A logo of a company

Description automatically generated A close-up of a logo

Description automatically generated

**Joint Tech Internship Community Program**

**Assignment: Predicting Population Growth in Urban Areas**

**Problem Statement:**

A government agency wants to predict the population growth in urban areas to better plan infrastructure, services, and resources. Understanding population trends can help in urban planning and policy-making. Your task is to develop a machine learning model that predicts the population growth rate in urban areas based on various demographic and socio-economic factors.

**Objective:**

Build a predictive model to estimate the population growth rate in urban areas. Evaluate the model using appropriate metrics and provide insights into the factors that most influence population growth.

**Dataset:**

You are provided with a dataset containing the following columns:

1. **CityID**: Unique identifier for each city.
2. **CityName**: The name of the city.
3. **Year**: The year of the data point.
4. **CurrentPopulation**: The current population of the city.
5. **PopulationDensity**: The number of people per square kilometer in the city.
6. **BirthRate**: The number of births per 1,000 people in the city.
7. **DeathRate**: The number of deaths per 1,000 people in the city.
8. **ImmigrationRate**: The net number of people moving into the city per 1,000 people.
9. **UnemploymentRate**: The percentage of the labor force that is unemployed.
10. **AverageIncome**: The average annual income per person in the city (in USD).
11. **EducationLevel**: The average education level in the city (e.g., High School, Bachelor, Master).
12. **HealthcareAccess**: A score representing access to healthcare services in the city.
13. **CrimeRate**: The number of reported crimes per 1,000 people in the city.
14. **PopulationGrowthRate**: The target variable representing the annual population growth rate as a percentage.

**Tasks:**

1. **Data Exploration and Preprocessing:**
   * Load the dataset and perform initial exploration to understand the data.
   * Identify and handle any missing values appropriately.
   * Convert categorical variables into numerical ones using techniques such as One-Hot Encoding or Label Encoding.
2. **Feature Engineering:**
   * Perform feature scaling (e.g., Standardization or Normalization) on continuous variables.
   * Create new features if relevant, such as a combined score of socio-economic factors.
   * Use dimensionality reduction techniques (e.g., PCA) to reduce the feature space if necessary.
3. **Model Building:**
   * Split the dataset into training and testing sets (e.g., 80-20 split).
   * Train different regression models (e.g., Linear Regression, Decision Trees, Random Forest, Gradient Boosting).
   * Perform hyperparameter tuning using techniques like Grid Search or Random Search to optimize model performance.
4. **Model Evaluation:**
   * Evaluate your models using appropriate metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), and R-squared.
   * Compare the performance of different models and select the best one.
   * Analyze feature importance to understand the most significant factors contributing to population growth.
5. **Insights and Recommendations:**
   * Provide insights based on your model analysis regarding the factors that influence population growth.
   * Suggest actionable strategies for urban planners to manage population growth effectively.
6. **Documentation:**
   * Document your process, including data exploration, preprocessing steps, model selection, and evaluation.
   * Include visualizations where necessary to support your findings.

**Deliverables:**

* A Jupyter notebook (or Python script) with the entire workflow.
* A report summarizing your findings, including the model's performance and recommendations for urban planning.

This assignment problem focuses on predicting population growth in urban areas, a linear regression problem in the population statistics domain. The problem involves data preprocessing, feature engineering, model training, and evaluation, with a focus on understanding the factors that influence population growth and aiding in effective urban planning.