Group Details

MBA/0042/61	MONAVARTHI SRI SAI AADARSH
MBA/0211/61	REDDI GNANESWAR NAIDU
MBA/0375/61	RAMAVATH BALAJI NAIK
MBA/0460/61	RAVOOR THARUN
MBA/0197/61	KATUKURI RAMAN

Business Context – Brief Background

India's agriculture sector relies on efficient use of land and water resources. Digital agriculture – combining data analytics with farming – is increasingly used to improve productivity and resource use while ensuring equity. In particular, understanding the distribution of farm holdings across different social groups (Scheduled Castes, Scheduled Tribes, Others, Institutional owners) and land size classes helps policymakers identify regions or communities with poor irrigation access. For example, data-driven irrigation management can optimize water use and increase crop yields, as studies note that applying data mining in farming "allows accurate management" of resources such as irrigation . In this context, the government can use analytics on this dataset to support marginalized farmers, plan irrigation projects, and improve agricultural outcomes

Business Objective

The goal of this project is to assist the Ministry of Agriculture in formulating targeted policies for irrigation and land development. By analyzing the holdings data by social group and land size, the government can identify disparities in irrigation coverage and allocate resources more equitably. Specifically, the project aims to pinpoint which farmer groups or districts lack irrigation infrastructure and to support interventions (such as building irrigation facilities or providing subsidies) where they are most needed. This aligns with national objectives to improve water use efficiency and support small/marginal farmers – data analysis will give evidence to "help make better decisions" about where to invest in irrigation and agricultural support

Analytics Objectives

We plan to use data mining methods to extract actionable insights from the holdings data. First, we will characterize the dataset: calculate distributions and summary statistics of holdings by social group, land size, and irrigation status. Second, we will segment and compare groups – for example, analyzing how the proportion of irrigated holdings varies among SC/ST versus Others, or how irrigation rates change with farm size. Third, we will identify patterns and clusters among districts or states based on irrigation coverage. Finally, we may build simple predictive models (e.g. decision trees or regression) to understand which factors (e.g. land size, social category) are most strongly associated

with irrigation levels. Overall, the analytics objective is to reveal trends, patterns, or disparities in irrigation and land use that inform government action

Specific Questions for Data Mining

How are operational holdings distributed by social group and land-size? For example, what share of SC/ST holdings are marginal vs. medium or large, and how does this compare to Others?
What is the irrigation status across groups?
Which social groups or size-classes have the highest proportion of wholly irrigated, partially irrigated, or unirrigated holdings?
Are disadvantaged groups (SC/ST) under-irrigated compared to Others? Geographic patterns: Which states or districts have the largest number of irrigated vs. unirrigated holdings, and can we cluster districts by similar irrigation profiles?
Are there regional pockets of low irrigation coverage? Correlation analysis: Is there a relationship between farm-size class and irrigation area? (For instance, do larger farms tend to have more irrigated area?) Predictive classification: Can we predict whether a holding receives irrigation (or how much irrigated area it has) based on its social category and size class? (Decision-tree or other classifiers can reveal the key factors.) These questions will guide the analysis and ensure the results address policy-relevant issues (e.g. equity and resource optimization)

Overview of the Dataset – Source, Records, Fields

The dataset comes from India's 10th Agricultural Census (2015–16) and is available via the National Data and Analytics Platform (NDAP). It was compiled by the Department of Agriculture, Cooperation and Farmers Welfare (DACFW) . The table provided is "Social Category wise Estimated Number of Operational Holdings by Size Classes and Irrigation Status" at the district level. It contains about 27,000 records with 16 fields (columns). Key columns include: State and District identifiers; Year; Social Group Type (Institutional, Others, Scheduled Caste, Scheduled Tribe); Land Area Size (categories like "Marginal (0.5 ha–1.0 ha)", "SemiMedium", etc.); Category of Holdings (Marginal, Small, Medium, Semi-Medium, Large); and irrigation breakdowns such as "Wholly Irrigated Holdings (Number)", "Wholly Unirrigated Holdings (Number)", "Partially Irrigated Holdings (Number)", along with corresponding irrigated and total areas in hectares. In other words, for each district/social-group/size-class, the data gives the counts and areas of holdings classified by their irrigation status. According to official metadata, this dataset "contains overview details of farm holdings in India" with fields for number and area irrigated . (Our loaded CSV shows 26,960 rows and 16 columns covering all states and districts of India.)

Planned Techniques and Rationale

- **Exploratory Data Analysis (EDA) and Visualization:** We will begin with bar charts, heatmaps, and maps to summarize holdings by category and irrigation. This will reveal major trends and outliers at a glance.
  Clustering (Unsupervised Learning): We will cluster districts or social categories by their irrigation profiles. Clustering can group together areas with similar irrigation deficits or surpluses, which is useful for regional policy planning. For example, k-means or hierarchical clustering on features like fraction of irrigated holdings could highlight high-risk clusters. According to reviews, classification and clustering are the two main categories of data mining used in agriculture analytics
  Classification (Decision Trees, Random Forest): We may treat "well-irrigated vs poorly-irrigated" as classes and use attributes like Social Group and Size Class as predictors. Decision trees can identify which factors (e.g. being "Large" vs "Marginal" landholding) most strongly determine irrigation status . This helps interpret the data.
  Correlation and Regression Analysis: We will compute correlations (e.g. Spearman/Pearson) between area size and irrigated area. Regression models might quantify how much irrigation area is expected for a given holding count.
  Statistical Tests: To validate findings, we could use hypothesis tests (e.g. chi-squared) to see if irrigation rates differ significantly between social groups. Data mining techniques are valuable here because they can "optimize the use of resources" in agriculture . In fact, digital agriculture often uses data mining to improve productivity while conserving resources . Our chosen methods (clustering, classification) are explicitly cited as the main used categories in agridata analysis , and they will help uncover non-obvious patterns in the dataset

How Results Will Help Solve the Business Problem

The analytics outcomes will directly inform government decision-making on irrigation and agriculture support. By identifying where and for whom irrigation is lacking, officials can prioritize infrastructure investments (e.g. canals, wells, drip irrigation subsidies) in those districts or among those social groups. For instance, if SC/ST marginal farmers in a region are found to have very low irrigation, targeted programs can be implemented. Clear visual and statistical evidence of disparities will also help monitor progress of government schemes. In line with the goals of digital agriculture, these insights give stakeholders "explicit information" to base decisions on . Ultimately, this project will translate raw census data into actionable knowledge—optimizing resource allocation and supporting smallholder farmers—helping the government achieve more equitable and efficient agricultural development