Maternal and Child Health Monitoring in LMICs

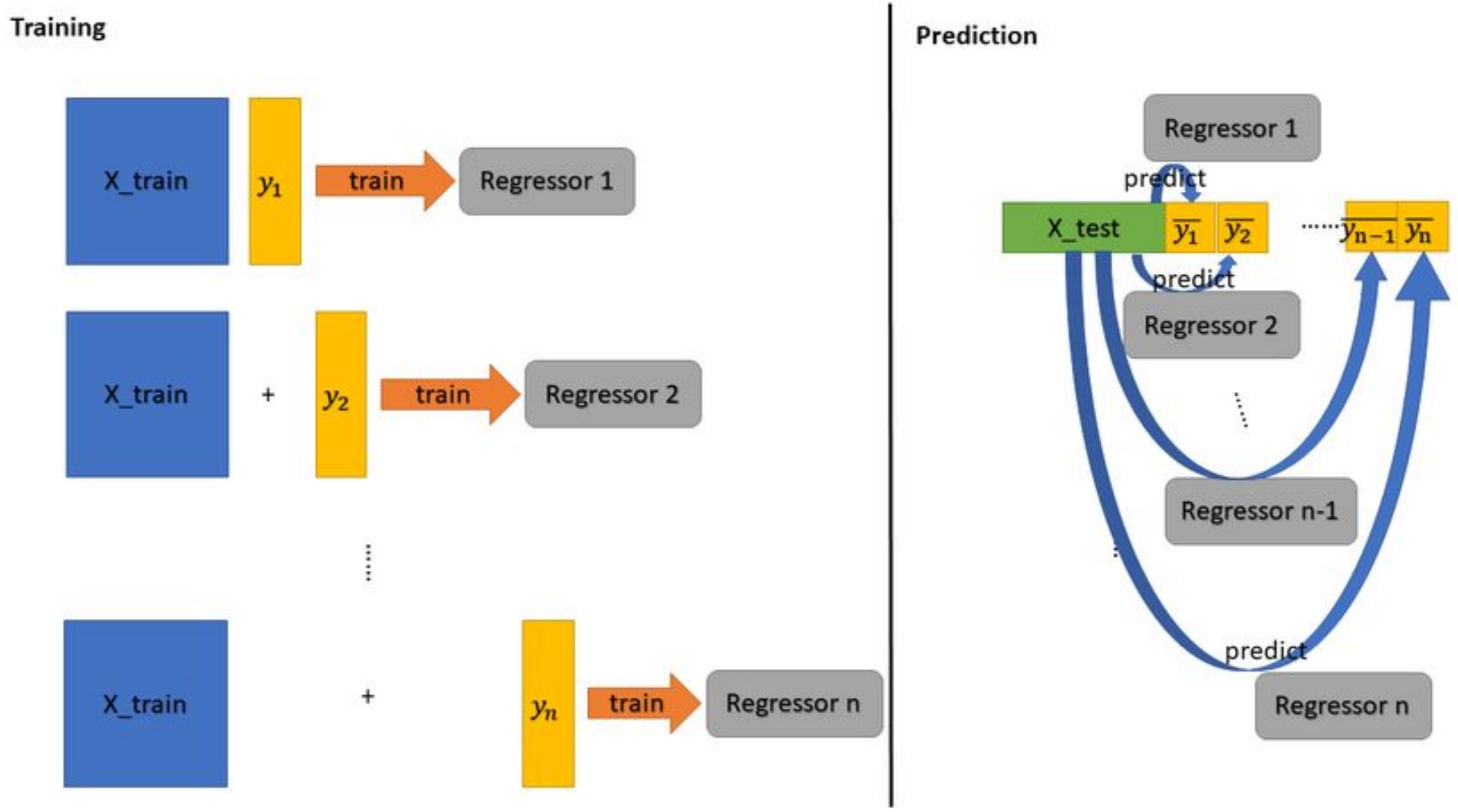
Using ML on satellite and geotagged data sources

- ***Group 4***

Speaker

- *Debajyoti Dasgupta (18CS30051)*
- *Rushil Venkateswar (20CS30045)*
- *Piran Karkaria (20IM3FP52)*

Problem Statement (Multi-Output Regression)



Interpretation of the Data

DHS Survey Data

Google Earth Engine

Tabular Data

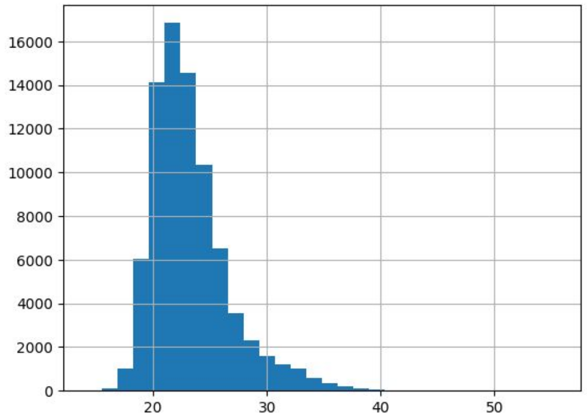
Interpretation of the Data (contd.)

	importance
URBAN_RURA_R	0.223737
Retrieved_Temperature_Profile_Mean_Mean_950_mean@MODIS/061/MOD08_M3&timestamped	0.054609
LONGNUM	0.040629
ozone_tropospheric_vertical_column_median@COPERNICUS/S5P/OFFL/L3_O3_TCL&timestamped	0.036747
Retrieved_Temperature_Profile_Mean_Mean_920_min_max@MODIS/061/MOD08_M3&timestamped	0.021586
gaugeRelativeWeighting_median@NASA/GPM_L3/IMERG_MONTHLY_V06&timestamped	0.018502
key3	0.017434
DHSYEAR	0.017129
Retrieved_Temperature_Profile_Mean_Mean_780_min_max@MODIS/061/MOD08_M3&timestamped	0.014571
Retrieved_Temperature_Profile_Mean_Mean_920_min_min@MODIS/061/MOD08_M3&timestamped	0.012517
randomError_max_min@NASA/GPM_L3/IMERG_MONTHLY_V06&timestamped	0.010403
ozone_tropospheric_mixing_ratio_median@COPERNICUS/S5P/OFFL/L3_O3_TCL&timestamped	0.008742
Retrieved_Temperature_Profile_Mean_Mean_780_min_min@MODIS/061/MOD08_M3&timestamped	0.008024
gaugeRelativeWeighting_mean@NASA/GPM_L3/IMERG_MONTHLY_V06&timestamped	0.007771
ADM1DHS	0.007671
Retrieved_Temperature_Profile_Mean_Mean_20_median@MODIS/061/MOD08_M3&timestamped	0.006344
gaugeRelativeWeighting_max_min@NASA/GPM_L3/IMERG_MONTHLY_V06&timestamped	0.006278
Cloud_Top_Temperature_Std_Deviation_Mean_median@MODIS/061/MOD08_M3&timestamped	0.005788
ozone_tropospheric_mixing_ratio_min_min@COPERNICUS/S5P/OFFL/L3_O3_TCL&timestamped	0.005735
ozone_tropospheric_mixing_ratio_min_max@COPERNICUS/S5P/OFFL/L3_O3_TCL&timestamped	0.005240

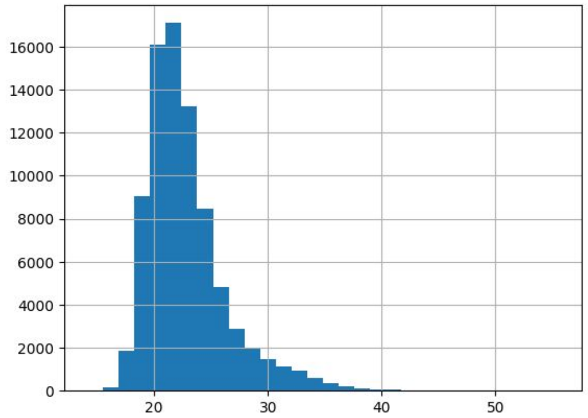
NaN values in training labels

```
Mean_BMI                         18517
Median_BMI                       18517
Unmet_Need_Rate                   1867
Under5_Mortality_Rate            29623
Skilled_Birth_Attendant_Rate     33279
Stunted_Rate                     58047
dtype: int64
```

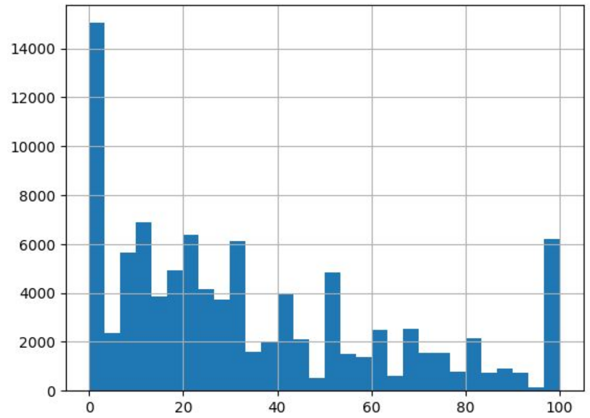
Interpretation of the Data (contd.)



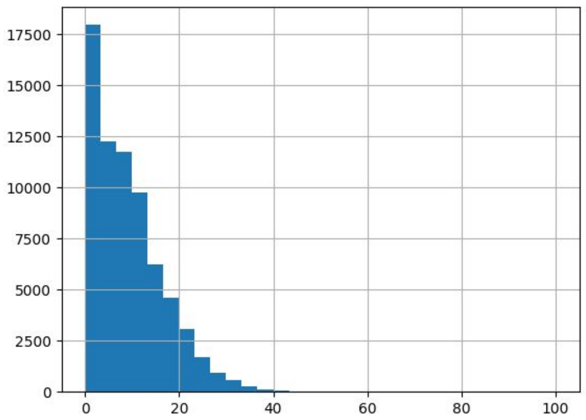
Mean_BMI



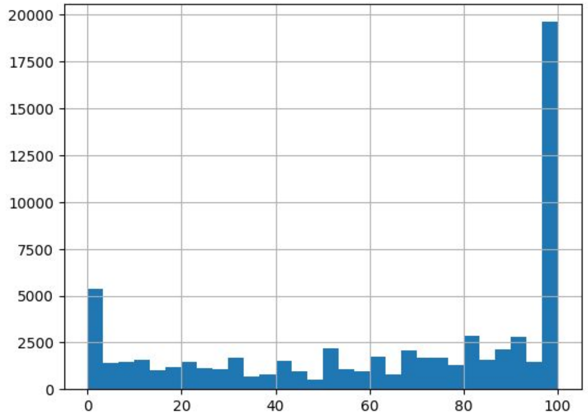
Median_BMI



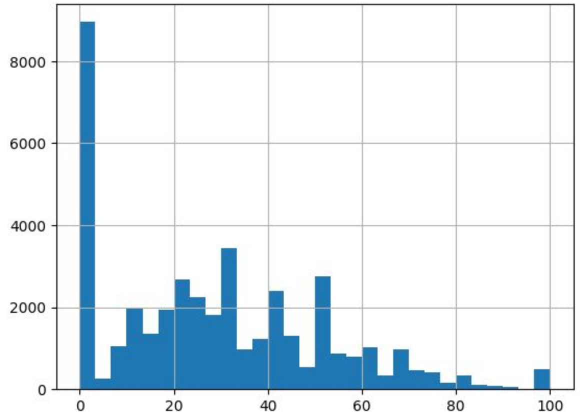
Unmet_Need_Rate



Under5_Mortality_Rate

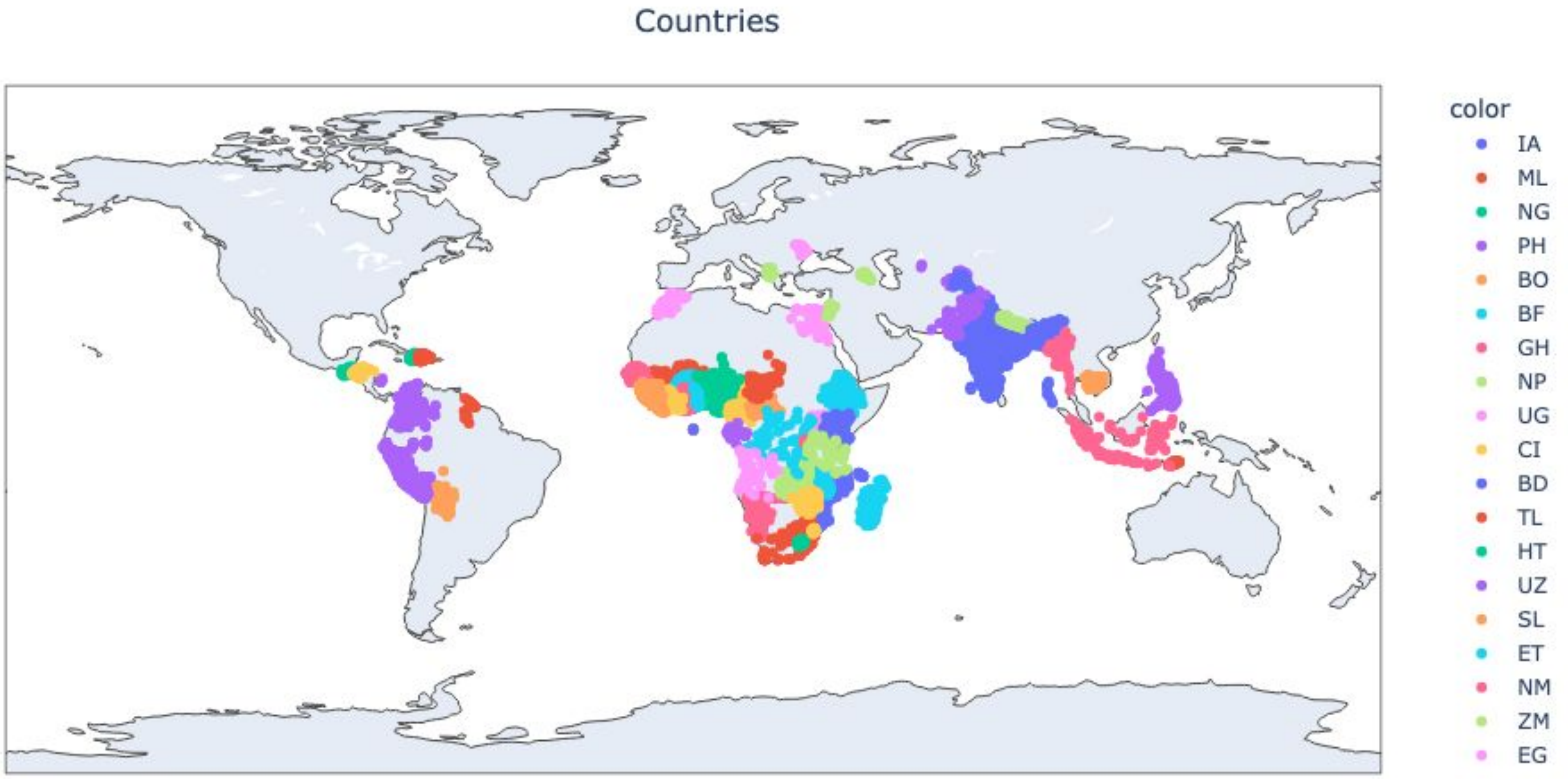


Skilled_Birth_Attendant_Rate

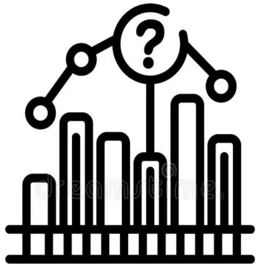


Stunted_Rate

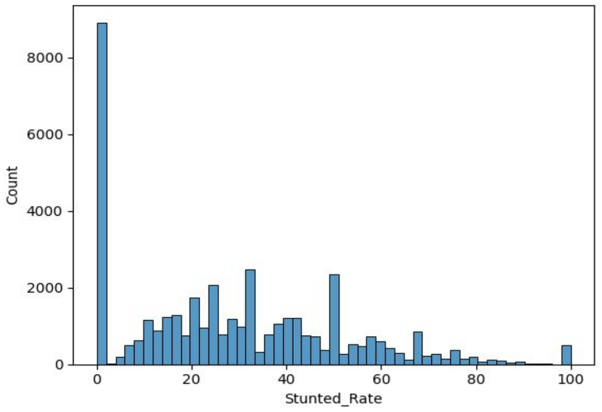
Interpretation of the Data (contd.)



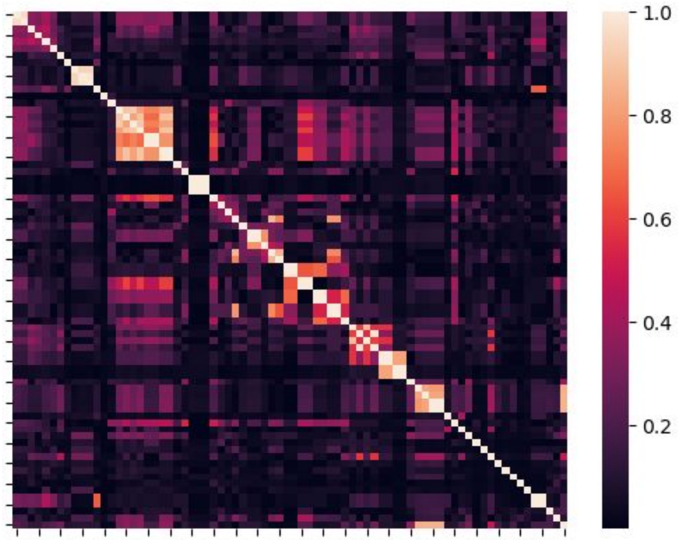
Issues with given Data



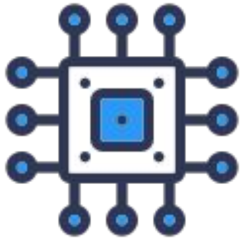
100% data missing in few columns



Data imbalance and missing data in some labels



Most columns have high correlation, only 2000 non-correlated columns



Loading 8GB dataset into memory infeasible on Kaggle

Data Preprocessing

DHSID	new_ind	F21	F22	F23	CCFIPS	ADM1FIPS	ADM1FIPSNA
IA201400110884	IA_2014_00110884	NaN	NaN	NaN	IN	NaN	NaN
IA201400051523	IA_2014_00051523	NaN	NaN	NaN	IN	NaN	NaN
IA201400150534	IA_2014_00150534	NaN	NaN	NaN	IN	NaN	NaN
ML200600000390	ML_2006_00000390	NaN	NaN	NaN	ML	ML06	Sikasso
NG200300000174	NG_2003_00000174	NaN	NaN	NaN	NI	NaN	NaN
...	...	...	...	...	...	...	...
PE200400001013	PE_2004_00001013	NaN	NaN	NaN	PE	PE18	Moquegua
EG200800000765	EG_2008_00000765	NaN	NaN	NaN	EG	EG18	Bani Suwayf
PE200000000348	PE_2000_00000348	NaN	NaN	NaN	PE	PE08	Cusco
PE200000000614	PE_2000_00000614	NaN	NaN	NaN	PE	PE13	La Libertad
DR200700001622	DR_2007_00001622	NaN	NaN	NaN	DR	NaN	NaN

12098 rows × 7 columns

- Drop the columns with completely distinct values, like in case of **new_ind**
- Drop the columns which contain any NaN value (except for prediction columns) like in case of **F21, F22, F23** etc.
    - The motivation behind this is that our preliminary analysis reveals that most of the columns that have at least one NaN value has almost >= 50% missing values
- Among the rows which are duplicated maintain only the last entry of the survey

Data Preprocessing

DHSID	DHSYEAR	DHSCLUST	LATNUM	LONGNUM	Mean_BMI	Median_BMI	Unmet_Need_Rate	Under
AL200800000001	2008	1.0	40.822652	19.838321	24.12	25.28	50.00	
AL200800000002	2008	2.0	40.696846	20.007555	23.04	21.98	7.69	
AL200800000004	2008	4.0	40.798931	19.863338	26.74	26.57	7.69	
AL200800000006	2008	6.0	40.711349	19.935309	27.58	28.08	0.00	
AL200800000010	2008	10.0	40.698522	19.950300	24.23	23.77	20.00	
...	...	...	...	...	...	...	...	
ZW201500000395	2015	395.0	-17.166506	29.718371	21.92	21.08	6.25	
ZW201500000396	2015	396.0	-17.915288	31.156115	23.16	22.14	33.33	
ZW201500000397	2015	397.0	-18.379501	31.872287	24.33	22.61	11.11	
ZW201500000398	2015	398.0	-16.660612	29.850649	23.70	21.44	10.53	
ZW201500000400	2015	400.0	-17.859114	31.797626	25.84	24.76	7.69	

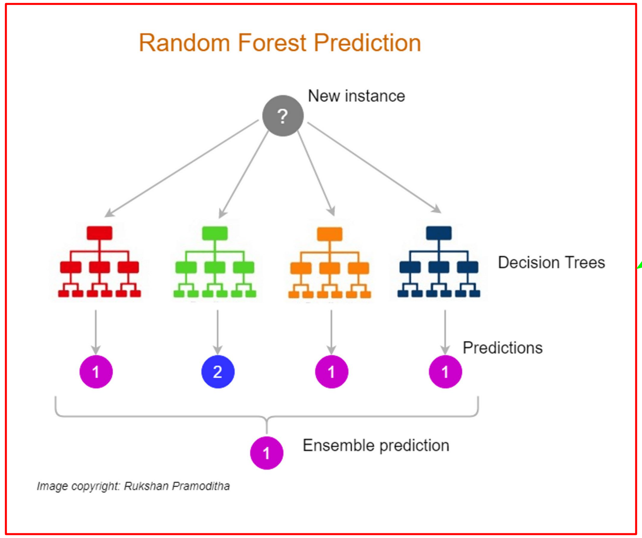
- Drop the columns from the training data which are already present in the `**gee_features**` files as they are not providing any additional information that we didn't already have
- Apply KNN Imputation on the prediction labels in the training data, not to avoid much of the data
    - Intuition behind this was that, half of the columns had more than 85% of the data available, imputation didn't make much effect on it
    - Rest half had less than 40% data and hence we were losing a lot of data, imputation improves the performance

Data Preprocessing

DHSID	DATUM	DHSCC	DHSREGNA	SOURCE	URBAN_RURA	key1
IA201400110884	WGS84	IA	Kachchh	GPS	U	IA
IA201400051523	WGS84	IA	Katihar	GPS	R	IA
IA201400150534	WGS84	IA	Deoghar	GPS	R	IA
ML200600000390	WGS84	ML	SIKASSO	GPS	U	ML
NG200300000174	WGS84	NG	North West	GPS	U	NG
...	...	...	...	...	...	...
PE200400001013	WGS84	PE	moquegua	GAZ	U	PE
EG200800000765	WGS84	EG	upper egypt rural	GPS	R	EG
PE200000000348	WGS84	PE	cusco	GPS	R	PE
PE200000000614	WGS84	PE	la libertad	GPS	R	PE
DR200700001622	WGS84	DR	iii	GPS	U	DR

- One hot encode the categorical columns, since the the machine learning model cannot easily handle text data
- Categorical data encoded in one hot form converts the textual labels into numerical form and improves the model performance
- This is also a very standard technique to handle text data

Methodology

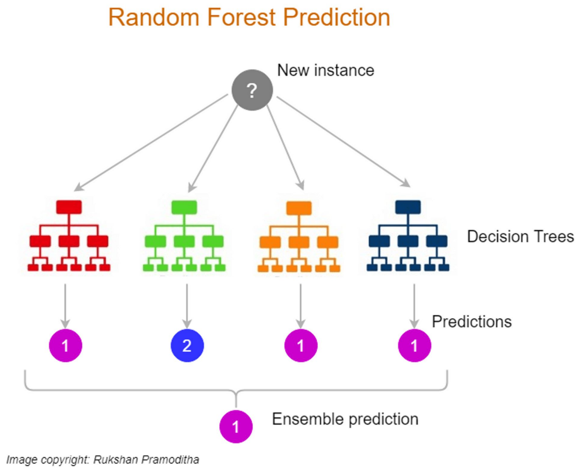


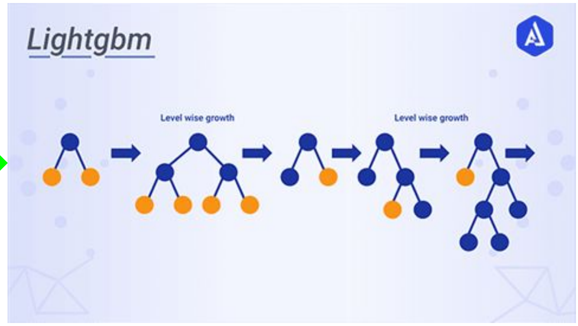
Base model for decision taking is built using the Random Forest Regressor

Random Forest Regressor

Methodology

Boost on Error 1





First level of Boosting of Error is performed using LightGBM which is a Gradient Boosting method which is based on leaf-wise tree growth

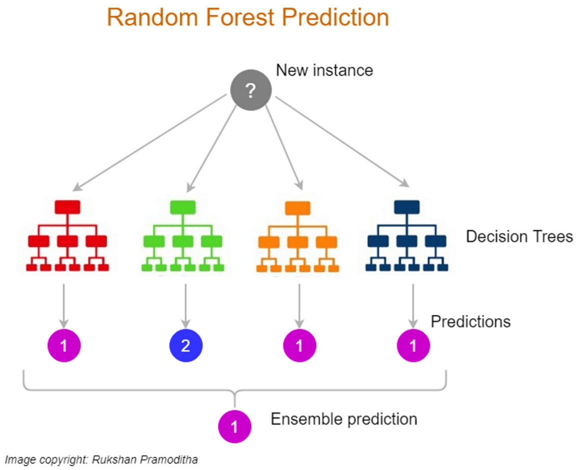
Random Forest Regressor

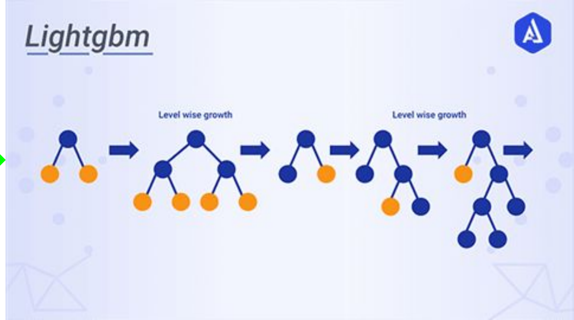
Light GBM Regressor

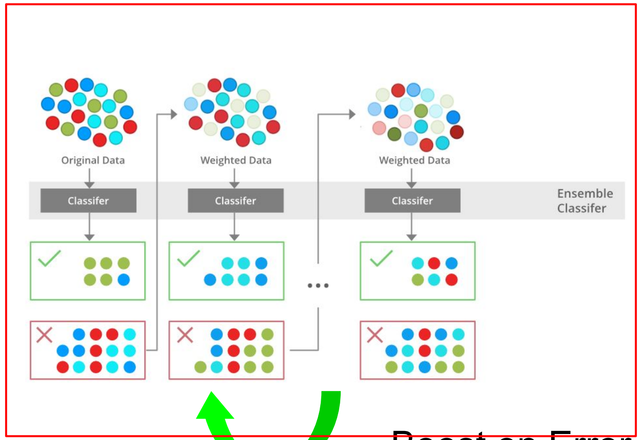
Methodology

Boost on Error 1

Boost on Error 2







Ensembled Output

Boost on Error 3

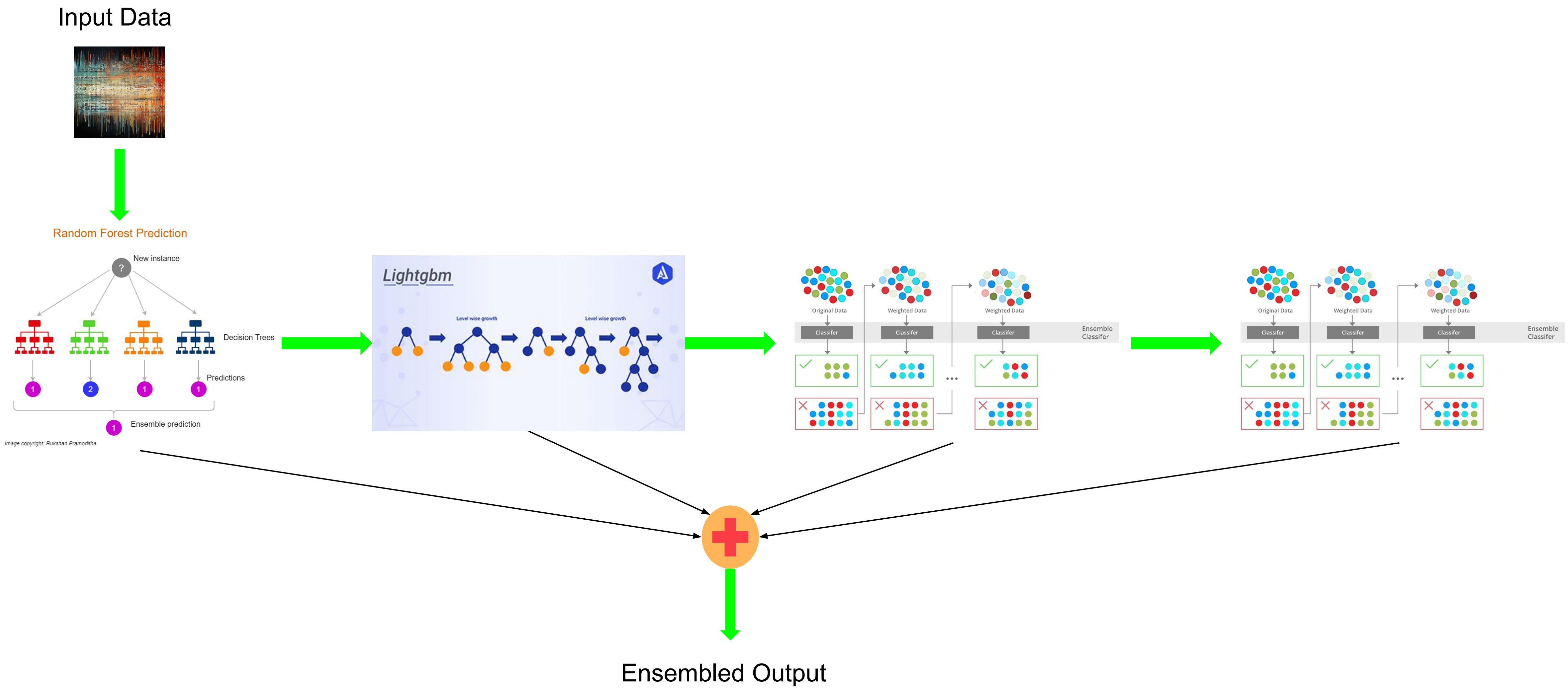
Random Forest Regressor

Light GBM Regressor

XGBoost Regressor

Finally 2 levels of boosting on errors is performed using XGBoost method

Methodology



Results and Inference

10.759

Error Prediction Ensemble Model

Model built to predict the residuals of regression using ensembling

- Reduced Bias
- Ensemble Averaging/ Regularization
- Compensating for model assumptions

10.959

Gradient Boosting Ensemble Model

Random Forest + Histogram Gradient Boosting + XGBoost + CatBoost + LightGBM

- Robustness to Noise and Outliers
- Regularization
- Grid Search to find best ensemble weights

11.110

Random Forest Regressor

Data pre-processing such as one-hot encoding, PCA and dropping of null values

- Filtering out most important features
- Manual fine-tuning for hyperparameters

12.273

CNN-based model

Utilizing the pre-trained ResNet-50 architecture to predict continuous values

- ResNet backbone to utilize trained filters
- Fully connected layer instead of classifier

Results and Inference

Public Private

This leaderboard is calculated with approximately 20% of the test data. The final results will be based on the other 80%, so the final standings may be different.

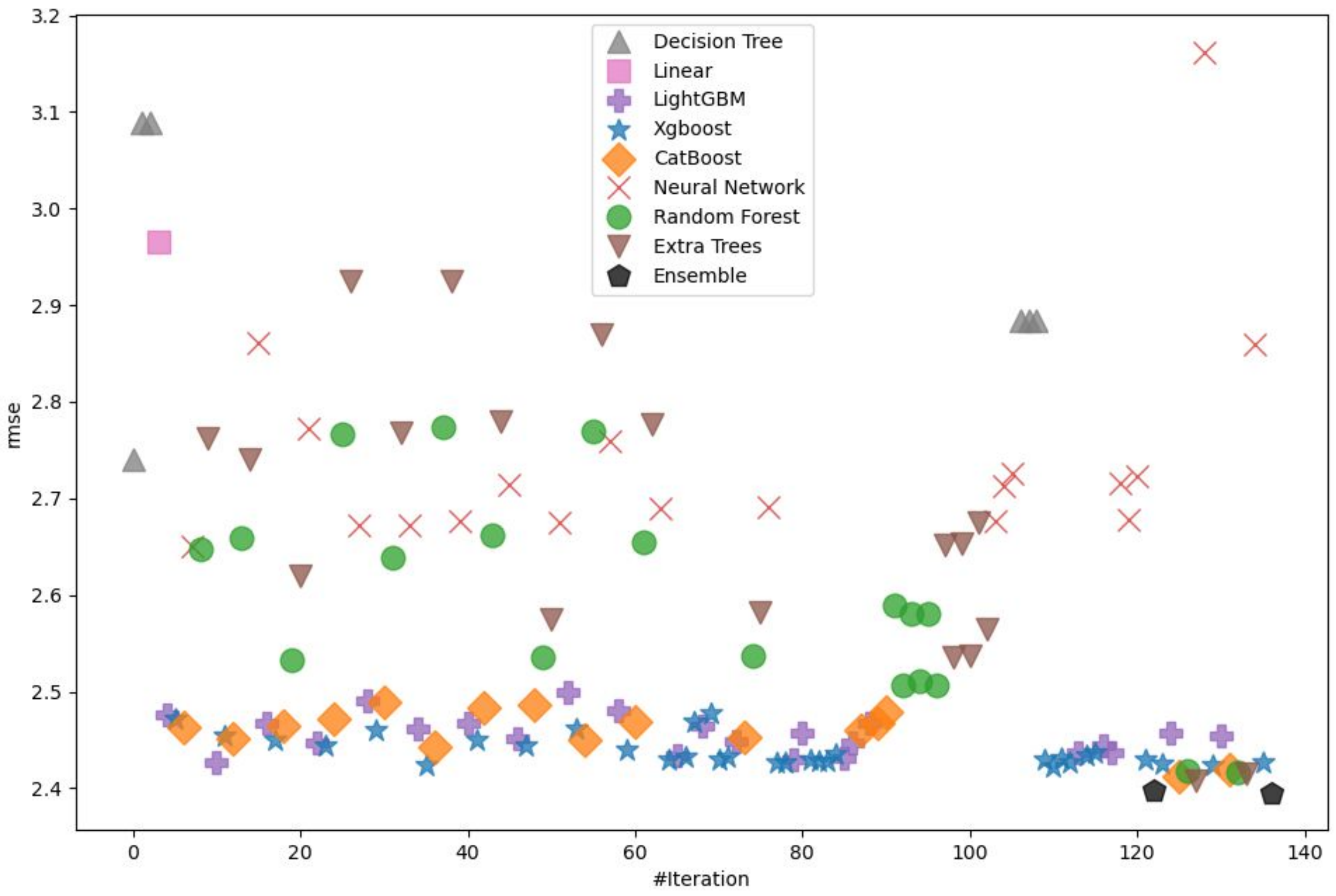
#	Team	Members	Score	Entries	Last	Join
1	**Team_4_T4**		10.75958	169	2h	

Your Best Entry!
Your submission scored 10.94818, which is not an improvement of your previous score. Keep trying!

#	Team	Members	Score	Entries	Last	Join
2	ZSK_T23		10.91713	103	19h	
3	DeepKGP_T38		10.95148	323	19h	
4	The Data Wraiths_T26	ZXR	11.00042	422	2h	
5	CeraVe_T57		11.00725	44	6h	
6	TrioML_T46		11.05028	41	1h	
7	GeoCare		11.06081	4	11d	
8	Byte Force		11.10429	50	12d	
9	MechML_T47		11.11084	172	4h	
10	LabRats_T25		11.16194	19	1mo	

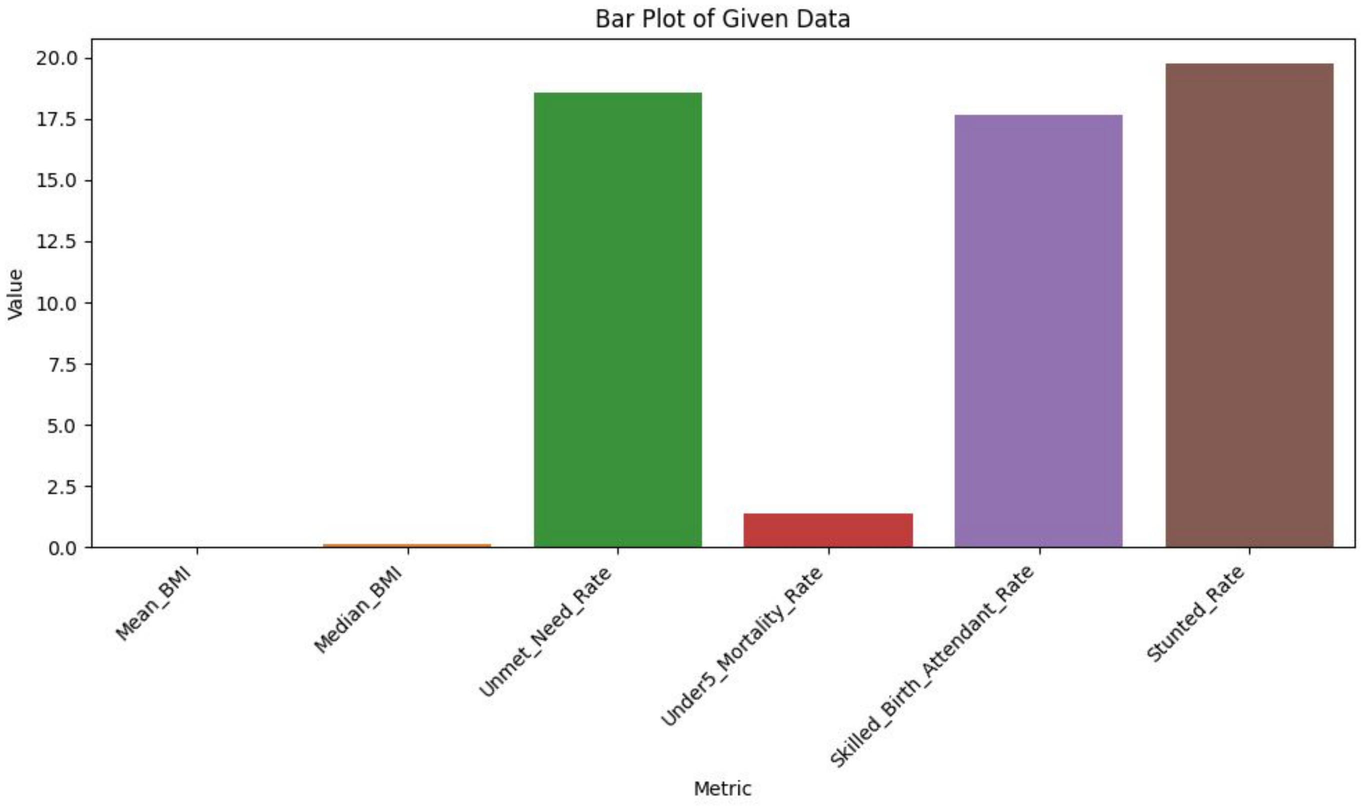
Model performance on less data column

- Even on the less performant columns, our ensemble model clearly performs better as compared to any of the standard model
- Availability of less data point for the following prediction labels significantly effect the generalization ability of the model on the various scenarios
  - Skilled_Birth_Attendant_Rate
  - Stunted_Rate
- For Unmet_Need_Rate around 68% of the data is either 0 or 100 and we have very little data which actually has a good distribution

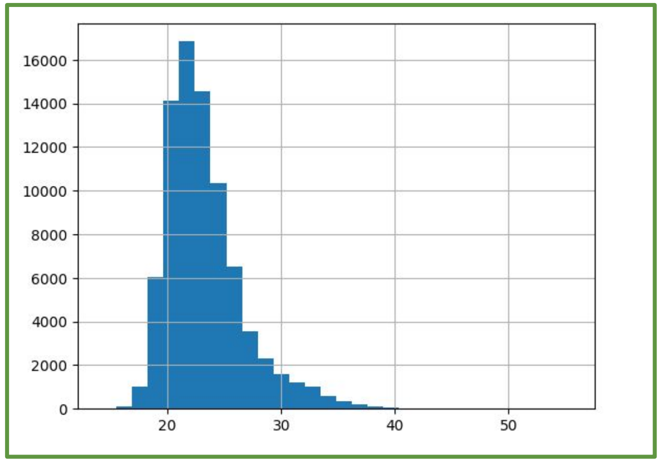


Model performance on all data column

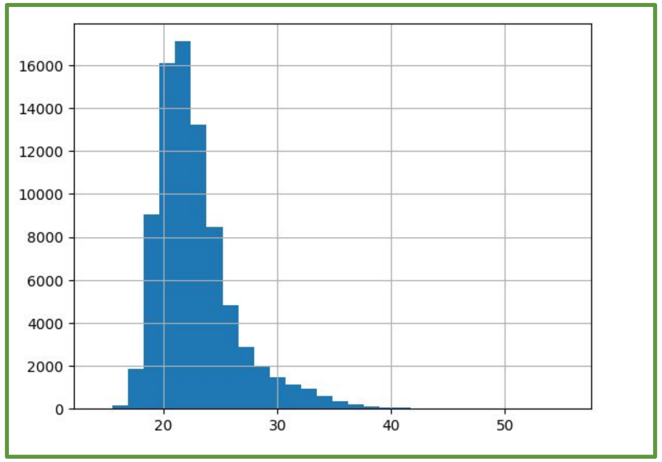
- From the model’s training performance on the validation set it is quite clear that our model performs quite well in case of the data labels where higher data points are available
- However it struggles a but in case of the labels with lesser data points
- The average of the training performance is near 9.89 which is close to the score achieved in the leaderboard, thus the test and the validation set **MCRMSE** score are quite close



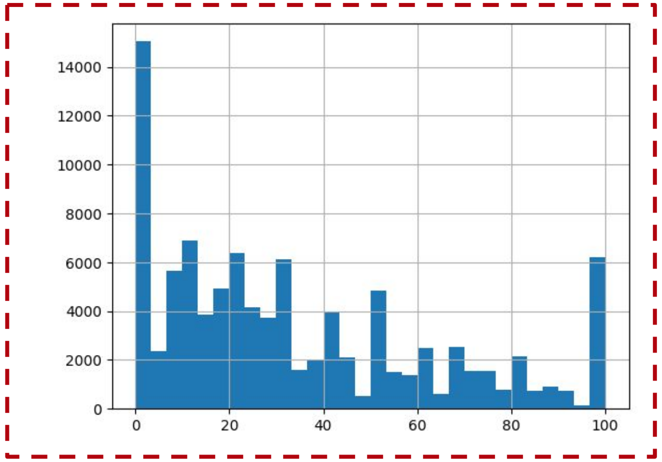
Observation (Model Performance and Data)



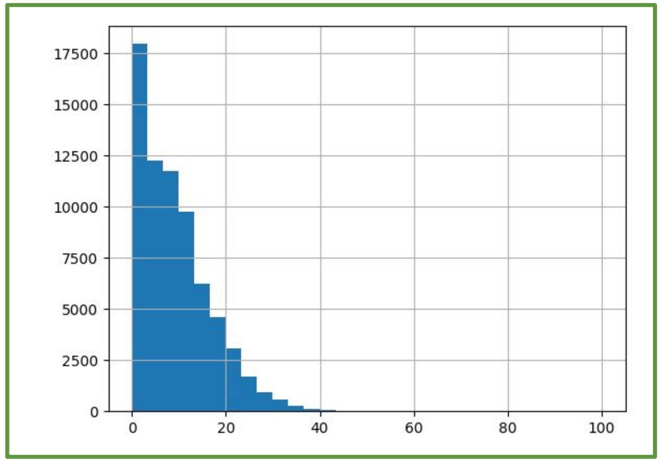
Mean_BMI



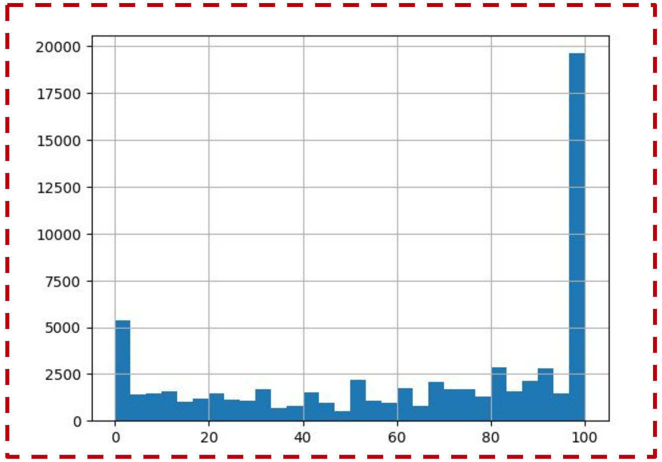
Median_BMI



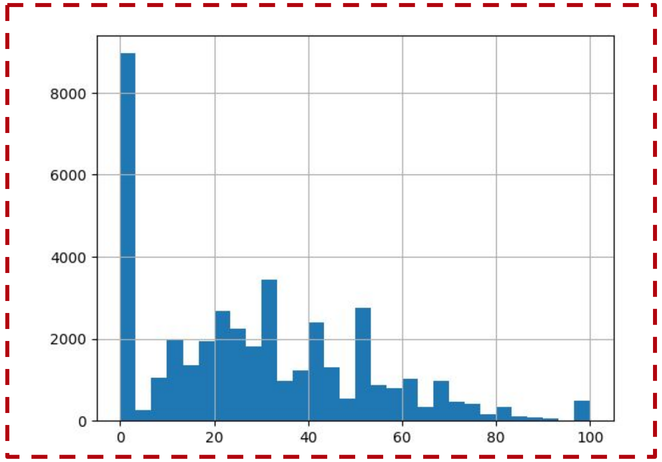
Unmet_Need_Rate



Under5_Mortality_Rate

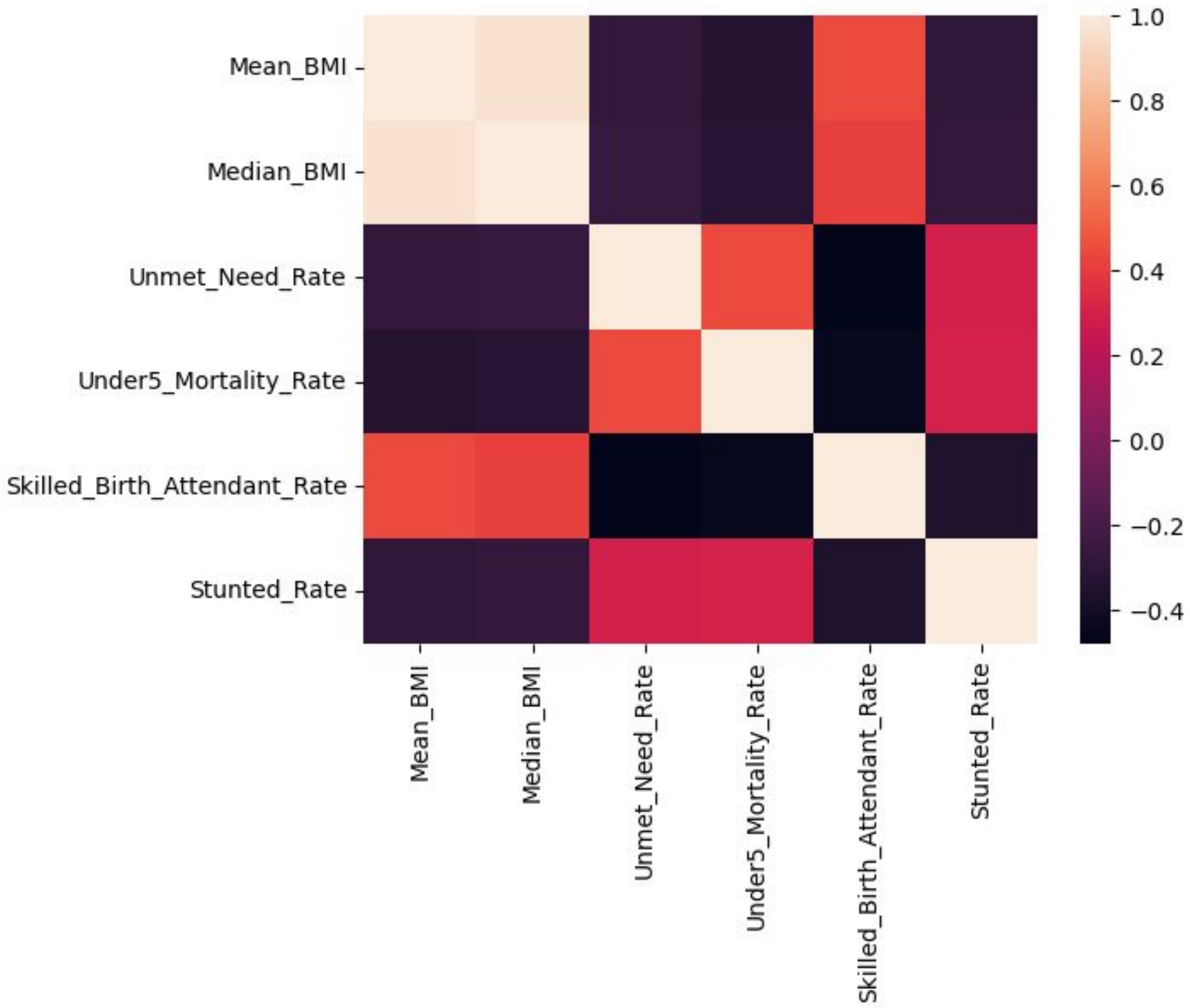


Skilled_Birth_Attendant_Rate



Stunted_Rate

Results and Discussions



Correlation Matrix

- We observe that the Mean BMI and the Median BMI values to be predicted are highly **correlated**
- Hence we train a model on the difference between both the values and set the values of the median BMI as follows
    - *Median BMI = Mean BMI + {Model Prediction}*
- We saw an improvement in the score from **10.761** to **10.759**, which is not much significant but it signifies that the high correlation between the mean and median BMI can be harnessed to improve predictions

[Future Scope] Satellite Data (NASA Earth Explorer)

Sparsely Populated Region

Densely Populated Region