Evaluating algorithms for missing value imputation in real battery data[[1]](#footnote-2)

Dauda Sheni1[0000-0002-4634-2538], Anton Basson1[0000-0001-7998-330X], Jacomine Grobler2[0000-0002-1868-0759]

# Appendix: Detailed Results

## Introduction

This document serves as an appendix to the paper titled “Evaluating algorithms for missing value imputation in real battery data”. This appendix begins with a description of the experiment carried out, the summarised training and cross-validation results, as well as the results from the Mann-Whitney U test. The k-NN algorithm emerges as the best performing, demonstrating efficient learning and generalization across all features This paper has shown that ML algorithms can be used for MVI and error correction, with k-NN and MICE providing efficient, and reliable estimates to missing values, with minimal training effort.

MVI performance is very good on features with lower standard deviation and cardinality, while it remains unreliable on others. The categorical features *'Material Weight'* and *'Material Part Number'*, modelled as continuous features, also provide good results, further demonstrating the versatility of the imputation methods. These observations are critical for understanding the applicability and limitations of different imputation techniques for error correction in the HVB dataset.

## Correlation Plot

Predictions are carried out for one categorical and nine continuous features. The correlation plot, shown in Figure 1, shows 3 pairs of highly correlated features: the mean and minimum capacity, capacity aging features, *‘Energy Throughput’* and *‘Capacity Throughput’.* A value of 1.00 indicates a strong positive correlation between two features (red), while a value of -1.00 represents a strong negative correlation (blue).



Figure 1: Correlation plot of HVB dataset after pre-processing

## Training Results

The training results in Figures 2-3 show the best performance for the *‘Minimum Capacity’* feature. The average error in the larger training set is higher than in the cross-validation set, indicating robustness in performance.

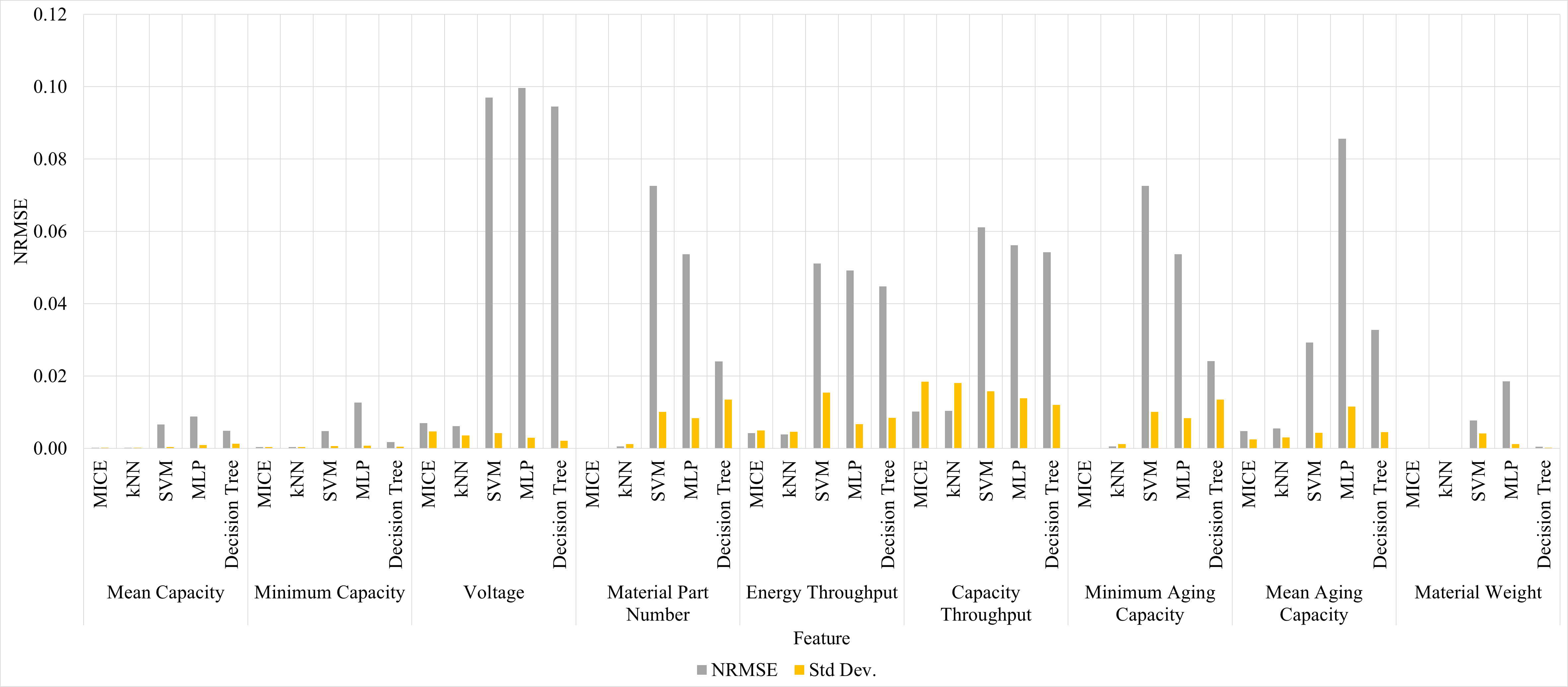


Figure 2: NRMSE across the continuous features during training.

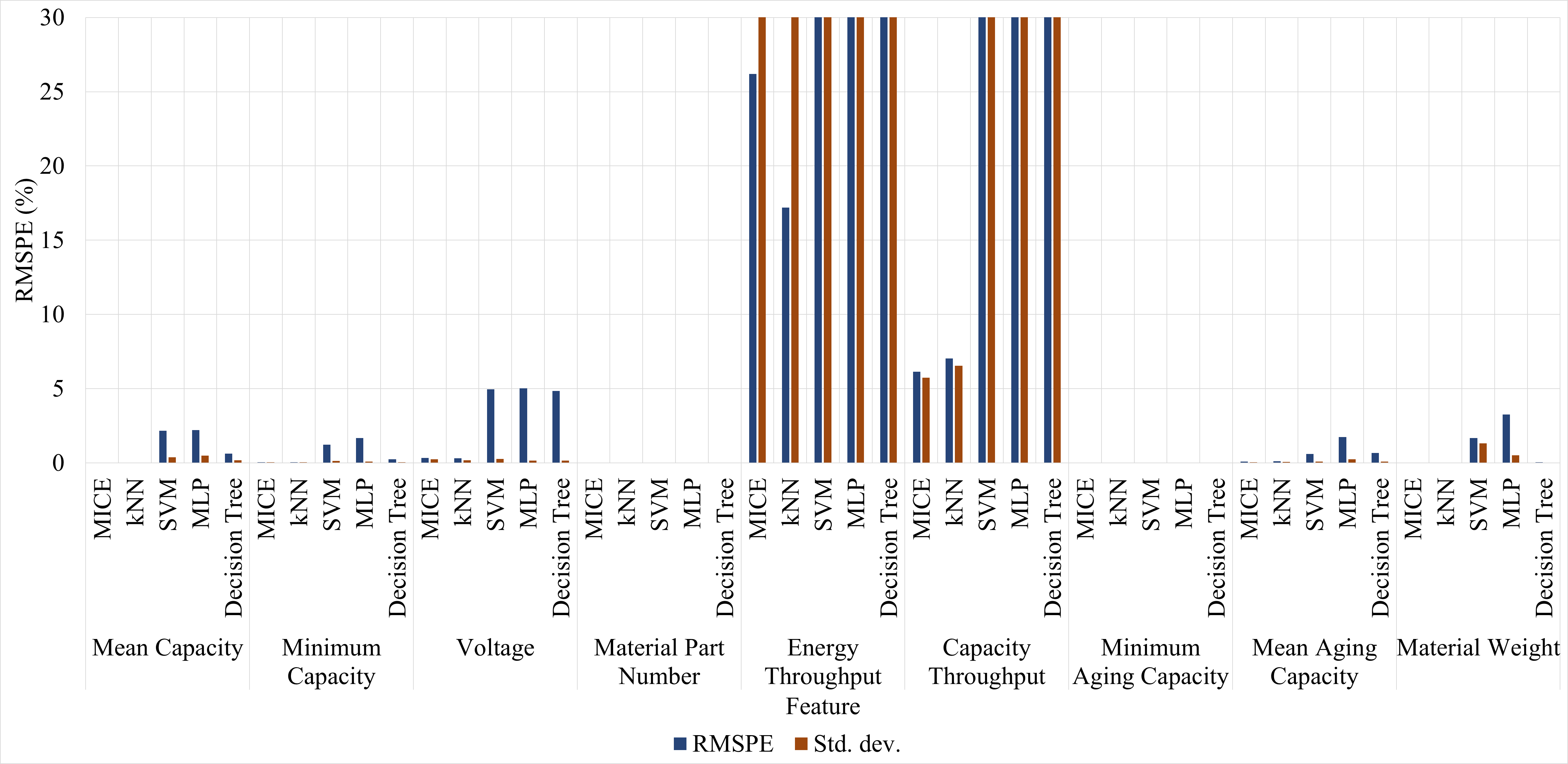


Figure 3: RMSPE across the continuous features during training.

## Cross-Validation Results

The cross-validation results are shown in Figures 4-5. The MICE and k-NN algorithms outperform the other algorithms, as in the training results. There is notable performance improvement in all algorithms.

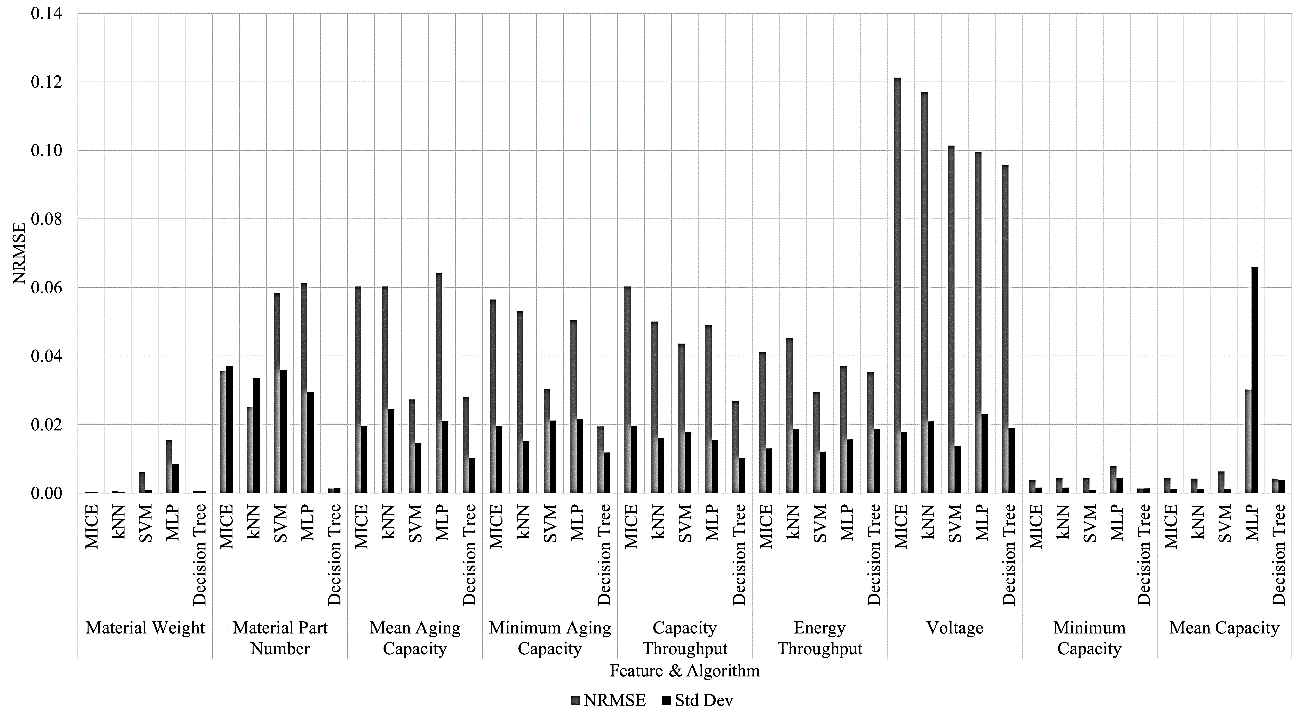


Figure 4: NRMSE across the continuous features in the dataset.

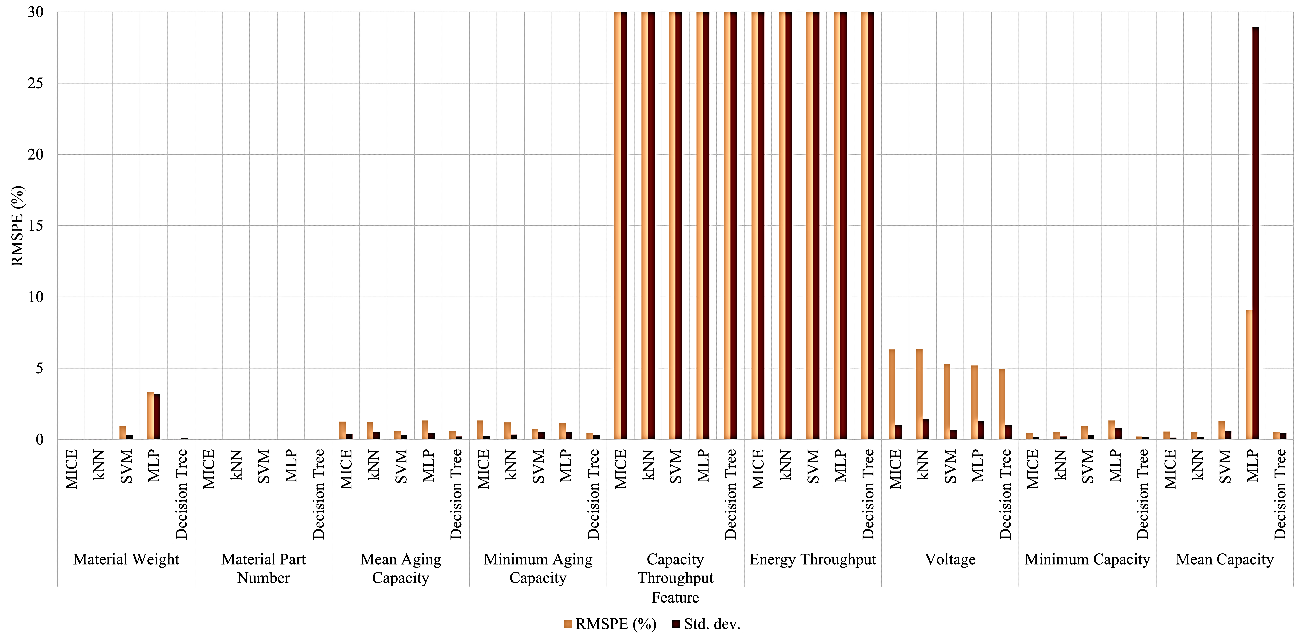


Figure 5: RMSPE across the continuous features.

## Mann-Whitney U Test

The Mann-Whitney U test revealed no statistical significance between the two imputation algorithms. In all other cases, the MICE and k-NN algorithm outperform the other three predictive algorithms. The results for the RMSPE and NRMSE metrics are shown in Tables 1 and 2, respectively.

Table 1: Mann-Whitney U Test on RMSPE for the continuous features

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| ***Material Part Number*** | | | | | | ***Voltage*** | | | | | |
|  | MICE | kNN | MLP | SVM | Decision Tree |  | MICE | kNN | MLP | SVM | Decision Tree |
| MICE | 0 | 0 | 1 | 1 | 1 | MICE | 0 | 0 | 1 | 1 | 1 |
| kNN | 0 | 0 | 1 | 1 | 1 | kNN | 0 | 0 | 1 | 1 | 1 |
| MLP | -1 | -1 | 0 | 1 | -1 | MLP | -1 | -1 | 0 | 0 | 0 |
| SVM | -1 | -1 | -1 | 0 | -1 | SVM | -1 | -1 | 0 | 0 | 0 |
| Decision Tree | -1 | -1 | 1 | 1 | 0 | Decision Tree | -1 | -1 | 0 | 0 | 0 |
| ***Capacity Min*** | | | | | | ***Capacity Mean*** | | | | | |
|  | MICE | kNN | MLP | SVM | Decision Tree |  | MICE | kNN | MLP | SVM | Decision Tree |
| MICE | 0 | 0 | 1 | 1 | 1 | MICE | 0 | 0 | 1 | 1 | 1 |
| kNN | 0 | 0 | 1 | 1 | 1 | kNN | 0 | 0 | 1 | 1 | 1 |
| MLP | -1 | -1 | 0 | 1 | -1 | MLP | -1 | -1 | 0 | 1 | -1 |
| SVM | -1 | -1 | -1 | 0 | -1 | SVM | -1 | -1 | -1 | 0 | -1 |
| Decision Tree | -1 | -1 | 1 | 1 | 0 | Decision Tree | -1 | -1 | 1 | 1 | 0 |
| ***Capacity Throughput*** | | | | | | ***Energy Throughput*** | | | | | |
|  | MICE | kNN | MLP | SVM | Decision Tree |  | MICE | kNN | MLP | SVM | Decision Tree |
| MICE | 0 | 0 | 1 | 1 | 1 | MICE | 0 | 0 | 1 | 1 | 1 |
| kNN | 0 | 0 | 1 | 1 | 1 | kNN | 0 | 0 | 1 | 1 | 1 |
| MLP | -1 | -1 | 0 | -1 | -1 | MLP | -1 | -1 | 0 | -1 | -1 |
| SVM | -1 | -1 | 1 | 0 | -1 | SVM | -1 | -1 | 1 | 0 | -1 |
| Decision Tree | -1 | -1 | 1 | 1 | 0 | Decision Tree | -1 | -1 | 1 | 1 | 0 |
| ***Capacity Aging Min*** | | | | | | ***Capacity Aging Mean*** | | | | | |
|  | MICE | kNN | MLP | SVM | Decision Tree |  | MICE | kNN | MLP | SVM | Decision Tree |
| MICE | 0 | 0 | 1 | 1 | 1 | MICE | 0 | 0 | 1 | 1 | 1 |
| kNN | 0 | 0 | 1 | 1 | 1 | kNN | 0 | 0 | 1 | 1 | 1 |
| MLP | -1 | -1 | 0 | -1 | -1 | MLP | -1 | -1 | 0 | -1 | -1 |
| SVM | -1 | -1 | 1 | 0 | -1 | SVM | -1 | -1 | 1 | 0 | 0 |
| Decision Tree | -1 | -1 | 1 | 1 | 0 | Decision Tree | -1 | -1 | 1 | 0 | 0 |

Table 2: Mann-Whitney U Test on NRMSE for the continuous features

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| ***Material Part Number*** | | | | | | ***Voltage*** | | | | | |
|  | MICE | kNN | MLP | SVM | Decision Tree |  | MICE | kNN | MLP | SVM | Decision Tree |
| MICE | 0 | 0 | 1 | 1 | 1 | MICE | 0 | 0 | 1 | 1 | 1 |
| kNN | 0 | 0 | 1 | 1 | 1 | kNN | 0 | 0 | 1 | 1 | 1 |
| MLP | -1 | -1 | 0 | 1 | -1 | MLP | -1 | -1 | 0 | -1 | -1 |
| SVM | -1 | -1 | -1 | 0 | -1 | SVM | -1 | -1 | 1 | 0 | 0 |
| Decision Tree | -1 | -1 | 1 | 1 | 0 | Decision Tree | -1 | -1 | 1 | 0 | 0 |
|  |  | **Minimum Capacity** | | |  | **Mean Capacity** | | | | | |
|  | MICE | kNN | MLP | SVM | Decision Tree |  | MICE | kNN | MLP | SVM | Decision Tree |
| MICE | 0 | 0 | 1 | 1 | 1 | MICE | 0 | 0 | 1 | 1 | 1 |
| kNN | 0 | 0 | 1 | 1 | 1 | kNN | 0 | 0 | 1 | 1 | 1 |
| MLP | -1 | -1 | 0 | -1 | -1 | MLP | -1 | -1 | 0 | -1 | -1 |
| SVM | -1 | -1 | 1 | 0 | -1 | SVM | -1 | -1 | 1 | 0 | -1 |
| Decision Tree | -1 | -1 | 1 | 1 | 0 | Decision Tree | -1 | -1 | 1 | 1 | 0 |
| Capacity Throughput | | | | | | Energy Throughput | | | | | |
|  | MICE | kNN | MLP | SVM | Decision Tree |  | MICE | kNN | MLP | SVM | Decision Tree |
| MICE | 0 | 0 | 1 | 1 | 1 | MICE | 0 | 0 | 1 | 1 | 1 |
| kNN | 0 | 0 | 1 | 1 | 1 | kNN | 0 | 0 | 1 | 1 | 1 |
| MLP | -1 | -1 | 0 | 0 | 0 | MLP | -1 | -1 | 0 | 0 | 0 |
| SVM | -1 | -1 | 0 | 0 | 0 | SVM | -1 | -1 | 0 | 0 | 0 |
| Decision Tree | -1 | -1 | 0 | 0 | 0 | Decision Tree | -1 | -1 | 0 | 0 | 0 |
| Minimum Aging Capacity | | | | | | Mean Aging Capacity | | | | | |
|  | MICE | kNN | MLP | SVM | Decision Tree |  | MICE | kNN | MLP | SVM | Decision Tree |
| MICE | 0 | 0 | 1 | 1 | 1 | MICE | 0 | 0 | 1 | 1 | 1 |
| kNN | 0 | 0 | 1 | 1 | 1 | kNN | 0 | 0 | 1 | 1 | 1 |
| MLP | -1 | -1 | 0 | -1 | -1 | MLP | -1 | -1 | 0 | -1 | -1 |
| SVM | -1 | -1 | 1 | 0 | -1 | SVM | -1 | -1 | 1 | 0 | 0 |
| Decision Tree | -1 | -1 | 1 | 1 | 0 | Decision Tree | -1 | -1 | 1 | 0 | 0 |

## Grid Search Results

The grid search results for the MLP algorithm are found in Table 3. The best solver is Adam with the ReLu function for the continuous features in the HVB dataset. The number of hidden neurons, and initial learning rate vary among the features, with similar features in the dataset having the same hyperparameters. The size of the grid space may have affected results, and increasing the grid search size may result in better MLP predictions. The grid search results for the decision tree and SVR algorithms are found in Tables 4 and 5, respectively.

Table 3: MLP Grid Search Results

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Target Column** | **Number of Hidden Neurons** | **Initial Learning Rate** | **Best CV Score (Neg MSE)** | **Test Score (MSE)** |
| Material Part Number | 55 | 0.1 | -7.54E-05 | 0.9990 |
| Material Weight | 70 | 0.01 | -1.17E-04 | 0.9982 |
| Mean Aging Capacity | 80 | 0.1 | -1.35E-04 | 0.9996 |
| Minimum Aging Capacity | 80 | 0.01 | -1.07E-04 | 0.9993 |
| Mean Capacity | 80 | 0.1 | -1.37E-04 | 0.9997 |
| Minimum Capacity | 80 | 0.1 | -1.19E-04 | 0.9997 |
| Energy Throughput | 80 | 0.05 | -1.19E-04 | 0.9987 |
| Capacity Throughput | 80 | 0.05 | -1.57E-04 | 0.9994 |
| Charges | 60 | 0.05 | -1.46E-04 | 0.9991 |
| Full Charges | 80 | 0.01 | -1.17E-04 | 0.9993 |
| Voltage | 70 | 0.01 | -9.38E-05 | 0.9997 |

The ‘best; splitting method is selected across all features for the decision tree algorithm. The ‘Poisson’ criterion is the most common splitting criterion.

Table 4: Decision Tree Grid Search Results

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Target Column** | **Criterion** | **Maximum Depth** | **Minimum Samples in Leaf** | **Best CV Score (Neg MSE)** | **Test Score (MSE)** |
| Material Part Number | Squared error | 10 | 5 | -0.0004 | 0.9632 |
| Material Weight | Friedman MSE | 10 | 10 | -0.0004 | 0.9603 |
| Mean Aging Capacity | Poisson | 30 | 5 | -0.0005 | 0.9481 |
| Minimum Aging Capacity | Squared error | 10 | 10 | -0.0004 | 0.9622 |
| Mean Capacity | Poisson | 10 | 10 | -0.0004 | 0.9623 |
| Minimum Capacity | Poisson | 10 | 10 | -0.0004 | 0.9637 |
| Energy Throughput | Poisson | 10 | 10 | -0.0004 | 0.9491 |
| Capacity Throughput | Poisson | 10 | 10 | -0.0004 | 0.9590 |
| Charges | Poisson | 10 | 10 | -0.0004 | 0.9561 |
| Full Charges | Squared error | 10 | 10 | -0.0004 | 0.9581 |
| Voltage | Friedman MSE | 10 | 10 | -0.0004 | 0.9609 |

The grid search results for the SVM (SVR) algorithm are shown in Table 5. There is less commonality with the hyperparameters for these features, but the *‘Mean’* and *‘Minimum Capacity’* features have the same hyperparameters.

Table 5:SVM Grid Search Results

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Target Column** | **C** | **Epsilon** | **Kernel** | **Best CV Score (Neg MSE)** | **Test Score (MSE)** |
| Material Part Number | 10 | 0.05 | RBF | -6.05E-03 | 0.8726 |
| Material Weight | 5 | 0.01 | RBF | -6.17E-05 | 0.9988 |
| Mean Aging Capacity | 10 | 0.01 | linear | -6.08E-05 | 0.9945 |
| Minimum Aging Capacity | 10 | 0.01 | linear | -5.41E-05 | 0.9942 |
| Mean Capacity | 5 | 0.01 | linear | -1.74E-05 | 0.9998 |
| Minimum Capacity | 2 | 0.01 | linear | -2.12E-05 | 0.9997 |
| Energy Throughput | 8 | 0.01 | RBF | -2.10E-05 | 0.9966 |
| Capacity Throughput | 5 | 0.01 | RBF | -2.49E-05 | 0.9964 |
| Charges | 8 | 0.01 | linear | -1.88E-04 | 0.7776 |
| Full Charges | 2 | 0.01 | poly | -2.81E-04 | 0.5428 |
| Voltage | 10 | 0.1 | RBF | -9.24E-03 | 0.7635 |

1. Sheni, D.N., Basson, A.H., Grobler, J.: Evaluating algorithms for missing value imputation in real battery data. (2024). Submitted to BCS-SGAI 2024. [↑](#footnote-ref-2)