Data-driven Governance: Unravelling and improving impact of Indian Administrators on policy execution using Machine Learning

Netaji Subhas University of Technology, Dwarka, Delhi, 110078, India

Department of Computer Science

Abstract

The core model of governance from earliest of administrations is based on the decision making based on statistics, evaluations of effects and the reiteration to improve the existing models[1][2][3]. It is imperative that this model of evaluation and action is one of the best models that can help us guarantee improvements or insights indefinitely. But in this paper, we touch on the potential benefits of using machine learning to get insights from our data. To do so, we have first come up with our own unique case study of Indian administrators. We have collected a unique dataset of key factors and indicators which we collected from multiple sources and then further used this dataset to draw key insights as well as catering predictions for future work. The purpose of this research is also based on promoting data dominant decision making so that Indian administrators can carry out their duties without political intervention or executives turmoil.

Keywords: Data-driven governance, District Magistrates, Indian Administrators

1. Introduction

With the announcement of open government data in June 2011 and further launch in October 2012 in compliance with NDSAP (National Data Sharing and Accessibility Policy) of India [1]. The scheme highlighted the use of data and it's value to the general public as well as the government institutions, which further encouraged them to release data and open the information available to use them both internally as well as externally for the decision making and scrutiny of the decisions made. The access to information to the public domain is an indicator of a healthy democracy where one can individually scrutinize the data and the claims made based on that data to refer them as true or not. In this paper, we will focus on the use of the data internally i.e. we take the data into account within an organization, for us that is Indian administration services or specifically district magistrates. We will try to predict an optimal service period within a district before a district magistrate or DM is transferred to a new location.

Email addresses: tanmay.nagori.ug21@nsut.ac.in (), aatish.malik.ug21@nsut.ac.in (), deepanshu-ug21@nsut.ac.in (), abhinav.tomar@nsut.ac.in (), gaurav.singal@nsut.ac.in (), vijay.bohat@nsut.ac.in ()

Let's look at why it is important to find an optimal tenure for a DM in a district. When a DM is transferred from one place to another it is imminent that they will need some time to adjust to their new district, let alone understanding it's culture, people and it's own unique challenges. Since, policy creation is almost same for the whole country by the legislation, it is the job of the administrators in policy execution of said policies. Hence, they can be directly credited for a good impact of a policy or an under performance of another. Such as, during the period of Covid pandemic with the same policy being enacted for whole country some districts were way better performer in vaccine drives while other were lacking. Although, multiple factors and reasons come into play, the main power over that jurisdiction for the enforcement of these policies lies within the hand of the sitting DMs in the office.

Now, it is important and pivotal that effective governance of districts ensures the well-being and development of communities. At the heart of district administration is DMs, they are entrusted with responsibility of managing facets of governance, law enforcement and public service delivery.

The decision to transfer a DM from one to another district is delicate balance. It is a balance between administrative continuity which helps a DM to work on his/her existing vision but there is also a need for new and fresh leadership. While frequent rotations can disrupt ongoing initiatives and hinder an important factor called institutional memory [2], on the other hand prolonged tenures are seen to be a reason for complacency and stagnation. With the amount of data we have to our disposal and by harnessing the power of data analytics and machine learning's predictive modeling. We aim to bridge this gap in this paper and further provide valuable insights into district administration for the policy makers, this facilitation of evidence-based decision-making and resource allocation will just further benefit the society.

The methodologies this paper seeks to leverage are data-driven used to address the long-standing challenge of optimizing DM tenures. By analyzing a rich dataset created by us. This dataset encompasses various socio-economic as well as demographic parameters. We aim to deliver the results using this technique to gain hidden insights even through a nuanced lens of understanding in the interplay between the administrations and it's district level local dynamics, we further endeavor to inform policy interventions aimed at enhancing governance effectiveness which in turn will foster sustainable government. In the further sections, we will delve into the methodologies employed for data collection, feature engineering and model development. We then present our findings and understandings through the lens of key highlights of predictive models. Finally, we discuss the implications of our research for policy formulation and administrative decision making, underscoring the informative potential of data driven governance[4][5].

2. Methodology

2.1. Research Type:

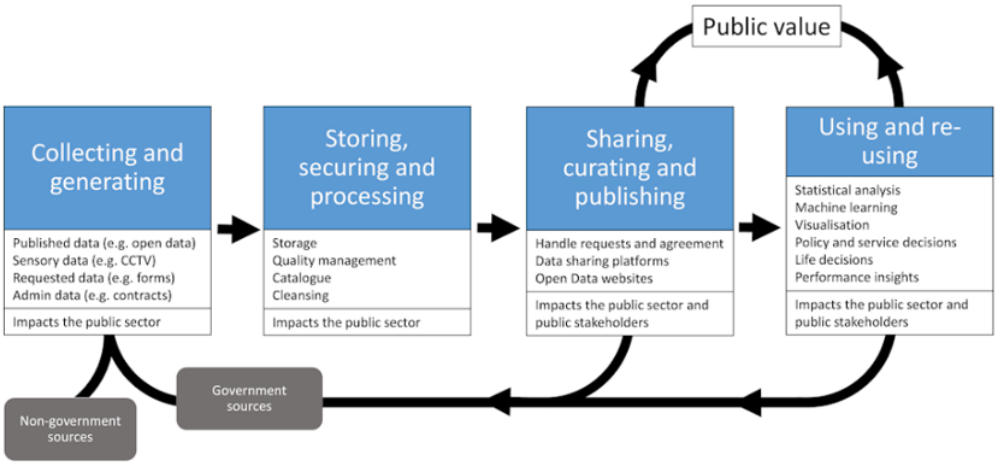

Figure 1: Government data value cycle.

The idea of our research stems from the government data value cycle, A government data value cycle involves the collection of both public and private sector data. It is the storing, pre-processing, analyzing, publishing and drawing out the useful insights out of it by applying machine learning and data visualization techniques. As shown in Figure-1 (A government data value cycle). We go through several phases of the government data value cycle. The phases we mitigated one by one are as follows:

1. The phase of collecting and generating is a phase which involves the collection of data from various sources such as published data like open data repositories in our case it is the open government data portal. Further requested data is another source where we used the data from different sources such as Niti Aayog and forms or surveys. Administrative data is also part of this data cycle. It is important to note that the data generation process is a crucial step to impact the public sector. An accurate representation is our stepping stone goal in this phase. So, that all levels of society can be accurately as well as independently represented.

2. After the data collection is completed, according to GDV cycle it is also important to store, secure and then process it. The next challenge we faced was creating an interface through which we could interact with our data smoothly and further store our findings in consistent and durable databases. We picked some rudimentary yet powerful techniques and libraries from python to do this job. This stage is more concerned with storage management, quality control, cataloging and data cleansing, ensuring that our data is accurate. This stage is also said to impact the public sector.

3. The next phase of stage was sharing, curating and publishing the data. This is concerned with handling requests and agreements, utilizing data sharing platforms and publishing data on open data websites. However, with our studies still in early drafts, we wanted to make sure that our data is further refined with future work in mind to share it in open sites. It is still evident to say that most of our data is derived from public sources and requested from the public sector.

4. The final phase of the GDV cycle is just the constant use and reuse of this data to draw conclusions for various purposes such as statistical analysis, machine learning and visualization. We have delved in all aspects of these, which we will further see in coming subsections. This four step procedure ensures data accuracy, reliability, analysis and further decision making to enhance the overall development process.

2.2. Data Collection:

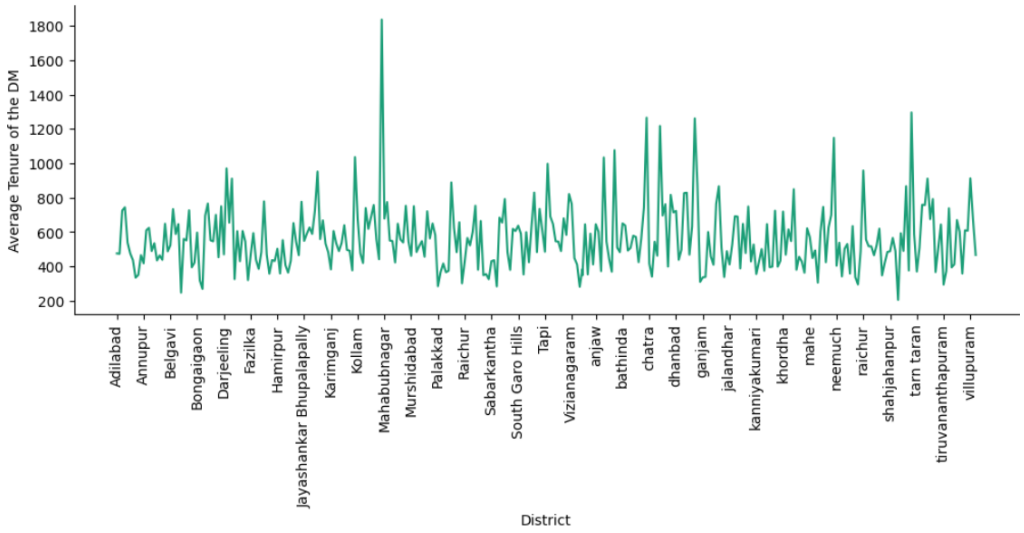

Figure 2: Average Tenure vs. Index

Commencing this research to enhance the efficiency of bureaucrats (i.e District Magistrates here) across the many districts of India needed immense attention to the step of data collection. Due to its huge size and the need to factor-in more than 100 attributes as we can see in Table-1, this required collection of data from various trustable sources to ensure data quality and consistency. The data about the average tenures of the District magistrates across different districts was collected from the officials at Niti Aayog as shown in Figure-2. The Niti Aayog serves as the main site for any new developments or policy implementation for the whole nation. This Government agency focuses solely on the developmental priorities of the nation in various sectors. The NFHS data is also surveyed by the organization officials.

Factors influencing district
1. Female population age 6 years and above who ever attended school (%)
2. Population below age 15 years (%)
3. Sex ratio of the total population (females per 1,000 males)
4. Sex ratio at birth for children born in the last five years (females per 1,000 males)
5. Children under age 5 years whose birth was registered with the civil authority (%)
6. Deaths in the last 3 years registered with the civil authority (%)
7. Population living in households with electricity (%)
8. Population living in households with an improved drinking-water source1 (%)
9. Population living in households that use an improved sanitation facility2 (%)
10. Households using clean fuel for cooking3 (%)
11. Households using iodized salt (%)
12. Households with any usual member covered under a health insurance/financing scheme (%)
13. Children age 5 years who attended pre-primary school during the school year 2019-20 (%)
14. Women who are literate4 (%)
15. Women with 10 or more years of schooling (%)
16. Women age 20-24 years married before age 18 years (%)
17. Births in the 5 years preceding the survey that are third or higher order (%)
18. Women age 15-19 years who were already mothers or pregnant at the time of the survey (%)
19. Women age 15-24 years who use hygienic methods of protection during their menstrual period5 (%)
20. Any method6 (%)
21. Any modern method6 (%)
22. Female sterilization (%)
23. Male sterilization (%)
24. IUD/PPIUD (%)
25. Pill (%)
26. Condom (%)
27. Injectables (%)
28. Total unmet need7 (%)
29. Unmet need for spacing7 (%)
30. Health worker ever talked to female non-users about family planning (%)
31. Current users ever told about side effects of current method8 (%)
32. Mothers who had an antenatal check-up in the first trimester (%)
33. Mothers who had at least 4 antenatal care visits (%)
34. Mothers whose last birth was protected against neonatal tetanus9 (%)
35. Mothers who consumed iron folic acid for 100 days or more when they were pregnant (%)
36. Mothers who consumed iron folic acid for 180 days or more when they were pregnant (%)
37. Registered pregnancies for which the mother received a Mother and Child Protection (MCP) card (%)
38. Mothers who received postnatal care from a doctor/nurse/LHV/ANM/midwife/other health personnel within 2 days of delivery (%)

Factors influencing district
39. Average out-of-pocket expenditure per delivery in a public health facility (Rs.)
40. Children born at home who were taken to a health facility for a check-up within 24 hours of birth (%)
41. Children who received postnatal care from a doctor/nurse/LHV/ANM/midwife/other health personnel within 2 days of delivery (%)
42. Institutional births (%)
43. Institutional births in public facility (%)
44. Home births that were conducted by skilled health personnel10 (%)
45. Births attended by skilled health personnel10 (%)
46. Births delivered by caesarean section (%)
47. Births in a private health facility that were delivered by caesarean section (%)
48. Births in a public health facility that were delivered by caesarean section (%)
49. Children age 12-23 months fully vaccinated based on information from either vaccination card or mother's recall11 (%)
50. Children age 12-23 months fully vaccinated based on information from vaccination card only12 (%)
51. Children age 12-23 months who have received BCG (%)
52. Children age 12-23 months who have received 3 doses of polio vaccine13 (%)
53. Children age 12-23 months who have received 3 doses of penta or DPT vaccine (%)
54. Children age 12-23 months who have received the first dose of measles-containing vaccine (MCV) (%)
55. Children age 24-35 months who have received a second dose of measles-containing vaccine (MCV) (%)
56. Children age 12-23 months who have received 3 doses of rotavirus vaccine14 (%)
57. Children age 12-23 months who have received 3 doses of penta or hepatitis B vaccine (%)
58. Children age 9-35 months who received a vitamin A dose in the last 6 months (%)
59. Children age 12-23 months who received most of their vaccinations in a public health facility (%)
60. Children age 12-23 months who received most of their vaccinations in a private health facility (%)
61. Prevalence of diarrhoea in the 2 weeks preceding the survey (%)
62. Children with diarrhoea in the 2 weeks preceding the survey who received oral rehydration salts (ORS) (%)
63. Children with diarrhoea in the 2 weeks preceding the survey who received zinc (%)
64. Children with diarrhoea in the 2 weeks preceding the survey taken to a health facility or health provider (%)
65. Prevalence of symptoms of acute respiratory infection (ARI) in the 2 weeks preceding the survey (%)
66. Children with fever or symptoms of ARI in the 2 weeks preceding the survey taken to a health facility or health provider (%)
67. Children under age 3 years breastfed within one hour of birth15 (%)
68. Children under age 6 months exclusively breastfed16 (%)
69. Children age 6-8 months receiving solid or semi-solid food and breastmilk16 (%)
70. Breastfeeding children age 6-23 months receiving an adequate diet16, 17 (%)
71. Non-breastfeeding children age 6-23 months receiving an adequate diet16, 17 (%)

Factors influencing district
72. Total children age 6-23 months receiving an adequate diet16, 17 (%)
73. Children under 5 years who are stunted (height-for-age)18 (%)
74. Children under 5 years who are wasted (weight-for-height)18 (%)
75. Children under 5 years who are severely wasted (weight-for-height)19 (%)
76. Children under 5 years who are underweight (weight-for-age)18 (%)
77. Children under 5 years who are overweight (weight-for-height)20 (%)
78. Women whose Body Mass Index (BMI) is below normal (BMI ¡18.5 kg/m^2)21 (%)
79. Women who are overweight or obese (BMI ≥ 25.0 kg/m^2)21 (%)
80. Women who have high risk waist-to-hip ratio (≥0.85) (%)
81. Children age 6-59 months who are anaemic (¡11.0 g/dl)22 (%)
82. Non-pregnant women age 15-49 years who are anaemic (¡12.0 g/dl)22 (%)
83. Pregnant women age 15-49 years who are anaemic (¡11.0 g/dl)22 (%)
84. All women age 15-49 years who are anaemic22 (%)
85. All women age 15-19 years who are anaemic22 (%)
86. Blood sugar level - high (141-160 mg/dl)23 (%)
87. Blood sugar level - very high (¿160 mg/dl)23 (%)
88. Blood sugar level - high or very high (¿140 mg/dl) or taking medicine to control blood sugar level23 (%)
89. Blood sugar level - high (141-160 mg/dl)23 (%)
90. Blood sugar level - very high (¿160 mg/dl)23 (%)
91. Blood sugar level - high or very high (¿140 mg/dl) or taking medicine to control blood sugar level23 (%)
92. Mildly elevated blood pressure (Systolic 140-159 mm of Hg and/or Diastolic 90-99 mm of Hg) (%)
93. Moderately or severely elevated blood pressure (Systolic ≥160 mm of Hg and/or Diastolic ≥100 mm of Hg) (%)
94. Elevated blood pressure (Systolic ≥140 mm of Hg and/or Diastolic ≥90 mm of Hg) or taking medicine to control blood pressure (%)
95. Mildly elevated blood pressure (Systolic 140-159 mm of Hg and/or Diastolic 90-99 mm of Hg) (%)
96. Moderately or severely elevated blood pressure (Systolic ≥160 mm of Hg and/or Diastolic ≥100 mm of Hg) (%)
97. Elevated blood pressure (Systolic ≥140 mm of Hg and/or Diastolic ≥90 mm of Hg) or taking medicine to control blood pressure (%)
98. Ever undergone a screening test for cervical cancer (%)
99. Ever undergone a breast examination for breast cancer (%)
100. Ever undergone an oral cavity examination for oral cancer (%)
101. Women age 15 years and above who use any kind of tobacco (%)
102. Men age 15 years and above who use any kind of tobacco (%)
103. Women age 15 years and above who consume alcohol (%)
104. Men age 15 years and above who consume alcohol (%)

Table 1: **Attributes considered that influence district**

The National Family & Health Survey (NFHS) in India is a comprehensive effort to collect data on a range of topics, including health issues, nutrition, family planning, domestic violence, and more. Surveys are conducted among a sample of Indian households from all states. The fifth edition of the NFHS, conducted from 2019 to 2021, produced reports that were publicly available in 2021. However, these reports are typically delivered in PDF format, which complicates computational processing.

One significant problem is extracting data from these PDFs, which are human-readable but difficult for machines to interpret efficiently. To solve this, district-specific reports were made, extracting data from tables, and converting it to machine-readable JSON format using parsing via beautiful soup. We attempted to verify the correctness of the parsed data but manually verifying all 704 PDF files and their outputs proved impossible. As a result, the resulting JSON files may have small inaccuracies.

The data retrieval technique begins with viewing the NFHS webpage, which serves as the primary source of this information. A csv file containing each state's data links is generated by extracting links to each state's unique page from this website. The python script to collect all district-specific file URLs is run and it saves all URLs in a csv file with each district link for convenient access.

The python script to download each districts' data is then run to begin downloading PDF files, resulting in download of the data district wise. Upon conclusion of the data collecting phase, 321 district data is collected. The following step is processing these PDF files to extract useful information. The parsing script plays an important role in this process, since it uses the Tabula and pdfminer.six libraries to parse PDF files efficiently. The retrieved data is then converted to JSON format and saved for further analysis and use.

Finally, to integrate all of the acquired data into a single coherent dataset, the information gained through the aforementioned steps is put into a single file which allows for simple access and reference in future endeavors.

2.3. Data analysis:

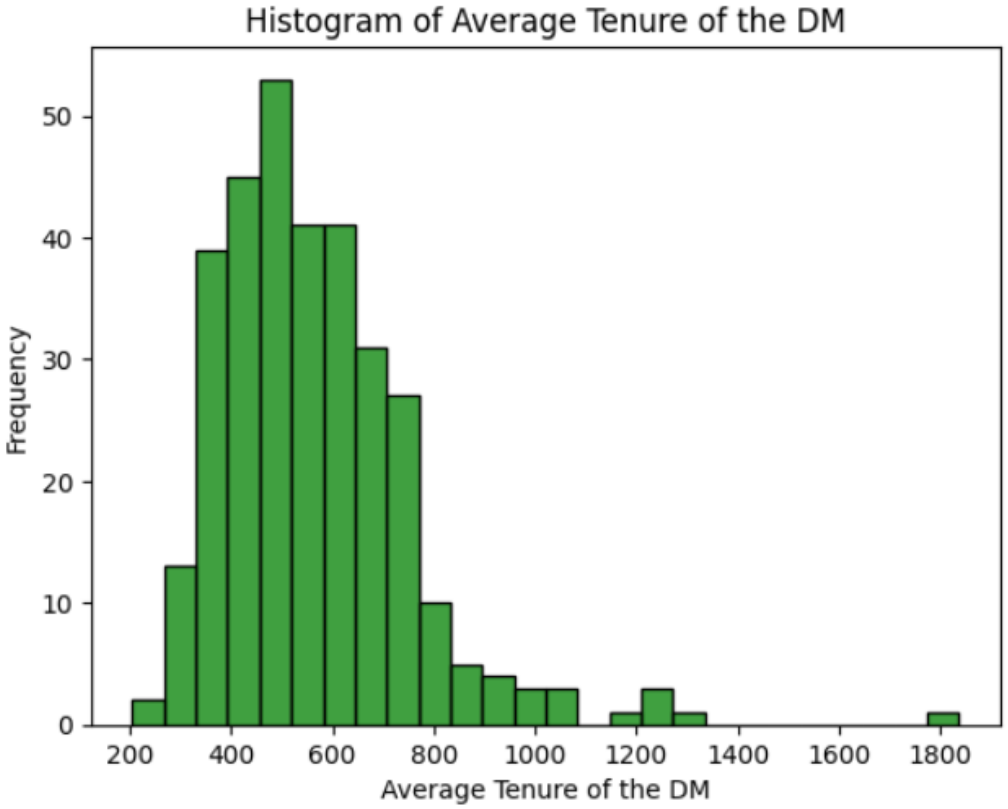

Figure 3: Histogram of Data

From Figure-3, it is evident that the highest frequency of average tenure is approximately **500 days**, which does not allow a District Magistrate to stabilize , adapt in the area or to find problems relating to that area. Conducting surveys of the district to provide analysis of the same typically requires around 9 to 12 months[6][7] amounting to 365 days. This is just a collection of data for the formation of policies, to formulate a policy for development of a district and then implementing it requires further time[8]. Then the DM is required to monitor and derive insights from the policy implementation. However based on the above given data, only 4-5 months are left for these activities.

In the data analysis step we begin by loading the relevant libraries like pandas,scikit-learn, numpy, and matplotlib to perform tasks like data manipulation, machine learning modeling, numerical calculations, and visualization. After that the extracted datasets are loaded into pandas dataframes. For data preparation a new column called “Improvement” is added to the dataframe which is calculated as the difference between the ”NFHS-5” and ”NFHS-4” columns. After that an inner join operation is performed on the dataframes using the “District” column resulting in a combined dataframe.This dataset now includes improvement on a district-wise as

well as attribute-wise basis[9][10][11].

The dataset is then organized by each attribute, allowing analysis on each individual indicator. This ensures that there is enough data for analysis. To ensure data integrity, records with missing values in the ”Improvement” column are removed.

Algorithm 1: Prediction of improvements based on average tenure

Input : List of indicator groups
Output: Average array of predicted improvements

```
Initialize an empty list all_predictions;
Define the range of average tenure values average_tenure_range;
foreach indicator, group in indicator_groups do
    if len(group) > 1 then
        Drop rows with NaN values in 'Improvement' column from group;
        if len(group) > 1 then
            Extract features (X) and target (y) from group;
            Reshape X to 2D array;
            Fit the SVR model on X and y;
            Reshape average_tenure_range to match the expected format;
            Use the trained model to make predictions;
            Store the predictions in all_predictions;
            Print predictions for the current indicator;
        end
        else
            Print ”Insufficient data for indicator. Skipping...”;
        end
    end
end
Stack the predictions vertically;
Calculate the column-wise average;
```

An empty list is initialized to store predictions for all attributes as we can see from algorithm-1. The range of average tenure values is defined using numpy's arange function, covering values from 365 to 1460 days in steps of 50 days. After data preparation, the machine learning modeling process is initiated and the dataset is divided into feature vectors (X) and goal variables (y), which indicate the average tenure of the District Magistrate (DM) and the improvement. The feature vector X is transformed into a 2D array, and the dataset is split into training and testing sets with the train_test_split function.Then the Support Vector Regression (SVR) model[12] [13] is created and trained using the training data. Predictions are produced on the test set, and the model's performance is measured using the Mean Squared Error metric.

Predicted improvements for each attribute are stored in a separate list and are shown for analysis. Then the predictions are stacked vertically, and the column-wise average is calculated to find a combined prediction.

The results are then visualized to better understand the data and result. Scatter plots are generated to show the association between improvements and average tenures. Regression lines giving SVR prediction are also generated. Predicted improvements are then visualized against average tenure showing the model's predicting abilities

Furthermore, we do column-wise averages of the expected progress which are generated and then visualized to help comprehend the overall trend across many metrics. These types of visualizations also help in uncovering the pattern, trends and links in data. It also allows for meaningful interpretations and conclusions. Throughout the study, we paid careful attention to our data quality, using model metrics we checked model performance and made changes on the basis of it. After result interpretation we derived solid and credible findings. Later to which we engaged in a debate and inference of our findings. The results are derived and published only after thought out reasoning.

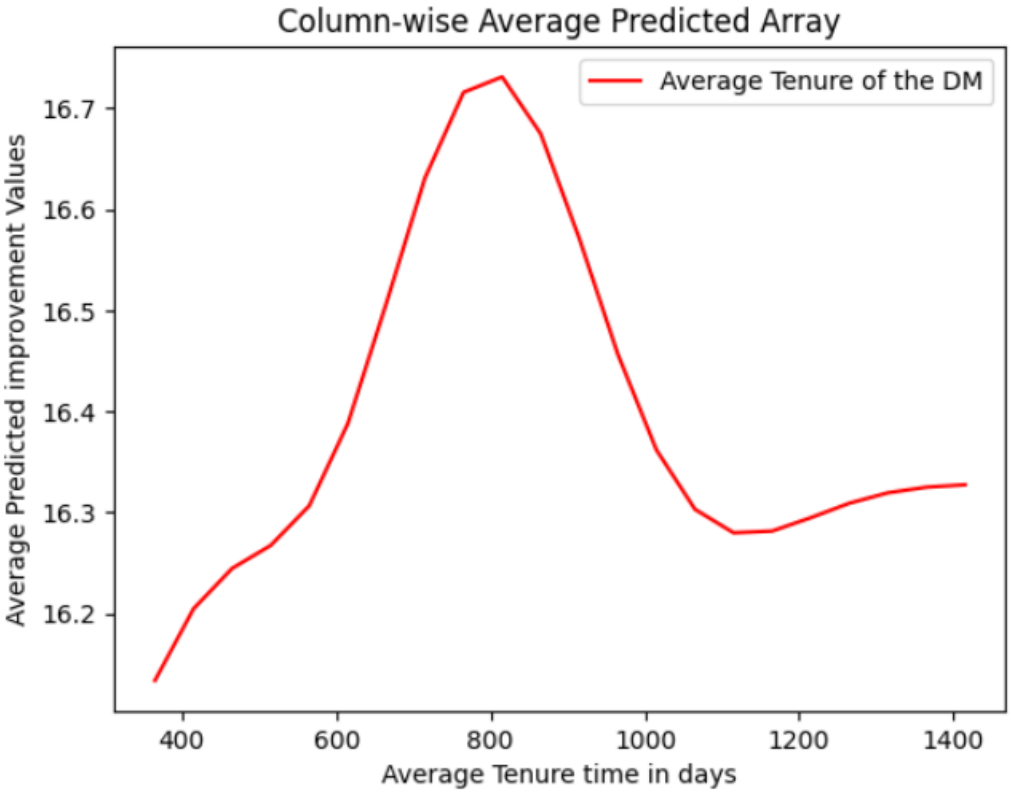

Figure 4: Predicted Improvement

Finally, we plotted the average predicted values against the average tenure range. This visualization shows the relationship between the average tenure and predicted improvements. The resulting plot provides insights into the expected improvements which correspond to different average tenure duration. This is visualized in Figure-4 shown above.

2.4. Net grade points:

Algorithm 2: Calculate Net Grade Points

Input: DataFrame df
Output: DataFrame net_grade_points_df

Function `calculate_net_grade_point(`*score_2019, score_2021*`)`
  improvement ← *score_2021* − *score_2019*;
  switch *score_2019* **do**
    case *10 to 19* **do**
      contribution ← 1 × *improvement*;
    end
    case *20 to 29* **do**
      contribution ← 1.5 × *improvement*;
    end
    case *30 to 39* **do**
      contribution ← 2 × *improvement*;
    end
    case *40 to 49* **do**
      contribution ← 2.5 × *improvement*;
    end
    case *50 to 59* **do**
      contribution ← 3 × *improvement*;
    end
    case *60 to 69* **do**
      contribution ← 3.5 × *improvement*;
    end
    case *70 to 79* **do**
      contribution ← 4 × *improvement*;
    end
    case *80 to 89* **do**
      contribution ← 4.5 × *improvement*;
    end
    case *90 to 100* **do**
      contribution ← 5 × *improvement*;
    end
    otherwise do
      contribution ← 0;
    end
  end
  return *score_2019, score_2021, contribution*;

grouped_df ← Group *df* by 'District' column;
Initialize an empty list *net_grade_points*;
foreach *group in grouped_df* **do**
  total_contribution ← 0;
  row_count ← 0;
  foreach *row in group* **do**
    score_2019, *score_2021*, *contribution* ← `calculate_net_grade_point(`*NFHS-5, NFHS-4*`)`;
    total_contribution ← *total_contribution* + *contribution*;
    row_count ← *row_count* + 1;
  end
  average_net_grade_point ← *total_contribution*/*row_count*;
  Append (*group*, *average_net_grade_point*) to *net_grade_points*;
end
net_grade_points_df ← Create DataFrame from *net_grade_points*;

Net grade point approach as mentioned in Algorithm-2, for determining the best performed districts and analyzing the Tenure data against it. This is another approach to support our result. In this, we define a function called calculate net grade point, designed to compute the net grade point based on the scores obtained from NFHS-4 and NFHS-5. This function takes as input the scores from both surveys and calculates the improvement as the difference between the NFHS-5 score and the NFHS-4 score. Later it is used to determine the contribution to the net grade point based on the score range observed in NFHS-4 and the magnitude of improvement observed. Higher grade points are assigned if the improvement occurs at the higher end of the score range. Finally, the function returns the scores from NFHS-4 and NFHS-5, along with the contribution to the net grade point.

We group the data by the 'District' column, creating a grouped DataFrame where each group represents data from a distinct district. This step helps the further computation of net grade points on per-district basis. We compute the net grade point for each district. First, we initialize an empty list to store the net grade points of each district. After which we iterate through each group in our DataFrame. Within this loop, we further iterate through each row in the group and apply the calculate net grade point function to compute the net grade point for each row. In the end we sum up the contributions from all rows within the district and track the count of rows. Now, we compute the average net grade point for the district by doing a division of the total contribution by the count of rows. After this appending the district name and its average net grade point to the list of net grade points gives us our final result.

Finally we created a new DataFrame from the list of net grade points, which contains the district names and their corresponding average net grade points. This DataFrame provides a comprehensive overview of the net average grade point for each district. We output the DataFrame containing the net average grade point for each district, thereby concluding the process of computing and organizing the net grade point data for further analysis and interpretation.

Note: The grade point calculation follows a formula where the contribution to the net grade point is determined by the improvement observed in the scores from NFHS-4 to NFHS-5. Higher improvement at the higher end of the score range results in a higher grade point, with a decreasing factor 'k' as the improvement occurs at lower ranges. For example, if the improvement is from 91 to 100, the grade point is calculated as 5 * improvement, while if the improvement is from 60 to 70, the grade point is calculated as 3.5 * improvement.

3. Challenges

Indian Administrative Services(IAS) across more than 350 districts in the country is quite challenging. It ranges from collecting data while ensuring its quality and consistency from different sources and formats to selecting the appropriate and best suited machine learning algorithm for the dataset, this endeavor comes with a lot of hurdles and challenges. This section delves into the various challenges encountered in the execution of this research which scales to more than 100 attributes and more than 30000 tuples. From scraping data from official sources to finding personal sources to facilitate the collection of data, each challenge poses a unique hurdle which needs a more innovative solution in overcoming it.

1. Understanding and addressing these challenges is crucial for successfully predicting and recommending the best possible timeframe aimed at improving and facilitating the development of the nation, starting with individual districts.
2. Data Collection: Collecting sufficient data required for machine learning techniques on individual districts' improvement based on several factors required use of various government sites like NFHS on which each of the districts data was given in PDF format and had to be downloaded for each district individually. Manually downloading data would require a huge amount of time and effort, so a solution to web scrape the data was thought of and executed. It allowed us to save our time by a significant amount.
3. Data Quality and Consistency: The data was collected for 2017 survey and 2021 surveys of NHFS in which some of the districts were missing from both of the surveys, this made the data inconsistent and incomplete. The districts for which data was not provided in both survey were removed to ensure the data remains consistent.
4. Data Processing: Analyzing 350+ districts amounting to 30000+ tuples with high dimensionality from the database required huge computing power which was not possible on a normal set of computer. So to mitigate this issue online distributed computing platforms such as Google Colab was used.
5. Machine Learning Algorithm: Choosing the best machine learning algorithm for this large of a database is a problem in itself as it requires a well thought out approach. First various algorithms like linear regression[14], KNN regressor[15], SVM regressor[16] were applied and after finding out their results, it was decided to implement a multi model approach[17][18] which gives a better and combined result.

4. Performance Evaluation

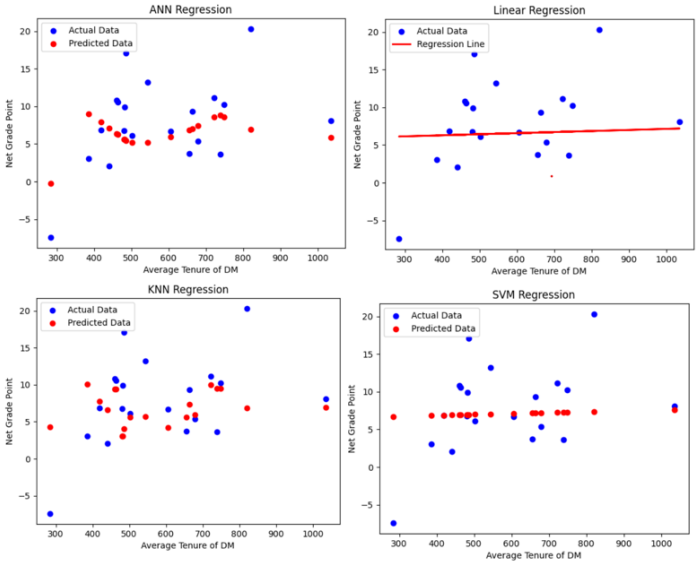

Figure 5: Performance Evaluation

Model	Mean Squared Error (MSE)
Linear Regression	32.98
ANN Regressor	**30.55**
KNN Regressor	36.56
SVM Regressor	32.01

Table 2: Mean Squared Error Comparison of Different Algorithms

To support our best average predicted tenure study, we used net average grade point score calculation on the basis of tenure days. For doing the prediction of an continuous dependent variable Net-Grade-Point on the basis of independent variable Tenure days, we used four different regression algorithms: SVR regressor, Linear regressor, KNN regressor, ANN regressor as mentioned in the Figure-5.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{1}$$

We compared the performance of the four algorithms using the evaluation matrix mean squared error as mentioned in equation-(1) and evaluated performance is mentioned in Table-2. Among these, the ANN Regressor achieved the lowest mean squared error (MSE) of 30.55, indicating its superior performance in predicting the target variable compared to the other algorithms. This superior performance of the ANN Regressor can be attributed to its ability to capture complex nonlinear relationships between the input features and the target variable. Unlike linear regression, which assumes linear relationships, the ANN Regressor can learn intricate patterns and dependencies in the data, making it more suitable for modeling real-world problems with complex data distributions.

In contrast, linear regression, SVM regressor, and KNN regressor exhibited comparatively higher MSE values, suggesting limitations in their predictive accuracy. Linear regression, for instance, operates under the assumption of linear relationships, which may not adequately capture the complexities present in the dataset. Similarly, SVM regressor and KNN regressor, while powerful in certain contexts, may struggle with capturing nuanced patterns and dependencies within the data, leading to sub-optimal predictions.

5. Result Analysis and Discussion

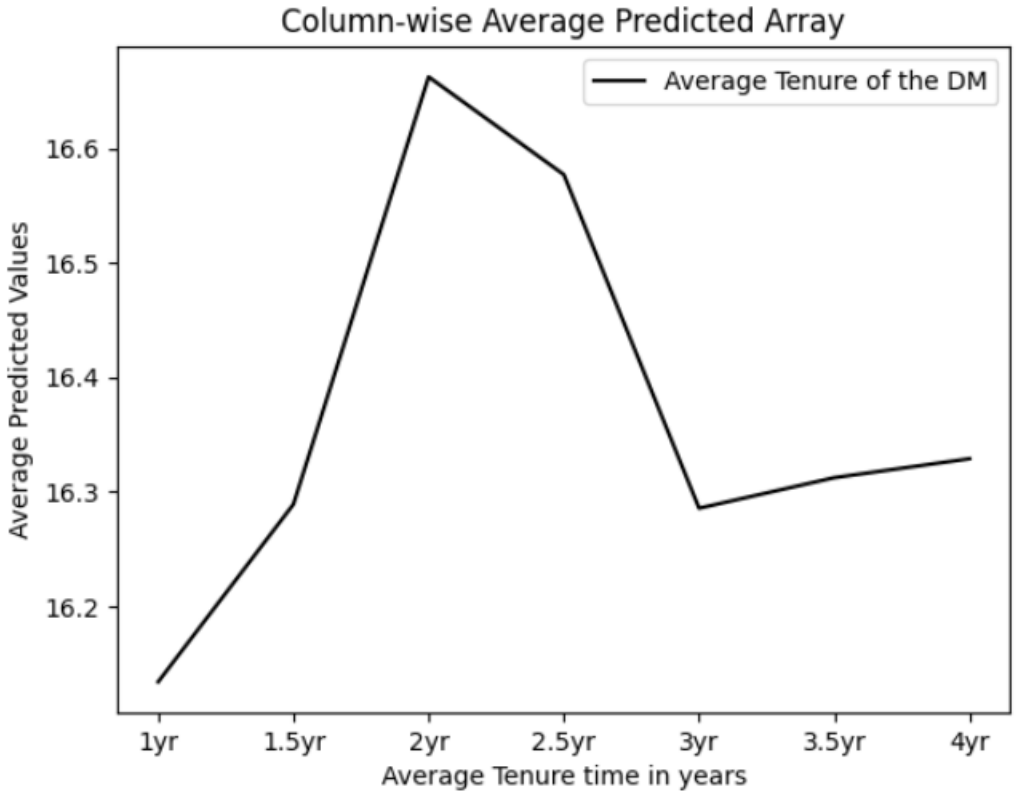

Figure 6: **Improvement**

The research predicts that the highest overall improvement should be achieved with a District Magistrate's tenure of approximately 730 days (Figure-6), amounting to 2 years, based upon given metrics. This shows that providing DMs with tenure of approximately 2 years may result in positive outcomes in general improvement and development of the desired areas

Index	Average Tenure of the DM	Net Grade Point
Valsad	820.800	20.30
Dahod	462.875	18.19
Surat	734.500	17.66
Tapi	485.250	17.10
The Dangs	646.714	16.81
Chirang	545.333	16.05

Table 3: **Best net grade points versus districts and their Tenures**

In Table-3, we can see the correlation between the best net grade points and district tenures. It

is evident that districts such as Valsad, Dahod, and Surat, having the highest grade points, have tenures of 820, 452, and 734 days, respectively. These findings further prove our prediction that a tenure of around 730 days, or 2 years, can show significant district improvement.

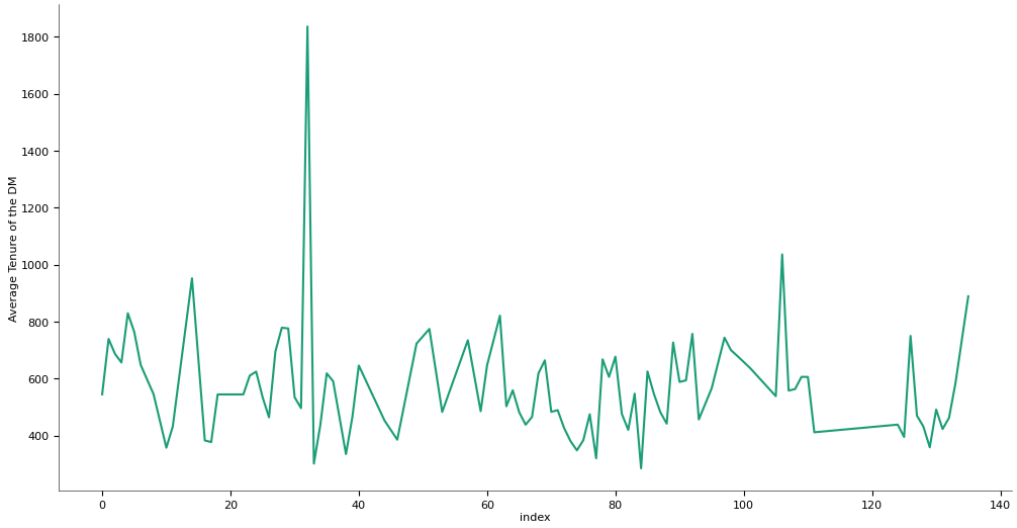

Figure 7: **Index (Represents Districts) vs Net Grade Point**

Understanding district conditions typically requires 9 to 12 months, which involve taking surveys the whole district [2]. After this, formulation and implementation of a policy take approximately 6 months, consisting of survey analysis, identification of under-performing areas, and establishing correlations between various factors such as sanitation and health outcomes, including malnutrition rates. The remaining 6 months can be used to conduct follow-up surveys to analyze the impact of implemented policies. Thus, the data presented in Figure-7 aligns with our hypothesis, showing that a tenure of around 730 days may give the most favorable outcomes for district improvement.

We can also evidentiate our results by plotting the net grade point versus the number of days data and seeing the most optimal days where the score for the district is maximised.

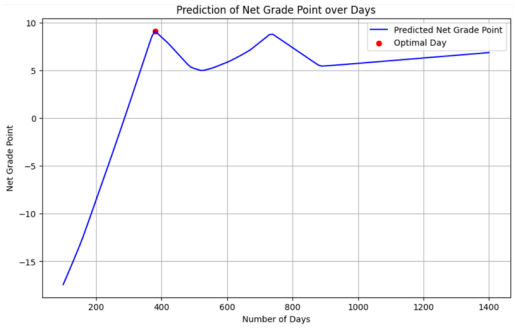

Figure 8: **Optimal Grade Point Image**

As shown in Figure-8, the grade point image which is plotted against the number of tenure days using the ANN regressor as mentioned in table-2, it gives least mean squared error. On analysing the image, it illustrates that number of days at which the net average grade point is highest comes around 380 days but after that graph again decreases that shows that may be after the survey, the things are not implemented optimally. Again the graph starts rising forming an rounding bottom pattern and gives hike at 740 days. This evident that our study in figure-5 which also give hike at nearly 2 year tenure time which is around 730 days correct.

Hence overall there are two evidences that supports our study.

6. Proposed Strategy for District Magistrates to Enhance Efficiency

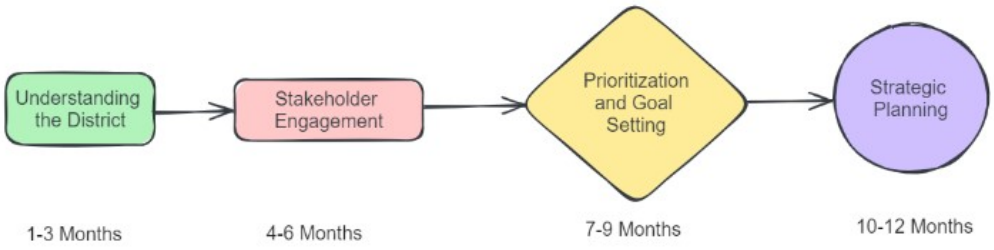

Figure 9: **Year-1 Plan**

Year 1 Analysis:

We can give an overall broader plan for best possible optimization of the DM tenure as illustrated in Figure-9. As the DM is posted in a district, analyzing, understanding and gathering the data on the demographics, economy, infrastructure, education, healthcare, agriculture, industries, employment, environmental factors, Understanding the political agenda of the region, the diverse mindset of people in a region etc is necessary. It takes nearly 1-3 months of time to properly go through the district's fundamental and socio-economic factors. Finally establishing the communication channel with stakeholders.

Further in the phase-2 of the first year i.e. from Month 4 to 6, stakeholder Engagement is a necessary task. Conducting meetings with local politicians, town halls, and focus groups with government officials, community leaders, NGOs, businesses, and residents. This gives insights on the needs, priorities, and aspirations of stakeholders.

In the phase-3 i.e. from 7 to 9 months, DM will identify the key areas of development on the basis of data analysis and stakeholders input. Further setting up the SMART goals for each priority area.

In phase-4 i.e. from 10 to 12th month, DM will develop a strategic plan according to goals set. Development of strategies, taking initiatives, and defining timelines for every area of development must be done. Proper Allocation of resources effectively and identifying potential

funding sources so that resources can be utilized in the best way and fast execution of plans can be done.

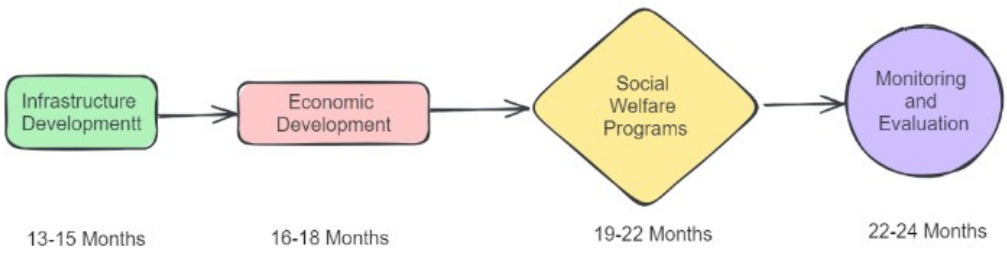

Figure 10: **Year-2 Plan**

For the execution on the policies made in year-1 based on analysis of district where the district is not developed enough in certain factors, year-2 can be utilized according to strategy of execution as mentioned in Figure-10. For the months 13 to 15, major focus should be on Infrastructure Development which includes improving roads, electricity, water supply, sanitation, and public transportation, investment in digital infrastructure for e-governance and connectivity. Development should be digitized which ultimately leads to improved delivery and efficiency of government services, improved government interactions with business and industry, citizen empowerment through access to information, more efficient government management, less corruption in the administration, increased transparency in administration.

From month 16 to 18, major focus should be on Economic development. Policies should be made in such a way that promotes entrepreneurship, SMEs (small and medium enterprises), and local industries. Developing tourist hotspots and investing in the key sectors such as agriculture which promotes the development of the poor and middle class[19][20]. Also invest in manufacturing which helps in development of medium to large scale business. Economic development overall leads to more employment.

From Month 19 to 21, Social Welfare Programs should be implemented targeting vulnerable populations. Policies should be made to Provide access to healthcare, education, skill development, and social security schemes.

From Month 22 to 24, Monitoring and Evaluation of the implemented plans and strategies should be done to track progress against predefined indicators. Conduct regular reviews and evaluations to assess effectiveness and impact. Evaluation of the implemented strategies should be done across the various district reports which overall promotes healthy competition and development.

7. Conclusion

The study's results underscore the importance of considering tenure duration as a significant factor in achieving improvements in various domains. By identifying the optimal average tenure time of 730 days (approximately 2 years), policymakers and stakeholders can make informed decisions regarding DM tenure durations. Moreover, the findings contribute to enhancing administrative effectiveness and governance by providing insights into the relationship between tenure duration and performance metrics. Future research could delve deeper into understanding the mechanisms underlying the observed association and explore additional factors influencing improvement outcomes. Overall, the study highlights the critical role of tenure management in fostering positive outcomes and underscores the need for evidence-based decision-making in governance.

8. Future work

Future works would be on enhancing the model with the adoption of additional algorithms and refinement of the parameters of the ones currently used. We can try out other machine learning techniques and fine-tune their parameters in a quest to find better patterns in the data that will result in better predictions. More elaborate pre-processing of datasets, including cleaning, feature engineering, and normalizing, may still improve the model to draw meaningful insights from the information available.

Future extensions that will involve data collection regarding several parameters influencing the growth of districts would be another line of work. Other datasets beyond the use of demographic indicators, administrative data, and previous family health surveys, education attainment, land use and industrial development metrics, and employment statistics would be used to provide further context. In general, these kinds of datasets will help to widen our comprehension of the multiple factors that are responsible at the district level and make our predictive models more accurate and relevant.

References

[1] A. R. Vittoria Biagi, Data model design to support data-driven it governance implementation, Techologies 10 (2022).
URL https://www.mdpi.com/2227-7080/10/5/106

[2] K. S. Ebenezer Agbozo, Establishing efficient governance through data-driven e-government, ICEGOV '18 11 (2018) 662–664.
URL https://dl.acm.org/doi/abs/10.1145/3209415.3209419?casa_token=XvjcFIXQuF4AAAAA:4ppTj3OWPkw4NmzUKYGR8EUcdrlk

[3] G. L. J. K. Patrick Mikalef, Maria Boura, The role of information governance in big data analytics driven innovation, Information and management 57 (2020).
URL https://www.sciencedirect.com/science/article/pii/S0378720620302998

[4] N. Oliver, Predicting gender from mobile phone metadata, FAT '18: Proceedings of the 6th ACM International Conference on Fairness, Accountability, and Transparency (2018).
URL https://www.nuriaoliver.com/papers/Oliver_FATEN.pdf

[5] E. G. Juenke, The tenure process and extending the tenure, Journal of Public Administration Research and Theory 19 (3) (2009) 445–459.
URL https://link.springer.com/article/10.1057/hep.2009.18

[6] The Hindu, Wait for land survey is as long as 9 months in some districts, accessed on 2024-05-01 (2024).
URL https://www.thehindu.com/news/national/karnataka/wait-for-land-survey-is-as-long-as-9-months-in-some-distri
[7] T. H. BUREAU, Rajya sabha panel pulls up union government for delay in framing rules for acts passed by the parliament, The HINDU (2023).
URL https://www.thehindu.com/news/national/rajya-sabha-panel-pulls-up-union-government-for-delay-in-framing-rul
[8] PRS India, How long can the central government take to frame rules?, accessed on 2024-05-01 (2024).
URL https://prsindia.org/theprsblog/how-long-can-the-central-government-take-to-frame-rules
[9] R. Sharma, K. R. Biedenharn, J. M. Fedor, A. Agarwal, Trends of infertility and childlessness in india: Findings from nfhs data, Journal of Human Reproductive Sciences 7 (2014) 159–170.
URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4188020/
[10] A. SENGUPTA, Gender inequality in well-being in india: Estimates from nfhs household-level data, Economic and Political Weekly 51 (2016) 43–50.
URL https://www.jstor.org/stable/44004046
[11] N. Dhirar, S. Dudeja, J. Khandekar, D. Bachani, Childhood morbidity and mortality in india – analysis of national family health survey 4 (nfhs-4) findings, Indian Pediatrics 55 (8) (2018).
URL https://link.springer.com/content/pdf/10.1007/s13312-018-1276-6.pdf
[12] G. G. Kemal Uçak, An adaptive support vector regressor controller for nonlinear systems, Soft Computing 20 (2016) 2531–2556.
URL https://link.springer.com/article/10.1007/s00500-015-1654-0
[13] A. raj, Unlocking the true power of support vector regression (2023).
URL https://towardsdatascience.com/unlocking-the-true-power-of-support-vector-regression-847fd123a4a0
[14] C.-L. T. Xiaogang Su, Xin Yan, Linear regression, Wires computational statistics 4 (2012).
URL https://wires.onlinelibrary.wiley.com/doi/abs/10.1002/wics.1198
[15] J. L. X. Z. Yunsheng Song, Jiye Liang, An efficient instance selection algorithm for k nearest neighbor regression, Neurocomputing 251 (2017) 26–34.
URL https://sciencedirect.com/science/article/abs/pii/S0925231217306884
[16] K. Uçak, G. Günel, An adaptive support vector regressor controller for nonlinear systems, Soft Computing 20 (2016) 247–256.
URL https://link.springer.com/article/10.1007/s00500-015-1654-0
[17] P. H. Verburg, B. Eickhout, H. van Meijl, The effect of agricultural trade liberalisation on land-use related carbon emissions: A global analysis, Environmental Science and Policy 10 (2007) 1–16.
URL https://link.springer.com/article/10.1007/s00168-007-0136-4
[18] D. S. Daoqiang Zhang, Multi-modal multi-task learning for joint prediction of multiple regression and classification variables in alzheimer's disease, Neuro Image 59 (2012) 895–907.
URL https://www.sciencedirect.com/science/article/abs/pii/S105381191101144X
[19] N. S. Krishna Kumar, Determinants of birth registration in india: Evidence from nfhs 2015–16 (2021).
URL https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0257014
[20] S. Barman, Socio-economic and demographic differentials of contraceptive usage in indian states: A study based on nfhs data, Journal of Human Ecology 42 (2013) 53–68.
URL https://www.tandfonline.com/doi/abs/10.1080/09709274.2013.11906581