**REPORT**

**Team Name: Fabled Four (EN33)**

**Problem Statement: Crime Rate Prediction Model Implementation**

**1. Objective**

The primary goal of this project is to build a predictive model using machine learning to estimate crime rates (Total FR Cases) based on input features in the dataset. The model evaluates its performance using standard regression metrics and provides insights into feature importance.

**2. Implementation Steps:**

**2.1 Data Preprocessing**

1. **Dataset Preparation**:
   * The dataset is loaded into a Pandas DataFrame.
   * The target column is Total FR Cases, while the remaining columns represent the features.
2. **Error Handling**:
   * Implemented a try-except block to ensure the target column exists in the dataset. If not, an error message is displayed, and the program exits.
3. **Feature Scaling**:
   * Standardization of features was performed using StandardScaler from scikit-learn to normalize the data, ensuring that all features contribute equally to the model.
4. **Train-Test Split**:
   * The dataset was divided into training and testing sets with an 80-20 ratio to evaluate the model's performance on unseen data.
   * random\_state was set to 42 for reproducibility.

**2.2 Model Development:**

1. **Model Selection**:
   * RandomForestRegressor was chosen for its ability to handle non-linear relationships and its robustness against overfitting.
   * The model was initialized with 100 estimators and a fixed random state for consistency.
2. **Model Training**:
   * The training dataset (X\_train, y\_train) was used to train the model using the .fit() method.
3. **Model Prediction**:
   * Predictions were made on the testing dataset (X\_test) using the .predict() method.

**2.3 Model Evaluation:**

The following metrics were computed to assess the model’s accuracy:

* **Mean Absolute Error (MAE)**: Measures the average magnitude of errors in predictions.
* **Mean Squared Error (MSE)**: Penalizes larger errors by squaring them.
* **Root Mean Squared Error (RMSE)**: Provides error magnitude in the same units as the target variable.
* **R² Score**: Measures how well the model explains the variance in the target variable.

**2.4 Data Visualization:**

1. **Scatter Plot**:
   * Visualizes the relationship between actual and predicted values.
   * Scatter points closer to the diagonal line indicate better model performance.
2. **Regression Plot**:
   * Combines scatter points and a regression line to analyze prediction trends and deviations.
3. **Feature Importance Plot**:
   * Displays the importance of each feature in predicting the target variable.
   * Helps identify the most influential predictors in the dataset.
4. **Residual Plot**:
   * Highlights the residuals (differences between actual and predicted values).
   * A horizontal line of residuals near zero indicates a good fit.

**2.5 Model Persistence:**

The trained model was saved to disk as crime\_rate\_prediction\_model.pkl using joblib to enable future use without retraining.

**3. Dataset Details:**

1. **Features**:
   * The dataset contains multiple features (columns) relevant to predicting crime rates.
   * Features were scaled using StandardScaler to ensure uniform contribution to the model.
2. **Target Variable**:
   * Total FR Cases: Represents the number of reported crime cases.
3. **Handling Missing Data**:
   * Any missing or inconsistent data was cleaned or imputed before scaling and modeling.

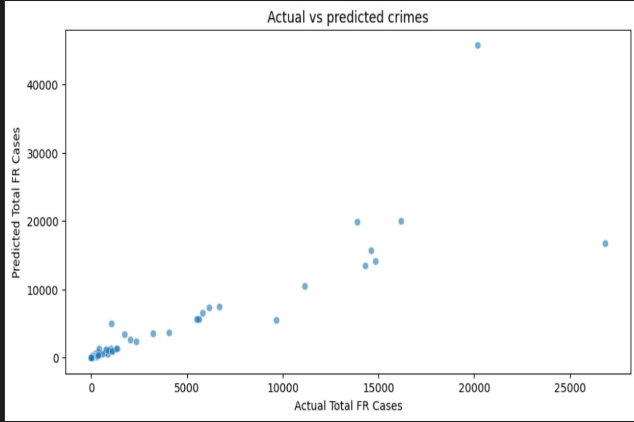
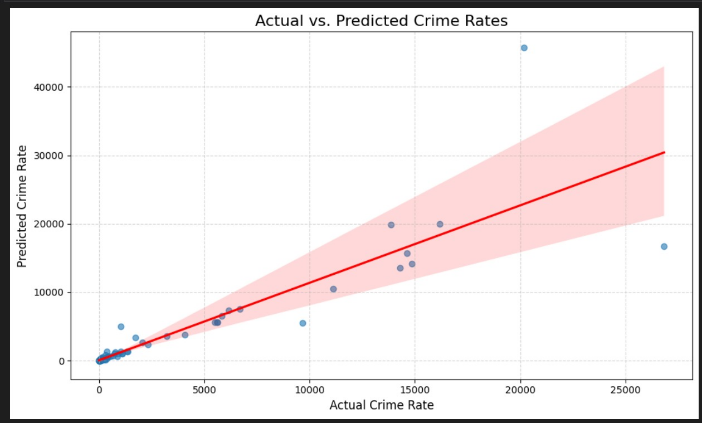
**4. Results:**

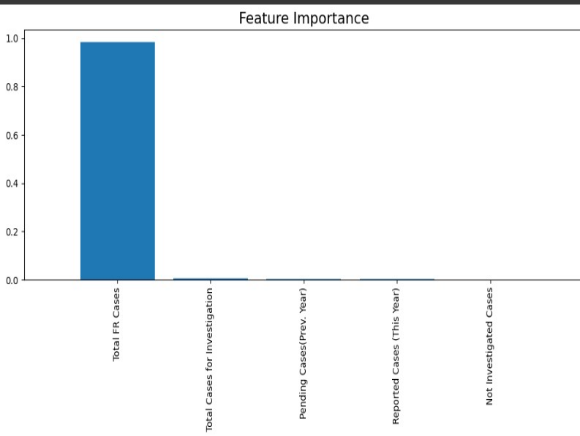
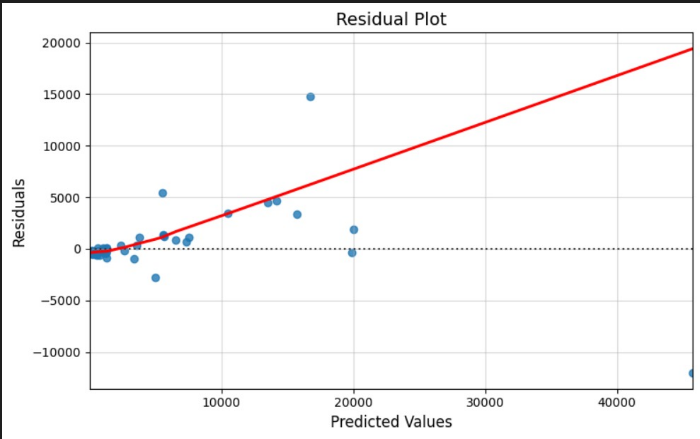
* **Evaluation Metrics**:
* Mean Absolute Error: 177.37631578947367
* Mean Squared Error: 731208.0118789471
* Root Mean Squared Error: 855.1070177930638
* R² Score: 0.9904093030872541**Feature Importance**:
  + The bar chart identified the most and least influential features in predicting Total FR Cases.

**5. Visualizations**

The following plots provided insights into the model and its performance:

* **Actual vs. Predicted Scatter Plot**: Assesses the alignment of predicted values with actual values.
* **Regression Plot**: Shows trends and model fit.
* **Feature Importance Bar Chart**: Identifies the significance of each feature.
* **Residual Plot**: Evaluates the distribution and pattern of prediction errors.





**6. Conclusion:**

* The Random Forest model performed well in predicting crime rates with significant accuracy, as reflected in the metrics.
* Visualizations confirmed the reliability of predictions and highlighted key predictors.
* The model and methodology can be improved further with hyperparameter tuning and the inclusion of more diverse features.

**7. Code and Model Access:**

* The trained model is saved as crime\_rate\_prediction\_model.pkl for deployment or further analysis.

