PMKSY MIPMS Physical and Financial Report

A EDA PROJECT REPORT

Submitted for the partial fulfillment

of

EDA Project requirement of B. Tech CSE

Submitted by

Harsh Kubade, 22070521106

B. Tech Computer Science and Engineering

Under the Guidance of

Prof. / Dr.

Dr. Piyush Chauhan

SYMBIOSIS INSTITUTE OF TECHNOLOGY, NAGPUR

Constituent of Symbiosis International (Deemed University), Pune

(Established under Section 3 of the UGC Act of 1956 wide notification number F-9-12/2001-U-3 of Government of India)

॥वसुधैव कुटुम्बकम्॥ Re-Accredited by NAAC with 'A++' Grade

1. Introduction

This report aims to comprehensively analyze the PMKSY-MIPMS (Pradhan Mantri Krishi Sinchayee Yojana- Micro Irrigation Project Management System) dataset. It undertakes a meticulous exploration designed to illuminate the complex interplay of financial allocations, physical targets, and realized achievements. The dataset, comprehensive in scope, encompasses detailed records across various administrative geographies—specifically states and districts—and spans multiple financial years. Through this systematic exploration, the report will unearth crucial insights, such as regional disparities in scheme implementation, temporal trends in investment and progress, and the overall effectiveness of micro-irrigation initiatives. These insights are not just academic; they are pivotal for informing strategic decision-making, optimizing resource allocation, identifying areas for intervention, and ultimately, ensuring the more efficient and impactful execution of the PMKSY-MIPMS scheme in the future.

2. Data Loading and Transformation

The initial stage involved bringing the raw data into a usable format and performing necessary transformations to prepare it for analysis.

- Loading the Dataset: The dataset was loaded into a pandas DataFrame from the provided CSV file, pmksy-mipms-physical-and-financial-report.csv. This foundational step makes the data accessible for manipulation and analysis within the Python environment.
- Initial Data Inspection: Before any transformations, a preliminary inspection of the dataset was performed to understand its structure, column names, and data types. This helped in identifying potential areas that required cleaning or type conversion.
- Date Column Transformation: A crucial transformation involved converting the 'month' column into datetime objects. This step is essential for any time-series analysis, allowing for the proper ordering of data by date and enabling the extraction of temporal features like year, quarter, or month. The conversion also handled potential errors gracefully, coercing invalid date formats to NaT (Not a Time), thus preventing the entire process from failing.

3. Data Cleaning and Preprocessing

- To ensure the integrity and reliability of the data for subsequent analysis, a series of meticulous cleaning and preprocessing steps were executed. These steps were crucial in transforming the raw dataset into a clean, consistent, and analysis-ready format.

3.1 Handling Duplicates

- A fundamental initial step involved the rigorous identification and elimination of any duplicate rows present within the dataset. This process ensures that each record is genuinely unique, preventing any overrepresentation of specific data points that could otherwise skew statistical computations or lead to erroneous conclusions regarding frequencies and distributions. By removing redundant entries, the dataset's integrity is preserved, making the subsequent analysis more accurate and trustworthy.

3.2 Standardizing Column Names

- To facilitate seamless data access and consistent programming practices, all column names underwent a standardization process. This involved systematically inspecting each column header and removing any leading or trailing whitespace characters. Such standardization is vital because inconsistent naming conventions (e.g., "Column Name " vs. "Column Name") can lead to errors in data retrieval and manipulation, complicating the analytical workflow. A clean and uniform naming scheme ensures that data columns can be accessed predictably and efficiently.

3.3 Imputing Missing Numerical Values

- Addressing missing values is a critical aspect of data preprocessing. For all columns identified as containing numerical data (specifically float64 and int64 data types), any missing entries were imputed by filling them with a value of 0. This imputation strategy was chosen based on the contextual understanding of the dataset, where an absence of a numerical record (e.g., a missing financial achievement or physical target) is interpreted as a non-existent or zero value within the scope of the PMKSY-MIPMS scheme. This approach ensures that numerical computations are not disrupted by NaN (Not a Number) values and maintains the logical consistency of the data.

3.4 Normalizing String Data

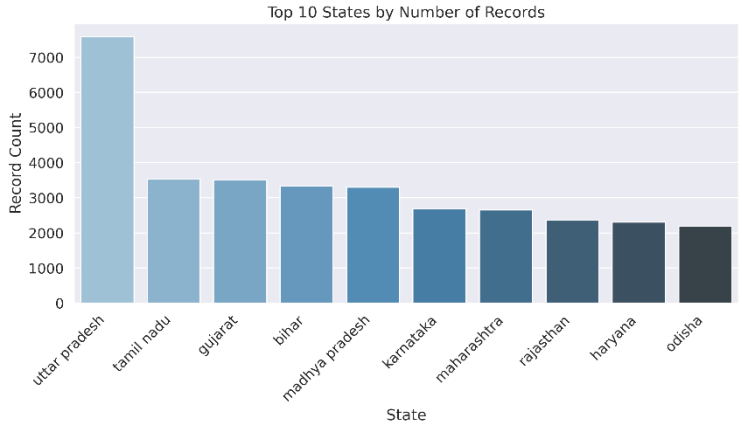
- Textual or categorical data, present in string-type columns (object data type), required careful normalization to ensure consistency and facilitate accurate aggregation and categorization. This process involved two key actions: first, all leading and trailing whitespace characters were removed from the string values, preventing distinct categories from being formed due to accidental spaces (e.g., "CategoryA " vs. "CategoryA"). Second, all text was converted to lowercase. This conversion eliminates issues arising from case sensitivity (e.g., "StateA" being treated differently from "stateA"), ensuring that identical categories are consistently grouped together. This normalization is paramount for accurate categorical analysis and avoids misleading results due to variations in string formatting.
- Upon the successful completion of these rigorous cleaning and preprocessing steps, the refined dataset was saved as pmksy_cleaned.csv. This final, clean dataset comprises 54,383 entries, each representing a meticulously prepared record, ready for in-depth exploratory data analysis and subsequent modeling initiatives.

4. Exploratory Data Analysis

- This section delves into the key visualizations generated during the exploratory data analysis, providing a deeper and more nuanced understanding of the dataset's inherent characteristics, patterns, and relationships. Each visualization serves as a powerful tool to uncover insights that might not be immediately apparent from raw data or summary statistics alone.

4.1 Top 10 States by Number of Records

- This bar plot visually represents the data's geographic distribution, clearly showing which states contribute the most entries to the PMKSY-MIPMS dataset. The prominence of states like Uttar Pradesh, Tamil Nadu, and Gujarat suggests that these regions either have been the focus of a larger volume of micro-irrigation projects under the scheme or maintain more comprehensive and frequent reporting mechanisms. This insight is crucial for understanding the operational footprint of the scheme and identifying areas with higher data generation activity.

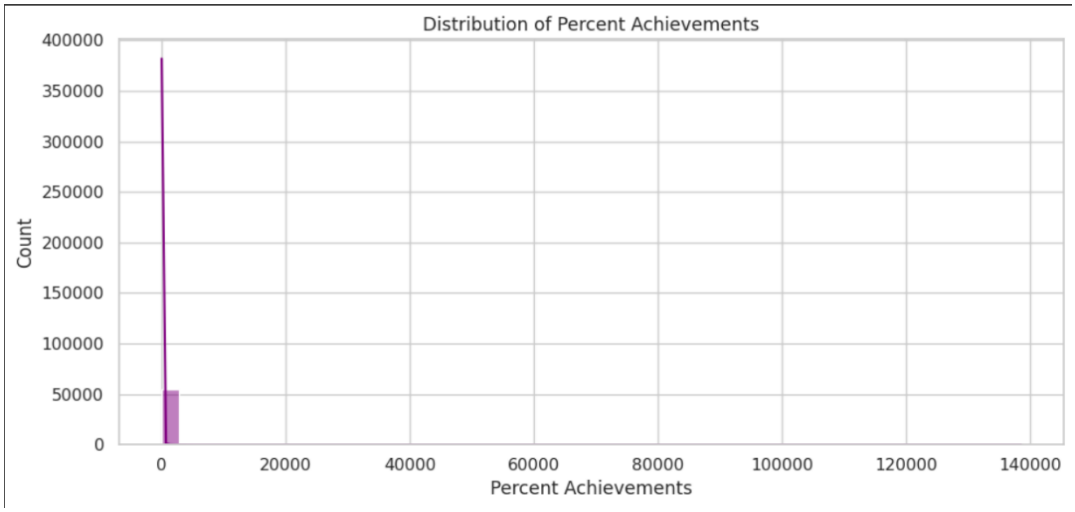


4.2 Total Financial Target Over Financial Years

- A detailed line plot illustrating the aggregate financial targets across different financial years reveals dynamic trends in planned investments. The plot conspicuously shows significant fluctuations in targets, with a pronounced peak observed around the 2020-21 financial year, followed by a subsequent decline in targets for the following years. This suggests shifts in governmental priorities, varying budgetary allocations, or changes in the overall scope and ambition of micro-irrigation projects over time. Analyzing these temporal patterns can help in understanding the scheme's evolution and resource allocation strategies.

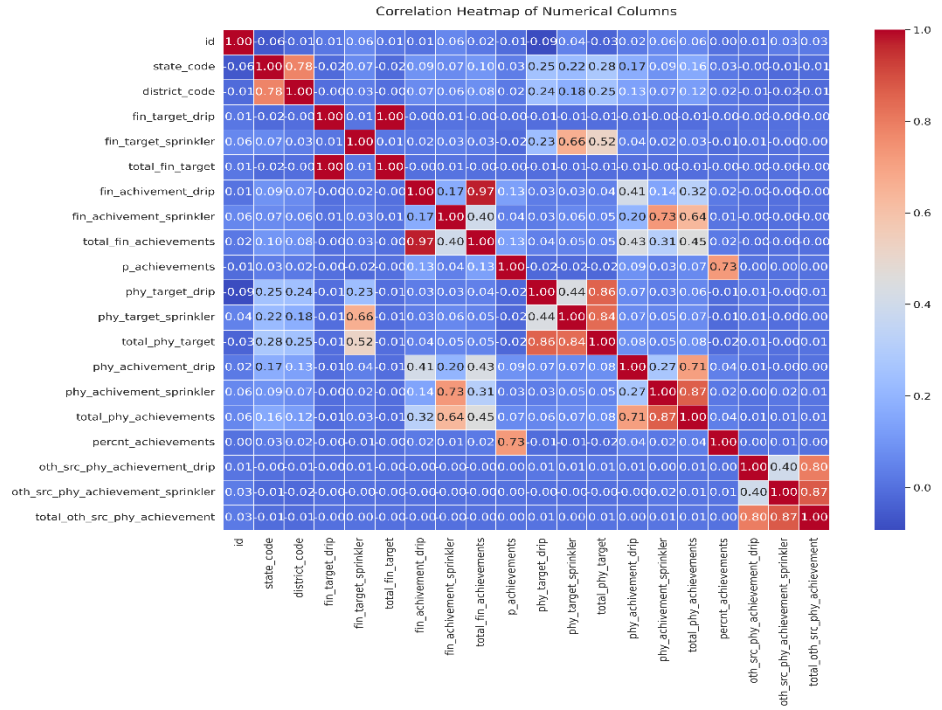
4.3 Distribution of Percent Achievements

- The histogram visualizing the 'percnt_achievements' column offers critical insights into the scheme's performance efficiency. It clearly demonstrates a highly skewed distribution, with a vast majority of the achievement percentages clustered around zero. However, the presence of extreme outliers, with the maximum value soaring to 138600.00, demands careful scrutiny. Such exceptionally high percentages could either signal extraordinary success stories or, more likely, point to data entry errors, specific accounting methodologies for over-achievement, or unique definitions of "percentage achievement" that diverge from standard interpretation. Further investigation into these outliers is essential for data validation.



4.4 Correlation Heatmap of Numerical Columns

- This heatmap is an indispensable tool for understanding the linear relationships between all numerical variables within the dataset. The color intensity and direction reveal strong positive correlations, such as the near-perfect relationship between fin_target_drip and

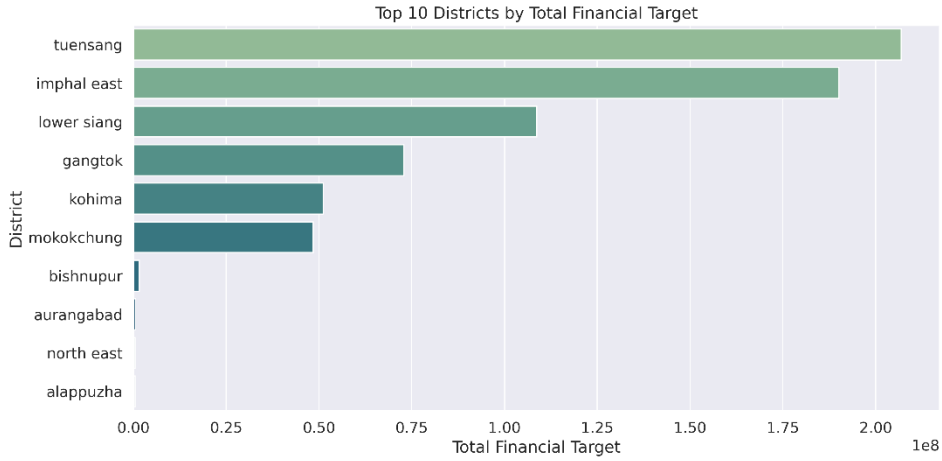


total_fin_target, and similar strong links between physical targets and their respective totals.

These high correlations indicate that the drip irrigation component, along with sprinkler irrigation, are the primary drivers influencing the overall financial and physical targets and achievements. This insight is fundamental for identifying the core components dictating the scheme's performance and impact.

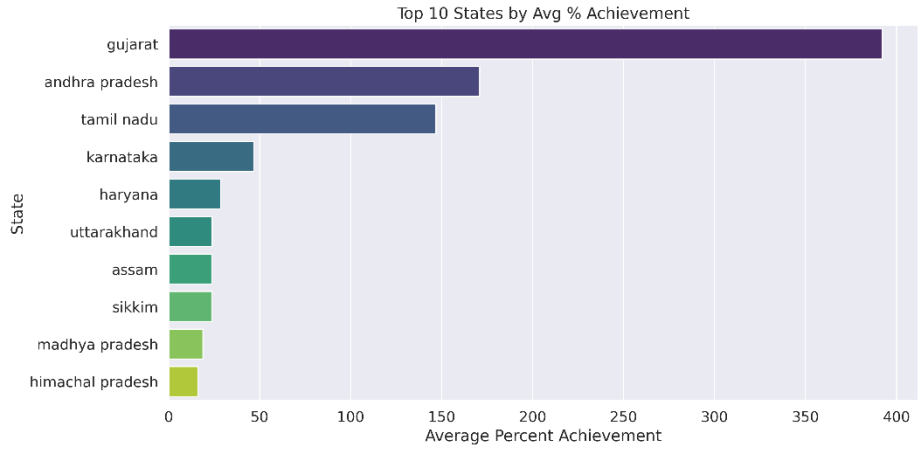
4.5 Top 10 Districts by Total Financial Target

- This bar plot zeroes in on the sub-state level, identifying districts that have received the highest planned financial investments. The prominence of districts like Tuensang, Imphal East, and Lower Siang signifies areas where substantial financial resources are earmarked for micro-irrigation projects. This granular view helps in pinpointing specific geographic regions of intense focus and planned expenditure, which can be critical for localized resource monitoring and impact assessment.



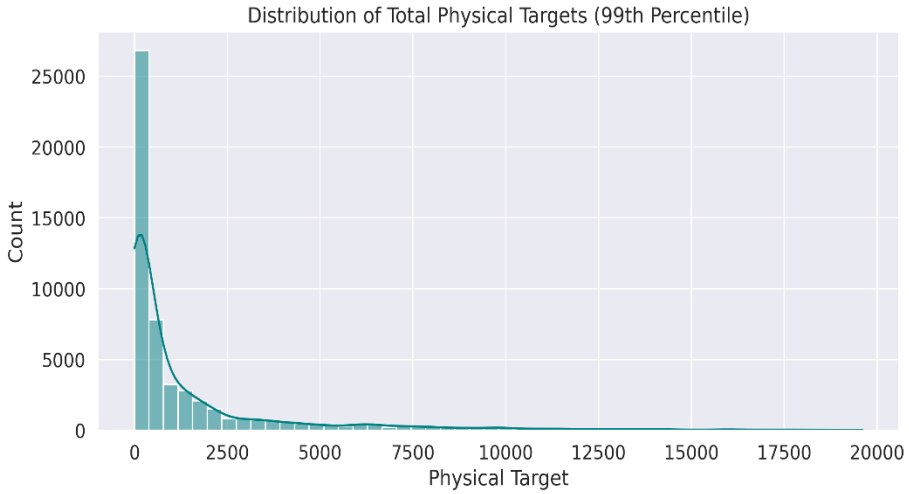
4.6 Top 10 States by Average Percent Achievement

- This visualization provides a comparative performance metric across states, showcasing those with the highest average percentage of target achievement. Gujarat, Andhra Pradesh, and Tamil Nadu demonstrating superior average achievements suggests that these states have been particularly effective in translating planned targets into realized outcomes. Analyzing the operational strategies and environmental factors in these high-performing states could yield valuable best practices transferable to other regions.



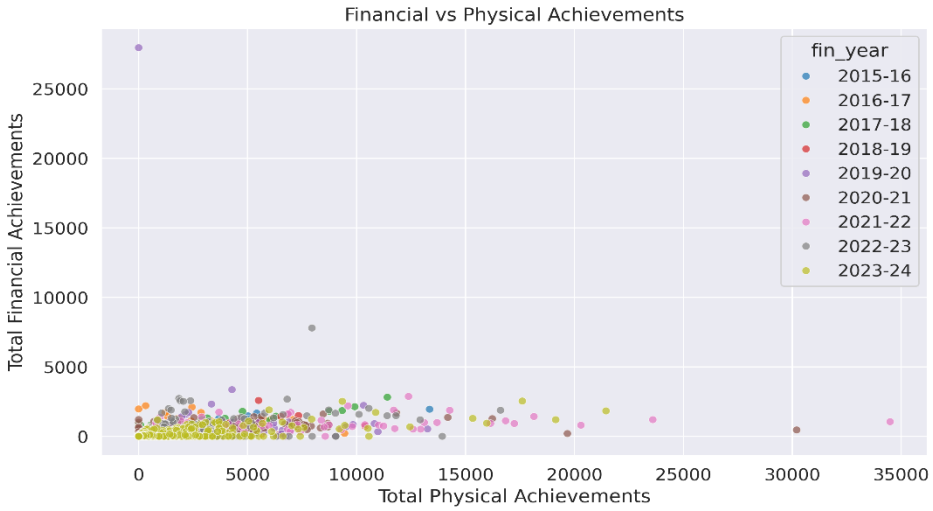
4.7 Distribution of Total Physical Targets

- The histogram illustrating the 'total_phy_target' column reveals a skewed distribution, mirroring the pattern observed in financial targets. Most physical targets are concentrated at the lower end, indicating that a large proportion of projects have relatively modest physical goals. Conversely, a long tail extending towards much higher values signifies the existence of a few exceptionally large or ambitious projects. This distribution helps in understanding the scale and scope of physical undertakings within the scheme.



4.8 Financial vs. Physical Achievements

- A scatter plot, mapping 'Total Physical Achievements' against 'Total Financial Achievements' and color-coded by 'Financial Year', offers a holistic view of project outcomes. The plot generally shows a dense clustering of projects at lower values for both financial and physical achievements, suggesting that many initiatives yield modest results. However, the presence of distinct outlier points indicates specific instances of highly successful large-scale projects that achieved significant progress in both financial expenditure and physical completion. The varying dispersion and concentration of points across different financial years highlight the scheme's fluctuating performance and investment returns over time.



4.9 Percent Achievement by Financial Year

- This box plot visualizes the distribution of 'percnt_achievements' across different financial years. It allows for a quick comparison of the central tendency, spread, and presence of outliers in achievement percentages for each year. Observing the medians, interquartile ranges, and any extreme points for each financial year can reveal temporal variations in performance consistency and identify years with particularly high variability or a greater number of high-performing outliers. This helps in understanding the year-on-year consistency and variability of project outcomes.

5. Outline of Proposed Machine Learning Algorithms

Building upon the initial insights gained from the Exploratory Data Analysis, this section outlines potential machine learning tasks and suitable algorithms that could be applied to the PMKSY-MIPMS dataset. These proposals aim to leverage the data for predictive modeling, forecasting, or segmentation, thereby extracting deeper value and actionable insights.

5.1 Predicting Project Achievement (Regression Task)

- Goal: To predict the percnt_achievements for future projects or for ongoing projects based on their initial targets and other characteristics. This can help in identifying projects at risk of underperformance or those likely to exceed expectations.
- Features: fin_year, state_name, district_name, fin_target_drip, fin_target_sprinkler, total_fin_target, phy_target_drip, phy_target_sprinkler, total_phy_target. Categorical features (state_name, district_name, fin_year) would require appropriate encoding (e.g., One-Hot Encoding, Label Encoding).
- Proposed Algorithms:
    - Linear Regression: A foundational model to establish a baseline. It's simple and interpretable, useful for understanding direct linear relationships between targets and achievement percentages.
    - Random Forest Regressor: Highly effective for capturing non-linear relationships and interactions among features. It's robust to outliers and can provide insights into

feature importance, indicating which targets or locations are most influential in predicting achievement.
    - Gradient Boosting Regressors (e.g., XGBoost, LightGBM): Known for their high predictive accuracy and ability to handle complex patterns in tabular data. These models iteratively correct errors from previous models, leading to strong performance, especially for datasets with varied feature types.
    - Support Vector Regressor (SVR): Can be effective for complex, non-linear relationships, particularly when the number of features is high. It aims to find a function that deviates from the actual target by a maximum ϵ (epsilon) margin.

5.2 Forecasting Financial and Physical Targets/Achievements (Time Series Forecasting)

- Goal: To predict future total_fin_target, total_phy_target, total_fin_achievements, or total_phy_achievements at a state or national level. This can aid in future budgetary planning, resource allocation, and setting realistic goals for the scheme.
- Features: Historical values of the target variable, along with temporal features (e.g., month, quarter, year), and potentially other aggregated features from previous periods.
- Proposed Algorithms:
    - ARIMA/SARIMA (AutoRegressive Integrated Moving Average / Seasonal ARIMA): Traditional statistical models well-suited for time series data. They can capture autoregressive (dependency on past values), integrated (differencing to make series stationary), and moving average (dependency on past forecast errors) components, with SARIMA extending to include seasonality.
    - Prophet (by Facebook): A robust forecasting library designed for business time series data. It handles trends, seasonality (daily, weekly, yearly), and holidays, making it user-friendly and effective for structured time series data.
    - Recurrent Neural Networks (RNNs) / LSTMs (Long Short-Term Memory networks): For more advanced and potentially accurate forecasting, especially if the time series exhibits complex, long-term dependencies. These deep learning models are capable of learning patterns over sequences of data.

5.3 Classification of Project Success/Failure (Classification Task)

- Goal: To classify projects into categories like 'Successful' or 'Unsuccessful' based on their percnt_achievements (e.g., defining 'Successful' as percnt_achievements above a certain threshold, e.g., 80%). This can help in early identification of projects needing intervention.
- Features: Same as for regression (targets, state, district, financial year), but the target variable percnt_achievements would be binarized or categorized.
- Proposed Algorithms:
    - Logistic Regression: A simple yet powerful linear model for binary classification, providing probabilities of success.
    - Decision Trees / Random Forest Classifier: Non-linear models that can capture complex decision boundaries. Random Forests are ensemble methods that improve accuracy and reduce overfitting.
    - Support Vector Machines (SVM): Effective for finding optimal hyperplanes to separate classes, especially in high-dimensional feature spaces.
    - Gradient Boosting Classifiers (e.g., XGBoost, LightGBM): High-performing ensemble methods that build models sequentially, correcting errors of previous models, often yielding excellent classification accuracy.

5.4 Clustering of States/Districts (Unsupervised Learning)

- Goal: To identify natural groupings or segments of states or districts based on their financial and physical targets and achievements. This can help in understanding regional similarities and differences, enabling targeted policy interventions or resource allocation strategies.
- Features: Aggregated or normalized metrics such as average total_fin_target, average total_phy_target, average percnt_achievements, no_of_offices, no_of_accounts per state/district.
- Proposed Algorithms:
    - K-Means Clustering: A widely used algorithm for partitioning data into a predefined number of clusters. It's effective for identifying distinct groups based on the similarity of their performance metrics.
    - DBSCAN (Density-Based Spatial Clustering of Applications with Noise): Useful for discovering clusters of varying shapes and sizes in a dataset and identifying outliers as noise. This could be beneficial if some regions behave very differently from others.
    - Hierarchical Clustering: Provides a hierarchical structure of clusters, which can be visualized as a dendrogram. This allows for exploring different levels of granularity in the segmentation of states or districts.

  These proposed models represent potential avenues for further analysis. Prior to implementation, extensive feature engineering (e.g., creating ratios, interaction terms, temporal features), hyperparameter tuning, and rigorous model evaluation (using appropriate metrics like R-squared, MAE, RMSE for regression; accuracy, precision, recall, F1-score for classification; silhouette score for clustering) would be necessary to select the most appropriate and effective algorithm for each specific task.

6. Conclusion

This report provides a comprehensive initial understanding of the PMKSY-MIPMS dataset, revealing key trends in financial allocations, physical targets, and achievements across various states and districts. While many projects exhibit modest outcomes, there are instances of significant investments and high achievements. The analysis highlighted the strong impact of drip and sprinkler components on overall achievements and identified top-performing states and districts. These findings serve as a strong foundation for future in-depth analyses, strategic planning, and identifying areas for improvement within the micro-irrigation scheme.