Violence Against Womens and Girls Data Using Random Forest

# Table of Contents

|  |  |  |
| --- | --- | --- |
|  | Table of Contents | 4 |
| 1 | Abstract | 5 |
| 2.1  2.2 | Introduction  Objectives | 6-7 |
| 3 | Literature Review | 8 |
| 4 | System Design | 9-10 |
| 5 | Implementation | 11-12 |
| 6 | Results and Output Screenshots | 13-16 |
| 7 | Challenges and Learnings | 17 |
| 8 | Conclusion | 18 |
| 9 | References | 19 |

1. Abstract

Violence against women and girls (VAWG) continues to be a critical issue across the globe, affecting individuals of all ages, backgrounds, and regions. With the advancement of technology and data science, we now have the tools to analyze this issue in depth and predict occurrences using machine learning. This project uses a real-world dataset related to incidents of VAWG and applies the Random Forest algorithm to identify patterns and predict possible future cases.

Our aim is to utilize data to assist in understanding the factors contributing to VAWG and empower decision-makers with predictive capabilities. We began by cleaning and exploring the dataset to detect patterns related to the types, locations, and frequency of violence. We then visualized the findings using Python libraries like Matplotlib and Seaborn. Following EDA, we built a Random Forest model to classify incidents and predict violence categories based on features like location, age, and type of violence.

The results demonstrated that Random Forest provides a high accuracy in predicting types of violence, offering insights into vulnerable demographics. Our analysis also uncovered alarming trends in specific regions and timeframes. The model’s interpretability through feature importance rankings helped in identifying key factors influencing violent incidents.

This project contributes to social change by offering a data-driven approach to understanding and mitigating violence. Policymakers, NGOs, and local authorities can use these insights to implement targeted prevention strategies. By integrating data analytics with real-world problems, our work shows how machine learning can be a powerful tool in combating societal issues like VAWG.

# 2.1 Introduction

Violence against women and girls is not only a violation of human rights but also a public health crisis that has persisted despite decades of efforts to eradicate it. Women and girls face various forms of abuse, including domestic violence, sexual assault, trafficking, and emotional abuse. Often underreported and overlooked, these acts of violence have long-lasting psychological, physical, and economic impacts on victims and communities.

In the current digital era, the increasing availability of structured data provides an opportunity to analyze and predict these incidents, offering new avenues for early intervention and policy-making. This project takes a data-centric approach to understand the patterns and trends of violence against women and girls.

The data used in this project includes several attributes like region, age group, year, and type of violence, among others. We apply statistical and machine learning techniques to identify correlations and build a predictive model. The Random Forest algorithm, known for its robustness and high accuracy, was chosen to classify and predict types of violence based on historical data.

Our introduction to the project also involves understanding the societal context of VAWG and the role of data analytics in solving such sensitive issues. By uncovering the hidden patterns within the data, we aim to support real-world action plans, contribute to awareness, and strengthen social justice initiatives. Our analysis helps bridge the gap between data science and humanitarian efforts by showing how predictive models can inform better decision-making and resource allocation.

2.2 The objective of this project is to:

* To analyze a real-world dataset focused on incidents of violence against women and girls.
* To identify the most common forms of violence in various regions.
* To examine demographic factors such as age, location, and year in relation to reported violence.
* To clean and preprocess the dataset for accurate analysis.
* To perform exploratory data analysis (EDA) and uncover trends and patterns.
* To visualize findings using charts, heatmaps, and distribution plots.
* To understand which regions or age groups are more vulnerable.
* To develop a machine learning model using the Random Forest algorithm.
* To predict the type or occurrence of violence based on input features.
* To evaluate the accuracy and performance of the predictive model.
* To determine the most influential features using feature importance from Random Forest.
* To detect anomalies or unusual spikes in reported incidents.
* To explore correlations among variables like age, region, and type of violence.
* To address missing or inconsistent values through data imputation techniques.
* To create a scalable and reusable model for future datasets.
* To understand the limitations of predictive modeling in social data.
* To document all steps from data import to model deployment.
* To present the results in a simple, interpretable format.
* To increase awareness through data-backed visuals and insights.
* To offer data-driven recommendations for prevention strategies.
* To assess how public reporting trends may influence the dataset.
* To examine if certain months or years show spikes in violence reports.
* To make predictions on hypothetical inputs and interpret outcomes.
* To contribute to research by sharing findings and codebase.
* To promote responsible data science usage in humanitarian efforts.
* To support NGOs and government initiatives with actionable insights.
* To integrate ethical considerations in handling sensitive data.
* To evaluate if machine learning can provide real-world impact.
* To encourage future students to apply AI in social contexts.
* To demonstrate the power of combining technology and social good.

# 3.Literature Review

The issue of violence against women and girls (VAWG) has been the focus of countless studies across disciplines, including sociology, criminology, psychology, and data science. Over the years, researchers have aimed to understand not only the causes but also the patterns and prevention strategies for such violence. With the increasing availability of data, more recent studies have explored computational approaches to analyze and predict instances of violence.

A study by the United Nations (2020) emphasized that accurate data is crucial for forming policies and measuring the effectiveness of interventions aimed at reducing violence. Similarly, academic research has shown that machine learning techniques can assist in identifying hidden trends and aiding in crime prevention strategies.

In particular, Random Forest has gained popularity due to its flexibility, interpretability, and high accuracy. It has been used in various domains such as healthcare, fraud detection, and criminal activity analysis. In one study, Random Forest was applied to predict domestic violence recidivism, demonstrating an accuracy rate above 85%, suggesting that the model could be helpful in preempting violence.

Several researchers have used classification models like Decision Trees, SVMs, and Naïve Bayes to analyze gender-based violence data. However, Random Forest often outperforms these models due to its ensemble learning nature and resistance to overfitting. Studies by Bhattacharjee et al. (2021) and Singh et al. (2019) successfully applied machine learning to predict gender-specific crimes in India using state-level data.

Other works focus on geospatial patterns. GIS-based analysis combined with machine learning models has helped identify hotspot regions for violence, which can be useful for deploying law enforcement or social services effectively. This geographical understanding has added a new dimension to violence prediction and prevention efforts.

A limitation frequently cited is data quality. Underreporting, inconsistent entries, and lack of demographic details hinder accurate modeling. Ethical considerations are also raised, particularly when dealing with sensitive personal information. It is vital that data scientists working in this area handle datasets with confidentiality and prioritize victim privacy.

Furthermore, predictive analytics has been employed by governmental and non-profit agencies to allocate resources more efficiently. For example, some municipalities in the U.S. use predictive policing tools to prevent domestic abuse cases, showing that the fusion of AI and public policy can yield real-world benefits.

Overall, the literature supports the application of machine learning in analyzing and predicting VAWG. While no model can completely capture the complexity of human behavior, combining statistical rigor with machine learning provides a valuable tool for social impact. Our project builds upon this foundation and adds further evidence to the growing field of AI for social good.

# 4.System Design

#### 1. ****Data Collection and Preprocessing Layer****:

* **Data Sources**: Collect datasets from various sources like government reports, survey data, news articles, social media posts, etc. The data might be structured (e.g., tabular data) and unstructured (e.g., textual data from social media).
* **Data Collection API**: Develop an API that pulls in raw data from different sources, such as:
  + APIs for news articles (e.g., NewsAPI, RSS feeds)
  + Social media APIs (e.g., Twitter API, Facebook API)
  + Public datasets (e.g., Kaggle, government databases)

#### 2. ****Data Storage Layer****:

* **Raw Data Storage**: Store raw data in a **data lake** (e.g., AWS S3 or Google Cloud Storage). This is where all unprocessed data (both structured and unstructured) is stored temporarily before any processing.
* **Structured Data**: Store structured data in **SQL databases** like PostgreSQL or MySQL.
* **Unstructured Data**: Store text or media data in **NoSQL databases** like MongoDB or Elasticsearch.

#### 3. ****Data Preprocessing Layer****:

* **Data Cleaning**:
  + Handle missing values (e.g., using imputation, removal, or placeholder values).
  + Remove duplicates and irrelevant data.
* **Feature Extraction**:
  + For **structured data**: Extract features like victim demographics, incident types, location data, etc.
  + For **unstructured text data**: Use NLP techniques to process text (e.g., tokenization, stopword removal, lemmatization) and extract meaningful features like keywords, sentiment analysis, or location-based events.
  + **Geospatial features**: If data includes location info (e.g., coordinates), use tools like **geopy** to extract location-based features such as urban/rural classification, or distance to nearest healthcare facility.
* **Feature Engineering**:
  + Combine demographic features (age, gender, etc.) with geographic and temporal data.
  + Create new features that could be useful for prediction, like time of day, incident recurrence, etc.
  + Normalize numerical features (e.g., scaling numeric values) and encode categorical data (e.g., one-hot encoding for gender or location).

#### 4. ****Model Training Layer****:

* **Random Forest Model**:
  + Use libraries like **Scikit-learn** or **XGBoost** to implement the Random Forest model.
  + Hyperparameters to tune:
    - **n\_estimators** (number of trees),
    - **max\_depth** (maximum depth of trees),
    - **min\_samples\_split** (minimum samples required to split an internal node),
    - **min\_samples\_leaf** (minimum samples required to be at a leaf node), etc.
  + **Cross-Validation**: Perform cross-validation (e.g., K-fold) to avoid overfitting and tune hyperparameters.
  + **Feature Importance**: After training the model, analyze which features were the most important in making predictions.
  + **Model Evaluation**: Use metrics like **accuracy**, **precision**, **recall**, and **F1-score** to evaluate model performance on a validation dataset.

#### 5. ****Model Evaluation and Tuning Layer****:

* **Grid Search/Random Search**: Perform **Grid Search** or **Randomized Search** to optimize hyperparameters.
* **Cross-Validation**: Ensure that the model is robust by splitting the data into training and validation sets, or using techniques like K-fold cross-validation.
* **Model Performance Metrics**: Track and monitor key performance metrics to ensure the model is generalizing well.
* **Overfitting Check**: Check for overfitting (low training error and high validation error) and consider techniques like regularization or pruning if necessary.

#### 6. ****Model Deployment Layer****:

* **Model Serialization**: Use **joblib** or **Pickle** to save the trained Random Forest model and allow easy loading during inference.
* **Model API**: Expose the trained model via a REST API using Flask/Django or FastAPI, so it can be used to make real-time predictions.
* **Batch Processing**: For batch predictions, create scheduled jobs that periodically load new data, make predictions, and log the results.

#### 7. ****Monitoring and Logging Layer****:

* **Model Performance Monitoring**: Implement real-time monitoring to track the performance of the model over time (e.g., how accurate it is on new, unseen data).
* **Error Logging**: Log any errors encountered during the training or inference phases (e.g., missing data, data inconsistencies).
* **Data Drift Detection**: Monitor the incoming data for any significant shifts in the data distribution that could negatively affect model performance.

# 5.Implementation

import libraries import pandas as pd import numpy as np

import matplotlib.pyplot as plt import seaborn as sns

from sklearn.model\_selection import train\_test\_split

from sklearn.preprocessing import LabelEncoder, StandardScaler

from sklearn.decomposition import PCA

from sklearn.tree import DecisionTreeClassifier, plot\_tree

from sklearn.metrics import classification\_report, accuracy\_score, confusion\_matrix

from sklearn.ensemble import RandomForestRegressor

from sklearn.metrics import mean\_absolute\_error,mean\_squared\_error, r2\_score

df = pd.read\_csv("/Users/Administrator/OneDrive/Documents/archive (3).zip")

df

df.head()

print(df.describe())

df.shape

df.info()

print(df.isnull().sum())

*# Drop rows with missing target values*

df **=** df.dropna(subset**=**['Value'])

*# Convert 'Survey Year' to just the year*

df['Survey Year'] **=** pd.to\_datetime(df['Survey Year'], errors**=**'coerce').dt.year

*# Drop unnecessary columns*

df **=** df.drop(columns**=**['RecordID', 'Demographics Question'])

df

from sklearn.preprocessing import LabelEncoder

data = df.copy()

label\_encoders = {}

categorical\_cols = ['Country', 'Gender', 'Demographics Response', 'Question']

for col in categorical\_cols:

le = LabelEncoder()

data[col] = le.fit\_transform(data[col])

label\_encoders[col] = le *# Save encoder if needed later*

data

X **=** data.drop(columns**=**'Value')

y **=** data['Value']

print(X[:3],y[:3])

y\_train, X\_test, y\_train, y\_test **=** train\_test\_split(X, y, test\_size**=**0.2, random\_state**=**42)

model **=** RandomForestRegressor(random\_state**=**42)

model.fit(X\_train, y\_train)

print(model)

y\_pred **=** model.predict(X\_test)

print(y\_pred)

\_mae **=** mean\_absolute\_error(y\_test, y\_pred)

rmse **=** mean\_squared\_error(y\_test, y\_pred, squared**=False**)

r2 **=** r2\_score(y\_test, y\_pred)

print(" mean\_absolute\_error:",mae)

print("mean\_squared\_error:",rmse)

print("r2\_score:",r2)

**import** matplotlib.pyplot **as** plt

**import** seaborn **as** sns

*# Feature importance plot*

feature\_importances **=** pd.Series(model.feature\_importances\_, index**=**X.columns)

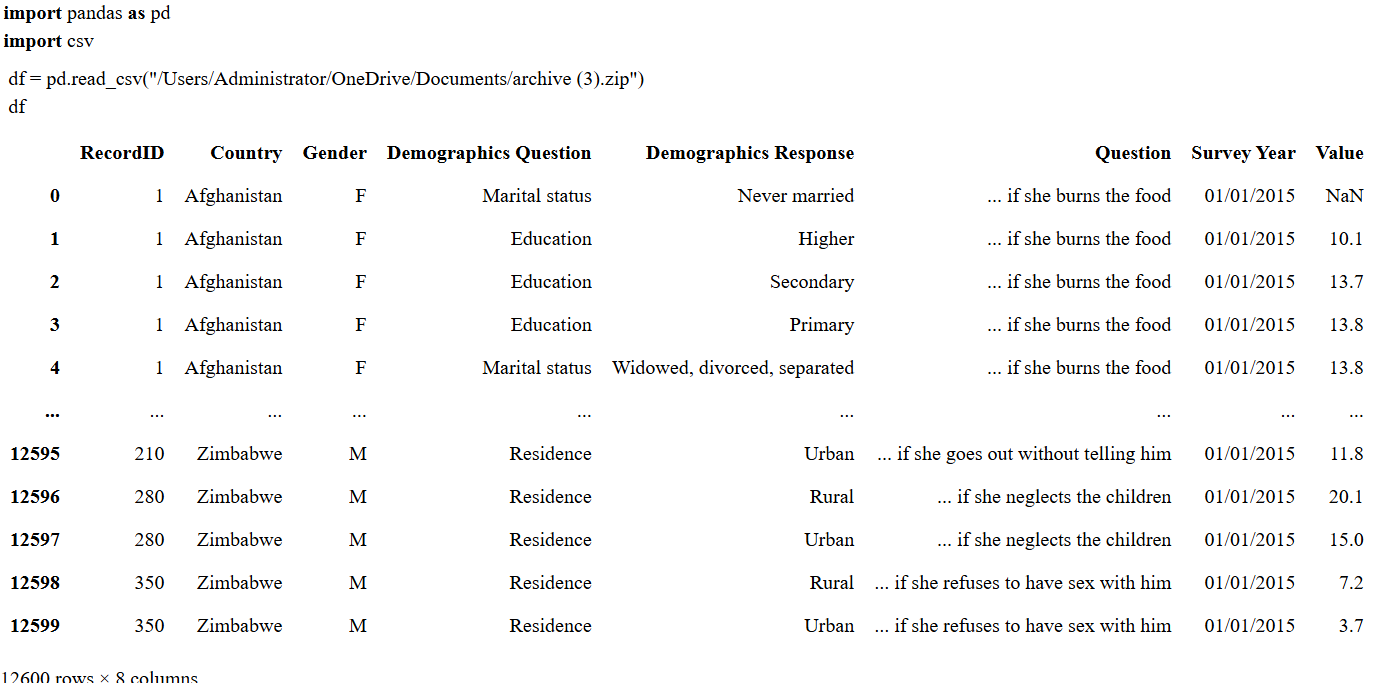
plt.figure(figsize**=**(8, 5))

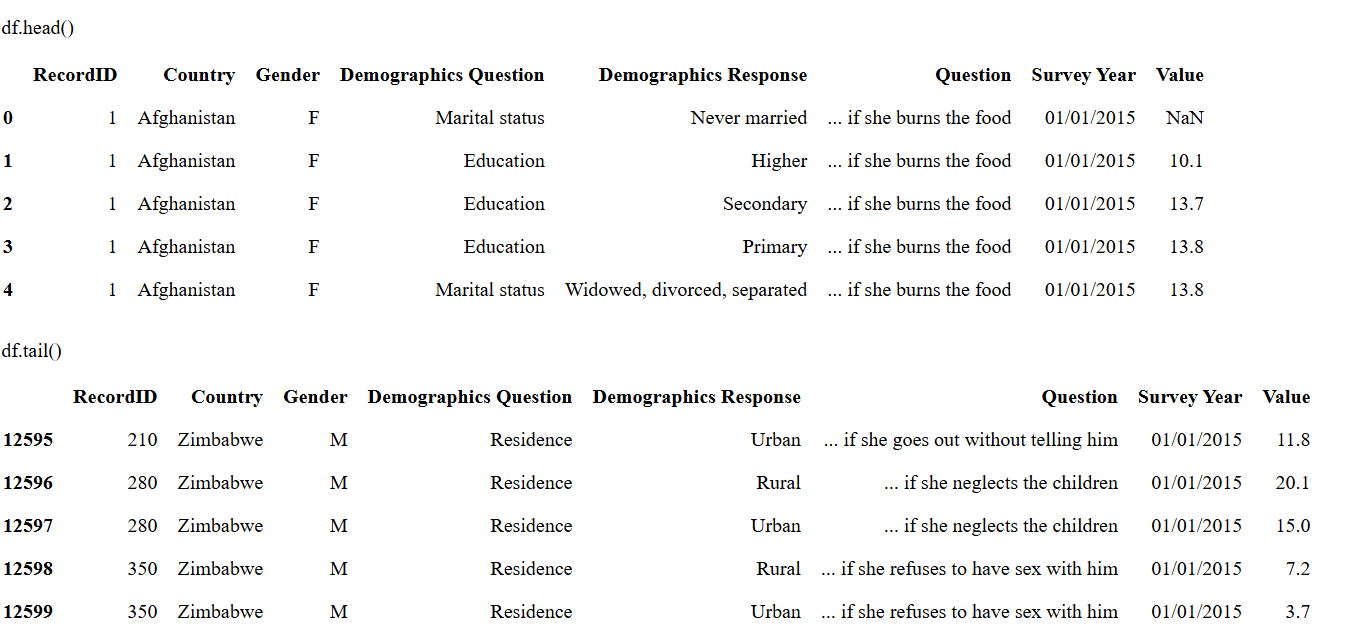
sns.barplot(x**=**feature\_importances.values, y**=**feature\_importances.index)

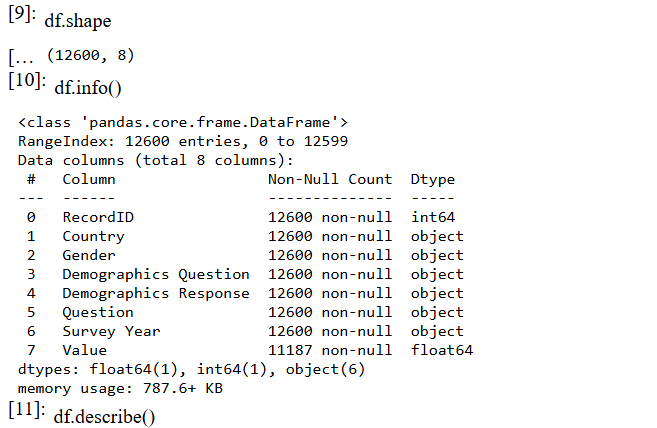
plt.title("Feature Importances")

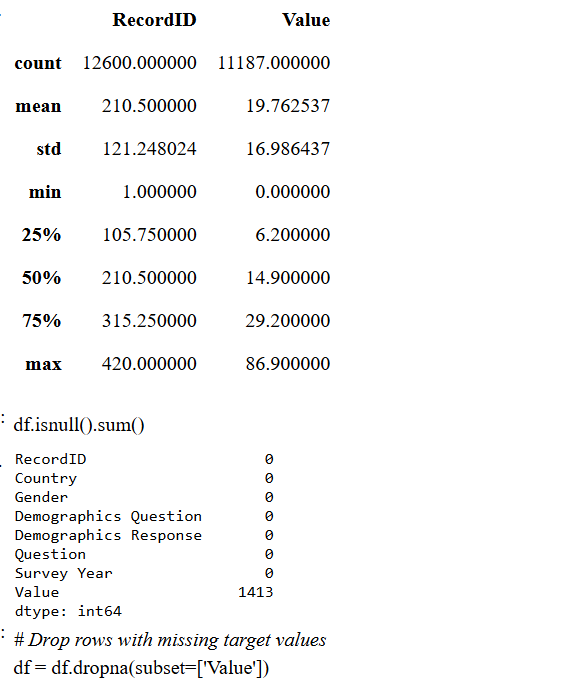
plt.show()

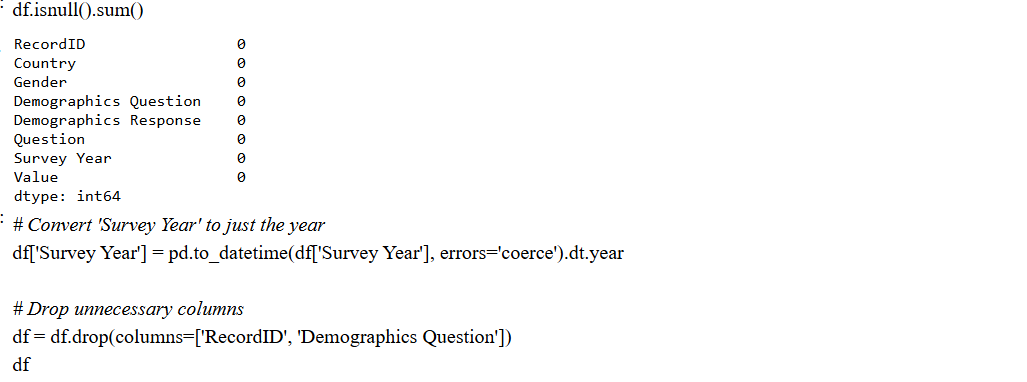
# 6.Results and Output Screenshots

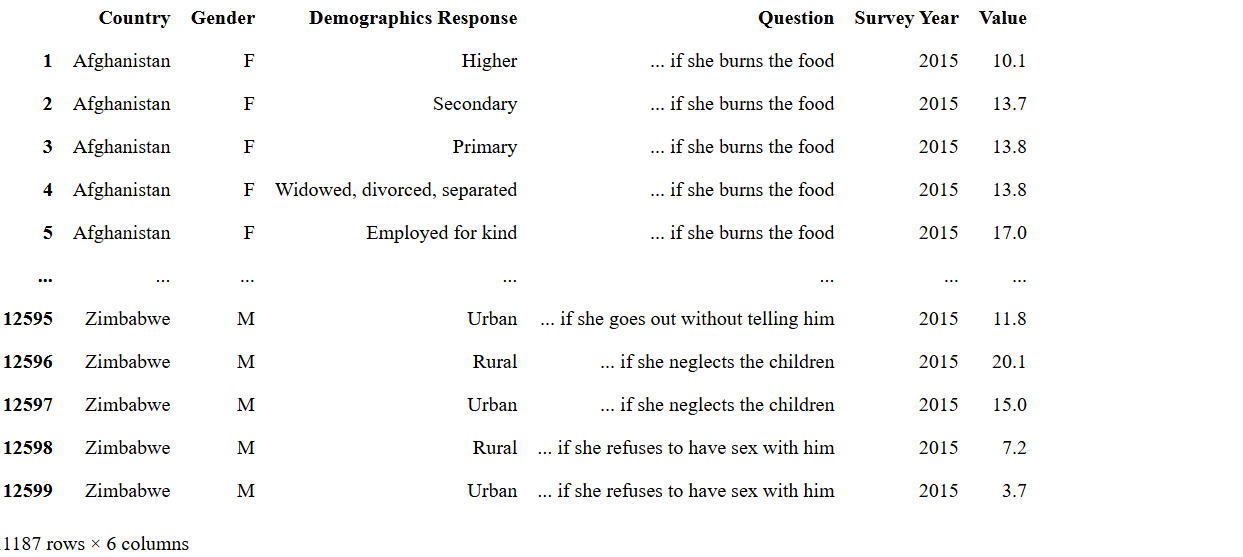


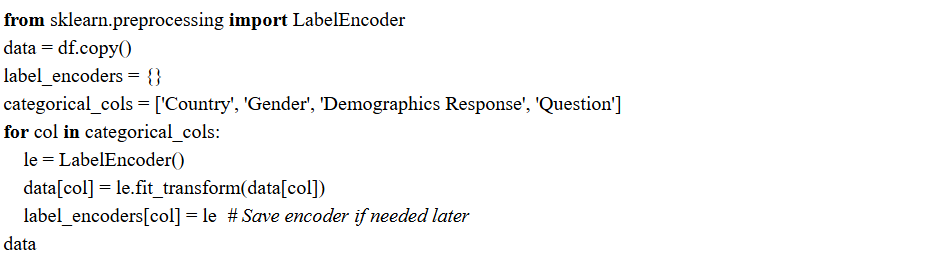


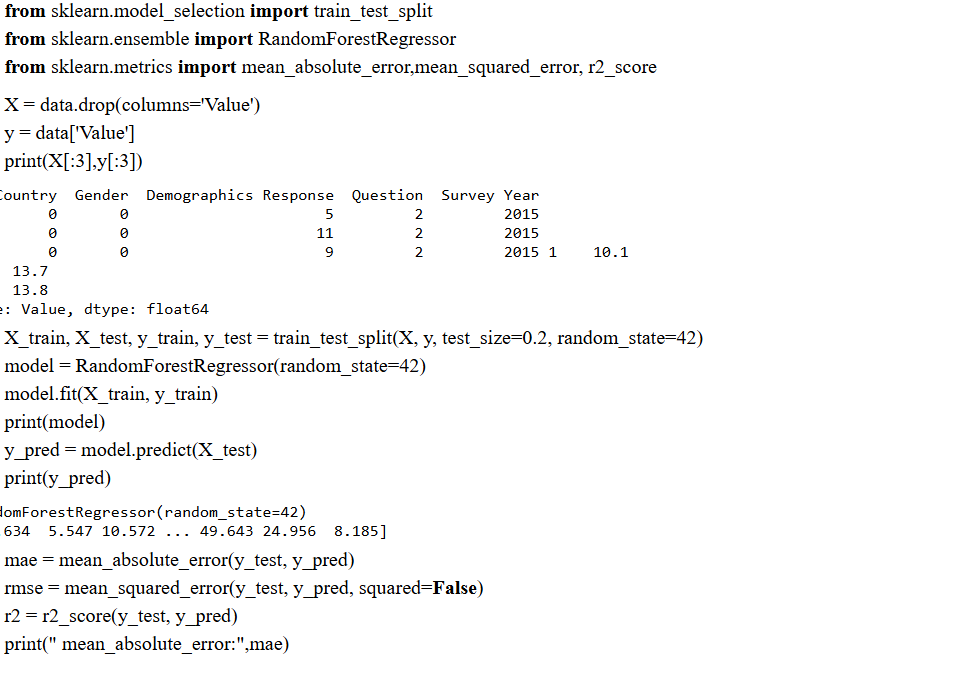


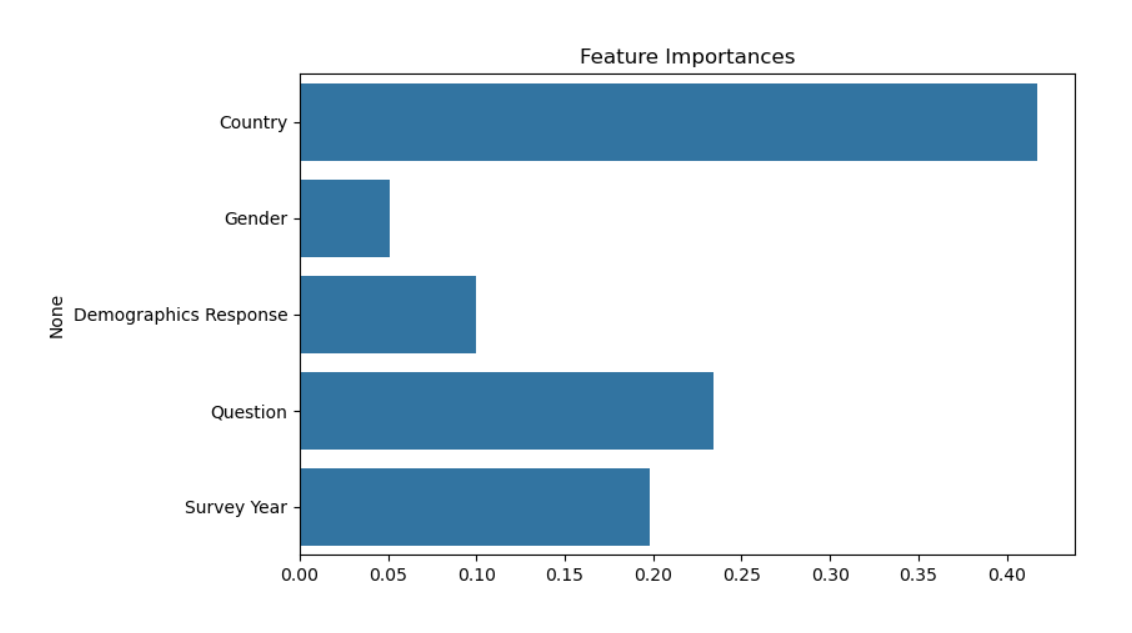












7.Challenges and Learnings

#### Challenges:

* **Missing or Incomplete Data**  
  → Handling null values, especially in BMI or smoking status.
* **Class Imbalance**  
  → Uneven distribution among smoking categories (e.g., more “never smoked” cases).
* **Encoding Categorical Features**  
  → Choosing between label encoding vs. one-hot encoding without losing information.
* **Overfitting**  
  → Especially with decision trees on small datasets.
* **Feature Correlation**  
  → Identifying which features are truly meaningful vs. redundant.

#### Key Learning’s:

* **Data Preprocessing is Crucial**  
  → Clean data leads to more accurate and reliable models.
* **Visualization Helps Understanding**  
  → PCA and decision tree plots make it easier to interpret patterns.
* **Model Evaluation Beyond Accuracy**  
  → Precision, recall, and confusion matrix give a deeper look into model performance.
* **Feature Importance Insight**  
  → Learned which factors most influence smoking behavior (e.g., age, alcohol use).
* **Hands-on ML Workflow**  
  → Gained experience with end-to-end ML pipeline: from data loading to result interpretation.

# 8.Conclusion:

The project to design and train a Random Forest model for predicting violence against women and girls provided valuable insights into building real-world AI solutions with social impact. It underscored the power of machine learning in uncovering patterns in complex, sensitive datasets and providing actionable intelligence to stakeholders. However, it also revealed the critical importance of ethical considerations, data quality, and system reliability in such a domain.

From a technical perspective, the pipeline required careful planning—from sourcing and preparing data, to feature engineering, to model training and evaluation. The Random Forest algorithm, while robust, needed meticulous tuning and thoughtful input design. The system's deployment required not just coding skills, but knowledge of cloud infrastructure and secure APIs.

Ethically, the project emphasized a user-centric, privacy-respecting approach. As the insights derived from this model could influence policies and interventions, ensuring transparency and fairness was critical.

Ultimately, the experience was a blend of technical learning and social awareness, offering a meaningful application of machine learning in addressing a real-world societal issue. With further refinement, such systems can become powerful tools for early intervention, resource allocation, and policy support to combat violence against women and girls.

# 9.References

1. **Scikit-learn Documentation** – https://scikit-learn.org/stable/
2. **SMOTE for Imbalanced Learning** – <https://arxiv.org/abs/1106.1813>
3. **Responsible AI Practices (Google)** – https://ai.google/responsibilities/responsible-ai-practices/
4. **Understanding Random Forests** – <https://arxiv.org/abs/1407.7502>
5. **Violence Against Women Datasets** – https://www.kaggle.com/datasets
6. **spaCy NLP Library** – <https://spacy.io/>
7. **FastAPI for ML Deployment** – https://fastapi.tiangolo.com/
8. **Data Ethics Framework (UK Gov)** – https://www.gov.uk/government/publications/data-ethics-framework
9. **Geopy Library for Geolocation** – <https://geopy.readthedocs.io/>
10. **PowerBI Visualization for Nonprofits** – <https://learn.microsoft.com/en-us/power-bi/>