**Methodology Plan: Labor Market Trends and ~~Ethnic~~ Conflict Prediction using Machine Learning (ML)**

**1. Research Objectives & Questions**

**Main Goal:**

* Identify key factors influencing **~~ethnic~~ conflicts** in the **Asia**.
* Predict ~~ethnic~~ conflicts (e.g., **2018–2023**) using **ML models trained on past data (1990–2017)**.

**Key Research Questions:**

1. What are the **most influential factors** driving ~~ethnic~~ conflict in Asia?
2. Can we **accurately predict** ~~ethnic~~ conflicts using machine learning?
3. How do **labor market conditions (ILOSTAT) and ~~ethnic~~ group characteristics (GTD)** interact with other drivers (e.g., political factors)?
4. Can ML-driven feature discovery reveal previously overlooked predictors of ~~ethnic~~ conflict? (feature selection using ML models)

**Preliminary Hypotheses:**

* H1: Higher informal employment rates → higher ~~ethnic~~ conflict risk.
* H2: Wage stagnation in certain sectors → increased risk of ~~unrest~~.
* H3: Labor market disruptions (e.g., trade shocks) → increased ~~ethnic/localized~~ conflicts.

**2. Data Collection & Preprocessing**

The followings are **candidate features** which will be harmonized into a **country-year** panel format for the period 1990–2022, with preprocessing steps including normalization, data imputation, and aggregation where needed.

1. **Background Features**

* Since this study is **exploratory**, the choice of background features prioritizes **broad socio-political and economic context** rather than well-known conflict drivers in conflict literature. The idea is to provide a **neutral contextual baseline** rather than reinforcing already known conflict predictors.
* **Broad Socio-political Features**:
  + World Bank Group – World Development Indicators (WDI)
    - Selected Features:
      * GDP per capita (PPP): GDP per capita, PPP (constant 2021 international $)
      * % Urban population: Urban population (% of total population)
      * Literacy rate: Due to the near-total absence of data for the “Literacy rate, adult total (% of people ages 15 and above)” in Asia, the “Compulsory education, duration (years)” indicator is used as a proxy for literacy rate in this region.
      * Internet access rates: Individuals using the Internet (% of population)
      * Government health expenditure: Domestic general government health expenditure (% of GDP)
  + Varieties of Democracy (V-Dem)
    - Selected Features:
      * Electoral democracy index: v2x\_polyarchy
      * Civil society participation index: v2x\_cspart
      * Freedom of expression index: v2x\_freexp
      * Local government power index: v2xel\_locelec
        + All of these selected indices use an interval scale ranging from 0 (low) to 1 (high).
  + UNDP Human Development Reports
    - Selected Features:
      * Human Development Index (HDI): hdi
      * Gender inequality index: gii
        + The above two selected indices use an interval scale ranging from 0 (low) to 1 (high).
  + a custom Multidimensional Poverty Index (MPI)
    - OPHI and World Bank MPI framework
    - Selected Indicators:
      * World Bank Group – World Development Indicators (WDI)
        + Access to clean fuels and technologies for cooking (% of population)
        + Access to electricity (% of population)
        + Mortality rate, under-5 (per 1,000 live births)
        + People using at least basic drinking water services (% of population)
        + People using at least basic sanitation services (% of population)
        + Prevalence of undernourishment (% of population)
      * World Bank Group – Education Statistic Indicators (EdStats)
        + School life expectancy, primary, both sexes (years)
        + Total net enrolment rate, primary, both sexes (%)
  + Quality of Government Institute (QoG Standard Dataset)
    - Selected Features:
      * Government Effectiveness, Estimate: wbgi\_gee (originally from World Bank Group – World Governance Indicators)
      * Rule of Law, Estimate: wbgi\_rle (originally from World Bank Group – World Governance Indicators)
      * Trust in Parliament: ess\_trparl
        + a score of 0 (no trust) – 10 (complete trust) how much people trust the parliament.
      * Governance & Institutional Quality
        + Control of Corruption (Estimate): wbgi\_cce
        + Government Effectiveness (Estimate): wbgi\_gee
        + Political Stability & Absence of Violence/Terrorism (Estimate): wbgi\_pve
        + Rule of Law (Estimate): wbgi\_rle
      * Civil Liberties & Human Rights
        + Right to Assembly and Association (0 = none, 2 = full rights): ciri\_assn
        + Government Disappearances: ciri\_disap
        + Political Killings: ciri\_kill
        + Physical Integrity Rights Index: ciri\_physint
      * Democracy & Political Freedom
        + Freedom House Civil Liberties Score (1 = most free, 7 = least free): fh\_cl
        + Freedom House Political Rights Score: fh\_pr
        + Freedom House Overall Status (0 = not free, 2 = free): fh\_status
        + Democracy Indicator (Boix-Miller-Rosato): bmr\_dem
        + Democracy with Female Suffrage (1 = yes): bmr\_demfsuf
        + Democratic Transition (-1 = breakdown, 1 = transition): bmr\_demtran
      * Regime Instability & Conflict Risk
        + Number of Coups: br\_coup
        + Foreign-Instigated Coups: br\_fcoup
        + Self-Coups: br\_scoup
        + Regime Change Indicator: br\_regch
        + Postponed Election Indicator: br\_elecpost
      * Regime Type & Historical Context
        + Was the regime communist? (1 = yes): br\_com
        + Presidential System Indicator: br\_pres
      * Socioeconomic Inequality
        + Income Share of Top 10%: top\_top10\_income\_share
        + Income Share of Top 1%: top\_top1\_income\_share
      * Health & Development
        + Total Life Expectancy (Years, all genders): ihme\_lifexp\_0102t
  + UN World Population Prospects
    - Selected Features:
      * ~~Youth population (% under 25)~~
      * Median age: Median Age (years)
      * Fertility rate: Total Fertility Rate (TFR) (live births per woman)
      * Population density: Population Density (persons per square km)
      * Net Migration Rate (per 1,000 population)
  + Freedom House – Freedom in the World
    - Selected Features:
      * Aggregate score (0–100)
  + International Telecommunication Union (ITU)
    - Selected Features:
      * Mobile cellular subscriptions per 100 people
      * Internet penetration
* **General Economic Features**:
  + World Bank – World Development Indicators (WDI)
    - Selected Features:
      * Economic Growth & Output
        + GDP (current US$)
        + GDP per capita (current US$)
        + GDP per capita, PPP (constant 2017 international $)
        + GDP growth (annual %)
        + GDP per capita growth (annual %)
        + Gross national income (GNI) per capita, Atlas method (current US$)
      * Inflation & Prices
        + Inflation, consumer prices (annual %)
        + Inflation, GDP deflator (annual %)
      * Trade & Openness
        + Exports of goods and services (% of GDP)
        + Imports of goods and services (% of GDP)
        + Trade (% of GDP) (i.e., exports + imports)
        + Current account balance (% of GDP)
      * Investment & Capital Formation
        + Gross capital formation (% of GDP)
        + Gross fixed capital formation (% of GDP)
        + Foreign direct investment, net inflows (% of GDP)
      * Finance & Credit
        + Domestic credit to private sector (% of GDP)
        + Bank nonperforming loans to total gross loans (%)
        + Broad money (% of GDP)
        + Domestic credit provided by financial sector (% of GDP)
      * Labor & Productivity
        + Employment to population ratio, 15+, total (%)
        + Labor force participation rate, total (% of total population ages 15+)
        + Value added per worker (constant 2010 US$)
        + Labor income share as percent of GDP, adjusted (%) (also available from ILOSTAT)
      * Remittances & Transfers
        + Personal remittances, received (% of GDP)
        + Net official development assistance (ODA) received (% of GNI)
      * Debt & Public Sector Finance
        + External debt stocks, total (DOD, current US$)
        + Total reserves (includes gold, current US$)
        + Government expenditure (% of GDP)
        + Revenue, excluding grants (% of GDP)
        + General government gross debt (% of GDP) (check IMF WEO for this)

2. **Main Feature**

* These are the **key variable** you hypothesize might influence conflict.
* **Labor Market Dataset:** 
  + ILOSTAT Labour Market Indicators:
    - Employment-to-Population Ratio
    - Informal Employment Rate
    - Jobs Gap Rate
    - Labour Force Participation Rate
    - Labour Income Share
    - Unemployment Rate
    - Working Poverty Rate
    - Youth NEET rate

3. **Target Feature**

* The dependent variable is **~~ethnic~~ conflict risk**: modeled as a **multi-variate target** to explore different conflict intensity patterns.
* **Conflict Dataset**: (GTD (for historical trend) & ACLED (for recent data)

**Expected outcome:** single-merged dataset (at least **500-1000+ rows** of **country-year**)

**Preprocessing Steps**

* **Data Merging & Alignment**: Convert all datasets to **a common unit of analysis** (e.g., country-year or subnational regions).
* **Feature Engineering**: Create **conflict history**, **economic shocks**, **~~ethnic~~ grievances**, and **climate change indicators**.
* **Handling Missing Data**: Use **imputation techniques (e.g., KNN, mean, multiple imputation)**.
* **Balancing Data**: If conflict cases are rare, use **oversampling (SMOTE) or weighting** to improve prediction accuracy.

**3. Machine Learning Framework**

**Step 1: Feature selection & importance analysis**

* **ML Models for feature importance:** **Random Forest, XGBoost, SHAP values** to rank the most important features influencing ~~ethnic~~ conflict.
  + Justification why these models for feature selection: Based on conflict prediction literature?
* ~~Compare feature importance~~ **~~between Global North & Global South~~**~~.~~
* **Expected outcome:** 10 – 30 highly relevant features (columns)

**Step 2: Conflict Prediction Model**

* **Training Period:** 1990–2017 (80%, might adjust later depends on dataset size)
* **Testing Period:** 2018–2022 (20%, might adjust later depends on dataset size)
  + The output data will later be tested on Step 3: Out-of-Sample Validation.
* **ML Model for prediction:** only **Boosted Regression Tree (BRT)** model (replicates a method of [Modelling armed conflict risk under climate change with machine learning and time-series data - PMC](https://pmc.ncbi.nlm.nih.gov/articles/PMC9123163/#notes3))
  + Robustness check: Cross-validation (just like what the sample article did)
  + More justification? More literature to justify why only one/ only this ML prediction model?
    - Stick to just one model to focus on prediction over comparison.
* **Expected outcome:** country-level, year-by-year predictions.

**Step 3: Model Evaluation & Validation** (to test the performance of the predictive model)

* **Evaluate with Accuracy Metrics:** Precision, Recall, F1 Score, and AUC-ROC (for conflict prediction)
* **Out-of-Sample Validation:** Predict 2018–2022 (testing period) and compare with real conflict data
  + To evaluate the model’s predictive accuracy.
* **~~Global North vs. Global South Comparison:~~** ~~How well do models perform in different regions?~~

**4. Robustness Checks** (ensuring generalizability)

* **Different Training Windows:** Check if results hold using different periods (e.g., train on 1995–2015, test on 2016–2023).
  + To ensure generalizability of model prediction in different time period (the prediction accuracy must not drop, also to ensure that important features are not just influential in certain time range).
* **Alternative Model Specifications:** Try different ML models and ensemble methods.
  + Just Step 1: Feature Selection & Importance Analysis, not Step 2: Conflict Prediction Model as it only uses 1 prediction model (will cross-validate as robustness check instead).
* **Causal Interpretation** (adding transparency to the research findings, explaining how variables influence conflict risk)
  + SHAP values – This provides individual interpretation for each feature and how much it affects ~~ethnic~~ conflict.
    - Example: How much does GDP influence the model's prediction of ~~ethnic~~ conflict.
  + Partial Dependence Plots (PDPs) – This averages out the effects of all other features to show the relationship between a single feature and the predicted outcome (~~ethnic~~ conflict).
    - Example: When averages out other features, does GDP have a linear or non-linear relationship with (if GDP is a direct or an indirect cause of) ~~ethnic~~ conflict.