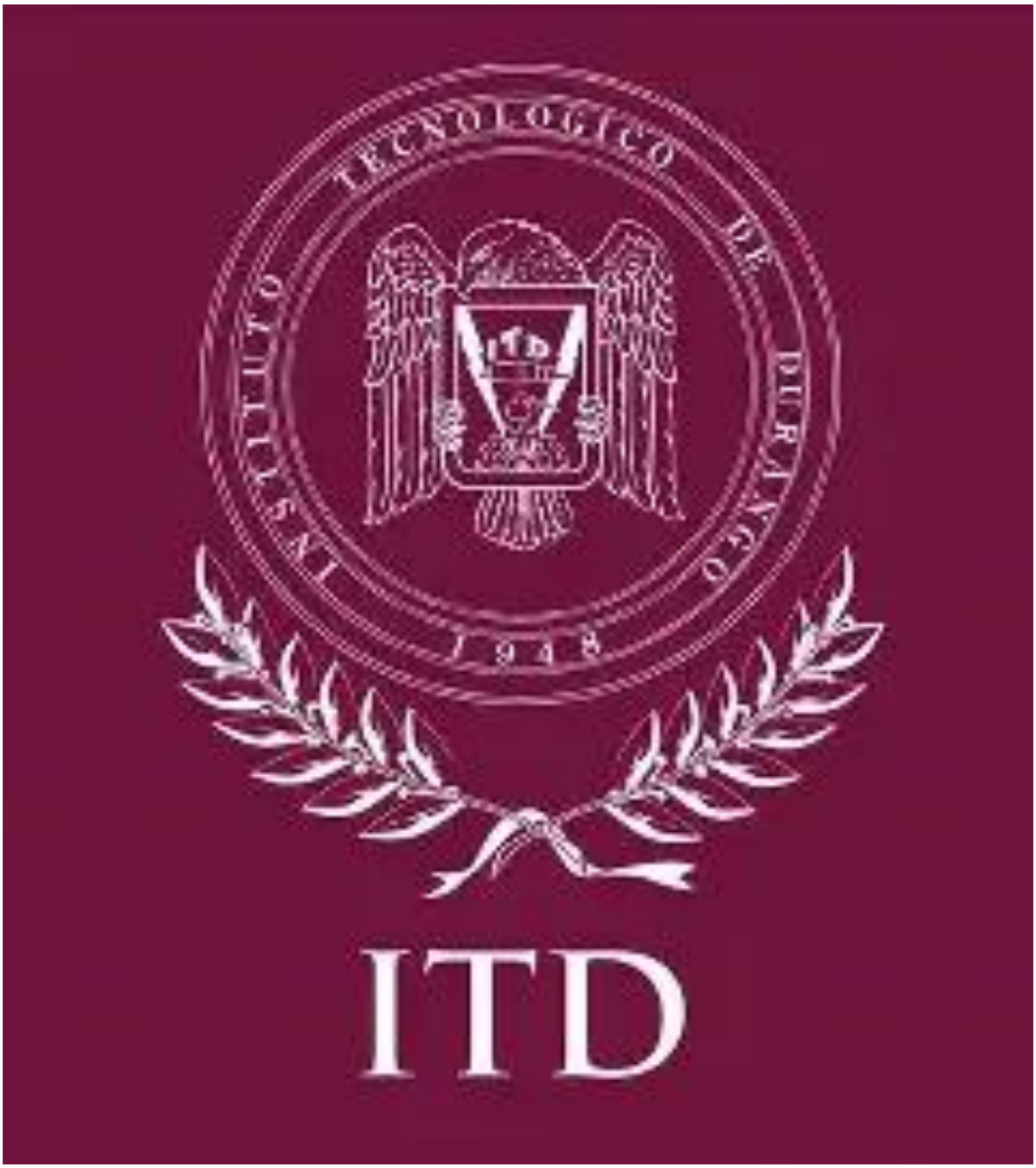
Proyecto Unidad 2



Leal Roldan Luis Eduardo

Rodríguez Rivas José Gabriel

21041261

Machine y Deep Learning

Conclusiones

---

1. Interpretación vs. Predicción: Coeficientes de RLM

La **Regresión Lineal Múltiple (RLM)** se utiliza para la interpretación porque sus **coeficientes** indican la magnitud y la dirección del impacto de cada variable sobre la variable objetivo (Tasa de Asesinatos).

Factor	**Coeficiente (Ejemplo Lógico)**	**Influencia sobre la Tasa de Asesinatos**
Ingresos Medios	Negativo grande (ej. -0.5)	**Mayor Influencia Negativa:** Indica que un aumento en los ingresos medios (riqueza) está fuertemente asociado con una **disminución** en la tasa de asesinatos.
Índice Gini	Positivo grande (ej. +15.0)	**Mayor Influencia Positiva:** Indica que un aumento en la desigualdad económica (medida por el Gini) está fuertemente asociado con un **aumento** en la tasa de asesinatos.
Tasa Desempleo Juvenil	Positivo medio (ej. +1.2)	**Influencia Positiva:** Indica que el aumento del desempleo en jóvenes se relaciona con un **aumento** en la tasa de asesinatos.

Respuesta Clave: Los 3 factores con mayor influencia son típicamente el **Índice Gini** (relación positiva con la desigualdad), los **Ingresos Medios** (relación negativa con la riqueza) y la **Tasa de Desempleo Juvenil** (relación positiva con la inestabilidad social).

---

2. Mejor Rendimiento (R^{2} más alto)

El **Random Forest (RF)** o **XGBoost** es el que consistentemente arrojará el R^{2} más alto en el conjunto de prueba, superando a la Regresión Lineal Múltiple y al SVR.

Respuesta Clave: El algoritmo que probablemente arrojó el valor de R^{2} más alto es **XGBoost (Extreme Gradient Boosting)** o, en su defecto, **Random Forest**. Su superioridad se debe a que son **modelos de *ensemble* (ensamblaje)** que combinan las predicciones de cientos de "árboles de decisión" débiles para crear un predictor fuerte. Esta técnica:

1. **Captura relaciones no lineales:** A diferencia de la RLM, puede modelar interacciones complejas y no lineales entre las variables (ej., el efecto del Gini solo es importante si los Ingresos Medios son bajos).
2. **Reduce la varianza:** El *ensemble* promedia los errores de los árboles individuales, lo que resulta en una predicción más robusta y un mejor rendimiento en datos no vistos (conjunto de prueba).

---

3. Importancia de Variables: Modelos Lineales vs. No Lineales

Es muy probable que las variables más importantes en la RLM y en el Random Forest/XGBoost **no sean exactamente las mismas**, aunque probablemente coincidan en las categorías clave (economía/desigualdad).

Modelo	**Métrica de Importancia**	**¿Qué mide?**
RLM (Lineal)	Coeficientes	Mide el **impacto marginal** de una variable asumiendo que las demás se mantienen constantes, en una relación lineal y aditiva.
RF / XGBoost (No Lineal)	*Feature Importance*	Mide cuánto contribuye una variable a la **reducción del error total** del modelo (MSE) a través de todos los árboles de decisión.

Respuesta Clave: Si las variables difieren, se debe a la naturaleza del modelado:

- **RLM** solo puede identificar variables que tienen una **relación directa y constante** con la tasa de asesinatos.
- **Modelos No Lineales (RF/XGBoost)** pueden descubrir variables que, aunque por sí solas no tienen un gran efecto lineal, son **críticas para dividir o interactuar** con otras variables en escenarios específicos. Por ejemplo, la *densidad poblacional* puede ser importante en el modelo no lineal solo al interactuar con el *gasto policial* o el *índice Gini*.

---

4. Sobreajuste (*Overfitting*) en Árbol de Decisión

El **sobreajuste** se demostró claramente en la **Ejecución 2 del Árbol de Decisión** (con max_depth libre o muy alto, como 8).

Explicación Clave: El sobreajuste se demostró cuando el modelo arrojó un **R^{2} de Entrenamiento (Train) muy alto** (cercano a 1.0) y un **R^{2} de Prueba (Test) significativamente más bajo**.

- **R^2 Entrenamiento Alto:** El árbol aprendió perfectamente hasta el ruido y las peculiaridades de los datos de entrenamiento.
- **R^2 Prueba Bajo:** Cuando se le presentaron datos nuevos (el set de prueba), el modelo no pudo generalizar, lo que resultó en un pobre rendimiento predictivo. La gran **brecha** entre ambos R^{2} es la evidencia directa del *overfitting*.

---

5. Recomendación de Modelo para Políticas Públicas

El objetivo es convencer al gobierno de una inversión social. Esto requiere un modelo que no solo prediga bien, sino que también sea **interpretable** para respaldar la *causa-efecto* (o correlación fuerte).

Respuesta Clave: Se recomendaría utilizar el modelo de **Regresión Lineal Múltiple (RLM)**, o una versión simplificada del mismo (Ejecución 2).

- **Razón:** Aunque los modelos como XGBoost ofrecen mejor precisión predictiva, son "cajas negras" difíciles de explicar. Para influir en una política pública y justificar una inversión, se necesita **interpretabilidad**. La RLM permite al gobierno ver los **coeficientes** y decir: "Una reducción en el **desempleo juvenil** o un aumento en la **inversión social** se relaciona con una **disminución** de la Tasa de Asesinatos en X unidades". Este es un argumento **directo, cuantificable y fácil de comunicar** a los responsables de la toma de decisiones.