Project Overview: Income Prediction and Clustering of Canadian Census Tracts Using Machine Learning

1. Objective

The goal of this project is to build a data-driven pipeline that predicts median household income across Canadian census tracts using engineered features derived from housing and demographic census data. Additionally, the project leverages unsupervised learning (clustering) to segment geographic areas into groups with similar socioeconomic characteristics. These insights can guide affordable housing policy, urban planning, and resource allocation.

2. Data Source and Context

- Data Source: 2021 Canadian Census microdata at the census tract level.
- Files Used:
    - CensusCanada2021Training.csv for training
    - CensusCanada2021Test.csv for prediction
- Target Variable: Median Household Income (Current Year $)
- The dataset contains detailed housing, demographic, and tenure-related information.

3. Feature Engineering and Data Enrichment

To extract more informative signals from the raw data, several domain-specific composite features were created:

- Tenure Metrics:
    - Pct_Owner and Pct_Renter: proportions of owner-occupied vs. rental dwellings.
    - Ratio_of_Renters_to_Owner: captures rental dominance relative to ownership.
- Housing Age:
    - Pct_Older_House: proportion of homes built before 1980.
    - Pct_New_House: proportion of homes built after 2006.
    - Ratio_of_Olderhouse_to_Newhouse: indicator of aging housing infrastructure.
- Dwelling Structure Type:
    - Pct_Structure_Houses vs. Pct_Structure_Apartment: built environment type as a proxy for wealth or density.

- Demographics:
    - Household_Size: derived as population per household, indicative of living density.

These variables were standardized, rounded, and cleaned for modeling. Infinite values were converted to NaN and imputed with column means.

4. Exploratory Data Analysis and Feature Selection

Correlation and Visualization:

- A correlation matrix revealed linear associations between income and housing indicators.
- Scatterplots between features and the income target helped assess nonlinear patterns and outliers.

Decision Tree Regressor:

- A Decision Tree model was trained to assess feature importance.
- Features contributing to 90% of cumulative importance were selected.
- This dimensionality reduction enhanced model interpretability and computational efficiency.

5. Clustering Analysis: K-Means and BIRCH

K-Means Clustering:

- Applied to scaled features to segment census tracts.
- Optimal number of clusters (k=2) determined via elbow method and second derivative analysis.
- Resulting clusters revealed two distinct groups based on housing composition and demographic indicators.

BIRCH Clustering:

- Used as a secondary clustering method for cross-validation.
- Silhouette scores determined BIRCH optimal k.
- Comparison with K-Means via confusion matrix and Adjusted Rand Index showed consistent segmentation.

Cluster Profiling:

- Each cluster's mean, standard deviation, and size were calculated.
- Histograms and box plots visualized intra-cluster distributions.
- Key differentiators: ownership rate, age of housing stock, and household size.

6. Regression Modeling per Cluster

Separate supervised learning models were trained for each cluster:

- **Random Forest Regressor:**
    - Ensemble model handling non-linearities and interactions.
    - Performed best across clusters with lowest RMSE and MAE.
- **Decision Tree Regressor:**
    - Provided interpretability but underperformed in accuracy.

Performance was evaluated using:

- **Root Mean Squared Error (RMSE)**: penalizes large errors.
- **Mean Absolute Error (MAE)**: interpretable average prediction error.

Results were saved per cluster to guide region-specific model deployment.

7. Test Data Prediction and Deployment

- Test set was preprocessed using the same feature engineering pipeline.
- K-Means model assigned each census tract to a cluster.
- Features were scaled and input into the trained Random Forest model.
- Predicted income values were exported for submission.

8. Conclusion and Key Insights

This project demonstrates a robust end-to-end pipeline for:

- Predicting income at a granular regional level using housing and demographic signals.
- Clustering census tracts into actionable segments for policy intervention.
- Evaluating models within subpopulations to capture heterogeneous income drivers.

Key insights include:

- High renter concentration and older housing stock correlate with lower income.
- Household size and dwelling structure types are strong differentiators.
- Segment-specific modeling improves predictive performance and interpretability.