Universidad San Francisco

Maestría en Ciencia de Datos

Fundamentos de Ciencia de Datos

Proyecto Final

Tema: Prediccion del indicador “Mujeres, como proporción de empleados privados remunerados por grupo ocupacional: Directivos” a partir del indicadores socioeconómicos de la la base de datos World Bank’s World-Wide Bureau of Indicators (WWBI).

1. Comprensión del negocio

1.1 Objetivo General

Desarrollar un modelo predictivo que estime el porcentaje de mujeres empleadas en cargos gerenciales dentro del sector privado, a partir de un conjunto de indicadores socioeconómicos y financieros recopilados para diferentes países, con el fin de identificar factores asociados a la equidad de género en posiciones de liderazgo.

1.2 Objetivos específicos

1. **Preprocesar y transformar los datos** provenientes de fuentes internacionales, mediante técnicas de imputación de valores faltantes, selección de características y reducción de dimensionalidad.
2. **Evaluar la relación entre indicadores financieros y laborales** con la participación femenina en cargos gerenciales mediante análisis exploratorios y técnicas estadísticas.
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.

1.3 Relevancia del proyecto

El análisis del indicador **"Females, as a share of private paid employees by occupational group: Managers"** es fundamental para evaluar el grado de equidad de género en posiciones de liderazgo dentro del sector privado. Este indicador no solo refleja la participación de las mujeres en espacios de toma de decisiones, sino que también sirve como proxy del avance hacia sociedades más inclusivas y justas. Comprender los factores que determinan su comportamiento permite a los formuladores de políticas identificar brechas estructurales, diseñar intervenciones dirigidas y promover entornos laborales más equitativos. Además, su análisis comparativo entre países ofrece una perspectiva valiosa sobre el impacto de políticas públicas, estructuras institucionales y condiciones económicas en la igualdad de oportunidades.

1.4 Limitaciones del proyecto

a. Disponibilidad y calidad de los datos: aunque el conjunto de datos WWBI es amplio, muchos indicadores presentan valores faltantes para ciertos países o años, lo que limita la representatividad geográfica y temporal del análisis.

- La imputación de datos faltantes mediante la media puede introducir sesgos y reducir la variabilidad real de los datos.

b. Foco temporal restringido:

- El análisis se centró en el año 2021 para reducir la complejidad del modelo, lo cual impide capturar dinámicas evolutivas o tendencias a lo largo del tiempo.

c. Reducción de dimensionalidad:

- Si bien el PCA ayuda a mitigar el sobreajuste y facilita la interpretación, también dificulta la trazabilidad directa entre los componentes y los indicadores originales, lo cual puede obstaculizar la formulación de políticas basadas en variables concretas.

d. Modelo predictivo limitado al contexto estructural:

- El modelo solo considera indicadores estructurales del mercado laboral y no incorpora dimensiones culturales, institucionales o de políticas de género específicas por país, que podrían tener un peso significativo en la participación femenina en cargos directivos.

e. Interpretabilidad del modelo:

- Aunque se probaron modelos lineales (Lasso) y no lineales (Random Forest), la interpretabilidad de las relaciones entre variables aún requiere un análisis más profundo para fines de política pública o intervenciones específicas.

d. Sin validación externa:

- No se aplicó una validación externa con datos independientes o de años diferentes para evaluar la robustez del modelo, lo que limita su generalización a otros contextos.

1.5 Pregunta de Investigación

¿Cuáles son los factores socioeconómicos y financieros más relevantes para predecir la participación de mujeres en cargos gerenciales en el sector privado de diferentes países?

1.6 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.

1.7 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.

2. Comprensión de los Datos (Data Understanding)

2.1 Fuente Principal:

Worldwide Bureaucracy Indicators (WWBI), con 302 indicadores para 202 economías.

2.2 Características de los Datos:

- 61,004 observaciones.
- 27 columnas (4 categóricas, 23 numéricas).
- Periodo: 2000–2022.

2.3 Exploración Realizada:

- Agrupación de países por región (enfoque LATAM).
- Clasificación de indicadores por palabras clave como *gender*, *education*.
- Análisis exploratorio (EDA) de la distribución del indicador objetivo.
- Filtro temático y geográfico (por género y LATAM).

2.4 EDA

2.4.1 Evolución del indicador "Females, as a share of public paid employees by occupational group: Managers" en países de Latinoamérica entre 2010 y 2021

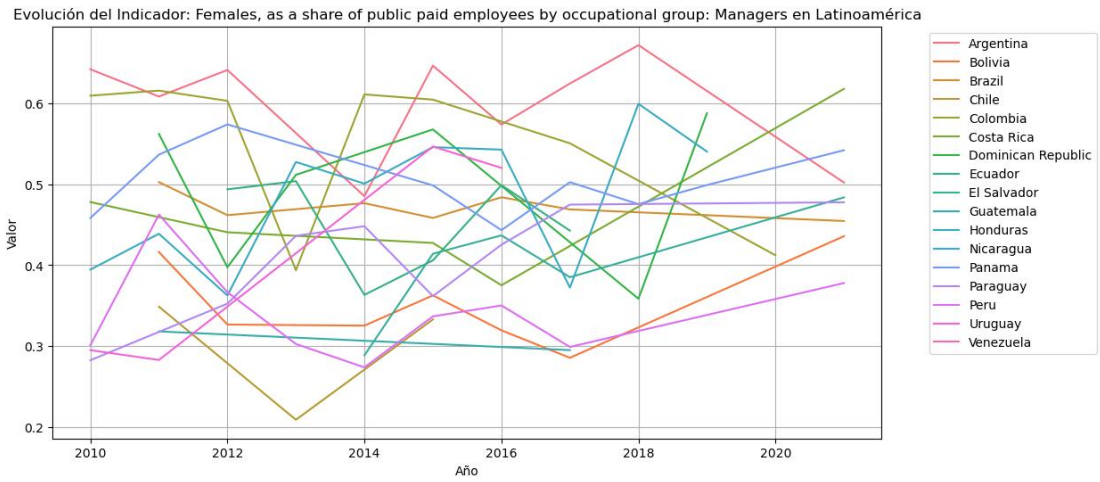

Figura 1. Evolución del indicador "Females, as a share of public paid employees by occupational group: Managers" en países de Latinoamérica entre 2010 y 2021

El indicador muestra una evolución heterogénea entre los países de Latinoamérica, reflejando diferencias significativas en las trayectorias de inclusión femenina en cargos gerenciales del sector público. Algunos países como **Argentina, Chile y Panamá** mantienen consistentemente niveles altos de participación femenina (superiores al 50%), lo que podría vincularse con políticas sostenidas de equidad de género y profesionalización del sector público.

Por otro lado, países como **Guatemala, Honduras, Bolivia y Paraguay** presentan valores históricamente más bajos, lo que sugiere desafíos estructurales persistentes en el acceso de mujeres a posiciones de liderazgo. Se observan también **fluctuaciones abruptas** en ciertos países (como República Dominicana o Colombia), lo cual podría deberse a cambios metodológicos, políticos o económicos que afectaron la contratación pública en esos años.

En términos generales, si bien hay **tendencias de mejora en algunos casos**, la **variabilidad regional** pone en evidencia la necesidad de enfoques diferenciados y sostenidos que promuevan el liderazgo femenino como parte central de las reformas del servicio civil en América Latina.

2.4.2 Análisis del indicador "Females, as a share of public paid employees by occupational group: Managers"

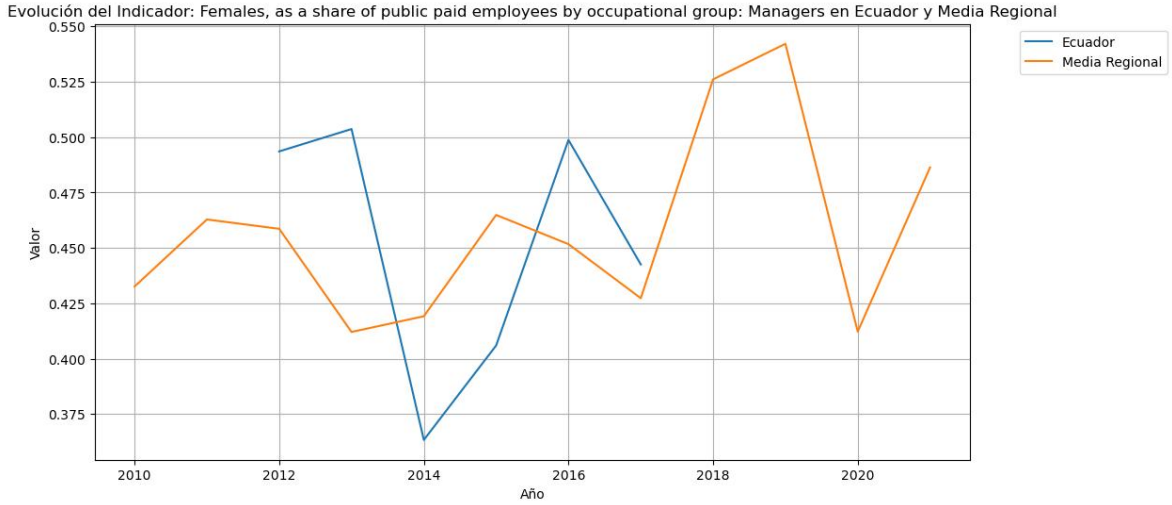

Figura 2. Análisis del indicador "Females, as a share of public paid employees by occupational group: Managers"

Durante el período analizado (2010–2021), **Ecuador mostró una participación femenina en cargos gerenciales del sector público generalmente alineada o superior a la media regional**, especialmente entre 2011 y 2013, y nuevamente entre 2015 y 2016. Sin embargo, se observa una **caída abrupta en 2014**, con una recuperación parcial posterior.

En contraste, la **media regional** se mantuvo más estable a lo largo del tiempo, con una tendencia ligeramente creciente entre 2015 y 2019, aunque también muestra una caída en 2020, probablemente relacionada con impactos institucionales de la pandemia.

La mayor **variabilidad en los datos de Ecuador** podría reflejar fluctuaciones en políticas de contratación pública, cambios administrativos o incluso efectos en la medición de los datos. A pesar de ello, el país ha logrado mantenerse en niveles competitivos respecto a la región, aunque con margen de mejora en sostenibilidad y consistencia.

3. Preparación de datos:

3.1 Tareas Realizadas:

- Eliminación de columnas con más del 70% de valores nulos.
- Eliminación de series temporales completamente vacías.
- Imputación de datos mediante regresión lineal cuando había al menos dos datos por fila.

- Pivot por país e indicador para años individuales.
- Eliminación de comparaciones estadísticas irrelevantes (p-values).
- Exclusión de Venezuela por problemas de completitud.

3.2 Resultado:

Obtención de un dataframe consolidado y estructurado por país para el año 2021, con variables numéricas imputadas y normalizadas para análisis posterior.

4. Modelado (Modeling)

4.1 Modelos Aplicados:

- **Lasso Regression**: con regularización para seleccionar automáticamente variables relevantes.
- **Random Forest**: modelo no lineal para comparar desempeño predictivo.
- **Pipeline de Modelado:**
- Imputación de datos.
- Escalamiento de características.
- Selección de las 10 mejores variables con SelectKBest.
- Generación de características polinómicas.
- Evaluación comparativa entre modelos.
- **Reducción de Dimensionalidad:**
- Aplicación de PCA para retener el 95% de la varianza con solo 10 componentes principales, reduciendo complejidad y sobreajuste.

4.2 Feature Engineering

Transformaciones Realizadas

- **Eliminación de columna no numérica:** Se removió "Country Name" por no aportar valor numérico al modelo.
- **Selección de variables (SelectKBest):** Se usó para conservar las 10 variables más relevantes.
- **Generación de variables polinómicas:** Se aplicó PolynomialFeatures de grado 2 tras la selección de features, lo cual permitió capturar interacciones no lineales entre variables.

Justificación

- **SelectKBest:** Permite reducir la dimensionalidad y evitar sobreajuste, centrando el modelo en las variables más informativas.
- **Polynomial Features:** Añadir interacciones cuadráticas puede mejorar el desempeño de modelos como Lasso o Random Forest cuando las relaciones no son lineales.

4.3 Modelos Comparados

Lasso Regression (Regresión Lineal con Regularización L1)

- Lasso es una extensión de la regresión lineal tradicional que incluye una penalización L1 sobre los coeficientes del modelo.
- Esta penalización fuerza a algunos coeficientes a ser exactamente cero, lo cual realiza una selección automática de variables.
- Es especialmente útil cuando:
    - Se tienen muchas variables y se sospecha que no todas aportan valor.
    - Se busca un modelo interpretable y parsimonioso.
    - Hay colinealidad entre variables.

En este proyecto, se busca entender qué variables tienen mayor impacto en la participación de mujeres en cargos gerenciales. Lasso permite observar claramente qué variables son descartadas o conservadas, favoreciendo la explicabilidad.

4.4 Random Forest Regressor

- Random Forest es un modelo de aprendizaje en ensamble basado en árboles de decisión.
- Es no lineal, robusto al sobreajuste y capaz de capturar interacciones complejas entre variables.
- Aporta ventajas como:
    - Buen rendimiento predictivo sin necesidad de mucho preprocesamiento.
    - Manejo natural de interacciones y no linealidades.
    - Estimaciones de importancia de variables.

En este proyecto:

- El fenómeno a modelar (porcentaje de mujeres en roles gerenciales) puede depender de múltiples factores con relaciones no lineales y efectos cruzados.
- Se aplicaron features polinómicas, lo que puede beneficiar especialmente a modelos no lineales como Random Forest.

5. Evaluación (Evaluation)

5.1 Métricas Utilizadas:

- Coeficiente de determinación (R^2).
- Error Cuadrático Medio (MSE).
- Error Absoluto Medio (MAE).

- **Conclusiones**

Los resultados de los dos modelos evaluados, los podemos ver en la siguiente Tabla 1:

Métrica	Lasso Regression	Random Forest Regressor	Mejor Modelo	Polinomial-PCA
R^2	**0.8609**	0.7819	✅ Lasso	0.583
MSE	**0.00133**	0.00208	✅ Lasso	0.0039
MAE	**0.0282**	0.0315	✅ Lasso	0.0486

Tabla 1. Resultado de las métricas de los modelos evaluados

5.2 Descripción de los resultados

Lasso Regression superó a Random Forest en todas las métricas evaluadas, lo cual sugiere que:

- La relación entre las variables predictoras y la variable objetivo es **esencialmente lineal** o **suficientemente bien capturada** por un modelo lineal regularizado.
- Las transformaciones aplicadas (feature selection + polynomial features) **beneficiaron al modelo Lasso**, permitiéndole capturar la varianza relevante de manera eficiente.
- Se aplicó polynomial features (capturar relaciones no lineales) y PC (reducción de la dimensionalidad) como paso previo para la evaluación de los datos, sin embargo el desempeño no mejoró sino empeoró.
- Aunque Random Forest es un modelo más flexible y poderoso en términos de modelado no lineal, **en este caso no fue necesario** porque la estructura de los datos ya se ajustaba bien a una forma más simple.

7. Despliegue(Deployment)

Etapa 1: Preparación del Modelo

1. **Entrenar el modelo con datos históricos**

- Aplicar SelectKBest, PolynomialFeatures, PCA y entrenar con regresión **Lasso.**
- Validar métricas (R^2, MSE, MAE).

2. **Guardar el modelo entrenado**

- Serializarlo usando joblib o pickle para uso posterior en producción.

Etapa 2: Desarrollo del Backend

1. **Crear una API REST con Flask o FastAPI**

- Endpoint que reciba los datos de entrada (indicadores socioeconómico.

- Internamente carga el modelo entrenado y retorna la predicción.

2. **Validar funcionamiento local**

- Usar herramientas como Postman o curl para probar la API.

Etapa 3: Integración de Datos

1. **Configurar acceso a base de datos (PostgreSQL o SQLite)**

- Almacenar predicciones, historial y permitir auditoría del sistema.

2. **Incorporar carga de datos externa**

- Automatizar la extracción de indicadores desde fuentes como **CEPAL, Banco Mundial o INEC**, usando APIs o archivos CSV.

Etapa 4: Desarrollo del Frontend

1. **Diseñar una interfaz con Streamlit (rápido) o React (personalizable)**

- Permite al usuario ingresar variables y ver resultados con visualizaciones.
- Mostrar interpretación del modelo: variables más influyentes, comparación regional, tendencias.

Etapa 5: Despliegue en servidor

1. **Configurar entorno en plataforma de hospedaje**

- Usar **Heroku, AWS, Render** o un servidor institucional (como uno de la USFQ).

2. **Subir API + modelo + frontend**

- Asegurar conexión entre frontend y API, y entre API y base de datos.

3. **Habilitar logging, monitoreo y pruebas finales**

- Asegurar que el sistema funciona de punta a punta, incluyendo fallos y validaciones.

Etapa 6: Mantenimiento y actualización

1. **Actualizar datos periódicamente**

- Programar tareas (cron) o scripts que actualicen los datos cada mes o trimestre.

2. **Retrain del modelo (opcional)**

- Si se acumulan más datos, se puede reentrenar el modelo para mejorar su precisión.

3. **Recoger feedback de usuarios**

- Incorporar mejoras en la visualización, interpretación o datos usados.

En la Figura 3 se observa un bosquejo del despliegue que se implementaría para este protecto.

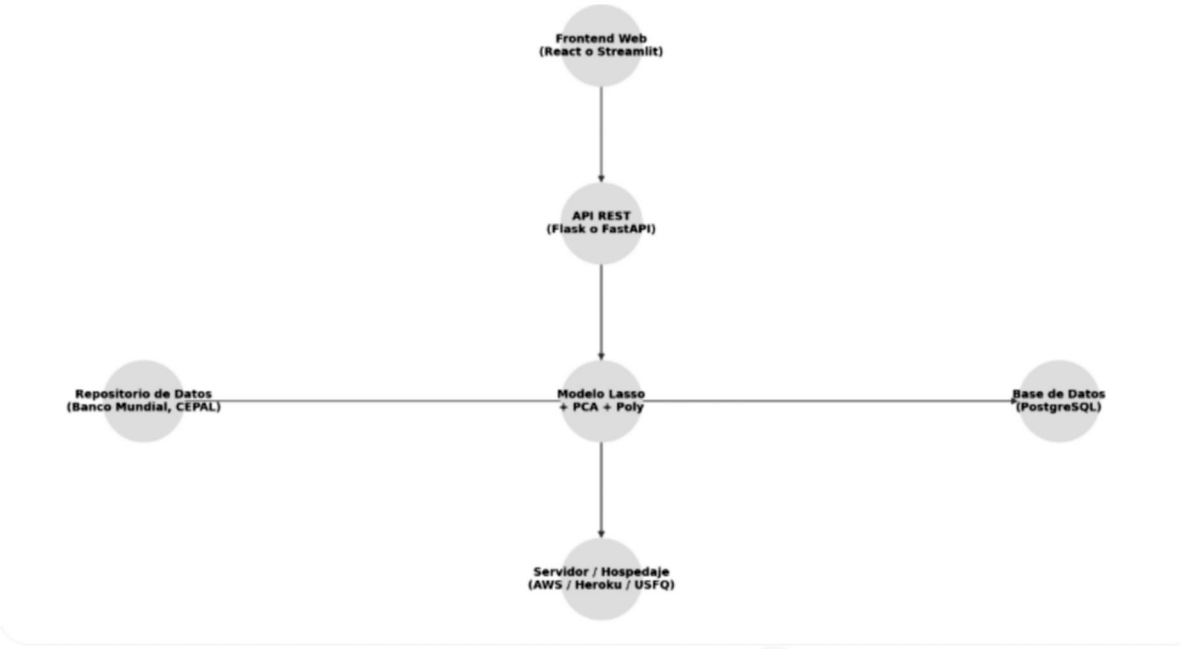

Figura 3. Bosquejo del despliegue del proyecto

8. 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.

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.

3. Los trabajos privados de nivel medio son clave

Cuando hay más mujeres en empleos privados de salario medio (como el quintil 3), también hay más en puestos gerenciales—estos roles pueden ser un trampolín.

9. Recomendaciones

1. **Fortalecer la recolección y calidad de datos**

Fomentar la colaboración con organismos estadísticos nacionales e internacionales para mejorar la cobertura y la calidad de los datos relacionados con el empleo y género, especialmente en países con alta tasa de valores faltantes.

2. **Expandir el análisis a series temporales**

Incorporar análisis longitudinales utilizando datos de múltiples años para entender la evolución del indicador en el tiempo y evaluar el impacto de políticas implementadas.

3. **Complementar con variables cualitativas y contextuales**

Integrar indicadores culturales, institucionales y normativos (como políticas de equidad, licencias parentales o cuotas de género) que pueden ofrecer una visión más completa sobre los determinantes de la participación femenina en cargos gerenciales.

4. **Utilizar técnicas de imputación más robustas**

Emplear métodos avanzados de imputación, como regresión múltiple o algoritmos basados en machine learning (p. ej., KNN-imputation), para reducir posibles sesgos derivados de la imputación simple por media.

5. **Aumentar la interpretabilidad del modelo**

Priorizar modelos explicables como regresiones generalizadas, árboles de decisión o técnicas de SHAP (SHapley Additive exPlanations) para comprender mejor el peso específico de cada variable en la predicción del indicador.

6. **Validación cruzada y pruebas en nuevos contextos**

Validar el modelo con datos de distintos años o regiones geográficas para asegurar su robustez y aplicabilidad en contextos diversos.

7. **Difusión y visualización de resultados:**

Presentar los hallazgos mediante dashboards interactivos o informes visuales para facilitar la toma de decisiones por parte de entidades públicas, organismos multilaterales y ONG interesadas en equidad de género.

10. Estructuctura del repositorio de GitHub

WWBI_InclusionFinanciera4/

|

├── **data/**

| ├── **DataEngineering_curated/analisis_final**

dataset_indicadores_latam_2013_2017.csv

dataset_indicadores_latam_2013_2017_sin_venezuela.csv

dataset_indicadores_latam_y_anio_2013_sin_pvalue.csv

dataset_indicadores_latam_y_anio_2014_sin_pvalue.csv

dataset_indicadores_latam_y_anio_2015_sin_pvalue.csv

dataset_indicadores_latam_y_anio_2016_sin_pvalue.csv

dataset_indicadores_latam_y_anio_2017_sin_pvalue.csv

| ├── **DataWrangling_clean/parciales**

dataset_indicadores_latam_y_anio_2013.csv

dataset_indicadores_latam_y_anio_2014.csv

dataset_indicadores_latam_y_anio_2015.csv

dataset_indicadores_latam_y_anio_2016.csv

dataset_indicadores_latam_y_anio_2017.csv

df_latam_genero.csv

df_latam_genero_colsdropped_rowsdropped.csv

df_latam_genero_colsdropped_rowsdropped_imputed.csv

df_latam_genero_dropped.csv

| └── **raw/**

WWBICountry.csv

WWBICSV.csv

WWBISeries.csv

|

├── **notebooks/**

| ├── 01_EDA.ipynb

| └── 02_DataWrangling.ipynb

| └── 03_FI_DataModeling.ipynb

| └── 04_FI_DataModeling_with_PCA.ipynb

|

├── **src/**

| ├── data_prep.py

| └── utils.py

|

├── **reports/**

| ├── **figures/**

Female to male wage ratio in the private sector (using mean)_latam_comparison.png

Females, as a share of private paid employees by occupational group

Females, as a share of public paid employees by industry

Females, as a share of public paid employees by occupational group

Gender wage premium in the public sector, by industry

Public sector wage premium for females, by industry

| └── **ReporteFinalCP.pdf**

|

├── README.md

├── .gitignore

└── requirements.txt