**Internship Project**: Analysis of Locality Performance Metrics Using Machine Learning and Statistical Techniques

**Objective**: To derive insights about the performance of various localities based on their attributes, and to identify key factors that influence business performance metrics. The goal is to determine what kind of localities perform well or poorly across different metrics and provide actionable recommendations to optimize business operations.

**Background**: The dataset contains locality-wise data, where each record represents a specific locality identified by its state, city, village ID, and geographic coordinates. Alongside these identifiers, the dataset includes various business performance metrics such as sales, margins, customer deliveries, and order counts. Additionally, the dataset contains a wide array of attributes describing the characteristics of each locality.

**Performance Metrics**:

* Sales Delivered Without Tax
* Sales Delivered With Tax
* Margin With GPR%
* Customers Delivered
* AOV (Average Order Value)
* Average Line Items Delivered
* Count of Unique Orders
* Count of Sales Invoices

**Attributes**: The dataset includes numerous attributes describing each locality, such as household characteristics, infrastructure quality, and asset ownership, among others.

**Key Tasks**:

1. **Data Preprocessing**:
   * Handle missing values and outliers.
   * Normalize and standardize numerical features.
2. **Descriptive Analysis**:
   * Summarize and visualize basic statistics of numerical features.
   * Analyze correlations between locality attributes and performance metrics.
3. **Exploratory Data Analysis (EDA)**:
   * Perform univariate and bivariate analysis to explore relationships between attributes and performance metrics.
   * Conduct segment analysis to group localities based on specific attributes and analyze their performance.
   * Use clustering techniques to identify homogeneous groups of localities.
4. **Feature Importance and Predictive Modelling**:
   * Identify key attributes influencing performance metrics using techniques like decision trees and random forests.
   * Build regression models to predict continuous performance metrics.
   * Build classification models to categorize localities into performance tiers (e.g., good, average, poor).
5. **Model Evaluation and Refinement**:
   * Evaluate models using appropriate metrics and cross-validation techniques.
   * Tune hyper-parameters and address model biases.
6. **Insights and Recommendations**:
   * Derive key insights and provide actionable recommendations to improve locality performance based on attributes.
7. **Reporting and Presentation**:
   * Prepare a comprehensive report summarizing findings and recommendations.
   * Present the findings to stakeholders.

**Expected Outcomes**:

* Identification of key attributes that significantly impact locality performance metrics.
* Classification of localities into performance categories based on their attributes.
* Actionable insights and recommendations to optimize business operations in different localities.

**Duration**: 7 Weeks

This project will leverage statistical analysis and machine learning techniques to provide a comprehensive understanding of the factors influencing business performance at the locality level, ultimately aiding in strategic decision-making and operational optimization.

**Link to Dataset**:<https://docs.google.com/spreadsheets/d/1QmsAUUwEB9Ay-et5GupbW4Xi1nlLkfTq/edit?usp=drive_link&ouid=100561164199308133372&rtpof=true&sd=true>