Universidad de Buenos Aires

Facultad de Ingeniería

Informe Final: Introduccion a la Ciencia de Datos

Autor/es:
Ambrosio, Facundo Nahuel
Arteaga, Lucas Ariel
Grondona, Juan Ignacio

Junio 2024

Introducción

El objetivo de este trabajo práctico integrador es aplicar los conocimientos adquiridos en la materia Introducción a la Ciencia de Datos, impartida en la Facultad de Ingeniería de la Universidad de Buenos Aires. En este proyecto grupal, desarrollado por estudiantes de Ingeniería en Petróleo, se emplea la metodología CRISP-DM para abordar un caso práctico de minería de datos.

El propósito principal es implementar esta metodología hasta la cuarta fase, correspondiente al Modelado, utilizando las herramientas y técnicas aprendidas a lo largo del cuatrimestre.

Fase 1: Entendimiento del negocio

Industria petrolífera

Los recursos hidrocarburíferos, tanto gas como petróleo, representan más del 50% de la matriz energética de la argentina, esto es así debido a que argentina posee múltiples zonas de explotación de este recurso, de las cuales una de ellas es la que se conoce coloquialmente como “vaca muerta”, siendo una de las más grandes del mundo en cuanto a recursos no convencionales se trate.

Es por ello que al ser la explotación de hidrocarburos una actividad difundida a lo largo del país, este no solo lleva consigo grandes inversiones de capital tanto en explorar zonas que contengan dichos recursos, como extraerlos y luego refinarlos para venderlos, sino también a la formación de múltiples regulaciones que permitan mantener una actividad lo menos nociva y contaminante para el medio ambiente posible y también (como se verá más adelante) para tener un seguimiento de cuánto están produciendo las empresas sea petróleo o gas en forma de declaración jurada.

Los hidrocarburos se producen mediante pozos, perforaciones que pueden tener un rango bastante amplio de profundidades, los más someros pueden ser de 1000 a 2000 metros y otros hasta 5000 o 6000 metros, estos pozos extraen los hidrocarburos que hay almacenados en las rocas reservorio a través de estos pozos, llegando hasta la superficie y luego son tratados para poder transportarlos a las refinerías y de ahí sacar todos los subproductos, como la nafta, kerosene, etc. Estos pozos no se ubican a cualquier profundidad arbitraria, sino que se hace un

estudio geológico de la zona para determinar si puede llegar a haber o no hidrocarburos, estos pueden almacenarse en lo que se conocen como formaciones geológicas, que no son más que capas de rocas en el subsuelo con características similares, con dimensiones definidas, que por lo general se acumulan una arriba de la otra. Un mismo pozo de petróleo puede perforar o atravesar varias formaciones geológicas (algunas productivas, otras no), hasta llegar a la de interés, volviendo al ejemplo de vaca muerta, esta última es una formación geológica que se presenta a un rango de profundidades específicas que puede abarcar a casi toda la provincia de Neuquén, y parte de La Pampa y Mendoza. Pero no siempre que se trate de explotar hidrocarburos en esa zona se va a querer perforar en la Formación Vaca Muerta, hay otras Formaciones comprendidas en esa misma zona que también son productivas, Por ejemplo, Los Molles.

La producción de una formación en específica puede estar dividida en yacimientos, que son zonas de características productivas en común, ya que no toda la formación productiva va a estar rellenada con hidrocarburos, como estos yacimientos pueden abarcar grandes extensiones de territorio, se las divide en concesiones, las cuales cada empresa va a estar habilitada para trabajar en dichas áreas.

Esta agrupación de formaciones geológicas forman parte de lo que se conoce como cuencas petrolíferas, en Argentina, hay principalmente 5 cuencas petrolíferas, Neuquina(Vaca Muerta, Los Molles), Golfo San Jorge, Austral, Cuyana, y del Noroeste, cada una de estas con extensiones kilométricas, abarcando varias provincias.

En términos simples, la subdivisión geográfica de un pozo de petroleo seria la siguiente:

CUENCA → PROVINCIA → FORMACION → YACIMIENTO → CONCESIÓN → POZO

Cuando uno produce hidrocarburos lo hace aprovechando la alta presión que hay en la formación, haciendo que esta empuje tanto el petróleo como el gas de la roca a el pozo, y de ahí hasta la superficie, en muchos casos cuando dicha presión baja por la declinación natural que esta presenta, se llega a emplear métodos de extracción artificial, para solventar esa falta de “empuje”, entre uno de ellos está el de emplear pozos inyectores, en los que se inyecta agua o algún otro fluido para “barrer” o “empujar” los hidrocarburos en la roca. tanto en un caso como en el otro los principales componentes que se obtienen son petróleo, gas y agua, en distintas proporciones, En un caso ideal lo mejor seria extraer la mayor cantidad de petróleo posible y la menor de agua, hay pozos en zonas en específico que por las características que

presentan son productivos en cuanto a petróleo pero tienen lo que se conoce como un alto corte de agua, sobre todo en los pozos no convencionales.

Un yacimiento de petróleo convencional es aquel que que presenta características físicas como porosidad y permeabilidad tales que una vez que se perfora la que seria la formación o la zona productiva de la roca, el petróleo pueda en un inicio producir por sí mismo, es decir es un yacimiento en el cual se puede aprovechar la energía natural del mismo para producir, en cambio en los no convencionales, si se emplea este mismo método no se va a lograr nada, esto es así debido a que las características que presenta la roca no permite el flujo de los hidrocarburos, por lo que se tienen que emplear otros métodos para poder lograrlo, por ejemplo, romper la roca que tiene almacenada los hidrocarburos para que estos puedan moverse y poder producirse en superficie.

Los reservorios de hidrocarburos pueden encontrarse en dos condiciones principales: subsaturados o saturados, y estas determinan cómo se distribuyen las fases de petróleo y gas en su interior. En el caso de un reservorio subsaturado, todos los fluidos presentes están en estado líquido, lo que significa que no existe gas libre dentro del reservorio, ya que todo el gas está “disuelto” en el petróleo. Sin embargo, dado que el petróleo es una mezcla multicomponente, cuando este se extrae y llega a la superficie, parte del gas se libera debido a la diferencia de presión entre el reservorio y la superficie. Por eso, aunque en el subsuelo solo haya petróleo, siempre se produce gas en superficie, en mayor o menor medida. Si la presión del reservorio disminuye lo suficiente durante la producción (hasta la conocida como presión de burbuja), el gas disuelto puede liberarse directamente en el reservorio, generando gas libre. Esto incrementa la cantidad de gas producido, ya que, desde un punto de vista técnico, el gas es más fácil de extraer que el petróleo debido a su menor viscosidad y densidad.

Por otro lado, en un reservorio saturado, dentro de la roca de almacenamiento ya coexisten el petróleo y el gas en fases separadas. En este caso, el gas está presente como gas libre en el reservorio desde el inicio. Durante la producción, se obtendrá tanto petróleo como gas, y la cantidad de gas será mayor que en un reservorio subsaturado, porque proviene tanto del gas que estaba disuelto en el petróleo como del gas libre ya presente en el reservorio.

Cada pozo tiene un tiempo de vida útil que depende de múltiples factores, principalmente productivos, ya que no tendría sentido seguir manteniendo un pozo que no produce, cuando estos lleguen a ese punto se decidirá abandonarlo o emplearlo para otros fines. Cada pozo está sujeto a revisiones periódicas y periodos de reparación.

Secretaria de energia

La Secretaría de Energía de la Nación Argentina, de cuya página proviene el dataset utilizado en este informe, es responsable de diseñar y ejecutar la política energética nacional, regulando el régimen de combustibles, servicios públicos energéticos y estructuras arancelarias. Además, supervisa el desarrollo tecnológico y fomenta la explotación racional de recursos naturales, la investigación de nuevas fuentes de energía y la cooperación internacional. También dirige la representación estatal en empresas energéticas, ejerce control sobre entes reguladores y se encarga de la política nuclear para fines pacíficos.

En el contexto de la industria petrolífera en argentina, según el artículo 1° y 2° de la resolución 319/1993 dictada el 21 de octubre de 1993, la secretaria de energia sería la encargada de recibir y almacenar datos relacionados a todo lo que tenga que ver con producción de pozos de petróleo, gas, u otros, sus datos geográficos (como la provincia a la que pertenece cada pozo), el estado operativo de los mismos (si está en producción, en reparación o abandonado), la identificación de cada pozo, el mes o año en el que se hizo la entrada de datos, la empresa operadora, o la profundidad a la que se está extrayendo los recursos, entre otros, todo esto con caracter de declaracion jurada.

cita textual de ambos artículos:

- **artículo 1**: Apruébanse las Normas y Procedimientos para la remisión de información estadística, datos primarios y documentación técnica a la SECRETARIA DE ENERGIA, a las que deberán ajustarse las empresas y/o consorcios Permisionarios de Exploración, Concesionarios de Explotación y de Transporte, Refinadoras y Comercializadoras de hidrocarburos.

- **artículo 2**: La información estadística y documentación técnica mencionadas en el artículo precedente, revestirá el carácter de declaración jurada, excepto cuando se indique lo contrario para temas específicos, debiendo suministrarse en los plazos y formas que para cada caso se determina en la presente resolución.

Toda esta información proporcionada por las empresas, se puede encontrar almacenada en la página oficial de la Secretaria de energia, en donde ofrecen múltiples archivos .cvs de datos de producción de pozos, o de listados de pozos, de series históricas de producción por cuenca,

etc. Esta naturaleza obligatoria de los datos suministrados garantiza un nivel de precisión y consistencia de datos tales que se pueda emplear técnicas de minería de datos.

Objetivos y criterios de éxito comercial

Como se indicó previamente, los fluidos principales que se extraen de un pozo son el petróleo, el agua y el gas. En general, el petróleo es el de mayor valor en la industria, ya que permite obtener una amplia variedad de productos derivados, como combustibles (nafta, kerosene, gasoil) y plásticos (polietileno, polipropileno, entre otros). El gas, aunque tiene cierto valor como fuente de energía en sectores domésticos, comerciales e industriales, suele quedar en segundo plano respecto al petróleo. De hecho, una parte importante del gas extraído de los pozos es utilizada en las mismas plantas de procesamiento para suministrar energía al tratamiento del crudo.

El agua, por su parte, es un fluido cuya presencia es inevitable en los pozos. No solo carece de valor comercial debido a que contiene hidrocarburos disueltos y sólidos en suspensión, sino que también representa un costo adicional para las empresas, que deben tratarla para separarla de los otros fluidos y darle algún uso. Esto puede implicar su reinyección en la formación para aumentar la productividad del pozo o su disposición en un acuífero. Por estas razones, aunque siempre está presente, el agua genera desafíos para las empresas productoras al representar un gasto adicional sin aportar beneficios económicos directos. Por lo que el principal objetivo comercial de la extracción de hidrocarburos sería la de extraer la mayor cantidad posible de petróleo y/o gas dependiendo de qué pozo se trate, mientras se produce la menor cantidad posible de agua y se reducen los costos operativos relacionados al tratamiento de esta.

Dicho esto, este informe tiene como objetivo principal identificar patrones de producción mediante minería de datos en los pozos de hidrocarburos, con el fin de orientar las inversiones de capitales hacia aquellos grupos de pozos que presenten una mayor rentabilidad basándose en los criterios antes mencionados, en particular, en la búsqueda de maximizar la producción de petróleo en las áreas con mayor productividad, considerando factores como la cuenca, provincia y características operativas de los pozos.

Además, se plantea como objetivo clave reducir los costos operativos asociados al manejo del agua, mediante la identificación de grupos de pozos que presenten una combinación favorable de altos niveles de producción de gas o petróleo, junto con bajos niveles de agua extraída de tal

forma de optimizar los recursos, mejorar la eficiencia operativa y minimizar gastos innecesarios.

Por último, el análisis segmentar las áreas productivas en función de características geográficas, técnicas y operativas, apoyando la toma de decisiones estratégicas. Esta segmentación servirá como base para la planificación de nuevos proyectos y priorización de esfuerzos en zonas de alto potencial productivo, alineándose con los objetivos de negocio a largo plazo.

Dicho todo esto, es importante tener en cuenta las siguientes restricciones y supuestos que limitan o condicionan el siguiente proyecto:

1. No se cuenta con datos comerciales como el costo por volumen de hidrocarburos o de tratamiento de aguas, por lo que los beneficios o pérdidas que se podrían llegar a obtener son meramente cualitativos, esta limitación se debe a la confidencialidad que manejan las empresas productoras.

2. Únicamente se cuenta con datos de producción de fluidos, geográficos, temporales, y de categorización de pozos.

3. Se asume que la información disponible es suficiente para identificar patrones consistentes tales que ofrezcan una base sólida para la toma de decisiones estratégicas de inversión sin tener en cuenta variables económicas cualitativas.

4. Si bien los datos que poseemos son todos pertenecientes a el año 2023, asumimos que las estimaciones y predicciones derivadas de ellos son extrapolables a años posteriores, siempre y cuando no haya cambios significativos en la industria petrolera argentina que afecten en la dinámica de la producción, como por ejemplo, que se descubran nuevas cuencas o yacimientos que generen un cambio en el paradigma productivo de la nación.

5. Se cuenta con una calidad de datos lejos de ser perfecta, por lo cual en el apartado de la fase 3 se detalla las correcciones y criterios tomados para solventar esta cuestión.

Fase 2: Entendimiento de los datos

Dataset

El Dataset en el cual nos enfocaremos en este informes es el “Producción de Pozos de Gas y Petróleo - 2023”, que en resumidas cuentas, almacena la producción mensual de todos los pozos registrados hasta diciembre de 2023, en donde puede haber tanto pozos que arrancaron su producción desde antes de enero de 2023, como aquellos que lo hicieron durante el transcurso de este. la pagina en la cual obtuvimos el dataset es la siguiente: https://datos.gob.ar/dataset/energia-produccion-petroleo-gas-por-pozo-capitulo-iv/archivo/energia_231c39b3-e81e-4398-af8d-b115807f2c25

Características

El dataset está compuesto por 972268 registros y 38 columnas, de los cuales comprenden datos del tipo, object, numéricos tanto enteros como float, y categorizar. en la siguiente tabla se muestran todos los datos con la cantidad de valores no nulos y el tipo de variable:

Índice	Nombre	Valores no nulos	Tipo de dato
0	idempresa	972268	object
1	anio	972268	int64
2	mes	972268	int64
3	idpozo	972268	int64
4	prod_pet	972268	float64
5	prod_gas	972268	float64
6	prod_agua	972268	float64
7	iny_agua	972268	float64
8	iny_gas	972268	float64
9	iny_co2	972268	float64
10	iny_otro	972268	float64
11	tef	972268	float64
12	vida_util	24630	float64
13	tipoextraccion	972227	object
14	tipoestado	972227	object
15	tipopozo	972227	object
16	observaciones	67502	object
17	fechaingreso	972268	object
18	rectificado	972268	object
19	habilitado	972268	object
20	idusuario	972268	int64
21	empresa	972268	object
22	sigla	972268	object
23	formprod	940108	object
24	profundidad	972268	float64
25	formacion	938753	object
26	idareapermisoconcesion	972268	object
27	areapermisoconcesion	972268	object
28	idareayacimiento	972268	object
29	areayacimiento	972268	object
30	cuenca	972232	object
31	provincia	972268	object
32	tipo_de_recurso	972268	object
33	proyecto	972268	object
34	clasificacion	768464	object
35	subclasificacion	768464	object
36	sub_tipo_recurso	42965	object
37	fecha_data	972268	object

Los valores máximos y mínimos, promedios, desviaciones estándar, y percentiles de cada campo se presenta a continuación:

Para el caso de las variables numéricas:

	anio	mes	idpozo	prod_pet	prod_gas	prod_agua	iny_agua
count	972268	972268	972268	972268	972268	972268	972268
mean	2023	6,495796426	101653,1114	38,34504049	49,2949393	353,8229819	357,2131339
std	0	3,444634064	42457,50812	247,643294	737,5020024	1199,242443	2057,653282
min	2023	1	212	0	0	-31,91	0
25%	2023	4	69950	0	0	0	0
50%	2023	6	109738	0	0	0	0
75%	2023	9	132204	11,1881225	0	21,1089625	0
max	2023	12	165014	15942,51505	77858,29	57112,8	181091,88
	iny_gas	iny_co2	iny_otro	tef	vida_util	idusuario	profundidad
count	972268	972268	972268	972268	24630	972268	972268
mean	0,009698499	0	0,627244215	10,49276486	0	375,6728371	1657,496467
std	2,936717312	0	122,483308	14,53220243	0	54,39932105	1700,278364
min	0	0	0	0	0	3	0
25%	0	0	0	0	0	334	1040
50%	0	0	0	0	0	345	1572
75%	0	0	0	29,67	0	420	2317
max	1588,05	0	72249,27	2387,01	0	478	378939

Para las variables categóricas y objetos:

	idempresa	tipoextraccion	tipoestado	tipopozo
count	972268	972227	972227	972227
unique	57	11	16	9
top	YPF	Sin Sistema de Extracción	Extracción Efectiva	Petrolífero
freq	472041	486754	325516	455330
	observaciones	fechaingreso	rectificado	habilitado
count	67502	972268	972268	972268
unique	245	4105	1	1
top	Cargado automáticamente como [Sin movimientos]	2023-12-11 12:13:44.220735	f	t
freq	24471	39339	972268	972268
	empresa	sigla	formprod	formacion
count	972268	972268	940108	938753
unique	57	74605	79	77
top	YPF S.A.	YPF.Nq.EL-11	BBAR	bajo barreal
freq	472041	60	109087	109087
	idareapermisoconcesion	areapermisoconcesion	idareayacimiento	areayacimiento
count	972268	972268	972268	972268
unique	359	358	1202	1086
top	ANG	ANTICLINAL GRANDE - CERRO DRAGON	HER	PUESTO HERNANDEZ
freq	67749	67749	30816	30816
	cuenca	provincia	tipo_de_recurso	proyecto
count	972232	972268	972268	972268
unique	8	11	4	2
top	GOLFO SAN JORGE	Santa Cruz	CONVENCIONAL	Sin Proyecto
freq	506700	279358	924739	958958
	clasificacion	subclasificacion	sub_tipo_recurso	fecha_data
count	768464	768464	42965	972268
unique	4	15	2	12
top	EXPLOTACION	DESARROLLO	SHALE	2023-11-30
freq	657824	585362	24522	81561

A continuación se explica que es cada campo del dataset:

- **idempresa**: Corresponde al identificador de la empresa que opera el pozo.

- **anio**: Año en el que se hizo la entrada de datos del pozo (2023)
- **mes**: Mes en el que se hizo la entrada de datos del pozo
- **idpozo**: Corresponde al identificador del pozo, cada pozo tiene un identificador único.
- **prod_pet**: Corresponde a la producción que hubo de petróleo durante ese mes, medida en metros cúbicos.
- **prod_gas**: Corresponde a la producción que hubo de gas durante ese mes, medida en miles metros cúbicos.
- **prod_agua**: Corresponde a la producción que hubo de agua durante ese mes, medida en metros cúbicos.
- **iny_agua**: Volumen de agua que se inyectó a la formación.
- **iny_gas**: Volumen de gas que se inyectó a la formación.
- **iny_co2**: Volumen de gas CO2 que se inyectó a la formación.
- **iny_otro**: Volumen de otros gases que se inyectaron a la formación.
- **tef**: Temperatura estabilizada de formación.
- **vida_util**: La proyección de vida útil del pozo, es decir, hasta cuando va a seguir produciendo.
- **tipoextraccion**: Qué método de extracción se está empleando para la producción de hidrocarburos, o si no se está empleando.
- **tipoestado**: Corresponde al estado productivo en el que está el pozo, si está en extracción efectiva, abandonado, a espera de abandono, si está en reparación, etc.
- **tipopozo**: Que tipo de pozo es, si es petrolero, gasífero, de inyección, etc.
- **observaciones**: Observaciones a tener en cuenta en conjunto con el atributo “tipoestado”.
- **fechaingreso**: Fecha de ingreso al dataset, o fecha de primer entrada de datos.
- **idusuario**: El identificador del usuario que cargó los datos del pozo.
- **empresa**: El nombre de la empresa operadora.

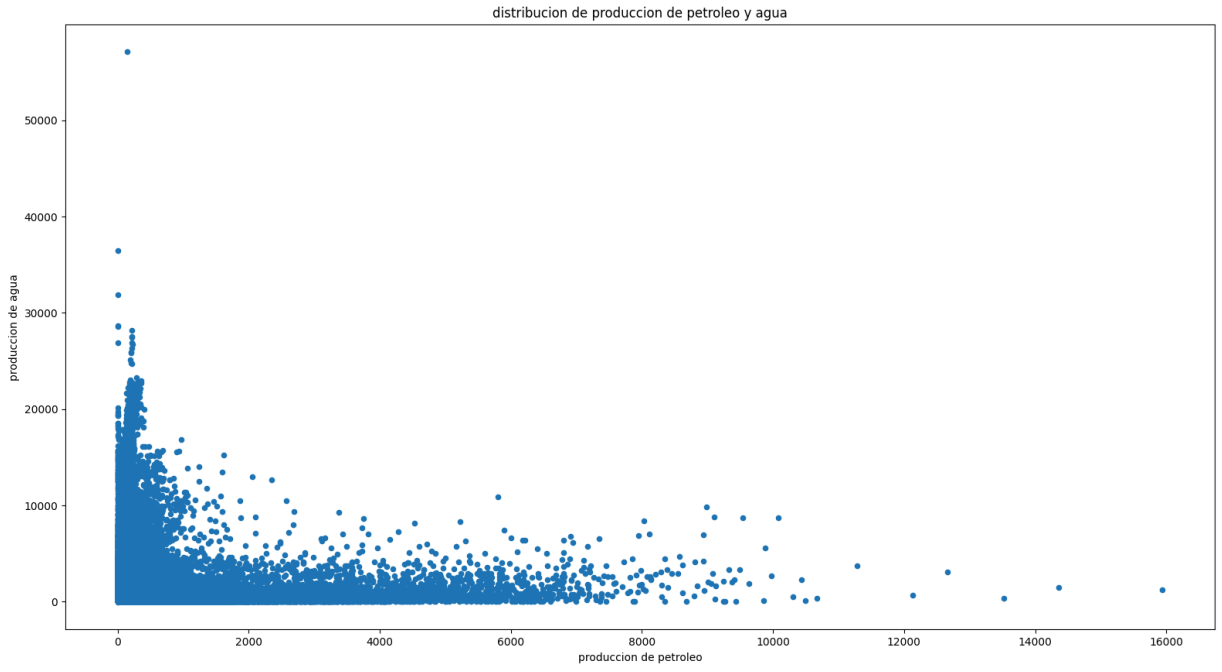
- **sigla**: Sigla de identificador alternativo al “idpozo”, contiene información sobre la empresa operadora en dicha sigla.
- **formprod**: Siglas de la formación productora a la que está produciendo.
- **profundidad**: profundidad del pozo.
- **formacion**: Nombre completo de la formación a la que pertenece el pozo.
- **idareapermisoconcesion**: Siglas de el area de concesion del pozo.
- **areapermisoconcesion**: Nombre completo del área de concesión del pozo.
- **idareayacimiento**: Siglas del yacimiento del pozo.
- **areayacimiento**: Nombre completo del yacimiento del pozo.
- **cuenca**: Nombre completo de la cuenca a la que pertenece el pozo
- **provincia**: Provincia en donde se encuentra el pozo.
- **tipo_de_recurso**: Si se trata principalmente de un recurso convencional o no convencional
- **proyecto**: Detalla si tiene algún proyecto activo en el pozo.
- **clasificacion**: En conjunto con el atributo “tipoestado”, si está en exploración, en explotación, si es de servicio.
- **subclasificacion**: en complemento con la anterior, proporciona más detalle sobre qué clasificación tiene el pozo, por ejemplo si es un pozo exploratorio profundo, o de explotación avanzada.
- **sub_tipo_recurso**: En complemento con el tipo de recurso, si es no convencional puede ser tight o shale.
- **fecha_data**: Corresponde hasta que fecha exacta se tuvo información de producción para ingresarlo al dataset.

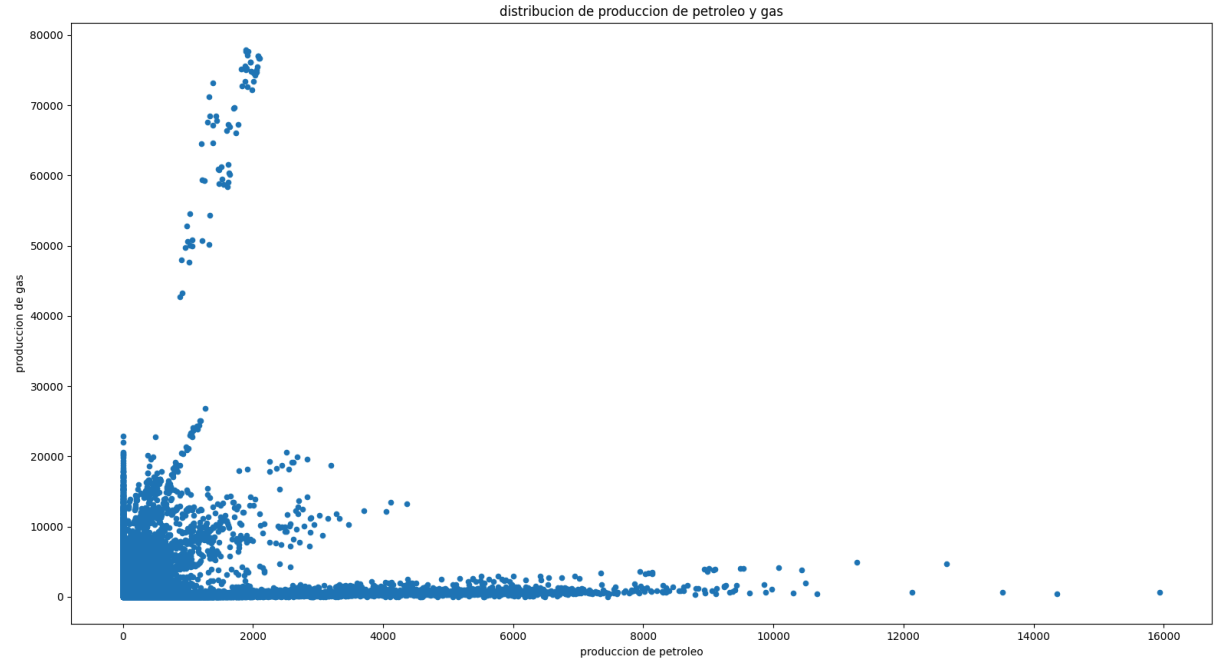
De todo este listado de columnas, se descartó la interpretación de “rectificado” y “habilitado” ya que la secretaria de energia no proporciona información al respecto de qué significan dichas variables o que interpretación se puede elaborar sobre el significado de ellas, además de que para el análisis posterior, el no tomarlas en cuenta no altera en nada el análisis o el proceso de

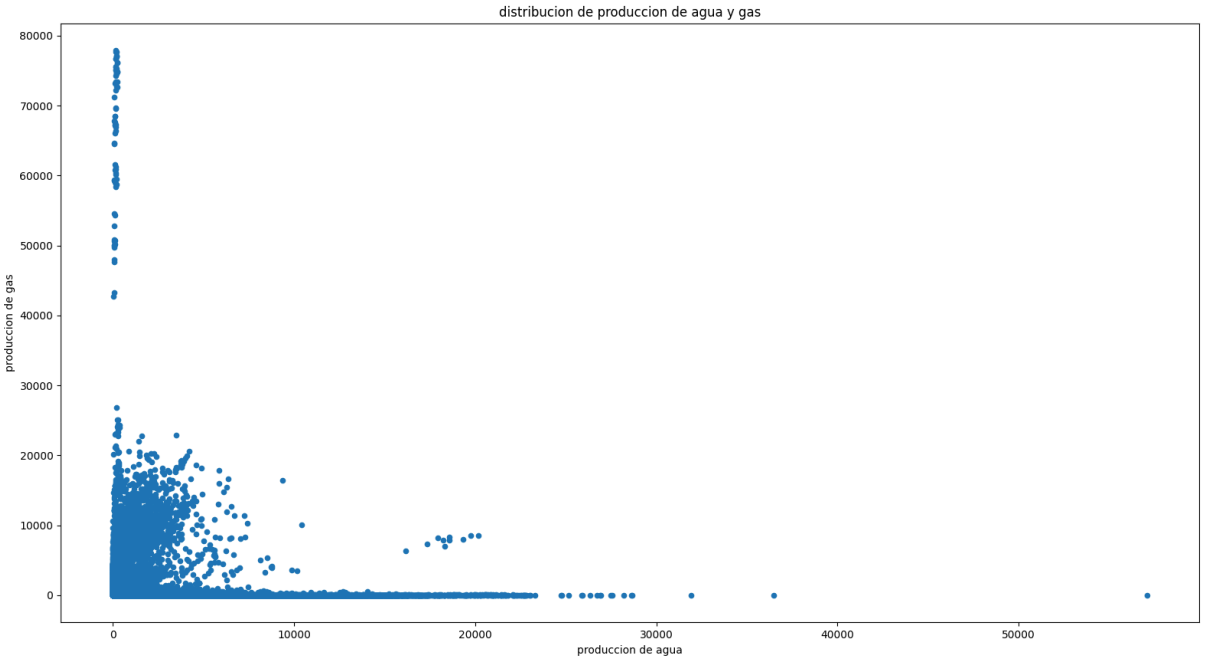
minería de datos, ya que los valores únicos que presentan cada columna son “f” o “t” respectivamente y aparecen en todos los registros.

El análisis inicial del dataset reveló que algunas columnas presentan valores faltantes, mientras que otras contienen datos inconsistentes o fuera de rango. Se pueden identificar una serie de valores fuera de lo normal en varios campos de lo cuales serán de interés en las fases posteriores, por ejemplo, los 3 valores de producción, en el caso del petróleo, valores de producción de 15 mil metros cúbicos por mes son absurdos, por lo general un pozo de petróleo puede producir en promedio entre 100 y 500 metros cúbicos en un mes, lo mismo se podría decir del gas y del agua, en el caso de este último destaca que la producción de agua mínima es un valor negativo, lo cual no tiene sentido, en el apartado de preparación de datos se detalla qué método se empleó para manejar estos valores fuera de lo común o outliers.

A continuación se presentan una serie de gráficos con los datos iniciales, tomando en cuenta variables de interés para el proyecto:





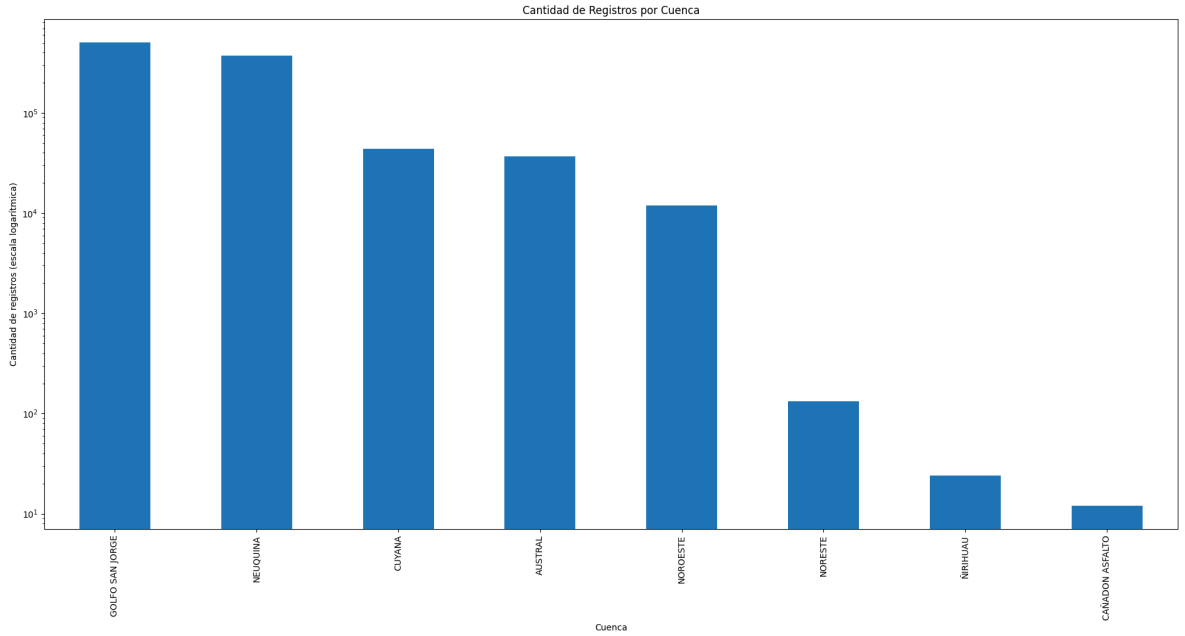


En estos dos últimos gráficos puede observarse la distribución de los registros de producción de petróleo, agua o gas.

Se puede observar por ejemplo en el primer gráfico (petróleo-agua) como hay una gran concentración de pozos que producen relativamente poco petróleo, y altas cantidades de agua, y como las cantidades de agua parecen converger a valores en común para producciones más altas de petróleo. Esto podría deberse a que los pozos maduros o cercanos al fin de su vida util empiezan a producir cada vez más agua.

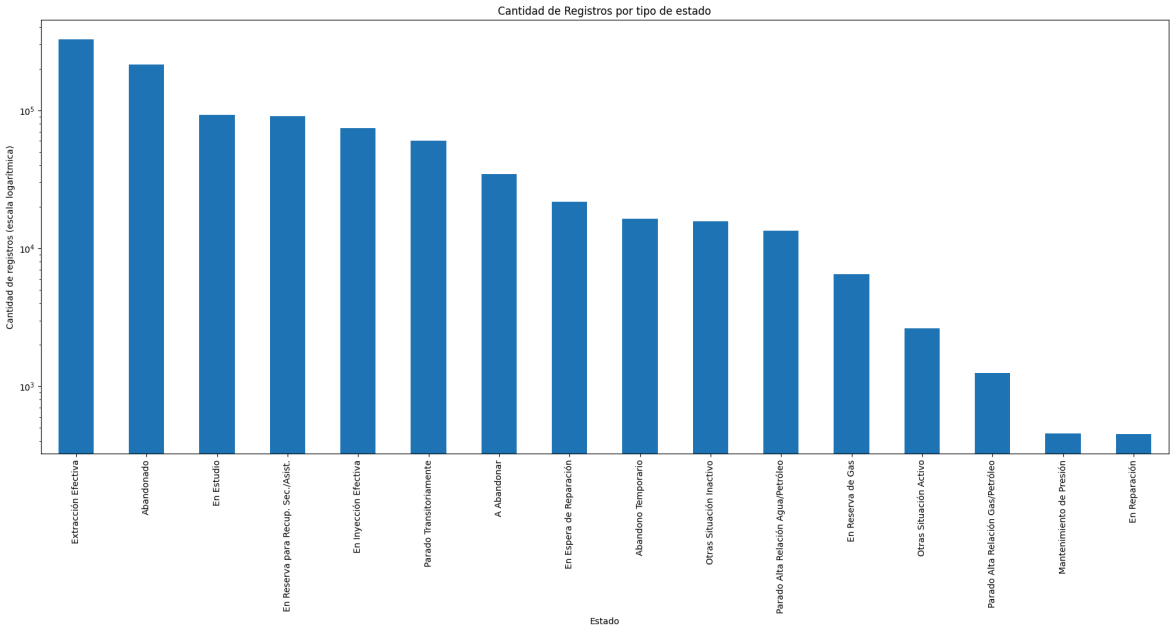
En el caso de agua-gas, se puede observar una gran concentración de registros en rangos de producción bajos de agua y gas, y la existencia de puntos más aislados tanto para gas como para agua, puntos que serán tratados en la fase 3 del informe.

Por último, para el caso de petroleo-gas, se observa mayores dispersiones de datos, en los que a mayor producción de petróleo disminuye la de gas, y viceversa. Esto en parte se debe a que en pozos de petróleo nuevos en un principio lo único que se produce es agua y petróleo, con nulas cantidades de gas como fase libre, y conforme se sigue explotando el pozo, y vaya bajando la presión del reservorio, la producción de gas en general irá en aumento.

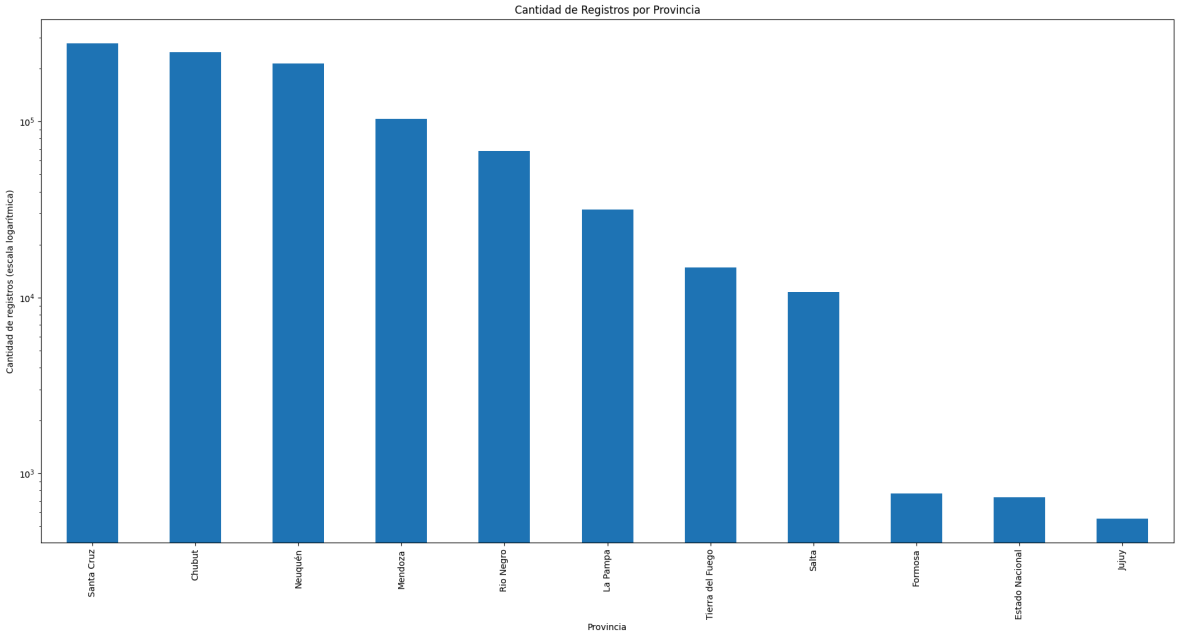


En el histograma se puede apreciar que la mayoría de los registros se concentran en las cinco principales cuencas productivas de Argentina, las cuales fueron descritas previamente en la fase 1 del informe. Entre ellas, destacan el Golfo San Jorge, ubicado en el sur del país, conocido por su producción predominantemente convencional, y la Cuenca Neuquina, famosa por su producción no convencional, en la que se encuentra la reconocida Formación Vaca Muerta.

Cuencas diferentes a las señaladas anteriormente, serán descartadas en la fase posterior.



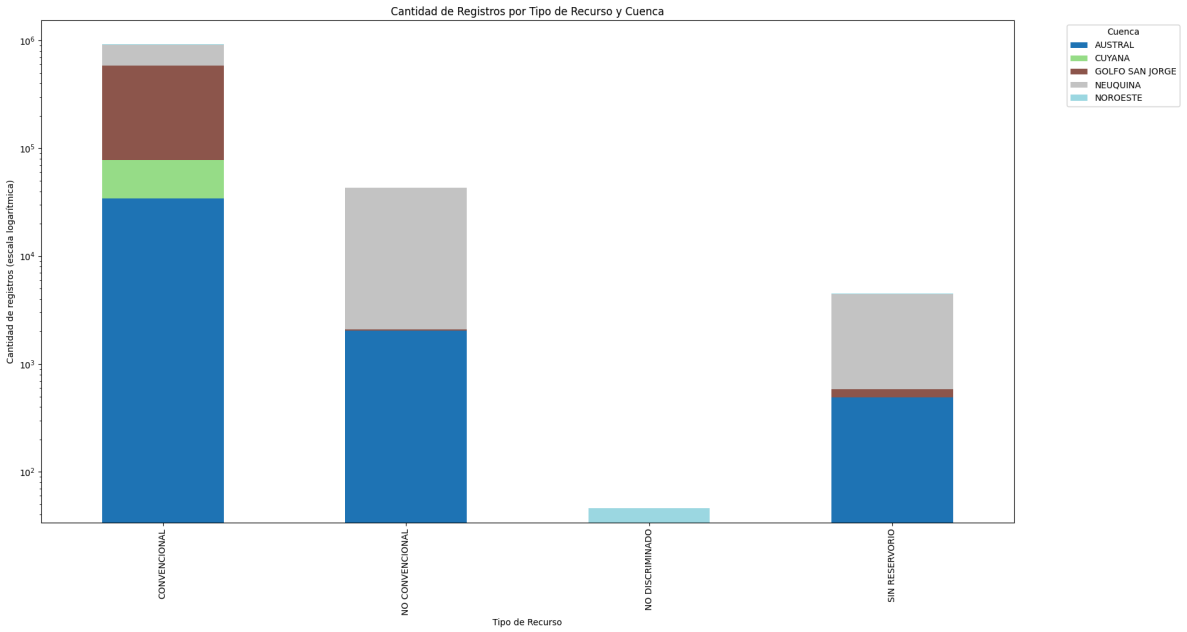
El gráfico que analiza los estados de los pozos muestra que las categorías más frecuentes son “Extracción efectiva”, que corresponde a pozos en plena producción y “Abandonado”. Los demás estados corresponden a clasificaciones más específicas, las cuales en la fase posterior del informe serán descartadas, y solo se utilizaran “Extracción Efectiva” y “Abandonado” como parámetros para determinar que un pozo es productivo o no, o dicho en términos comerciales y haciendo alusión al objetivo del informe, si conviene o no invertir en tal pozo.



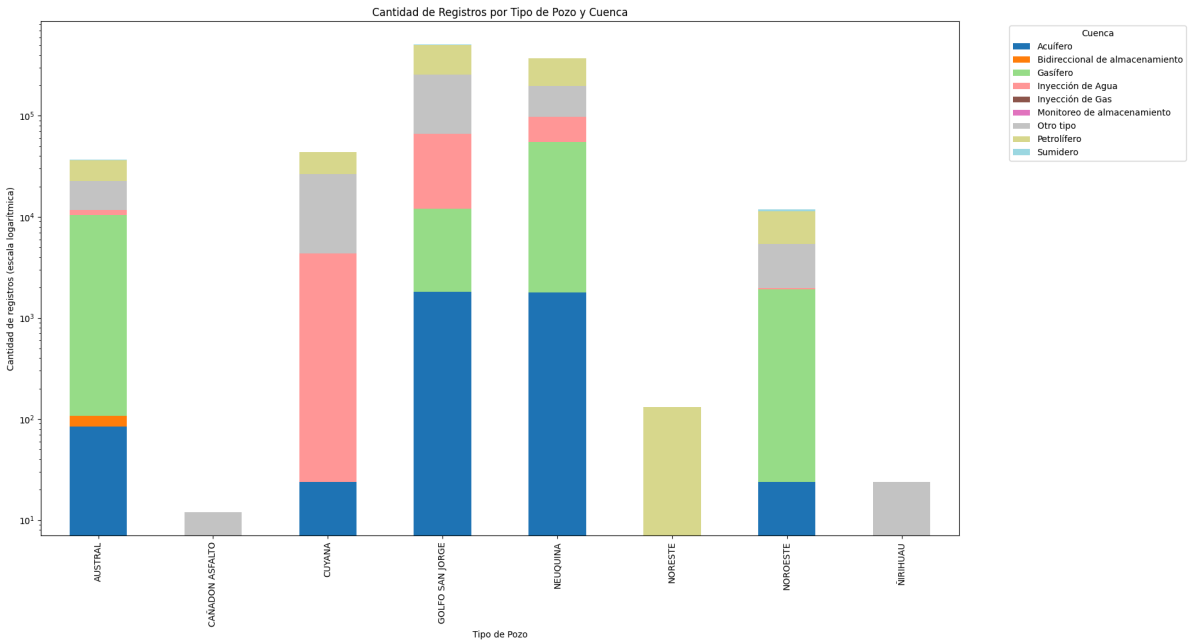
En cuanto a la distribución por provincias, se observa que las principales productoras de hidrocarburos, es decir, las que concentran la mayor cantidad de registros, son Santa Cruz y

Chubut, situadas en el sur del país y asociadas a las cuencas Austral y del Golfo San Jorge, respectivamente. Les siguen Neuquén y Mendoza, vinculadas a las cuencas Neuquina y Cuyana, y Salta, relacionada con la cuenca del Noroeste.

Cabe aclarar que no es cantidad de pozos por provincia, sino cantidad de registros, ya que como el dataset abarca la producción de todos los pozos de la Argentina durante el 2023, cada pozo puede tener más de un registro asociado, es decir, si el pozo es “nuevo”, y haya empezado su producción en diciembre de 2023, tendrá un solo registro, mientras que si es un pozo más antiguo, tendrá como máximo 12 registros.



El análisis de los registros según el tipo de recurso y la cuenca indica que la Cuenca Austral es predominante tanto en producción convencional como no convencional. Esto tiene sentido debido a su relevancia como una de las cuencas más productivas de Argentina. Sin embargo, cabe señalar que en conjunto con lo que se señaló en el párrafo anterior, este gráfico considera todos los registros, incluyendo pozos activos, inactivos, en reparación o en otros estados, por lo que el análisis no es del todo riguroso. Para el tipo de recurso, se tomará únicamente Convencional y no convencional como tipos de estado para la fase posterior.



En este último gráfico se observa como hay una predominancia clara de pozos de gas respecto a los de petróleo en las 5 principales cuencas de la argentina, y sobretodo de pozos acuíferos, la existencia de estos pozos nace también de la necesidad de extraer agua de alguna fuente para emplearla en alguna otra área, por ejemplo, se puede utilizar el agua como método de recuperación de hidrocarburos inyectandola devuelta a la formación para “barrer” con el petróleo restante, o en el empleo de métodos no convencionales de producción, esto se ve claramente en la cuenca neuquina, que en su mayoría son pozos de este tipo.

Fase 3: Preparación de los datos

Inclusión / Exclusión de Datos

La selección de variables y registros se basó en los objetivos comerciales definidos en la fase 1, que incluyen maximizar la producción de petróleo y/o gas, reducir los costos asociados al manejo del agua y optimizar la rentabilidad de las inversiones en pozos productivos. Por ello, se prioriza incluir datos que permitan identificar patrones relevantes de producción y características operativas de los pozos.

Se identificaron las siguientes columnas necesarias para el análisis.

- Producción: prod_pet, prod_gas y prod_agua, que representan las cantidades producidas de petróleo, gas y agua respectivamente.

- Información Descriptiva: tipopozo, tipo_de_recurso, tipoestado, provincia y cuenca que brindan detalles sobre el tipo de pozo, recurso, estado y la ubicación geográfica.
- Temporalidad: mes.

Estas variables fueron seleccionadas ya que las consideramos necesarias para identificar patrones de producción en función de (principalmente) las características geográficas y temporales, así como para explorar diferencias entre tipos de pozos y recursos.

Por otro lado, se optó por excluir todas las demás columnas que no aportan información relevante al análisis, o que pudiesen complejizar el mismo, tal es el caso de “vida_util”, “observaciones” o “proyecto”, en el que predominan no sólo una gran cantidad de valores faltantes como el caso de “vida_util” u “observaciones”, sino que en el caso de proyecto la gran mayoría de registros como se ve en el apartado de la fase 2, tienen como valor “sin proyecto”, otros como “rectificado” y “habilitado” si bien no tienen valores faltantes, tienen un único valor en dicha columna en todos los registros, los cuales no se pudo obtener una clara interpretación de lo que representan ya que no hay información al respecto que haya proporcionado la secretaria de energia.

El qué empresa esté operando ese pozo no es relevante a priori para determinar la rentabilidad de un pozo, tampoco el metodo de extraccion (tipoextraccion), ya que un pozo (en terminos simples) es rentable o no si da mas o menos hidrocarburos, y el incluir si tiene algun metodo de extraccion para mejorar esa rentabilidad puede complejizar el análisis, esto aplica de igual forma a las columnas de inyección de fluidos, como agua, CO2, u otros.

En cuanto a factores geográficos, excluimos la formación productiva, la concesión y el yacimiento, ya que el enfoque del informe está centrado a nivel provincial/cuenca, el incluir la gran lista de concesiones diferentes que puede haber en cada cuenca, haría más inexacto el análisis, y podría llegar a dar lugar a incoherencias en los resultados, como por ejemplo, que un grupo de pozos se ubique en una cuenca en específico o alrededores, y que la concesión corresponda a otra provincia u otra cuenca más distante, además de que consideramos que la categorización geográfica existente en los valores de cuenca y provincia son más que suficientes para elaborar las observaciones que estamos buscando.

La temperatura estabilizada de formación, la profundidad , el subtipo de recurso, las fechas de ingresos de datos, o de inicio de producción e identificadores, no aportan prácticamente nada al análisis, por lo cual serán descartadas.

Variables finales incluidas:

- mes
- prod_pet
- prod_gas
- prod_agua
- tipoestado
- tipopozo
- cuenca
- provincia
- tipo_de_recurso

Variables excluidas:

- idempresa
- anio
- idpozo
- iny_agua
- iny_co2
- iny_otro
- tef
- vida_util
- tipoextraccion
- observaciones
- fechaingreso
- rectificado

- habilitado
- idusuario
- empresa
- sigla
- formprod
- profundidad
- formacion
- idareapermisoconcesion
- areapermisoconcesion
- idareayacimiento
- areayacimiento
- proyecto
- clasificacion
- subclasificacion
- sub_tipo_recurso
- fecha_data

Estas reglas aseguran que el análisis se centre en pozos con operaciones relevantes y en áreas con potencial productivo significativo.

Limpieza de Datos

El dataset original contenía registros que podían introducir ruido o inconsistencias en el análisis. Para garantizar la relevancia de los datos, se aplicaron los siguientes filtros:

- Solo se incluyeron registros con valores mayores o iguales a 0 en las columnas (prod_pet, prod_gas, prod_agua), esto es debido a que columnas como la de agua, presentaban valores negativos.

- Se descartaron registros que no tengan ningún valor en los campos de cuenca, o provincia, o que tengan valores como “SIN RESERVORIO”.
- Se descartaron registros que tuviesen en el campo “cuenca”, valores diferentes a los de las 5 principales cuencas productivas (Neuquina, Cuyana, Golfo San Jorge, Austral, y Noroeste) ya que la producción de otras cuencas distintas a éstas son prácticamente despreciables.
- Solo se incluyeron pozos en estados de "Extracción Efectiva" y en “Abandono” de esta forma podemos categorizar tanto a pozos que sean productivos como no productivos en base a esas dos categorizaciones de una manera simple, sin entrar en otras mas especificas, como “En reparacion” o “A espera de abandono”.
- Se seleccionaron pozos clasificados como "petrolifero" o "gasífero".
- Se incluyeron solo recursos "Convencionales" y "No Convencionales".
- Se eliminaron registros con valores nulos en todas las columnas clave (prod_pet, prod_gas, prod_agua)
- Se utilizaron percentiles de 95% en cada uno de los campos de producción para filtrar valores mayores a estos, esto se tomó como una alternativa a plantear un valor máximo posible de manera arbitraria, de esta forma, valores mayores a los adoptados por estos percentiles se descarta ya que no se consideran típicos de la producción de hidrocarburos en la argentina.

Esto garantiza que los datos reflejan tendencias generales sin ser influenciados por registros anómalos.

Derivación de Atributos

La creación de nuevos atributos permite descubrir relaciones y patrones que no serían evidentes con las variables originales.

- **Trimestre**: Se generó a partir de la columna correspondiente al mes. Con el fin de poder simplificar en períodos específicos del año.
- **Segmentación por rangos de producción:** Usando los percentiles del 95%, se los dividió en tres rangos de iguales proporciones, en “baja”, “media” y “alta”. Es decir, sea “p95” el valor numérico del percentil 95 del campo prod_pet, si un registro tiene un valor de

producción de petróleo menor o igual a “p95”/3, tendrá un rango de producción bajo. De esta forma se generaron 3 nuevas columnas que corresponden a los rangos de producción de petróleo, gas, y agua y descartamos el uso de las columnas de producción.

Formateo de Datos

En este paso se reestructuró el dataset de tal forma que tenga compatibilidad con el modelo de SOM. Y así lograr minimizar la disparidad de los registros sin que generen “ruido”. Permite que cada variable tenga la misma influencia en el mapa SOM, sin que se generen sesgos.

Codificación de variables categóricas:

Se transforman todas las variables categóricas a numéricas para que el modelo (SOM) entienda las relaciones entre categorías. Mediante el siguiente código:

```
label_encoders = {}
for column in ['tipo_de_recurso', 'tipoestado', 'tipopozo', 'cuenca', 'provincia', 'tipo_de_recurso', 'rango_p', 'rango_g', 'rango_a']:
    le = LabelEncoder()
    info[column] = le.fit_transform(info[column])
    label_encoders[column] = le
```

quedando los datos transformados de la siguiente manera:

Columna	Categoría	Valor Numérico
tipo_de_recurso	convencional	0
	no convencional	1
tipoestado	Abandonado	0
	Extracción Efectiva	1
tipopozo	Gasífero	0
	Petrolífero	1
cuenca	AUSTRAL	0
	GOLFO SAN JORGE	1
	NEUQUINA	2
	CUYANA	3
	NOROESTE	4
provincia	Tierra del Fuego	0
	Santa Cruz	1
	Chubut	2
	Rio Negro	3
	Neuquén	4
	La Pampa	5
	Mendoza	6
	Salta	7
	Jujuy	8
	Formosa	9
rango_p	baja	0
	media	1
	alta	2
rango_g	baja	0
	media	1
	alta	2
rango_a	baja	0
	media	1
	alta	2

Esta tabla de datos, es la que se usará como referencia para interpretar los resultados obtenidos en la fase de modelado.

Normalización:

Todas las variables fueron normalizadas para que tengan media 0 y desviación estándar 1, utilizando “StandardScaler”, todo esto para garantizar que todas tengan la misma importancia en el modelo, y no haya diferencias marcadas en las escalas de valores de cada atributo.

El formateo asegura que el modelo SOM interprete correctamente las relaciones entre las variables sin sesgos.

```
#Normalización de los datos
scaler = StandardScaler()
data_normalized = scaler.fit_transform(data_selected)
```

```
# Crear un DataFrame con los valores normalizados
data_normalized_df = pd.DataFrame(data_normalized, columns=data_selected.columns)
```

Descripción del Dataset Obtenido

El dataset final contiene:

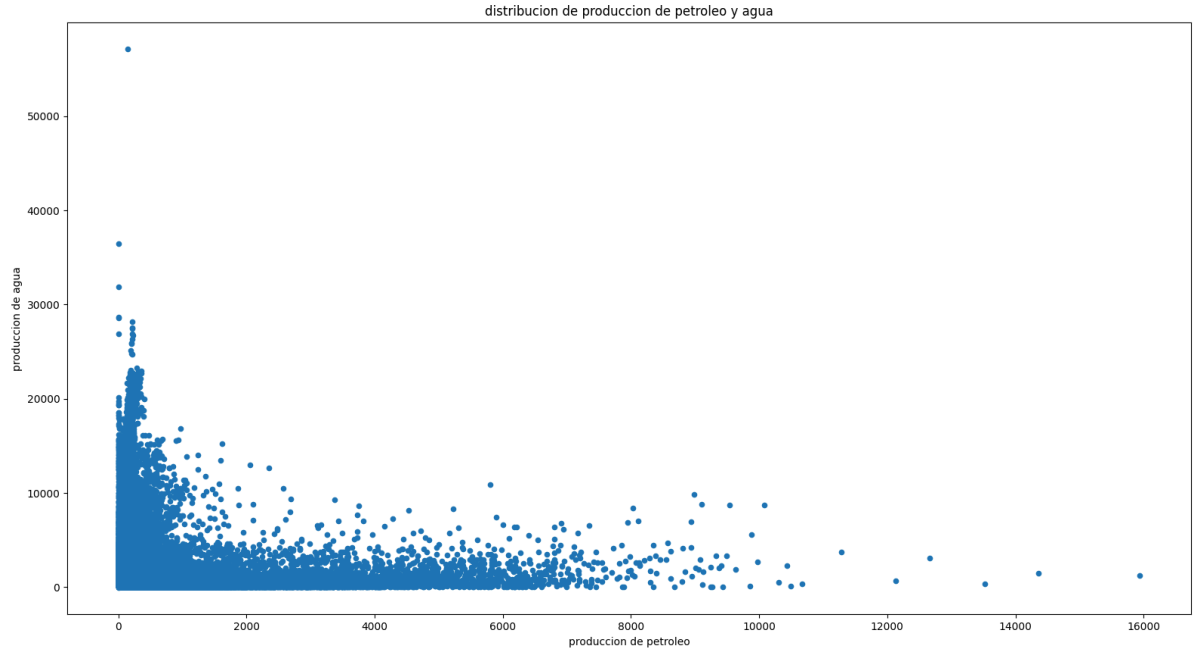
- **Número de registros:** 306,165
- **Número de columnas:** 9 columnas principales.
- **Variables clave:** Rangos de producción de petróleo (rango_p), gas (rango_g), agua (rango_a), trimestre, provincia, cuenca, tipo_de_recurso, tipoestado, tipopozo.

Indice	Nombre	Valores no nulos	Tipo de dato
0	mes	306165	int64
1	tipoestado	306165	object
2	tipopozo	306165	object
3	cuenca	306165	object
4	provincia	306165	object
5	tipo_de_recurso	306165	object
6	trimestre	306165	int64
7	rango_p	306165	object
8	rango_g	306165	object
9	rango_a	306165	object

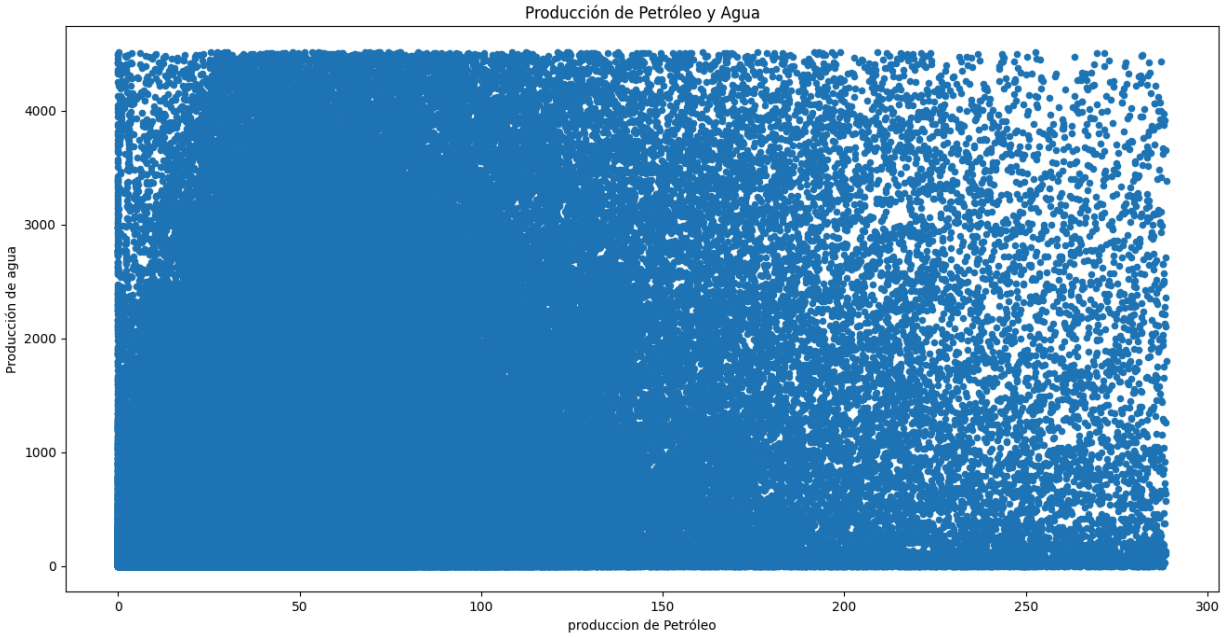
	rango_p	rango_g	rango_a	trimestre	cuenca	tipoestado	po_de_recur	tipopozo	provincia
count	306165,00	306165,00	306165,00	306165,00	306165,00	306165,00	306165,00	306165,00	306165,00
mean	-0,05	0,15	-0,05	-0,06	-0,01	0,43	0,00	0,10	0,03
std	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00
min	-0,35	-0,20	-0,41	-1,33	-2,22	-3,01	-0,25	-2,67	-1,62
25%	-0,35	-0,20	-0,41	-1,33	-0,72	0,33	-0,25	0,37	-1,05
50%	-0,35	-0,20	-0,41	-0,44	-0,72	0,33	-0,25	0,37	-0,47
75%	-0,35	-0,20	-0,41	0,46	0,78	0,33	-0,25	0,37	0,68
max	4,35	6,71	3,42	1,35	3,77	0,33	3,97	0,37	3,55

A continuación vamos a demostrar el Antes y Después luego de la preparación de los datos con el fín de poder resaltar los resultados para el posterior análisis en la etapa de modelado:

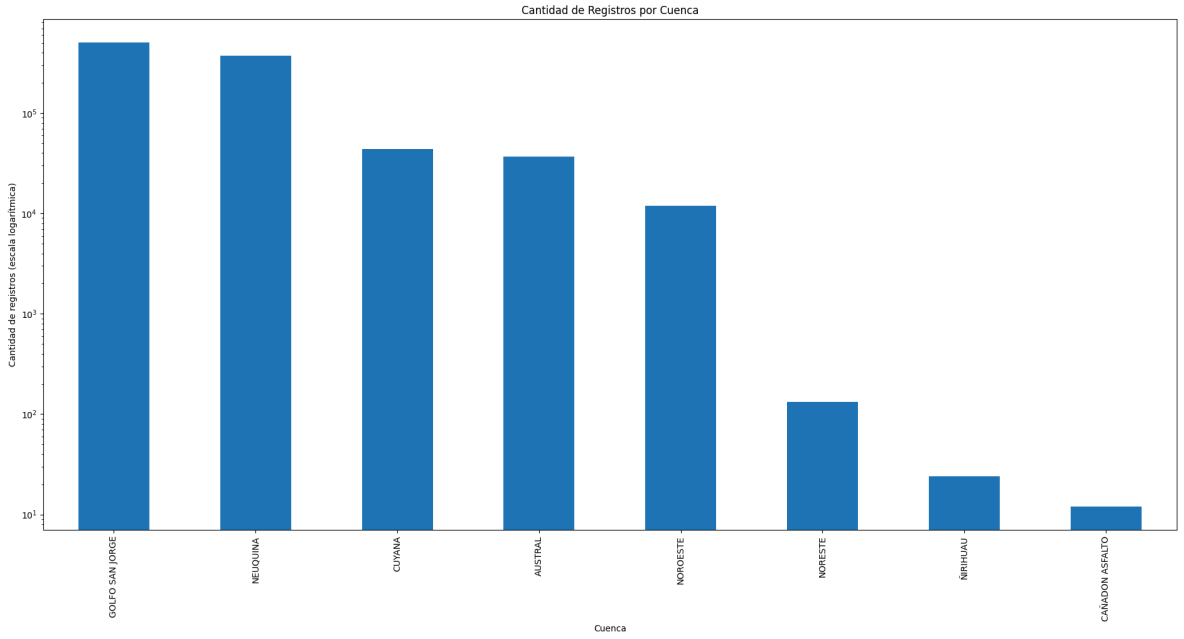
Antes del Filtrado de Registros - Limpieza de Datos:



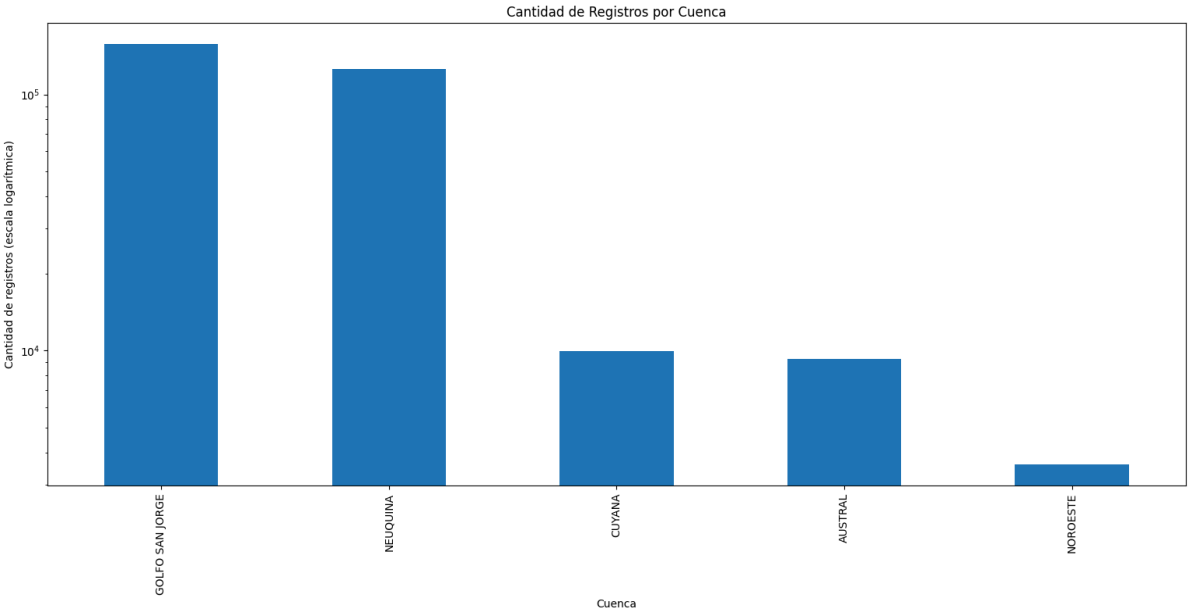
Después del Filtrado de Registros - Limpieza de Datos:



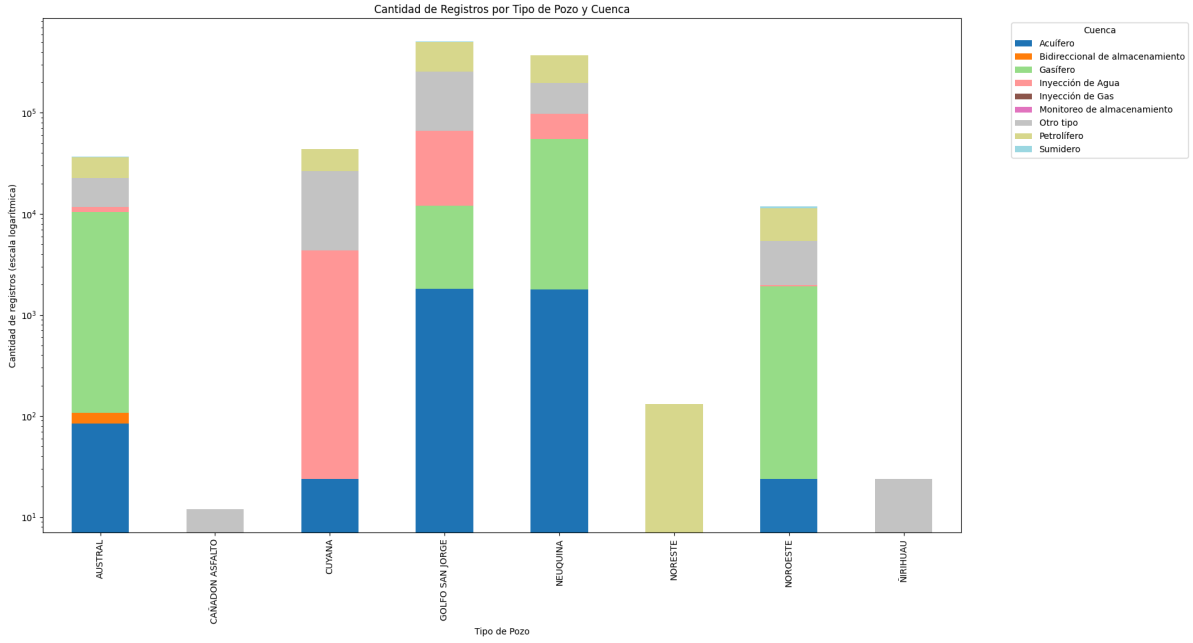
Antes de Inclusión/ Exclusión de Datos - Limpieza de Datos:



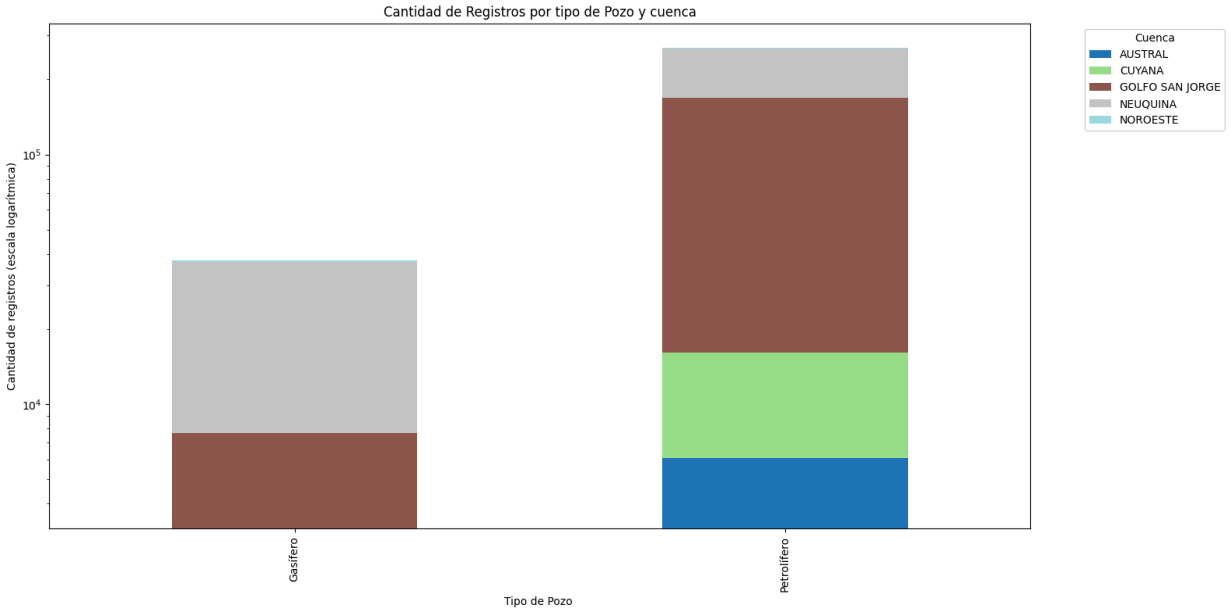
Después Inclusión/ Exclusión de Datos - Limpieza de Datos:



Antes Inclusión/ Exclusión de Datos - Limpieza de Datos:



Después Inclusión/ Exclusión de Datos - Limpieza de Datos:



Fase 4: Modelado

Selección del Modelo

Siguiendo con la metodología CRISP, una vez determinado el DataSet definitivo, se procede a seleccionar el modelo a emplear.

En este caso vamos a aplicar el algoritmo de Self Organization Maps (SOM), el cual es un tipo de red neuronal de aprendizaje no supervisado competitivo. En otras palabras, la red no se comporta de manera correcta o incorrecta dado que no hay ningún atributo clase hacia el cual la red deba tender, sino que su finalidad es descubrir rasgos comunes, regularidades, y categorías en los datos de entrada.

Las neuronas de la red se organizan en función de los estímulos externos. Es un aprendizaje competitivo porque las neuronas compiten para activarse, quedando una como neurona vencedora y teniendo como objetivo categorizar los datos que se introducen en la red.

Los valores similares se clasifican dentro de la misma categoría, por lo que todos ellos activarán a la misma neurona de salida siendo las clases o categorías creadas por la red.

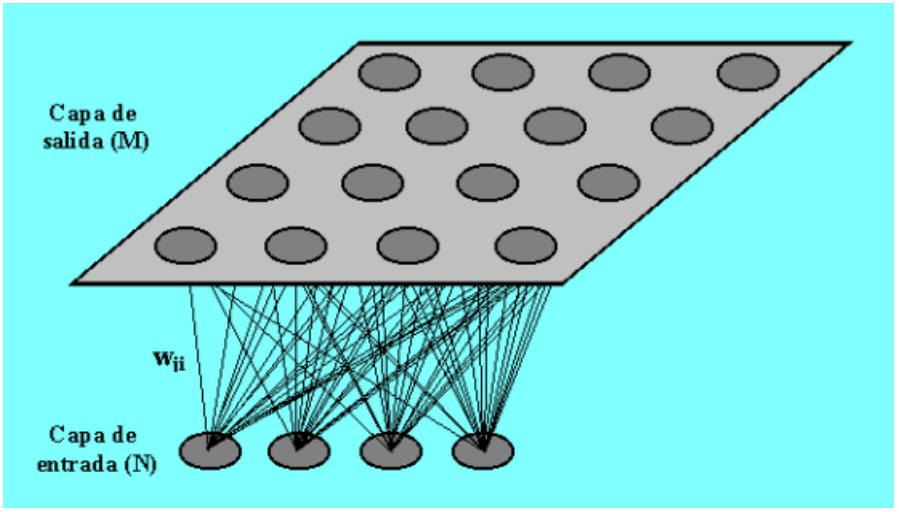
En cuanto a la arquitectura, el algoritmo se dividen dos capas:

- La primera capa está formada por los nodos de aprendizaje, la cual contiene la información acerca de la representación resultante, manteniendo los datos asociados a los datos de entrada.

- La segunda capa la forman los nodos de entrada, que servirán de alimentadores de los nodos de aprendizaje por medio de los datos de entrada (vectores originales) durante el proceso de entrenamiento.

Desde el punto de vista de redes neuronales, diríamos que es una capa densa. Además, todos los elementos de la primera capa están conectados con todos los elementos de la segunda capa. Estas conexiones tienen un peso que puede modificarse para que los nodos de aprendizaje aprendan la información de los nodos de entrada.

La siguiente figura esquematiza la arquitectura para un entrenamiento SOM, donde la red de aprendizaje viene representada por los nodos superiores (capa de salida), y los vectores de entrenamiento (capa de entrada) vienen representados en el sector inferior.



Las conexiones son siempre hacia adelante, es decir, desde la capa de entrada hacia la de salida. Cada neurona de la capa de entrada “i” está conectada con otra de la capa de salida “j”, mediante un peso Wji. De esta forma la neurona de la capa de salida tiene asociado un vector de pesos W.

Descripción de Parámetros Configurados

En cuanto a los parámetros evaluados del algoritmo se destacan los siguientes:

1) **Tamaño del mapa SOM**

   - Define las dimensiones del mapa SOM como una cuadrícula.
   - Cada nodo representa un clúster que agrupa puntos similares del conjunto de datos.
   - Cuanto mayor sea el tamaño, más granular será la representación. Es decir que el algoritmo puede capturar patrones más específicos pero requiere más datos y tiempo de entrenamiento.
   - Vimos conveniente usar una cuadrícula de 20x20 ya que es un tamaño estándar y además de que si sumamos más nodos, no solo no contribuye significativamente al análisis de resultados sino que complejiza la interpretación de los mismo de sobremanera.

2) **Dimensión de Entrada**

   - Se refiere a la cantidad de características que presentan los datos de entrada.

- Asegura que cada nodo tenga un vector de pesos de la misma dimensión que las entradas.

3) Radio de Vecindad Inicial

- Controla cuánto influyen los nodos vecinos en la actualización de un nodo durante el entrenamiento.
- Un valor alto implica que más nodos serán afectados inicialmente, lo que ayuda a estabilizar el entrenamiento.
- A medida que el SOM se entrena, este radio se reduce gradualmente.

4) Tasa de Aprendizaje

- Define cuánto cambian los pesos de un nodo en respuesta a un ejemplo.
- También disminuye con el tiempo para asegurar que el SOM converge hacia una solución estable.

5) Pesos de los Nodos

- Es un paso fundamental en la secuencia del aprendizaje y consiste en asignar valores iniciales a los pesos para que después vayan variando al ir aumentando el número de iteraciones.

6) Cantidad de Clusters

- Define la cantidad de grupos en los cuales se agrupan los diferentes comportamientos de los datos. A menor número de clusters, más datos son agrupados, y la información que se puede extraer de cada uno es más generalizada, mientras que a mayor número de estos permite captar más matices en los datos, y patrones más específicos, al igual que en el caso de la elección del número de nodos, creímos conveniente utilizar 8 clusters ya que presenta un balance entre generalidad y practicidad en el análisis.

7) Cantidad de Iteraciones

- Cada iteración selecciona un ejemplo aleatorio del conjunto de datos y realiza las siguientes operaciones: primero encuentra el nodo más cercano, luego calcula la distancia euclidiana entre el ejemplo y todos los nodos del SOM para

finalmente seleccionar el nodo cuya distancia sea mínima. A continuación, actualiza los pesos y sus vecinos.

A modo de resumen, a continuación se muestran los parámetros que fueron utilizados en la tabla siguiente:

Parámetro	Valor
Tamaño del Mapa	20x20
Dimensión de Entrada	9
Radio de Vecindad Inicial	1
Tasa de Aprendizaje	0,5
Pesos de los Nodos	Aleatorios
Cantidad de Clusters	8
Cantidad de Iteraciones	1000

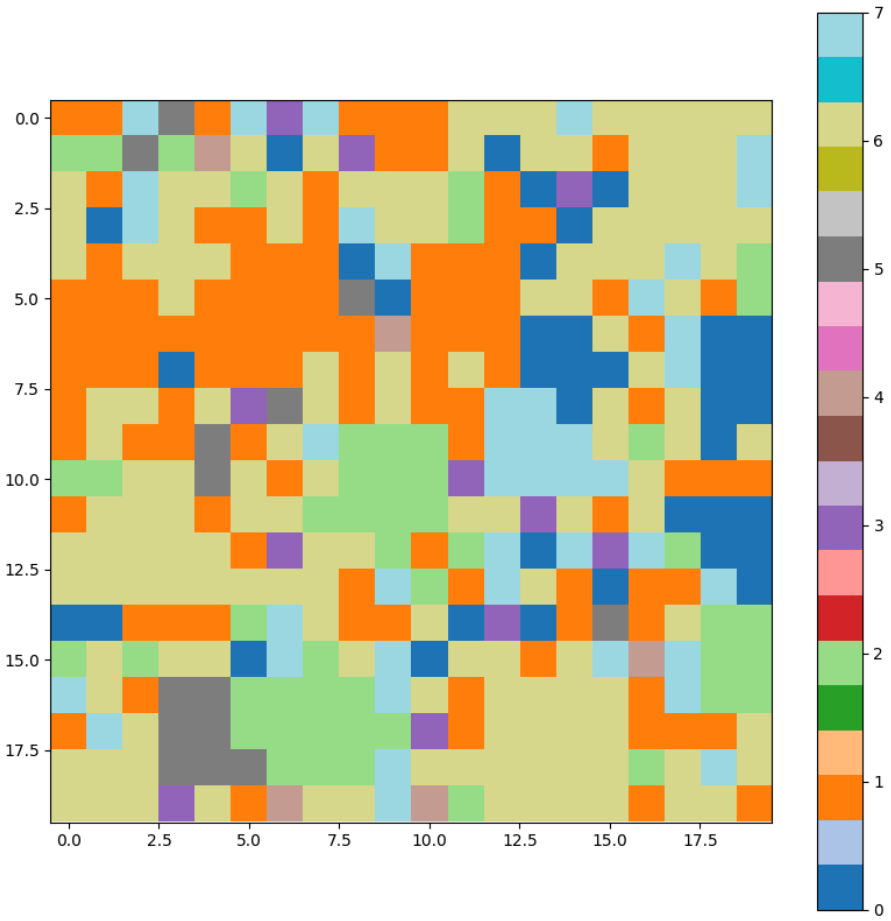
Resultados y Evaluación del Modelo

Para comenzar con el análisis de los resultados obtenidos, en primer lugar se procederá a analizar un mapa de calor, donde cada celda representa una neurona del SOM. Las celdas están coloreadas para indicar las categorías o agrupaciones detectadas durante el proceso de aprendizaje.

Dado que el mapa es de 20x20, el modelo presenta 400 neuronas en total, lo que permite detectar una buena variedad de patrones y subgrupos. Cada color representa un grupo o clúster al que pertenecen los datos después de ser procesados por el SOM.

Las regiones con el mismo color o colores similares indican que los datos en esas áreas tienen características similares y en consecuencia, se agruparán en un clúster.Mientras que las transiciones entre colores muestran cambios en las propiedades de los datos.

A continuación se presenta el mapa de calor obteniendo al emplear el algoritmo:



- ➔ Como podemos observar en la cuadrícula situada a la derecha, se determinan 8 clúster para el conjunto de datos.

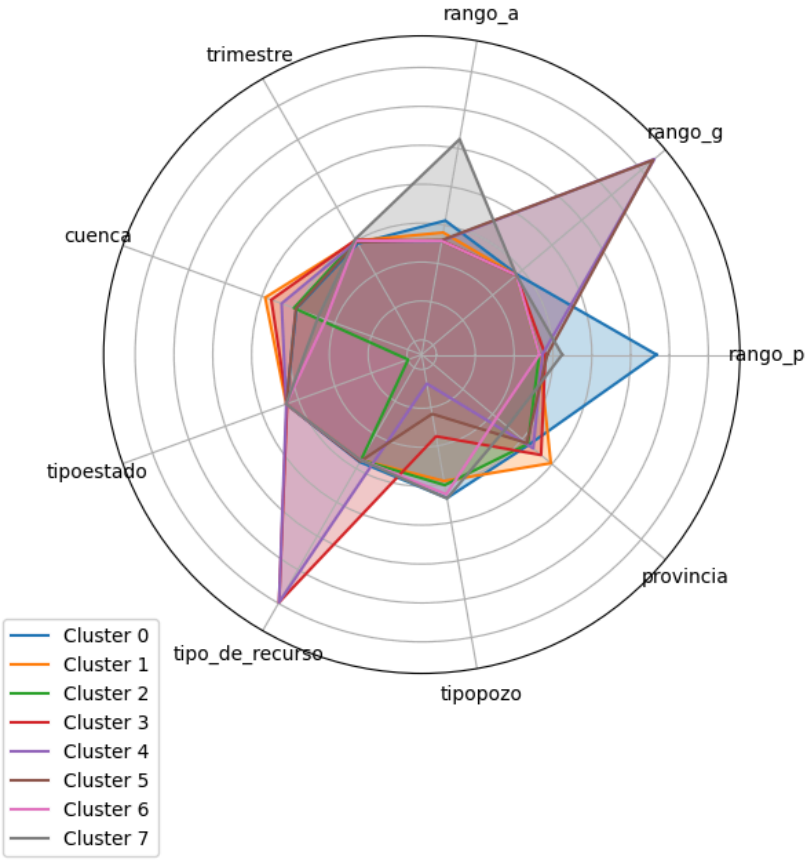
- ➔ En la parte superior izquierda del gráfico puede verse una fuerte predominancia del color naranja (1) y en otras el color verde claro (6). Estos dos colores son los que abarcan una mayor proporción en la imagen.

Como siguiente paso se procede a realizar la construcción de un gráfico radial, el cual es de gran utilidad para comparar múltiples variables a través de diferentes categorías o clusters. En ellos cada eje radial corresponde a una variable, los diferentes colores representan clusters generados por el modelo y las líneas muestran cómo cada cluster se comporta en relación con dichas variables. Para el caso en que se presente una superposición de clusters sugiere similitudes entre ciertos grupos en variables específicas.

A continuación se muestra el gráfico mencionado anteriormente con datos normalizados, lo que asegura que las variables tengan un peso equitativo en el análisis de los clusters y evita

que alguna con valores más grandes domine el resultado. Con la normalización ya aplicada, el comportamiento que se observa en el gráfico radial refleja verdaderas diferencias relativas entre los clusters en función de las variables.

Para la interpretación de los valores normalizados, las puntuaciones en cada eje representa qué tan lejos está el cluster del valor promedio en cada variable. Así, los valores cercanos al borde del gráfico indican que ese cluster tiene un comportamiento más extremo en la variable correspondiente, mientras que los que están cerca del centro sugieren que el cluster tiene un comportamiento más equilibrado en esa dimensión.



Analizando el gráfico rápidamente podemos concluir los siguientes puntos:

- ➔ Los valores extremos para los rangos de producción de petróleo, gas y agua se dan para los clústers 0, 4 y 7 respectivamente. Esto indica que hay grupos de pozos con comportamientos muy diferentes respecto de dichas variables.
- ➔ En la variable trimestre hay una superposición significativa entre clusters. Esto podría significar que esta variable no es tan determinante para separar los grupos, o bien que representa patrones comunes en el Dataset.

Para abordar el análisis de los resultados obtenidos con mayor profundidad se creará una tabla con los valores promedio en la escala original para cada atributo según los clústers obtenidos al emplear el algoritmo.

Cluster	rango_p	rango_g	rango_a	trimestre	cuenca	tipoestado	tipo_de_recurso	tipopozo	provincia
0	1,28	0,01	0,28	2,4	1,52	1	0,01	1	3,21
1	0	0	0,11	2,49	2,09	1	0	0,85	4,49
2	0	0	0	2,49	1,56	0	0	0,89	3,15
3	0,07	0	0	2,52	1,98	0,99	1	0,47	3,92
4	0,03	1,34	0	2,47	1,79	1	1	0,02	3,47
5	0,09	1,33	0,01	2,44	1,53	1	0,01	0,28	3,09
6	0	0	0	2,49	0,98	1	0	0,96	1,33
7	0,25	0	1,39	2,54	1,12	1	0	1	1,96

Clúster 0: Corresponde al agrupamiento de pozos con mayor producción de petróleo, con una producción de agua relativamente alta respecto a los demás agrupamientos. Según las variables tipo_de_recurso y tipopozo es del tipo convencional y petrolífero respectivamente. Respecto de la ubicación geográfica de este clúster, destacamos que la mayor parte de estos pozos pertenecen a la Cuenca Neuquina aunque también se destaca una presencia de la Cuenca del Golfo San Jorge. Respecto a las provincias en cuestión tenemos gran preponderoncia por parte de Río Negro.

Clúster 1: Resalta el agrupamiento de pozos, en mayor proporción de petróleo y del tipo convencional que presentan una elevada producción de agua respecto a los demás clústers. Geográficamente, este comportamiento corresponde a la Cuenca Neuquina y en mayor detalle a las provincias de Neuquén y La Pampa. Podríamos decir que representa un grupo de pozos de muy baja productividad, y por consiguiente de poca o nula rentabilidad, debido a sus bajos valores de producción.

Clúster 2: agrupa aquellos pozos convencionales abandonados, en mayor proporción de petróleo, que están presentes en la Cuenca del Golfo San Jorge y Neuquina. En mayor proporción son de la provincia de Río Negro. Al igual que el cluster 1, son grupos de pozos de rentabilidad nula ya que en su mayoría son pozos abandonados.

Clúster 3: Corresponde a los pozos con un nivel de producción de petróleo en su mayoría bajos, respecto a los demás agrupamientos, per con algunos registros que mejoran ligeramente el valor promedio de producción de petróleo del conjunto, pudiendo asumir que en términos generales, este cluster es mas productivo en términos de este ultimo recurso que el cluster 1 y 2. Corresponde a la Cuenca Neuquina, tanto a pozos gasíferos como petrolíferos, y de

reservorios no convencionales. En gran proporción se trata de pozos que provienen de la provincia de Neuquén.

Clúster 4: Se trata de los pozos con mayor producción de gas, siendo en mayor proporción pozos gasíferos y con bajo contenido de agua. Según el tipo de recurso y los campos geográficos se trata de reservorios no convencionales de la cuenca neuquina y en mayor proporción de la provincias de Río Negro y Neuquén.

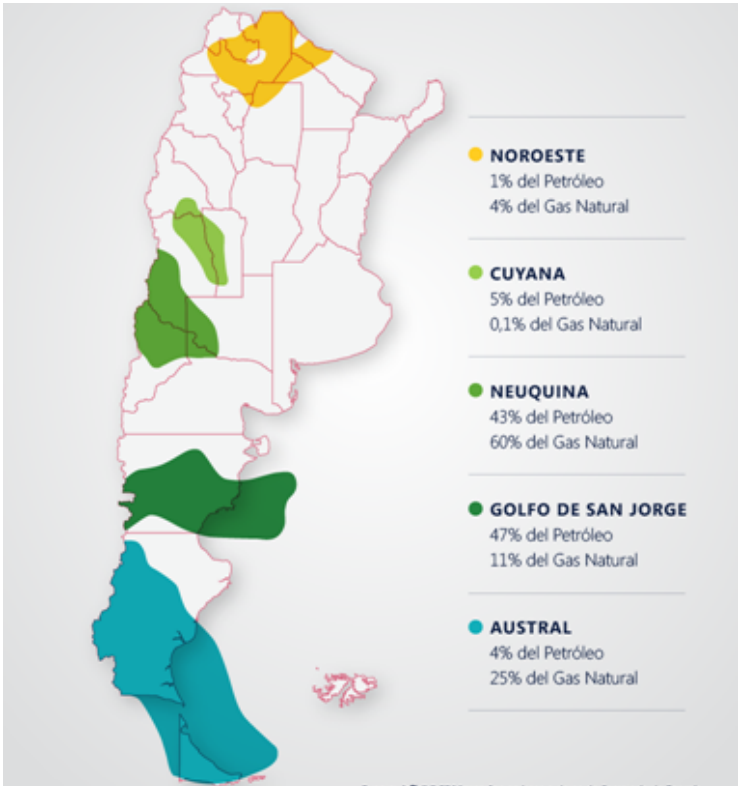
Clúster 5: Contiene pozos con una gran producción de gas, quedando en segundo lugar como el clúster de mayor producción de gas, y con bajo contenido de agua. En cuanto al tipo de pozo, en mayor proporción se encuentran pozos gasíferos pero además hay pozos de petróleo agrupados en este clúster. Según el tipo de recurso y los campos geográficos se trata de reservorios convencionales de las Cuencas Neuquina y del Golfo San Jorge, mientras que en mayor medida los hidrocarburos se concentran en la provincia de Río Negro.

Clúster 6: notamos que se trata de pozos convencionales de petróleo en la Cuenca del Golfo San Jorge o Austral, más precisamente en la provincia de Santa Cruz que presentan bajos niveles de producción. Este hecho se corresponde con lo que sucede en la actualidad dado que es de público conocimiento el Proyecto Andes impulsado por YPF de ceder las concesiones de dichos campos a otras empresas de menor escala por dejar de ser rentables para la compañía. Estos valores hacen que se sitúe junto con los clusters 1 y 2 como los menos rentables de todos.

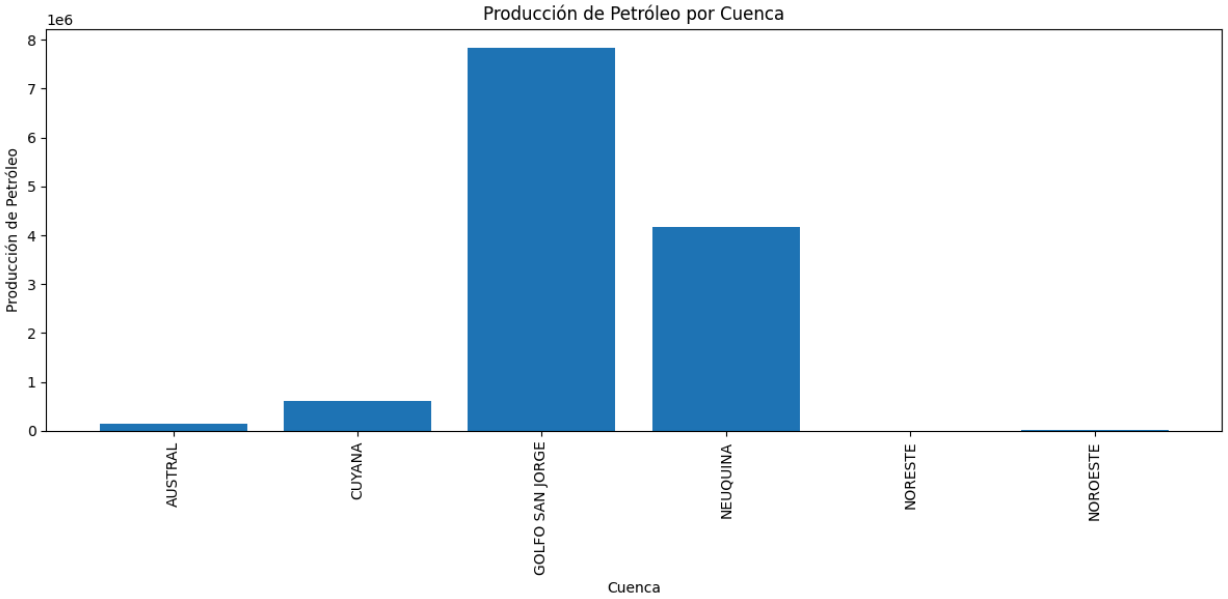
Clúster 7: Se destaca por ser el clúster con el segundo nivel más alto de producción de petróleo respecto de los anteriores y sumado a esto con elevados volúmenes de producción de agua. Corresponde a pozos petrolíferos del tipo convencional y en términos geográficos a la Cuenca del Golfo San Jorge, provincia de Chubut. Si bien los valores de producción de agua son medios con tendencia a altos, la rentabilidad de este grupo de pozos está justificada por los valores de producción de petróleo que son lejos de tan bajos como los anteriores (exceptuando el cluster 0), por lo que este alto rango de agua se podría solventar con el empleo de de esta misma en otras áreas de la producción, como la reinyección a la formación.

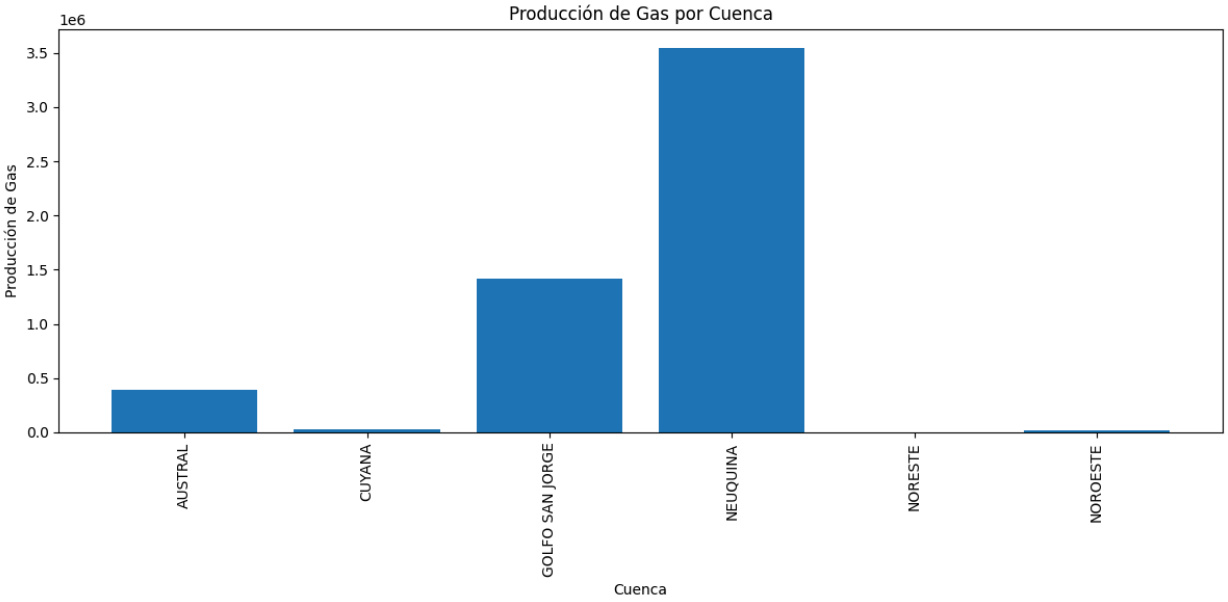
Estas observaciones pueden verse reflejadas y justificadas en los gráficos que se presentan a continuación:

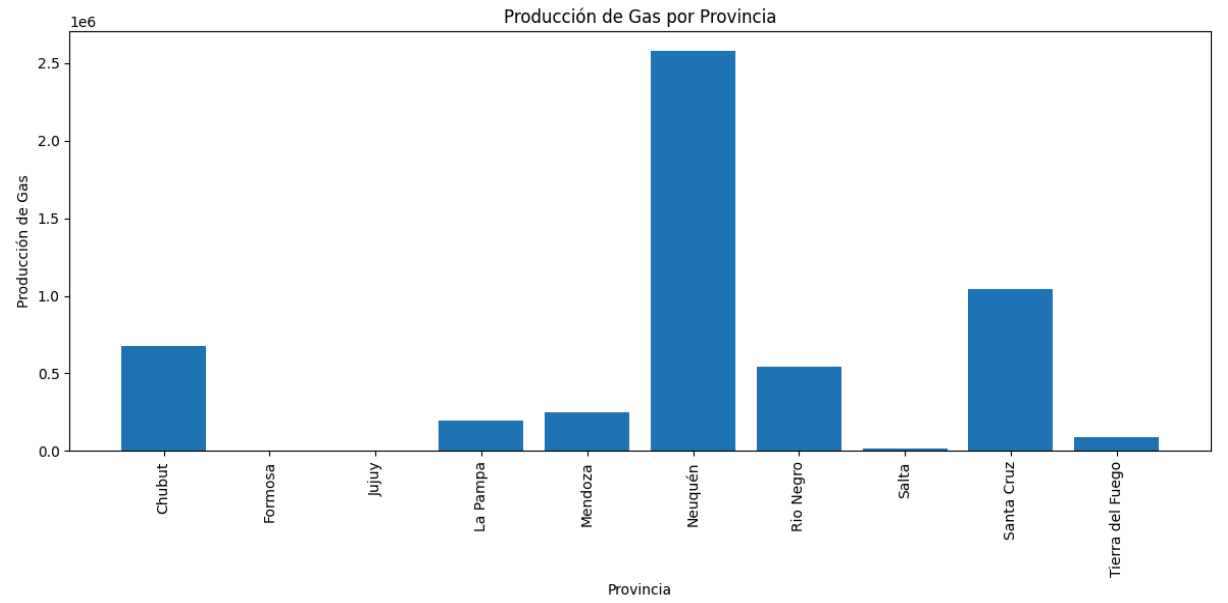
Cuencas de Hidrocarburos en el Territorio Argentino:

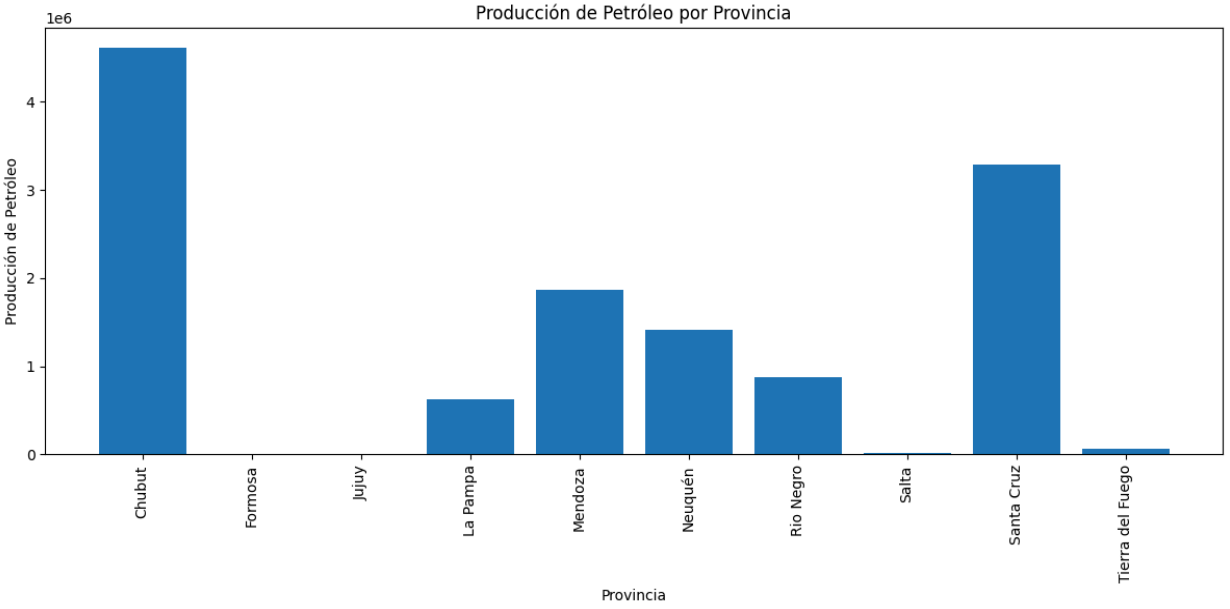


A continuación se muestran los valores de producciones acumuladas en el año 2023 por cuenca y por provincia empleando los datos que provienen del Dataset.


Producción de Petróleo por Cuenca
1e6
Producción de Petróleo
0
1
2
3
4
5
6
7
8
AUSTRAL
CUYANA
GOLFO SAN JORGE
NEUQUINA
NORESTE
NOROESTE
Cuenca


Producción de Gas por Cuenca
1e6
Producción de Gas
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
AUSTRAL
CUYANA
GOLFO SAN JORGE
NEUQUINA
NORESTE
NOROESTE
Cuenca





Como podemos ver en los gráficos de barras queda en evidencia que la mayor parte del volumen de producción de petróleo se concentra en las Cuencas del Golfo San Jorge y la Neuquina. Sumado a esto, vemos que ampliamente Chubut es la provincia que tiene los mayores niveles de producción en términos absolutos pero tal como vimos en el gráfico radial y su análisis, podemos concluir que en términos relativos son de mayor productividad los pozos de las provincias de la Cuenca Neuquina tales como Río Negro y Neuquén.

Respecto de la producción de gas ampliamente la más productiva es la Cuenca Neuquina, lo cual se ve reflejado en la actualidad al ver los proyectos y planes que tienen las compañías de la industria con el objetivo de exportar gas a distintos países en el mundo a fin de aprovechar la gran capacidad que tiene el país de generar este recurso.

Como posibles mejoras para implementar el modelo de datos, sería la de no analizar la totalidad de las cuencas, sino más bien centrarse en una sola, o en una sola provincia, y trabajar con los datos de yacimientos, formaciones, y quizás incluir las áreas de concesión. Esto debido a que la mayoria de los clusters se situan en las cuencas de Golfo San Jorge y Neuquina, habiendo una predominancia de provincias como Rio negro, y Neuquén, ademas de que este nuevo modelo daria como resultado una ubicacion geografica de clusters mas especifica, por lo que la cuestion de donde invertir capitales para la extraccion de hidrocarburos, no se limitaria a una cuenca u otra, sino a algo mas particular y exacto, como en que region de la cuenca (si al norte, sur, etc), en cual concesion o zonas cercanas, hasta incluso en que formacion productiva se deberia perforar. En el caso de plantear el enfoque del modelo como se describió, sería ideal también incluir series históricas, ya que al enfocarse en una cuenca sola, podría generar que el volumen de datos se reduzca considerablemente, haciendo al modelo más inexacto, además de que tendría un potencial predictivo de producciones futuras en los pozos existentes.