DSA 210- Final report

• Motivation:

Female Genital Mutilation (FGM) is the non-medical cutting or injury of female genital organs.It is done mainly due to cultural traditions, social pressure, and beliefs about purity or marriage.**FGM harms over 200 million women** and girls worldwide, causing severe pain, infections, childbirth complications, psychological trauma, severe health problems, and sometimes death.

I first learned about this practice through a documentary I watched three years ago, which sparked my interest in understanding the issue more deeply. Since then, I have been motivated to investigate the underlying factors that sustain FGM. The practice is carried out in more than **30 countries,** primarily across parts of Africa, the Middle East, and Asia, and is internationally recognized as a violation of human rights.

Despite global awareness and international efforts to eliminate FGM, progress remains uneven and slow. I chose to work on this topic out of both curiosity about the social, cultural, and economic drivers behind the practice and a strong desire to contribute to raising awareness through data-driven insights.

• Data source:

I first decided to focus on **four countries with different cultural backgrounds and varying levels of FGM prevalence**: **The Gambia, Sierra Leone, Guinea, and Kenya**. Selecting countries with diverse characteristics allows for a more meaningful comparison and a better understanding of how social and cultural factors relate to the practice.

After determining the countries of interest, I searched for **reliable and trustworthy data sources** and chose the **UNICEF Data Center**. Through this platform, I accessed the **Multiple Indicator Cluster Surveys (MICS)** conducted in the selected countries.

MICS is an internationally standardized household survey program developed by UNICEF to collect high-quality data on women and children. The **Multiple Indicator Cluster Surveys (MICS)** are well suited for data science analysis because they provide **large-scale, standardized, and nationally representative datasets** across multiple countries. The surveys are nationally representative and include detailed information at both the **household and individual levels**. For this project, the datasets include variables related to **FGM status**, as well as **demographic, educational, socioeconomic, and household characteristics**, which provide a rich feature set for analysis.

- # Data analysis:

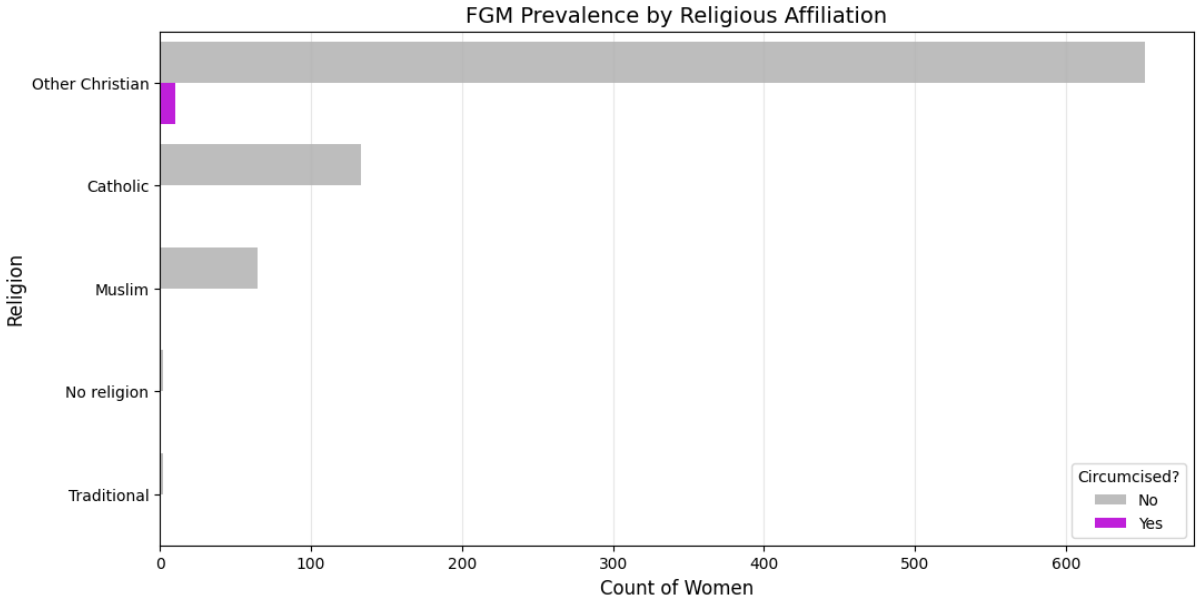
After obtaining the MICS datasets, I merged multiple files to create a unified analytical dataset. Specifically, I combined **`wm.sav`** (women's data, including education variables), **`fg.sav`** (FGM-related data, such as circumcision status), and **`hh.sav`**(household-level data, including region and religion).

Following the merge, I identified key variables—such as **education level, religion, and household characteristics**—that were relevant to the analysis. I selected comparable variables for each country to ensure consistency in the analysis.

These variables were then analyzed in relation to **FGM prevalence** to examine potential associations.

To give an example I will show the data analysis steps I conducted with Kenya Analysis:

1. Religion and FGM Prevalence



Hypothesis Testing (Chi-Square Test)

- **Null Hypothesis (H_0):**
  There is **no statistically significant association** between **religion** and **FGM prevalence** among women in Kenya.
- **Alternative Hypothesis (H_1 / H_a):**
  There **is a statistically significant association** between **religion** and **FGM prevalence** among women in Kenya.

I applied a **chi-square test of independence** to assess whether there was a statistically significant association between the two variables.

Chi-Square P-Value (Religion): 0.8294154741026055
RESULT: NOT SIGNIFICANT ($p > 0.05$)
Interpretation: Religion has NO statistical impact on FGM prevalence.

Since the resulting p value is more than the significant level , I have accepted the null hypothesis.

2. Ethnicity and FGM Prevalence

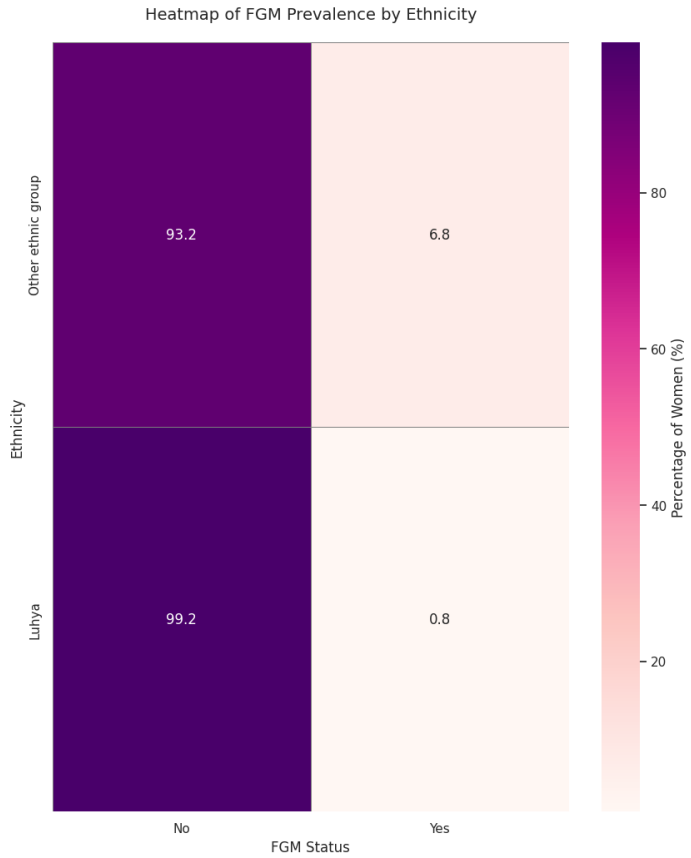
Hypothesis Testing (Chi-Square Test) — Religion vs FGM (Kenya)

• **Null Hypothesis (H_0):**
There is no statistically significant relationship between **ethnicity** and **FGM status**.

• **Alternative Hypothesis (H_1):**
Ethnicity and FGM status are dependent.

```
--- Chi-Square Test Results ---
Chi-square statistic: 15.201761303723654
Degrees of freedom: 1
p-value: 9.661326752652377e-05
```

Since the resulting p value is below the significant level , i have accepted the Alternative hypothesis



For a better visualization I have used a heatmap next to the boxplot.

3. Education and FGM Prevalence

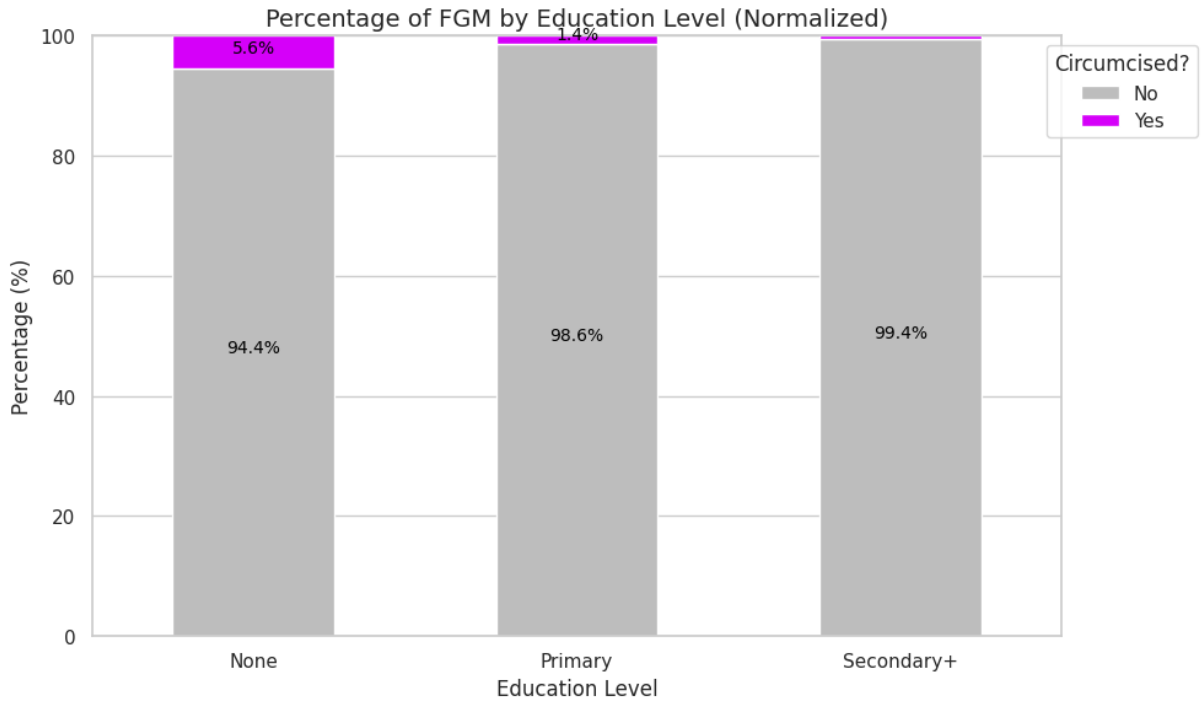
• **Null Hypothesis (H_0):**
There is no statistically significant association between **education level** and **FGM status** among women in Kenya.

• **Alternative Hypothesis (H_1):**
There is a statistically significant association between **education level** and **FGM status** among women in Kenya.

Chi-Square P-Value (Education): 0.04282754759237658

Interpretation: Education level has a **statistically significant impact** on FGM prevalence.

Conclusion: Since the resulting p-value is **below the significance level**, the **null hypothesis (H_0) was rejected**, and the **alternative hypothesis (H_1) was accepted**.



- ## Machine Learning Methods

Machine learning methods were employed to move beyond descriptive and statistical analysis and to **evaluate the relative predictive power of multiple factors simultaneously**. While chi-square tests identify pairwise associations, they cannot capture **nonlinear relationships**, **feature interactions**, or **relative importance** when multiple demographic and cultural variables are considered together.

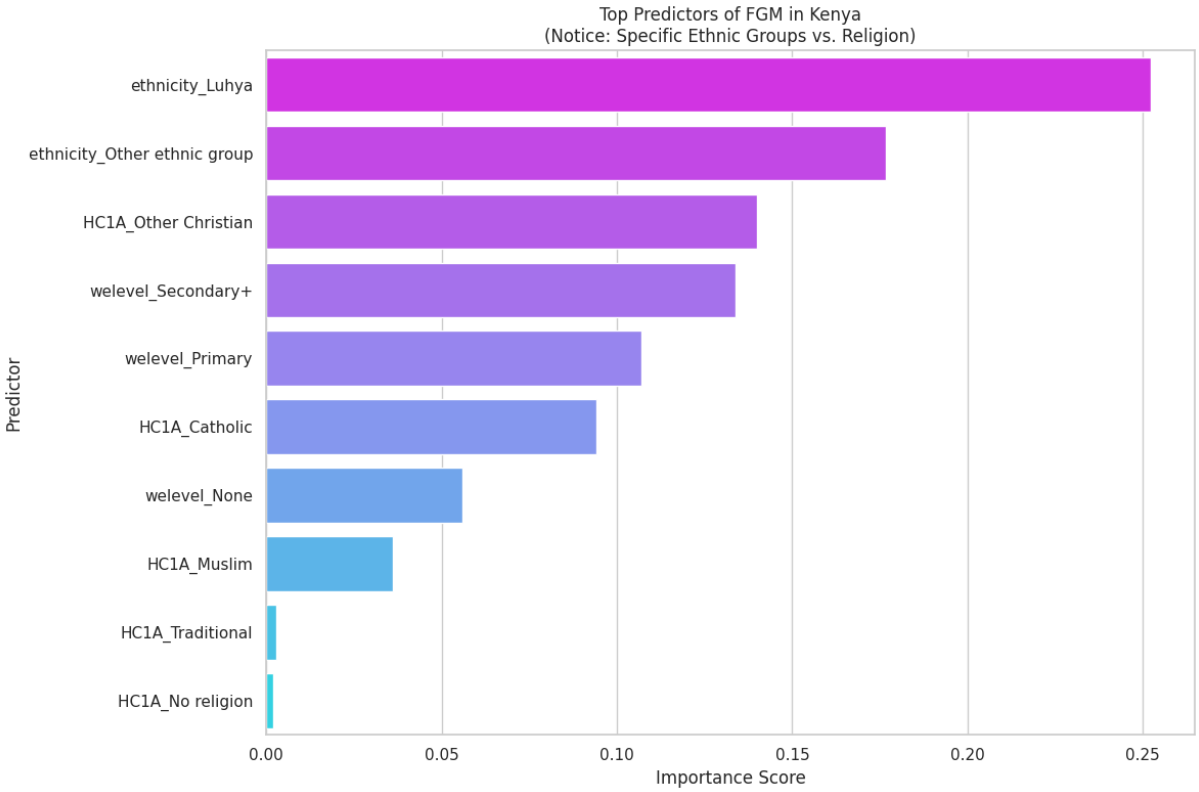
In the context of FGM in Kenya, machine learning enables a **data-driven comparison of competing hypotheses**, particularly whether **ethnicity or religion** is a stronger determinant of FGM status. The goal was to determine which variables most effectively predict FGM outcomes when all factors are evaluated jointly.

1. Random Forest Classifier

A **Random Forest** classification model was used to predict **FGM status (Yes / No)**. This model was selected due to its ability to:

- Handle **high-dimensional categorical data**
- Capture **nonlinear patterns**
- Reduce overfitting through ensemble learning
- Provide **feature importance scores**, allowing interpretability

The model was trained using demographic, cultural, and socioeconomic variables, including **ethnicity indicators, religion, and education level**.

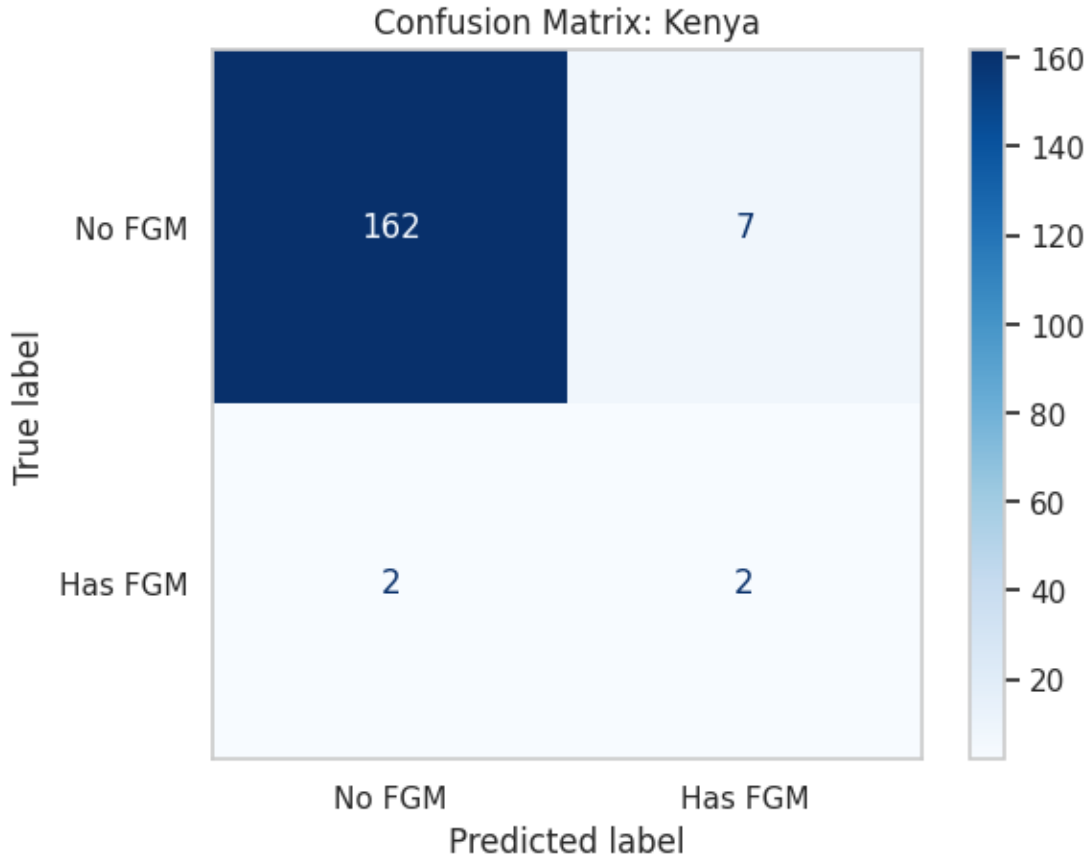


2. Confusion Matrix Analysis

To evaluate model performance beyond overall accuracy, a **confusion matrix** was constructed. This allowed for:

- Assessment of **true positives, true negatives, false positives, and false negatives**
- Evaluation of the model's ability to correctly distinguish between **FGM-practicing and non-practicing groups**
- Detection of potential **class imbalance or prediction bias**

The confusion matrix provides a transparent view of classification performance and supports the reliability of the Random Forest results.



- **Findings: The Drivers of FGM in Kenya**

Based on exploratory data analysis, statistical testing, and machine learning results, two primary drivers of FGM prevalence in Kenya were identified: **ethnicity** as the dominant cultural determinant and **education** as a key socio-economic mitigating factor.

1. Primacy of Ethnicity (Culture > Religion)

Initial analysis focused on religious affiliation; however, both statistical tests and machine learning models indicate that **ethnicity is the dominant confounding variable**.

FGM is not practiced uniformly within any single religion. Instead, it is highly concentrated within specific ethnic groups (such as Somali, Kisii, and Maasai), regardless of whether these groups are predominantly Muslim or Christian. Random Forest feature importance further supports this finding, as ethnic indicators ranked above religious variables in predictive strength.

Implication: Interventions focused solely on religious leadership are likely insufficient unless they directly address **tribal and cultural norms** within high-prevalence ethnic groups.

2. Education as a Protective Factor

Education shows a statistically significant negative association with FGM prevalence (**$p < 0.05$**). Women with no formal education exhibit substantially higher FGM rates (5.6%) compared to those with secondary or higher education (0.6%).

Both statistical testing and machine learning results confirm that education acts as a **protective factor**, reducing the likelihood of FGM even after accounting for ethnicity.

Implication: Education functions as a "social vaccine," weakening adherence to harmful traditional practices and offering an effective pathway for long-term prevention.

Final Insight for Kenya

FGM in Kenya is **not a universal religious practice** but a **culturally specific tradition tied to certain ethnic groups**, with **formal education playing a critical role in mitigation**. Data-driven evidence shows that ethnicity determines risk, while education significantly reduces it.

The results presented in this report focus primarily on Kenya. Detailed country-level findings are documented in the respective **Jupyter Notebook (IPYNB) files**

- # Limitations and Future Work

Limitations

Second, the analysis relies on a **restricted set of features** available and comparable across countries. Due to variations in survey design and feature availability between countries, not all potentially relevant variables (such as local enforcement of anti-FGM laws, urban–rural migration patterns, or generational attitudes) could be consistently included.

Third, the data are **cross-sectional**, which prevents causal inference. While statistical and machine learning methods identify strong associations and predictive patterns, they cannot establish causality or temporal dynamics in FGM practices.

Finally, despite strong model performance in some countries, **class imbalance** and cultural heterogeneity may affect model generalizability, particularly in settings where FGM prevalence is either extremely high or extremely low.

Future Work

Future research could expand this study by incorporating **additional countries**, enabling broader regional comparisons and stronger generalizability of findings. Including a **wider range of features**, such as parental education, media exposure, legal enforcement indicators, and intergenerational variables, would allow for a more comprehensive understanding of the drivers of FGM.

Moreover, applying **longitudinal or repeated survey data** could help capture trends over time and assess the impact of education and policy interventions. From a methodological perspective, future work may explore **advanced models** such as gradient boosting or causal inference techniques to better quantify the mechanisms underlying observed patterns.

Overall, a more comprehensive, multi-country, and feature-rich analysis has the potential to generate deeper insights and support more targeted, data-driven intervention strategies.