Estimación del tiempo restante de explotación de yacimientos de turba en Tierra del Fuego mediante Aprendizaje Automático.

ENTREGA 3

Presentación del Modelo
y Análisis de Resultados

Cátedra: Aprendizaje Automático – 1C 2025
Docente: Martín Mirabete

Alumno
Sergio Andrés SANCHEZ

Presentación de Objetivo del Proyecto

Desarrollar un modelo de Aprendizaje Automático que permita estimar el tiempo de vida restante de explotación de un yacimiento de turba en Tierra del Fuego, promoviendo una gestión sostenible del recurso.

El problema que se busca resolver es de regresión, ya que la variable objetivo será una estimación continua del tiempo (en meses, años o décadas) restante hasta que se agote el yacimiento, en función del volumen total estimado de turba y del ritmo de extracción y comercialización histórico.

Importancia y Relevancia del problema abordado.

La extracción de turba en la provincia de Tierra del Fuego representa una actividad productiva clave, especialmente en zonas rurales. La turba es un recurso natural no renovable en escalas humanas de tiempo y su extracción está condicionada por factores climáticos, estacionales y económicos.

Con una estacionalidad marcada (más actividad en verano que en invierno), la explotación turbera requiere una planificación precisa para evitar la sobreexplotación y promover la sostenibilidad del recurso. Actualmente, las decisiones sobre tiempos de explotación se basan en cálculos aproximados o estimaciones estáticas. Un modelo predictivo que estime cuánto tiempo resta de explotación puede apoyar políticas públicas, decisiones empresariales y controles ambientales.

Repositorio Git:

Todo el proyecto se encuentra en un repositorio Git de acceso público:

https://github.com/SanchezSergioAndres/EstimacionTiempoRestanteExplotacionTurbaTDF

Para la organización del mismo se utilizó la herramienta CookieCutter.

Descripción del Dataset

Origen: El Dataset proviene de una base de datos en Postgre de la Dirección General de Desarrollo Minero dependiente del Ministerio de Producción y Ambiente del Gobierno de la Provincia de Tierra del Fuego, Antártida e Islas del Atlántico Sur.

Fecha de disponibilidad del Dataset: Viernes 06 de junio de 2025 – Segunda Versión Dataset.

El Dataset provisto contempla expedientes tramitados por explotación de Turba Rubia y Negra por cada uno de los yacimientos desde enero de 2020 a diciembre de 2024.

Este Dataset cuenta con 1635 filas y 27 columnas con las siguientes características:

Variable	***Tipo***	*Definición*
Area	Numérica	Superficie del yacimiento, en hectáreas
Vol_N	Numérica	Volumen de turba negra a 2023
Vol_R	Numérica	Volumen de turba rubia a 2023
Vol_total	Numérica	Volumen total de turba a 2023
Fecha	Numérica	Año de cuantificación (2023)
id	Numérica	valor de identificación (no tener en cuenta)
*_uid_*	Numérica	valor de identificación (no tener en cuenta)
poligono	Numérica	número del polígono
id_2	Numérica	valor de identificación (no tener en cuenta)
producto	Categórica	tipo de mineral extraído
num_expte	Numérica	número de expediente de regalías mineras
ano	Numérica	año de declaración jurada
mes	Numérica	mes de declaración jurada
volumen_produccion	Numérica	volumen de mineral extraído
trimestre	Numérica	trimestre de la declaración jurada

regalias	Numérica	importe de regalías mineras liquidado
tasa_insp_fisc	Numérica	importe de tasas de fiscalización liquidado
ano_comer	Numérica	año de comercialización territorio nacional continental
mes_comer	Numérica	mes de comercialización a territorio nacional continental
productor_comer	Numérica	número de productor
producto_comer	Numérica	producto comercializado
volumen_comercializado	Numérica	volumen comercializado
valor_fob_usd	Numérica	valor fob expresado en dólares (valor de mineral puesto en aduana)
valor_fob_ars	Numérica	valor fob expresado en pesos (valor de mineral puesto en aduana)
tasas_comercial	Numérica	tasas abonadas en concepto de emisión de certificado exportación
tasa_cambio	Numérica	relación peso dólar, al momento de emitir el certificado de exportación
sin_comercializacion	Categórica	valor TRUE (no comercializó) valor FALSE (comercializó)

De este Dataset anonimizado, se estimará el tiempo de vida de cada uno de los yacimientos en base a saber su volumen total actual y las extracciones realizadas en los periodos que van desde Enero 2020 a Diciembre 2024 (5 años exactos)

Para el análisis a realizar, tendremos en cuenta solo aquellos yacimientos que han tenido producción en el periodo mencionado, es decir, que hayan tenido o estén en actividad informada.

Desarrollo del Modelo

Para este modelo se utilizó:

- Algoritmo: Random Forest Regressor
- Tipo de Modelo: Aprendizaje supervisado, Regresión
- Variable Objetivo: meses_restantes_estimados (vida útil en meses)
- Variable predictora principal: extraccion_prom_mensual (media móvil de extracción de los últimos 6 meses).
- Transformación aplicada: Log-transformación de la variable objetivo para estabilizar la varianza y mejorar la distribución.

Hiperparámetros utilizados:

n_estimators = 100: número de árboles en el bosque.

random_state = 42: para reproducibilidad.

Se utilizó el valor por defecto de max_depth para permitir que cada árbol crezca libremente, maximizando el ajuste local.

Métricas utilizadas para evaluar el modelo:

MAE (Error Absoluto Medio)

RMSE (Raíz del Error Cuadrático Medio)

R^2 (Coeficiente de Determinación)

Análisis de resultados:

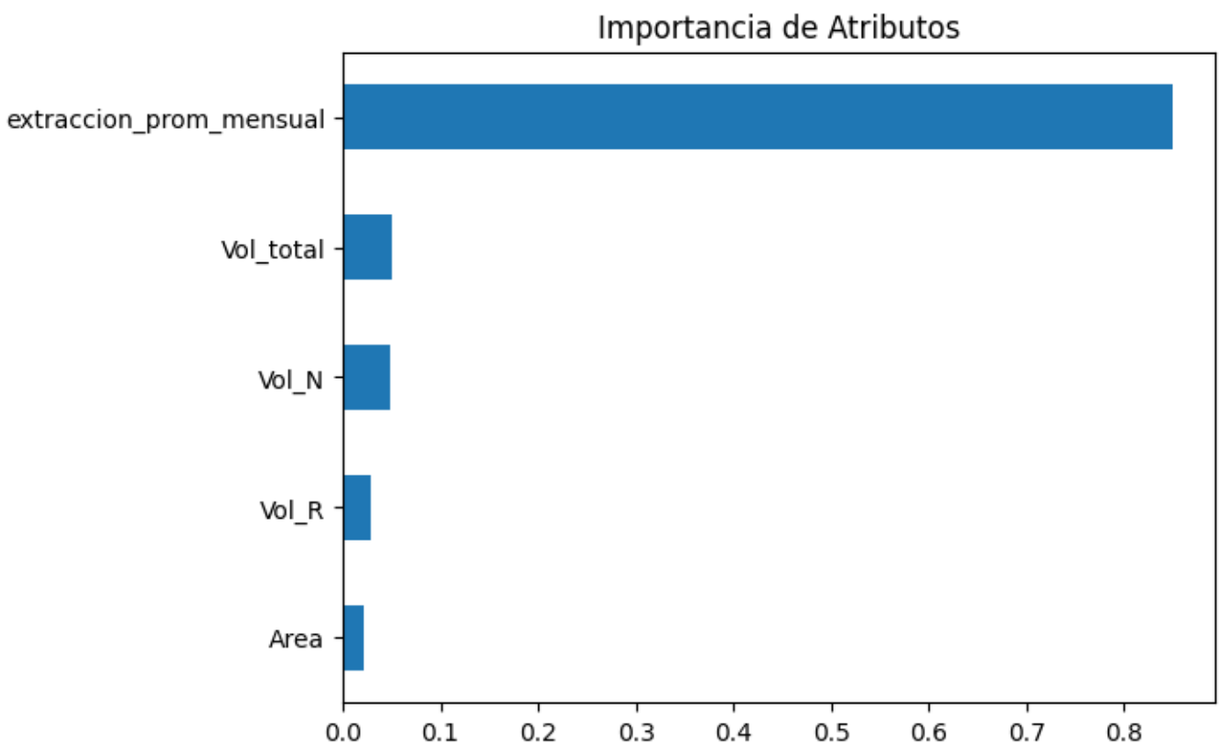
Análisis Exploratorio de datos:

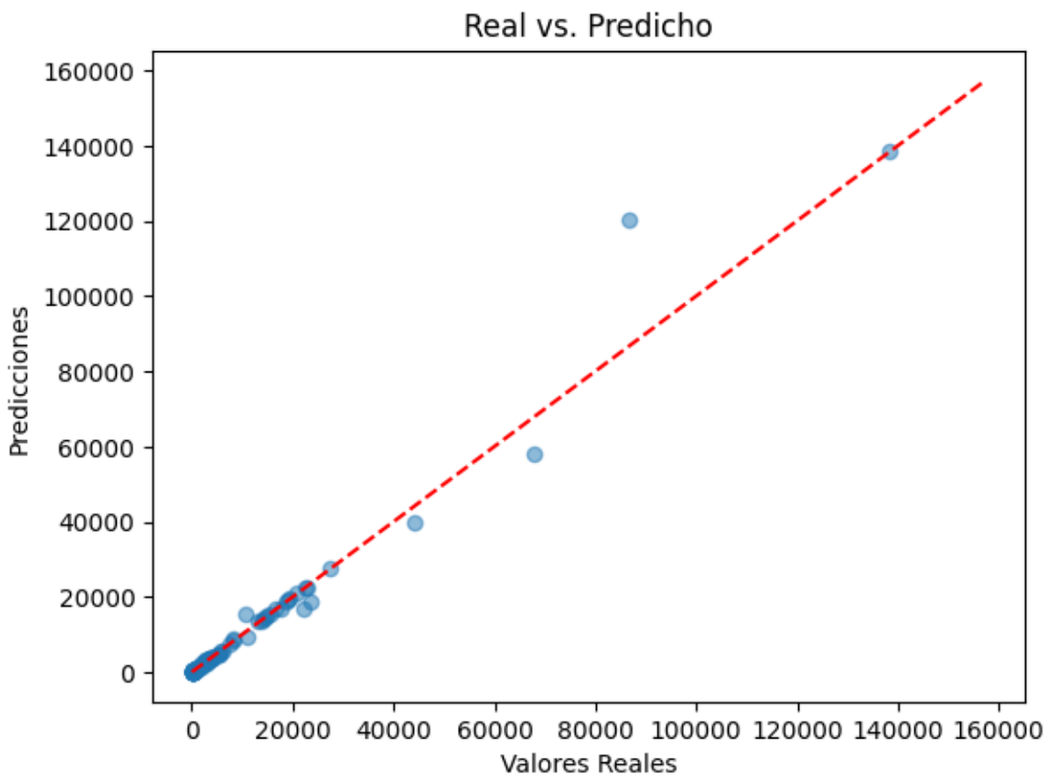
- Se tienen 1635 filas y 27 columnas.
- Se descartan 19 Columnas que no son de utilidad para nuestro análisis, siendo estas las siguientes:
- "Fecha", "id", "_uid_", "id_2", "producto", "num_expte", "trimestre", "regalias", "tasa_insp_fisc", "ano_comer", "mes_comer", "productor_comer", "producto_comer", "volumen_comercializado", "valor_fob_usd", "valor_fob_ars", "tasas_comercial", "tasa_cambio" y "sin_comercializacion".
- Lo que nos lleva a quedarnos con 8 columnas:

- “Area”, “Vol_N”, “Vol_R”, “Vol_Total”, “polígono”, “ano”, “mes” y “volumen_producción”
- Del resultado restante se procede a eliminar las filas donde polígono es 0 o NaN, lo que nos lleva a eliminar 11 filas, quedándome 1624 filas.
- Luego se convierten las columnas “Area”, “Vol_N”, “Vol_R”, “Vol_Total” y “volumen_producción” a tipo Float , las columnas “poligono2, “ano” y “mes” a tipo entero (int64) y creo la columna “fecha” de tipo periodo con los datos de “ano” y “mes”.
- A posterior se eliminan las filas en las que “volumen_producción” = 0, ya que si no produjeron no serán tenidas en cuenta para el modelo. Se eliminan 1120 Filas.
- Todo este proceso lleva a quedarme con 504 Filas y 10 Columnas.
- Ya con la reducción del Dataset, procedo a evaluar los Outliers. Lo que me da que tengo Outliers que procedo a eliminar.
- Luego agrupo por “polígonos” y “ano_mes”.
- Calculo la extracción acumulada "volumen_acumulado", el "volumen_restante" y "extraccion_prom_mensual" y estimo los meses restantes.
- A este momento tengo 447 Filas y 11 Columnas.

Contruyo el Modelo - RandomForestRegressor

- Al Modelo lo creo sin los "meses_restantes_estimados" y "extraccion_prom_mensual" que tengan NaN, ya que tampoco me serán de utilidad.
- Divido en Entrenamiento 80% y prueba 20%.
- Confirmo que no tenga nulos.
- Entreno el modelo con RandomForestRegressor.
- Hago la Predicción.
- Evaluo MAE, RMSE y R^2 y obtengo:
  - MAE: 837.0908775674981
  - RMSE: 3840.657154548537
  - R^2 Score: 0.9596689800910517



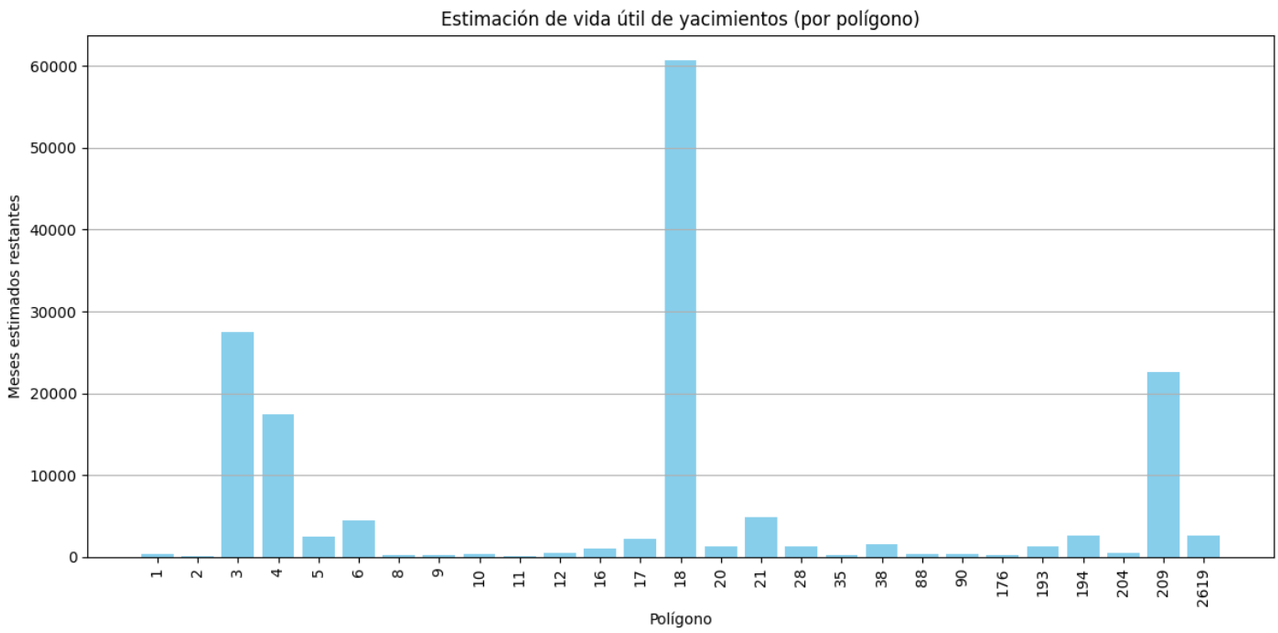


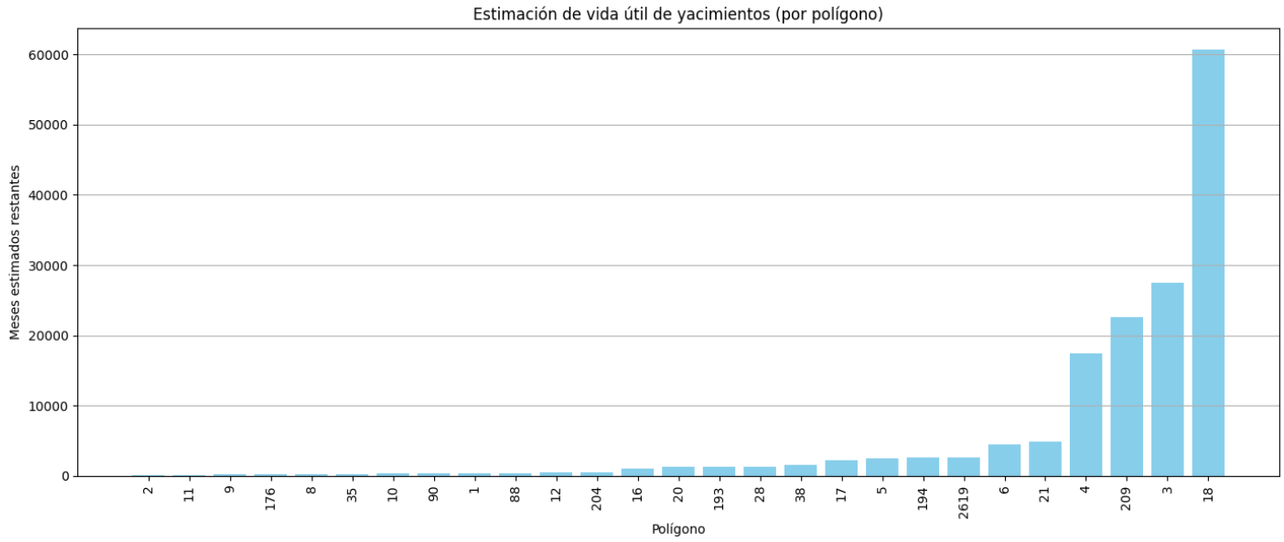
- Se puede observar que salvo algunos puntos el modelo se ajusta muy bien a la predicción, por eso el R^2 es de casí un 96%.
- Analizo los polígonos extremos.
- Observo que los polígonos con más datos son: [209, 2, 204, 8, 17]
- Visualizo el impacto:

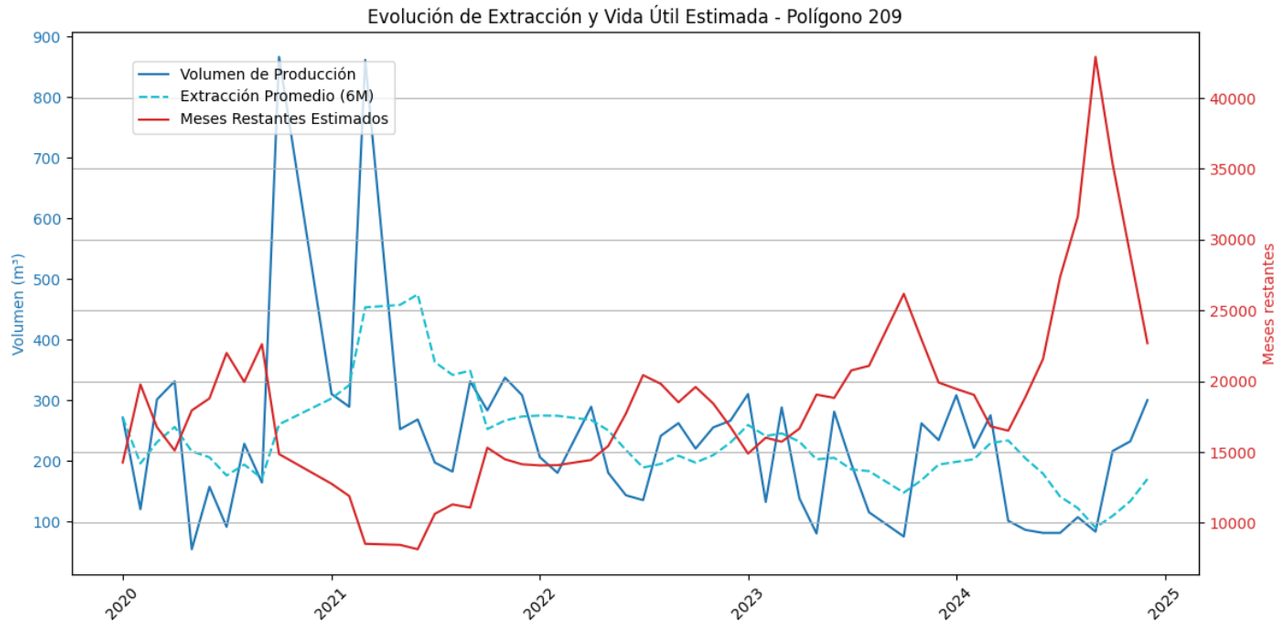
	meses_restantes_estimados	meses_estimados_modelo	error_meses
count	447.000000	447.000000	447.000000
mean	8515.542346	8506.338186	385.112295
std	22349.245128	22208.215179	2276.363074
min	36.624348	40.690531	0.005968
25%	444.247029	461.253184	4.752634
50%	1110.280716	1109.181371	18.883226
75%	4948.577875	4857.436082	78.992783
max	157290.381818	151162.990130	33632.275280

- Obtengo la tabla de resumen de polígonos:

	poligono	ano_mes	volumen_restante	extraccion_prom_mensual	meses_restantes_estimados	meses_estimados_modelo	error_meses
15	1	2024-12	231729.69	608.333333	380.925518	391.363340	10.437822
46	2	2024-11	33265.70	488.300000	68.125538	77.941643	9.816105
64	3	2024-02	576541.40	21.666667	26609.603077	27533.802406	924.199329
74	4	2024-03	1758183.80	106.833333	16457.258658	17463.076309	1005.817651
99	5	2024-12	2097020.34	840.840000	2493.958827	2502.492512	8.533685
122	6	2024-12	1773770.40	397.500000	4462.315472	4494.997119	32.681647
149	8	2024-12	322789.62	1078.583333	299.271841	305.077971	5.806130
155	9	2021-07	39503.60	211.338333	186.921130	211.767373	24.846243
161	10	2023-02	63350.21	180.898333	350.197864	349.946588	0.251276
170	11	2022-03	39259.01	426.533333	92.042068	93.532367	1.490300
188	12	2024-12	222561.05	435.800000	510.695388	521.784549	11.089162
195	16	2024-04	97369.16	86.136667	1130.403158	1064.072213	66.330944
221	17	2024-09	452235.12	222.600000	2031.604313	2204.345538	172.741225
224	18	2021-01	7469643.33	116.666667	64025.514257	60726.388001	3299.126257
236	20	2023-08	490487.40	376.000000	1304.487766	1326.358569	21.870803
246	21	2022-12	555194.24	118.333333	4691.782310	4899.739592	207.957283
269	28	2024-03	786154.18	577.833333	1360.520646	1360.217625	0.303021
278	35	2024-12	122759.63	344.750000	356.083046	306.878260	49.204785
285	38	2023-03	720921.31	478.333333	1507.152564	1527.592985	20.440421
294	88	2024-12	172556.19	523.120000	329.859669	415.798866	85.939197
302	90	2024-12	222731.89	638.166667	349.018370	371.716666	22.698296
324	176	2024-12	290053.84	1178.333333	246.156017	248.065682	1.909665
334	193	2024-03	898274.28	629.666667	1426.586998	1346.731616	79.855382
340	194	2024-12	266373.11	101.666667	2620.063377	2589.754621	30.308756
368	204	2024-12	528761.13	891.616667	593.036391	566.691284	26.345107
423	209	2024-12	3853472.10	169.833333	22689.727772	22541.088418	148.639354
446	2619	2024-12	2888473.19	1076.923333	2682.153038	2622.558411	59.594626



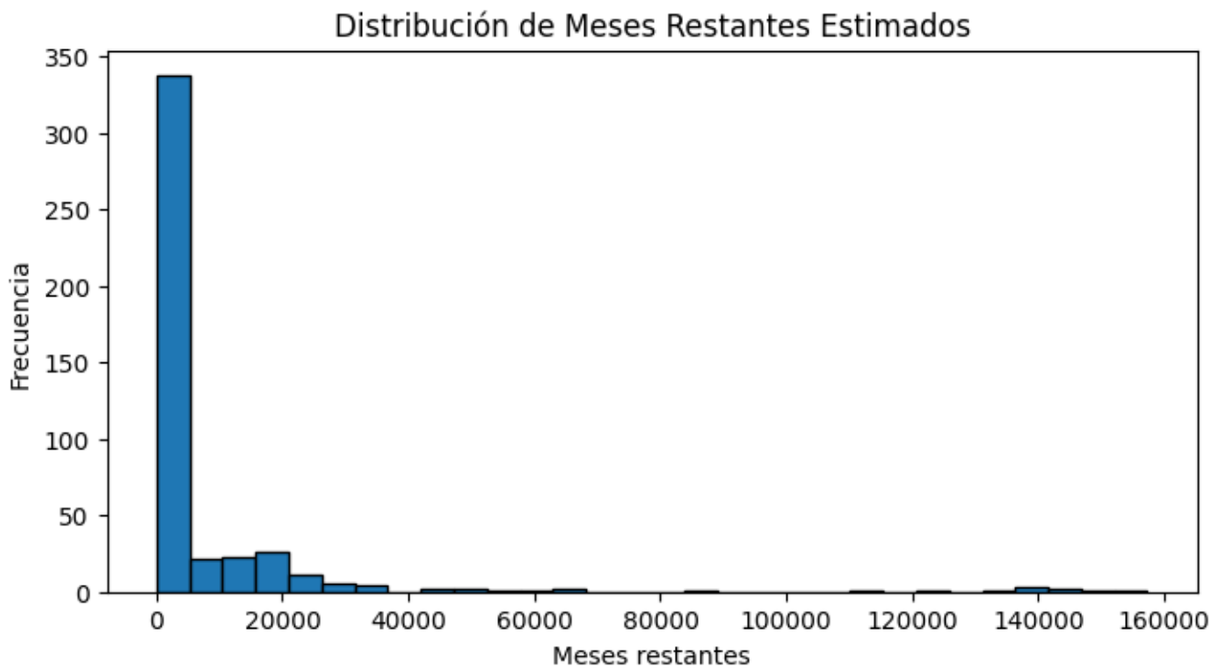




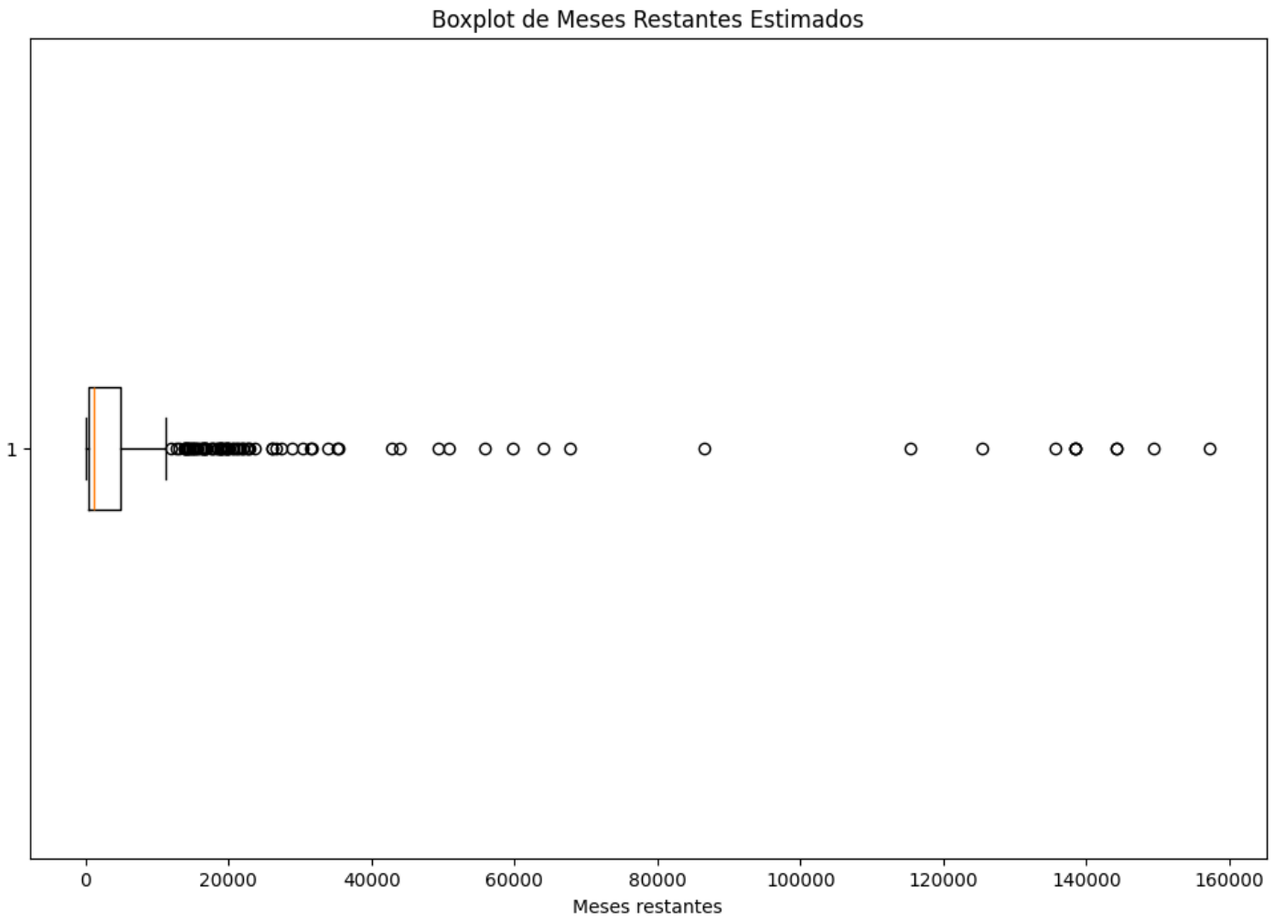
En este gráfico se puede observar como cuando la producción (línea azul continua) se ve incrementada, el tiempo de meses restantes estimados baja, dando a entender que el yacimiento se terminará antes.

También podemos observar en una línea punteada celeste el promedio mensual de extracción de los últimos 6 meses.

Saco unas estadísticas de la variable objetivo "meses_restantes_estimados"



Genero un Box-Plot de la variable objetivo



Y observo que existen outliers, por lo que procedo a sacar medidas y eliminar los innecesarios (83), lo que me d

ja con 364 Filas sin los outliers.

Contruyo el Modelo de Transformación Logarítmica

Para intentar mejorar el modelo Genero un modelo de Transformación Logarítmica y vuelvo a calcular MAE, RMSE y R^2, obteniendo:

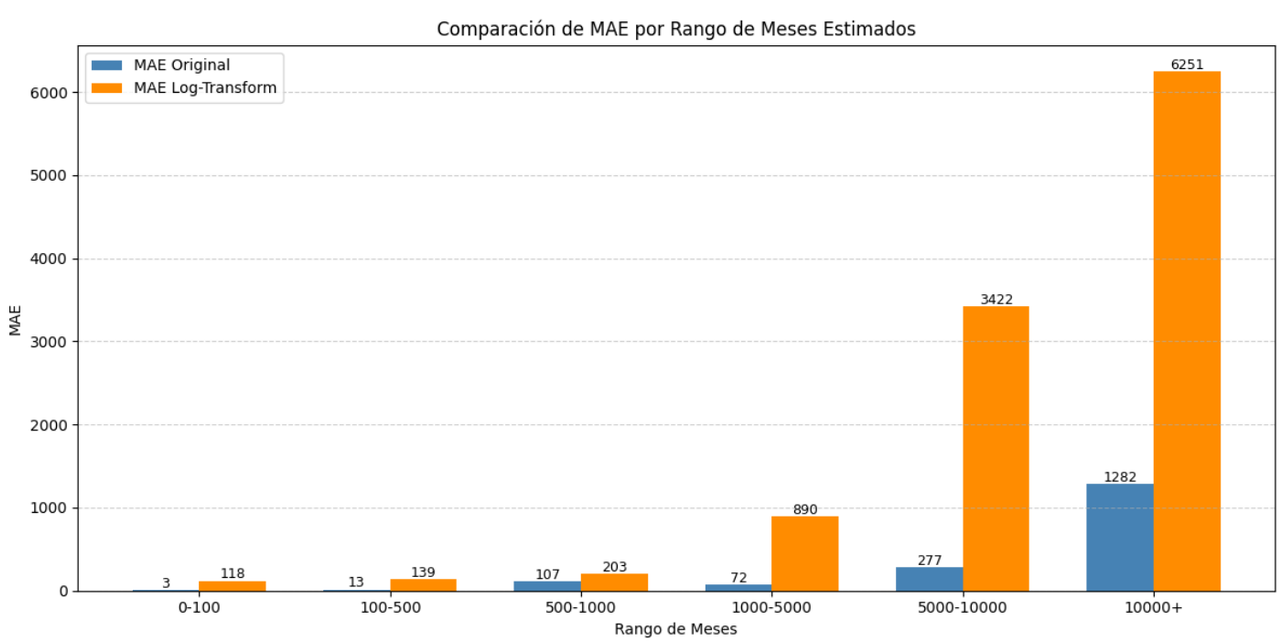
- MAE después de log-transform & back-transform: 710.9698775230927
- RMSE después de log-transform & back-transform: 1432.9253672516036
- R^2 para predicción log-transformada y revertida: 0.5681

Y vuelvo a generar la tabla por polígonos:

	poligono	ano_mes	volumen_restante	extraccion_prom_mensual	meses_restantes_estimados	meses_estimados_log_modelo	error_log
15	1	2024-12	231729.69	608.333333	380.925518	489.289181	108.363664
46	2	2024-11	33265.70	488.300000	68.125538	331.087183	262.961646
73	4	2023-02	1758309.80	165.833333	10602.873166	5607.218383	4995.654782
99	5	2024-12	2097020.34	840.840000	2493.958827	2140.465256	353.493571
122	6	2024-12	1773770.40	397.500000	4462.315472	1117.412162	3344.903310
149	8	2024-12	322789.62	1078.583333	299.271841	412.847464	113.575623
155	9	2021-07	39503.60	211.338333	186.921130	728.099460	541.178331
161	10	2023-02	63350.21	180.898333	350.197864	430.175989	79.978124
170	11	2022-03	39259.01	426.533333	92.042068	175.139673	83.097606
188	12	2024-12	222561.05	435.800000	510.695388	578.773857	68.078470
195	16	2024-04	97369.16	86.136667	1130.403158	2004.259267	873.856109
221	17	2024-09	452235.12	222.600000	2031.604313	2692.471859	660.867547
236	20	2023-08	490487.40	376.000000	1304.487766	1242.008087	62.479679
246	21	2022-12	555194.24	118.333333	4691.782310	2762.561299	1929.221011
269	28	2024-03	786154.18	577.833333	1360.520646	970.434253	390.086393
278	35	2024-12	122759.63	344.750000	356.083046	597.871448	241.788402
285	38	2023-03	720921.31	478.333333	1507.152564	1810.370554	303.217989
294	88	2024-12	172556.19	523.120000	329.859669	253.927305	75.932364
302	90	2024-12	222731.89	638.166667	349.018370	654.476331	305.457961
324	176	2024-12	290053.84	1178.333333	246.156017	515.700821	269.544804
334	193	2024-03	898274.28	629.666667	1426.586998	1056.392068	370.194930
340	194	2024-12	266373.11	101.666667	2620.063377	2898.954625	278.891248
368	204	2024-12	528761.13	891.616667	593.036391	521.275975	71.760416
386	209	2021-09	3860977.80	348.500000	11078.845911	2833.985352	8244.860559
446	2619	2024-12	2888473.19	1076.923333	2682.153038	1276.660466	1405.492572

En la cual no solo tengo 2 polígonos menos (3 y 18), sino que los valores si bien parecen mejorar tanto en MAE como en RMSE, el R^2 empeoró muchísimo.

Y al realizar una comparativa del MAE por rango de meses obtengo:



En este gráfico se puede observar para concluir que:

- Para los rangos bajos (0-100, 100-500, etc.), el modelo original tiene errores mucho menores que el modelo log-transformado.
- A medida que aumenta el rango, especialmente desde los 1000 meses en adelante, el modelo log-transformado comienza a producir errores mucho mayores.

El Modelo de Transformación Logarítmica puede suavizar la influencia de valores extremos en el entrenamiento, pero al revertirla puede amplificar los errores en las predicciones grandes. Esto sugiere que el modelo original está mejor calibrado para todo el rango, especialmente para valores grandes, mientras que el log-transformado puede no estar capturando correctamente la magnitud de las predicciones en rangos altos.

Conclusión Final

El modelo de Random Forest Regressor sin transformación logarítmica fue el que presentó el mejor rendimiento global, con un R^2 de 0.96, lo que indica una excelente capacidad explicativa de la variabilidad de los datos. Además, mostró un error absoluto medio (MAE) de aproximadamente 837 meses y un RMSE de 3840, valores coherentes con la escala y dispersión de la variable objetivo.

Por otro lado, el modelo con transformación logarítmica logró un MAE más bajo (711) y una notable reducción del RMSE (1432), pero a costa de una fuerte caída en el R^2 (0.57). Esta baja en el poder explicativo, sumada al análisis por rangos de meses restantes, donde el modelo original superó consistentemente al transformado, refuerza la decisión de mantener el modelo sin logaritmo como el más confiable.

Sin embargo, es importante remarcar que el dataset cuenta con solo cinco años de información histórica (Enero 2020 a Diciembre 2024), lo cual resulta escaso frente a estimaciones que se expresan en escalas de cientos o incluso miles de meses. Esta limitación temporal reduce la capacidad del modelo para aprender patrones a largo plazo y afecta su fiabilidad cuando proyecta escenarios muy alejados en el tiempo.

A pesar de la mejora en algunos indicadores puntuales, la transformación logarítmica no aportó beneficios generalizados al rendimiento del modelo. El modelo original de Random Forest, sin transformación, es el más robusto y balanceado para estimar la vida útil restante de los yacimientos de turba. No obstante, debe considerarse que la brevedad del histórico disponible condiciona la precisión de las estimaciones, especialmente en horizontes temporales largos.