**Femicide Machine Learning Model Presentation Notes**

**Executive Summary (2-3 minutes)**

(Adrian)

* **Problem:**
  + Femicide, or instances of women being killed on account of their gender, is a brutal crime that occurs globally and is typically underreported by governments. The United Nations (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.
* **Audience:**
  + Our audience is the UN, specifically the UN Office on Drugs and Crime, UN Women, and UN Foundation. These organizations are the primary offices within the UN that combat femicide. The intent of our project is to communicate key femicide statistics and causes of femicide to the UN, so they can allocate resources to address femicide and identify new areas of research.
* **Proposed Solution (our model):** 
  + Use proxies on violent crimes, sexual crimes, and access to criminal justice systems to predict global rates of femicide.
  + Concerns for Implementation: This model acts as a broad overview of the global issue of femicide, but requires more and better data collection to make a stronger model. This model is biased (due to the subregion grouping and low representation of women in crime statistics within certain countries). However, it is a good pilot study and can be more effective with better data.
* **Findings/Recommendations:** 
  + Across all models subregion (specifically Sub Saharan Africa and Latin America) and robbery are most important predictors of femicide.
    - According to classification trees, 26% of the subregions have “critical” rates of femicide. These regions are Central Asia, Latin America and the Caribbean, South Asia, and Sub Saharan Africa.
  + There has been a global decrease in femicide over time.
  + In all models, the relationship between femicide and sexual violence varied.
    - In Ridge, LASSO, PCR, and PLS models, sexual violence typically had a negative correlation with femicide rates.
    - In simple linear regression and classification models, greater sexual violence was positively correlated or led to a higher classification of femicide.
    - Rape was not a significant predictor for classification models.
  + The efficacy of a justice system (i.e., number of individuals arrested, prosecuted, and convicted) was not a significant predictor in regression models. However, the number of arrests and prosecutions were significant predictors in classification models.
  + **Recommendation:** The UN needs to gather better country level data over time, including gender specific data and proxies that help predict femicide. Only then can researchers build more robust models that have greater prediction power.

**Data Background (2-3 minutes)**

(Rebecca)

**Summary:**

* The United Nations Office on Drugs and Crime (UNODC) tracks and analyzes global trends related to drugs and criminal activities. National figures on intentional homicide rates, sexual and violent crimes, and the access and functioning of the justice system are submitted to UNODC by member states through the United Nations Survey of Crime Trends and Operations of Criminal Justice Systems (UN-CTS). The accuracy of the data is dependent on UN member states’ measuring and reporting capabilities and practices and some crimes, especially sexual crimes, may be underreported.
* We limited our analysis to measure the effects of our control variables on femicide rates from 2010 to 2020 by subregion. Our additional control variables included crime rates (kidnapping, robbery, serious assaults, sexual violence, and rape) and the strength and access to the judicial system (measured by arrests, conviction rates, and prosecution rates). Rates were measured per 100,000 people.
* **Source:** 
  + United Nations Office on Drugs and Crime (UNODC) [Main Webpage](https://dataunodc.un.org/)
  + National figures: intentional homicide rates, sexual and violent crimes, and the access and functioning of the justice system.
  + Population Data: United Nations Department of Economic and Social Affairs (only used in rate calculation).
* **Limitations:**
  + Only a decade’s worth of data (2010 to 2020) was used due to missing data prior to 2010.
  + Sexual crimes are likely to be underreported, with 100+ missing observations over time and across the majority of subregions.
  + Accuracy of data depends on Member State. Due to the lack of Member State data over time, the research team had to group predictors by subregion, which may introduce bias into our results.
  + Other proxies of femicide contained large amounts of missing data (i.e., the context of intimate or family relationships, domestic violence, women's social/legal/education vulnerabilities, and cultural/social factors).
* **Variables:**
  + These variables were selected due to the lack of missing data.
  + Predictors: Subregion, Year, Kidnapping Rates, Robbery Rates, Serious Assault Rates, Sexual Violence Rates, Rape Rates, Arrests, Conviction Rates, Prosecution Rates.
  + Response: Femicide Rate
    - Additionally, there is a femicide rate with categorical predictors (Critical, High, Medium, and Low) based on the quartiles of femicide rates. Since there are no international standards on the severity of femicide, the research team created four groups to help prioritize subregions for policymakers.
* Sub Saharan Africa and Latin America have multiple years when femicide rates were over 2.5, which is the highest rate among all of the subregions.

**Linear Models: Numerical Rate of Femicide (3-4 minutes)**

(Rebecca/Barbara)

* **Baseline Full Linear Model:** 
  + In our baseline full linear model, we see that it does a fairly good job of accounting for the variance in femicide rates, explaining 87.87% of the variability. The most significant variables at the 0.05 level are regions (Central Asia, Eastern Asia, Eastern Europe, Latin America, Northern Europe, South Asia, Sub-Saharan Africa and Western Europe), year, and several crimes (serious assault and rape).
  + There is a negative association of femicide rates over time, which reflects the time-trend bar plot. The two significant crimes (rape and serious assault) have a positive correlation with femicide, which is to be expected because the murder of women is an escalatory crime that is often associated with assaults and rapes. All of the regions with a significant association in predicting femicide rates have a positive association, with Sub Saharan Africa and Latin America having the highest coefficients (3.731 and 2.760 respectively).
* **Models Used:** 
  + LASSO, Ridge, Principal Component Regression, Partial Least Squares (all with cross-validation).
  + We used these models to reduce the dimensions of the original linear model, as well as tune to model to see if we could determine a lower MSE and higher percentage of response variance explained.
  + There was a minor issue with multicollinearity, specifically involving crime rates, sexual violence, and robbery.
  + LASSO and Ridge models were used to improve validity and reduce overfitting.
  + Principal Components Regression and Partial Least Squares Regression were used to reduce multicollinearity.
* **Best Model**:
  + The base linear model has the lowest MSE of all of the models, and explains 87.87 percent of the variability in femicide rates. However, there are several independent variables which are highly correlated with other independent variables (sexual violence and robbery).
  + The Principal Components Regression model has the lowest MSE and highest variance explained of the machine learning models. While it has a higher MSE than the base model, it reduces multicollinearity and thus serves as a better model to predict future femicide rates.
* **Findings:**
  + Across all models, we see that Sub Saharan Africa and Latin America have a greater positive association with predicting femicide rates than most of the other countries. This may be because these regions have higher femicide rates than the rest of the world. The other regions in the data are generally negatively correlated with an increase in femicide rates, but the degree of their association varies with each model.
  + Among the types of crimes, robbery had the greatest association with predicting future femicide rates, and was positive, suggesting that countries with higher rates of burglary are more likely to have higher rates of femicide. Sexual violence was included in each model, although the degree of its association varied. It usually had a negative correlation with femicide rates, suggesting that as sexual violence rates increase, rates of femicide may decrease.
  + Interestingly, the measures of the functioning of the justice system (arrests, conviction, and prosecution) were not very influential in any of our models and were completely eliminated from the LASSO model. This suggests that there is a more complicated relationship between the strength of legal enforcement and the reduction of femicide rates.

| **Method** | **Predictors (P)** | **MSE** | **Variance Explained** |
| --- | --- | --- | --- |
| Full Linear Model | 22 | 0.190 | 0.8787 |
| Ridge | 22 | 0.5209654 | 0.6864701 |
| Lasso | 4 | 0.5690844 | 0.6575109 |
| **PCR** | **14** | **0.2803102** | **0.9515** |
| PLSR | 3 | 0.3210804 | 0.9098 |

**Categorical Models: Categorical Rate of Femicide (3-4 minutes)**

(Adrian)

* **Models Used:**
  + Classification Trees and K Nearest Neighbors (KNN) were used to classify the intensity of femicide in a subregion. The goal of these methods is to predict femicide by using the best thresholds to maximize the classification rate. The classification trees used K-Fold Cross-Validation, Bagging, and Random Trees to create more accurate estimations of error and improve prediction accuracy. KNN used tuning to find the optimal K to improve prediction accuracy.
* **Best Model:** 
  + The best model is Pruned and Tuned Trees with subregions (prediction classification rate of 75.38%, ~2-3% better than KNN). We recommend using Classification Trees because:
    - Provides a visual for our audience and is a more useful model for policy decisions than KNN.
    - Highest prediction classification rate when using Pruned and Tuned Trees/Random Trees.
    - K for KNN is too little (K = 1, suggesting prediction will be unreliable; based on a small set of data which could lead to overfitting).
* **Findings:**
  + 26% of the subregions have “critical” rates of femicide. These regions are Central Asia, Latin American and the Caribbean, South Asia, and Sub Saharan Africa. Subregion and robbery are most critical.
    - In the remaining regions (Australia and New Zealand, East Asia, Eastern Europe, North America, Southeast Asia, Southern Europe, and Western Asia), rates are medium if the rate of robbery is greater than or equal to 7.2 per 100,000 people.
  + Without subregion, the following variables are most critical: rates of robbery, serious assault, number of individuals arrested, sexual violence, and number of individuals convicted.
    - 14% of “critical” subregions have rates of robbery greater than or equal to 111 per 100,000 people. 5% of “critical” subregions have:
      * Serious assault is less than 182 per 100,000 people.
      * Robbery is greater than or equal to 40 per 100,000 people.
      * Conviction rates are less than 194 per 100,000 people.
  + Significant predictors for the regression models were not significant for the classification models (i.e., rape, year).
  + Predictors not significant for the regression models were significant for the classification models (i.e., number of people arrested and convicted).

| **Method** | **Parameters** | **Prediction Classification Rate** |
| --- | --- | --- |
| Trees (with subregion) | 2 | Pruned Trees: 75.38%  Random Trees: 75.19% |
| Trees (without subregion) | 5 | Pruned Trees: 58.46%  Random Trees: 72.09% |
| KNN | K = 1 | 73.85% |

**Summary of Findings and Policy Suggestions (2-3 minutes)**

(Barbara)

* **Findings:**
  + There seems to be agreement between both classification and regression models that robbery, Central Asia, Latin America, and Sub Saharan Africa are the most influential predictors. Robbery, Central Asia, and Sub Saharan Africa appears to be the most important predictor overall. PCR and PLS showed that these three variables explained a large majority of variance. With Latin America and Sub Saharan Africa being so important, it is clear that subregion is an important variable to predict femicide.
* **Limitations:**
  + Country Reporting & Accuracy
    - International data is not independently collected and relies on countries to report their numbers, which means researchers may or may not have accurate reporting of femicide. Our model is limited because the numbers are biased. The reported rates of femicide in some regions are likely significantly lower than what they actually are.
  + Time Issues & Dropped Information
    - Some countries have not provided data consistently over the years, resulting in incomplete data. As such, this analysis is limited and only based on the information we were able to use that did not have missing observations. Meaning this model may not be limited in its ability to predict femicide because when countries did not provide data, it affects the overall subregion rate of femicide.
* **Policy Implications:**
  + First, these models show that there is a significant correlation between femicide and robbery. Robbery is a relatively common crime that should be accurately reported (unlike crimes like sexual violence, serious assaults, and rape which are more difficult to report or may not be reported by some countries). Therefore, it is a proxy for crime in general. So this trend may suggest that an increase in crime leads to an increase in femicide.
  + Second, the regression models show there is a statistically significant difference between different subregions, suggesting that certain subregions (like Sub Saharan Africa and South America) experience greater rates of femicide than other subregions. This means the UNODC can use our models to determine which subregions have the highest rates of femicide and begin to allocate resources and money to those areas to address the problem.
  + Finally, We suggest that future data collection, ML models, or studies should collect country or district level data on homicide rates by gender in subregions with high femicide rates, such as Latin America and Sub Saharan Africa, to better predict where femicide is occurring. The UNODC can then allocate resources more efficiently to combat femicide.

**Codebook: (25 total variables)**

| **subregion** | UNODC subregion names (16 total) |
| --- | --- |
| **year** | Years spanning from 2010-2020 |
| **femicide** | Average annual intentional homicide rates for females (measured in per 100,000 people) *[UNODC data]* |
| **femicide\_class** | Categorical predictors (Critical, High, Medium, and Low) based on the quartiles of femicide rates |
| **kidnapping** | Average annual kidnapping rates (measured in per 100,000 people) *[UNODC data]* |
| **robbery** | Average annual robbery rates (measured in per 100,000 people) *[UNODC data]* |
| **serious\_assault** | Intentional or reckless application of serious physical force inflicted upon the body of a person resulting in serious bodily injury. [annual average rates per 100,000 people] *[UNODC data]* |
| **sexual\_violence** | Unwanted sexual act, attempt to obtain a sexual act, or contact or communication with unwanted sexual attention without valid consent or with consent as a result of intimidation, force, fraud, coercion, threat, deception, use of drugs or alcohol, or abuse of power or of a position of vulnerability. [annual average rates per 100,000 people] *[UNODC data]* |
| **rape** | Sub-indicator of sexual\_violence: Sexual penetration without valid consent or with consent as a result of intimidation, force, fraud, coercion, threat, deception, use of drugs or alcohol, abuse of power or of  a position of vulnerability, or the giving or receiving of benefits.[annual average rates per 100,000 people] *[UNODC data]* |
| **arrested** | Average annual number of persons arrested by local law enforcement *[UNODC data]* |
| **prosecuted** | Average annual number of alleged offenders against whom prosecution commenced in the reporting year.  Persons may be prosecuted by the public prosecutor or the law enforcement agency responsible for prosecution. *[UNODC data]* |
| **convicted** | Average annual number of Persons found guilty by any legal body authorized to pronounce a conviction under national criminal law, whether or not the conviction was later upheld. Persons receiving a sentence after plea-bargaining, or in an abbreviated court procedure, are counted as persons convicted. *[UNODC data]* |
| **AUZ\_NZ** | A dummy variable indicating that the observation occurred in Australia and New Zealand |
| **C\_Asia** | A dummy variable indicating that the observation occurred in Central Asia |
| **E\_Asia** | A dummy variable indicating that the observation occurred in Eastern Asia |
| **E\_Europe** | A dummy variable indicating that the observation occurred in Eastern Europe |
| **LA\_Carib** | A dummy variable indicating that the observation occurred in Latin America and the Caribbean |
| **N\_Africa** | A dummy variable indicating that the observation occurred in Northern Africa |
| **N\_America** | A dummy variable indicating that the observation occurred in Northern America |
| **N\_Europe** | A dummy variable indicating that the observation occurred in Northern Europe |
| **SE\_Asia** | A dummy variable indicating that the observation occurred in Southeast Asia |
| **S\_Asia** | A dummy variable indicating that the observation occurred in Southern Asia |
| **S\_Europe** | A dummy variable indicating that the observation occurred in Southern Europe |
| **SubS\_Africa** | A dummy variable indicating that the observation occurred in Sub-saharan Africa |
| **W\_Africa** | A dummy variable indicating that the observation occurred in Western Africa (not used) |
| **W\_Europe** | A dummy variable indicating that the observation occurred in Western Europe |

**Countries by Subregion:**

| **Subregion** | **Number of Countries** | **Countries** |
| --- | --- | --- |
| Australia and New Zealand | 2 | Australia and New Zealand |
| Central Asia | 5 | Kazakhstan; Kyrgyzstan; Tajikistan; Turkmenistan; Uzbekistan |
| East Asia | 6 | China; China, Hong Kong; China, Macau; Japan; Mongolia; Republic of Korea |
| Eastern Europe | 11 | Belarus; Bulgaria; Czechia; Hungary; Poland; Republic of Moldova; Romania; Russian Federation; Slovakia; Ukraine |
| Latin America and the Caribbean | 46 | Anguilla; Antigua and Barbuda; Argentina; Aruba; Bahamas; Barbados; Belize; Bolivia (Plurinational State of); Brazil; British Virgin Islands; Cayman Islands; Chile; Columbia; Costa Rica; Cuba; Curaçao; Dominica; Dominican Republic; Ecuador; El Salvador; French Guiana; Granada; Guadeloupe; Guatemala; Guyana; Haiti; Honduras; Jamaica; Martinique; Mexico; Montserrat; Nicaragua; Panama; Paraguay; Peru; Puerto Rico; Saint Kitts and Nevis; Saint Lucia; Saint Martin; Saint Vincent and the Grenadines; Suriname; Trinidad and Tobago; Turks and Caicos Islands; United States Virgin Islands; Uruguay; Venezuela |
| Melanesia | 5 | Fiji; New Caledonia; Papua New Guinea; Solomon Islands; Vanuatu |
| Micronesia | 5 | Kiribati; Guam; Marshall Islands; Micronesia; Palau |
| Northern Africa | 4 | Algeria; Egypt; Morocco; Tunisia |
| Northern America | 5 | Bermuda; Canada; Greenland; Saint Pierre and Miquelon; United States of America |
| Northern Europe | 11 | Denmark; Estonia; Finland; Iceland; Isle of Man; Ireland; Latvia; Lithuania; Norway; Sweden; United Kingdom |
| Southeastern Asia | 10 | Brunei Darussalam; Cambodia; Indonesia; Malaysia; Myanmar; Philippians; Singapore; Thailand; Timor-Leste; Vietnam |
| Southern Asia | 9 | Afghanistan; Bangladesh; Bhutan; India; Iran; Maldives; Nepal; Pakistan; Sri Lanka |
| Southern Europe | 17 | Albania; Andora; Bosnia and Herzegovina; Croatia; Gibraltar; Greece; Holy See; Italy; Kosovo; Malta; Montenegro; North Macedonia; Portugal; San Marino; Serbia; Slovenia; Spain |
| Sub-saharan Africa | 36 | Angola; Botswana; Burkina Faso; Burundi; Cabo Verde; Cameroon; Djibouti; Eritrea; Eswatini; Ethiopia; Ghana; Guinea-Bissau; Kenya; Lesotho; Liberia; Madagascar; Malawi; Mauritania; Mauritius; Mayotte; Mozambique; Namibia; Niger; Nigeria; Reunion; Rwanda; Saint Helena; Sao Tome and Principe; Seychelles; Sierra Leone; South Africa; South Sudan; Uganda; United Republic of Tanzania; Zambia; Zimbabwe |
| Western Asia | 18 | Armenia; Azerbaijan; Bahrain; Cyprus; Georgia; Iraq; Israel; Jordan; Kuwait; Lebanon; Oman; Qatar; Saudi Arabia; State of Palestine; Syrian Arab Republic; Turkiye; United Arab Emirates; Yemen |
| Western Europe | 9 | Austria; Belgium; France; Germany; Liechtenstein; Luxembourg; Monaco; Netherlands; Switzerland |
| **Total Regions: 16** | **Total: 199** |  |