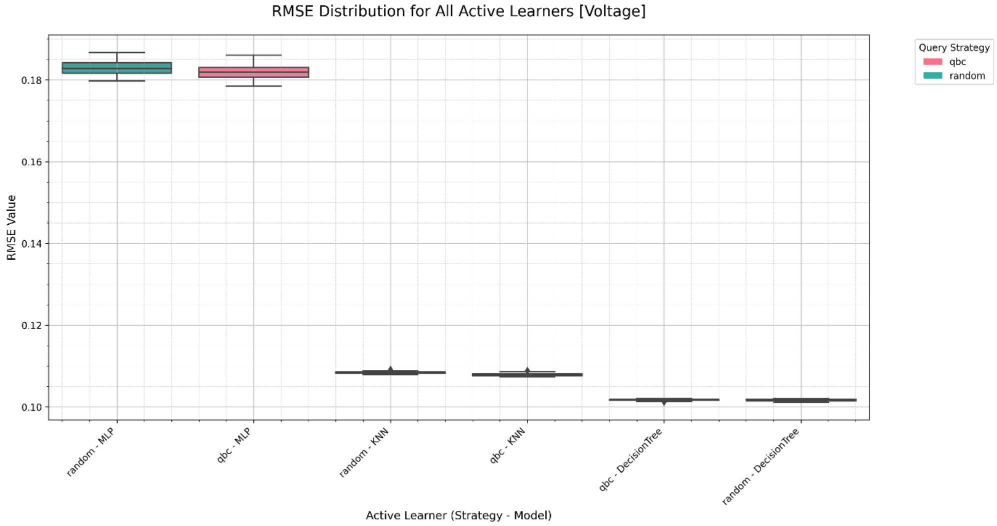

Figure 13: RMSE results for all active learners on the “Voltage” feature.

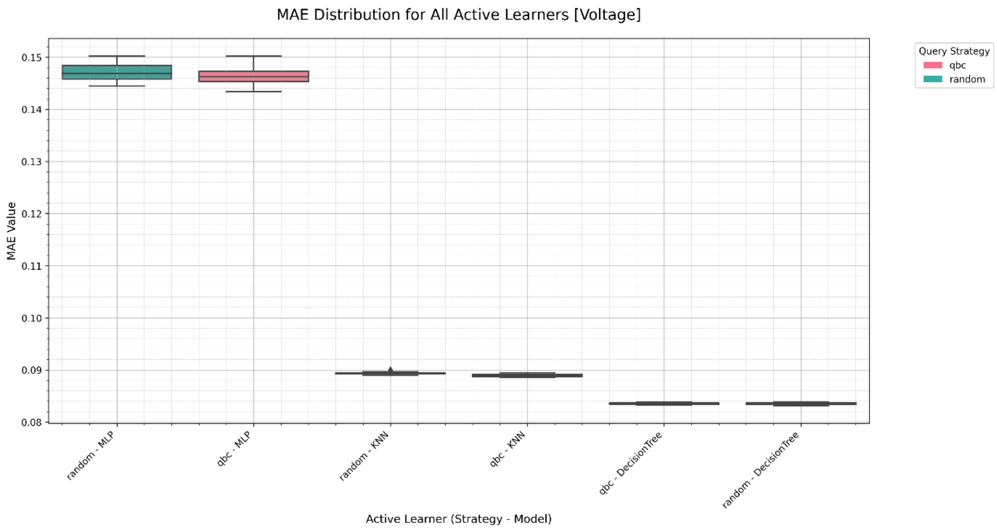

Figure 14: RMSE results for all active learners on the “Voltage” feature.

Figure 16 shows the clear improvement in R^2 score due to the QBC strategy.

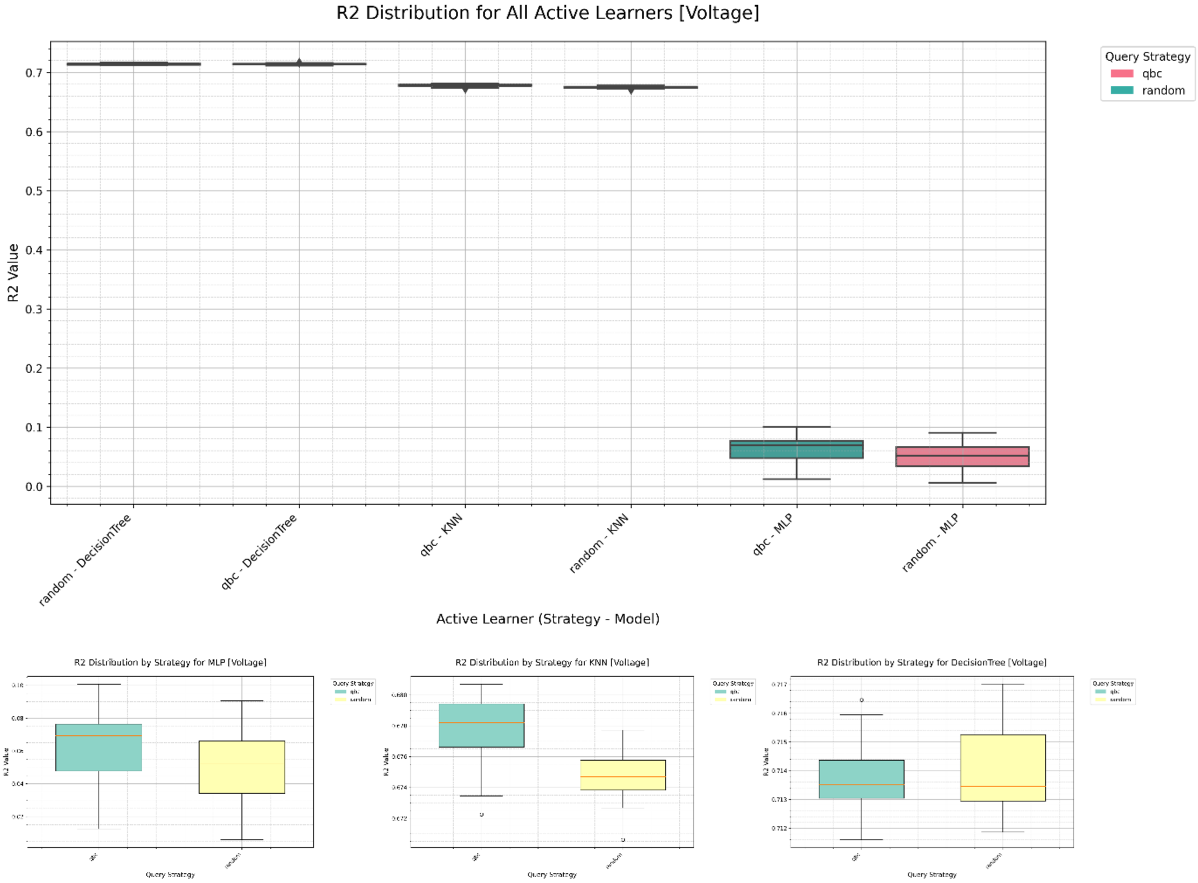

Figure 15: R2 results from all active learners (top) and per model type (bottom) on the Voltage feature from the HVB dataset

4.5 Material Weight

KNN (QBC) achieved the highest performance score, winning all five of its pairwise comparisons. KNN (Random) followed closely, outperforming both decision tree learners. Decision tree (QBC) and (Random) tied, with modest performance, while both MLP strategies ranked lowest.

Table 9: Mann-Whitney U-test ranking results on the "Material Weight" feature from the HVB dataset

Model (Strategy)	Total Wins	Total Losses	Total Draws	Performance Score
KNN (QBC)	5	0	0	1
KNN (Random)	4	1	0	0.8
Decision Tree (QBC)	2	2	1	0.5
Decision Tree (Random)	2	2	1	0.5
MLP (QBC)	1	4	0	0.2
MLP (Random)	0	5	0	0

Figures 20, 21, and 22 display RMSE, MAE, and R² distributions, respectively. The improvements caused by the QBC strategy are evident in model types, except for the decision tree learners.

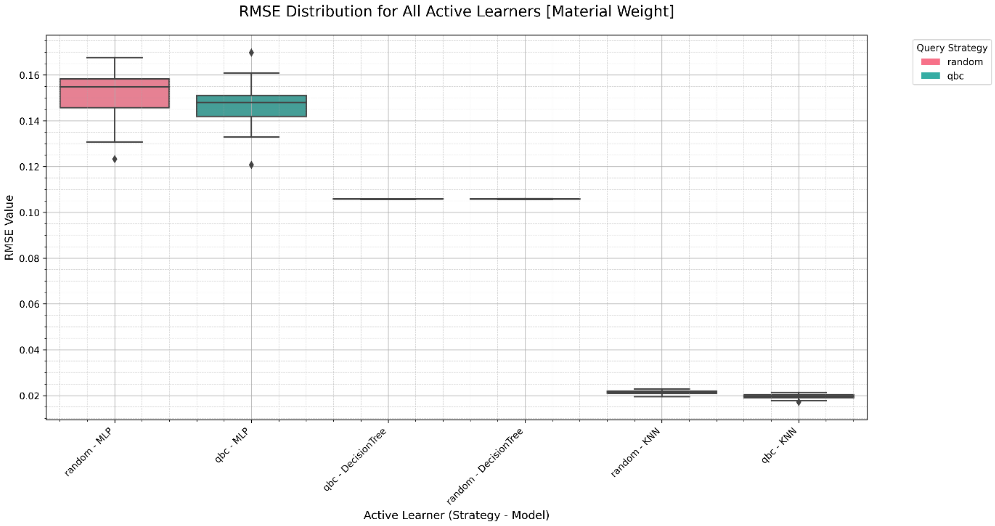

Figure 16: RMSE results for all active learners on the "Material Weight" feature.

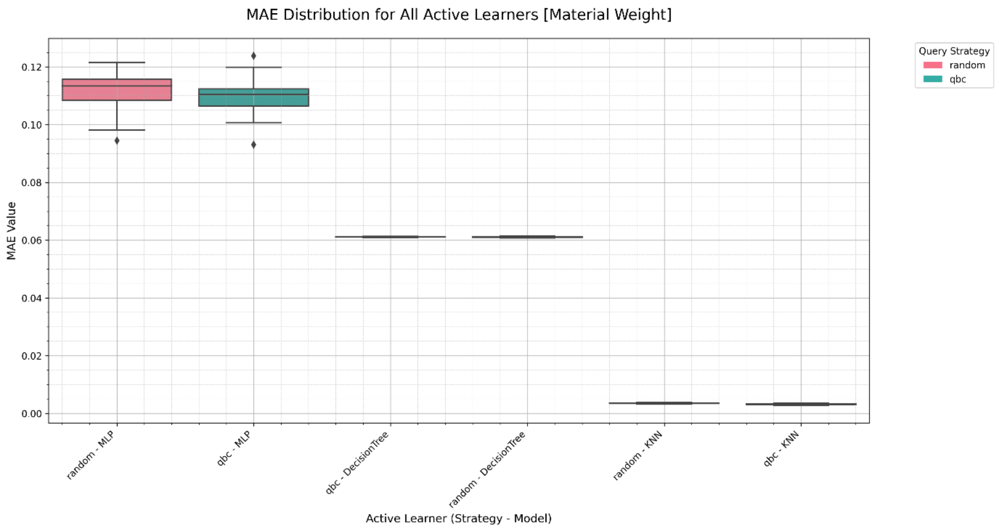

Figure 17: MAE results for all active learners on the "Material Weight" feature.

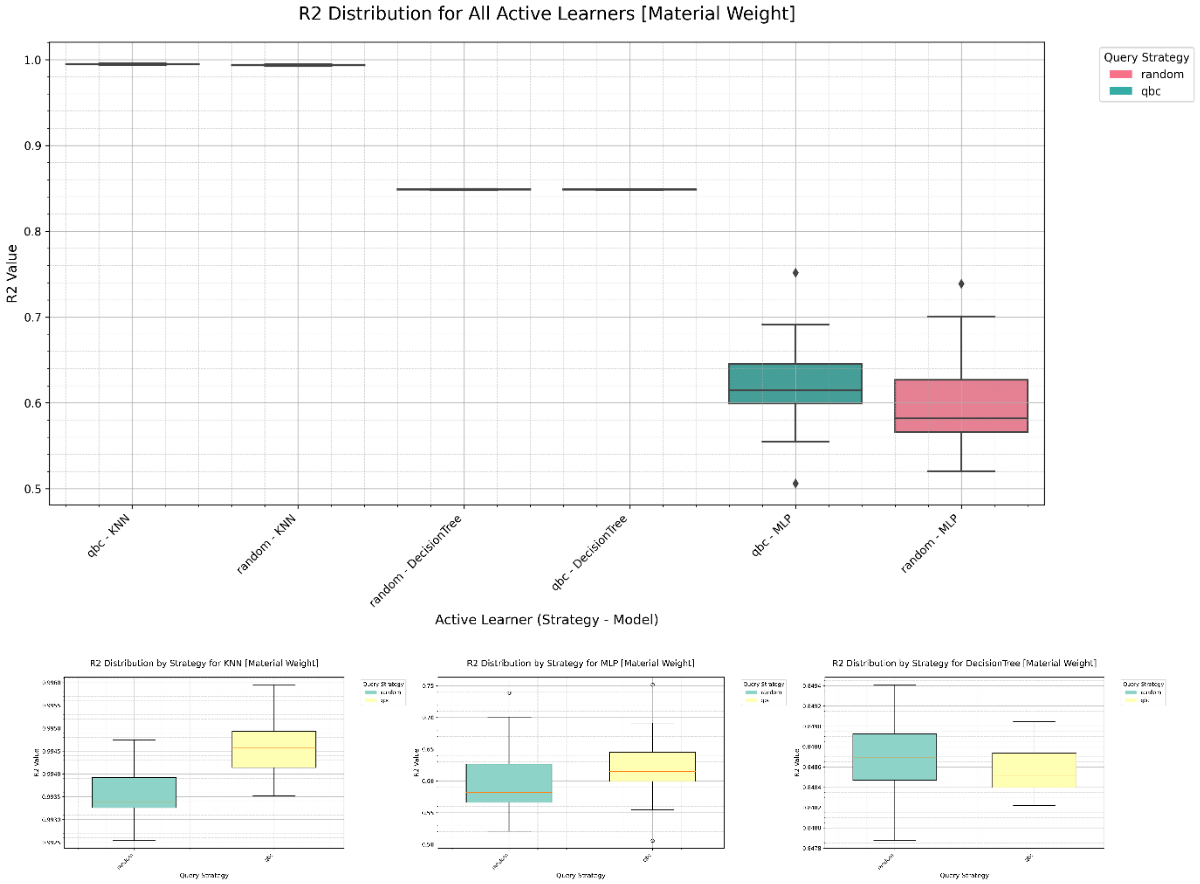

Figure 18: R2 results from all active learners (top) and per model type (bottom) on the Material Weight feature from the HVB dataset

4.6 Material Part Number

Similar to the *"Capacity Throughput"* feature, the decision tree and k-NN QBC learners outperform their random sampling variants. The MLP learners exhibit low performance on this feature, seen by the negative R^2 scores.

Table 10: Mann-Whitney U-test ranking results on the "Material Part Number" feature from the HVB dataset

Model (Strategy)	Total Wins	Total Losses	Total Draws	Performance Score
Decision tree (QBC)	5	0	0	1
Decision tree (Random)	4	1	0	0.8
KNN (QBC)	3	2	0	0.6
KNN (Random)	2	3	0	0.4
MLP (QBC)	0	4	1	0.1
MLP (Random)	0	4	1	0.1

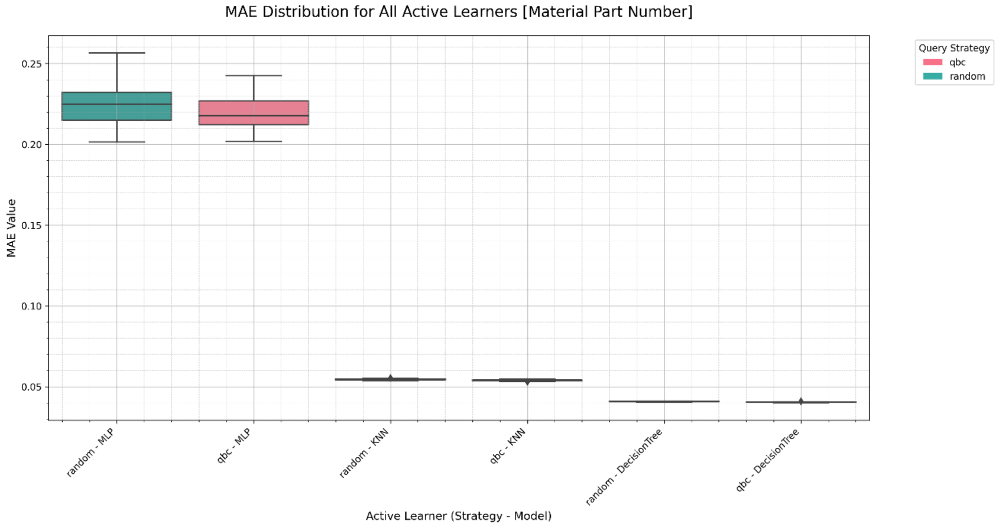

Figure 19: MAE results for all active learners on the "Material Weight" feature.

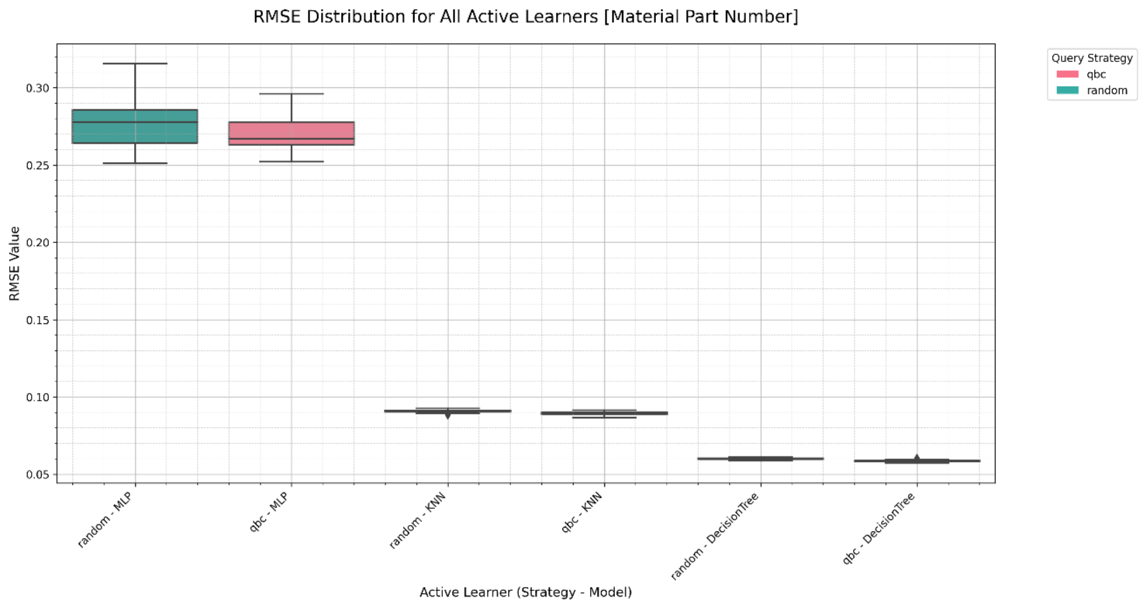

Figure 20: RMSE results for all active learners on the “Material Weight” feature.

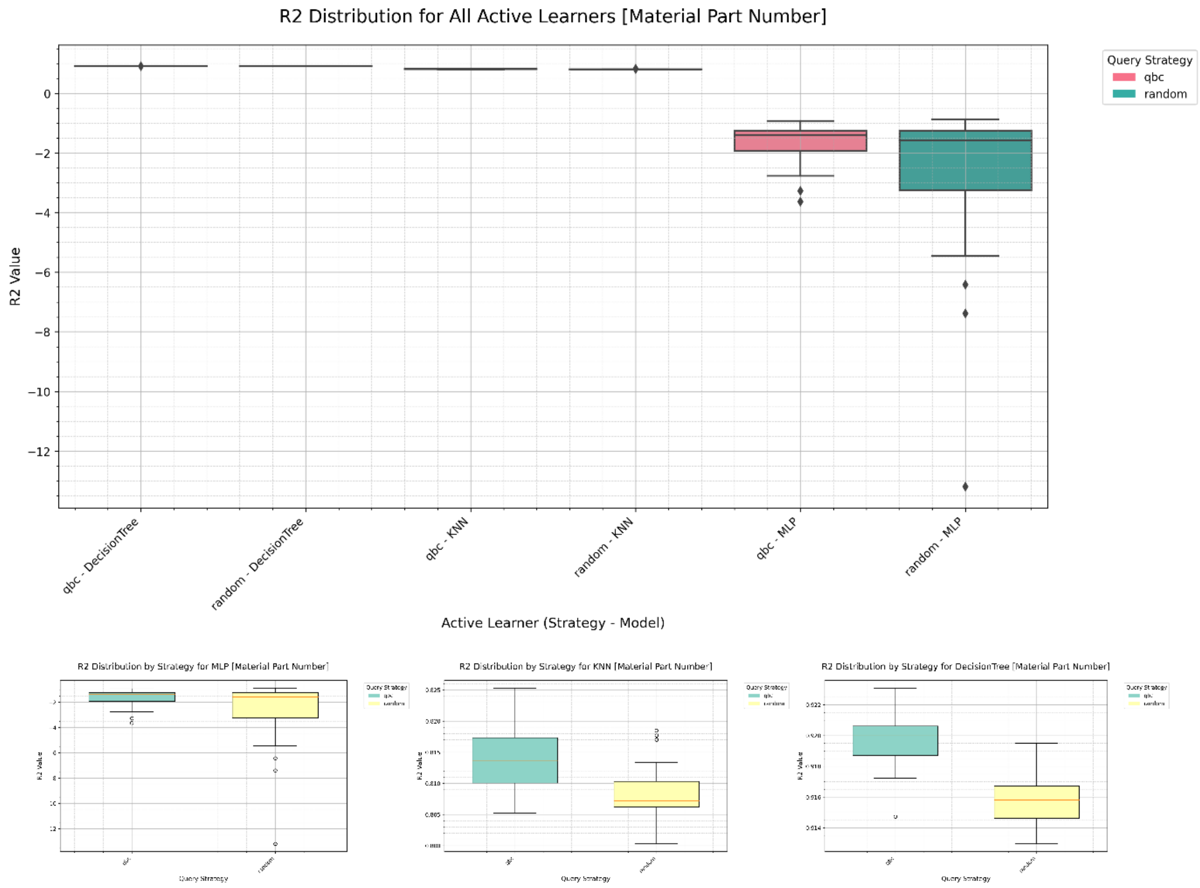

Figure 21: R2 results from all active learners (top) and per model type (bottom) on the Material Part Number feature from the HVB dataset