

Estimación de Duración de Proyectos de Ciencia, Tecnología e Innovación

Generación y Evaluación de Múltiples Modelos de Regresión

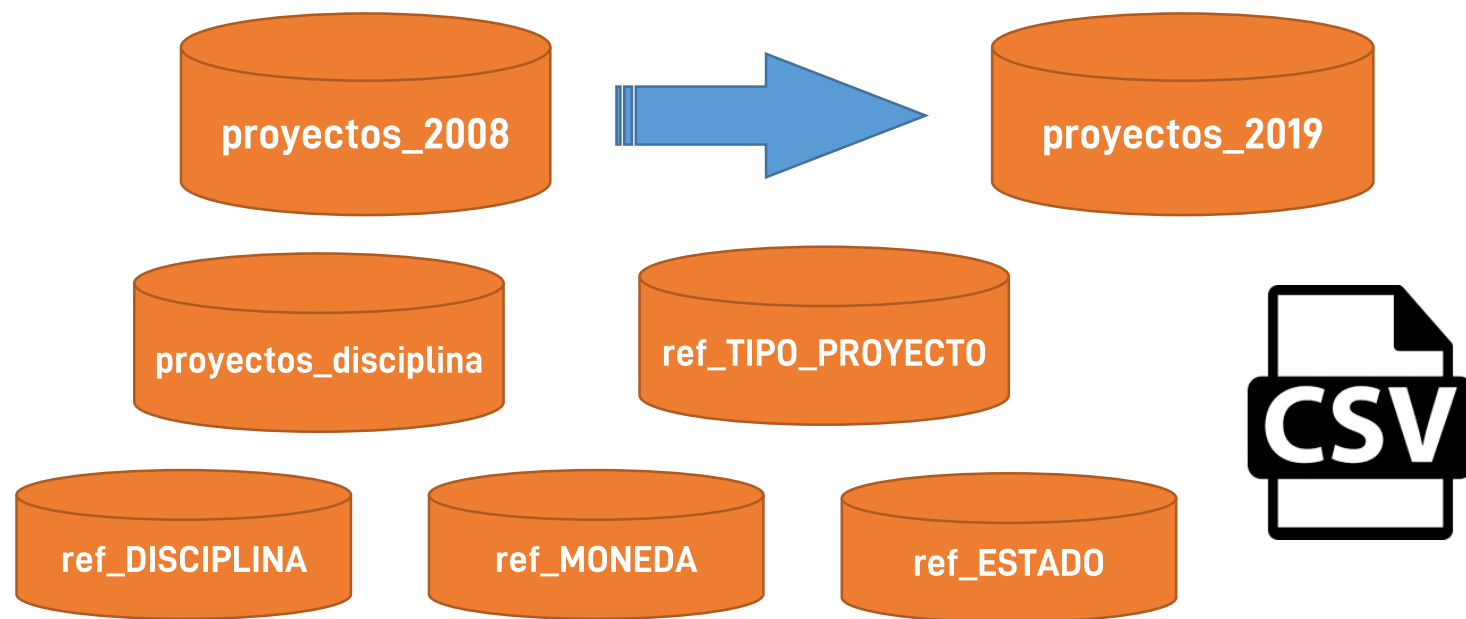
➤ Proyectos de ciencia, tecnología e innovación

Son unidades estadísticas que incluyen los proyectos de I+D y/o de innovación (conjunto de actividades que se llevan a cabo para crear resultados CyT y/o innovativos en un tiempo determinado).

datos.gob.ar



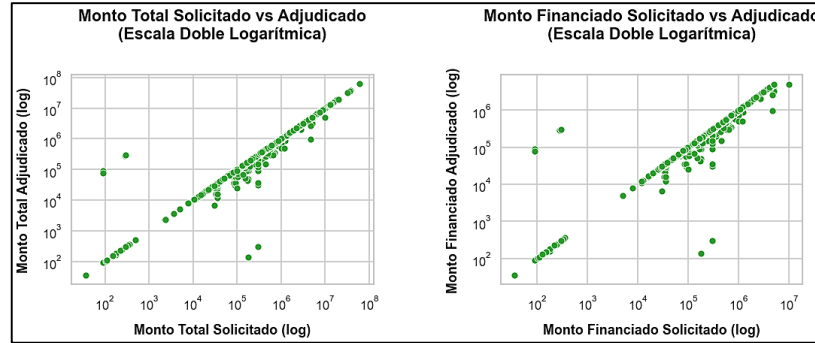
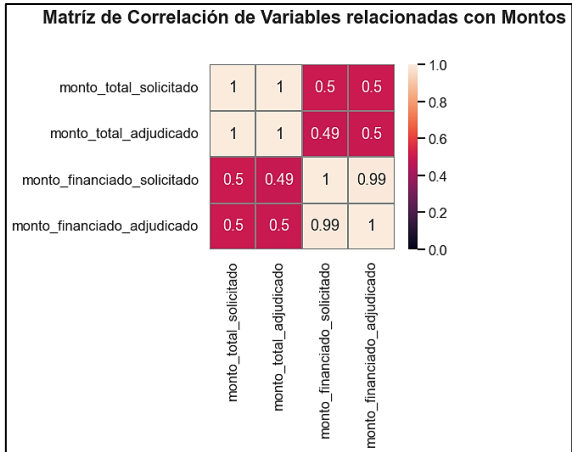
PORTAL DE INFORMACIÓN
DE CIENCIA Y TECNOLOGÍA
ARGENTINO



<https://datos.gob.ar/dataset/mincyt-proyectos-ciencia-tecnologia-e-innovacion>

19 columnas
(2) 'datetime64[ns]', (7) 'float64',
(3) 'int64', (7) 'object'
19.266 registros TOTAL

ANÁLISIS EXPLORATORIO DE DATOS (EDA)



- **Utilizamos Criterio de P99**

- 'duración_días'
- 'año_inicio'

Unificar Tablas de Proyectos

Eliminar Columnas Poco Útiles

Limpieza y Vinculación de Tablas Auxiliares

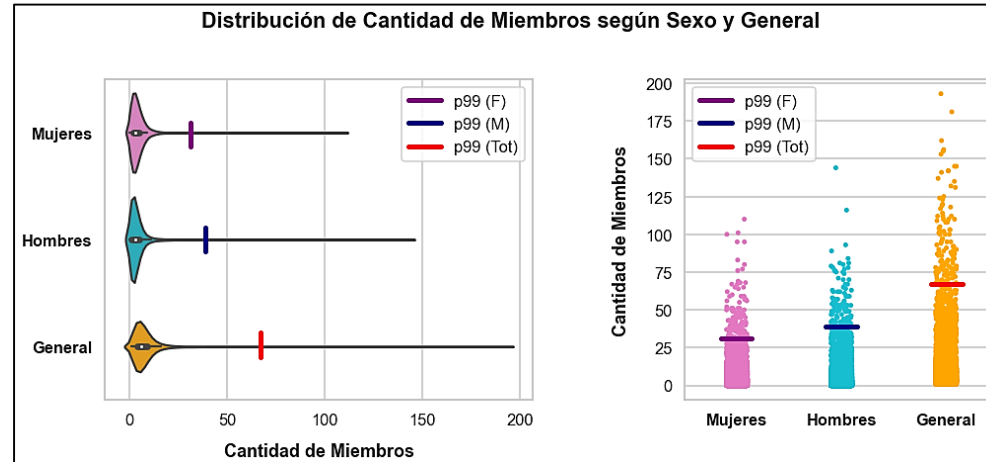
Análisis y Limpieza de NaNs

Análisis y Eliminación de Out-Liers

Generación de Variables Calculadas

Análisis Gráfico y Numérico de Datos

- **Variables Categóricas con Alta Cardinalidad**
- **Variables Poco Informativas**

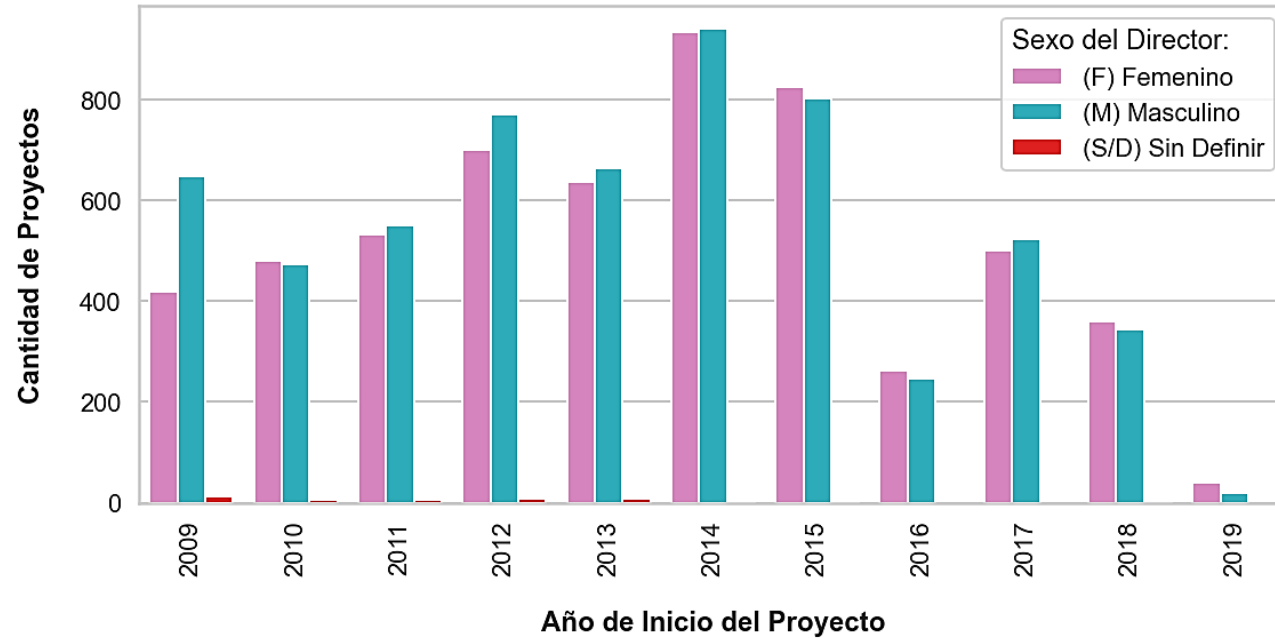


Pipeline

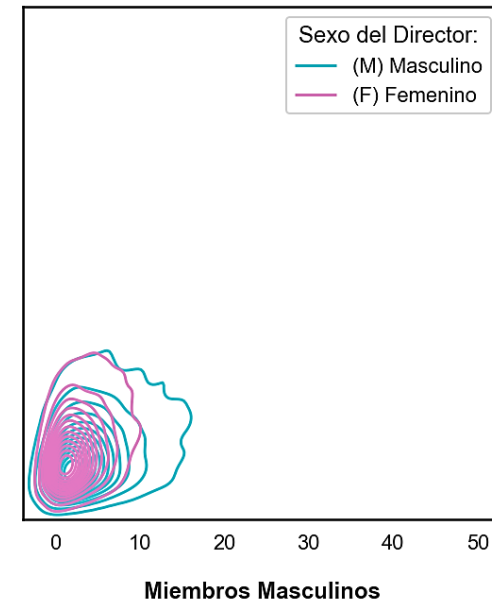
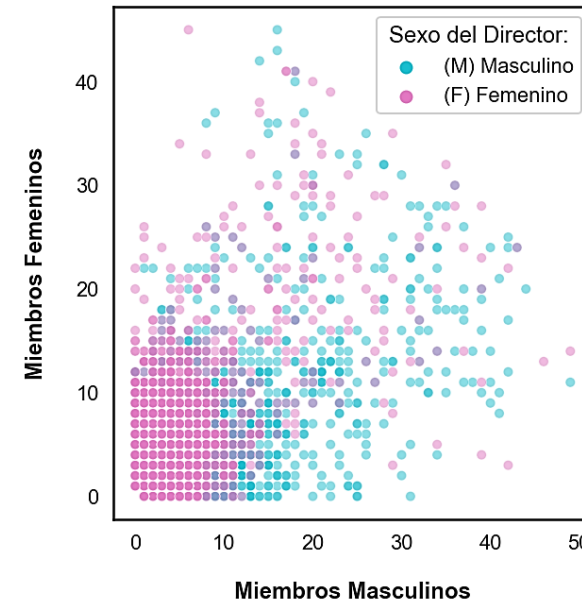
ANÁLISIS EXPLORATORIO DE DATOS (EDA)

➤ Análisis Gráfico

Cantidad de Proyectos por Año de Inicio según Sexo del Director

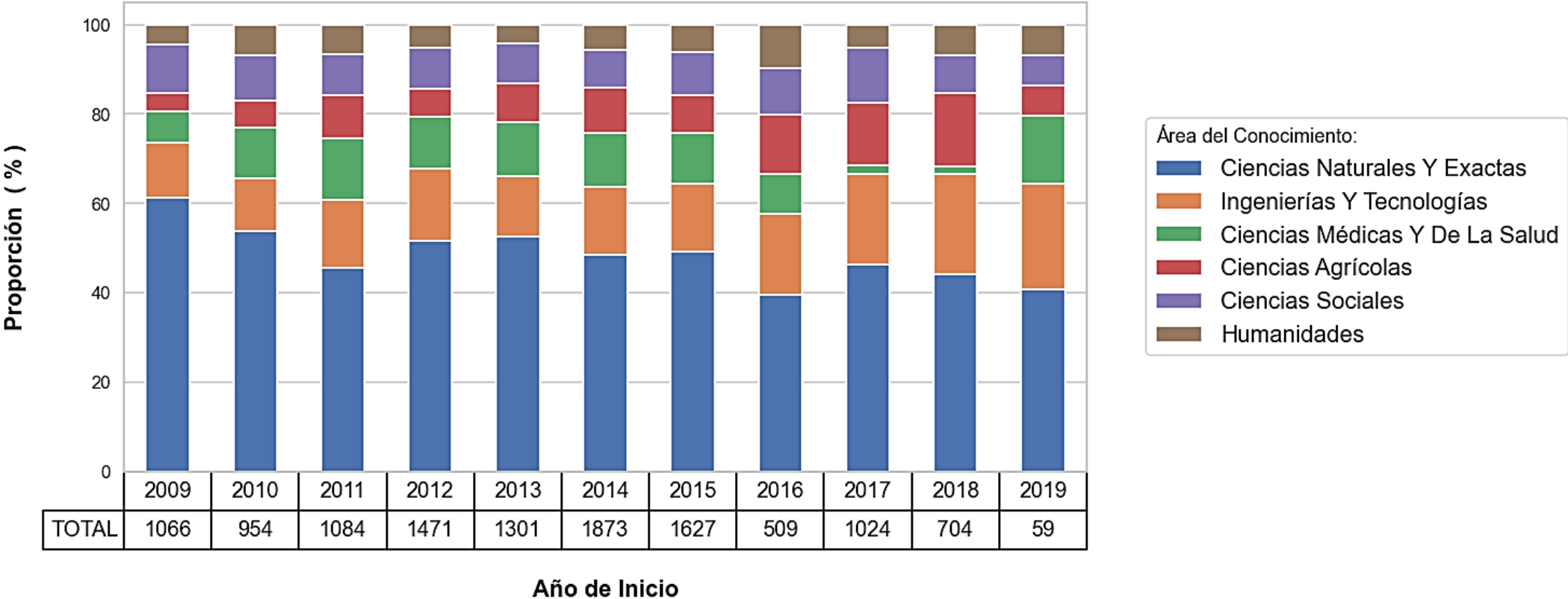


Análisis de Preferencia de Miembros por su Sexo.
(En función del Sexo del Director del Proyecto)



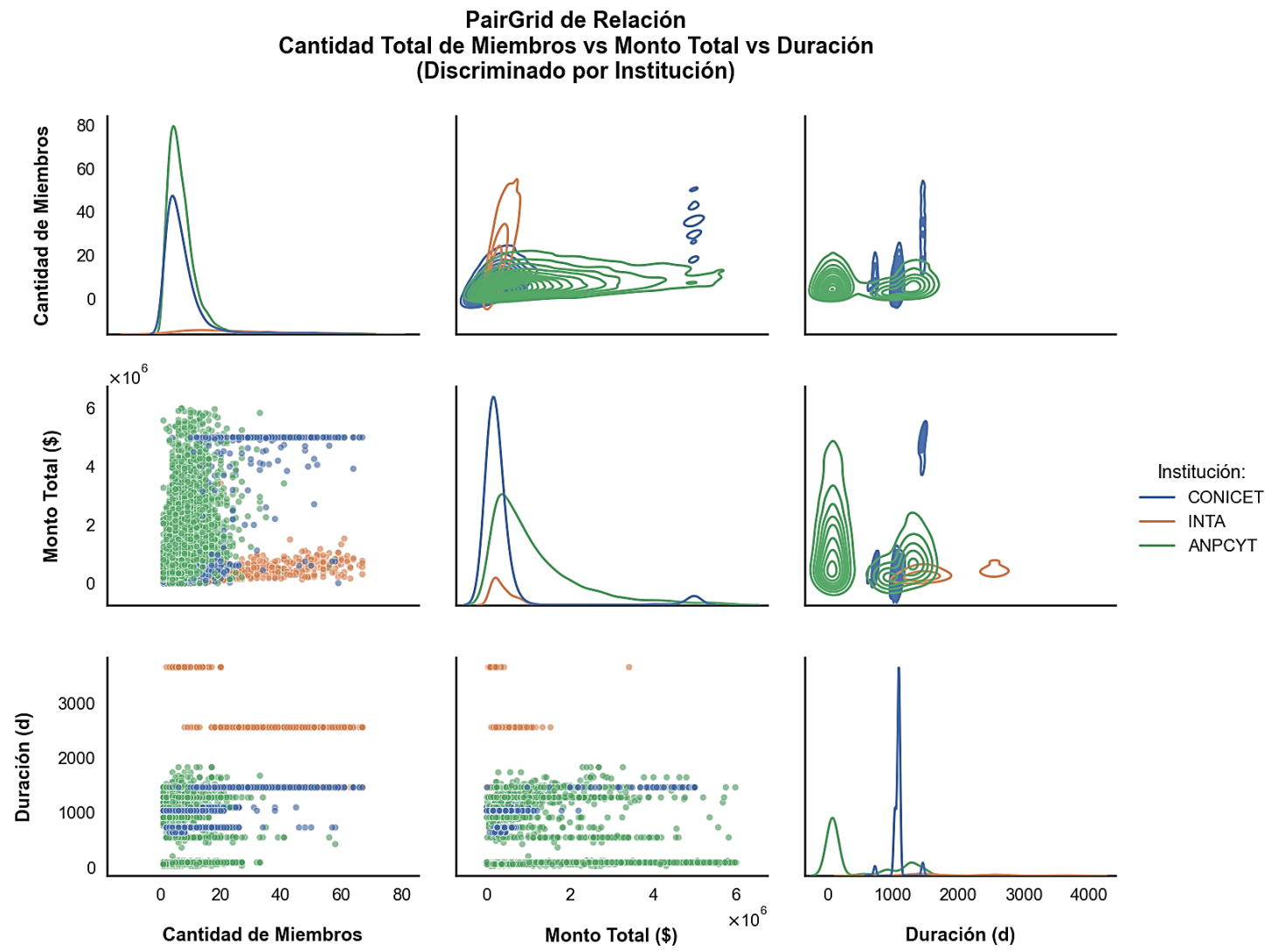
➤ Análisis Gráfico

Proporción de Proyectos de acuerdo a Área del Conocimiento y Año de Inicio

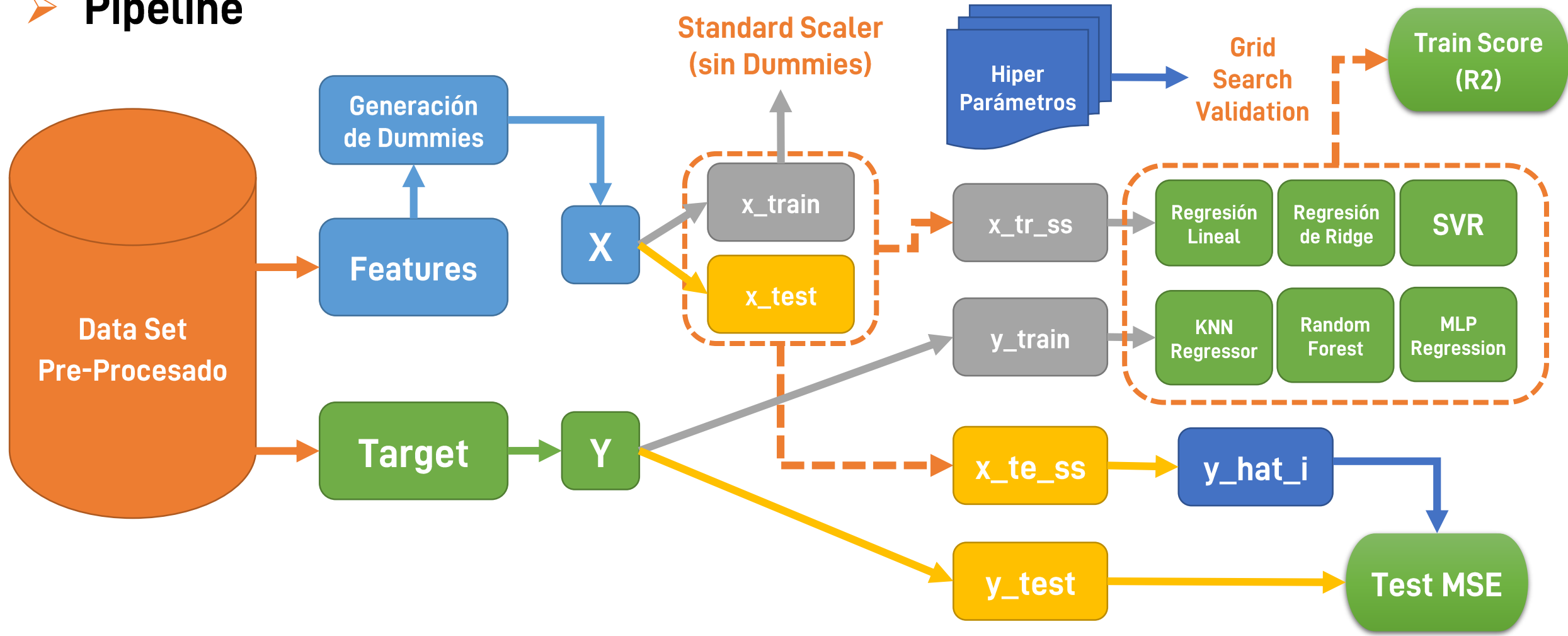


ANÁLISIS EXPLORATORIO DE DATOS (EDA)

➤ **Análisis Gráfico**

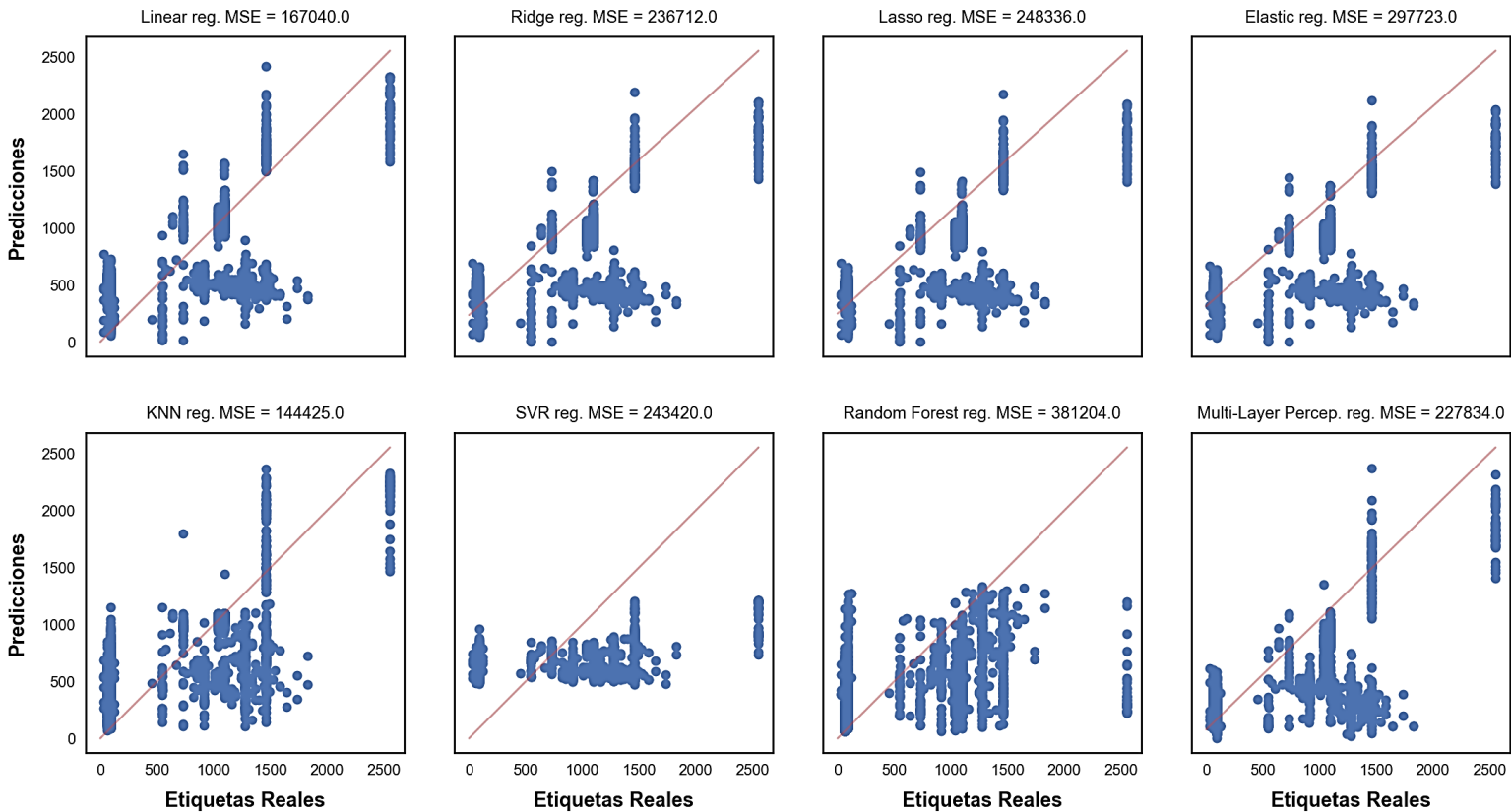


➤ Pipeline



Resultados

Visualización de Predicciones de Cada Modelo Propuesto



	Train Score (R2)	Test MSE
Linear Regression	0.461494	167040.0
Ridge Regression	0.461409	236711.0
Lasso Regression	0.461444	248336.0
Elastic Regression	0.461396	297722.0
KNN Regression	0.541587	144424.0
SVR Regression	0.387586	243420.0
Random Forest Regression	0.637317	381204.0
Multi-Layer Percep. Regression	0.558808	227833.0

CONCLUSIONES

- ✓ Ninguno de los modelos ha conseguido generar una predicción que esté alineada con la recta de correspondencia con los valores reales.
- ✓ Contrastar la performance de un modelo en train con su performance en test evidencia si un modelo consigue generalizar o no. **No siempre el modelo con mejor performance en train tiene capacidad de generalizar en test.**
- ✓ Esto se puede deber, o bien a que **el caso en estudio es demasiado complejo** para los modelos propuestos, **o no contamos con suficientes variables / datos informativos** para conseguir un mejor resultado.
- ✓ De tener que optar por uno de los modelos propuestos más arriba, nos quedaremos con KNN Regressor, pero lo ideal sería continuar trabajando con los datos y proponer un modelo que pueda abordar mejor el problema.

¡ MUCHAS GRACIAS !