Creator **Olusanmi Hundogan** Created **Jan 11, 2024, 23:49** Last updated **Jan 12, 2024, 00:14**

# Files

- notebooks/analyze_missing_value_imputers_kcluster.py
- experiments/experiment_missing_value_imputers_kcluster.py
-

# Motivation

This experiment acts as a preliminary study for the experiment which compares different imputation methods. Instance based models like KNN, KMeans and KMedoid rely on the hyperparameter K and strongly depend on this hyperparameter.

To narrow the search space we conduct this preliminary study to choose the best hyperparameters for this task.

For this experiment we tested 3 instance based methods that rely on K.

- KNNImputer: Uses KNN regression to impute the missing values.
- KMeans: Imputes missing values with the centroid values of the dataset
- KMedoid: Imputes missing values using a medoid ("median" data point in the dataset)

# Design

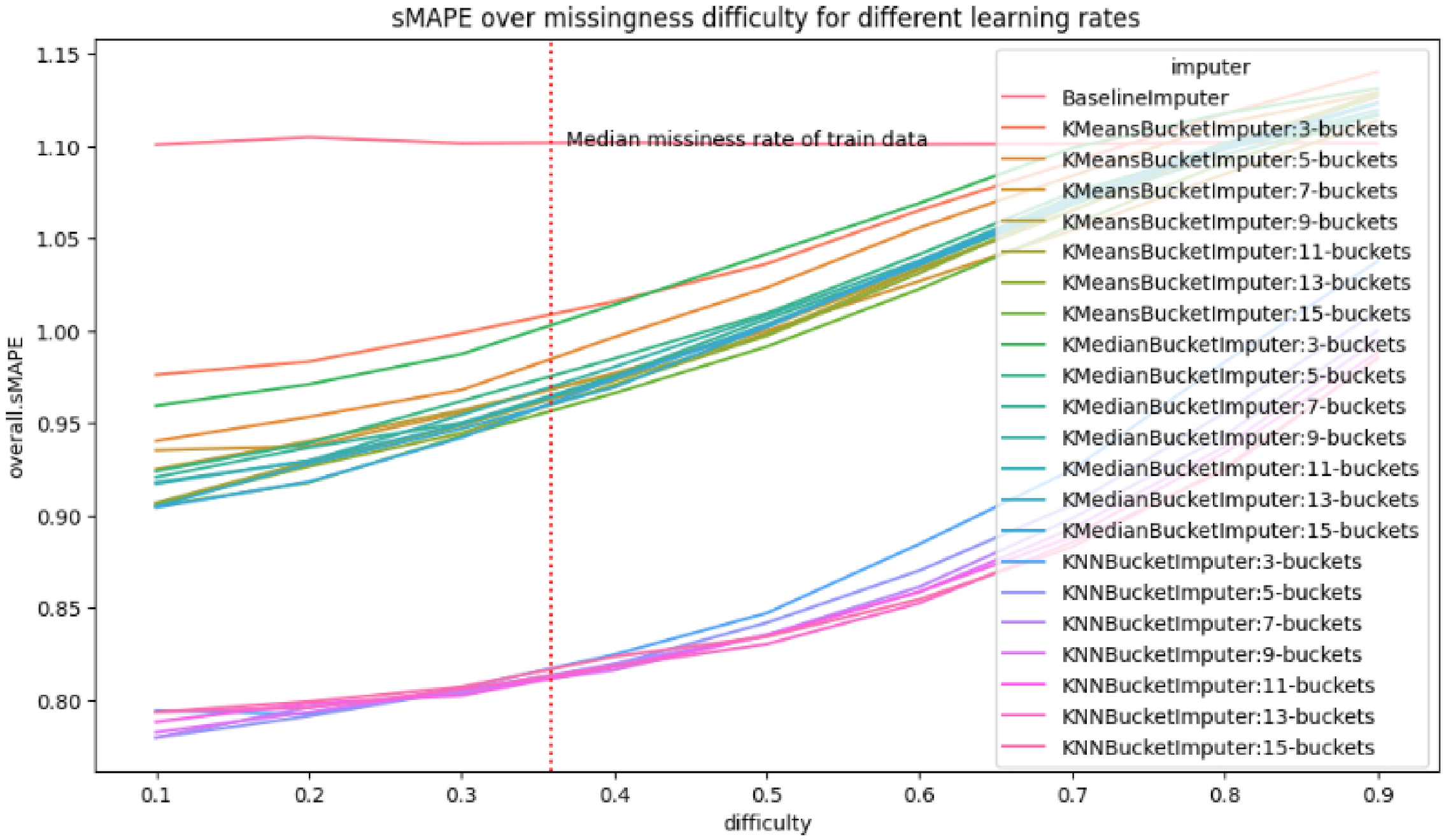
Within the experiment we explored different learning rates and n_estimators. The K values were [3, 5, 7, 9, 11, 13, 15].

We evaluate the imputer by removing values from a dataframe gradually and measuring the sMAPE of reconstructing the values. We test difficulties from 0.1 to 0.9 in even steps. A difficulty of 0.1 corresponds to the artificial removal of 10% of known values. (The number is not completely accurate because the 10% applies to the dataframe, but the dataframe already contains missing values.)

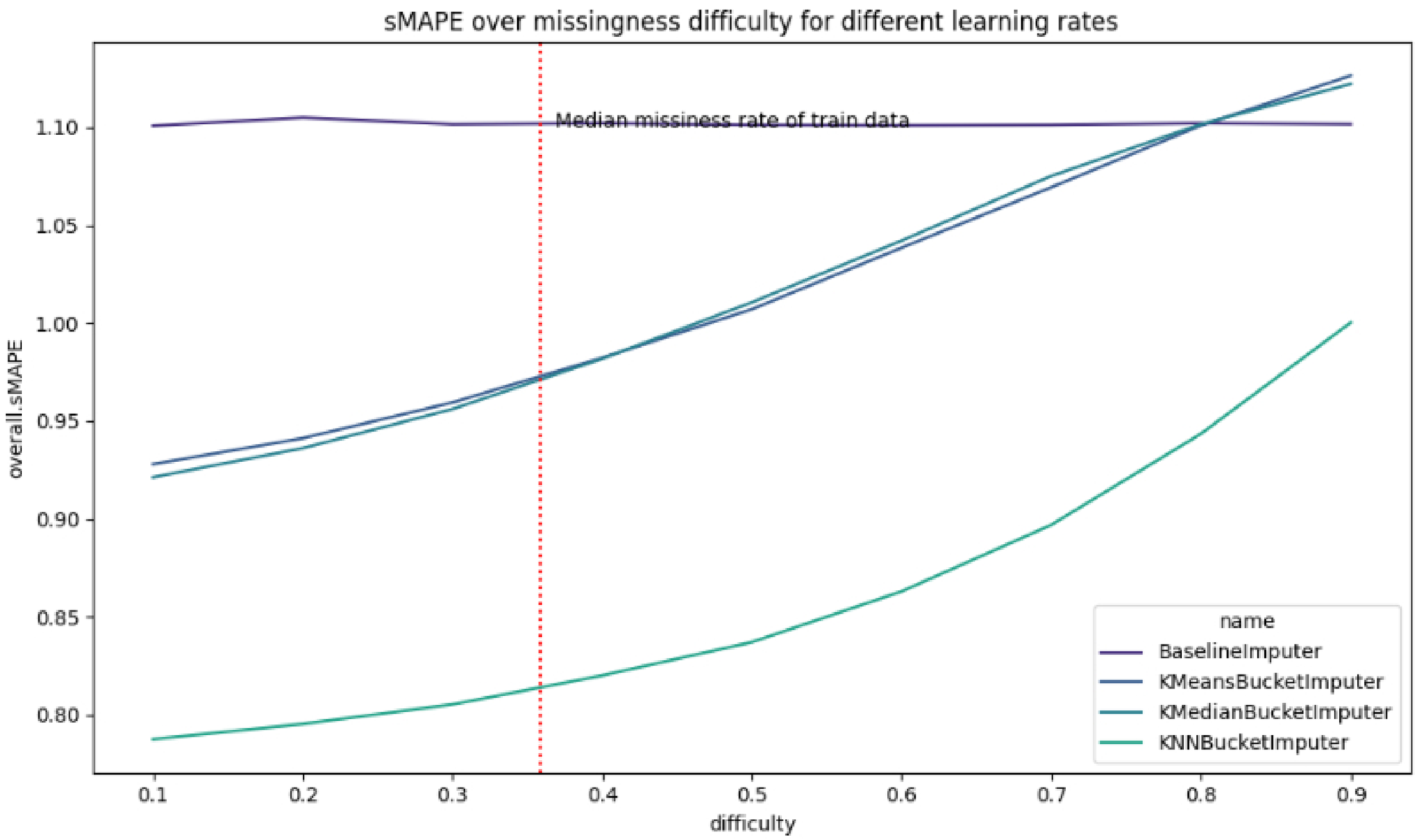
Fruthermore, we only impute on features that have a rate of missingness below 30% across the data set. After applying this feature selection 50% of the data rows have a missingness of below 10%.

# Results and Insight

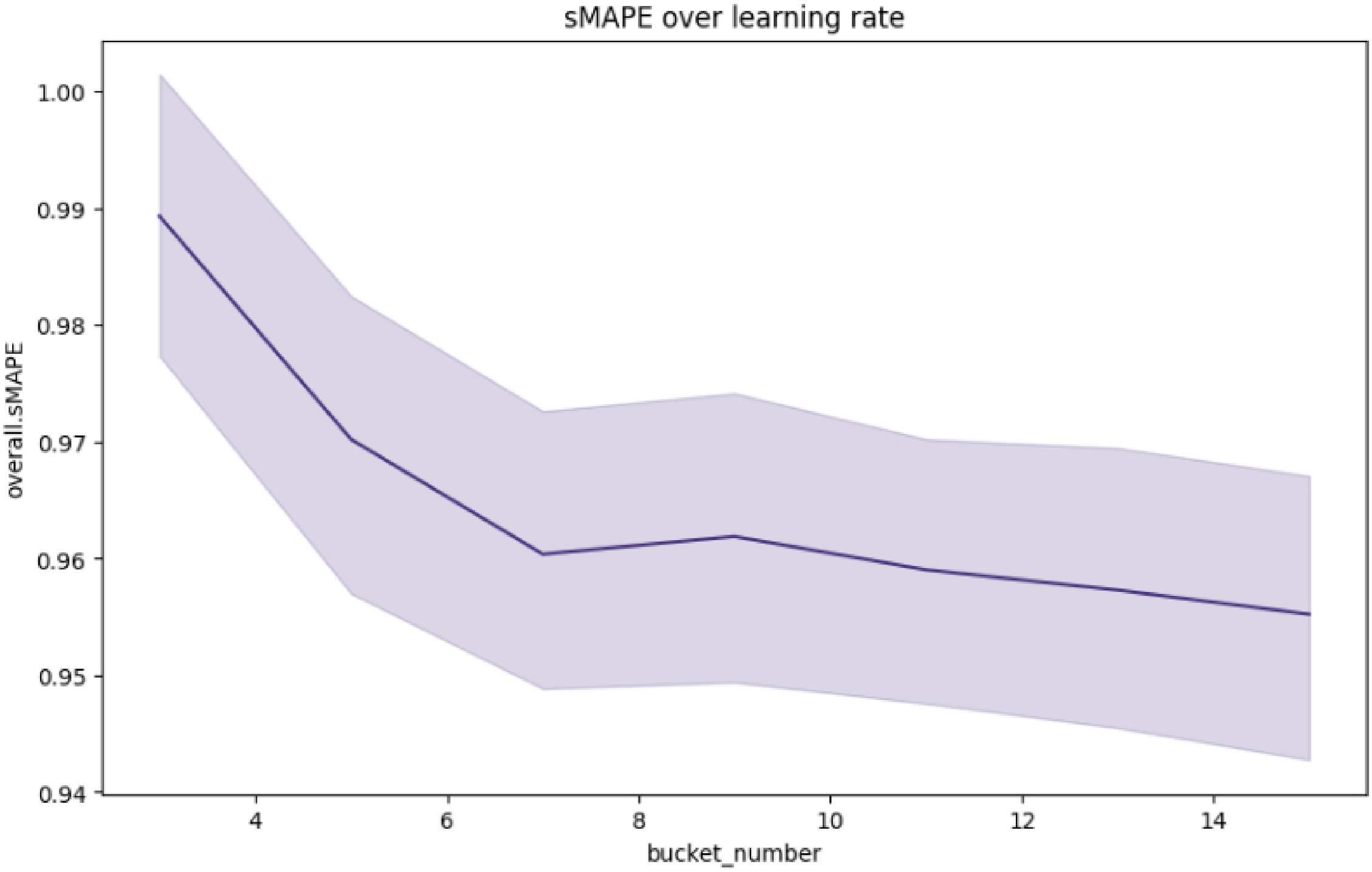
The plot shows all the different configurations tested across all difficulty levels. The results reveal a strong effect of the learning rate and the number of estimators on the performance.



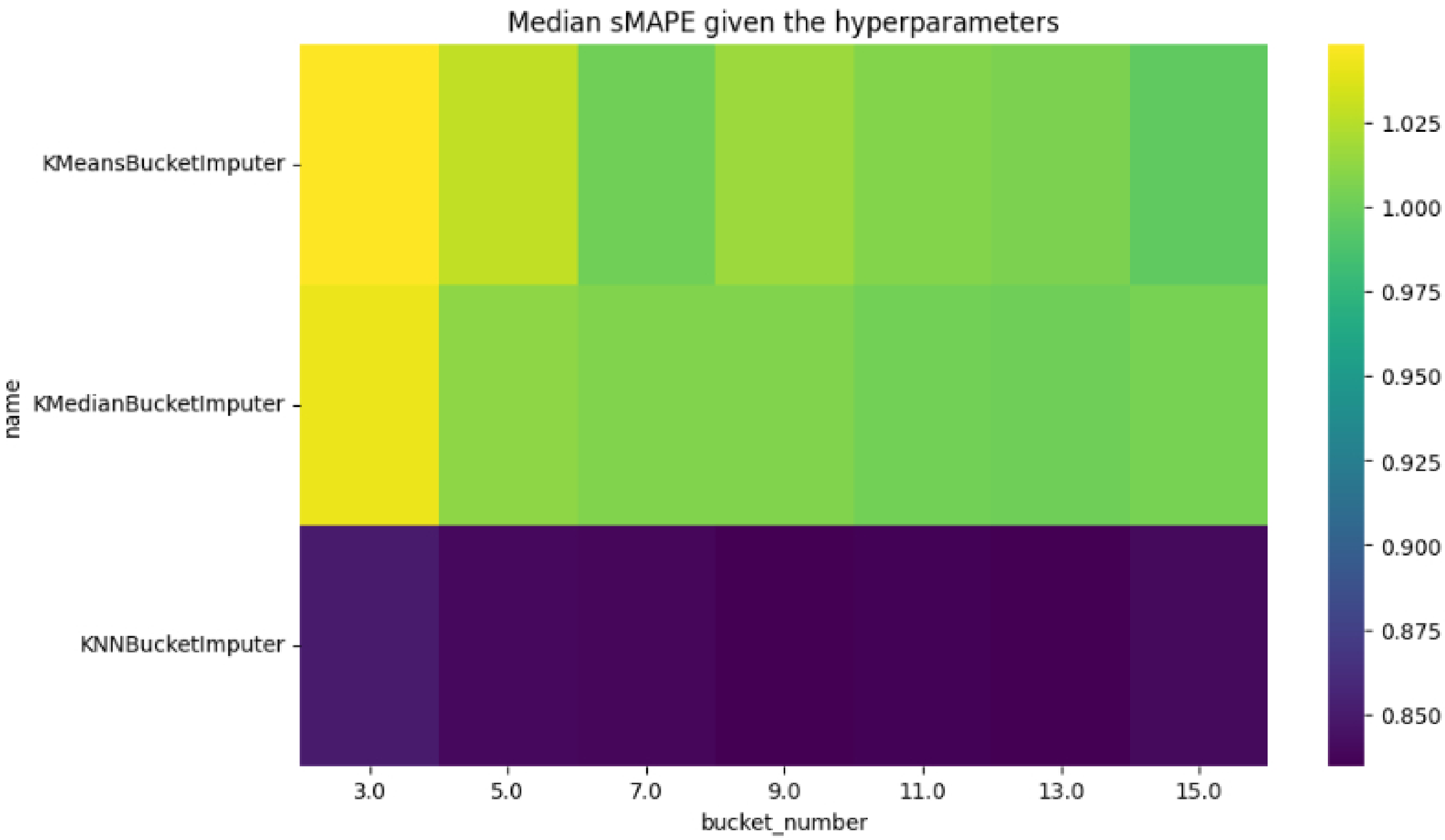

We can also see that the KNNImputer visibily outperform KMeans and KMedoids.




The bucket number is a strong factor influencing the the model performance. It converges after around 9 buckets.




Eachmodel has slightly different peaks for K. Kmeans peaks at K being 7 and 15 , Kmedian at K being 13 and KNN at K of roughly 9 to 13.



# Decision

## Update : 09.01.2024

For the final experiment we choose K as 9, 13 and 7 for KNN, Kmedian and Kmean, respectively.