Universidad San Francisco de Quito

Maestría en Ciencia de datos

Fundamentos de Ciencia de Datos

Carla Parra

Abril 2025

CRISP-DM-Phase 1: Understanding of business (I)

- **Objetivo general:** El presente proyecto tiene como objetivo principal desarrollar un modelo predictivo que permita estimar el porcentaje de mujeres que ocupan cargos gerenciales en el sector privado, a partir de diversos indicadores socioeconómicos y laborales
- **Objetivos específicos:**

1. **Preprocesar y transformar los datos.**
2. **Evaluar la relación entre indicadores financieros, de educación y laborales** para predecir la participación femenina en cargos gerenciales en el sector privado en LATAM.
3. **Construir y comparar modelos de regresión multivariada,** como Lasso y Random Forest para identificar los mejores predictores y evaluar su capacidad explicativa y predictiva.
4. **Implementar técnicas de reducción de dimensionalidad (PCA)** para optimizar el desempeño del modelo y facilitar la interpretación de los resultados.

CRISP-DM-Phase 1: Understanding of business (II)

- **Pregunta de investigación:**

¿Qué indicadores socioeconómicos y laborales que se asocian con una mayor representación femenina en roles de liderazgo en el sector privado?

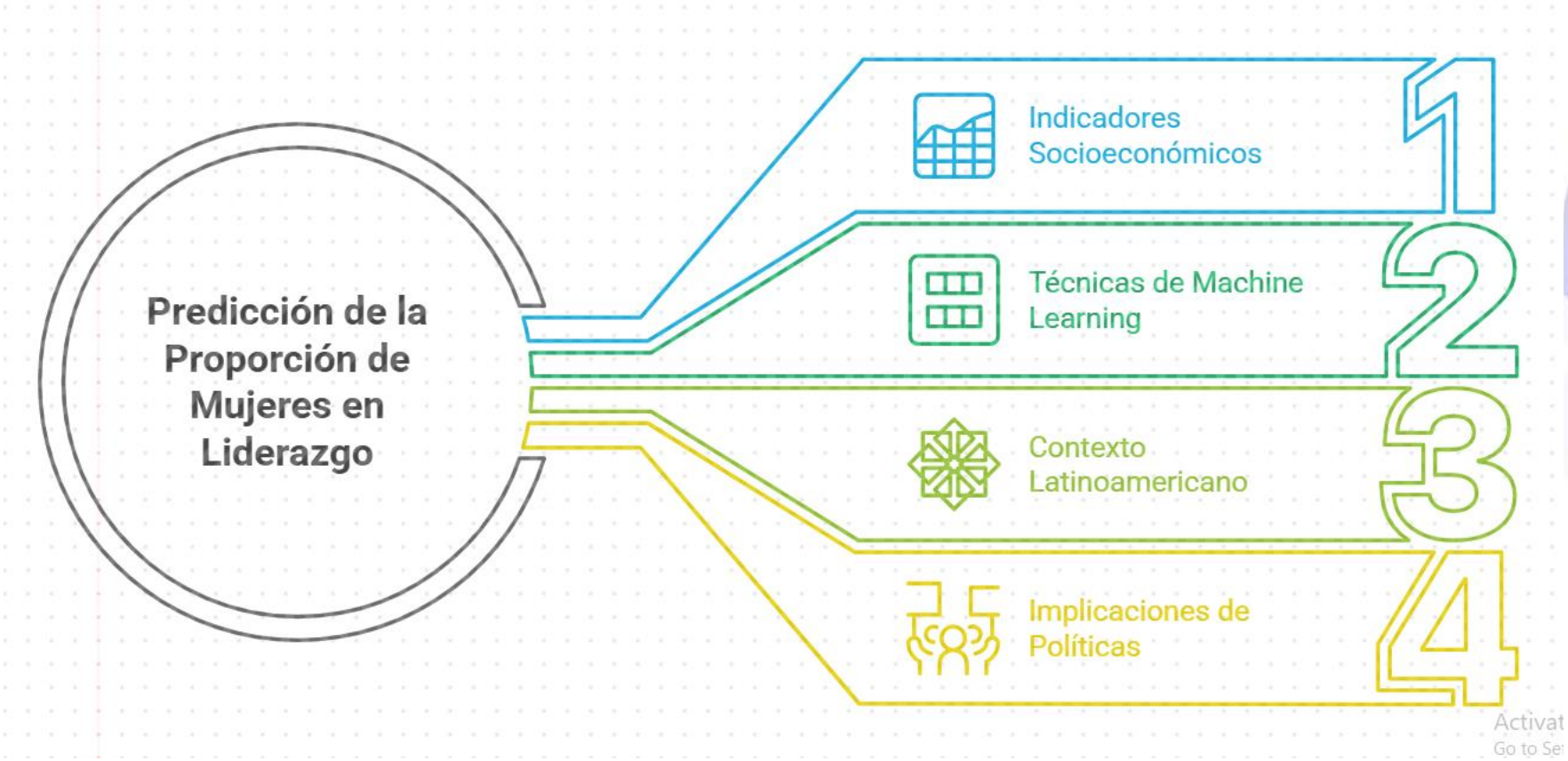
- **Hipótesis principal:**

La participación de mujeres en cargos gerenciales en el sector privado está significativamente influenciada por indicadores de inclusión financiera, nivel educativo y condiciones laborales del país.

- **Hipótesis secundaria:**

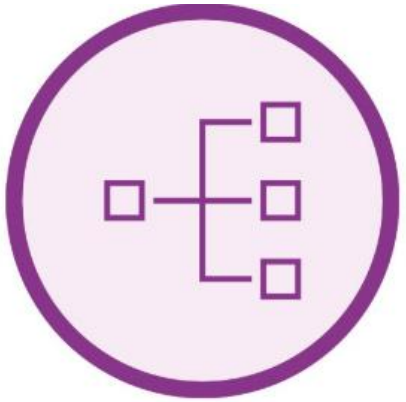
Modelos de regresión regularizada como Lasso pueden superar a modelos basados en árboles (como Random Forest) en la predicción de este indicador cuando las relaciones entre variables son principalmente lineales.

CRISP-DM-Phase 1: Understanding of business -Alcance (III)



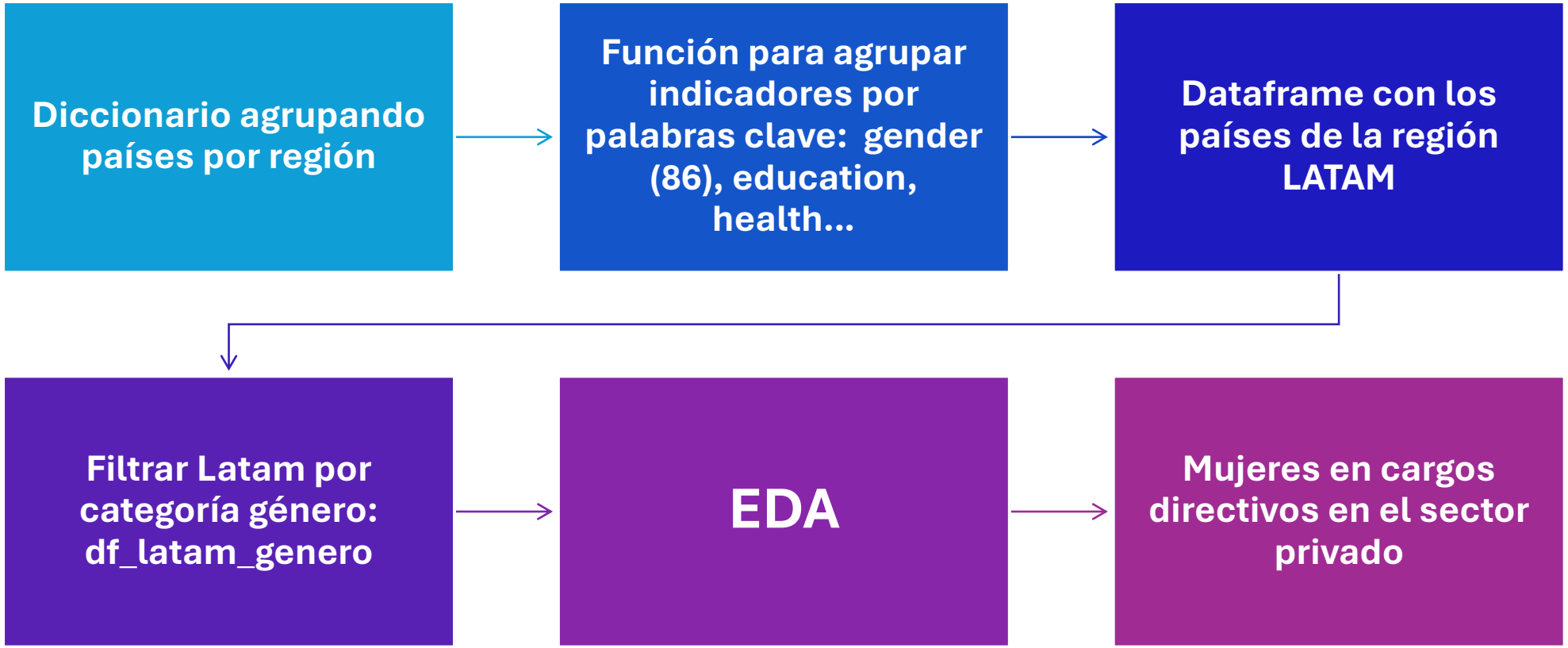
CRISP-DM-Phase 1: Understanding of business - Limitaciones (IV)

Phase 2: Understanding of data

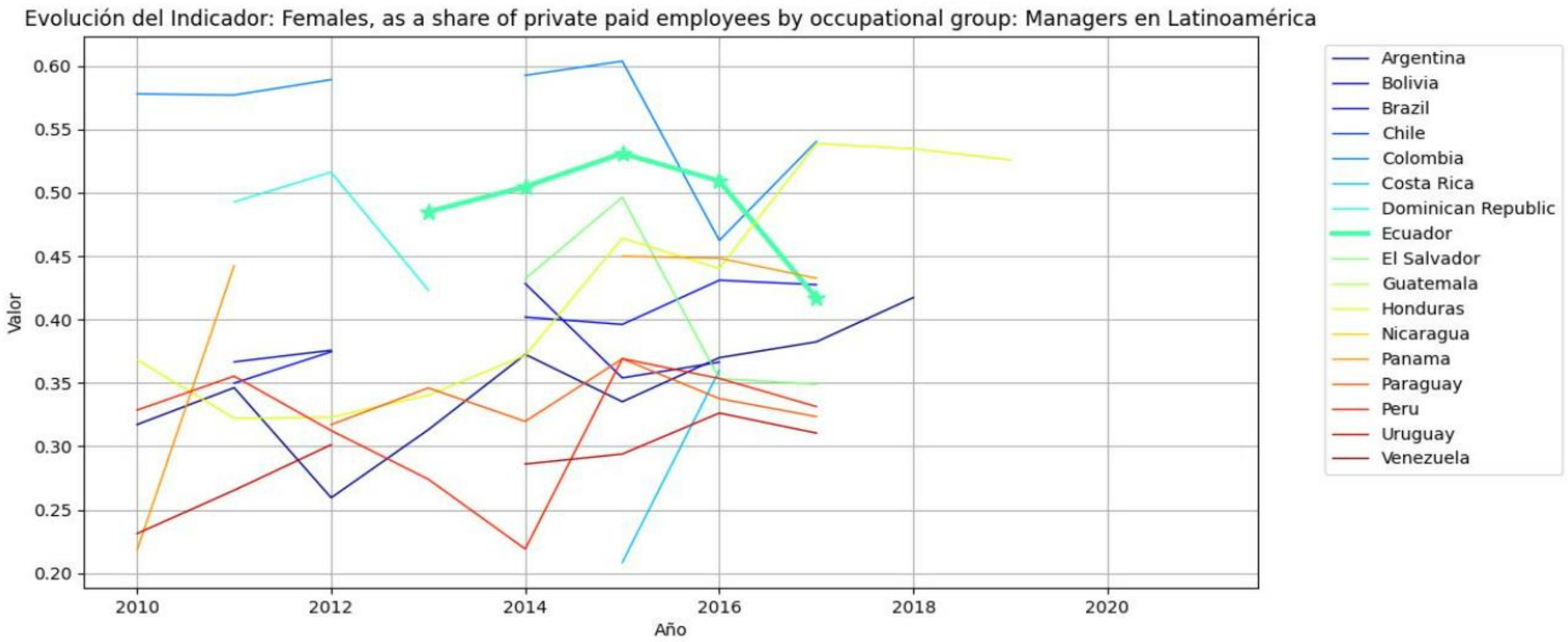


- Fuente principal: **Worldwide Bureaucracy Indicators (WWBI) (Enlace:** https://datacatalog.worldbank.org/search/dataset/0038132)
- Incluye 302 indicadores para 202 economías, construidos a partir de encuestas de hogares, fuerza laboral y datos administrativos. Estos abarcan características demográficas, brechas salariales, equidad de género, y la estructura del gasto público en salarios.
- **Filas (observaciones):** 61,004
- **Columnas (variables):** 27
- **Tipo de datos:**
  - **4 columnas categóricas:**
    - Country Name, Country Code, Indicator Name, Indicator Code
  - **23 columnas numéricas (años del 2000 al 2022)**

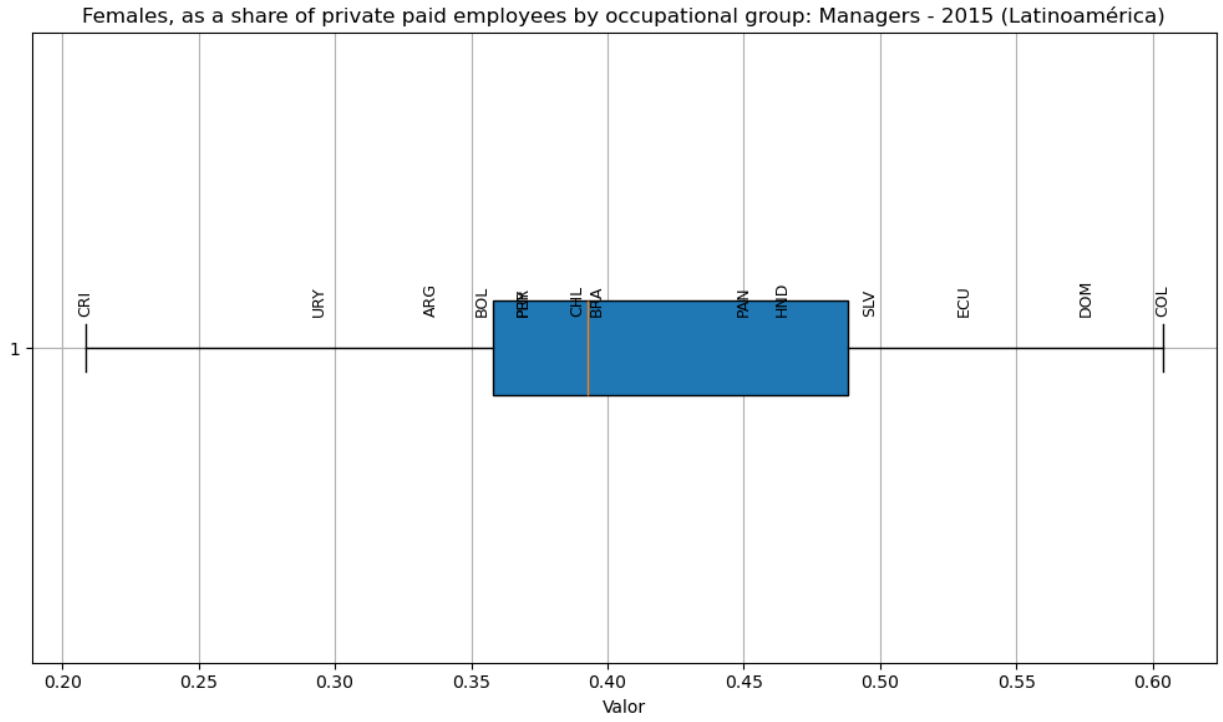
Phase 2: Understanding of data

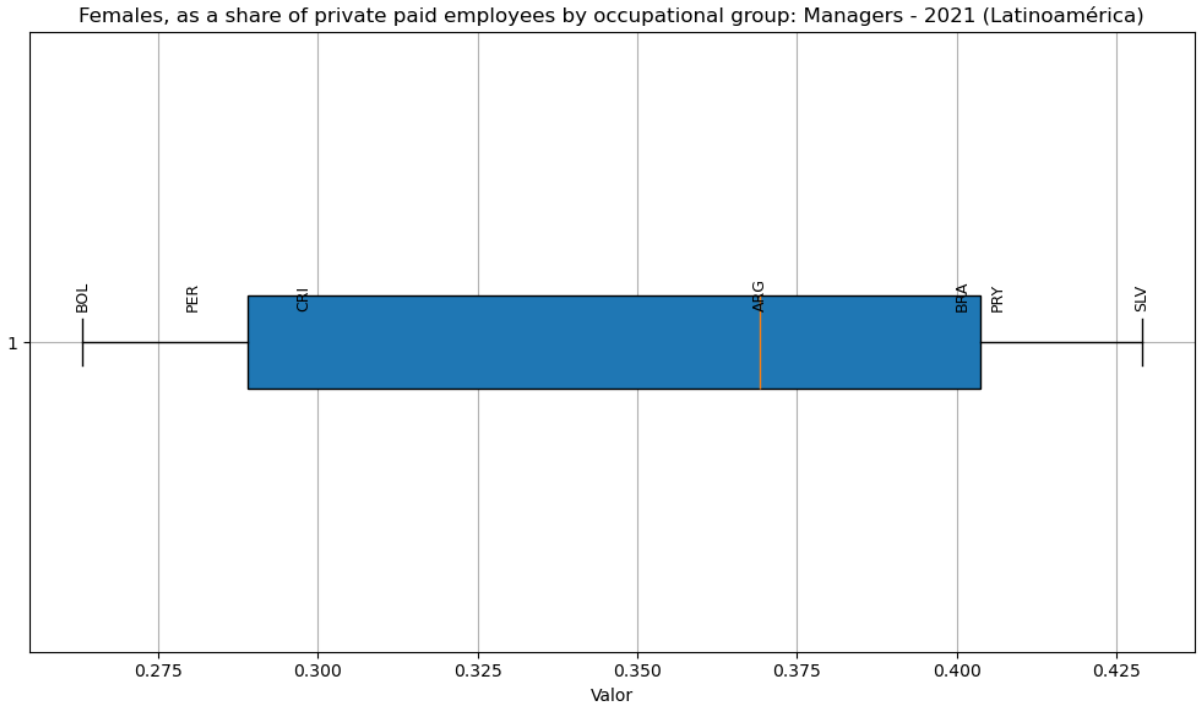


Phase 2: Understanding of data

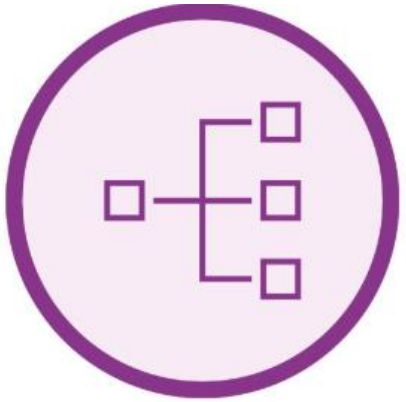


Phase 2: Understanding of data





Phase 3: Data preparation



DataWrangling

Eliminar columnas con más del 70% de valores NaN

Eliminar filas con series de tiempo completamente vacías

Imputación de valores faltantes usando regresión lineal por fila (serie temporal)

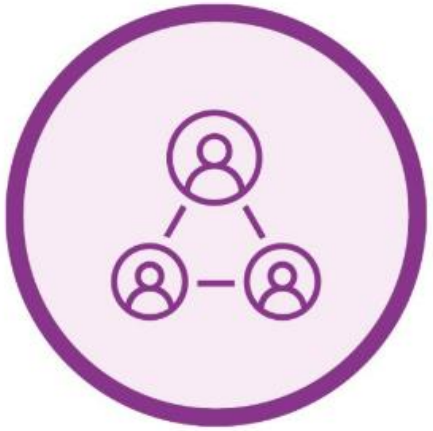
Si la serie tiene al menos dos valores conocidos, se aplica una regresión lineal

Pivotear por país e indicadores para un año específico (dataframes del 2013 al 2017)

Eliminar los p-values (comparaciones entre distribuciones)

Eliminar los datos de Venezuela por tener datos faltantes

Phase 4: Modeling



Split Dataset

Training and test sets created

Feature Selection

SelectKbest
Irrelevant features removed
Polynomial feature (no linear relations)

Standardization

Data normalized using z-score
PCA

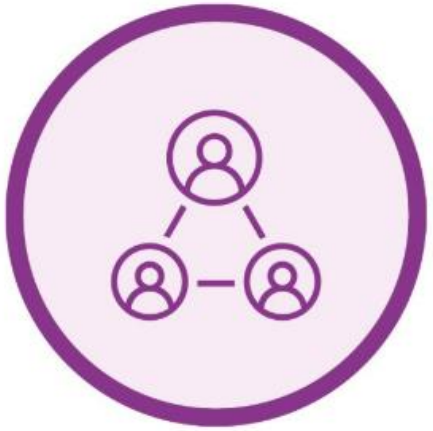
Model Regression

Lasso or Random Forest applied

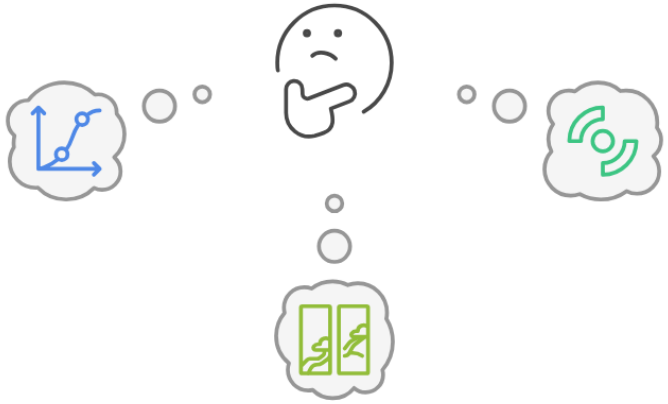
Evaluate Configurations

Performance across setups compared

Phase 4: Modeling



Which pipeline configuration yields the best model performance?



Include Polynomial Features

Expands feature space, potentially capturing non-linear relationships

Include PCA

Reduces dimensionality, potentially improving computational efficiency and reducing overfitting

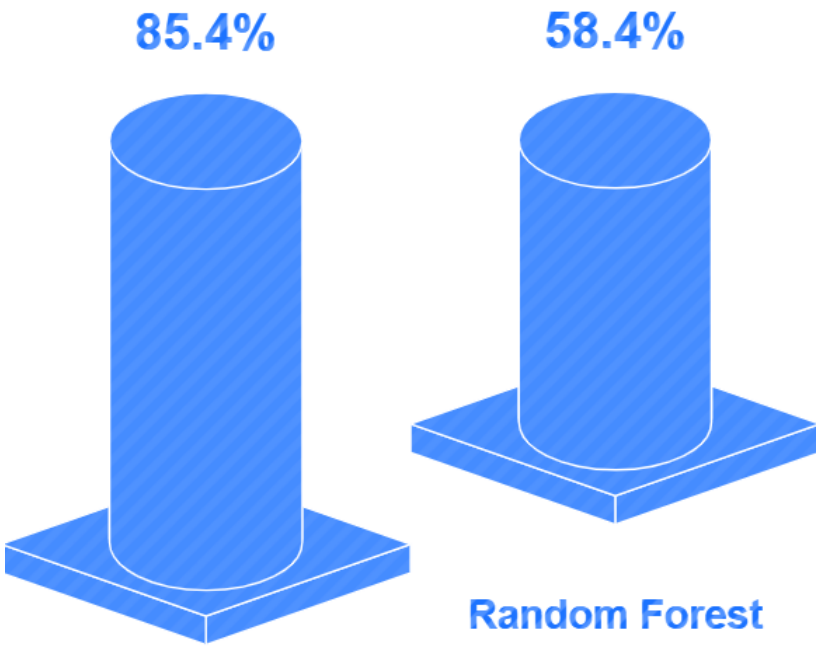
Exclude Both

Maintains original feature space, ensuring simplicity and effectiveness with proper selection

Activate Windows

Phase 5: Evaluation

Comparison of R^2 Values for Lasso and Random Forest



Lasso Regression

Captures underlying data structure effectively

Random Forest

Less effective in explaining data variance

Which model provides more reliable predictions?

VS

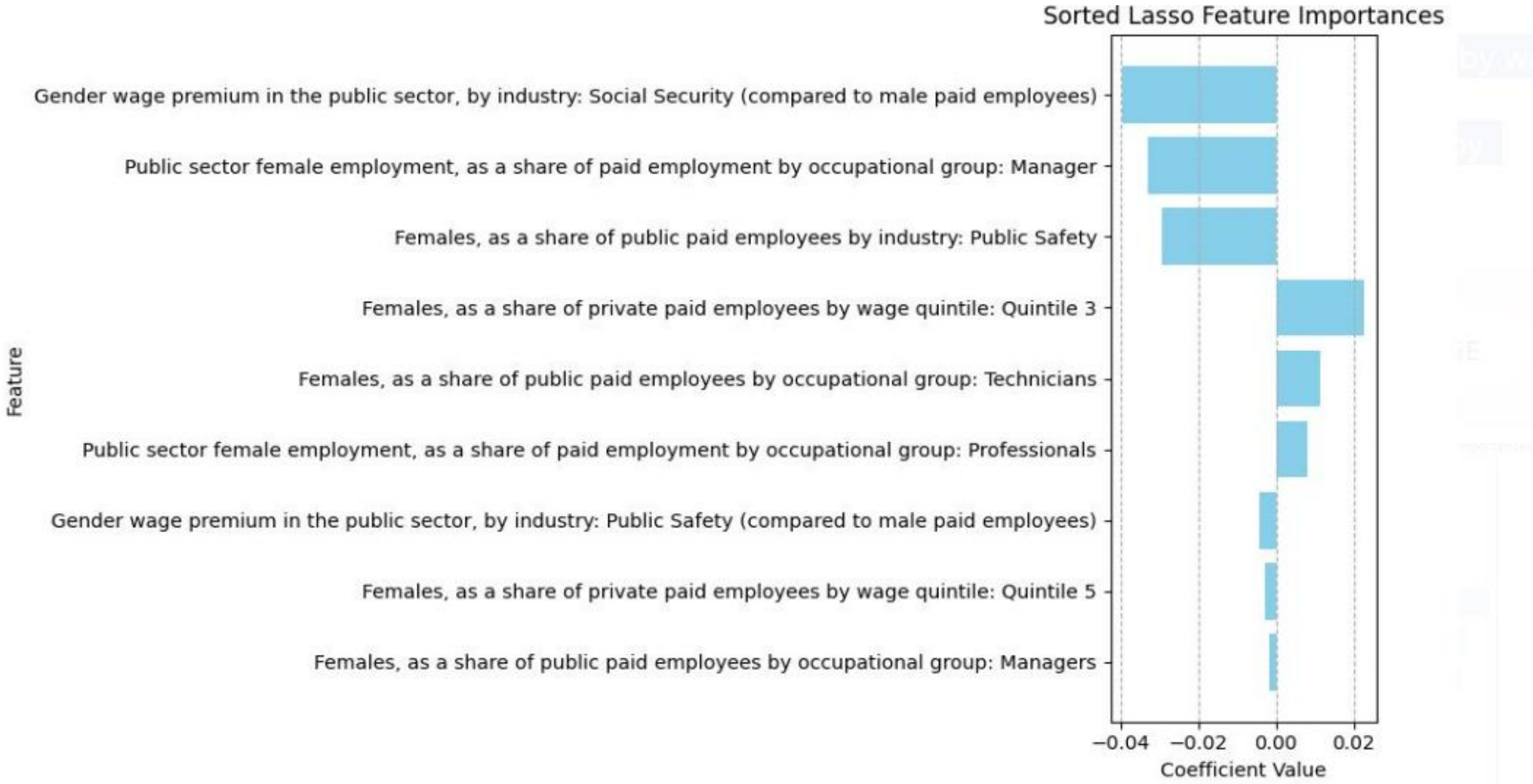
Lasso Regression

Achieves lower MSE and MAE, indicating higher accuracy

Random Forest

Higher MSE and MAE, suggesting less precision

Resultados:


Sorted Lasso Feature Importances
Gender wage premium in the public sector, by industry: Social Security (compared to male paid employees)
Public sector female employment, as a share of paid employment by occupational group: Manager
Females, as a share of public paid employees by industry: Public Safety
Females, as a share of private paid employees by wage quintile: Quintile 3
Females, as a share of public paid employees by occupational group: Technicians
Public sector female employment, as a share of paid employment by occupational group: Professionals
Gender wage premium in the public sector, by industry: Public Safety (compared to male paid employees)
Females, as a share of private paid employees by wage quintile: Quintile 5
Females, as a share of public paid employees by occupational group: Managers
Feature
−0.04
−0.02
0.00
0.02
Coefficient Value

Conclusiones:

1.Las brechas salariales en el sector público afectan al sector privado

→ Donde los hombres ganan mucho más en el sector público, hay menos mujeres como gerentes en el sector privado.

Brecha salarial de género en el sector público: Seguridad Social (-0.039)

Brecha salarial de género en el sector público: Seguridad Pública (-0.004)

2.El sector público atrae a las líderes mujeres

→ Más mujeres en cargos directivos públicos suele significar menos en el sector privado—puede haber competencia entre sectores.c

Mujeres en el quintil 3 de salarios en el sector privado (+0.023)

3.El sector público atrae a las líderes mujeres

→ Más mujeres en cargos directivos públicos suele significar menos en el sector privado—puede haber competencia entre sectores.

Empleo femenino en el sector público como porcentaje de gerentes (-0.033)

Mujeres en cargos de gerencia en el sector público (-0.0015)

Phase 6: Deployment