National Rural Health Mission

Guide: Gopinath Panda

Dhirubhai Ambani Institute of Information and Communication Technology

July 3, 2024

Contents

Introduction

Introduction

- NRHM stands for the National Rural Health Mission, which was a program launched by the Government of India in 2005.
- It aimed to provide accessible, affordable, and quality healthcare to rural populations, especially focusing on maternal and child health.
- The NRHM aimed to strengthen healthcare infrastructure, improve human resource capacity, and enhance the delivery of healthcare services in rural areas.

Figure: National Rural Health Mission

Tools and Technologies

Tools and Technologies

We used the following tools and technologies to analyse the datasets of NHRM in an efficient manner. These are efficient and easy-to-use librariues in Python.

1. Numpy
2. Pandas
3. Matplotlib
4. Seaborn
5. Scikit-learn

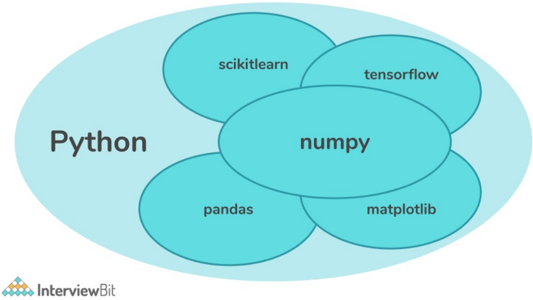


Figure: Python libraries

Methodology

Methodology

We followed similar strategy to analyse both the given datasets related to NRHM, which included:

1. Data Collection and Loading the Data
2. Data description and Data Cleaning
3. Data Visualization
4. Data Preprocessing
5. Feature Engineering and Feature Selection
6. Predictions based on Model fitting

Dataset-1 : NHM Budget from 2015-16 to 2022-23

Dataset-1 : NHM Budget from 2015-16 to 2022-23

- The Budget Allocation Dataset comprises of various aspects such as approved budgets, proposed allocations, expenditure patterns, and fund utilization under the NRHM.
- It offers a comprehensive overview of the financial resources allocated to different healthcare programs and initiatives aimed at improving rural healthcare services.
- First 5 rows of the dataset is:

```
# Displaying top 5 rows
df1.head()
```

	State_UT	State_UT_Code	Fiscal Year	Budget Proposed by the States/UTs	Budget Approved for the States/UTs	Opening Balance with the States/UTs	Release of Government of India's Fund	Total Expenditure Reported (Including States' Share)	Extent of Budget Approved Against Budget Proposed	Extent of Funds Utilised Against Budget Approved	...	Unnamed: 14	Unnamed: 15	Unnamed: 16	Unnamed: 17	Unr
0	Andhra Pradesh	1.0	2015-16	2380.02	1336.13	NaN	659.04	1105.70	56.14	82.75	...	NaN	NaN	NaN	NaN	
1	Arunachal Pradesh	2.0	2015-16	346.46	195.94	NaN	163.80	147.41	56.55	75.23	...	NaN	NaN	NaN	NaN	
2	Assam	3.0	2015-16	2275.32	1853.40	NaN	997.59	1212.25	81.46	65.41	...	NaN	NaN	NaN	NaN	
3	Bihar	4.0	2015-16	3874.98	2672.45	NaN	1269.67	1731.85	68.97	64.80	...	NaN	NaN	NaN	NaN	
4	Chhattisgarh	5.0	2015-16	1528.15	1215.09	NaN	423.31	769.33	79.51	63.31	...	NaN	NaN	NaN	NaN	

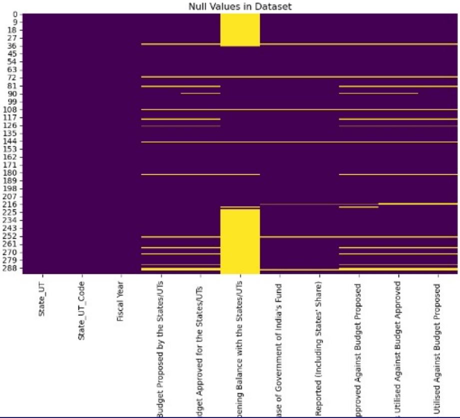
5 rows × 24 columns

Figure: First 5 rows of the Budget Allocation Data

Data Cleaning

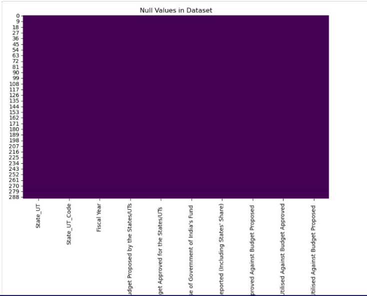
Data Cleaning

- Firstly, we checked for null value using the function in the pandas library, "df.isnull().sum()", which display number of NULL/NaN values in each column.
- Dropping the rows which had NULL values in all the columns.
- Visualizing NULL values with the help of heatmap.



Data Cleaning

- We can clearly see that 40% values are missing in the Column 'Opening Balance', and since, our dataset is relatively small, it's better to drop the column.
- For other missing values, for columns such as Budget Proposed and Budget Approved, Release of Govt Expenditure, Total Expenditure Reported, we impute NULL values by the Mean for the corresponding state over the fiscal years.
- After performing these operations, we got our corresponding cleaned dataset. It can be visualized using the following heatmap.



Visualization

Budget Approved for States/UTs over the fiscal years

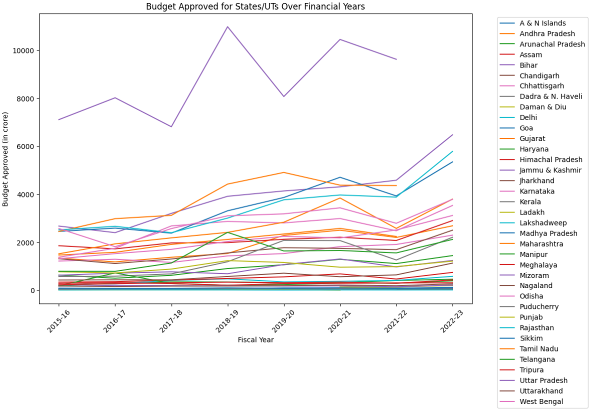
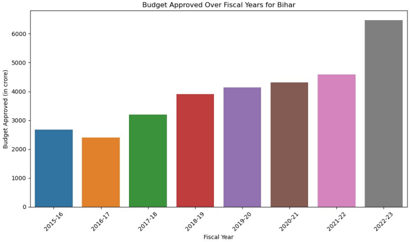
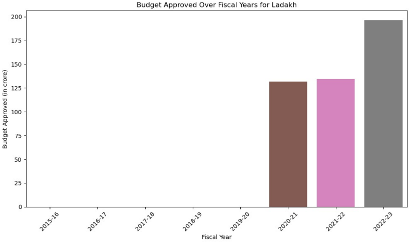


Figure: Line Chart

- Uttar Pradesh can consistently received higher budgets as compared to other States/UTs.

Budget Approved for particular States/UTs over the fiscal years





- We can clearly observe that government is increasing budget consistently over the years.

Total Budget Approved per Fiscal Years

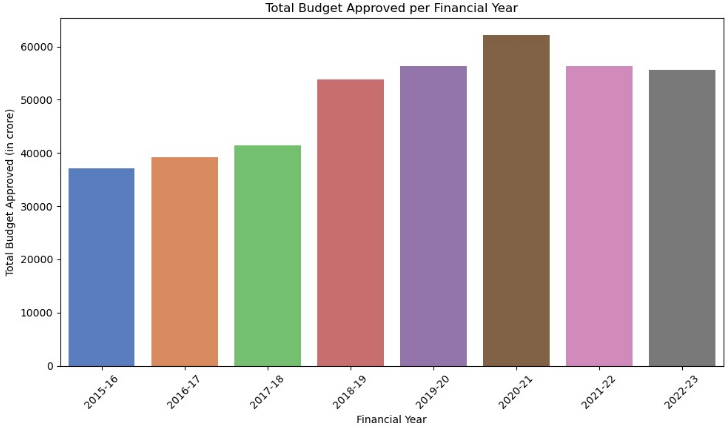
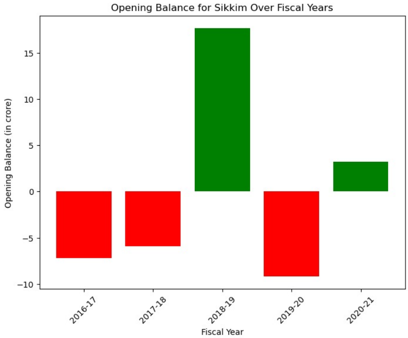
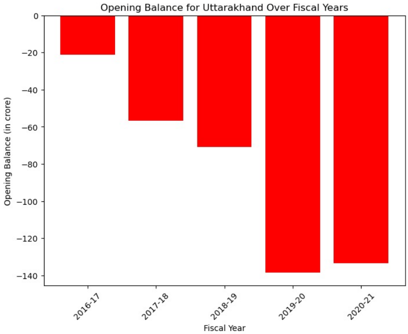


Figure:

- The spike in the year 2020–21 may be due to COVID-19.

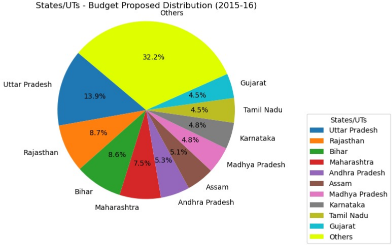
Opening Balance for States/UTs over fiscal years

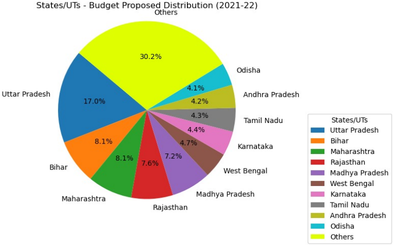




- The red bar indicates the opening balance is negative, while the green bar indicates a positive opening balance.

Budget Proposed Distribution over the years





- As we can clearly observe from the pie chart, Uttar Pradesh hold majority of the budget proportion consistently over the years.

Comparison of Proposed vs Approved Budget for particular states over the years

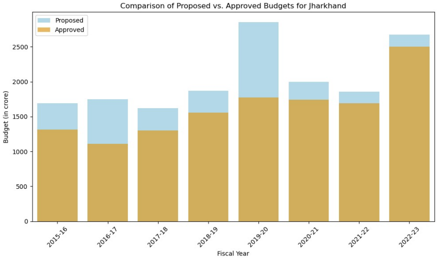


Figure: Stacked Bar Chart

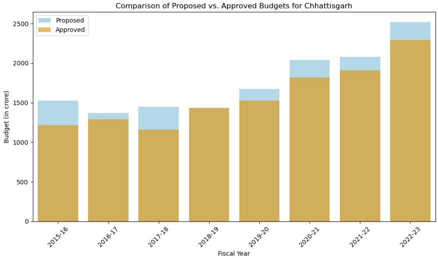


Figure: Stacked Bar Chart

Variation of Parameters over the Fiscal Years

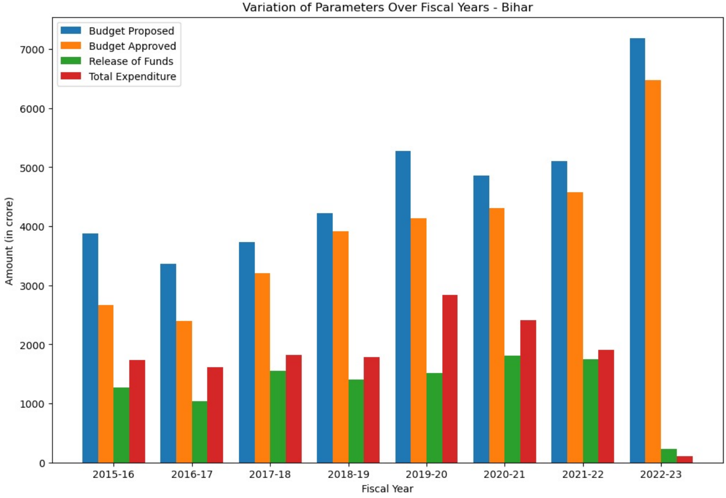


Figure: Side Bar Chart

Data Preprocessing

Encoding Categorical Variables

- There are two types of Encoding schemes, Label Encoding and One-Hot Encoding. We used one-hot encoding, as our data is not ordinal but categorical.
- One-hot encoding is a technique used in machine learning and data processing to represent categorical data numerically.
- Machine learning algorithms require numerical input data. When dealing with categorical features (such as the names of states /UTs), one hot encoding can be applied to convert these categorical variables into numerical format.

Encoding Categorical Variables

- The following changes have been reflected after performing One-Hot Encoding.

Extent of Budget Approved Against Budget Proposed	Extent of Funds Utilised Against Budget Approved	Extent of Funds Utilised Against Budget Proposed	State_UT_Andhra Pradesh	State_UT_Arunachal Pradesh	...	State_UT_Uttar Pradesh	State_UT_Uttarakhand	State_UT_West Bengal	Fiscal Year_2016-17	Fiscal Year_2017-18
56.14	82.75	46.46	1	0	...	0	0	0	0	0
56.55	75.23	42.55	0	1	...	0	0	0	0	0
81.46	65.41	53.28	0	0	...	0	0	0	0	0
68.97	64.80	44.69	0	0	...	0	0	0	0	0
79.51	63.31	50.34	0	0	...	0	0	0	0	0

Feature Scaling

- Normalization and standardization are both techniques used in data preprocessing to scale numerical features.
- Many machine learning algorithms assume that the features follow a Gaussian distribution. Standardization helps to achieve this assumption by centering the distribution at 0 and scaling it to have a standard deviation of 1.

Budget Proposed by the States/UTs	Budget Approved for the States/UTs	Release of Government of India's Fund	Total Expenditure Reported (Including States' Share)
0.382582	-0.055853	0.054749	0.150831
-0.636169	-0.690335	-0.604436	-0.614089
0.330131	0.231992	0.505373	0.235881
1.131511	0.687769	0.867522	0.650632
-0.044178	-0.123208	-0.259018	-0.117664

ns

Feature Engineering

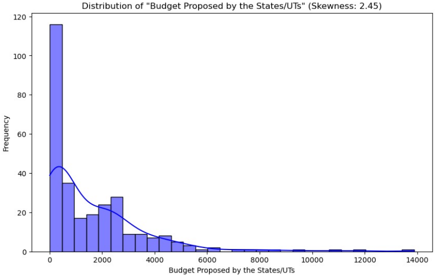
Creating new features

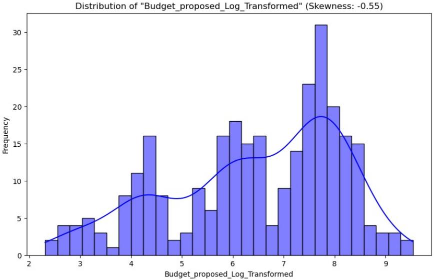
- We divided the values of 'Budget Proposed' into three categories based on their values.

```
# Convert numerical feature into categorical bins
# Dividing the budget proposed by the states into low , medium, and high
df1['Budget_Category'] = pd.cut(df1['Budget Proposed by the States/UTs'], bins=3, labels=['Low', 'Medium', 'High'])
```

Feature Transformation

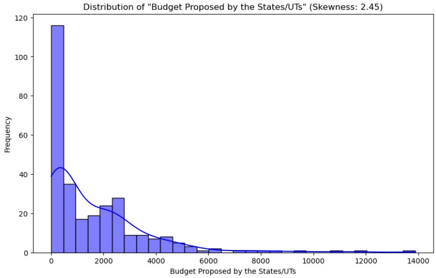
- We applied log-transformation, due to which skewness decreased drastically from 2.45 to -0.55.

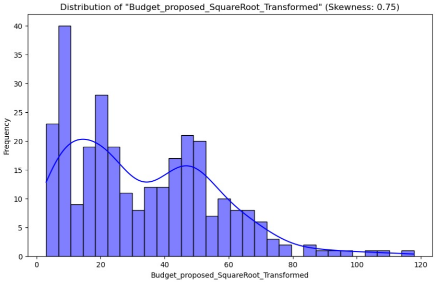




Feature Transformation

- We applied square root transformation, due to which skewness decreased from 2.45 to 0.75.





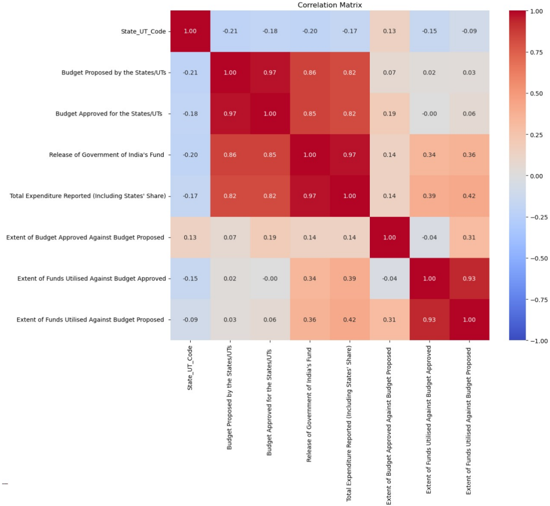
- We can state that, according to the values of the column, Log transformation seems to be better than Square Root transformation, as -0.55 is nearer to 0.

Feature Selection

Feature Selection using Correlation Matrix

- Feature selection is the process of choosing a subset of relevant features from the original set of features to improve model performance, reduce overfitting, and enhance interpretability.
- We use correlation matrix for feature selection, which is a common technique to identify highly correlated features and remove redundant ones.

Feature Selection using Correlation Matrix



Feature Selection

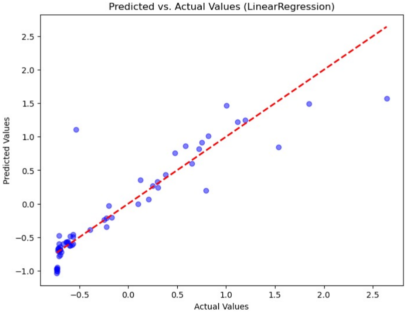
- We set the target_variable as 'Total Expenditure Reported.'

```
In [61]: # Identify features with highest absolute correlation with the target variable
         target_correlation = correlation_matrix["Total Expenditure Reported (Including States' Share)"].abs().sort_values(ascending=False
         important_features = target_correlation[target_correlation >= 0.5].index.tolist()
         print("Important Features based on Correlation:")
         print(important_features)

         Important Features based on Correlation:
         ["Total Expenditure Reported (Including States' Share)", "Release of Government of India's Fund ", 'Budget Proposed by the Stat
         es/UTs', 'Budget Approved for the States/UTs  ']
```

- Important Features based on Correlation: ["Total Expenditure Reported (Including States' Share)", "Release of Government of India's Fund ", 'Budget Proposed by the States/UTs', 'Budget Approved for the States/UTs ']

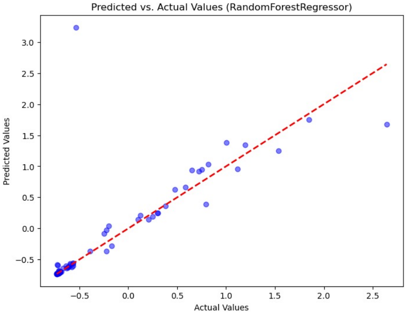
Model Fitting

Model Fitting using Linear Regression



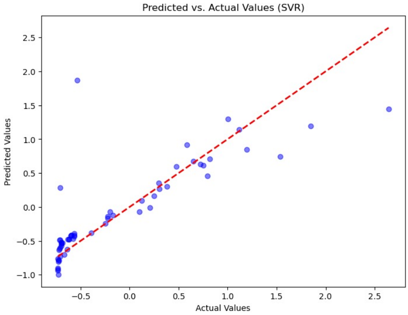
- Based on the graph visualized above, R-squared value is 0.8255 and MSE is 0.1021.
- We can infer from the values that the linear regression model explains approximately 82.55% of the variance in the target variable.
- The mean squared error, which measures the average squared difference between the actual and predicted values, is relatively low at 0.1021.

Model Fitting using Random Forest Regressor



- Based on the graph visualized above, R-squared value is 0.5347 and MSE is 0.2723.
- The random forest regressor model explains approximately 53.47% of the variance in the target variable, which is lower compared to the linear regression model.
- The mean squared error is higher at 0.2723, indicating a higher level of prediction error compared to linear regression.

Model Fitting using Support Vector Regressor



- Based on the graph visualized above, R-squared value is 0.6946 and MSE is 0.1787.
- The SVR model explains approximately 69.46% of the variance in the target variable, which falls between the R-squared values of linear regression and random forest regressor.
- The mean squared error is moderate at 0.1787.

Observations based on Model Fitting

- Linear regression performs the best in terms of R-squared and MSE, indicating a good fit to the data.
- Support Vector Regressor performs moderately well, with an R-squared value between linear regression and random forest regressor.
- Random forest Regressor performs the worst among the three models, with the lowest R-squared value and highest MSE, suggesting a weaker fit to the data compared to the other models.

Dataset-2 : Status of Healthcare Infrastructure in Rural Areas

Dataset-2 : Status of Healthcare Infrastructure in Rural Areas

- The infrastructure facilities dataset provides detailed information on the availability and quality of healthcare infrastructure across different rural healthcare facilities.
- It includes data on the presence of essential amenities such as thenumber of PHCs functioning, PHCs functioning on 24x7 basis, PHCs with labour room and many more.
- First 5 rows of the dataset is:

	S.No.	State/UT	Number of PHCs Functioning	PHCs functoning on 24X7 basis	With Labour Room	With OT	With at least 4 beds	Without Electric Supply	Without Regular Water Supply	With Telephone	Fiscal Year
0	1	Andhra Pradesh	1142	1142	1134	1129	1142	0	0	0	2021-22
1	2	Arunachal Pradesh	126	42	67	1	46	11	18	18	2021-22
2	3	Assam	920	280	721	64	339	17	0	0	2021-22
3	4	Bihar	1492	630	465	254	612	326	374	374	2021-22
4	5	Chhattisgarh	770	477	752	63	610	17	8	8	2021-22

Figure: First 5 rows of the Infrastructure Facilities Data

Data Cleaning

Data Cleaning

- From the below heatmap visualization of NULL values, we can see that for some states, there is data missing for some particular year, and it has been filled with 0.
- To address the issue of missing or zero values in specific columns for certain states and fiscal years, we replace these zeros with NaN (NULL) values.

Data Cleaning

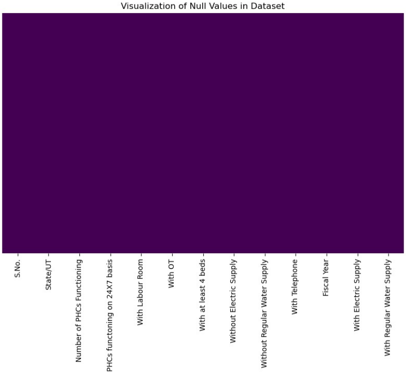


Figure: Before

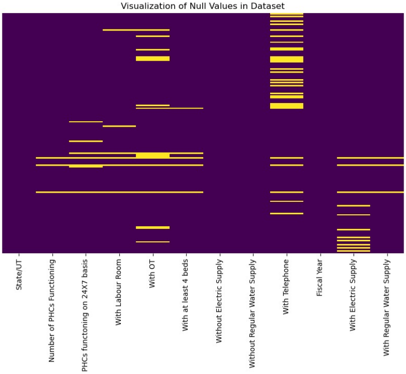
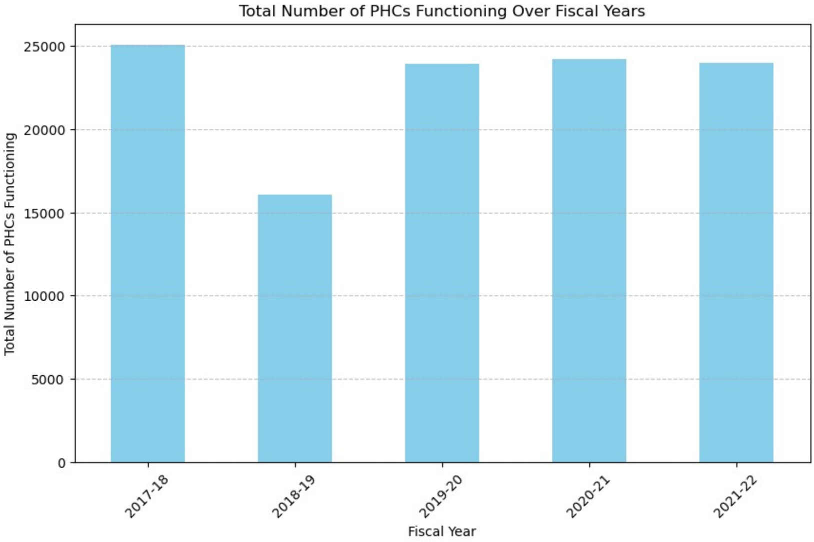


Figure: After

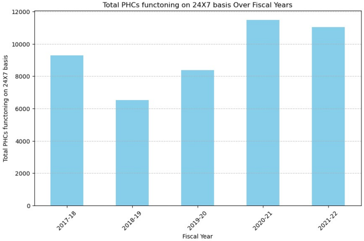
- We now fill these missing values with the mean of the corresponding data for the same state over fiscal years. This process helps to impute missing values based on the average behavior of the data within each state.

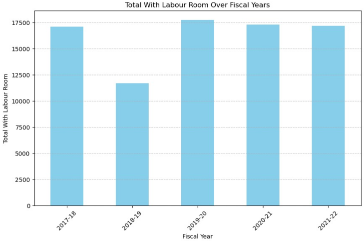
Visualization

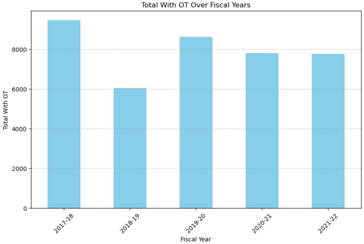
Total number of PHCs functioning over each fiscal year

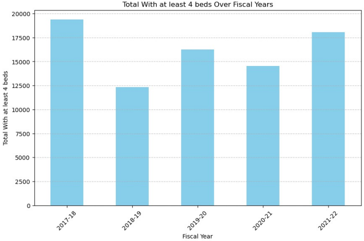


Variation in Infrastructure facilites over each fiscal year

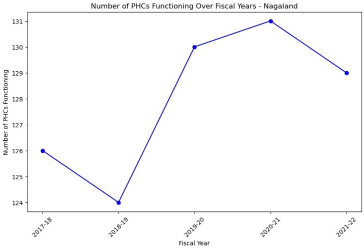


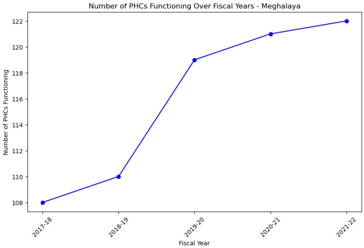






Number of PHCs functioning over Fiscal Years for each State/UTs





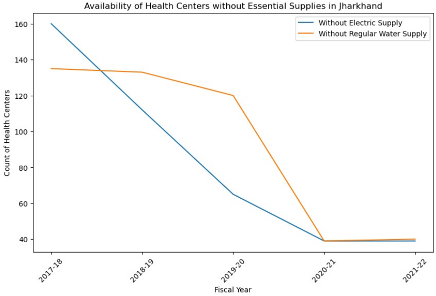
- The number of PHCs is usually increasing over the years.

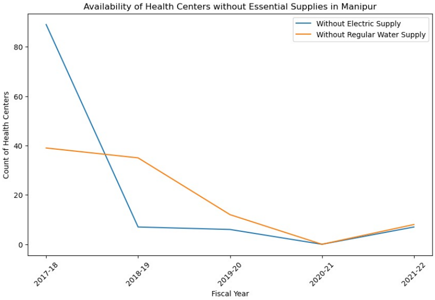
States with less than 50% of PHCs with labour room for each fiscal year

```
For Fiscal Year 2017-18:
- State/UT: Bihar, Percentage with Labour Room: 41.86%
- State/UT: Delhi, Percentage with Labour Room: 20.00%
- State/UT: Himachal Pradesh, Percentage with Labour Room: 28.47%
- State/UT: Kerala, Percentage with Labour Room: 7.30%
- State/UT: Odisha, Percentage with Labour Room: 47.90%

For Fiscal Year 2018-19:
- State/UT: Delhi, Percentage with Labour Room: 20.00%
- State/UT: Goa, Percentage with Labour Room: 40.94%
- State/UT: Himachal Pradesh, Percentage with Labour Room: 19.08%
- State/UT: Kerala, Percentage with Labour Room: 9.14%
- State/UT: Maharashtra, Percentage with Labour Room: 49.96%

For Fiscal Year 2019-20:
- State/UT: Bihar, Percentage with Labour Room: 32.20%
- State/UT: Delhi, Percentage with Labour Room: 20.00%
- State/UT: Goa, Percentage with Labour Room: 23.64%
- State/UT: Himachal Pradesh, Percentage with Labour Room: 34.40%
- State/UT: Kerala, Percentage with Labour Room: 5.48%
- State/UT: Odisha, Percentage with Labour Room: 49.84%
- State/UT: West Bengal, Percentage with Labour Room: 47.10%

For Fiscal Year 2020-21:
- State/UT: Bihar, Percentage with Labour Room: 29.30%
- State/UT: Delhi, Percentage with Labour Room: 20.00%
- State/UT: Himachal Pradesh, Percentage with Labour Room: 31.83%
- State/UT: Kerala, Percentage with Labour Room: 0.64%
- State/UT: Odisha, Percentage with Labour Room: 45.34%
- State/UT: Punjab, Percentage with Labour Room: 49.29%
- State/UT: West Bengal, Percentage with Labour Room: 36.39%

For Fiscal Year 2021-22:
- State/UT: Bihar, Percentage with Labour Room: 31.17%
- State/UT: Delhi, Percentage with Labour Room: 20.00%
- State/UT: Himachal Pradesh, Percentage with Labour Room: 29.66%
- State/UT: Kerala, Percentage with Labour Room: 5.51%
- State/UT: Odisha, Percentage with Labour Room: 46.20%
- State/UT: Punjab, Percentage with Labour Room: 49.53%
- State/UT: Uttarakhand, Percentage with Labour Room: 33.33%
- State/UT: West Bengal, Percentage with Labour Room: 33.88%
```

- From this insight, the government can improve Labour Room facilities by focusing on particular states. According to this data, Bihar is a state with most of the PHCs without having Labour Room.

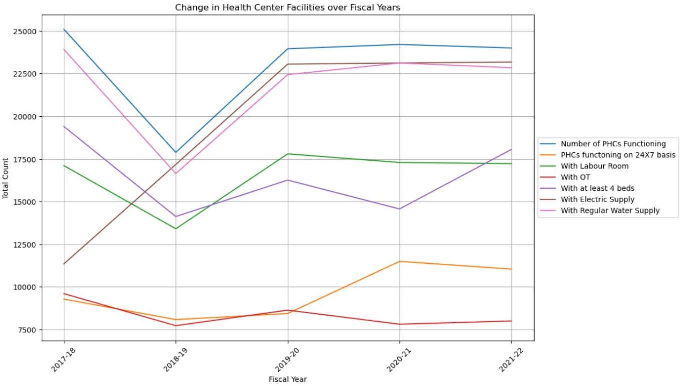
Health Centers without both electric supply and regular water supply





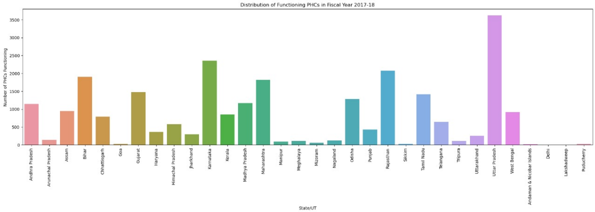
- From this, we can see that PHCs without having Electric Supply and Water Supply are decreasing year-by-year.

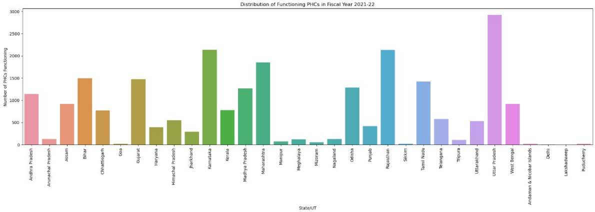
Variation in Health Centre facilities over fiscal years



- We can conclude that overall healthcare facilities in PHCs are improving.

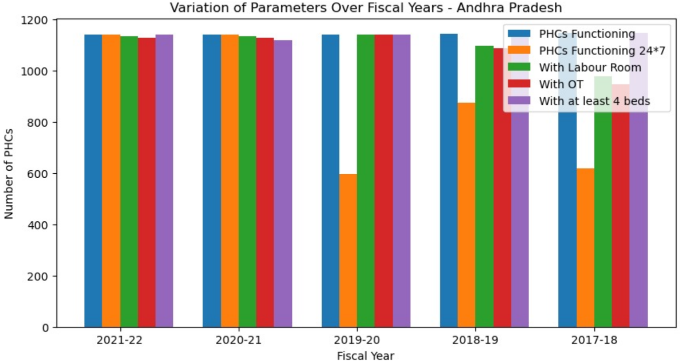
Distribution of PHCs over fiscal years





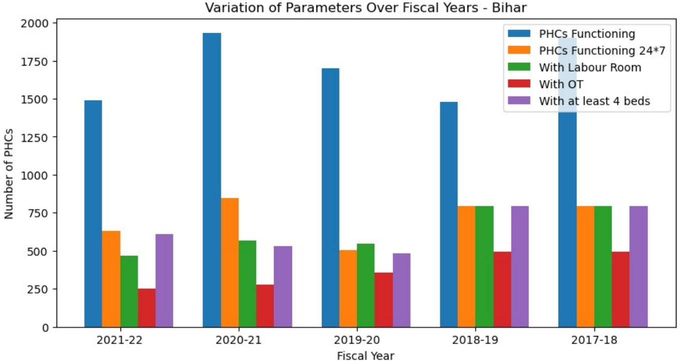
- We can observe from the above visualization that maximum number of PHCs are established in Uttar Pradesh, followed by Karnataka.

Variation of Parameters in particular States over Fiscal Years



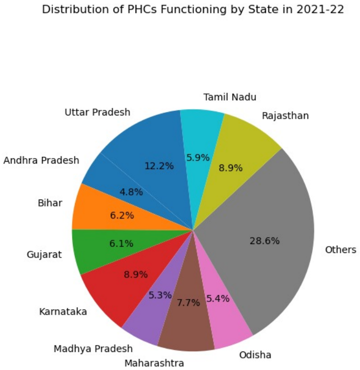
- From this, it can be inferred that most of the PHCs in Andhra Pradesh are working with decent infrastructure facilities.

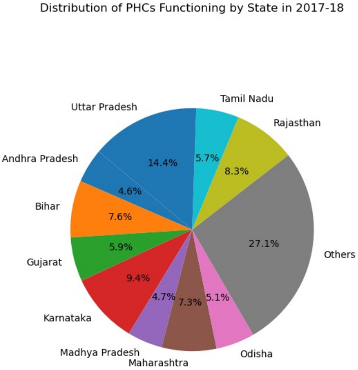
Variation of Parameters in particular States over Fiscal Years



- From this, it can be inferred that most of the PHCs in Bihar are working with poor infrastructure facilities.

Distribution of PHCs functioning over the fiscal years





- States with majority number of PHCs are Uttar Pradesh, Karnataka, Rajasthan, and Maharashtra.

Data Preprocessing

Encoding Categorial Variables and Feature Scaling

- Similar to the previous dataset, we performed one-hot encoding on the categorical columns and standardization on the numerical columns to get the following output.

	Number of PHCs Functioning	PHCs functioning on 24X7 basis	With Labour Room	With OT	With at least 4 beds	Without Electric Supply	Without Regular Water Supply	With Telephone	With Electric Supply	With Regular Water Supply	...	State/UT_Tamil Nadu	State/UT_Telangana	State
0	0.557370	2.299522	1.030414	2.151284	0.928342	-0.329785	-0.500445	1.436125	0.773568	0.640943	...	0	0	
1	-0.783160	-0.711813	-0.753071	-0.644421	-0.694228	-0.304451	-0.267115	-0.614661	-0.724313	-0.796383	...	0	0	
2	0.264459	-0.060269	0.340086	-0.488278	-0.260457	-0.290633	-0.500445	-0.180886	0.424986	0.332349	...	0	0	
3	1.019167	0.897883	-0.087816	-0.017370	0.143705	0.421015	4.347644	0.217813	0.808572	0.607582	...	0	0	
4	0.066546	0.479033	0.391903	-0.490756	0.140744	-0.290633	-0.396743	-0.638045	0.206211	0.112719	...	0	0	

5 rows × 46 columns

Feature Engineering

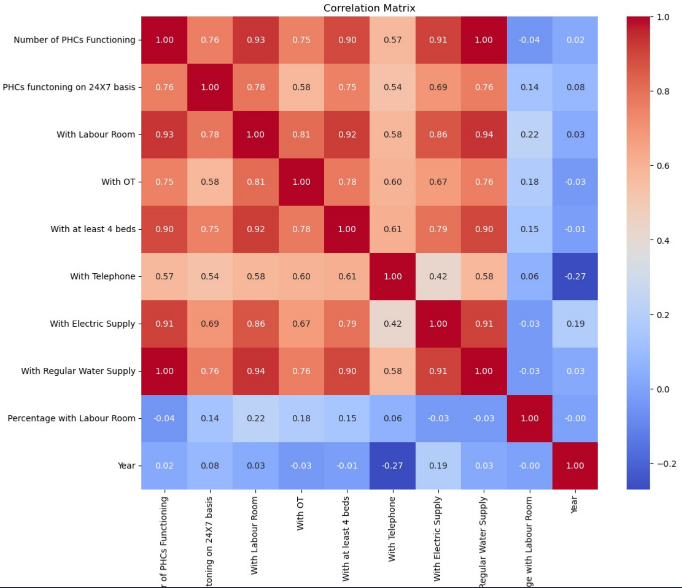
Creating new features

- Now, we group all the States/UTs into regions based on geographical location, like North, South, West, North East.

	State/UT	Number of PHCs Functioning	PHCs functoning on 24X7 basis	With Labour Room	With OT	With at least 4 beds	Without Electric Supply	Without Regular Water Supply	With Telephone	Fiscal Year	With Electric Supply	With Regular Water Supply	Percentage with Labour Room	Region	Year
0	Andhra Pradesh	1142.0	1142.0	1134.0	1129.0	1142.0	0	0	895.0	2022-01-01	1142.0	1142.0	99.299475	South India	2022
1	Arunachal Pradesh	126.0	42.0	67.0	1.0	46.0	11	18	18.0	2022-01-01	115.0	108.0	53.174603	Northeast India	2022
2	Assam	920.0	280.0	721.0	64.0	339.0	17	0	203.5	2022-01-01	903.0	920.0	78.369565	Northeast India	2022
3	Bihar	1492.0	630.0	465.0	254.0	612.0	326	374	374.0	2022-01-01	1166.0	1118.0	31.166220	East India	2022
4	Chhattisgarh	770.0	477.0	752.0	63.0	610.0	17	8	8.0	2022-01-01	753.0	762.0	97.662338	Central India	2022

Feature Selection

Feature Selection using Correlation Matrix

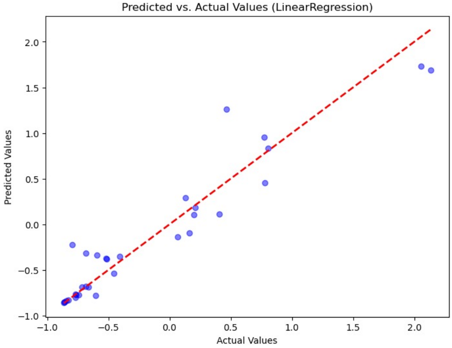


Feature Selection using Correlation Matrix

- Wet set the target_variable as 'With Labour Room'.
- Important features based on correlation analysis: ['With Labour Room', 'With Regular Water Supply', 'Number of PHCs Functioning', 'With at least 4 beds', 'With Electric Supply', 'With OT', 'PHCs Functioning on 24X7 basis', 'With Telephone']

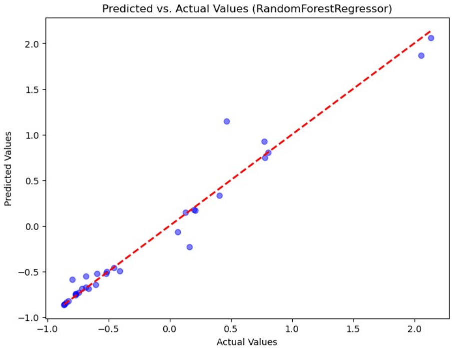
Model Fitting

Model Fitting using Linear Regression



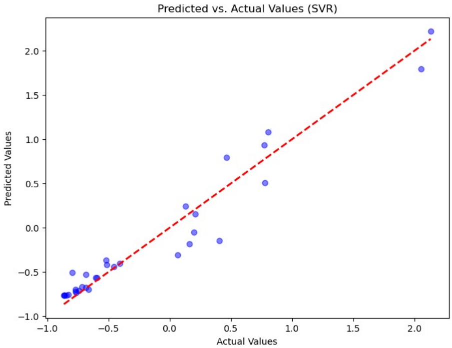
- Based on the graph visualized above, R-squared value is 0.9049 and MSE is 0.0599.
- The linear regression model explains approximately 90.49% of the variance in the target variable.
- The mean squared error is relatively low at 0.0599, indicating a small average squared difference between the actual and predicted values.

Model Fitting using Random Forest Regressor



- Based on the graph visualized above, R-squared value is 0.9613 and MSE is 0.0244.
- The random forest regressor model explains approximately 96.13% of the variance in the target variable, which is higher compared to linear regression.
- The mean squared error is also lower at 0.0244, indicating a smaller prediction error compared to linear regression.

Model Fitting using Support Vector Regressor



- Based on the graph visualized above, the R-squared value is 0.9406 and the MSE is 0.0374.
- The SVR model explains approximately 94.06% of the variance in the target variable.
- The mean squared error is moderate at 0.0374, falling between linear regression and random forest regressor.

Observations based on Model fitting

- Random Forest Regressor performs the best among the three models, with the highest R-squared value and lowest MSE, suggesting the best fit to the data and the smallest prediction error.
- Linear regression and Support Vector Regressor perform well, with high R-squared values and relatively low MSE values.
- Linear regression explains slightly less variance compared to Support Vector Regressor but has a smaller MSE.

Conclusion

Conclusion

- It's evident that the National Rural Health Mission (NRHM) has made significant efforts to improve healthcare infrastructure and facilities across India.
- The allocation of substantial funds and the focus on enhancing the functionality of PHCs with improved infrastructure have contributed positively to healthcare accessibility.
- However, the persistence of disparities in the distribution of healthcare facilities, with Uttar Pradesh consistently ranking higher while northeastern states lag behind, highlights the ongoing challenges in achieving equitable healthcare access nationwide.
- Despite the successes, there's recognition that there's still room for improvement.

Contribution

- We are Group-2, who presented the data analysis of National Rural Health Mission. We are a team of three: **Shreyans Jain** (202103009), **Yash Garg** (202101006) and **Kalp Shah** (202103003).
- Equal Contribution
- Everyone on our team helped make the project successful.
- We worked together and shared the work, which helped us finish our goals quickly and well.

Thank you!