

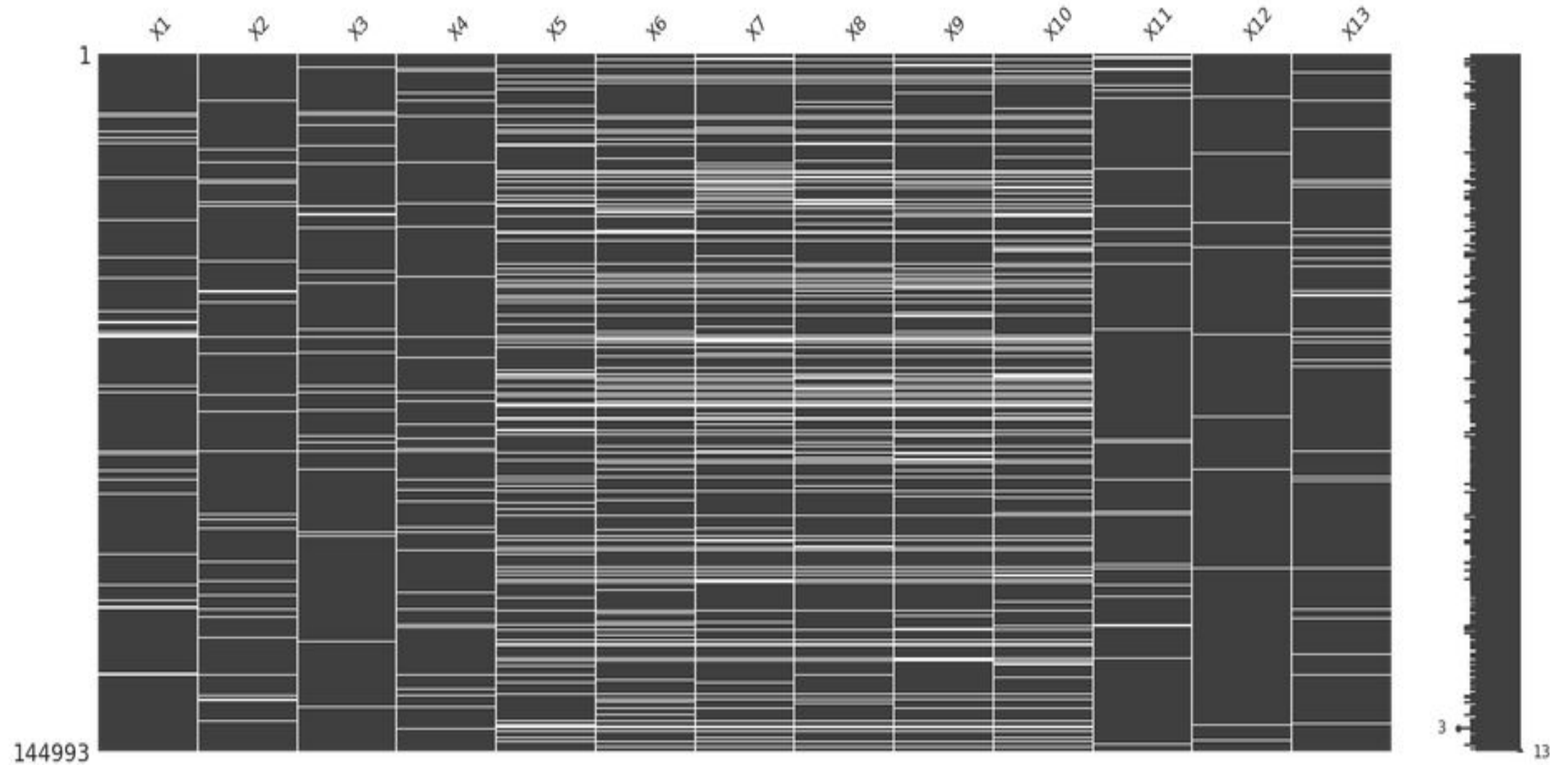
Missing Clinical Data Imputation

Group: 12

Anusha Manur (amanur@ncsu.edu)

Kartik Shah (kcshah@ncsu.edu)

DATASET



DATASET

Analyte	% Missing Values
X1	5.31
X2	5.47
X3	5.53
X4	5.39
X5	16.53
X6	19.15
X7	19.29
X8	18.62
X9	18.84
X10	19.42
X11	4.88
X12	4.83
X13	6.85

Methods Used for Imputation:

- 1) Mean Imputation
- 2) Back Fill
- 3) Back Fill Front Fill
- 4) 4Mean BackFrontFill
- 5) Multiple Imputation by Chained Equations with KNN Regressor and Linear Regressor
- 6) Deep Learning
- 7) K-Nearest Neighbors

1) Mean imputation

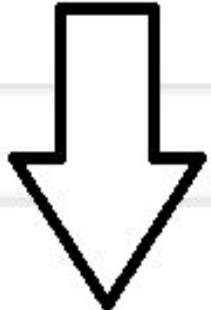
	col1	col2	col3	col4	col5
0	2	5.0	3.0	6	NaN
1	9	NaN	9.0	0	7.0
2	19	17.0	NaN	9	NaN

`mean()`

	col1	col2	col3	col4	col5
0	2.0	5.0	3.0	6.0	7.0
1	9.0	11.0	9.0	0.0	7.0
2	19.0	17.0	6.0	9.0	7.0

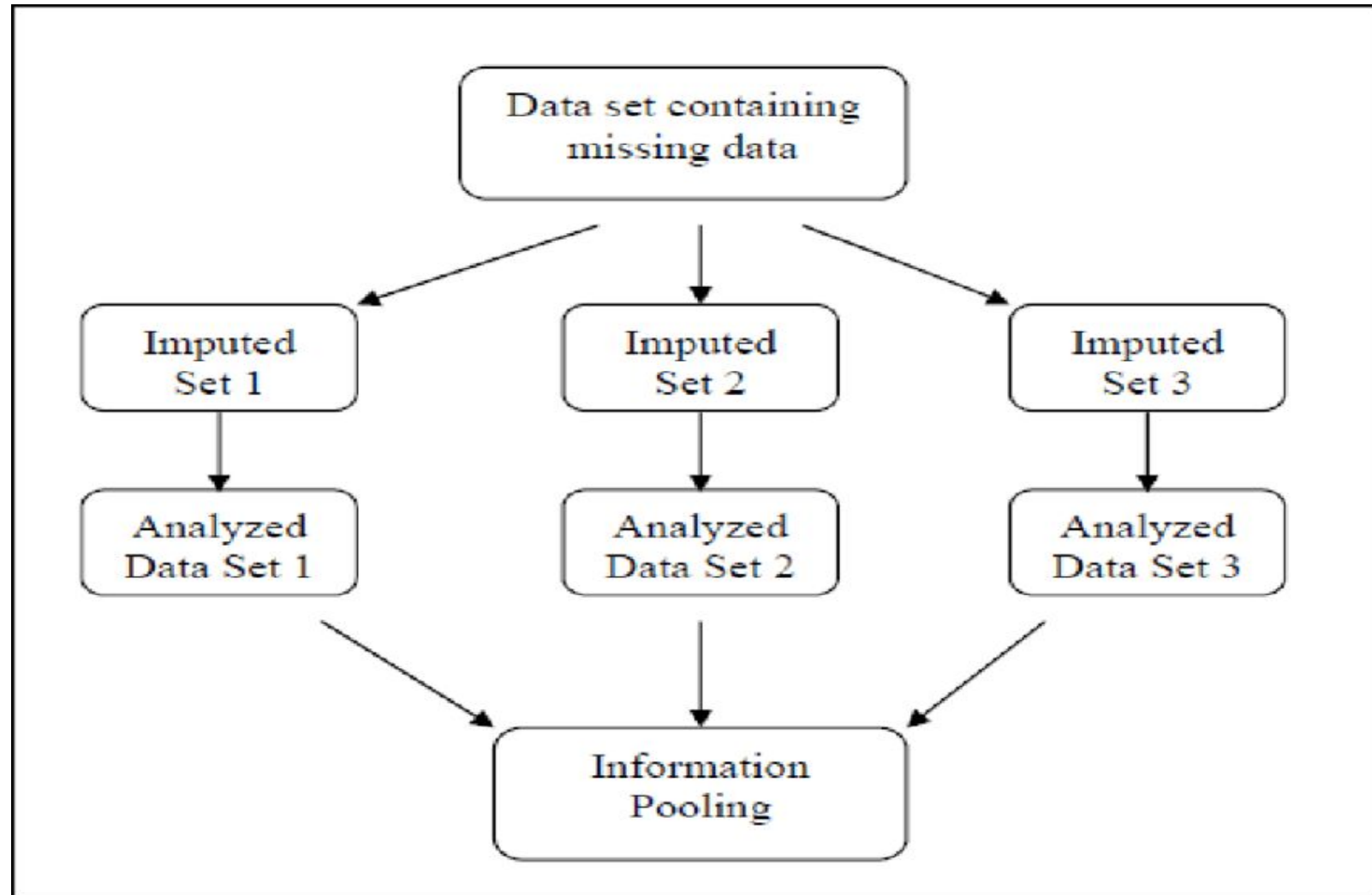
2) Back Fill

0	300
1	NaN
2	150

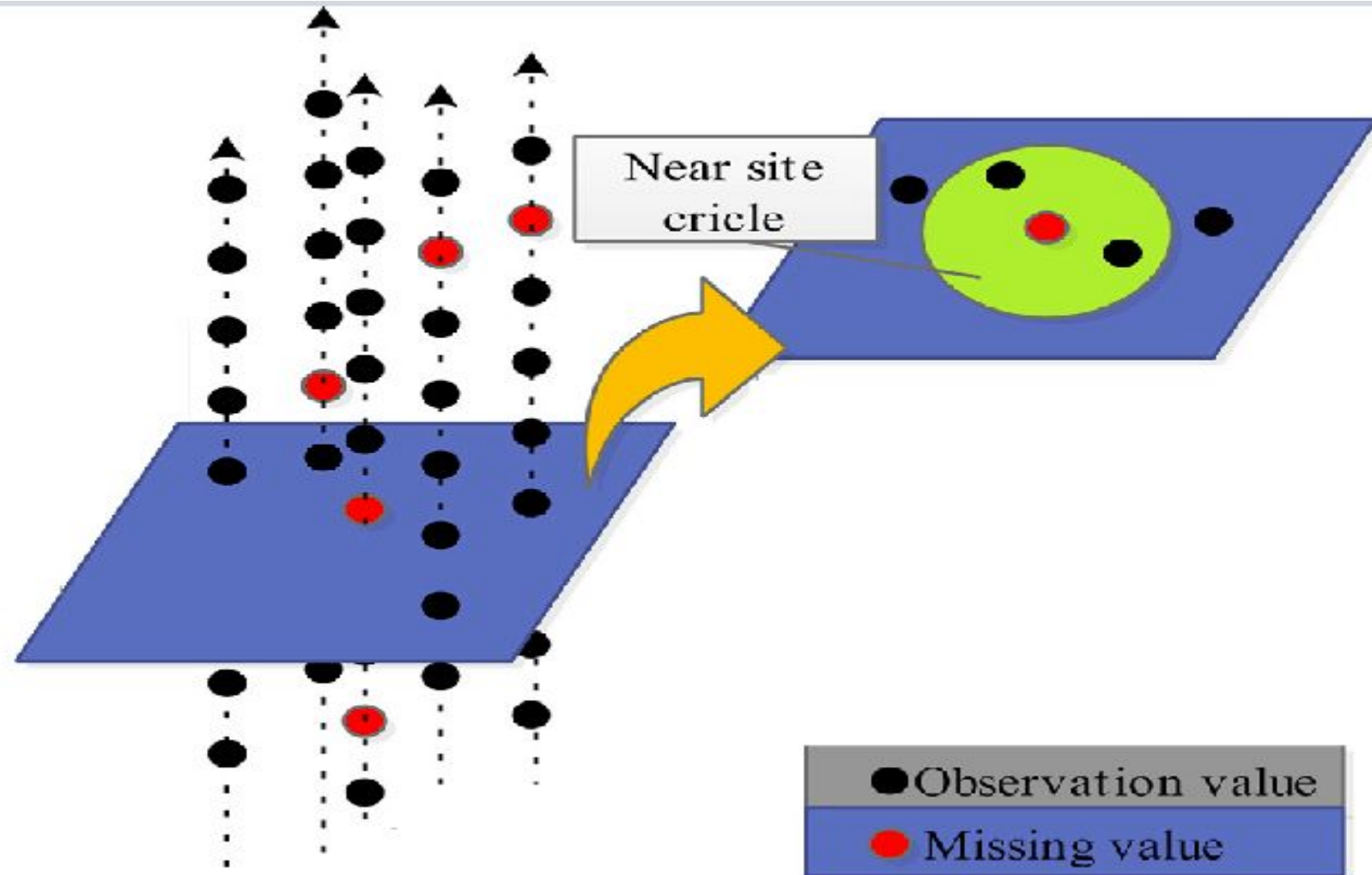


0	300
1	150
2	150

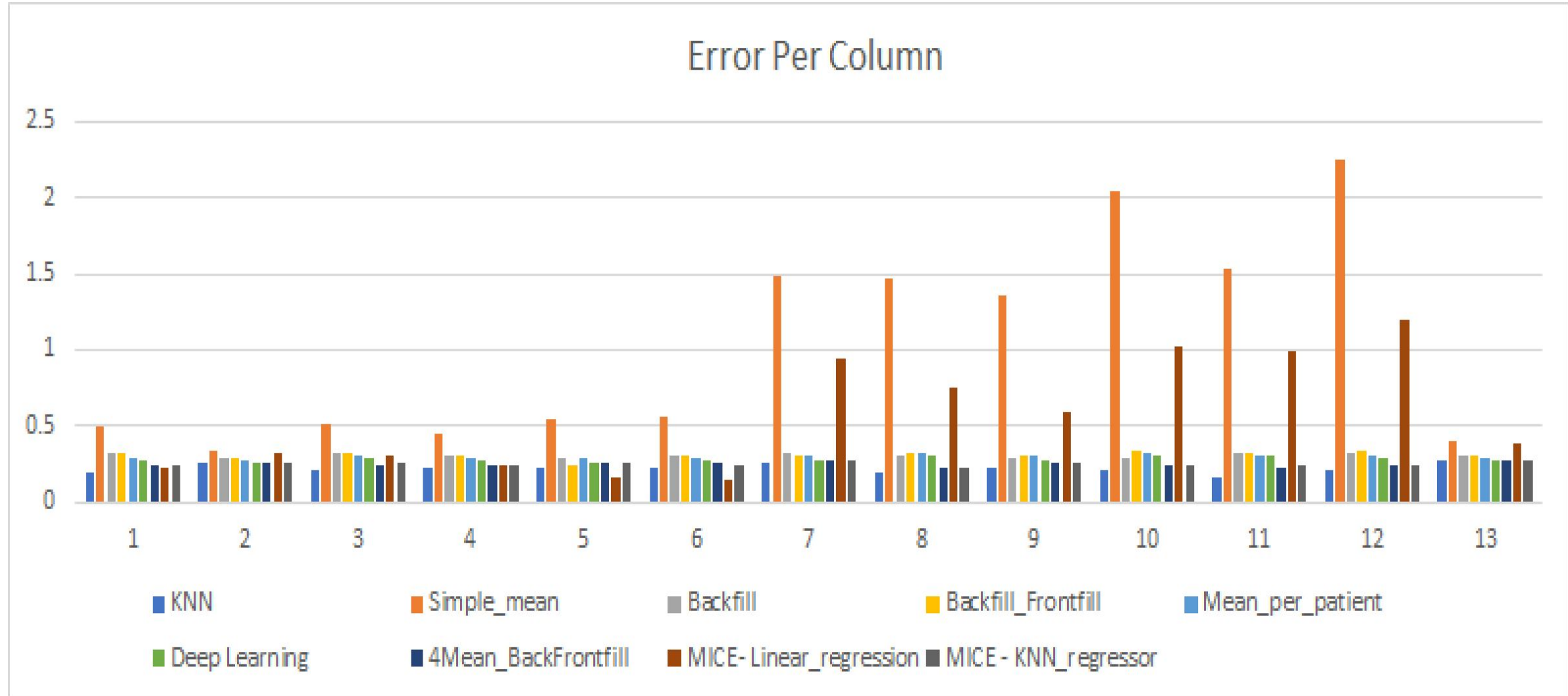
MICE



KNN

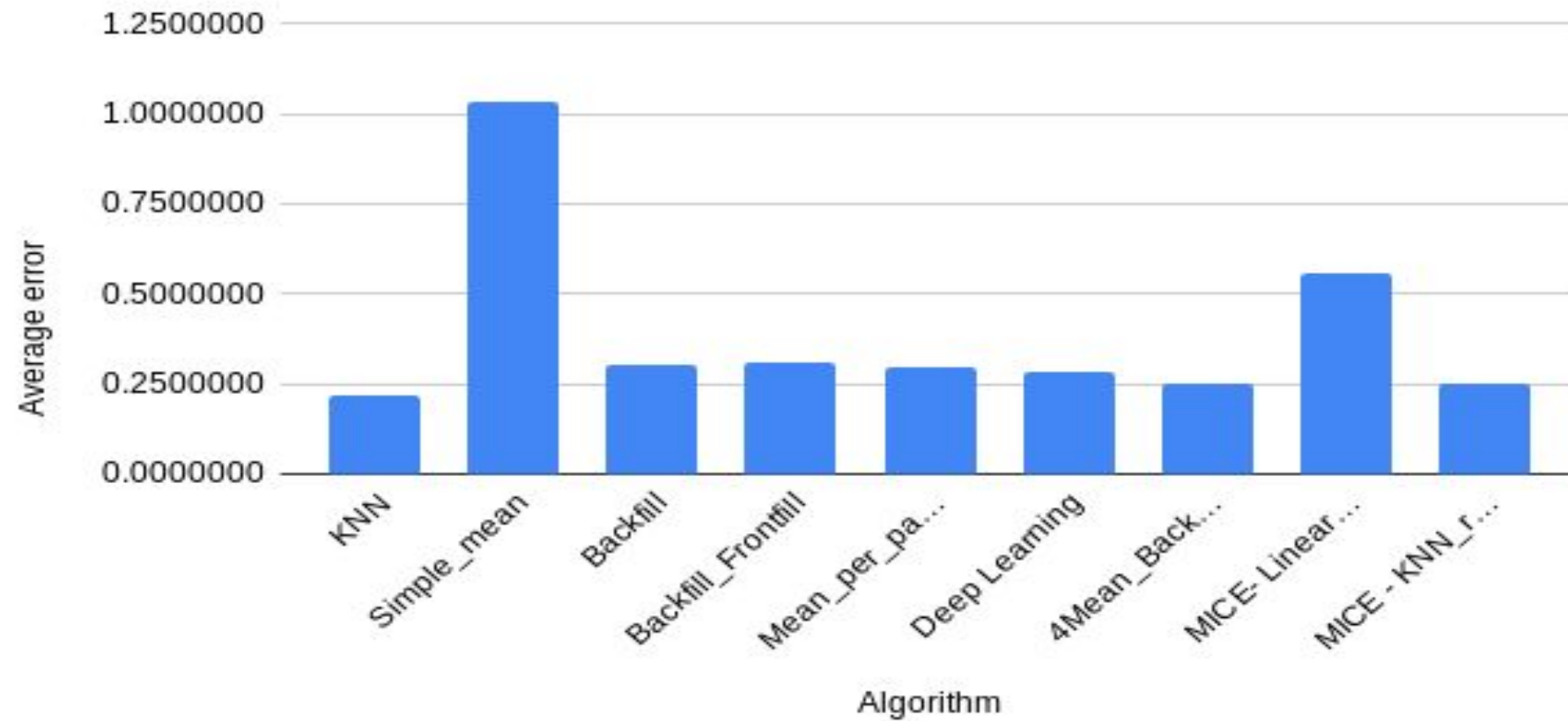


Error Results Per Column



Overall error

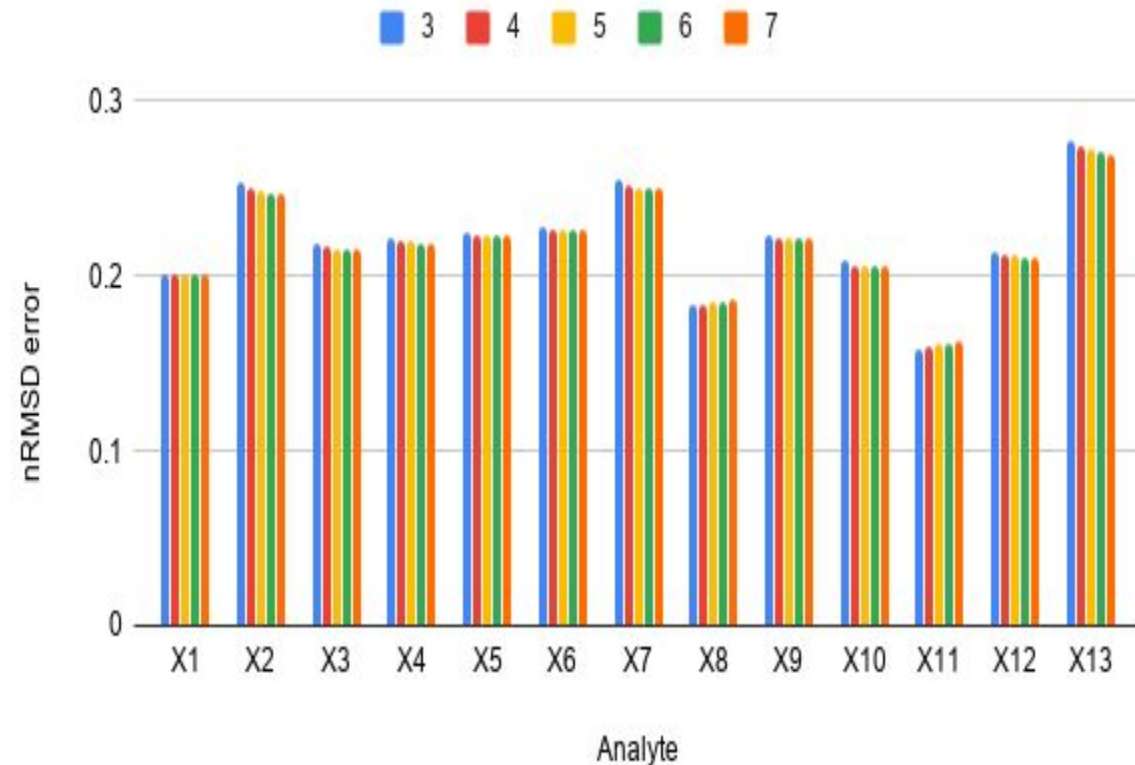
Average RMSD overall error



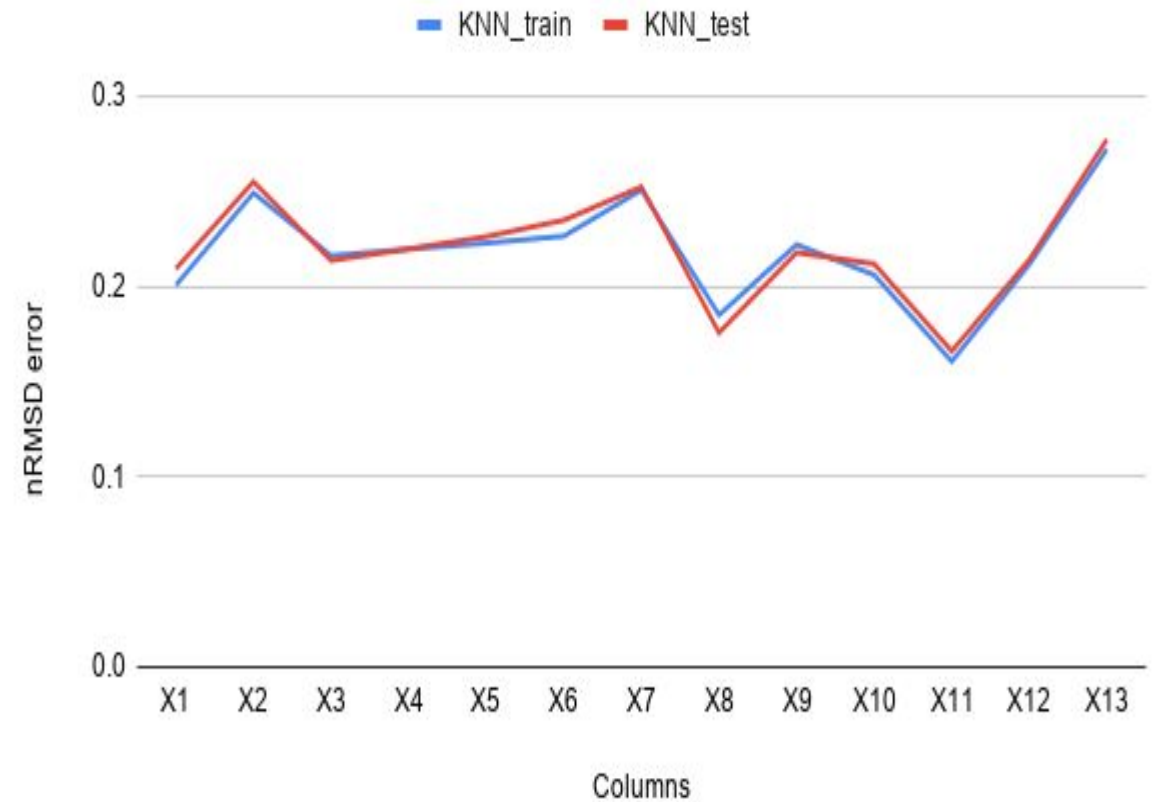
KNN Implementation analysis

KNN error per column

Varying number of nearest neighbors



KNN -Training vs Testing error performance



Conclusion

- In the training dataset, we vary K, the number of nearest neighbors to tune the KNN algorithm, using mean to impute missing values. We chose optimal value of K=5 which results in minimal error. The total error we got 0.2185.
- In testing dataset we got an average nRSMD error of 0.2209.

Leaderboard Results

- Our model performed better compared to 3D-MICE
- We rank 8 on the leaderboard