Testing New Data Imputation Method

a) Data Imputation Method

The dataset obtained contained a great deal of missing values as can be seen in Figure I and II. Therefore, the author decided to explore another method of data imputation.

DATEPRD	0.000000 %
WELL_BORE_CODE	0.000000 %
NPD_WELL_BORE_CODE	0.000000 %
NPD_WELL_BORE_NAME	0.000000 %
NPD_FIELD_CODE	0.000000 %
NPD_FIELD_NAME	0.000000 %
NPD_FACILITY_CODE	0.000000 %
NPD_FACILITY_NAME	0.000000 %
ON_STREAM_HRS	1.822950 %
AVG_DOWNHOLE_PRESSURE	42.561085 %
AVG_DOWNHOLE_TEMPERATURE	42.561085 %
AVG_DP_TUBING	42.561085 %
AVG_ANNULUS_PRESS	49.533069 %
AVG_CHOKE_SIZE_P	42.951260 %
AVG_CHOKE_UOM	41.403352 %
AVG_WHP_P	41.441730 %
AVG_WHT_P	41.499296 %
DP_CHOKE_SIZE	1.880517 %
BORE_OIL_VOL	41.403352 %
BORE_GAS_VOL	41.403352 %
BORE_WAT_VOL	41.403352 %
BORE_WI_VOL	63.502622 %
FLOW_KIND	0.000000 %
WELL_TYPE	0.000000 %

Figure I : Missing Values in Volve Dataset

Wellbore ID	0.000000 %
Date	0.000000 %
Hours Online	0.168447 %
Av. WHT (Deg C)	1.309077 %
Av. WHP (bar)	7.339494 %
Av. DHT (Deg C)	3.104245 %
Av. DHP (bar)	3.214939 %
Platform Choke %	72.408316 %
Oil (m3)	0.168447 %
Gas (m3)	0.168447 %
Produced Water (m3)	0.168447 %

Figure II : Missing Values in Kyle Master Dataset

The new data imputation method is described in Figure III. The Volve and Kyle Master dataset would be split into two smaller datasets – empty values dataset and filled values dataset. The filled values dataset is the dataset where all the empty values were dropped whereas the empty values dataset contains any row that contain empty values. The empty values dataset is then scrutinized further to eliminate the rows where every value is empty. On the other hand, the filled values dataset is used to train a gradient boosting and random forest model. Afterwards the trained model is used to fill in the missing values.

b) Evaluating the New Data Imputation Method

This new method was then evaluated the same way as the other data imputation methods were as shown in Figure IV. The metric used to evaluate the performance is RMSE. The lower the value of this metric, the better the performance of the model. The results of the evaluation are shown in Table I, Table II, and Table III.

Table I : Forward Filling Method

	Gradient Boosting	Random Forest
	RMSE (m3)	
Oil Prediction	128	183
Gas Prediction	52,701	78,021

Table II : Median Imputation Method

	Gradient Boosting	Random Forest
	RMSE (m3)	
Oil Prediction	160	184
Gas Prediction	67,436	84,612

Table III : New Data Imputation Method

	Gradient Boosting	Random Forest
	RMSE (m3)	
Oil Prediction	157	204
Gas Prediction	73,905	95,762

Oil Prediction

When the gradient boosting model is used, the new data imputation method is shown to give better results than the median imputation method. However, it performed worse when compared to the forward filling method. On the other hand, the new method performed worse than the median imputation and forward filling method when the random forest model is used.

Gas Prediction

The new data imputation method performed worse than the other two methods when both models were used.

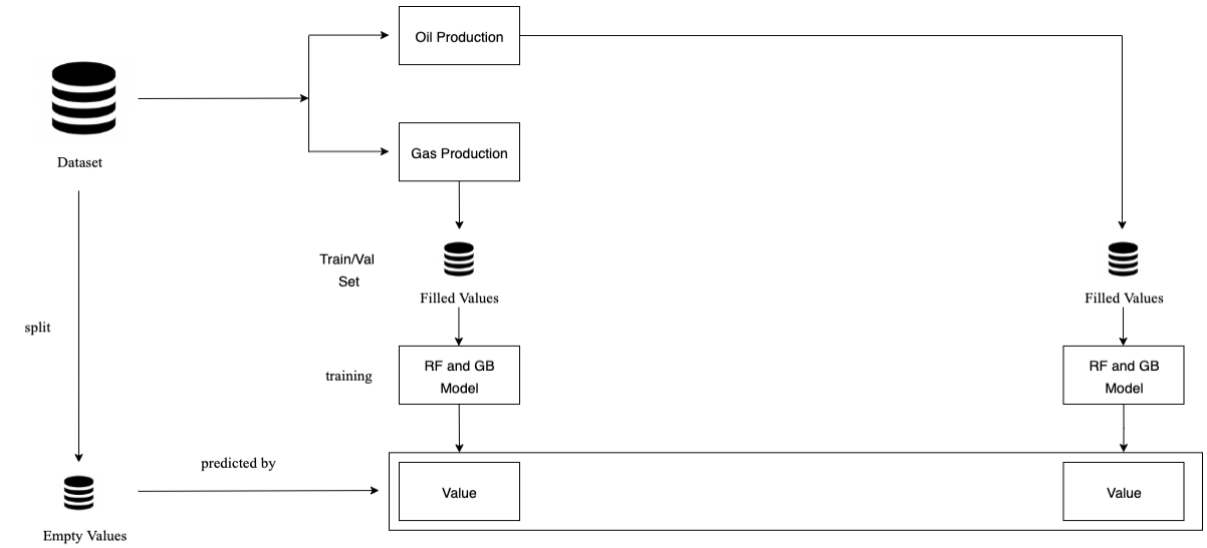

Figure III : New Data Imputation Methodology

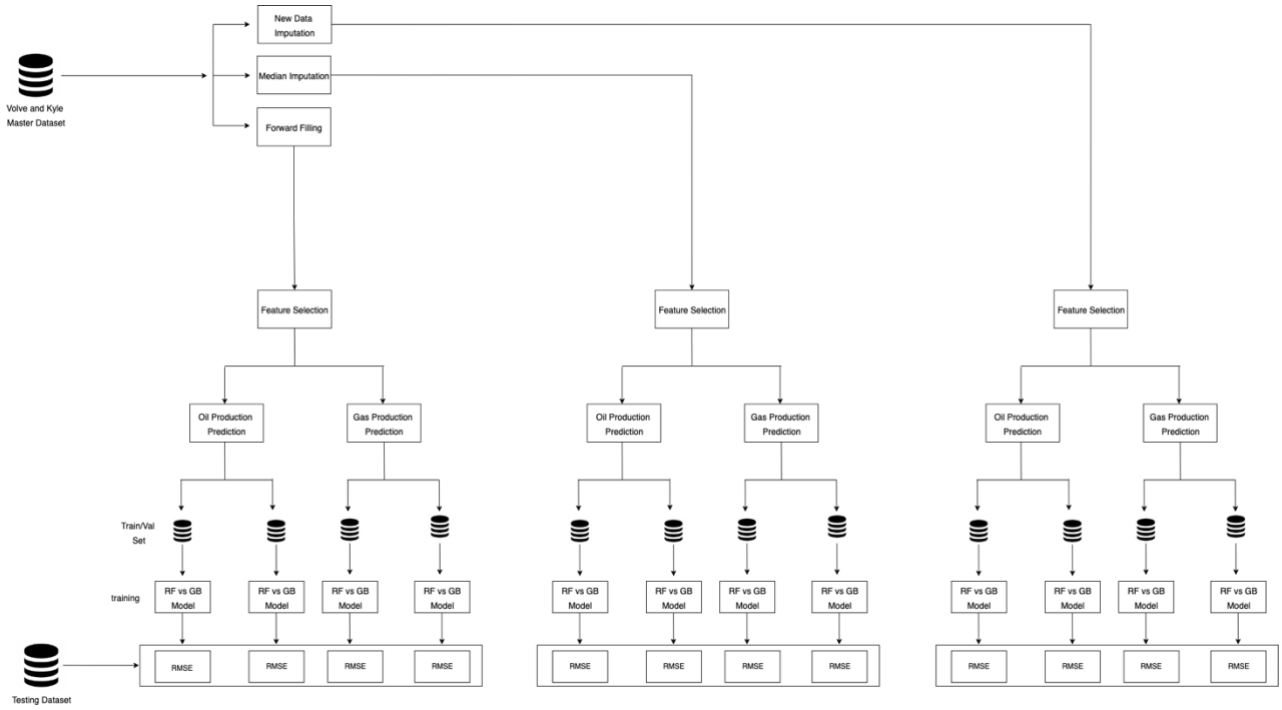

Figure IV : Methodology