📊 Women Crime Analysis: Complete Data Science Report

🔹 1. Introduction

This report presents a comprehensive analysis of the **'womencrimes.csv'** dataset using the **complete data science pipeline**. The goal is to apply preprocessing, transformation, feature selection, modeling, and evaluation to build regression models that can predict:

1. Total number of persons under trial
2. Number of persons convicted

We compare several models and optimization strategies to determine the best predictive approach.

🔹 2. Data Cleaning and Preprocessing

- **Dataset Summary**:
  - Rows: 2765
  - Columns: 17
  - No missing values were detected.
- **Data Types**:
  - **Numerical Columns**: `Year`, `Persons_Acquitted`, `Persons_Arrested`, etc.
  - **Categorical Columns**: `Area_Name`, `Group_Name`, `Sub_Group_Name`
- **Steps Applied**:
  - Label Encoding on categorical columns using `LabelEncoder`

- Feature scaling of numerical columns using `StandardScaler`

🔹 3. Feature Engineering and Exploration

- Feature Importance was calculated using `RandomForestRegressor`.
- High correlation was observed between **'Persons_under_Trial_at_Year_beginning'** and **'Total_Persons_under_Trial'**.
- Categorical features contributed less to model performance and were considered for elimination during optimization.

🔹 4. Model 1: Predicting Total Persons Under Trial

🔸 Model Used: `RandomForestRegressor`

a. Performance with All Features:

- **MAE**: 0.0068
- **RMSE**: 0.0052
- **R^2 Score**: 0.9962

b. After Feature Selection (4 Important Features):

- **Features kept**:
  - `Total_Persons_under_Trial`
  - `Persons_under_Trial_at_Year_beginning`
  - `Persons_in_Custody_or_on_Bail_during_Trial_at_Year_End`
  - `Persons_in_Custody_or_on_Bail_during_Investigation_at_Year_end`

- **MAE**: 0.0059
- **RMSE**: 0.0037
- **R^2 Score**: 0.9973

c. Final Optimized Model (3 Features Only):

- **MAE**: 0.0054
- **RMSE**: 0.0033
- **R^2 Score**: 0.9976

✅ **Conclusion**: Simpler model, better accuracy, less noise.

🔹 5. Model 2: Predicting Number of Persons Convicted

🔸 Initial Model: `RandomForestRegressor`

a. Features Used:

- Arrest, Chargesheeted, Trial Progress, Under Trial, etc.

b. Initial Performance:

- **MAE**: 0.102
- **RMSE**: 0.380
- **R^2 Score**: 0.477

c. Dynamic Feature Selection:

- **Selected Features (9)**:
  - `Persons_Trial_Completed` (highest importance = 0.80)
  - `Persons_Released_or_Freed_before_Trial`

    - Trial and custody-related features

- **Final Performance**:
    - **MAE**: 0.101
    - **RMSE**: 0.355
    - **R^2 Score**: 0.512

⚠️ **Conclusion**: While performance improved slightly, R^2 remains low due to the weak correlation between inputs and convictions.

🔹 6. Model 3: Linear Regression with PCA

🔸 Goal: Predict `Total_Persons_under_Trial` using `LinearRegression` after dimensionality reduction

- Applied **PCA (95% Variance)**: Features reduced from 16 → 4
- Model: `LinearRegression`

Performance:

- **MAE**: 0.129
- **RMSE**: 0.423
- **R^2 Score**: 0.869

⚠️ **Conclusion**: PCA + Linear Regression underperformed compared to Random Forest, indicating non-linear relationships in the data.

🔹 7. Final Comparison of Models

Task	Model	R^2 Score	MAE	RMSE
Predict Total Under Trial	RandomForest (3 feat.)	**0.9976**	0.0054	0.0033
Predict Conviction	RandomForest (9 feat.)	0.512	0.101	0.355
Predict Under Trial (PCA + LR)	LinearRegression	0.869	0.129	0.423

🔹 8. Final Inference and Deployment Strategy

- ✅ **Best Performing Model**: `RandomForestRegressor` with feature selection for predicting total persons under trial.
- ✅ **Strengths**:
  - High accuracy
  - Low error
  - Interpretable through feature importance
- 🧠 **Saved Assets** for Deployment:
  - Final Model `.pkl`
  - Scaler, Encoder, and Important Features

🌟 Summary

- Data cleaning and preprocessing ensured no missing values.
- Feature engineering revealed core drivers of trial volume.
- Multiple models were compared; feature selection and optimization enhanced performance.
- Future Work: Use classification models or advanced boosting (XGBoost) to improve conviction prediction.