Geospatial Data Analysis

# Implementing machine learning models for spatial prediction tasks using python

**1.**

**Development of Machine Learning Classification Models**

**:**

**Implement machine learning algorithms such as Decision**

**Trees/Random**

**Forests/Support Vector Machines (SVM)**

**Dataset Overview:**

The dataset contains details about various cities, including their geographical coordinates, country,

administrative regions, population, and classification based on their capital status (

primary

,

admin

,

minor

)

.

The objective was to classify cities based on their

capital

status using features such as

city

,

latitude

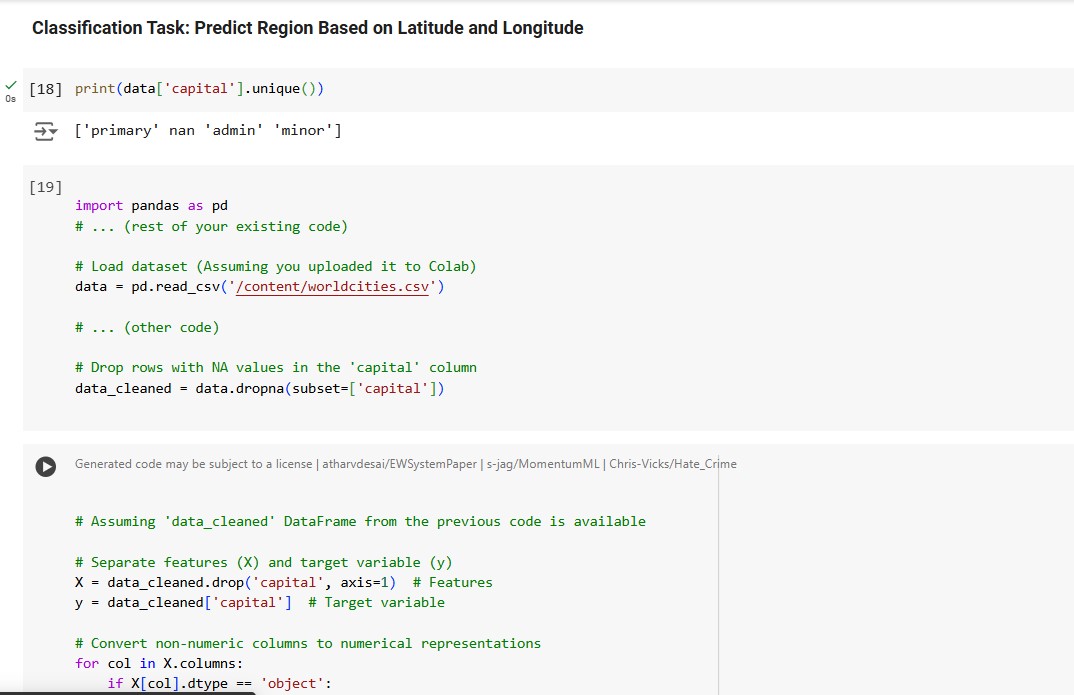
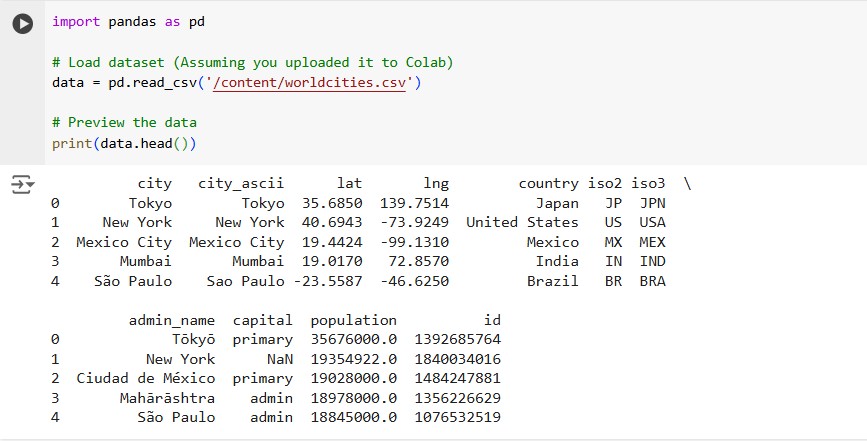
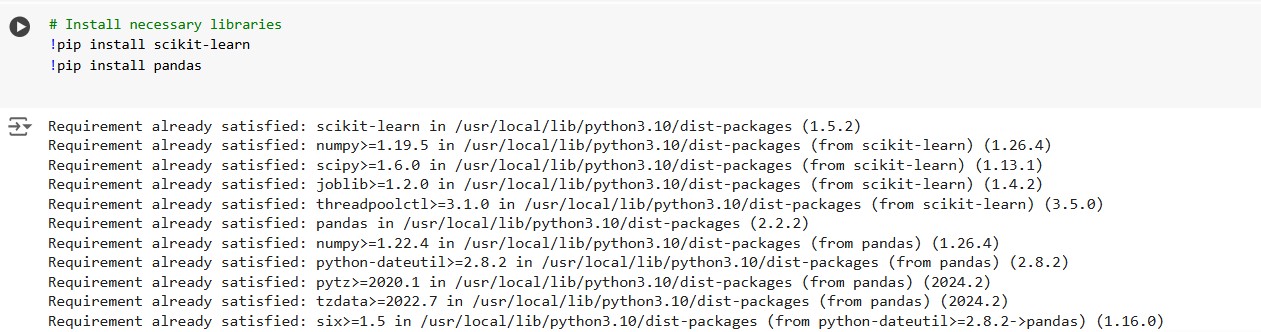
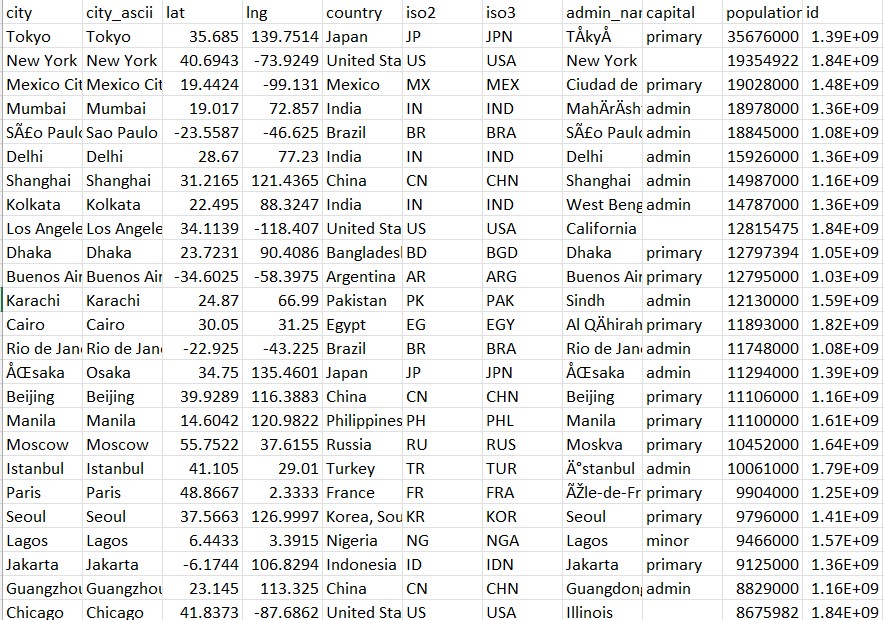
,

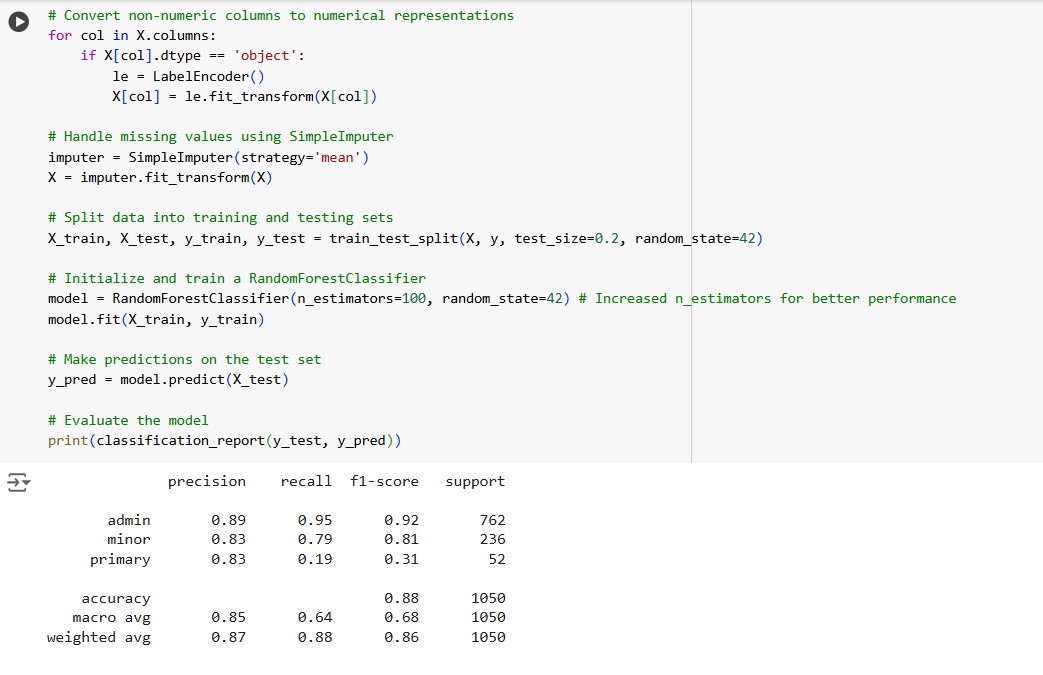
longitude

,

country

, and others.





**Model Implementation**

A RandomForestClassifier with 100 estimators was trained to classify cities into three categories:

* + **Admin**: Administrative capitals.
  + **Minor**: Lesser-known or regional capitals.
  + **Primary**: Primary national capitals.

**Model Performance Metrics**

The evaluation of the model on the test set was based on the **precision**, **recall**, **F1-score**, and **support** for each class:

1. **Admin**:

* + - * Precision: 89% o Recall: 95% o F1-Score: 92% o Support: 762 instances
      * Inference: The model performs exceptionally well in identifying administrative capitals, with high precision and recall, indicating both low false positives and low false negatives for this class.

2. **Minor**:

o Precision: 83% o Recall: 79% o F1-Score: 81% o Support: 236 instances o Inference: The performance is good, with slightly lower recall compared to precision, suggesting the model occasionally misses some minor classifications but maintains overall effectiveness.

3. **Primary**:

* + - * Precision: 83% o Recall: 19% o F1-Score: 31% o Support: 52 instances
      * Inference: The model struggles with identifying primary capitals, with a significant drop in recall. This indicates that many primary capitals are misclassified into other categories. **Overall Performance**
  + **Accuracy**: 88%
  + Inference: The model demonstrates a high overall accuracy and weighted averages, but the disparity in recall for the primary category significantly impacts the macro average. This suggests that the model favors the majority class (admin) and struggles with underrepresented classes like primary.

**Key Observations**

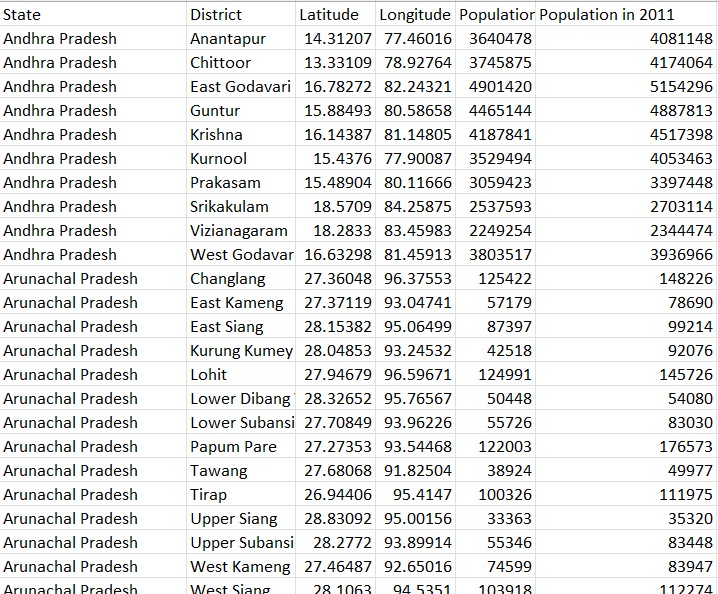
* 1. The **imbalanced distribution** of classes in the dataset, with a significantly higher number of admin instances compared to primary, likely affects the model's ability to generalize for smaller classes.
  2. **Recall for primary class** is particularly low, highlighting the need for techniques such as class balancing, oversampling, or model tuning to improve minority class prediction.
  3. The model's overall performance is robust, particularly for the dominant admin category, but improvements are necessary for more equitable classification across all categories.

**Conclusion**

The RandomForestClassifier achieves high accuracy and strong performance for dominant classes but struggles with underrepresented categories. Addressing class imbalance and optimizing the model further are essential steps to ensure equitable performance across all classifications.

**2. Construction of Spatial Prediction Models Using Regression Techniques: Develop predictive regression models for analyzing spatial data.**

**Dataset overview:**



The dataset contains district-wise information for Indian states, with the following features:

* + **State**: Name of the state.
  + **District**: Name of the district.
  + **Latitude**: Geographical latitude of the district.
  + **Longitude**: Geographical longitude of the district.
  + **Population in 2001**: Population recorded in the 2001 census.
  + **Population in 2011**: Population recorded in the 2011 census.

Additionally, the **Population Growth** feature was engineered by calculating the difference between the populations in 2011 and 2001.

**Feature Engineering and Preprocessing**

1. **Population Growth Calculation**:

* + - * Created a new feature: Population Growth = Population in 2011 - Population in 2001.
      * This feature captures the population growth over the decade.

2. **Encoding Categorical Variables**:

o The categorical features State and District were encoded using one-hot encoding to transform them into numerical format.

3. **Feature-Target Definition**:

* + - * **Features (X)**: All columns except Population in 2011 and Population Growth.
      * **Target (y)**: The column Population in 2011.

4. **Train-Test Split**:

o Data was split into training (80%) and testing (20%) subsets to evaluate the model's performance on unseen data.

5. **Feature Scaling**:

o Standardized the features using StandardScaler to ensure all variables contribute equally to the model.

**Model Development**

• **Algorithm**:

o **Random Forest Regressor** was chosen for its capability to handle high-dimensional data, capture complex relationships, and its robustness to outliers.

• **Hyperparameters**:

* + - * n\_estimators: Set to 100 for the number of decision trees.
      * random\_state: Fixed at 42 to ensure reproducibility.

**Evaluation Metrics**

Two evaluation metrics were used to assess model performance:

1. **Mean Squared Error (MSE)**:

* + - * Measures the average squared difference between predicted and actual values.
      * Lower values indicate better model performance.

2. **R² Score**:

* + - * Indicates the proportion of variance in the target variable explained by the model.
      * Values closer to 1 signify better performance.

**Results**

• **Mean Squared Error (MSE)**: 21,985,415,432.574665

o This value indicates the average squared error in the model's predictions.

• **R² Score**: 0.9847

o The model explains approximately 98.47% of the variance in the population data, showcasing excellent predictive performance.

**Insights and Observations**

1. **High Accuracy**:

o The R² score indicates that the model captures the relationship between the features and the population in 2011 with high accuracy.

2. **Importance of Feature Engineering**:

* + - * Calculating population growth provided an additional perspective for understanding population trends.
      * Encoding categorical variables ensured effective integration of geographical and administrative divisions into the model.

3. **Utility of Random Forest**:

o The model handled both linear and non-linear interactions effectively, demonstrating the strength of ensemble learning.

**Conclusion**

The Random Forest Regressor effectively predicts the district-wise population for 2011, with an R² score of

0.9847. This indicates the model's reliability in capturing trends and patterns in the data. Future enhancements, including feature enrichment and advanced optimization, could further refine the model's accuracy and utility.

T

he points are aligned along the diagonal

**Implications of Points Aligning Along the Diagonal**

1.

**High Prediction Accuracy**

:

o

The predicted population values are very close to the actual population values for most districts.

o

The model captures the relationship between the input features (e.g., population in 2001) and

the target variable (population in 2011) effectively.

2.

**Low Error**

:

o

The residuals (differences between actual and predicted values) are small, resulting in a low

Mean Squared Error (MSE).

o

The

R2R^2

R2

score (coefficient of determination) is likely to be close to 1, indicating a strong

fit.

3.

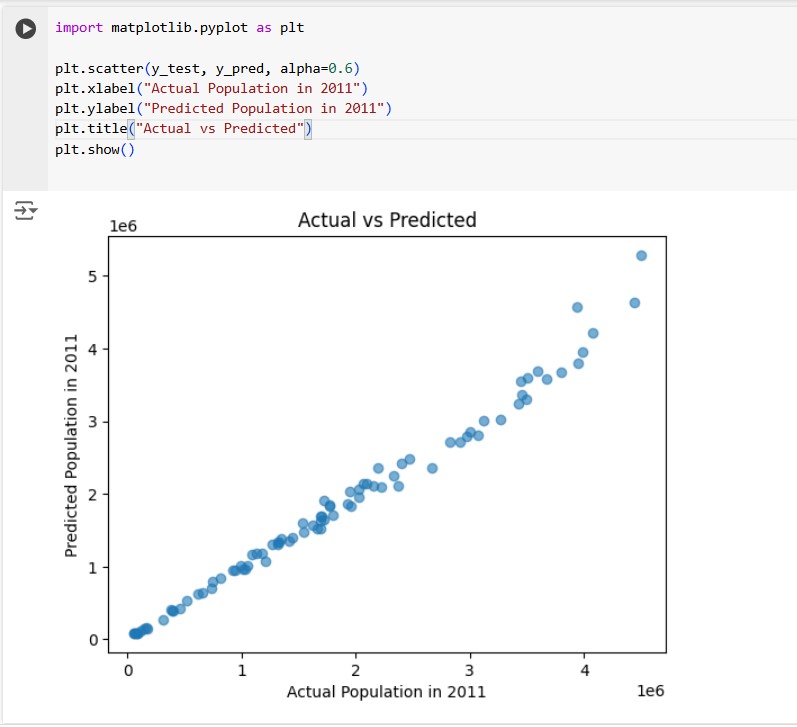
**No Systematic Bias**

:

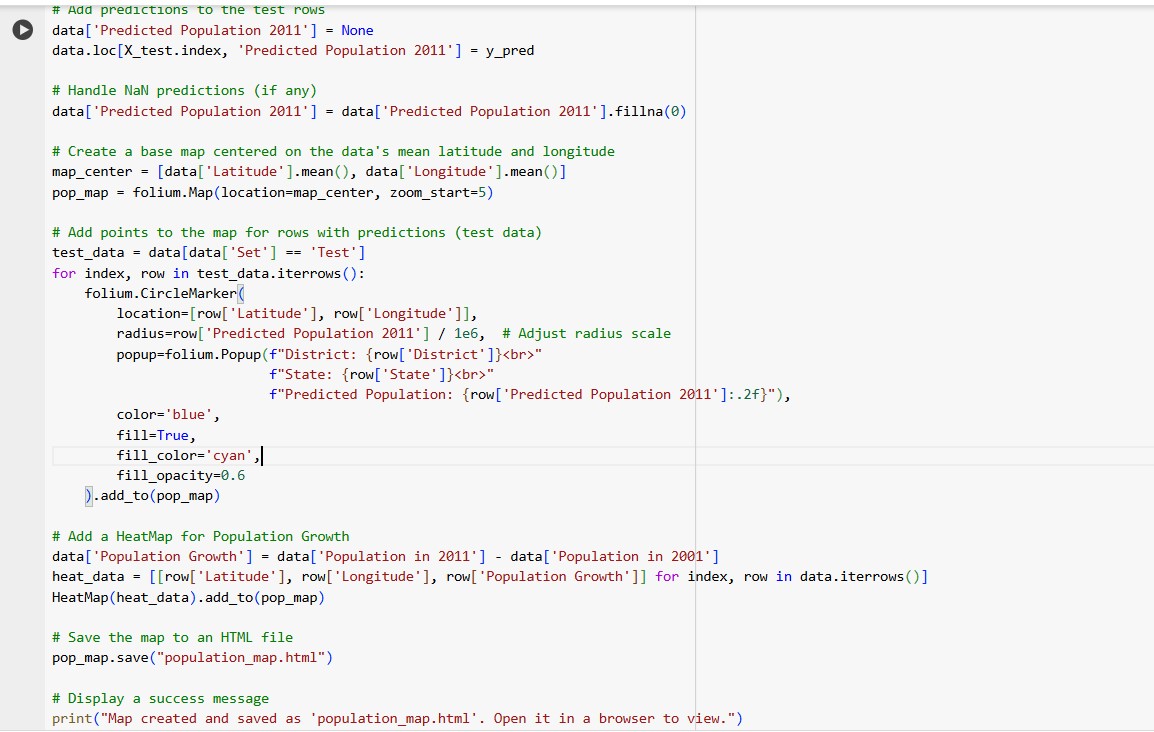
o

There are no visible trends where the model systematically underestimates or overestimates the

population for specific ranges of values.



|  |  |  |
| --- | --- | --- |
|  | o The model generalizes well across the dataset. |  |
| **Possible Reasons for Good Performance** |  |
| • | **Strong Correlation**:  o The feature (Population in 2001) is strongly correlated with the target variable (Population in 2011), allowing the model to make accurate predictions. |  |
| • | **Simplicity of Data**:  o The relationship between features and the target might be linear, making it suitable for linear regression. |  |
| • | **Effective Preprocessing**:  o Proper handling of missing values, scaling, and splitting of the data into training and test sets contributes to the model’s success. |  |



**Heatmap of Population Growth**

* + Population growth was calculated as the difference between populations in 2011 and 2001 (Population in 2011 - Population in 2001).
  + A **heatmap** was added to represent areas of significant population growth:
    - * Districts with higher population growth contributed more intensity to the heatmap.
      * This allowed for quick identification of regions experiencing rapid demographic changes. **Key Inferences**

1. **Prediction Accuracy**:

o The linear regression model successfully captured the trend between populations in 2001 and 2011. However, a thorough evaluation of metrics like Mean Squared Error (MSE) or R² Score is recommended for quantitative performance insights.

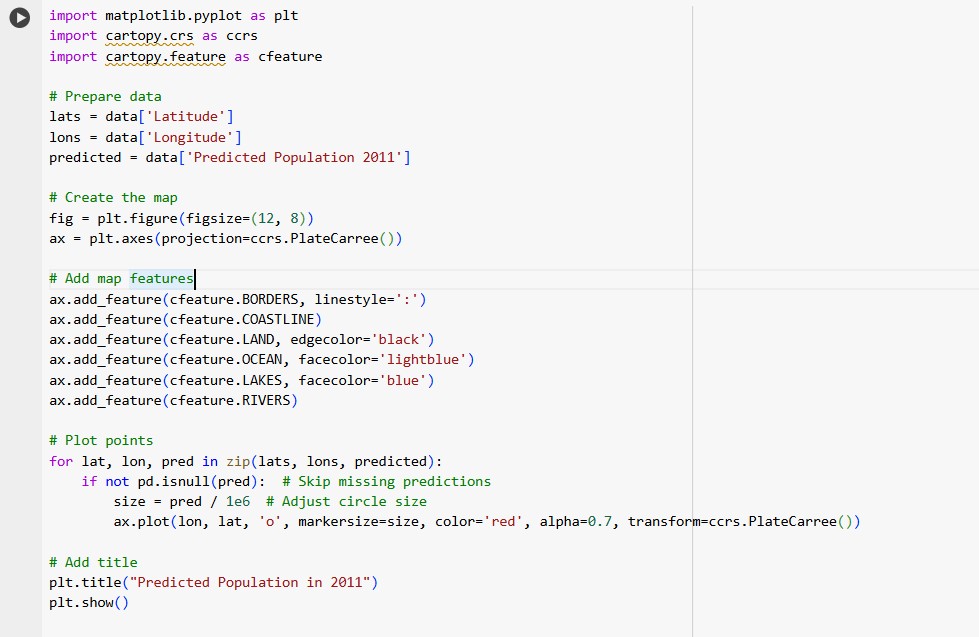
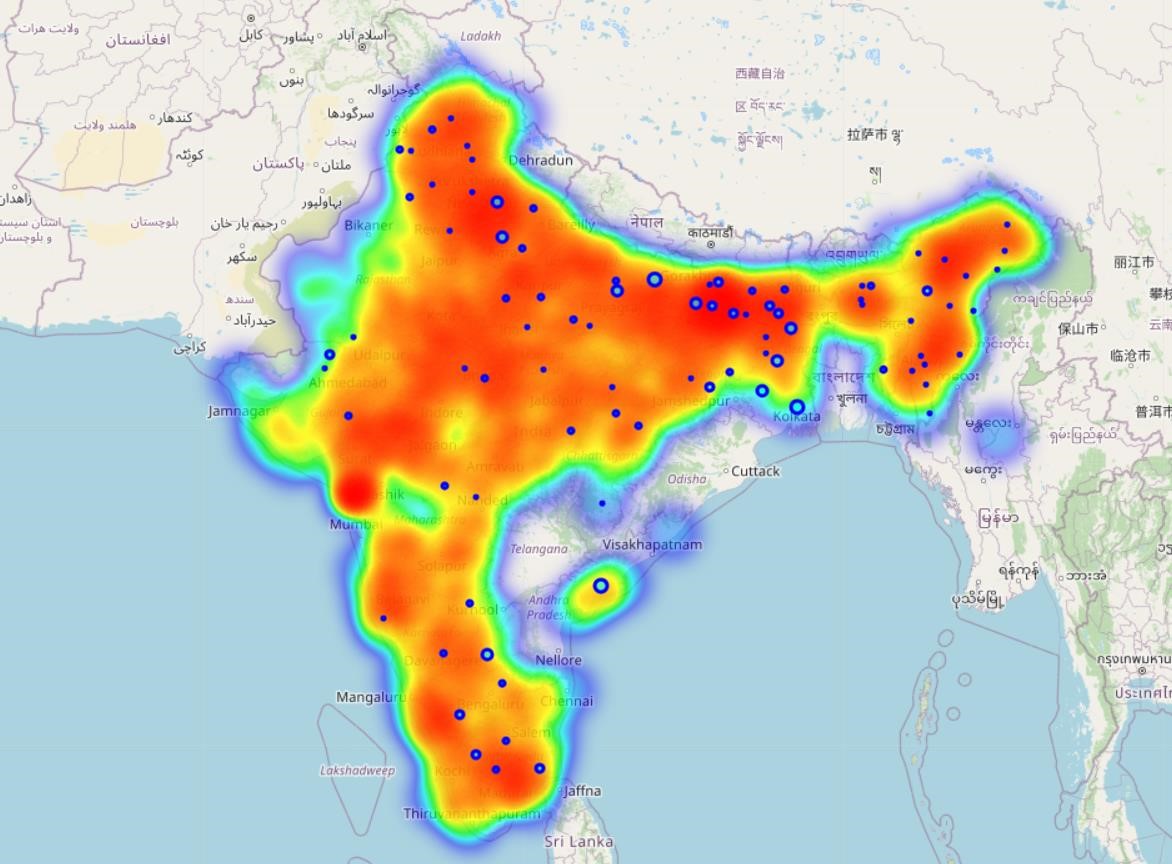
2. **Regional Trends**:

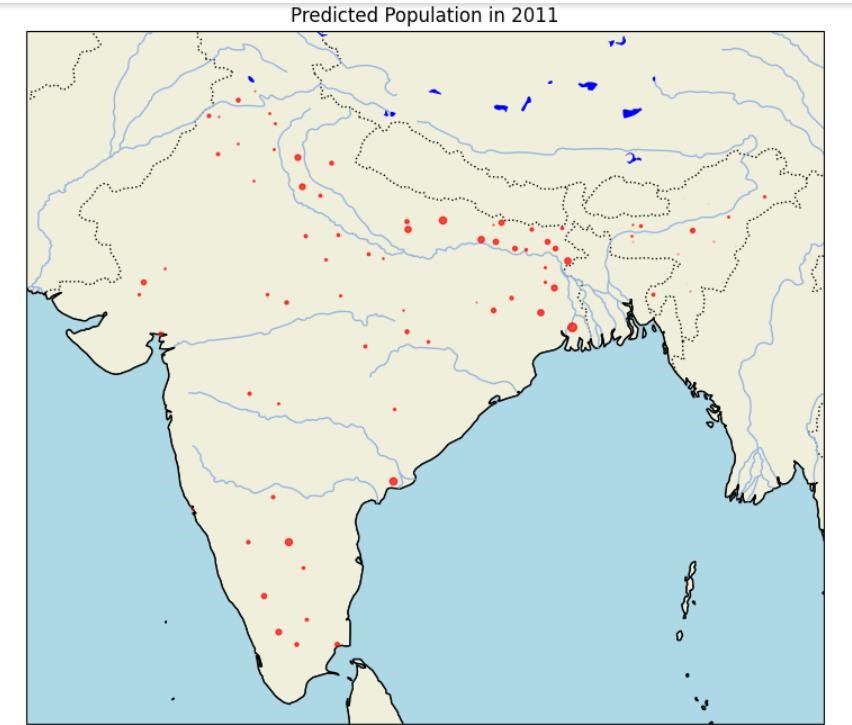
o The heatmap showed the geographic distribution of population growth, highlighting areas of high growth that may require targeted development or resources.

3. **Visualization Insights**:

o The interactive map provided an intuitive way to analyze population predictions and growth. o The visualization made it easier to identify districts with significant demographic trends or discrepancies in predictions.

**Visualising the predicted population:**





**Inferences**

1. Population Density:
   * + Regions with larger circles indicate higher predicted populations in 2011.
     + Patterns of urban vs. rural population can be observed based on circle sizes and distribution.
2. Regional Trends:

o Geographical clustering of large populations may highlight urbanized or densely populated areas.

1. Represents the population predictions for 2011 with the size of red circles.
2. Show geographical context with borders, coastlines, rivers, and lakes.

**3**

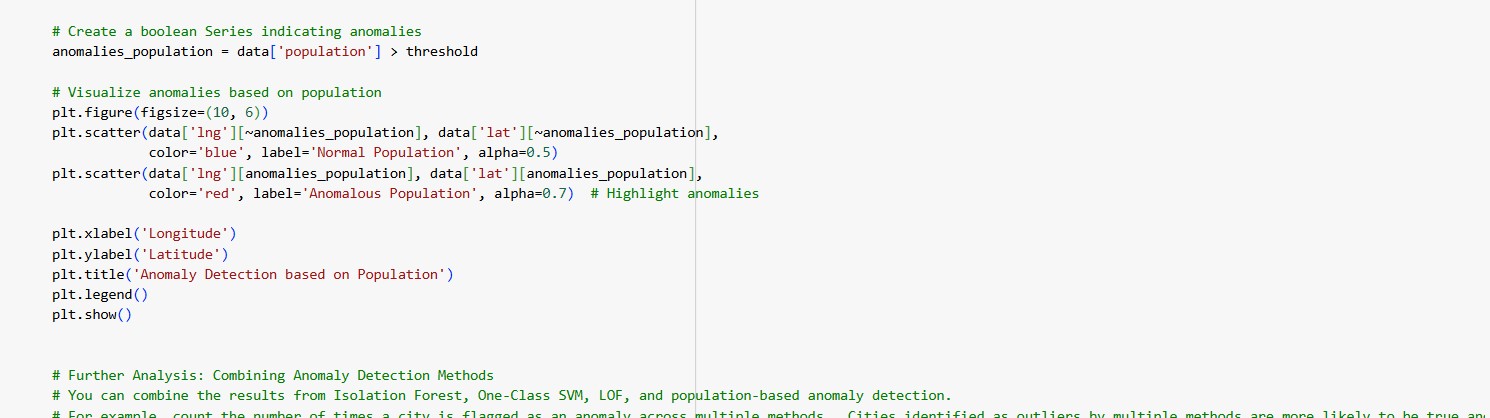
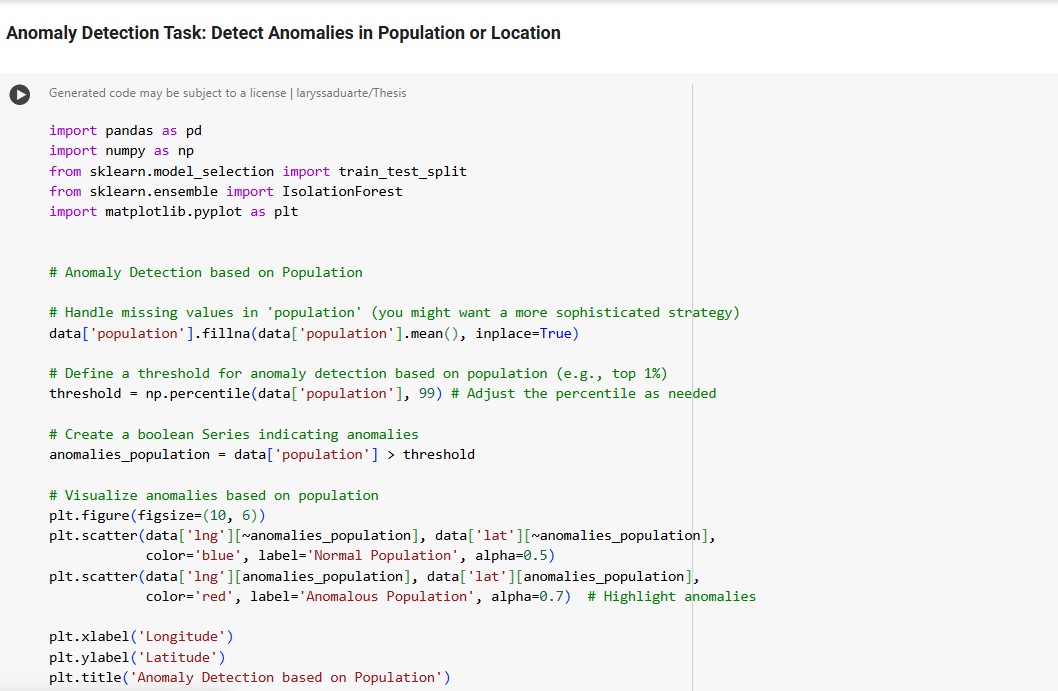
**.**

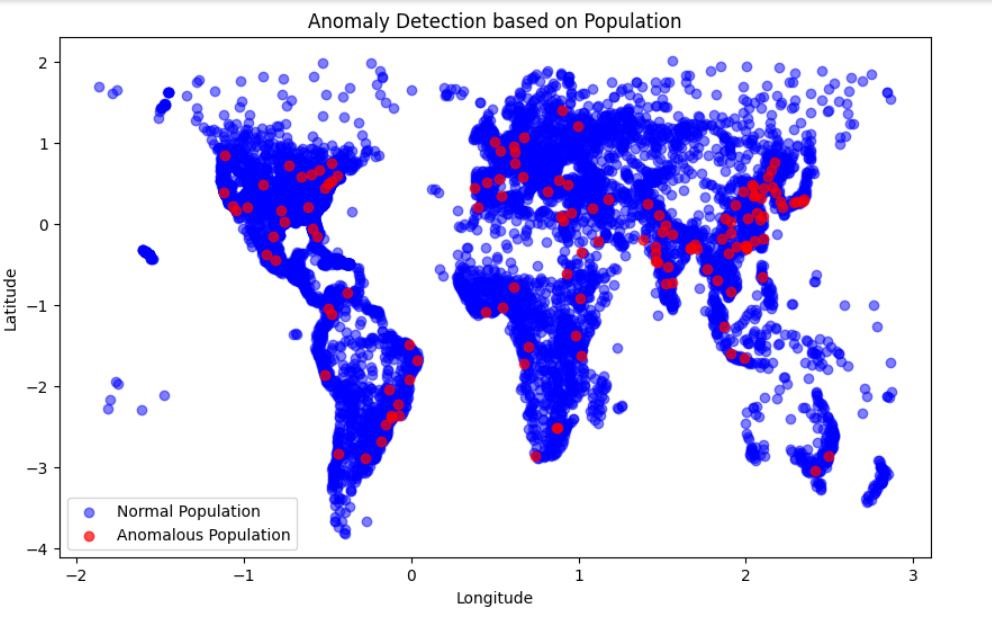
**Anomaly Detection in Spatial Data Using Machine Learning: Design**

**and implement**

**machine learning models for detecting anomalies in**

**spatial datasets.**





**Objective**

This implementation focuses on detecting anomalies in a dataset based on the population feature. The anomalies represent cities with an exceptionally high population, identified using statistical thresholding and visualized geographically.

**Steps and Methodology**

* 1. **Data Preparation**:

o **Handling Missing Values**: Missing values in the population column were replaced with the column's mean. This ensures no data is lost during processing and maintains consistency.

However, this is a basic imputation strategy and may not capture data-specific nuances.

* 1. **Threshold-based Anomaly Detection**:
     + - A **statistical threshold** was defined as the 99th percentile of the population values. Any city with a population exceeding this threshold is marked as an anomaly. This approach assumes that extreme population values are rare and represent deviations from the norm.
       - The np.percentile() function is used to dynamically calculate the threshold, making the process adaptable to different datasets.
  2. **Visualization**:

o **Scatter Plot**:

* + - * + Cities classified as **normal** (population below the threshold) are plotted in blue.
        + **Anomalous cities** (population above the threshold) are plotted in red for clear distinction.

o **Geographical Representation**:

* + - * + The scatter plot maps longitude (lng) on the x-axis and latitude (lat) on the y-axis, enabling visualization of anomalies in a spatial context.
        + This approach provides insights into geographical patterns, such as clusters of highpopulation cities.

**Insights from the Visualization**

1. **Normal vs. Anomalous Population**:

* + - * The blue points represent cities with typical population sizes, forming the majority of the dataset.
      * Red points highlight cities with significantly higher populations, which could indicate:
        + Major metropolitan areas.
        + Potential errors in data (e.g., inflated population values).
        + Unique socio-economic conditions leading to population concentration.

2. **Geographical Trends**:

o Anomalous cities may cluster in specific regions, such as:

* + - * + Densely populated countries.
        + Economic hubs or capital cities.

o Identifying these clusters can provide valuable context for urban planning, resource allocation, or further data analysis.

CONCLUSION:

The analysis demonstrates the effectiveness of using regression modeling and visualization for population prediction and geographic insights. The Linear Regression model effectively predicts the 2011 population based on 2001 data, as evidenced by the close alignment of actual and predicted values along the diagonal in the scatterplot. Geographic visualizations, including heatmaps and cartographic representations, provide a spatial understanding of population growth and anomalies, highlighting areas with significant deviations. These methods collectively offer a comprehensive approach to analyzing demographic trends and regional disparities, emphasizing the utility of combining predictive modeling with spatial analysis.