Universidad San Francisco

Parte 2 del Proyecto: Modelado de Datos y Feature Engineering

1. Objetivo General

Predecir el valor de la variable objetivo: **"Females, as a share of private paid employees by occupational group: Managers"**utilizando las demás columnas del dataset, excepto "Country Name".

2. Feature Engineering

2.1 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.

2.2 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.

3. Model Research (Modeling)

3.1 . Modelos Comparados

3.1.3 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.

3.1.4 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 mucha 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.

2. Análisis

- El modelo **Random Forest** superó ampliamente al modelo Lasso en todas las métricas, demostrando ser capaz de capturar relaciones complejas no lineales.
- El uso de Pipeline asegura reproducibilidad y claridad en el flujo de procesamiento.

4. Conclusiones

Los resultados de los dos modelos evaluados lo podemos ver en la siguiente tabla:

Métrica	Lasso Regression	Random Forest Regressor	Mejor Modelo
R^2	**0.8609**	0.7819	✅ Lasso
MSE	**0.00133**	0.00208	✅ Lasso
MAE	**0.0282**	0.0315	✅ Lasso

Tabla 1. Resultado de las métricas de los modelos evaluados

4.1 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.

- 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.