Pakistan Sub-Division Population Analysis & Prediction

1. Problem Definition

The purpose of this project is to analyze the population distribution across Pakistan's sub-divisions and predict whether a given sub-division is urban-dominated or rural-dominated. This can support government planning, resource allocation, and development projects by identifying urbanizing areas.

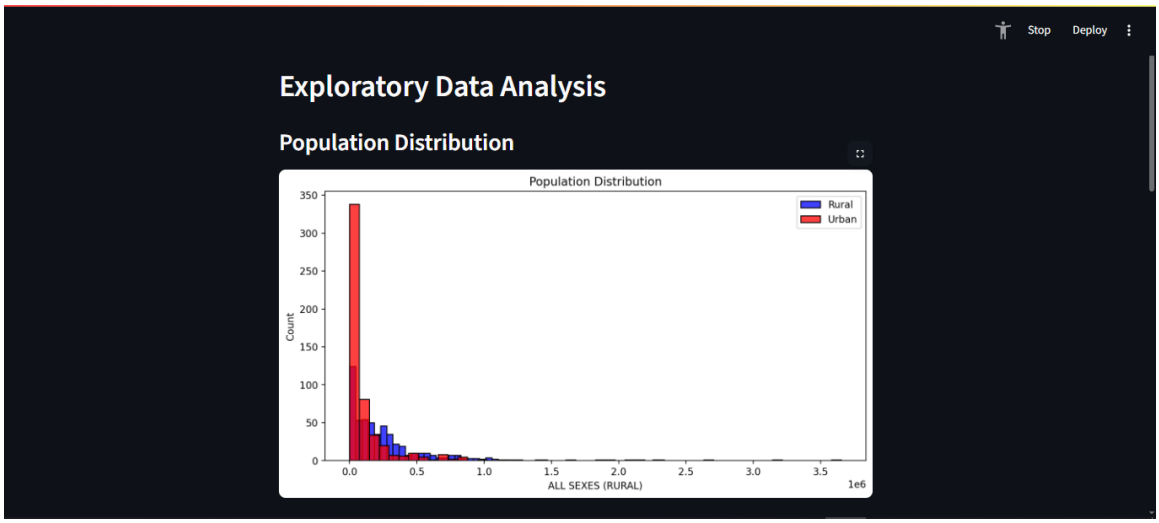
2. Data Preprocessing

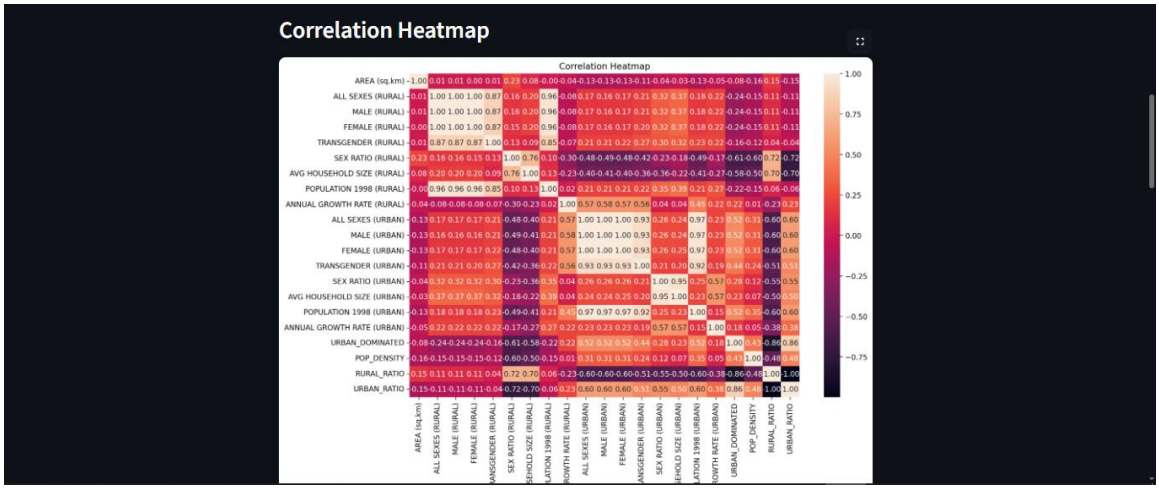
The dataset contains demographic and geographic attributes of sub-divisions in Pakistan, including population figures for rural and urban areas, area size, sex ratio, and household sizes. The following steps were taken during preprocessing:

- Removed leading/trailing spaces from column names.
- Replaced missing values with 0.
- Engineered new features:
  * URBAN_DOMINATED: Binary flag if urban > rural population.
  * POP_DENSITY: Total population divided by area.
  * RURAL_RATIO and URBAN_RATIO: Proportional values of population distribution.

3. Exploratory Data Analysis (EDA)

Visual analysis was performed using histograms and heatmaps. Histograms showed the distributions of rural and urban populations, while correlation heatmaps highlighted relationships between variables such as area, population, sex ratios, and derived ratios.





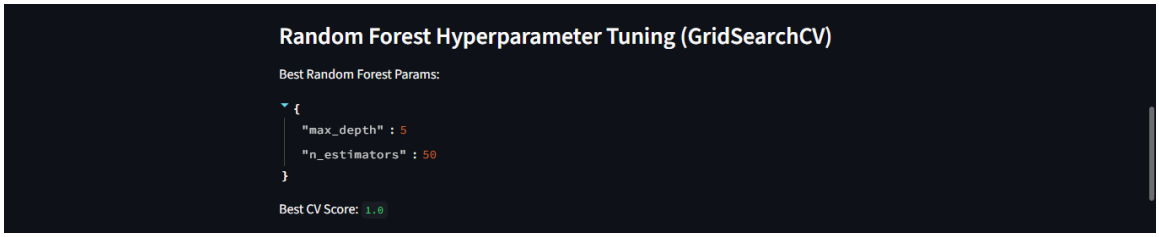
4. Model Building

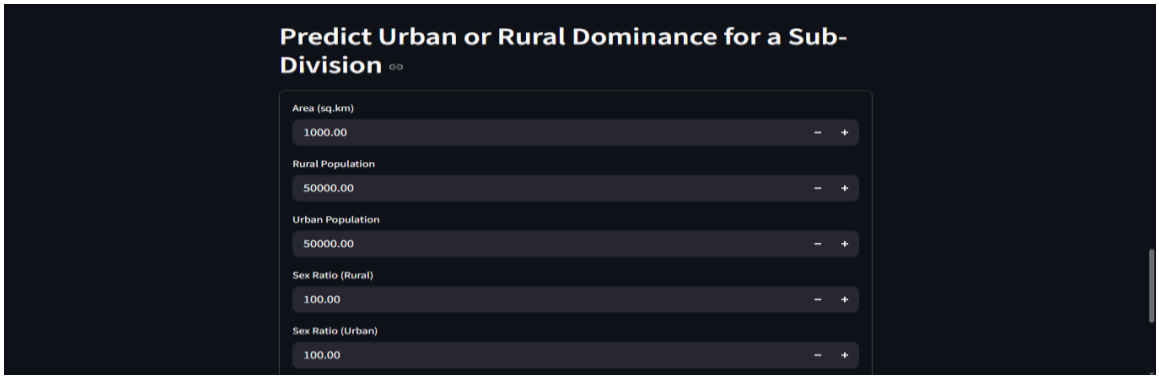
Two models were trained to predict whether a sub-division is urban or rural dominated:
- Logistic Regression: A linear model used on standardized data.
- Random Forest Classifier: An ensemble method trained on original feature values.

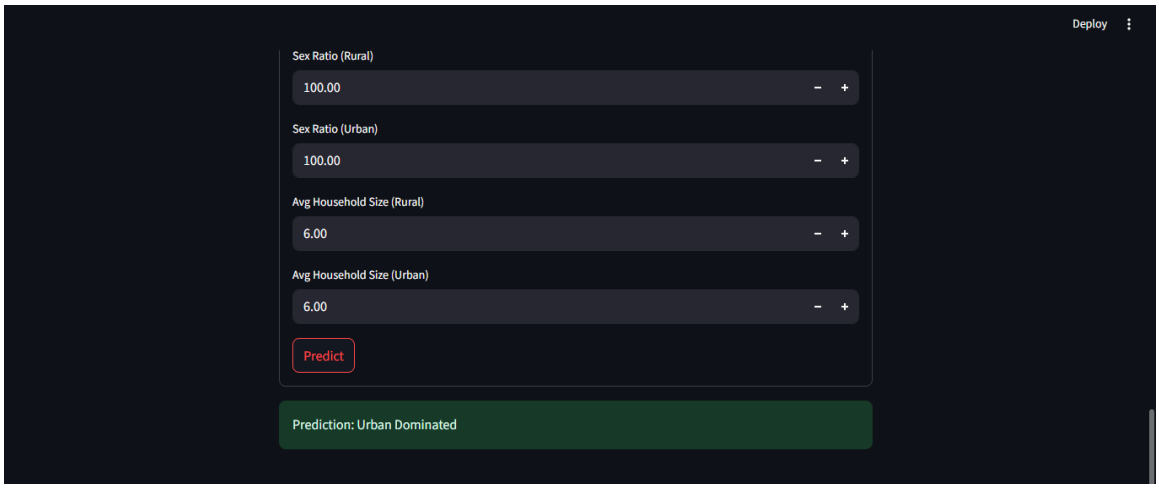
Features used:
- Area, Rural and Urban Populations
- Sex Ratios (Rural & Urban)
- Avg Household Sizes (Rural & Urban)
- POP_DENSITY, RURAL_RATIO, URBAN_RATIO

5. Model Evaluation

Classification reports were generated for both models, including precision, recall, and F1-score. Additionally, GridSearchCV was used to tune hyperparameters for the Random Forest model, optimizing for accuracy. Best parameters and cross-validation score were displayed in the Streamlit app.







6. Deployment with Streamlit

A user-friendly Streamlit web app was developed, allowing users to input sub-division statistics and receive a prediction on whether the region is urban or rural dominated. The app includes data visualization, model comparison, and real-time prediction.