Pre-Analysis Plan: Analyzing the Mexican Labor Market

1. Introduction:

I am considering three distinct models, each addressing a specific research question(s).

**Model 1: Predict Employment Status (Classification)**

**Objective:** Develop a classification model to predict whether an individual is employed or unemployed based on demographic, educational, and economic factors.

* + Logistic Regression
  + Decision Tree
  + Random Forest
  + SVM (Support Vector Machine)
  + Gradient Boosting (XGBoost, LightGBM)

Research Question:

* What factors are most predictive of employment status in the Mexican labor market?

**Model 2: Predict Income Level (Regression)**

**Objective:** Develop a regression model to understand the factors influencing income levels and track how these factors evolve over time.

* + Lineal Regression
  + Ridge/Lasso Regression
  + Random Forest Regressor
  + Gradient Boosting Regressor

Research Questions:

* How do education, job sector, and demographics influence income levels?
* **How has the importance of education, experience, and other factors in determining income evolved from 1990 to 2020?**
* **Did education become more or less important over time compared to other factors like experience or economic sector?**

**Model 3: Segment Workers Based on Labor Market Profiles (Clustering – Unsupervised Learning)**

**Objective:** Identify distinct worker profiles in the Mexican labor market using clustering techniques, uncovering potential patterns in employment characteristics.

* + K-Means (to segment workers by similar characteristics)
  + Latent Class Analysis (LCA) (to identify job profiles)
  + PCA (Principal Component Analysis) (to reduce dimensions)

Research Question:

* Can we identify distinct worker profiles based on their labor market characteristics?

2. Data Overview

The dataset is sourced from [Banco de México’s Local Labor Market Database](https://www.banxico.org.mx/DataSetsWeb/dataset?ruta=LLM&idioma=en), combined with data from the Mexican Census (INEGI). It spans from 1990 to 2020 and is based on the **Population and Housing Censuses** from 1990, 2000, 2010, and 2020, as well as the **2015 Intercensal Survey**. It does not include the COVID-19 period. To ensure comparability, data from all these years has been standardized across the mentioned periods. Each of this years has around 8,000,000 observations.

The unit of analysis is individuals participating in the labor market.

Key Variables:

- Outcome Variable (Y)

* Employment Status (LLAVE\_SITTRA)
* Income (INGRESO)
* Worker Segments (Clustering Outcome)

- Predictor Variables (X) for reference, but not restricted to.

* **Demographics:** EDAD (Age), LLAVE\_SEXO (Gender), LLAVE\_SITUACONYUGAL (Marital Status)
* **Education:** ESCOLARIDAD\_ACUMULADA (Years of Education), LLAVE\_NIVACAD (Highest Academic Level)
* **Geography:** LLAVE\_ENTIDAD\_RES5A (State of Residence), LLAVE\_MUNICIPIO\_RES5A (Municipality)
* **Socioeconomic status:** , LLAVE\_TAMLOC (Locality Size)
* **Job-related:** LLAVE\_OCUPACION (Occupation), LLAVE\_ACTECONOMICA (Economic Activity), HORAS\_TRABAJADAS(Hours Worked), MERCADO\_TRABAJO\_LOCAL(Local labor market)

Can we track the ids?

Different snapshots 🡪

weight of each year and can we tell if its education more relevant as time passes or not?

**Regressions** / random forest : relevance of coefficients among splits

Relevance of each variable through time