Using Machine Learning to Predict Femicide Globally

Barbara Broussard, Adrian Bogart, and Rebecca Rogers

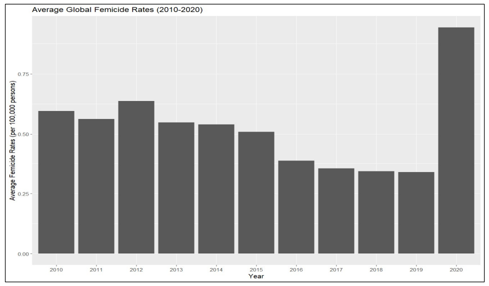
Introduction/Context

- **Scope:** Use data from the United Nations (UN) to build ML models on femicide, or instances of women being killed on account of their gender.
- **Research Question:** Can a machine learning application predict future rates of femicide when accounting for violent crimes, sexual crimes, and access to a criminal justice system?
- **Importance of Research:** Femicide is a brutal crime that occurs globally and is typically underreported by governments. The UN is dedicated to improving gender equality and security of women globally in its Sustainable Development Goals. However, it lacks data and statistical modeling on femicide, which means resources cannot be allocated globally to address this problem.

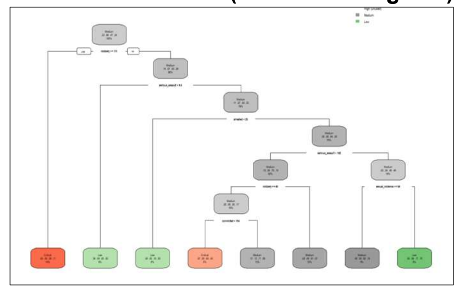
Approach

- Use ***Ridge Regression***, ***LASSO Regression***, ***Principal Components Regression (PCR)***, and ***Partial Least Squares Regression (PLS)*** to predict a subregion's femicide rates. The goal of these methods is to improve prediction power by reducing multicollinearity, variance, and the complexity of models to prevent overfitting.
- Use ***Classification Trees*** and ***K Nearest Neighbors (KNN)*** to classify the intensity of femicide in a subregion. The goal of these methods is to predict femicide by using the best thresholds maximize the classification rate.
- Use ***K-Fold Cross-Validation*** to create more accurate estimations of error and improve prediction accuracy.

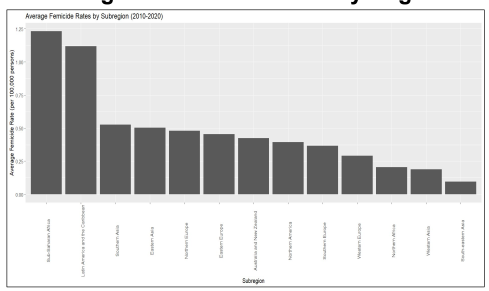
Average Rates of Femicide by Year



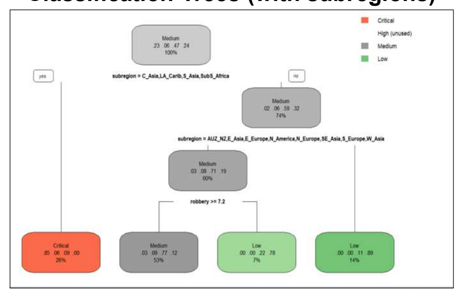
Classification Trees (without subregions)



Average Rates of Femicide by Region



Classification Trees (with subregions)



Regression			
Method	**Parameters**	**MSE**	**Variance Explained**
Baseline Linear Model	22	0.190	87.87%
Ridge	22	0.5209654	68.65%
LASSO	4	0.5690844	65.75%
PCR	14	0.2803102	95.15%
PLS	3	0.3210804	90.98%
Classification			
Method	**Parameters**	**Classification Rate**	
Classification Trees	5	75.38%	
KNN	K = 1	73.85%	

Results/Implications

- Across all models, subregion (especially Sub Saharan Africa and Latin America) and robbery were the most significant predictors of femicide.
- 26% of subregions have "critical" rates of femicide: Central Asia, Latin America and the Caribbean, South Asia, and Sub Saharan Africa.
- The relationship between femicide and sexual violence varied between models. In regression models, high rates of sexual violence resulted in lower rates femicide. In classification models, high rates of sexual violence led to a higher classification of femicide.
- The efficacy of a justice system (arrests, prosecutions, and convictions) were not a significant predictor of femicide.

Assumptions/Limitations/ Challenges or Secondary Results

- **Ethical Implications:**
  - Representation Bias: UN data is not gender specific and there is limited data of violent or sexual crimes affecting females.
  - Measurement Bias: UN did not have enough data on other proxies that are necessary to measure femicide (i.e., intimate relationships, domestic violence, etc.)
- **Constrains in Approach:**
  - Lack of country specific data, so researchers had to group data by subregion. This introduces bias into our models.

Data

Source: United Nations Office on Drugs and Crime (UNODC)

Size/Scale: 130 observations, 11 variables

Predictor: Femicide

Response: Subregion, Year, Kidnapping, Robbery, Serious Assault, Sexual Violence, Rape, Arrests, Prosecutions, and Convictions

*All variables except Subregion and Year are rates per 100,000 people.

Conclusion and Future Research

- The UN can use these models to determine which subregions have the highest rates of femicide and allocate resources to combat femicide.
- The UN needs more gender and femicide specific data at the country or district level to better predict where femicide occurs.

Primary References

https://dataunodc.un.org/

Suggested Reading

https://www.unodc.org/documents/data-and-analysis/gsh/Booklet1.pdf