**UNIVERSIDAD NACIONAL DE INGENIERÍA**

**FACULTAD DE INGENIERÍA DE PETRÓLEO,**

**GAS NATURAL Y PETROQUÍMICA**

****

**TESIS**

**“ALGORITMOS DE APRENDIZAJE AUTOMÁTICO PARA MEJORAR EL CÁLCULO DEL CAUDAL INICIAL DE PRODUCCIÓN DE UN POZO EN UN CAMPO MADURO EN ETAPA DE DESARROLLO”**

**PARA OBTENER EL TÍTULO PROFESIONAL DE:**

**INGENIERO DE PETRÓLEO Y GAS NATURAL**

**ELABORADO POR:**

**PATRICK KARUSSO, TANTA PUJADA**

**ASESOR:**

**MSc. VICTOR HUERTA QUIÑONES**

**LIMA – PERÚ**

**2022**

# RESUMEN

El cálculo del flujo inicial de un pozo en un campo maduro disminuye la incertidumbre cuando se perfora la cantidad de pozos aumento, debido a la caracterización del comportamiento de producción de la historia. No obstante, esta regla se cumple en la mayoría

casos, la heterogeneidad del campo juega un papel importante todavía.

Este estudio se desarrolla en un campo de la Cuenca de Talara, donde el coeficiente determinante (r2) del caudal real inicial y la calculada con la ecuación de flujo es de aproximadamente 0.14, debido a la alta heterogeneidad de campo

La razón principal por la que consideramos la ecuación de flujo como variable de estudio es la fuerte correlación entre reservas y flujo inicial real de un pozo. Luego, estimar un flujo inicial más preciso para los pozos genera como consecuencia obtener reservas con menor sesgo, incentivando a la empresa a tomar decisiones de inversión.

La ingeniería nos dice que, para obtener un flujo inicial preciso, necesitamos recopilar datos para estimar mejores parámetros involucrados en la ecuación de flujo, pero esto requiere una inversión adicional que con frecuencia no es recuperar. Considerando el vasto historial de producción y pozos perforados de un campo maduro en la Cuenca Talara, el uso de algoritmos estadísticos para encontrar patrones es adecuado para esta tarea. Además, la aplicación de esta técnica es de bajo costo ya que utilizamos los datos de entrada disponibles y los resultados ayudarán a aumentar confianza para el pronóstico de la campaña de perforación.

El objetivo de este estudio es utilizar algoritmos de aprendizaje automático utilizando los mismos datos de entrada que utilizan la metodología tradicional para disminuir la incertidumbre en el cálculo del caudal inicial.

Utilizamos varios algoritmos para modelar datos, estableciendo una canalización con el algoritmo de mejor

rendimiento, automatizando este cálculo y poniéndose a disposición de todos los que requerían el

estimación de esta técnica.

Finalmente, utilizando las estimaciones de la ecuación de flujo y los algoritmos de aprendizaje automático, realizamos un

evaluación económica de cada método para comparar con los resultados económicos reales.

# ABSTRACT

Initial flow calculation of a well in a mature field decrease uncertainty when quantity of wells drilled increase, due to the characterization of history production behavior. Despite, this rule fulfill in most cases, the field’s heterogeneity play an important role yet.

This study takes place in a field of Talara Basin, where the determinant coefficient (r2) of real flow initial and the calculated with flow equation is approximately 0.14, because of the high heterogeneity of field.

The main reason we consider flow equation as study variable is the strong correlation between final reserves and real initial flow of a well. Then, estimating a more accurate initial flow for wells generate as consequence obtain reserves with less bias, encouraging company to take investment decision

The engineering tells us to get an accurate initial flow, we need to collect data in order to estimate better parameters involved in flow equation, but this require extra investment that frequently is not retrieve. Considering the vast history production and wells drilled of a mature field in Talara Basin, the use of statistical algorithms to find patterns is adequate for this task. Furthermore, the application of this technique is low cost as we use input data available and the results will help to increase confidence for drilling campaign prognosis.

The goal of this study is to use machine learning algorithms using the same input data that use the traditional methodology in order to decrease the uncertainty in initial flow calculation.

We used several algorithms for modelling data, establishing a pipeline with the algorithm of better performance, automating this calculation and becoming available for everybody that required the estimation of this technique.

Finally, using the estimations of flow equation and machine learning algorithms, we perform an economic evaluation for every method to compare with the real economic results.

# PRÓLOGO

En el presente trabajo, Tesis de grado titulado *Algoritmos de Aprendizaje Automático para mejorar el cálculo del caudal inicial de producción de un pozo en un campo maduro en etapa de desarrollo,* estudio y evalúo la factibilidad de usar información de producción y reservorio histórica para aumentar la confiabilidad en la estimación del caudal inicial de producción de un pozo en la cuenca de Talara, donde justamente abunda este tipo de información debido a sus largas campañas de perforación. En los primeros capítulos se describe, limpia y acondiciona los datos que serán usados para el modelado con algoritmos de predicción, este proceso sigue la estructura de una metodología que se sigue para la aplicación de estos algoritmos.

Posteriormente, comparé los caudales estimados usando el método tradicional y los estimados usando algoritmos de aprendizaje automático con los resultados reales para obtener márgenes de error que puedan ser comparados a través de gráficos sencillos.

Finalmente, realice una evaluación económica con los caudales obtenidos de los diferentes métodos usando las curvas de declinación como una constante de cada pozo para ambos métodos.

Esta tesis tiene como fin demostrar que es una opción viable usar datos históricos y algoritmos de aprendizaje automático para validar los resultados obtenidos a través de la ingeniería. Ya que su aplicación no tiene mayores costes y puede ser totalmente automatizado.

**INDICE**

[RESUMEN i](#_Toc120004481)

[ABSTRACT iii](#_Toc120004482)

[PRÓLOGO v](#_Toc120004483)

[INDICE v](#_Toc120004483)i

[LISTA DE FIGURAS](#_Toc120004483) IX

[LISTA DE TABLAS X](#_Toc120004483)II

[CAPITULO I: INTRODUCCIÓN 1](#_Toc120004484)

* 1. [Antecedentes 1](#_Toc120004485)
  2. [Problemática 2](#_Toc120004486)
  3. [Formulación del problema 3](#_Toc120004487)
     1. [Problema General 3](#_Toc120004488)
     2. [Problemas específicos 3](#_Toc120004489)
  4. [Objetivos………………………. 4](#_Toc120004490)
     1. [Objetivo General 4](#_Toc120004491)
     2. [Objetivos Específicos 4](#_Toc120004492)
  5. [Justificación de la Investigación 4](#_Toc120004493)
  6. [Hipótesis y Variables 5](#_Toc120004494)
     1. [Hipótesis General 5](#_Toc120004495)
     2. [Hipótesis específicas 5](#_Toc120004496)
  7. [Identificación de Variables 6](#_Toc120004497)
     1. [Variable independiente: 6](#_Toc120004498)
     2. [Variable dependiente: 6](#_Toc120004499)
  8. [Operacionalización de variables 6](#_Toc120004500)
  9. [Matriz de consistencia 8](#_Toc120004501)

[CAPÍTULO II: MARCO TEÓRICO Y CONCEPTUAL 11](#_Toc120004502)

[2.1. Marco Teórico 11](#_Toc120004503)

[2.2. Marco conceptual 11](#_Toc120004504)

[2.2.1 Ecuación de la difusividad 11](#_Toc120004505)

[2.2.2 Curvas de declinación 13](#_Toc120004506)

[2.2.3 Prueba de Minifrac 16](#_Toc120004507)

[2.2.4 PVT…………………..…………………………………………………………………………………………………………………….…………18](#_Toc120004508)

[2.2.5 Petrofísica……………………………………………………………………………………………………………………………….…………26](#_Toc120004509)

[2.2.6 Estadística……………………………………………………………………………………………………………..………………………….28](#_Toc120004510)

[2.2.7 Aprendizaje automatizado 35](#_Toc120004511)

[Preparación de datos 36](#_Toc120004512)

[Modelamiento/ Aprendizaje: 37](#_Toc120004513)

[Evaluación: 37](#_Toc120004514)

[2.2.8 Aprendizaje automático supervisado 38](#_Toc120004515)

[2.2.9 Aprendizaje Profundo. 44](#_Toc120004516)

[2.2.10 Equilibrio entre el sesgo (bias) y la varianza 47](#_Toc120004517)

[CAPÍTULO III: DESARROLLO DEL TRABAJO DE INVESTIGACIÓN 51](#_Toc120004518)

[3.1. Metodología. 5](#_Toc120004519)1

[3.1.1. Tipo de investigación 5](#_Toc120004520)1

[3.1.2. Población y muestra 5](#_Toc120004521)1

[3.2. Instrumentos de recolección de datos 5](#_Toc120004522)1

[3.3. Análisis e interpretación de la información 5](#_Toc120004523)2

[3.4. Descripción de la data 5](#_Toc120004524)3

[3.5. Obtención de la data: 5](#_Toc120004525)5

[3.5.1 Estimación de Permeabilidad a través de Ley K-Phi 5](#_Toc120004526)5

[3.5.2 La Prueba de Minifrac 5](#_Toc120004527)6

[3.5.3 Estimación de parámetros PVT (Sw, Bo, Uo, Ct, API) 5](#_Toc120004528)9

[3.5.4 Estimación de Permeabilidad a través de Ley de Timur 5](#_Toc120004529)9

[3.5.5 Índice de Hidrocarburos](#_Toc120004530) 60

[3.5.6 Curvas de declinación para el cálculo de las reservas de petróleo. 60](#_Toc120004531)

[3.5.7 Cálculo de la gradiente de presión 6](#_Toc120004532)2

[3.6. Preparación de datos 6](#_Toc120004533)3

[3.6.1 Eliminando valores nulos 6](#_Toc120004534)7

[3.7. Representación de datos 6](#_Toc120004535)8

[3.7.1 Exploración de datos (EDA) 6](#_Toc120004536)8

[3.7.2 Eliminando valores atípicos 7](#_Toc120004537)3

[3.7.3 Transformación de Distribuciones 7](#_Toc120004538)7

[3.7.4 Escalamiento de datos 7](#_Toc120004539)7

[3.7.5 Codificar Atributos No numéricos 7](#_Toc120004540)9

[3.8. Modelamiento. 80](#_Toc120004541)

[3.9. Análisis de resultados. 8](#_Toc120004541)4

[CAPÍTULO IV: ANÁLISIS Y DISCUSIÓN DE RESULTADOS 8](#_Toc120004542)6

[CAPÍTULO V: ANÁLISIS ECONÓMICO](#_Toc120004543) 90

[CAPÍTULO VI: CONCLUSIONES Y RECOMENDACIONES 9](#_Toc120004545)3

[CAPÍTULO VII: REFERENCIAS BIBLIOGRÁFICAS 9](#_Toc120004546)5

[ANEXOS 95](#_Toc120004547)

**LISTA DE FIGURAS**

[Figura 1. Clasificación de curvas de declinación (Arps J.J. “Estimación de preservas de recuperación primaria”) 15](#_Toc120257742)

[Figura 2. Análisis de a derivada de la función G – Mecanismos Leakoff 17](#_Toc120257743)

[Figura 3. Variación de la viscosidad en función de la presión 19](#_Toc120257744)

[Figura 4. Variación del factor volumétrico en función de la presión 22](#_Toc120257745)

[Figura 5. Formas que adopta la compresibilidad según la variación del volumen con la presión 25](#_Toc120257746)

[Figura 6. Ilustración de la posición los parámetros involucrados en la ecuación de flujo 28](#_Toc120257747)

[Figura 7. Distribución de valores típicos de porosidad 28](#_Toc120257748)

[Figura 8. Ubicación aproximada de la media en una distribución 29](#_Toc120257749)

[Figura 9. Ubicación exacta de la mediana en una distribución 29](#_Toc120257750)

[Figura 10. Ubicación aproximada de la moda en una distribución 30](#_Toc120257751)

[Figura 11. Ubicación de los límites que determinan el rango 30](#_Toc120257752)

[Figura 12. Ubicación aproximada de los percentiles 25 y 75 30](#_Toc120257753)

[Figura 13. Módulo de la varianza de una distribución 31](#_Toc120257754)

[Figura 14. Gráfica típica de la porosidad versus la permeabilidad 32](#_Toc120257755)

[Figura 15. Valores típicos del coeficiente de distribución 33](#_Toc120257756)

[Figura 16. Gráfico de histograma 34](#_Toc120257757)

[Figura 17. Gráfico de boxplot con los valores de la porosidad 34](#_Toc120257758)

[Figura 18. Gráfico de Quantile-Quantile con los valores de la porosidad 35](#_Toc120257759)

[Figura 19. Flujo de trabajo típico en la aplicación de algoritmos de aprendizaje automático 35](#_Toc120257760)

[Figura 20. Ilustra cómo funciona los algoritmos basados en Bagging 42](#_Toc120257761)

[Figura 21. Ilustra la forma en cómo trabajan los algoritmos de Stacknet 44](#_Toc120257762)

[Figura 22. Ilustra cómo funciona una red neuronal en su forma más básica 45](#_Toc120257763)

[Figura 23. Modelos ajustados pertenecientes a un mismo conjunto de datos con bajo sesgo 47](#_Toc120257764)

[Figura 24. Modelos ajustados pertenecientes a un mismo conjunto de datos con baja varianza 48](#_Toc120257765)

[Figura 25. Ilustra el equilibrio entre el sesgo y varianza 49](#_Toc120257766)

[Figura 26. Grafica K vs Phi de una de las Formación X. 54](#_Toc120257767)

[Figura 27. Visualización de curvas Time – Caudal – Presión 56](#_Toc120257768)

[Figura 28. La derivada de la función G 56](#_Toc120257769)

[Figura 29. Análisis de Minifrac después del cierre (ACA) 57](#_Toc120257770)

[Figura 30. Índice de Hidrocarburos (pies) vs Caudal de producción real (bls) 59](#_Toc120257771)

[Figura 31. Gráfico realizado para estimar los parámetros b y D a través de la comparación con pozos vecinos, en un pozo candidato 60](#_Toc120257772)

[Figura 32. Grafico que ilustra cómo se estima la gradiente actual de un reservorio 61](#_Toc120257773)

[Figura 33. Captura de la dimension de la tabla con la data de entrada 65](#_Toc120257774)

[Figura 34. Tipos de datos y contador de nulos en la data de entrada 66](#_Toc120257775)

[Figura 35. Estadística descriptiva de los valores numéricos de la data de entrada. 67](#_Toc120257776)

[Figura 36. Gráfico de la kurtosis para cada atributo numérico 72](#_Toc120257777)

[Figura 37. Gráfico de la asimetría para cada atributo numérico 73](#_Toc120257778)

[Figura 38. Gráfico de porosidad antes y después de eliminar los valores atípicos 73](#_Toc120257779)

[Figura 39. Gráfico de la viscosidad antes y después de eliminar los valores atípicos 74](#_Toc120257780)

[Figura 40. Gráfico del factor volumétrico antes y después de eliminar los valores atípicos 74](#_Toc120257781)

[Figura 41. Asimetría de cada variable de entrada de nuestro modelo 75](#_Toc120257782)

[Figura 42. Comparación de una misma variable cruda y normalizada 76](#_Toc120257783)

[Figura 43. Histograma de los atributos numéricos luego de ser escalados 77](#_Toc120257784)

[Figura 44. Distribución del caudal en la data de entrenamiento y de prueba 79](#_Toc120257785)

[Figura 45. Rendimiento de los modelos con sus parámetros por defecto. 80](#_Toc120257786)

[Figura 46. Ilustra la forma en cómo funciona la validación cruzada 81](#_Toc120257787)

[Figura 47. Rendimiento de los modelos con sus parámetros y usando la validación cruzada 82](#_Toc120257788)

[Figura 48. Rendimiento de los modelos con sus parámetros optimizados y usando validación cruzada 85](#_Toc120257789)

[Figura 49. Ilustra el método que más se aproxima a la línea recta negra que representa al caudal real. 86](#_Toc120257790)

[Figura 50. Captura de la tabla con los datos necesarios para el pronóstico de producción 87](#_Toc120257791)

[Figura 51. Comparación del VAN de cada pozo. 88](#_Toc120257792)

[Figura 52. Comparación del TIR de cada pozo. 88](#_Toc120257793)

[Figura 53. Porcentaje general de los datos usados en entrenamiento y prueba 89](#_Toc120257794)

**LISTA DE TABLAS**

[**Tabla 1.** Tipos de declinación según Arps 11](#_Toc106183113)

[**Tabla 2.** Valores de coeficientes de la correlación de Vásquez y Beggs por rangos. 19](#_Toc106183114)

[**Tabla 3.** Definiciones y formas de las funciones de activación. 41](#_Toc106183115)

[**Tabla 4.** Operacionalización de variables 47](#_Toc106183116)

[**Tabla 5.** Matriz de consistencia 49](#_Toc106183117)

[**Tabla 6.** Breve descripción de los atributos que se usaran para el modelo 66](#_Toc106183118)

[**Tabla 7.** Descripción de los atributos principales en base al analisis univariado 73](#_Toc106183119)

[**Tabla 8.** Muestra los algoritmos corridos con sus principales características. 89](#_Toc106183120)

[**Tabla 9.** Los coeficientes que más se acercan a la unidad pertenecen al obtenido con el algoritmo. 94](#_Toc106183121)

# CAPITULO I: INTRODUCCIÓN

## Antecedentes

2020: OTC-30854-MS: Metodología de pronóstico y acondicionamiento de datos mediante aprendizaje automático en datos de producción para una plataforma de pozo.

Este estudio usa algoritmos de aprendizaje automático de ensamble para condicionar y pronosticar data de producción, entre los modelos que se usaron están el perceptrón multicapa (MLP), análisis de componentes principales (PCA) y soporte de regresión vectorial (SVR). La data de producción tenía 23 tipos de información, pero solo fueron elegidas 12 variables (fecha, hora, temperatura, presión, etc.), las cuales fueron usadas para explorar la complejidad de la no linealidad de las variables y estimar la producción con modelos de aprendizaje automático y profundo. El trabajo concluyo con un modelo que tenía un coeficiente de determinación de 98% usando 70% como data de entrenamiento.

2019: SPE-195022-MS: Aplicación de algoritmos de aprendizaje automático para la predicción multivariable de la producción de un pozo.

Este estudio usa modelos de aprendizaje automático para maximizar la producción usando la data disponible de 470 pozos horizontales, usando en total 15 variables de tipo numéricas (longitud de intervalo estimulado, caudal de producción) y categóricas (formación, tipo de propante, etc.). El problema requería algoritmos de regresión y los de mayor desempeño fueron los de ensamble (220/GBM vs. 20/RF). El trabajo mostro la aplicabilidad y el desempeño de estos modelos.

## Problemática

Actualmente en los campos de la cuenca Talara se sigue usando el método de las curvas de declinación de Arps para la estimación de reservas de petróleo, esto debido a que la confiabilidad del análisis mejora con la calidad y cantidad de los datos de producción. Y justamente esa condición suelen cumplir los campos maduros, lográndose un buen ajuste al seleccionar bien los parámetros b y Di. Pero esta buena selección de parámetros no es suficiente, ya que tener una buena aproximación a las reservas reales de un pozo solo es posible luego del primer o segundo mes de producción del pozo, ya que las reservas calculadas con las curvas de declinación se ven más influenciadas por el caudal inicial de producción. El método tradicional para la estimación del caudal inicial de producción de un pozo antes de perforarlo, es a través de la ecuación de flujo modificada de acuerdo a las características del reservorio y campo, el resultado es calibrado variando la permeabilidad, la gradiente actual de presión o alguna otra variable que se haya podido observar o inferir una vez el pozo ya entra en producción, esto con el objetivo de que se ajuste a la producción real máxima de los primeros meses. Entonces, para que la ubicación de un pozo sea seleccionada y entre en plan de ejecución, el caudal inicial debe ser estimado previamente usando la ecuación de flujo, para luego estimar las reservas con el análisis de las curvas de declinación. Sin embargo, eso nos lleva a resultados erróneos porque el caudal inicial estimado casi siempre requiere de ajuste. La incertidumbre crece aún más cuando en los campos maduros de Talara se suelen producir los pozos en comingle, ósea de varias formaciones al mismo tiempo, ya que ocurre una mezcla de comportamientos de cada reservorio objetivo. La posibilidad de mitigar este problema y evaluar mejor el potencial de cada formación sería posible probando cada intervalo, pero en muchos casos el costo-beneficio de evaluar cada formación no se satisface. De este modo, para aproximar un valor estimado, lo que se hace es relacionarlo directamente con el flujo k\*h para obtener la segregada de producción para cada formación.

La problemática radica cuando se quiere maximizar la producción del campo, y tenemos un cálculo de caudal inicial sesgado, esto disminuye la probabilidad de seleccionar las mejores ubicaciones para perforar. Ya que nos conduce a una estimación errónea de las reservas de un pozo, llevándonos a tener rendimientos negativos en un contexto de campañas de perforación en campos maduros, desestimando de esta manera la intención de inversión en la recuperación primaria y disminuyendo el ingreso de divisas al país por medio de las regalías. Se debe tener en cuenta que la mayor inversión para la obtención de más datos que alimenten a nuestros modelos matemáticos pueden convertir a un pozo en no rentable.

## Formulación del problema

### Problema General

¿Cómo es posible tener una aproximación más acertada del caudal inicial de producción de un pozo en un campo maduro en etapa de desarrollo usando algoritmos de aprendizaje automático?

### Problemas específicos

* ¿Qué información y que herramientas necesito para la implementación de los algoritmos de aprendizaje automático?
* ¿Qué tanto por ciento de error actual disminuiré al aplicar este método?
* ¿Qué tanto por ciento varía la viabilidad económica de un proyecto de perforación al usar algoritmos de aprendizaje automático para el cálculo del caudal inicial de producción de un pozo en un campo maduro?

## Objetivos

### Objetivo General

Estimar un caudal de producción inicial de un pozo en un campo maduro en etapa de desarrollo, de menos error que la ecuación de flujo, con respecto al valor real, a través de algoritmos de aprendizaje automático.

### Objetivos Específicos

* Compilar y limpiar la data disponible de los parámetros del pozo antes y después de ser perforado, usando Python.
* Entrenar, evaluar y comparar el desempeño de los modelos de aprendizaje automático (lineales y no lineales) a través de métricas de regresión en Python.
* Evaluar el comportamiento de la métrica del modelo cuando variamos la proporción de la data de entrenamiento y de prueba, a través de las curvas de aprendizaje.
* Evaluar el sentido físico a través de la proporcionalidad de las variables de entrada para el modelo.
* Comparar la variación de error porcentual del resultado de la ecuación de flujo y del modelo de aprendizaje automático con lo real.
* Comparar la variación porcentual de los indicadores de la viabilidad real de un proyecto de perforación con el método convencional y el que usa algoritmos de aprendizaje automático, para estimar el caudal inicial de producción de un pozo un campo maduro.

## Justificación de la Investigación

El presente trabajo es pertinente dentro de un contexto de proyectos de perforación en campos maduros cuando la diferencia que existe entre el cálculo del caudal inicial de producción de un pozo con la ecuación de flujo antes de su perforación y lo que observamos después de los primeros meses de producción ,no es despreciable y llega a ser relevante cuando influye en gran medida al cálculo de reservas por el análisis de las curvas de declinación provocando un error de estimación en las variables que miden la viabilidad del proyecto. A través de la aplicación de los algoritmos de aprendizaje automático se busca estimar un caudal inicial de producción de menos error con respecto a lo real observado usando la misma data de entrada que para el cálculo con la ecuación de flujo. El flujo de trabajo para la aplicación de este método consiste en la limpieza, transformación y escalamiento de los datos, para luego continuar con el modelado, optimización de los parámetros del modelo y evaluación de métricas para maximizar la eficiencia de predicción. Para esto, solamente usaremos la data disponible y Python, el cual es un software libre. Esta aplicación puede ser mejorada considerando modelos de aprendizaje automático más complejos o mejorando la cantidad o calidad de los datos de entrada, y también busca incentivar el uso de estos algoritmos para solucionar problemas y mejorar la toma de decisiones aprovechando la data disponible recolectada en la historia del campo como entrada para los modelos.

## Hipótesis y Variables

### Hipótesis General

La aplicación de algoritmos de aprendizaje automático disminuirá el error del método convencional en el cálculo del caudal inicial de producción de un pozo en un campo maduro en etapa de desarrollo.

### Hipótesis específicas

* El uso de algoritmos de aprendizaje automático toma ventaja de la data recolectada y obtiene una mejor estimación, que la ecuación de flujo, del caudal inicial de producción de un pozo en un campo maduro en etapa de desarrollo.
* El error calculado al estimar el caudal inicial de producción de un pozo de un campo maduro en etapa de desarrollo usando algoritmos de aprendizaje automático es menor al estimado con la ecuación de flujo.
* El porcentaje de error de los indicadores de viabilidad real de un proyecto de perforación con la estimada antes de su ejecución disminuye cuando usamos algoritmos de aprendizaje automático para estimar el caudal inicial de producción del pozo en un campo maduro en etapa de desarrollo.

## Identificación de Variables

### Variable independiente:

El método de cálculo para el caudal inicial de producción de un pozo en un campo maduro en etapa de desarrollo. El performance será medido a través de métricas que evalúan el error de los métodos de predicción comparado con los valores reales.

* MAE: Error absoluto medio
* MSE: Error cuadrático medio
* R2: Coeficiente de determinación

### Variable dependiente:

Según la certidumbre del método que se use para pronosticar la producción de petróleo, la viabilidad de un proyecto de perforación variará y podemos medirlo usando dos indicadores:

* VAN (Valor actual neto)
* TIR (Tasa interna de retorno)

## Operacionalización de variables

1. Operacionalización de variables

|  |  |  |  |
| --- | --- | --- | --- |
| Variable | Tipo de Variable | Definición | Indicador |
| Variable Independiente: Algoritmos de aprendizaje automático. | Cuantitativa | Son un conjunto de algoritmos matemáticos útiles para resolver problemas de clasificación y regresión debido a su habilidad para reconocer patrones no mejor que los humanos, pero si más escalables. | El performance de los modelos construidos se miden a través de las métricas de tipo regresión, usado para predecir variables continuas. |
| Variable  Dependiente:  Indicadores de la viabilidad. | Cuantitativa | Son herramientas que miden la viabilidad económica de un proyecto. | El VAN es el valor actual nominal y mide la rentabilidad, mientras que el TIR es la tasa interna de retorno y mide el tiempo que pasara hasta que se recupere la inversión inicial. |

## Matriz de consistencia

8

1. Matriz de consistencia

|  |  |  |
| --- | --- | --- |
| Formulación del Problema | Objetivos | Hipótesis |
| ¿ Es posible tener una aproximación más acertada del caudal inicial de producción de un pozo en un campo maduro en etapa de desarrollo usando algoritmos de aprendizaje automático? | Estimar un caudal de producción inicial de un pozo en un campo maduro en etapa de desarrollo, de menos error que la ecuación de flujo, con respecto al valor real, a través de algoritmos de aprendizaje automático. | La aplicación de algoritmos de aprendizaje automático disminuirá el error del método convencional en el cálculo del caudal inicial de producción de un pozo en un campo maduro en etapa de desarrollo. |
| Problemas específicos | Objetivos específicos | Hipótesis específica |
| ¿Qué información y que herramientas necesito para la implementación de los algoritmos de aprendizaje automático?  ¿Qué tanto por ciento de error actual disminuiré al aplicar este método?  ¿Qué tanto por ciento varía la viabilidad económica de un proyecto de perforación al usar algoritmos de aprendizaje automático para el cálculo del caudal inicial de producción de un pozo en un campo maduro? | Compilar y limpiar la data disponible de los parámetros del pozo antes y después de ser perforado, usando Python.  Entrenar, evaluar y comparar el desempeño de los modelos de aprendizaje automático (lineales y no lineales) a través de métricas de regresión en Python.  Evaluar el comportamiento de la métrica del modelo cuando variamos la proporción de la data de entrenamiento y de prueba, a través de las curvas de aprendizaje.  Evaluar el sentido físico a través de la proporcionalidad de las variables de entrada para el modelo.  Comparar la variación de error porcentual del resultado de la ecuación de flujo y del modelo de aprendizaje automático con lo real.  Comparar la variación porcentual de los indicadores de la viabilidad real de un proyecto de perforación con el método convencional y el que usa algoritmos de aprendizaje automático para estimar el caudal inicial de producción de un pozo en proyectos de desarrollo de un campo maduro. | El uso de algoritmos de aprendizaje automático toma ventaja de la data recolectada y obtiene una mejor estimación, que la ecuación de flujo, del caudal inicial de producción de un pozo en un campo maduro en etapa de desarrollo.  El error calculado al estimar el caudal inicial de producción de un pozo de un campo maduro en etapa de desarrollo usando algoritmos de aprendizaje automático es menor al estimado con la ecuación de flujo.  El porcentaje de error de los indicadores de viabilidad real de un proyecto de perforación con la estimada antes de su ejecución disminuye cuando usamos algoritmos de aprendizaje automático para estimar el caudal inicial de producción del pozo en un campo maduro en etapa de desarrollo. |

# CAPÍTULO II: MARCO TEÓRICO Y CONCEPTUAL

## Marco Teórico

Existen varios métodos para estimar el caudal inicial de producción de un pozo nuevo que produce de multi-reservorios, entre los que tienen mayor fundamento y soporte teórico están el balance de materiales y la simulación numérica, estos métodos suelen tener un error asociado dependiendo de las variables que se asumirán para satisfacer la data de entrada para estos modelos. La aplicación de las mismas toma un tiempo no despreciable que no te garantiza un buen ajuste y mejor predicción.

El aprendizaje automatizado es un conjunto de algoritmos matemáticos útiles para resolver problemas de clasificación y regresión debido a la habilidad para reconocer patrones no mejor que los humanos, pero si más escalables, esto quiere decir cuando tenemos gran cantidad de data relacionada con el comportamiento de alguna variable. Existen 3 tipos de estos algoritmos, el aprendizaje supervisado donde el algoritmo aprende del ejemplo, el aprendizaje no supervisado donde se busca obtener grupos con comportamientos diferentes y el aprendizaje reforzado que usa la retroalimentación para mejorar la respuesta de un modelo.

El análisis de curvas de Arps han demostrado ser el método más fácil y eficiente para estimar reservas en campos maduros cuando se tiene el caudal inicial del pozo y el comportamiento de declinación de los pozos productores vecinos.

## Marco conceptual

### Ecuación de la difusividad

Es la ecuación que describe el comportamiento del petróleo en el reservorio, resulta de la combinación de 3 ecuaciones, la ecuación de la continuidad producto de la ley de la conservación de la masa, la ecuación de flujo empírica desarrollada por Darcy y la ecuación de estado que rige el estado de los fluidos.

Donde:

k: permeabilidad, milidarcys

r: posición radial, pies

p: presión, psia

ct: compresibilidad total,

t: tiempo, horas

: porosidad, fracción

: viscodidad, centipoise

Los campos del noroeste peruano llegan a presentar más de un fluido, por lo tanto, deben ser considerados en la siguiente formula:

Donde co, cw y cg representan la compresibilidad del petróleo, agua y gas respectivamente, y So, Sw y Sg representan la fracción de la saturación de estos fluidos.

Cuando un pozo está fluyendo con un caudal constante bajo un estado transiente, el comportamiento de la presión del pozo será como si el reservorio fuera infinito en tamaño. La diferencia de presión actual en este periodo está representada por la suma de la variación de presión ideal del reservorio con la producida por el skin.

Adicionalmente tomando en cuenta el principio de superposición que es aplicable debido a que los pozos de los Campos de Talara suelen estar rodeados de pozos que producen en estado transiente. El concepto de superposición nos dice que la caída total de presión en cualquier punto del reservorio es la suma de los cambios de presión en ese punto causado por el flujo de cada uno de los pozos del reservorio.

La ecuación que resulta de tomar en cuenta los anteriores conceptos es la siguiente:

)

Donde:

k = permeabilidad, md

t = tiempo, horas

rw = radio de pozo, ft

s = factor de daño

La geología del noroeste peruano está compuesta predominantemente por arenas apretadas con una permeabilidad mínima del orden de 0.001 milidarcy, por lo tanto, se considera que aún no llega a las fronteras y está justo en el periodo descrito anteriormente. Entonces para el cálculo de nuestro caudal inicial usaremos esta ecuación para aproximar su valor.

### Curvas de declinación

En la vida productiva de un pozo se pueden identificar patrones de comportamiento en la data de producción con respecto al tiempo y también a la producción acumulada. Estos patrones siguen comportamientos que se pueden agrupar en curvas tipo, las cuales han sido establecidas producto del estudio de la relación entre las variables antes mencionadas en diferentes pozos. Para que estas curvas tengan consistencia, se debe asumir que los factores que la afectaron al inicio de su declinación, están presentes en toda la vida productiva.

Las curvas de declinación están descritas principalmente por 3 factores:

* Caudal inicial de producción
* Curvatura de declinación
* Razón de declinación

Algunas de las condiciones que incrementan la confiabilidad en su uso según Ikoku (1984):

La producción debe ser estable en el periodo que ha sido analizado, para el caso del noroeste peruano esta condición si se cumple ya que la mayoría de pozos producen a través del bombeo mecánico con las mismas condiciones invariables (presión fluyente constante) y controlando el nivel de fluido. Salvo excepciones cuando una causa externa que origine cambios en las condiciones de producción (por ejemplo, mantenimiento del sistema de bombeo mecánico), la data de producción luego de estos eventos puede significar un cambio en la curva que al ser extrapolada junto con la de antes del evento, puede cuantificar la magnitud de ganancia por haber realizado la inversión en algún servicio de pozo o workover.

Este método usado para la estimación de reservas puede llegar a tener mejor desempeño que un cálculo volumétrico o balance de materiales cuando tenemos un campo con gran historia de producción con patrón de comportamiento definido.

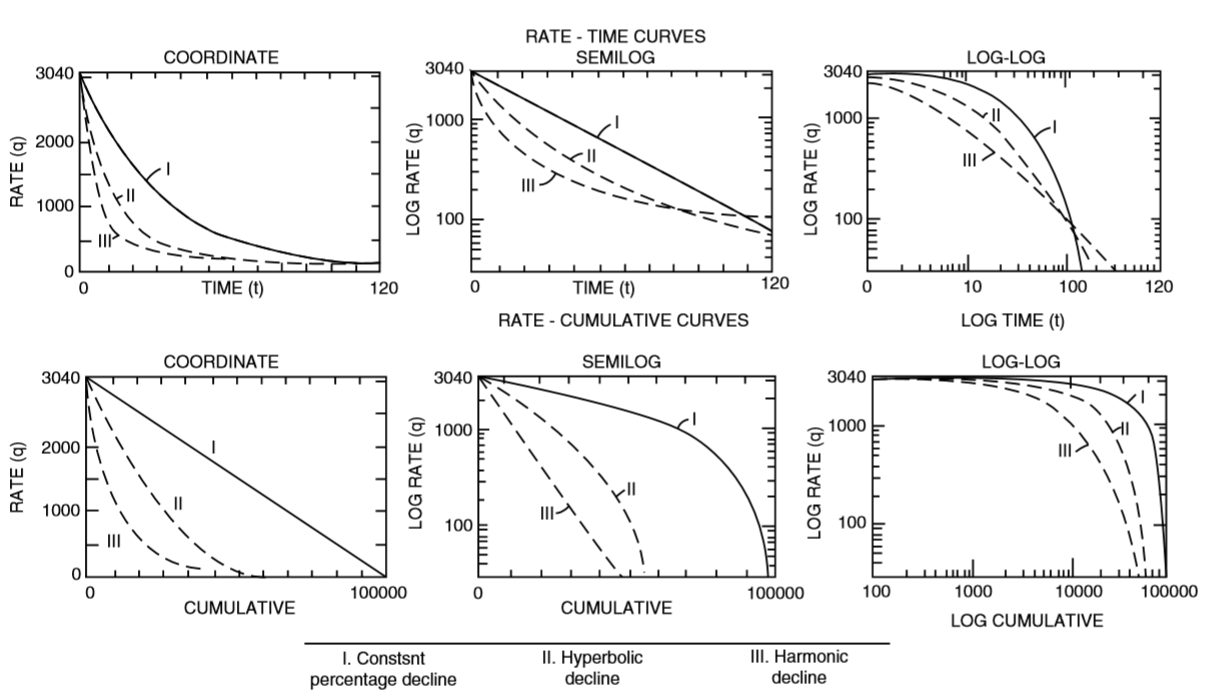
Arps estudio la relación entre el caudal de producción y el tiempo de un pozo, obteniendo lo siguiente:

Donde b y D son constantes determinados empíricamente. Integrando la ecuación antes expuesta obtenemos:

Arps (1945) propuso 3 tipos de comportamientos de declinación, diferenciados matemáticamente por los parámetros de curvatura (b) y razón de declinación (D):

* Declinación Exponencial
* Declinación Armónica
* Declinación Hiperbólica

1. Clasificación de curvas de declinación (Arps J.J. “Estimación de preservas de recuperación primaria”)



Fuente: Analysis of Decline Curves by J. J. Arps (1994)

A continuación, diferenciaremos los tipos de comportamiento según el valor de b, cabe mencionar que el valor de la constante de declinación (D) es calculada mediante ajuste de la curva con la producción real.

1. Tipos de declinación según Arps

|  |  |  |
| --- | --- | --- |
| Declinación exponencial | Declinación hiperbólica | Declinación harmonica |
| b = 1 | 0 < b < 1 | b = 1 |
|  |  |  |

*Nota.* Esta tabla muestra las 3 formas que puede adoptar la declinación de un pozo, según las curvas de declinación de Arps(1994).

### Prueba de Minifrac

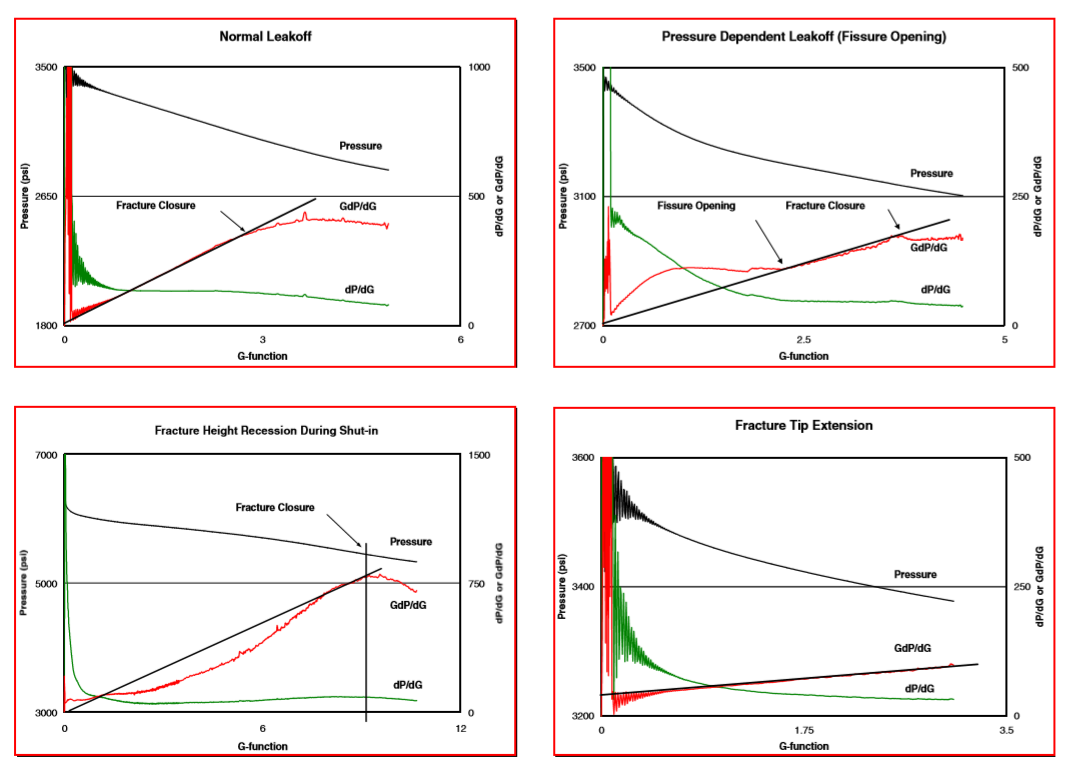
Es uno de los métodos con mejor costo beneficio para estimar la capacidad de flujo del reservorio (kh), para reservorios de baja permeabilidad el beneficio es mayor debido a que estos necesitan de fracturamiento hidráulico para poder producir. Esta prueba nos da información para diseñar una fractura y caracterizar el reservorio.

La prueba consiste en la inyección de un fluido, que puede ser agua, a una presión mayor a la presión de fractura de la formación, hasta llegar a fracturar la formación. Luego la inyección se detiene y se observa el comportamiento de la presión. Finalmente podemos evaluar los parámetros físicos medidos de la prueba para determinar el mecanismo de leakoff e identificar el cierre de la fractura.

#### Función G

Barre y Mukherjee presentaron el análisis de la derivada de la función G para identificar el mecanismo de leakoff siguiendo una prueba de inyección para el diagnóstico de fracturas. Para el análisis de la derivada de la función G se requiere un gráfico de la presión de fondo, la derivada de la presión (dP/dG) y la superposición derivada (GdP/dG) versus Función G. El tipo de leakoff es identificado por su forma característica de la derivada y la superposición de curvas derivadas. A continuación, se muestra la gráfica de la derivada de la función G para los cuatro casos de leakoff observados en formaciones de baja permeabilidad.

1. Análisis de a derivada de la función G – Mecanismos Leakoff



Fuente: David P. Craig and Michael J. Eberhard, Robert D. Barree. (2000) Adapting High Permeability Leakoff Analysis to Low Permeability Sands for Estimating Reservoir Engineering Parameters

#### Normal leakoff

Comportamiento que ocurre cuando el área de la fractura es constante durante el cierre y leakoff, a través de una roca matriz homogénea. Con el análisis de la derivada de la función G, podemos identificar una línea recta que representa una derivada constante y la superposición de la derivada de la función G en la línea recta a través del origen. El cierre de la fractura es identificado cuando la superposición de la derivada usando la data se desvía por la parte inferior de la línea recta.

#### Pressure-dependent leakoff

Se caracteriza por indicar la existencia de fracturas secundarias intersectando la fractura principal y puede ser identificado por una joroba en la derivada de la función G sobre una línea recta extrapolada a través de la data de un normal leakoff. La presión de abertura de la fisura es identificada el final de la joroba cuando la superposición de la derivada de la data llega a la línea recta extrapolada.

#### Fracture-height recession

Se identifica cuando la superposición de la derivada de la función G cae por debajo de la línea recta extrapolada de un leakoff normal, siguiendo una tendencia cóncava. Lo cual es una señal de que el líquido se escapa con menor velocidad que en una fractura normal sugiriendo la posible presencia de una presión de soporte.

#### Fracture-tip extensión

Se puede identificar cuando la fractura crece incluso después de que se detiene la inyección y se cierra el pozo. Este tipo de comportamientos suele ocurrir en reservorios con baja permeabilidad. Debido a la fuga, se liberaría la energía que se extiende a los extremos de la fractura, extendiéndola la punta de la misma.

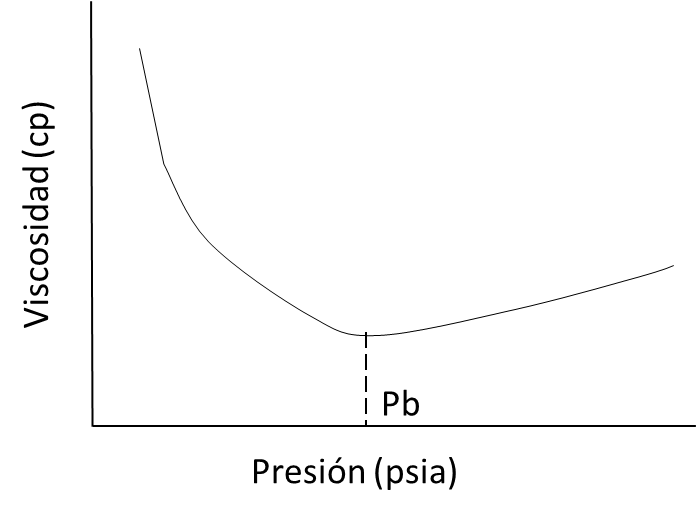
### PVT

En todo modelo que intente reproducir la estimación respecto al performance de un reservorio, necesita caracterizar los fluidos por medio de los valores físicos medidos en pruebas de presión, volumen y temperatura. Estas mediciones son aproximadas a lo más exacto posible debido a su influencia en el cálculo del caudal de producción. En los campos de talara las pruebas PVT realizadas son muy antiguas y el costo-beneficio de tomar alguna en la actualidad vuelve a influir en la decisión final. Afortunadamente, existen correlaciones que suelen ser muy precisas al ser producto de estudios estadísticos y ensayos físicos que con un margen porcentual de error y unos intervalos de confianza se usaran para estimar los valores que usaremos para nuestro modelo.

#### Viscosidad de petróleo

Es la resistencia interna que ofrece el petróleo a deformarse en orden de fluir en una dirección. Mide el flujo del petróleo a través de la roca porosa y se suele expresar en centipoise. Esta propiedad es directamente proporcional con la presión e inversamente proporcional con la temperatura del reservorio.

1. Variación de la viscosidad en función de la presión



Fuente: Propia

##### Para petróleo saturado

Cuando el reservorio tiene la presión inicial menor o igual que la presión de burbuja (Pi <= Pb), esto quiere decir que el reservorio presenta dos fases (petróleo y gas).

###### Correlación Chew-Connaly:

Donde:

*Rango de confianza:*

* Presión: 132 - 5645 psia
* Temperatura: 72 - 292 °F
* Solubilidad del Gas: 51 - 3544 Pcs/Bf
* Viscosidad del petróleo muerto: 0.377 - 50 cp

###### Correlación Beggs-Robinson:

Donde:

*Rango de confianza:*

* Presión: 132 - 5265 psia
* Temperatura: 72-295 °F
* Solubilidad del Gas: 20 - 2070 Pcs/Bf
* °API = 16 - 58

##### Para petróleo bajo saturado

Cuando el reservorio tiene la presión inicial mayor que la presión de burbuja (Pi > Pb), esto quiere decir que solo está presente una fase (petróleo) y el gas aun esta disuelto en el.

###### Correlación Beal:

)

Donde:

###### Correlación Khan:

###### Correlación Vásquez-Beggs:

*Rango de confianza:*

* Presión psi: 141-9.515
* Rs Pc/Bbl: 9.3-2.199
* Viscosidad: 0.117-148
* °API = 15.3-59.5

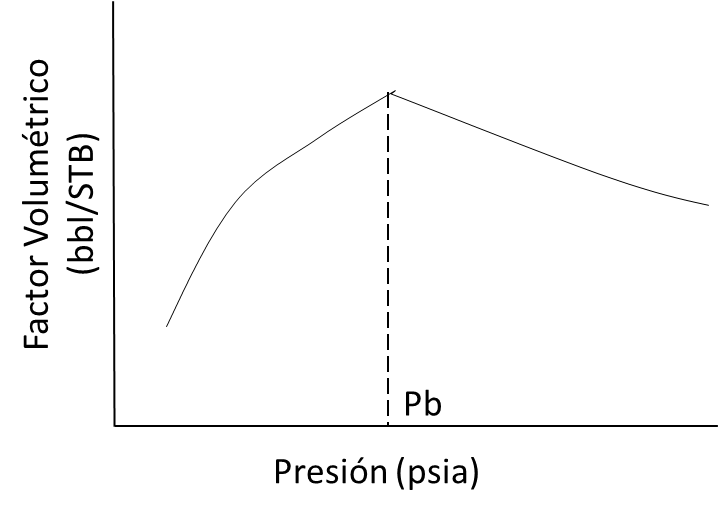
#### Factor Volumétrico de petróleo

Es la relación entre el volumen de petróleo incluyendo el gas en solución en condiciones de reservorio y el volumen de petróleo a condiciones estándar.

Donde:

En el contexto de un petróleo bajo saturado (Pi > Pb), el factor volumétrico se comportaría de la siguiente manera:

1. Variación del factor volumétrico en función de la presión



Fuente: Propia

Para su cálculo usaremos correlaciones empíricas:

##### Para petróleo saturados

###### Correlación de Standing (1981):

Basado en la evaluación de 105 datos experimentales en 22 diferentes sistemas de hidrocarburos.

Donde:

###### Correlación de Glasso:

Donde:

###### Correlación de Vásquez y Beggs:

Basado en la evaluación de 6000 medidas de Bo a diferentes presiones.

Donde:

1. Valores de coeficientes de la correlación de Vásquez y Beggs por rangos.

|  |  |  |
| --- | --- | --- |
| Coeficiente | API <= 30 | API > 30 |
|  | 4.677 × 10−4 | 4.670 × 10−4 |
|  | 1.751 × 10−5 | 1.100 × 10−5 |
|  | −1.811 × 10−8 | 1.337 × 10−9 |

*Nota.* Esta tabla muestra los valores que asumen los coeficientes de la correlación, cuando el API es mayor o menor que 30.

##### Para petróleo bajo saturados

Cuando el reservorio tiene la presión inicial mayor que la presión de burbuja (Pi > Pb), en este caso el factor volumétrico disminuye gracias a la compresibilidad

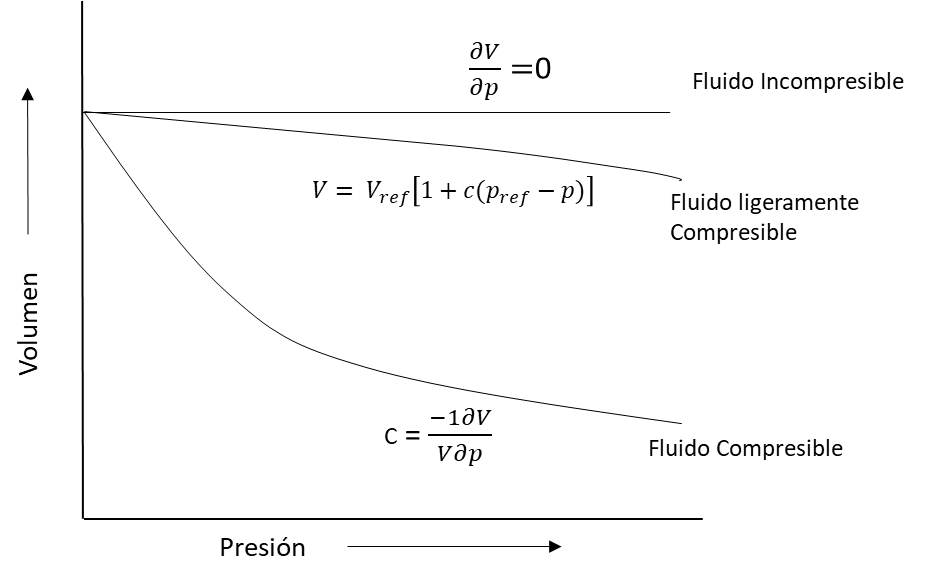
###### Correlación de Vásquez y Beggs

Donde

#### Compresibilidad del petróleo

Es la propiedad responsable del cambio de volumen del petróleo cuando tiene gas en solución debido a la variación de la presión y temperatura. Esto considerando que la compresibilidad de únicamente el petróleo liquido es despreciable. Se mide e cambio de unidad volumétrica por cambio de presión.

1. Formas que adopta la compresibilidad según la variación del volumen con la presión



Fuente: Tarek Ahmed. (2006) Reservoir Engineering. Tercera edición

La compresibilidad del petróleo expresada según el factor volumétrico del petróleo es:

Para su cálculo usaremos correlaciones:

##### Para petróleo saturados

Donde:

##### Para petróleo bajo saturados

###### Correlación de Vásquez y Beggs

### Petrofísica

La naturaleza de las rocas del reservorio determina principalmente la cantidad de fluido que contiene la roca en el espacio entre sus partículas sólidas y la habilidad de estos fluidos para fluir a través de las rocas. Llamamos porosidad de la roca a la medición del espacio vacío y permeabilidad a la habilidad de la roca de transmitir fluidos a través de sí misma. Tener conocimiento de estas dos propiedades es fundamental antes de pasar al tipo de fluido, cálculo de flujo y estimación del fluido recuperable.

#### Porosidad

La porosidad de la roca es la fracción del volumen del espacio entre las partículas sólidas de la roca y el volumen total de la roca. Este espacio está conformado por todos los poros, vugs, espacios inter e intra cristalinos, etc. Su valor expresado en fracción va en el rango de 0 a 1, mientras que expresado en fracción oscila entre 0 y 100%. Para los cálculos posteriores usaremos siempre la fracción.

Donde:

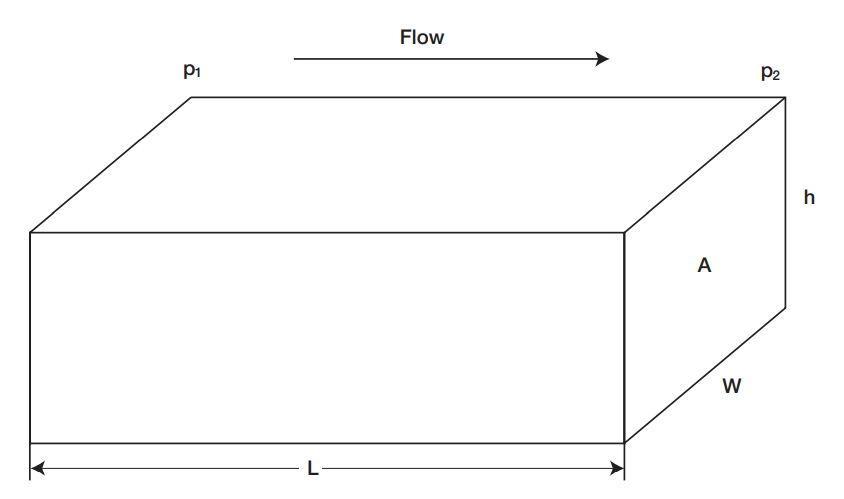
Hay varias formas de obtener la porosidad, usando registros de densidad, neutrónico, sónicos. Cada registro tiene su alcance y su grado de sensibilidad a ciertos contextos, tales como la naturaleza del líquido, la litología. Así como también, el uso de dos o más registros en conjunto puede ser determinantes para detectar petróleo o gas. En campos maduros con una amplia historia de producción ya existen modelos geológicos confiables que estiman estos valores por yacimiento, por reservorio y por zona.

#### Permeabilidad

Es una propiedad de la roca que mide la capacidad y habilidad de la formación para transmitir fluidos a través del medio poroso del mismo. Esta propiedad es la que más influye cuando se estima el caudal inicial de producción, ya que controla la dirección y la cantidad de fluidos del reservorio que se moverán. Darcy desarrollo una ecuación que mide el flujo de los fluidos.

Donde:

1. Ilustración de la posición los parámetros involucrados en la ecuación de flujo



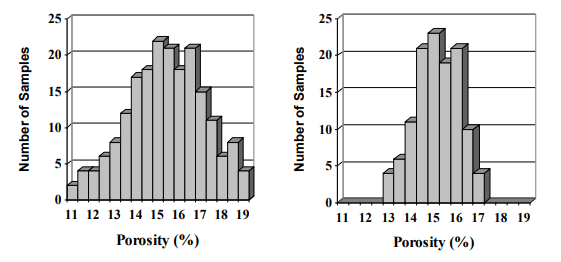
Fuente: Tarek Ahmed. (2006) Reservoir Engineering. Tercera edición Estadística.

#### Análisis univariado.

En este análisis se evalúan los parámetros estadísticos de un solo atributo, estas medidas estadísticas son:

* *Valor típico o central:* indicador que representa a la mayoría de datos.
* *Dispersión del valor central:* indicador de concentración de los datos con respecto a las medidas de tendencia central.
* *Simetría:* indicador de forma de la distribución de datos.

1. Distribución de valores típicos de porosidad



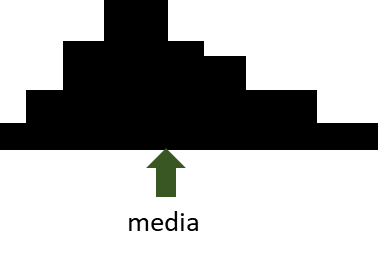
Fuente: Deepak Devegowda (2018) *Data Mining Course. The University of Oklahoma*

##### Medidas de centralización

Son valores centrales de la población de una muestra.

Media: Es la suma de valores dividido por el tamaño muestral, ósea es el promedio aritmético de todos los valores de una variable. Esta medida es sensible a valores atípicos.

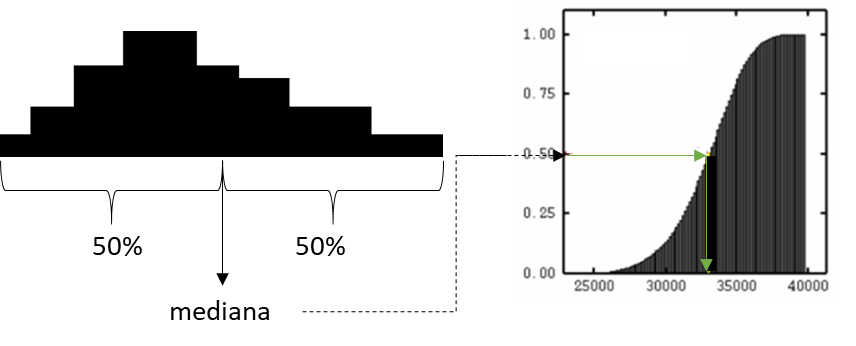
1. Ubicación aproximada de la media en una distribución



Fuente: Propia

##### Mediana: Valor que ocupa la posición central, ósea coincide con la división de las observaciones en dos grupos con el mismo número de observaciones para cada lado. Si los datos estuvieran ordenados en orden ascendente, la mitad de los valores están por debajo y la otra mitad por encima de la mediana.

1. Ubicación exacta de la mediana en una distribución



Fuente: Propia

Moda: Valor que ocurre con más frecuencia, donde esta alcanza su máximo valor. Se puede tener más de un valor para la moda.

1. Ubicación aproximada de la moda en una distribución



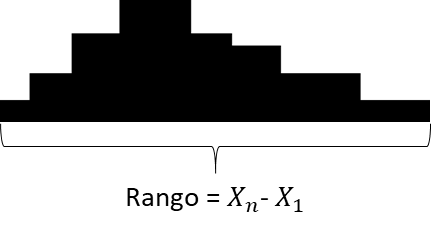
Fuente: Propia

##### Medidas de dispersión

Son valores que nos indican lo dispersos o agrupados que están los datos entre si y respecto al valor de la media.

Rango: Diferencia o recorrido entre dos observaciones, inferior y superior. Mide la dispersión total de entre todos los elementos. Nos sirve como medida general.

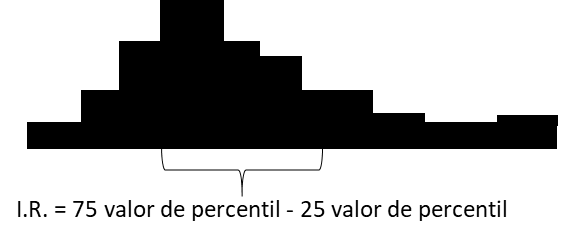
1. Ubicación de los límites que determinan el rango



Fuente: Propia

Rango intercuartil: Distancia entre el primer y tercer cuartil, mide la dispersión de la mediana. Se parece al rango, pero la diferencia es que esta solo toma el 50% de las observaciones. 25% por encima y 25% por debajo de la mediana, esto permite que sea menos sensible a los outliers.

1. Ubicación aproximada de los percentiles 25 y 75

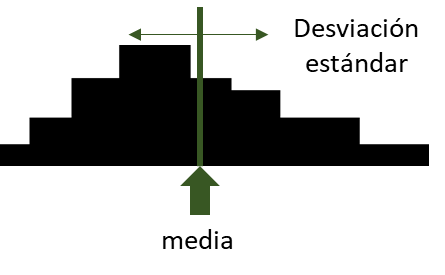


Fuente: Propia

Desviación estándar: Nos indica que tan separados están nuestros datos. Es el resultado del promedio de la diferencia que vamos a encontrar entre los valores absolutos de cada uno de los elementos con la media aritmética (m). Equivale a la raíz cuadrada de la varianza, esta expresada en las mismas unidades que la data.

Varianza: Resultado del promedio de las diferencias con respecto a la media (m) al cuadrado. Se usa el denominador (n-1) para observaciones menores a 30, caso contrario se usa n.

1. Módulo de la varianza de una distribución

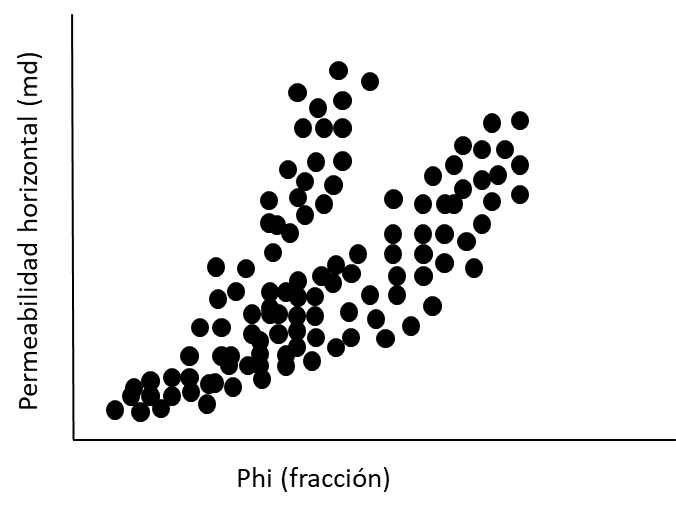


Fuente: Propia

#### Análisis bivariado.

En este análisis se analiza la relación entre variables, nos permite entender la causa del comportamiento de una variable. Usando este análisis podemos entender la relación que existe entre las propiedades del reservorio con algún parámetro geológico y en consecuencia predecir su comportamiento.

1. Gráfica típica de la porosidad versus la permeabilidad



Fuente: Propia

##### Covarianza

Medida estadística que mide la correlación entre todas las observaciones de 2 variables, su magnitud depende de la magnitud de las dos variables. Este valor nos ayuda a comprender el tipo de correlación lineal (directa, inversa o nula). Es sensible a los outliers.

Medida aritmética de los productos de las desviaciones de cada una de las variables respecto a sus medidas respectivas.

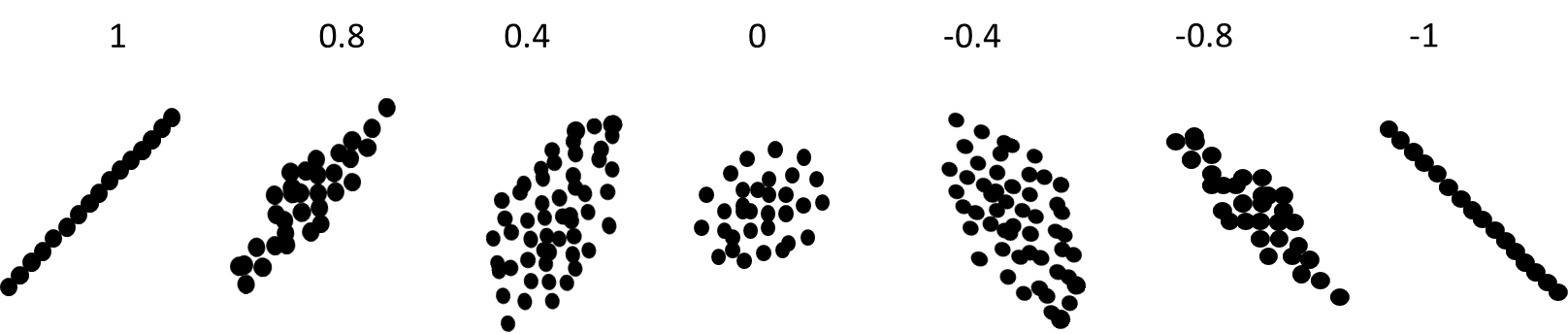
: media del atributo X

: media del atributo Y

##### Coeficiente de correlación

Nos ayuda a describir cómo es la relación existente entre dos variables. Una variable independiente y otra dependiente. Este coeficiente es un valor cuantitativo y es el resultado de la relación de dos o más variables. Su valor está en el rango de -1 a +1, toma el valor de +1 si la proporcionalidad es directa o positiva, el valor de -1 si la proporcionalidad es inversa o negativa y finalmente el valor de 0 cuando no existe relación alguna.

1. Valores típicos del coeficiente de distribución



Fuente: Propia

El coeficiente de correlación no depende de la pendiente. Se define como el cociente entre la covarianza de la distribución y el producto de las desviaciones típicas de cada una de las variables, ósea como una versión normalizada de la covarianza.

Sx: desviación estándar del atributo X

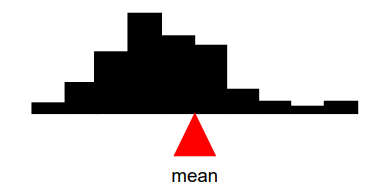
Sy: desviación estándar del atributo Y

#### Gráficos Estadísticos.

##### Histograma.

Agrupa data en intervalos y la frecuencia de cada intervalo es graficado con una barra centrada en el punto medio del intervalo, es el gráfico más usado en análisis univariado ya que nos permite identificarlas medidas de tendencia central y de dispersión, permitiéndonos conocer la naturaleza de nuestra data.

1. Gráfico de histograma

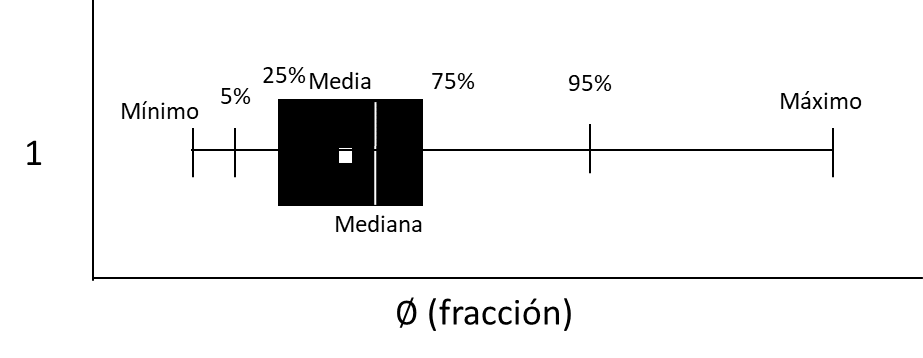


Fuente: Propia

##### Boxplot

Grafica que nos da una visión de donde está concentrada la data, así como también, nos muestra el origen de los valores extremos. Se construye usando el valor mínimo, el primer cuartil, la mediana, el tercer cuartil y el valor máximo. Esta gráfica también nos ayuda a ubicar visualmente donde se encuentran los valores extremos.

1. Gráfico de boxplot con los valores de la porosidad

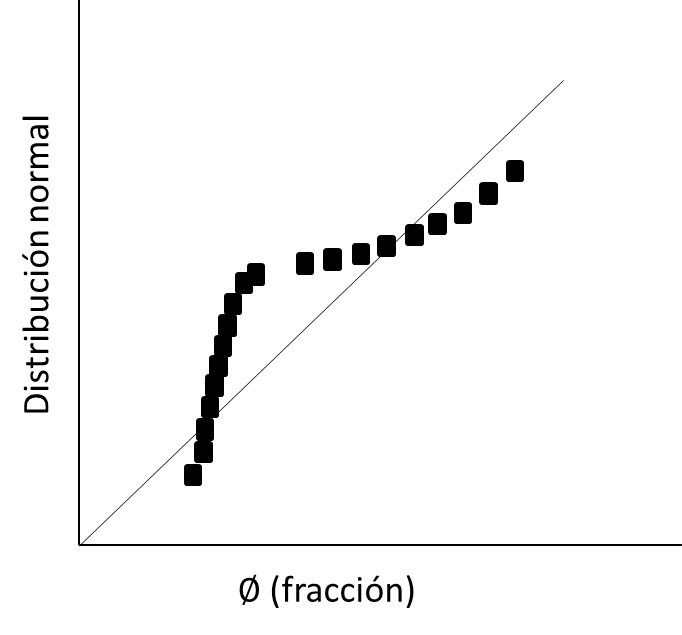


Fuente: Propia

##### Quantile-Quantile plot (Q-Q)

Grafica que nos permite verificar la normalidad de un conjunto de datos o para comprar distribuciones en busca de similitudes. Para este método no se requiere que las muestras sean del mismo tamaño.

1. Gráfico de *Quantile-Quantile con los valores de la porosidad*



Fuente: Propia

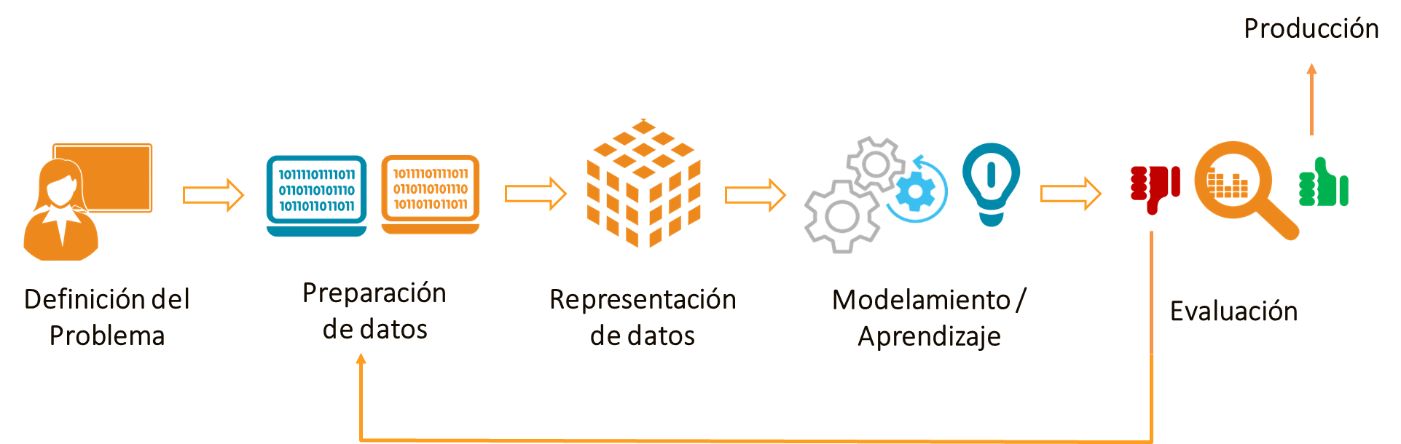
### Aprendizaje automatizado

El aprendizaje automatizado es un conjunto de algoritmos matemáticos útiles para resolver problemas de clasificación y regresión debido a la habilidad para reconocer patrones no mejor que los humanos, pero si más escalables, esto quiere decir cuando tenemos gran cantidad de data relacionada con el comportamiento de alguna variable.

#### Ciclo de Trabajo del Aprendizaje automatizado

Los algoritmos de aprendizaje automático son solo una etapa dentro de la cadena, todos los pasos son importantes para conseguir buenos resultados. El ciclo consiste en trabajar con 5 etapas e ir iterando conforme a los resultados de cada una.

1. Flujo de trabajo típico en la aplicación de algoritmos de aprendizaje automático



Fuente: Propia

#### Definición del Problema:

En esta etapa nos toca definir el problema que queremos resolver y cuáles son las condiciones que determinan el éxito de su solución. Primero establecemos que tipo de algoritmo de aprendizaje automático nos ayudara a resolver nuestro problema, debemos elegir entre aprendizaje automático supervisado, no supervisado y reforzado, luego según los recursos definimos la variable objetivo. Finalmente, se elige la métrica que usaremos para medir la calidad de cada posible solución

#### ***Preparación de datos***:

Aquí empezamos con la obtención de la data, según el conocimiento de la naturaleza del problema a resolver, elegimos la data que será usada para esta etapa. Se tratará de abarcar la mayor cantidad de variables que estén relacionadas con la variable objetivo definida anteriormente, ya que en una etapa posterior se discernirá cuáles de ellos serán usados para el modelo final. Normalmente la data se encuentra en varias tablas, por lo cual es necesario hacer una compilación de las mismas a través de la automatización. Finalmente, sabemos que no se suele trabajar directamente con una base de datos producto de la recopilación y mediciones, sino que siempre se le hace una limpieza donde modificamos y transformamos la data para crear otra base de datos más limpia que nos permita sacarle mayor provecho.

#### Representación de datos:

Con la data compilada y limpia realizamos la ingeniería de datos, esta etapa es considerada la que más influye en el rendimiento del algoritmo y la que te permite llegar al estado de arte, ya que elegimos las columnas o las características de cada fila, que serán usadas como input para el modelo. Esta etapa empieza con el análisis exploratorio, el cual es el primer acercamiento y un paso fundamental antes de usar algoritmos de aprendizaje automático, nos ayuda entender cómo vamos a resolver el problema haciendo uso de la programación. Para esto nos apoyamos de herramientas como Python para visualizar los datos con el objetivo de obtener conclusiones que direccionen las siguientes etapas del ciclo. Para la elección de las columnas usaremos distintos algoritmos que nos ofrece las librerías de Python (entre ellas sklearn).

#### ***Modelamiento/ Aprendizaje:***

En esta etapa haremos la selección del modelo que usaremos, el tipo de modelos que se probaran para llegar al indicado los describiremos en los siguientes apartados del capítulo. El modelo se ajustará a nuestros datos y arrojará un score de desempeño en la data de prueba separada con anterioridad.

#### ***Evaluación:***

En esta etapa usaremos el score del modelamiento para determinar si el modelo es bueno o no. Se usará la validación cruzada para aumentar la robustez del score y disminuir la probabilidad de sobre ajustar nuestro modelo, de esta manera contribuimos a que el modelo sea independiente de la muestra. También se hará uso de algoritmos de automatización para iterar en un rango de posibles valores para los hiperparámetros de los modelos que probaremos. Realizaremos esto hasta que lleguemos a un resultado que satisfaga nuestras necesidades de negocio, para finalmente llevarlo a producción.

Para esta etapa haremos énfasis en seguir el principio de Parsimonia, el cual nos dice que debemos optar por la solución más simple, empezando con lo más rápido y menos complejo, para finalmente ir agregando la complejidad.

Existen 3 tipos de estos algoritmos, el aprendizaje supervisado donde el algoritmo aprende del ejemplo, el aprendizaje no supervisado donde se busca obtener grupos con comportamientos diferentes y el aprendizaje reforzado que usa la retroalimentación para mejorar la respuesta de un modelo.

### Aprendizaje automático supervisado

Es uno de los más utilizados y requiere la intervención humana en el aprendizaje. Para su aplicación debemos obtener observaciones con el valor objetivo de ejemplo. Resuelven problemas de clasificación, cuando queremos predecir valores categóricos, y también de regresión, cuando queremos predecir algún valor numérico. A continuación, se mostrar los modelos que usaremos para lograr nuestro objetivo.

#### Modelos lineales.

##### Regresión Simple

Realiza predicciones haciendo una suma ponderada de las variables de entrada y una constante. Para calcular los pesos usamos una función de coste que mide el error de nuestro modelo, el objetivo es optimizarlo al mínimo valor, ya que de esta manera obtendremos el modelo con menor error de predicción. Se puede inferir que el peso de cada variable nos indicara el grado de influencia en la predicción de la variable de interés.

Los siguientes 3 modelos son lineales, a los cuales se les aplica restricciones o penalizaciones en los pesos de sus coeficientes para reducir el efecto de sobreajuste. Esto nos evita construir modelos complejos que memoricen la data de entrenamiento y así obtener un mejor desempeño del modelo en la data de prueba. Existen 3 maneras de aplicar estas restricciones.

##### Regresión Ridge

Para este tipo de regresión, agregamos el término de penalización “ “ a la función de coste. Esto provoca que el algoritmo no solo ajuste la data, sino que también mantenga los pesos del modelo lo más pequeños posible. El parámetro alfa controla la intensidad para regularizar el modelo. Si alfa es igual a 0, esta se convierte una regresión linear, si alfa es grande, entonces todos los pesos terminaran siendo cercanos al cero y el resultado será una línea recta atravesando el promedio de los datos. El valor óptimo de alfa se alcanza evaluando el error del modelo en la data de prueba para cada valor de alfa.

*Función de coste de la regresión de Ridge*

##### Represión Lasso

Básicamente, es otra regresión creada para regularizar la regresión linear simple, pero a diferencia de la antes visto, esta añade la norma L1 del vector de pesos en lugar del cuadrado medio de la norma L2 visto en la regresión Ridge. En esta regresión se elimina completamente el peso de las variables con menos importancia.

*Función de coste de la regresión de Ridge*

##### Regresión de Red Elástica

Es un término medio entre la regresión Ridge y Lasso. El termino de regularización es una mezcla de ambos términos usados respectivamente en cada regresión, con el detalle de que se puede controlar la relación de mezcla. Cuando r = 0; L red elástica es equivalente a la regresión Ridge, y cuando r = 1, es equivalente a la regresión Lasso.

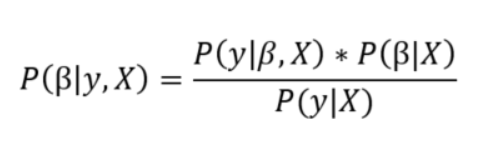
*Función de coste de la regresión de red Elástica*

##### Regresión de Soporte Vectorial

Usa los mismos principios que la máquina de soporte vectorial para clasificación, salvo unas pequeñas diferencias, las cuales radican en la amplia gama de valores de salida que puede adoptar al ser una variable continua. El hiperplano que genera el SVR cuenta con dos vectores paralelos de soporte, la diferencia de los valores que están fuera de la región limitada por los vectores es el error absoluto o épsilon (), mientras que el vector “w” representa los coeficientes que generan el hiperplano. Lo que se busca es encontrar un hiperplano que minimice su margen con un error mínimo que pueda predecir nuestra variable de interés.

##### Regresión Lineal Bayesiana

Es una regresión lineal que hace uso de la estadística bayesiana para modelar. Este modelo usa el conocimiento a priori o evidencia y la data para obtener una distribución posterior de los parámetros del modelo. Este tipo de modelos es efectivo cuando se tienen poca data.



#### Modelos No Lineales

##### Árboles de decisión.

Usa la división binaria recursiva para dividir el espacio en regiones que no se superponen, y que forman subconjuntos que en su versión final sirven para dar la media como un estimado para alguna observación de la data de prueba que caiga en uno de estos subconjuntos que fueron creadas inicialmente con la data de entrenamiento. El algoritmo usará la división binaria recursiva e iterará usando cortes de umbral para reducir lo máximo posible el error (RSS).

Siendo el R1 y R2 las regiones donde se separan las observaciones acordes al predictor j y el valor s. De esta manera se elige la mejor opción en el momento, pero no toma en cuenta si las futuras divisiones nos llevaran a un mejor desempeño del modelo.

Esto en algún momento nos conlleva a sobre ajustar la data de entrenamiento en muchos casos, pero se da la opción de limitar su propio crecimiento y podar el árbol final para convertirlo en uno menos profundo.

Ventajas:

* Fácil de interpretar a pesar de la relación compleja entre las variables de predicción.
* Los árboles manejan variables numéricas y categóricas, sin tener que crear dummies o usar one-hot-enconding. Esto depende de la implementación del algoritmo de la librería elegida para su uso.
* Al ser no paramétricos no necesita que las variables cumplan con una distribución.
* No se ven muy influenciados por outliers.
* Manejan bien los valores perdidos.
* Útiles para determinar la importancia de predicción de cada variable.
* Se pueden aplicar a problemas de clasificación y regresión.
* Pertenecen al conjunto de técnicas supervisadas no paramétricas que consiguen segmentar el espacio de los predictores en regiones más simples.

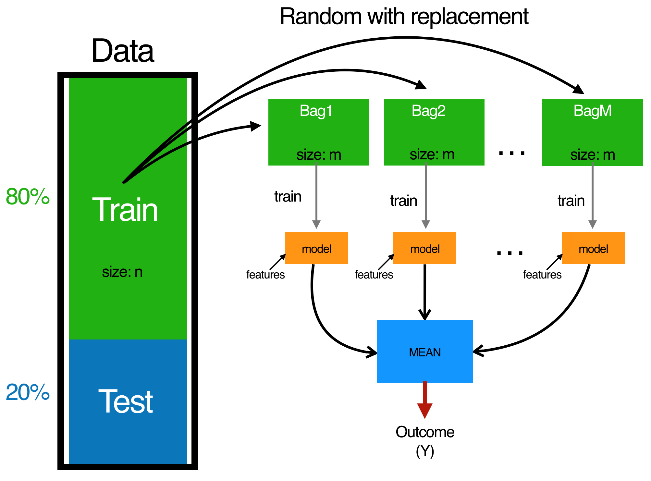
##### Modelos de ensamble.

Son modelos que combinan diferentes algoritmos de aprendizaje automático con diferentes arreglos y aplican un método para lograr un consenso. Cuando los comparamos con un solo modelo, estos modelos tienen una eficiencia y exactitud mejorada.

###### Bagging

Es el promedio de un conjunto de versiones del mismo modelo para obtener un resultado más generalizado mediante el uso del boostrapping, el cual consiste en crear subconjuntos de observaciones del set de datos que pueden repetirse, con esto logramos disminuir la varianza del modelo.

1. Ilustra cómo funciona los algoritmos basados en Bagging



Fuente: Propia

Bosques Aleatorios: modelo formado por un conjunto de árboles de decisión, en la cual cada uno de ellos se entrena con la misma data de entrenamiento, pero modificada por el método de bootstraping (muestreo aleatorio con reemplazo). El resultado del modelo se obtiene agregando las predicciones de todos los árboles individuales que forman al modelo.

###### Boosting

Es un método que consiste en el promedio ponderado de modelos, donde cada modelo es construido secuencialmente tomando en cuenta y corrigiendo el error del modelo anterior, con esto logramos reducir el bias de nuestro modelo.

###### Potenciación del Gradiente

Formados por un conjunto de árboles de decisión individuales, entrenados de forma secuencial, de forma que cada nuevo árbol trata de mejorar los errores de los árboles anteriores. La predicción de una nueva observación se obtiene agregando las predicciones de todos los árboles individuales que forman el modelo.

###### AdaBoost.

Método en el que todas las observaciones parten con la misma importancia o peso, estos pesos se van actualizando en función al error de clasificación de cada variable predictora generada por un un árbol de decisión de un nodo y dos hojas, también llamados stump, para determinar el stump de referencia que se usara para la actualización de pesos se usa un coeficiente que mida la efectividad o pureza del árbol, para nuestro caso el coeficiente Gini. Al ser un algoritmo boosting se entiende que cada stump toma los errores de los anteriores del mismo.

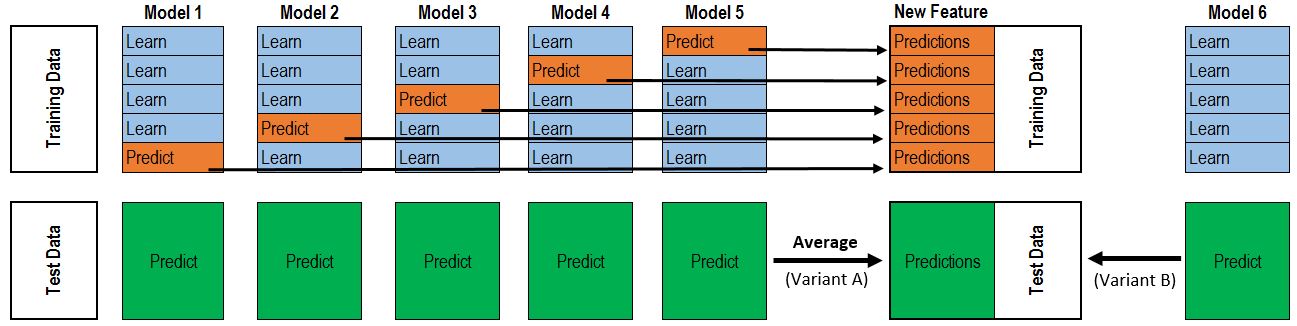
###### XGBoost

Es uno delos más usados actualmente por su poder predictivo cuando tenemos gran cantidad de datos heterogéneos. Consiste en construir árboles en distintos subespacios aleatorios, pero a diferencia de los bosques aleatorios, el xgboost sigue el principio de corregirse a sí misma en cada iteración. Además, en su implementación posee características como la computación paralela, regularización y poda de árboles.

###### Stacknet

Es un método de meta modelado que combina diferentes modelos. Consiste en hacer predicciones con diferentes modelos en la data de entrenamiento para luego agregarlos en un nuevo modelo. Este método no necesita la data de entrada, sino que necesita saber cómo han trabajado los modelos históricamente para encontrar el mejor camino para combinarlos. Permitiendo incrementar la fuerza predictiva de nuestro modelo.

1. Ilustra la forma en cómo trabajan los algoritmos de Stacknet



Fuente: How to Win a Data Science Competition: Learn from Top Kagglers Course,

National Research University Higher School of Economics.

Sub campo del aprendizaje automático, que hace uso de las redes neuronales con arquitecturas complejas para modelar data lineal y no lineal de alta complejidad.

#### Redes Neuronales

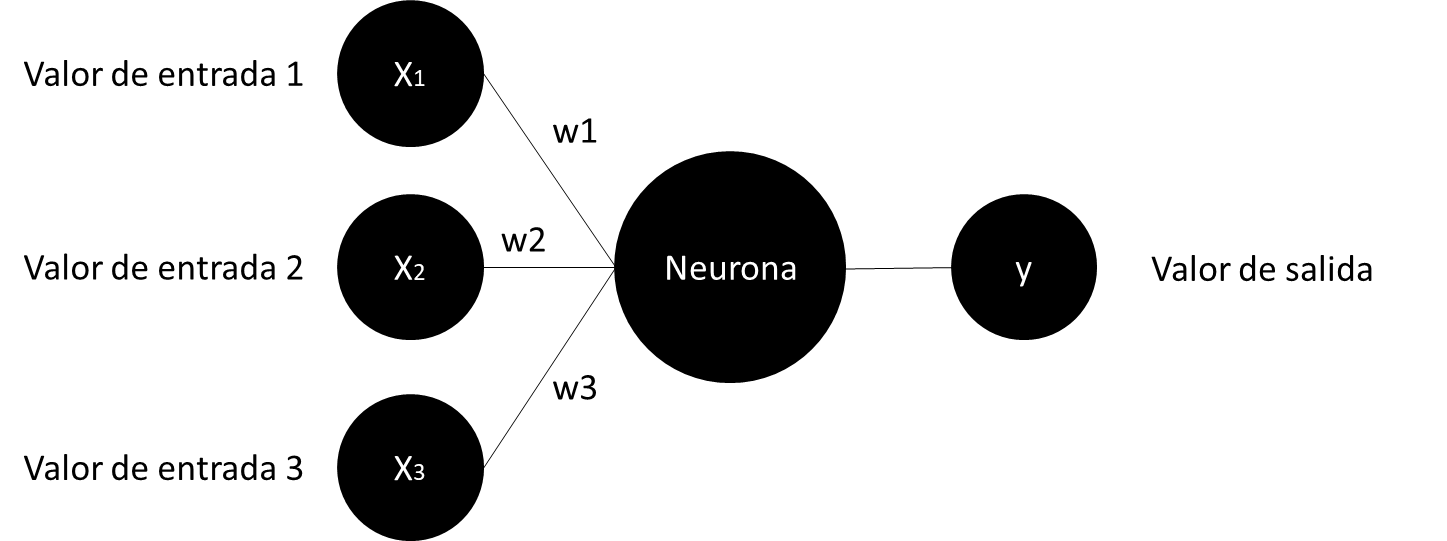
Conjunto de algoritmos matemáticos desarrollados para aprender patrones en la data de entrenamiento.

##### La neurona:

Unidad básica de procesamiento dentro de una red neuronal, trata de imitar el comportamiento de la neurona del cerebro humano. Así como las neuronas humanas tienen señales de entrada producto de estímulos externos y una señal de salida en respuesta, la neurona artificial tiene valores de entrada producto de observaciones (independientes) y un valor de salida producto de un cálculo interno realizado por la neurona.

El cálculo que realiza la neurona internamente es una suma ponderada que representa una ecuación lineal, a cada observación se le asigna un peso para darle forma a la ecuación.

1. Ilustra cómo funciona una red neuronal en su forma más básica



Fuente: Propia

El resultado de la suma ponderada debe pasar por un filtro llamado función de activación cuya función es de distorsionar el plano generado por la neurona. Una neurona por si misma solo puede predecir correctamente cuando la solución es tan simple como graficar una ecuación lineal, pero cuando una ecuación lineal no es la solución entonces se necesita combinar dos o más neuronas.

##### Función de activación:

Función que nos permite encadenar neuronas para producir un resultado jerarquizado o profundo. Esta función modifica el valor de salida de la suma ponderada.

1. Definiciones y formas de las funciones de activación.

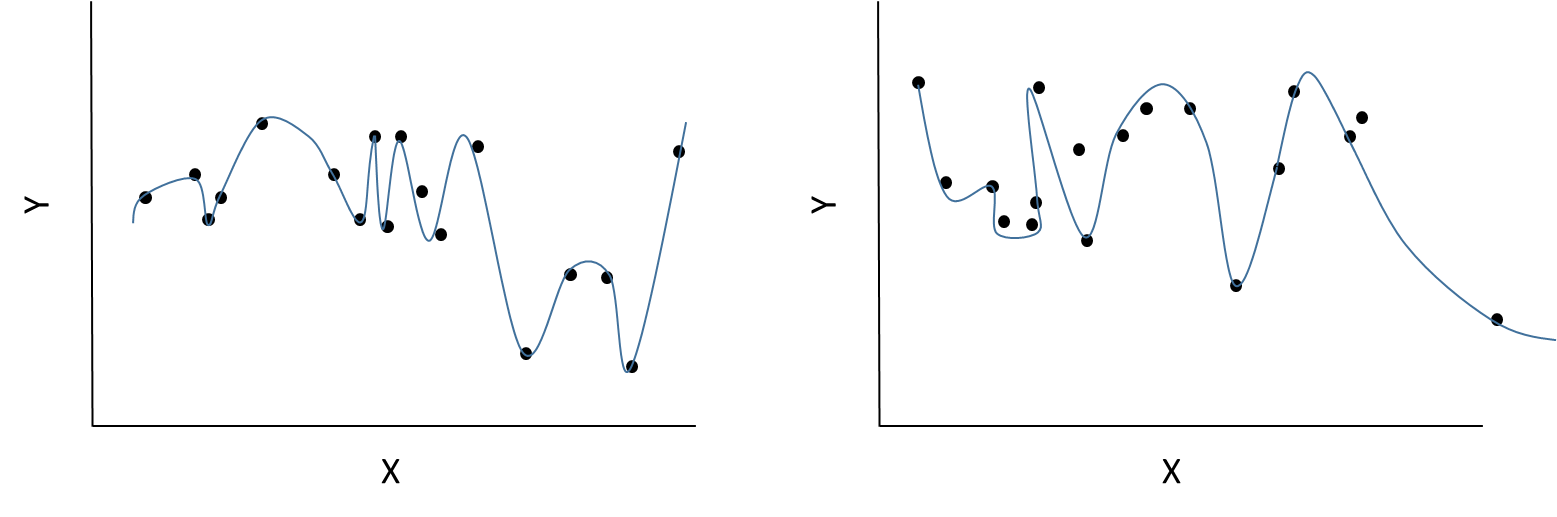
|  |  |  |
| --- | --- | --- |
| Nombre | Definición | Gráfica |
| Umbral | Es la más simple, el valor de entrada es la suma ponderada de los valores de entrada de la neurona, ósea el valor de salida de la misma. Los valores de salida están en el rango de 0 a 1. | Artificial Neural Network |
| Sigmoide | Usado en los modelos de regresión logística cuando se quiere clasificar datos en dos categorías. Cuando la salida de la neurona es menor que 0, esta función obtiene un valor aproximado a 0 y cuando está por encima de 0, se aproxima a 1. | Redes Neuronales |
| Rectificadora | También llamada función ReLU, es una función parcialmente lineal, se hace 0 cuando los valores de entrada son menores a 0 y caso contrario el valor de entrada es el mismo al de salida.  Es el más usado por su rápida convergencia. | Redes Neuronales |
| Hiperbólica | Se parece a la función sigmoidal con la única diferencia que la esta vez la salida está en rango de -1 a 1. De esta manera resuelve el problema de asimetría de la función sigmoidal. | Activation Functions in Neural Networks | LaptrinhX |

*Nota.* Esta tabla muestra las funciones de activación más conocidas y ampliamente usadas en el desarrollo de una red neuronal.

### Equilibrio entre el sesgo (bias) y la varianza

El sesgo es el error en el rendimiento de un modelo, puede ser calculado por las diferentes métricas disponibles, por otro lado, la varianza mide la variación de un modelo al variar el set de entrenamiento. En la siguiente gráfica se muestra dos modelos ajustados, el primero corresponde al modelo generado al entrenar los índices pares de un set de datos de entrenamiento, mientras que en el segundo se usaron los de índice impar. Observamos que entre un modelo y otro hay mucha diferencia, a pesar de pertenecer al mismo conjunto de datos.

1. Modelos ajustados pertenecientes a un mismo conjunto de datos con bajo sesgo

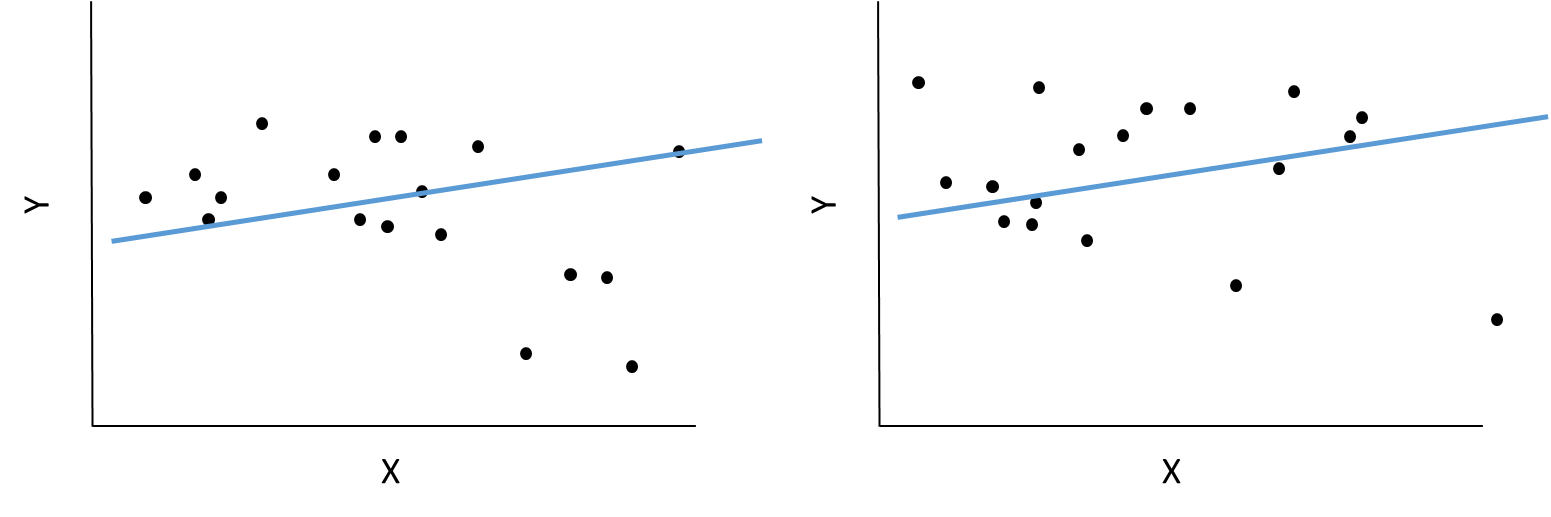


Fuente: Propia

Sin embargo, si calculamos el error de los modelos, obtenemos un error muy bajo. Esto quiere decir que el sesgo es bajo mientras que, al observar la diferencia notoria en la forma del modelo entre el primer y segundo ajuste, decimos que la varianza es alta. Cuando el modelo se ajusta más a los datos de entrenamiento, mayor es la varianza del modelo al variar los datos de entrenamiento.

El caso contrario sería que, al tener un mayor sesgo, tendríamos una menor varianza. En la siguiente grafica mostramos diferentes modelos ajustados, que tienen muy poca variación, pero a su vez implica un alto sesgo.

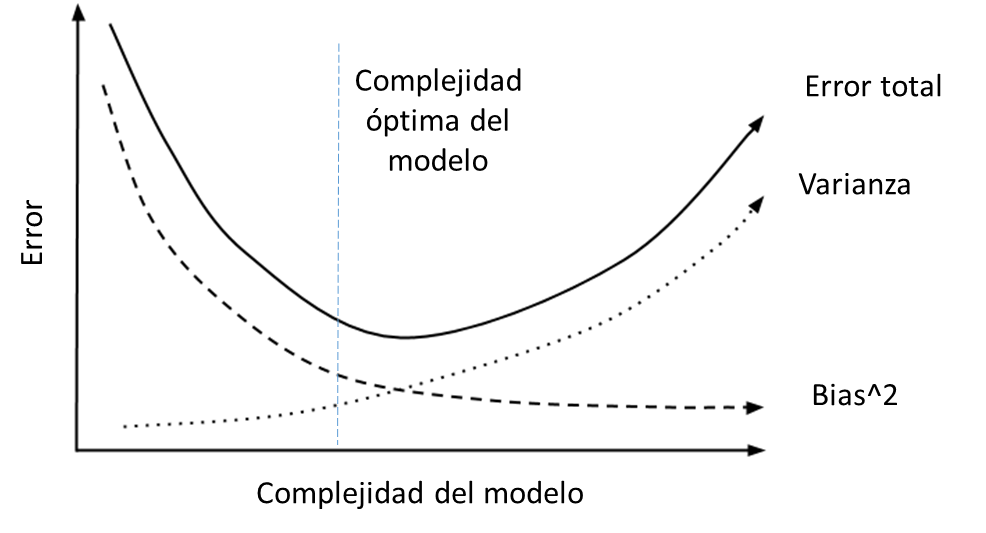
1. Modelos ajustados pertenecientes a un mismo conjunto de datos con baja varianza



Fuente: Propia

De lo expuesto, podemos decir que un incremento del sesgo o la varianza implica un inminente incremento en el error del modelo. Por lo tanto, lo ideal es conseguir reducir estos valores a sus mínimos posibles.

1. Ilustra el equilibrio entre el sesgo y varianza



Fuente: Propia

El modelo óptimo sería el que tenga un bajo sesgo y varianza, pero ya que no es posible se aspira a tener un modelo lo suficientemente complejo para tener un sesgo y varianza intermedio.

# CAPÍTULO III: DESARROLLO DEL TRABAJO DE INVESTIGACIÓN

## Metodología

### Tipo de investigación

Este estudio es de tipo exploratorio, ya que busca medir la efectividad de la aplicación de los algoritmos de aprendizaje automatizado para mejorar el cálculo del caudal inicial de producción de un pozo en un campo maduro en etapa de desarrollo viéndose reflejado en la variación en el error del cálculo de este caudal y, en consecuencia, en los parámetros de viabilidad de un proyecto de perforación en un campo maduro.

### Población y muestra

4.2.1 Población:

Los pozos perforados en las campañas de perforación pasadas, en un campo maduro de Talara, con información de locación, parámetros petrofísicos y del reservorio, de antes y después de la ejecución.

4.2.2 Muestra:

Son 367 pozos de producción multi-reservorio, con la información requerida, distribuidos en un campo maduro en Talara.

## Instrumentos de recolección de datos

* Fuentes Bibliográficas de aprendizaje automatizado con Python.
* Librerías de uso libre en GitHub.
* Información petrofísica de todos los reservorios del pozo que pertenezcan a los intervalos recomendados antes y después de la perforación.
* Información de locación, gradientes de presión y propiedades de fluidos de todos los reservorios del pozo que pertenezcan a los intervalos recomendados antes y después de la perforación.

## Análisis e interpretación de la información

En el estudio se recopilará y compilará toda la información de las recomendaciones de todos los pozos que tengan información petrofísica y del reservorio, de esta manera se tendrá la data estructurada para su uso más ágil. Luego se aplicará el flujo de trabajo para la aplicación de algoritmos de aprendizaje automatizado y se establecerá un modelo para el cálculo del caudal inicial de producción usando la data previamente preparada. Finalmente se estimarán las reservas y se evaluarán los indicadores de viabilidad de un proyecto de perforación usando el método propuesto para comparar si se aleja más o menos de la rentabilidad real, que el método convencional.

## Descripción de la data

En el set de datos que usaremos, las columnas categóricas están enumeradas para guardar confidencialidad. También se debe mencionar que las variables usadas están relacionadas de forma directa e indirectamente con la variable objetivo.

Hay variables dependientes e independientes, donde las dependientes también son llamadas sintéticas, esto quiere decir que son producto de alguna fórmula que involucra otras variables y se basan en el fundamento teórico, pero que dentro de la metodología para la aplicación del aprendizaje automático puede ser considerado como resultado de un paso más en la ingeniería de datos.

* **Id:** Numeración de filas
* **Pozo:** Nombre del pozo
* **Yacimiento:** Nombre del Yacimiento
* **Formación:** Nombre de Formación
* **Base**: Base de intervalo baleado (pies)
* **Tope:** Tope de intervalo baleado (pies)
* **Hbruto:** Espesor bruto de intervalo baleado (pies)
* **Hneto:** Espesor neto de intervalo baleado (pies)
* **Porosidad:** Porosidad de la formación baleada (fracción)
* **Saturación de agua:** Saturación de agua de la formación baleada (fracción)
* **KTI:** permeabilidad de Timur (md)
* **Gradiente Inicial:** gradiente de presión inicial de la formación baleada (psi/ft)
* **Gradiente actual:** gradiente de presión actual de la formación baleada (psi/ft)
* **Uo:** Viscosidad del petróleo (cp)
* **Bo:** Factor de formación volumétrica (bls/stb)
* **Cfluidos:** Comprensibilidad del fluido (1/psia)
* **API:** Grados API del petróleo de la formación baleada

Cada uno de estos parámetros debe ser estimado durante la etapa de planificación del proyecto, en esta etapa recolectamos la suficiente cantidad información para garantizar la rentabilidad de los pozos. Cuando se tiene esta información, se calculan las reservas que se espera obtener de cada pozo y se comparan en el aspecto financiero para que puedan pasar a la etapa de ejecución. Luego, se hace un monitoreo y control de los pozos ya perforados y se obtienen conclusiones del contraste entre lo que se esperaba (pronóstico) y lo que realmente ha producido el pozo.

Todo este contexto nos lleva a esforzarnos por maximizar la confiabilidad de los datos que se usan para el cálculo de reservas, el cual tiene como parámetro más influyente al caudal inicial de producción.

Cabe mencionar, que estos datos deben ser obtenidos antes de la perforación del pozo, a continuación, mencionaremos como se obtiene cada uno de estos atributos.

Atributos como el id y el nombre de pozo son identificadores del pozo, estos no pueden faltar en una tabla que se integrara a una base de datos relacional para una mejor administración de la información. El yacimiento y la formación son propios de la ubicación del pozo producto de la evaluación geológica y recomendación del geólogo, esta toma en cuenta factores como el espaciamiento entre pozos y la desviación óptima para maximizar el número de formaciones objetivo a perforar; sin embargo, todo proyecto tiene su alcance y esta selección de ubicaciones debe cumplirlas a criterio de los intereses de la operadora.

Los atributos como la base, tope y el espesor neto y bruto, delimitan y miden el intervalo de donde se espera producir las reservas, estas se obtienen producto de las correlaciones entre las formaciones de los pozos vecinos, producto de este análisis es que se pueden evitar las fallas que suelen desplazar parcial o totalmente el objetivo principal de un pozo. Para estas correlaciones se usan los registros eléctricos y secciones disponibles de cada pozo vecino lo suficientemente cerca para que influya en el comportamiento productivo, el análisis se basa en la identificación de patrones de continuidad entre las respuestas de los registros eléctricos (GR, SP y Resistividad) y la sección de los pozos que ya han sido analizados previamente y actualizados a la fecha. Cabe mencionar, que a pesar de tener correlaciones confiables que identifican la ubicación de las fallas y su dirección, aún existen algunas que no llegan a ser mapeadas y esto es producto de la complejidad geológica propio del campo.

La porosidad y la saturación de agua se obtienen promediando el valor obtenido en pozos vecinos a través de sus registros eléctricos para cada formación respectivamente.

La permeabilidad de Timur se estima de una formula obtenida de forma empírica que tiene por argumento a las variables petrofísicas de porosidad y saturación de agua promedio calculadas anteriormente.

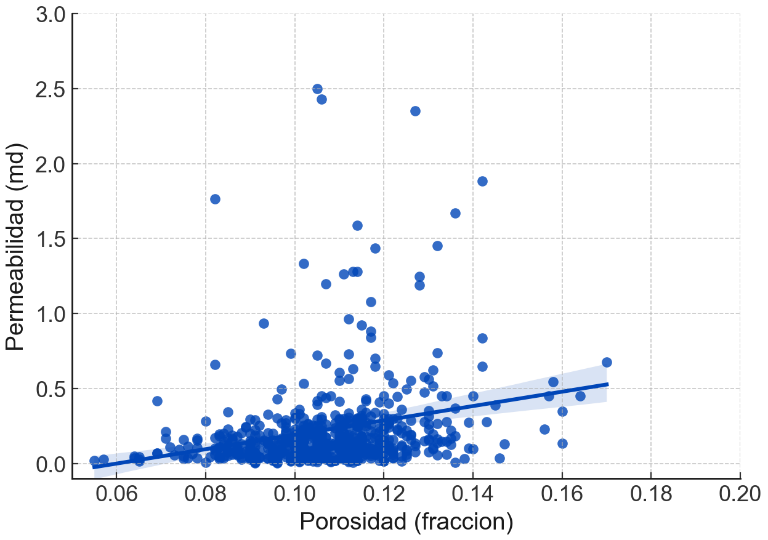
Las gradientes de presión inicial y actual son estimadas usando un plot de producción acumulada versus los datos de presión que se han tomado históricamente llevados a un datum de referencia, de esta manera identificamos la tendencia del comportamiento de la curva presión para poder extrapolar su comportamiento.

## Obtención de la data:

### Estimación de Permeabilidad a través de Ley K-Phi

Para la estimación de la permeabilidad, se hizo uso de datos de núcleos por formación y se calculó una Ley K-Phi para cada una de las formaciones objetivo del pozo. Además, usando los perfiles eléctricos de pozos vecinos podemos estimar una porosidad que a través de la fórmula que nos provee la ley k-phi se convierte en un estimado empírico de la permeabilidad, el cual es un parámetro fundamental en el cálculo del caudal inicial.

1. Grafica K vs Phi de una de las Formación X.



Fuente: Propia

Esta permeabilidad empírica se ajusta, pero usando información que obtenemos luego de perforar y completar el pozo, como por ejemplo haciendo uso de los parámetros estimados producto de la interpretación de la prueba mini frac, la cual se suele practicar en los pozos del noroeste peruano usando como fluido base el agua.

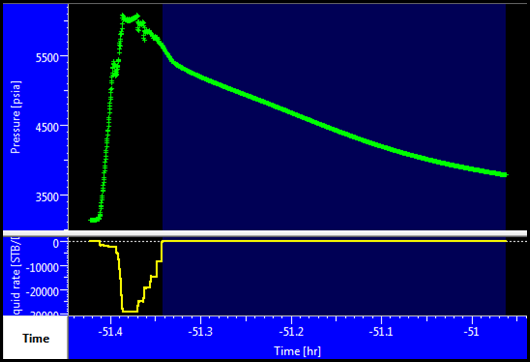
### La Prueba de Minifrac

En los campos del noroeste peruano las pruebas de minifrac se realizan en cada una de las etapas perforadas del pozo. La interpretación de estas pruebas se hace con el objetivo de aproximar la permeabilidad absoluta del petróleo a partir de la permeabilidad del agua obtenida, y en algunos casos era posible estimar la presión actual del reservorio dependiendo si la presión se había llegado a estabilizar después del cierre. El procedimiento típico consiste en:

Los sensores arrojan una raw data (.dat) que se transforma al formato requerido dependiendo del tipo de software que se usara para la interpretación, típicamente es el formato “.txt”. El primer paso es cargar la data principalmente de presión y caudal para su limpieza y visualización. Esto incluye hacer la conversión correspondiente, eliminar los outliers y corregir la data de salida del sensor usando sus valores cercanos.

Posteriormente se eligen el modelo y los parámetros que la caracterizan para su inicialización.

1. Visualización de curvas Time – Caudal – Presión

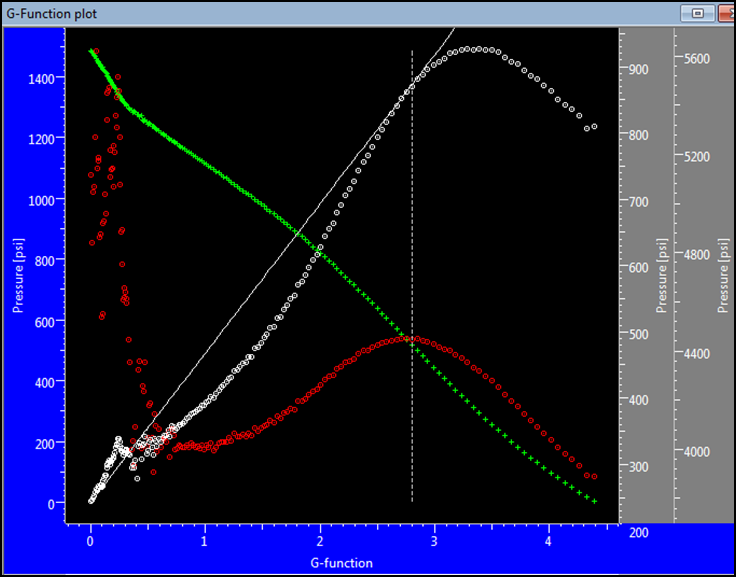


Fuente: Propia

Siguiente a esto se obtiene la derivada de la función G en el periodo de caída de presión cuando el pozo se cierra.

Esto nos permitiría diagnosticar el tipo de comportamiento a través del análisis de la forma de la “Función G” y el “Square Root Plot”. Existen 4 curvas tipo que nos permiten identificar el tipo de comportamiento de la fractura. A continuación, se muestra la función G.

1. La derivada de la función G



Fuente: Propia

Finalmente se estima la permeabilidad usando la viscosidad y el modelo de Young de la formación analizada. La ecuación usada para la estimación de la permeabilidad fue:



Donde:

k: Permeabilidad efectiva, milidarcys

: Viscodidad, centipoise

Pz: Presión neta

ct: Compresibilidad total,

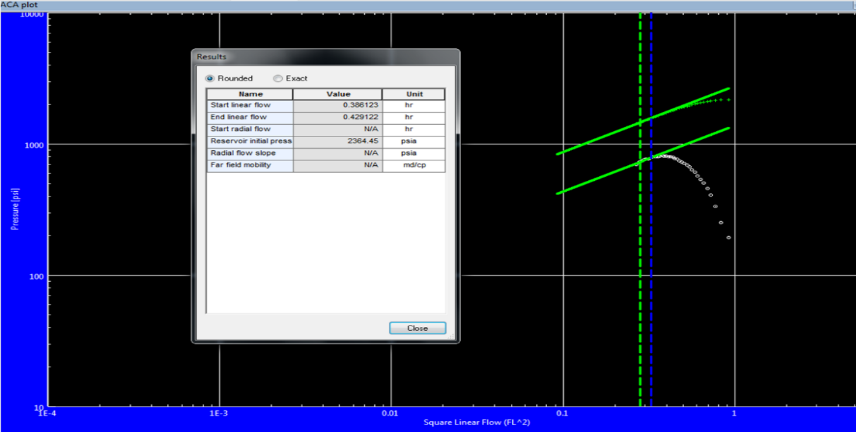
: Porosidad, fracción

E: Modulo de Young, MMpsi

rp = Altura de leakoff o relación de espesor de fractura, ft

En caso se hubiese llegado al cierre, usamos el ACA plot para evaluar si se llegó a identificar algún flujo, con el objetivo de estimar la presión actual del reservorio. A continuación, se muestra la estimación de la presión del reservorio a partir de la identificación de un flujo en el ACA plot.

1. Análisis de Minifrac después del cierre (ACA)



Fuente: Propia

### Estimación de parámetros PVT (Sw, Bo, Uo, Ct, API)

Para la estimación de los parámetros de PVT se usaron correlaciones o modelos de referencia predeterminados definidos en el fundamento teórico del presente trabajo. Teniendo curvas que siguen una ecuación o comportamiento podemos ajustarlas a nuestros datos puntuales más confiables, muchas veces no basta el ajuste con menor estimado de error medio, sino que se toma en cuenta el comportamiento en modelos más complejos donde el modelo de fluidos es un valor de entrada determinante.

### Estimación de Permeabilidad a través de Ley de Timur

Teniendo en cuenta el efecto de la saturación de agua en la permeabilidad, se realizaron estudios que involucraban formulas empíricas que tomaran en cuenta la saturación de agua irreducible (Swi) y la porosidad en la permeabilidad absoluta. Según Timur este efecto se considera en la siguiente formula:

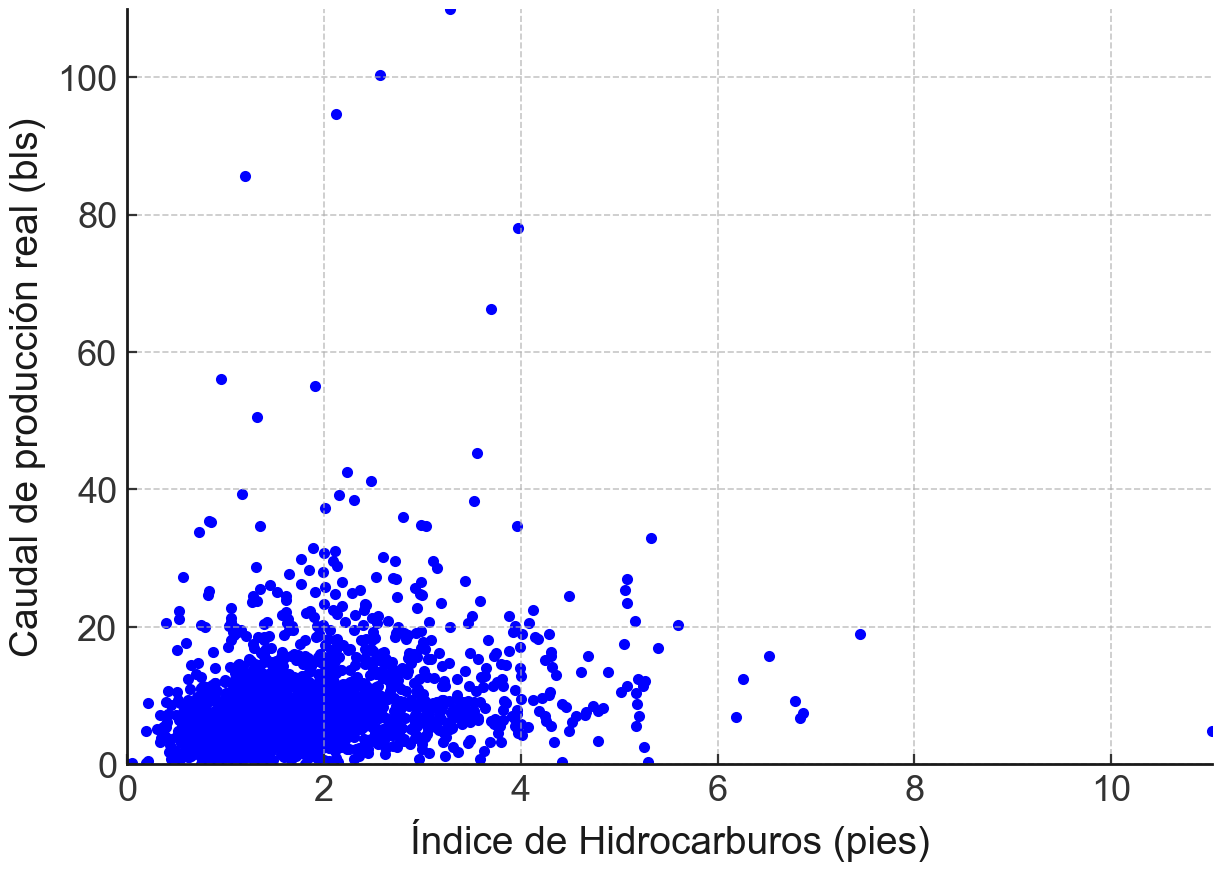
Donde Swi es la saturación de agua irreducible y es la porosidad, ambos expresados en porcentaje. Cabe destacar que en la fórmula no se toma en cuenta el tipo de hidrocarburo presente en la roca.

### Índice de Hidrocarburos

Es un parámetro calculado que resulta del producto del espesor neto, porosidad y saturación, este valor sintético es buen indicador para discriminar zonas con alto potencial productivo.

Donde es la saturación de petróleo, es la porosidad, ambos expresados en porcentaje y H es el espesor neto de interés y está en pies.

1. Índice de Hidrocarburos (pies) vs Caudal de producción real (bls)



Fuente: Propia

### Curvas de declinación para el cálculo de las reservas de petróleo.

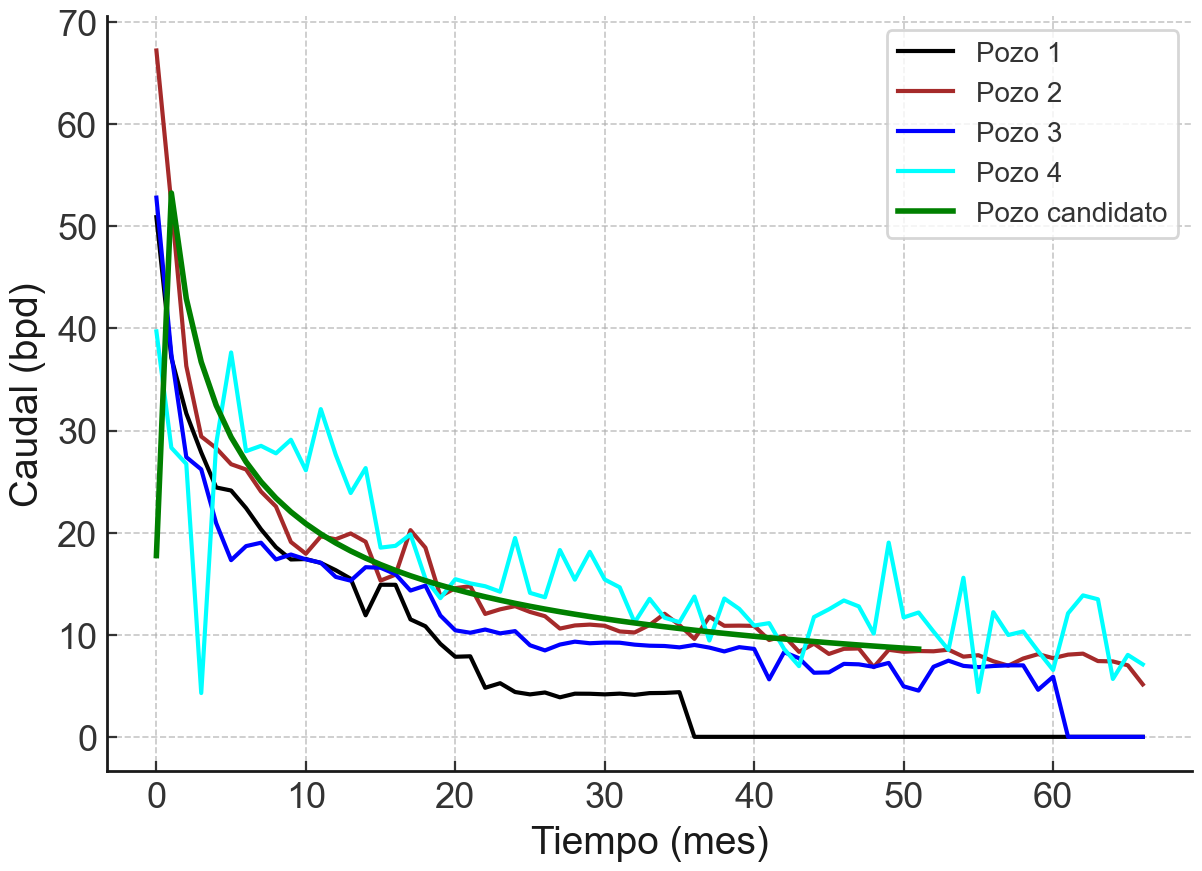
Tomando en cuenta que la mayor cantidad de data acumulada en los campos de Talara provienen de su historia de producción, es confiable usar las curvas de Arps para estimar las reservas usando como referencia los pozos vecinos. Los rangos de valores que siguen los parámetros b y D de los pozos de Talara, no pertenecen a ninguno de los 3 tipos de curva estándar, ya que el valor b supera la unidad.

Estos rangos son variables, pero se pueden estimar usando como referencia los parámetros b y D de los pozos vecinos, estos parámetros deben ser ajustados en el momento en el que se está elaborando la recomendación para el cálculo del caudal inicial, sino no podríamos saber el estado actual de la declinación en la zona potencial para perforar.

Cabe mencionar que este tipo de análisis se suele realizar por reservorio o formación productiva. A continuación, se mostrará un ejemplo de cómo se aproximan estas curvas.

Se selecciona la ubicación de un pozo, y se quiere evaluar el comportamiento productivo de los pozos vecinos que hayan producido de las mismas formaciones.

1. Gráfico realizado para estimar los parámetros b y D a través de la comparación con pozos vecinos, en un pozo candidato



Fuente: Propia

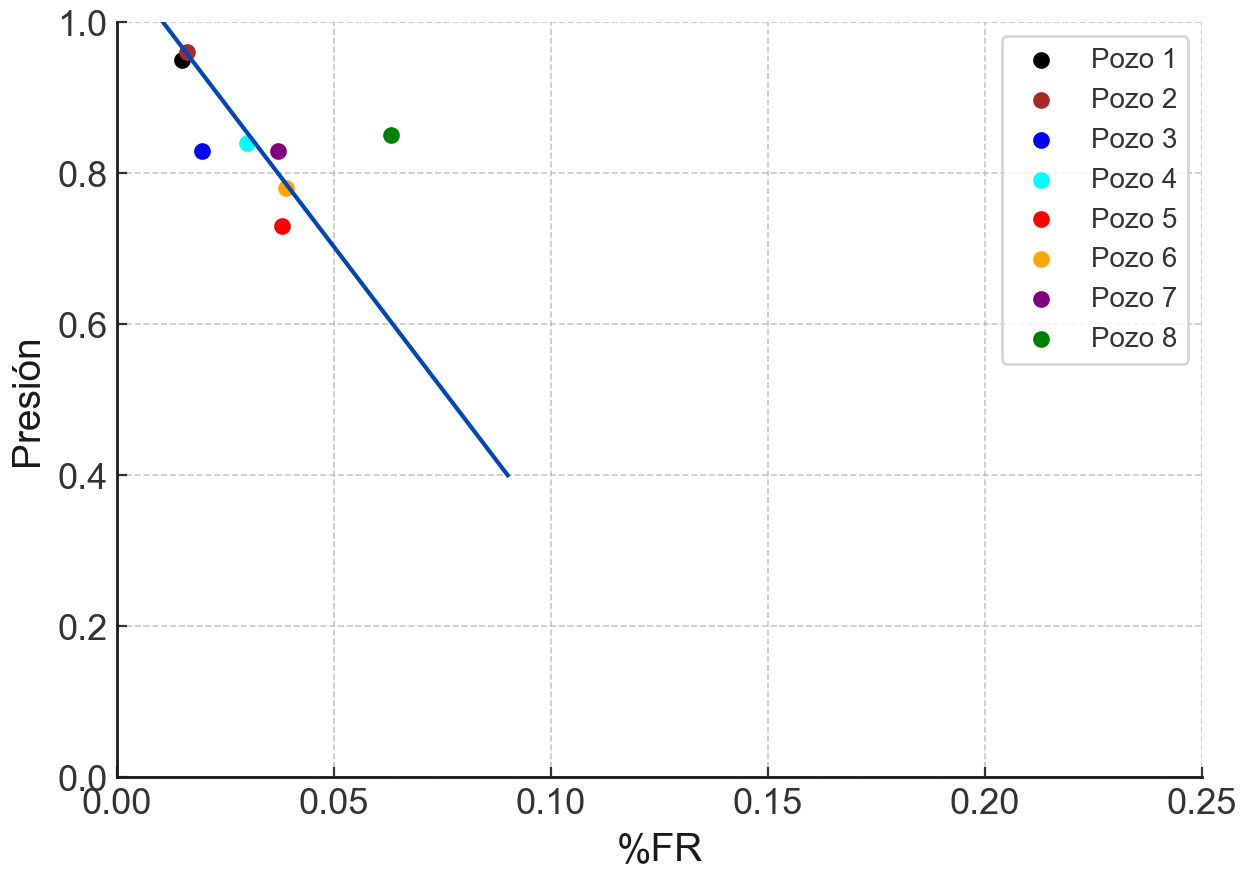
Como podemos observar en la anterior figura la linea verde es el pronóstico de producción en base al ajsute de la curva de declinacion con sus parametros b y D. La intencion de plotearlas juntas es emular la tendencia en el comportamiento productivo del reservorio. Al graficar el comportamiento productivo de varios pozos tambien podemos evaluar en que grado los pozos estan interconectados y nos brinda informacion para estimar la gradiente de presion del pozo en base al acumulado producido.

### Cálculo de la gradiente de presión

Para calcular la gradiente de presión de un pozo infill es necesario saber el comportamiento de la presión en el bloque, actualmente con la producción de los pozos de Talara no justifica la toma de pruebas de presión y considerando que en la mayoría de pozos solo se ha tomado una sola vez la prueba de presión graficaremos las presiones tomadas en diferentes tiempos con la producción acumulada del reservorio en esas fechas donde se tomó la prueba, de esta manera identificaremos la tendencia de la presión para cada reservorio.

Una vez identificada la tendencia, con el dato del factor de recobro actual, nosotros podemos extrapolar la presión actual.

1. Grafico que ilustra cómo se estima la gradiente actual de un reservorio



Fuente: Propia

## Preparación de datos

En esta etapa creamos un set de datos que sera resultado de una limpieza de los mismos datos recolectados anteriormente. Cabe destacar que encriptaremos el nombre de los pozos y formaciones productivas.

En la preparacion de los datos, buscamos estandarizar el formato de cada atributo para que todos sigan el mismo patrón y sea escalable, con el fin de poder llegar a interpretar nuestros datos. Siendo esta parte fundamental, ya que si usamos datos sesgados para construir nuestro modelo, obtendremos un modelo de la misma calidad.

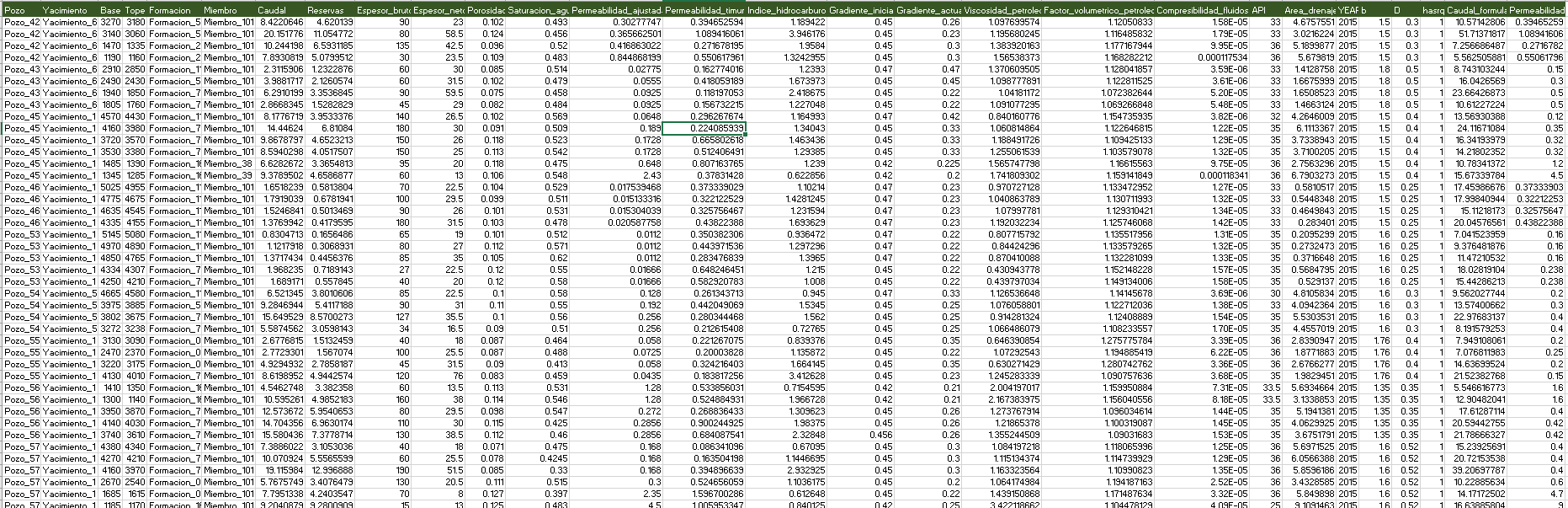
La evaluacion de potenciales ubicaciones para perforar se hace de manera individual y para cada posible ubicación se extrae todos los datos que se presentan a continuación:

1. Breve descripción de los atributos que se usaran para el modelo

|  |  |  |
| --- | --- | --- |
| Atributo | Tipo de variable | Resumen |
| Pozo | Categórica | Identificador de un pozo. Resultado de la concatenacion de la palabra ‘pozo’ con el id. |
| Yacimiento | Categórica | Delimitación de bloque estructural. Resultado de la concatenacion de la palabra ‘yacimiento’ con el id. |
| Base | Numérica | Limite inferior de posible zona productiva según correlaciones e historia de la zona. Todos medidos en pies. |
| Tope | Numérica | Limite superior de posible zona productiva según correlaciones e historia de la zona. Todos medidos en pies. |
| Formación | Categórica | Identificador para el tipo de formación. Resultado de la concatenacion de la palabra ‘Formacion’ con el id. |
| Miembro | Categórica | Identificador de miembros en la formación. Resultado de la concatenacion de la palabra ‘Miembro’ con el id. |
| Caudal | Numérica | Volumen de petróleo por dia que produce el pozo. Resultado de la ecuación de difusividad. |
| Reservas | Numérica | Total de volumen de petróleo remanente en el pozo. Variable continua función de b, Di y Qinicial. |
| Espesor bruto | Numérica | Espesor de la formacion productiva. Variable resultado de la resta entre el tope y la base. |
| Espesor neto | Numérica | Espesor neto de la formacion productiva. Variable resultado de la multiplicacion del espesor bruto por un factor entre 0 y 1. |
| Porosidad | Numérica | Es la fracción de espacio poral de la formación productiva. Variable promedio de pozos vecinos |
| Saturacion de agua | Numérica | Relación de la cantidad de agua en el espacio poroso de la formación productiva. Variable promedio de pozos vecinos |
| Permeabilidad | Numérica | Variable que mide la capacidad de fluir de la formación. Variable promedio de pozos vecinos. |
| Permeabilidad (Timur) | Numérica | Permeabilidad estimada a traves de una ecuacion empírica. Variable calculada usando la porosidad y saturación de agua |
| Indice de hidrocarburos | Numérica | Indicador que busca medir la calidad de las zonas para la producción de petróleo. Resultado de la multiplicación del espesor neto, la porosidad y la saturación de petroleo (HPhiSo) |
| Gradiente inicial | Numérica | Fracción de la presión inicial |
| Gradiente actual | Numérica | Fracción de la presión actual |
| Viscosidad de petroleo | Numerica | Variable viscosidad resultado de una correlación en cp |
| Factor Volumetrico del petroleo | Numérica | Variable PVT resultado de una correlacion |
| Factor de compresibilidad de fluido | Numérica | Variable cf que se obtiene de correlaciones |
| Grados API | Numérica | Variable promedio de la zona |

El análisis de cada ubicación y la información descrita anteriormente, se almacenan en tablas dentro de las planillas de excel las cuales son compiladas aprovechando el modelo estándar de las mismas.

1. Captura de la dimension de la tabla con la data de entrada



Fuente: Propia

La limpieza de datos para nuestro dataset consistirá en eliminar obervaciones que tengan valores nulos en la mayoria de sus atributos e identificar valores que se alejen del concentrado de datos para cada atributo numérico.

### Eliminando valores nulos

En un dataset compilado simpre existiran observaciones que tienen ausencias en algunos atributos, esto se debe a que no existe la informacion exacta para ese intervalo evaluado o algun problema de optimización en la compilación y extracción de datos, que deberá ser corregido en lo posible. La información que obtenemos de nuestra tabla compilada es:

1. Tipos de datos y contador de nulos en la data de entrada



Fuente: Propia

Tambien identificamos que el atributo miembro tiene menos del 50% de datos no nulos, una cantidad importante considerando la diversidad de la misma al ser subgrupo de una formación. Debido a esto eliminamos esta columna de nuestra tabla.

## Representación de datos

### Exploración de datos (EDA)

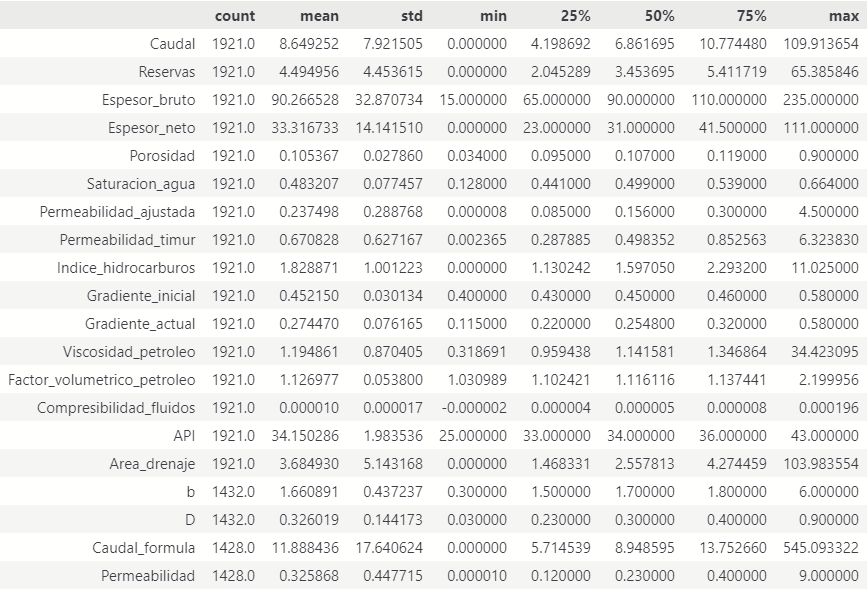
Con la data compilada y limpia pasamos a la exploracion de datos donde usaremos gráficos estadísticos para análisis univariado y bivariado, esto nos permitira identificar outliers y depurarlos, ya que sesgan el proceso de aprendizaje de nuestro algoritmo.

Detectar inconsistencias, probar, determinar y entender tipos de relaciones entre los atributos, generar hipótesis e identificar patrones son ventajas de este análisis que nos sirven para ejercer el mejor criterio de selección de variables para nuestro modelo.

Nuestro dataset luego del proceso de limpieza tiene las siguientes dimensiones: (1921 observaciones, 17 atributos).

El primer paso en el EDA es la estadística descriptiva de las variables mostradas en la siguiente tabla.

1. Estadística descriptiva de los valores numéricos de la data de entrada.



Fuente: Propia

Para lograr esto realizaremos una descripción univariada de la data usando gráficos estadísticos. El análisis univariado consiste en identificar los valores de la data dividida en diferente cantidad de partes, para los atributos de nuestro dataset usaremos los percentiles y los identificaremos a traves de graficar la distribución de cada dato usando histogramas y diagrama de caja para mayor visualización.

En el gráfico anterior identificamos los valores estadísticos para cada atributo numerico e identificamos que la compresibilidad de fluidos tiene un valor mínimo negativo, lo cual contradice la física, esto puede ser producto de la correlacion que se haya usado para obtener esta variable PVT para las condiciones del intervalo o algun cambio en las posiciones de las variables que origino este cambio de signos. Se identificó que fue un error en el cálculo y el verdadero valor para esos valores negativos es el mismo pero en positivo. Habiendo superado esa disconformidad con los datos, observaremos la distribucion de cada atributo a traves del uso de graficas para mejor visualización. La linea punteada roja representa la mediana, mientras que la verde representa la media de la distribucion.

1. Descripción de los atributos principales en base al analisis univariado

|  |  |  |
| --- | --- | --- |
| Atributo | Observación | Gráfica |
| Caudal (bopd) | Se observa una distribucion unimodal, y los datos estan concentrados hacia la izquierda, lo cual nos indica que posee una asimetrica positiva y tambien observamos que la frecuencia esta por encima de la normal, lo cual nos indica un apuntamiento de tipo leptocurtica. |  |
| Espesor Bruto (pies) | Se observa una distribucion unimodal, asimetrica ligeramente positiva y apuntamiento muy ligeramente de tipo leptocurtica casi normal. |  |
| Espesor Neto (psi) | Se observa una distribucion unimodal, asimetrica ligeramente positiva y apuntamiento muy ligeramente de tipo leptocurtica. |  |
| Porosidad (fracción) | Se observa una distribucion unimodal, asimetrica positiva y apuntamiento fuertemente de tipo leptocurtica. |  |
| Saturación de agua (fracción) | Se observa una distribucion unimodal, asimetrica ligeramente negativa y apuntamiento ligeramente de tipo leptocurtica. |  |
| Indice de Hidrocarburos (pies) | Se observa una distribucion unimodal, asimetrica positiva y apuntamiento de tipo leptocurtica. |  |
| Gradiente inicial de presion (psi/pies) | Se observa una distribucion unimodal, asimetrica positiva y apuntamiento de tipo leptocurtica. |  |
| Gradiente actual de presion (psi/pies) | Se observa una distribucion unimodal, asimetrica ligeramente positiva y apuntamiento ligeramente de tipo leptocurtica. |  |
| Viscosidad de petroleo (cp) | Se observa una distribucion unimodal, asimetrica positiva y apuntamiento fuertemente de tipo leptocurtica. |  |
| Factor Volumetrico de petroleo (bbl/stb) | Se observa una distribucion unimodal, asimetrica positiva y apuntamiento fuertemente de tipo leptocurtica. |  |
| Compresibilidad de fluido (1/psi) | Se observa una distribucion unimodal, asimetria positiva. Por otro lado el apuntamiento es de tipo leptocurtica. |  |
| Gravedad API | Se observa una distribucion unimodal, asimetrica ligeramente negativa y apuntamiento ligeramente de tipo leptocurtica. |  |

### Eliminando valores atípicos

Son valores extremos que sobrepasan el rango normal de lo esperado y se alejan de los otros datos. Estos valores serán removidos a criterio con el objetivo de elevar la precisión del modelo de aprendizaje automático.

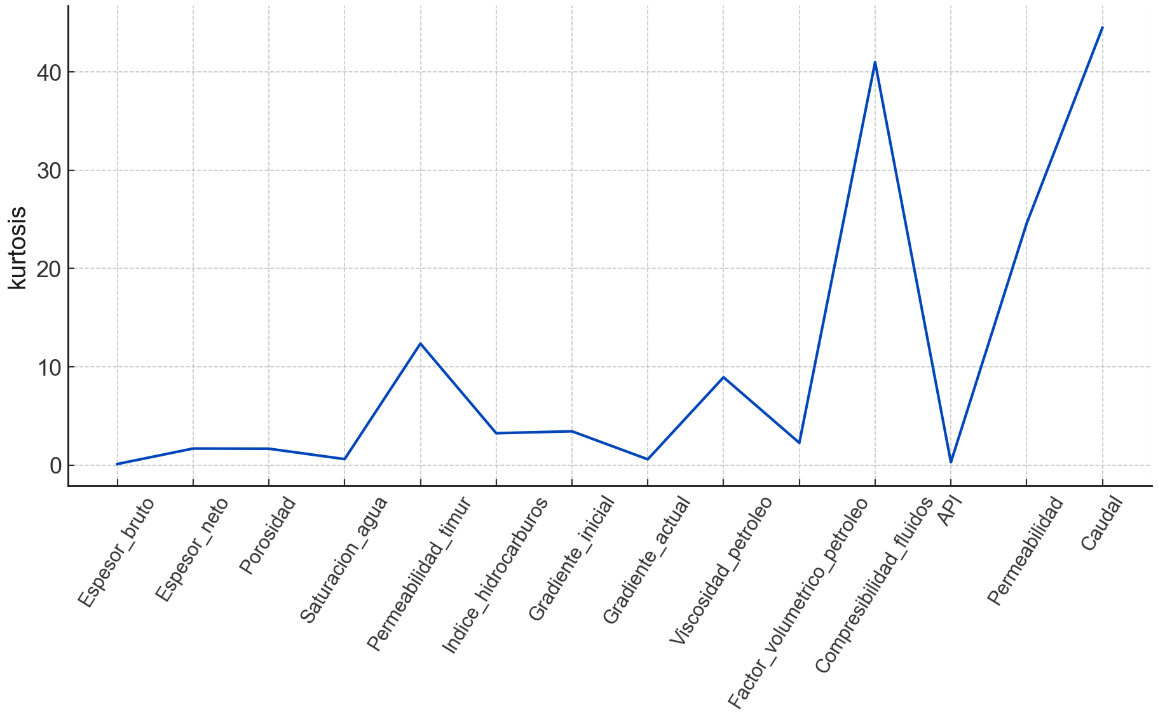
Cabe mencionar que estos valores atípicos no siempre sesgan los resultados del modelo, algunos pueden delimitar los valores de un atributo o ser puntos de interés para el objetivo final.

Los valores atípicos se identifican teniendo conocimiento de la naturaleza de las variables en cuestión, y tambien existen herramientas matemáticas que nos ayudan a identificarlos. Estas herramientas dependen de la dimension del análisis, es decir, podemos identificar valores atípicos de las medidas un atributo (1 dimension) hasta múltiples atributos usando algoritmos de aprendizaje no supervisado que miden la distancia entre los datos (KNN) y tambien algoritmos de reduccion de dimensiones (PCA).

Otro de los metodos para identificar valores atípicos es a través de gráficos convenientemente elegidos. Para el caso de nuestra data usamos específicamente gráficas de análisis bivariado para identificar los valores atípicos, ya que conocemos bien la naturaleza de nuestros atributos.

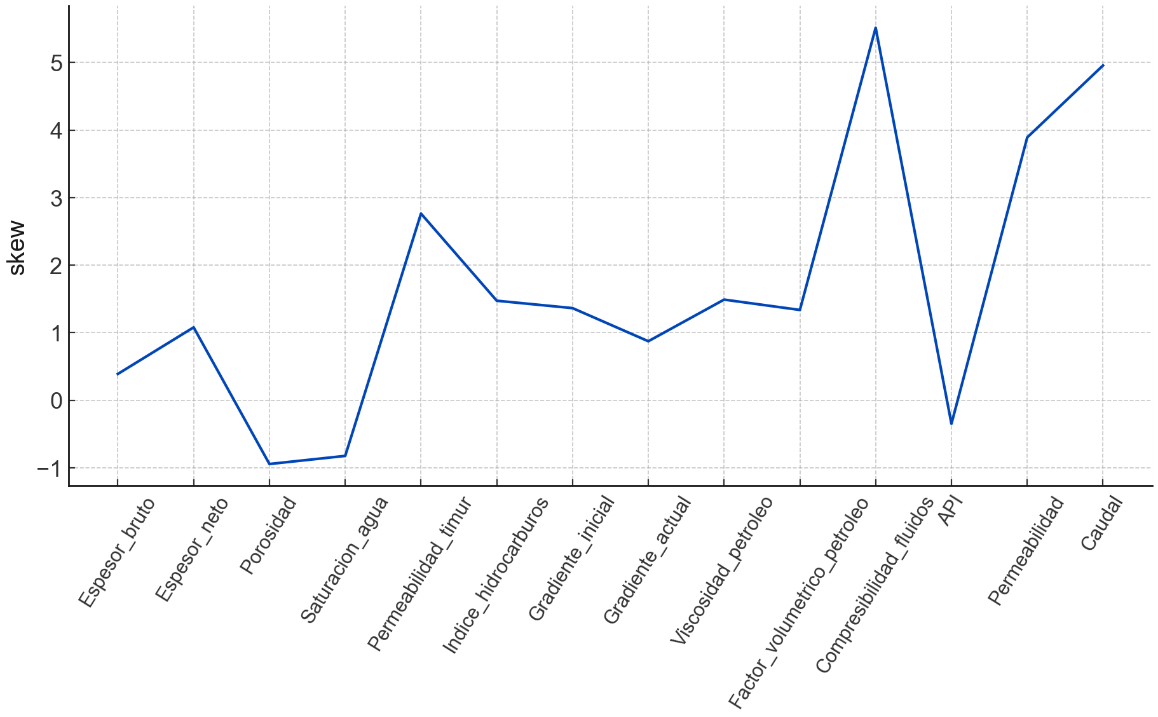
Para darnos una idea general de los atributos en los que debíamos poner especial énfasis usamos el coeficiente de asimetría y kurtosis.

1. Gráfico de la kurtosis para cada atributo numérico



Fuente: Propia

1. Gráfico de la asimetría para cada atributo numérico

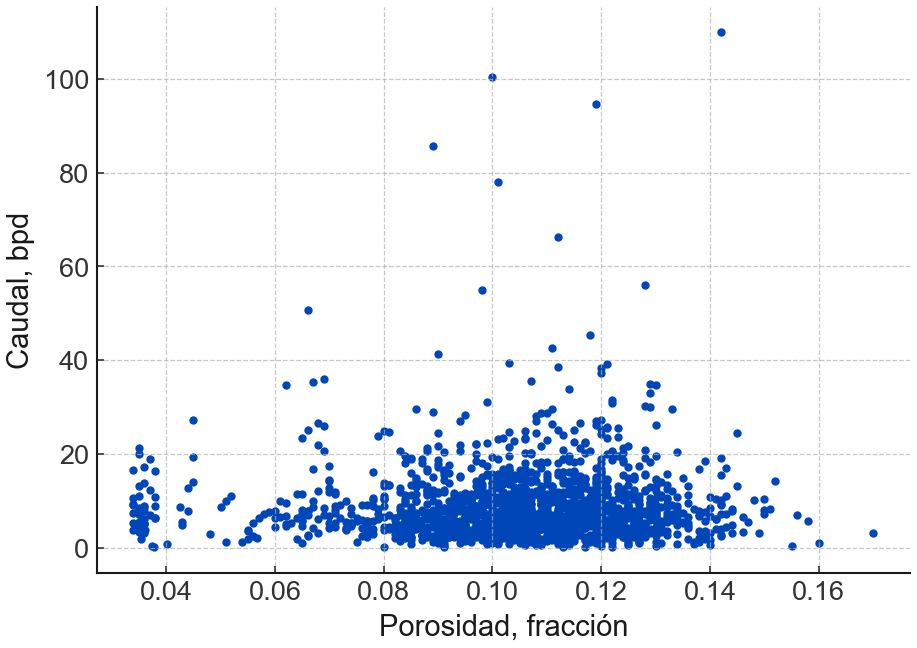
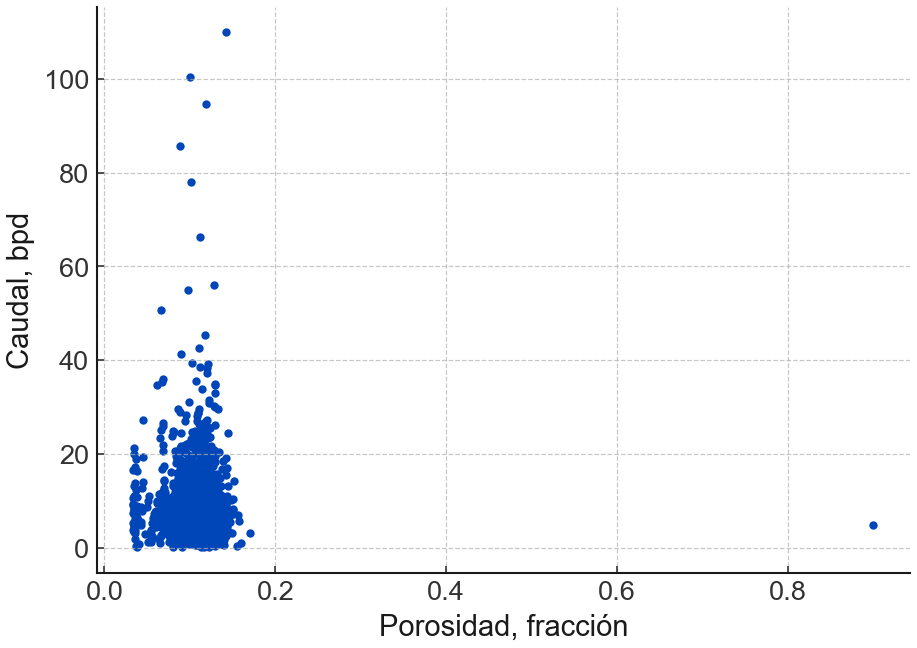


Fuente: Propia

En estos graficos observamos que los atributos con alta asimetria y kurtosis son las elegidas ya que la asimetria denota una tendencia a la concentracion de los datos de un solo lado, dejando los valores de la cola como atipicos. Asimismo, la kurtosis mide la concentracion de datos en rangos acotados generando valores atipicos en los extremos. Entonces se graficó estas variables versus la variable target y se acoto el rango de valores a criterio, obteniendo lo siguiente:

#### Para la porosidad:

1. Gráfico de porosidad antes y después de eliminar los valores atípicos



Fuente: Propia

#### Para la viscosidad

1. Gráfico de la viscosidad antes y después de eliminar los valores atípicos

#### :

Fuente: Propia

#### Para el factor volumétrico.

1. Gráfico del factor volumétrico antes y después de eliminar los valores atípicos

#### 

Fuente: Propia

Realizando el acotamiento del rango de valores de un atributo con valores atípicos se pudo observar que valores atípicos identificados en un atributo también lo eran en otros atributos.

### Transformación de Distribuciones

El tipo de distribución de nuestras variables de entrada influyen significativamente en el desempeño de nuestro algoritmo de aprendizaje automático, por ello es recomendable evaluar la distribución de cada una de nuestras variables y discriminar si deben o no ser normalizadas. Se prefiere que la distribución modelo deba asemejarse a la gaussiana, usando la función logarítmica.

Se usó el valor de asimetría de cada variable para establecer visualmente que los que de valor superior a 0.75 son aquellos que serán transformados usando la función logarítmica antes mencionada.

1. Asimetría de cada variable de entrada de nuestro modelo

#### 

Fuente: Propia

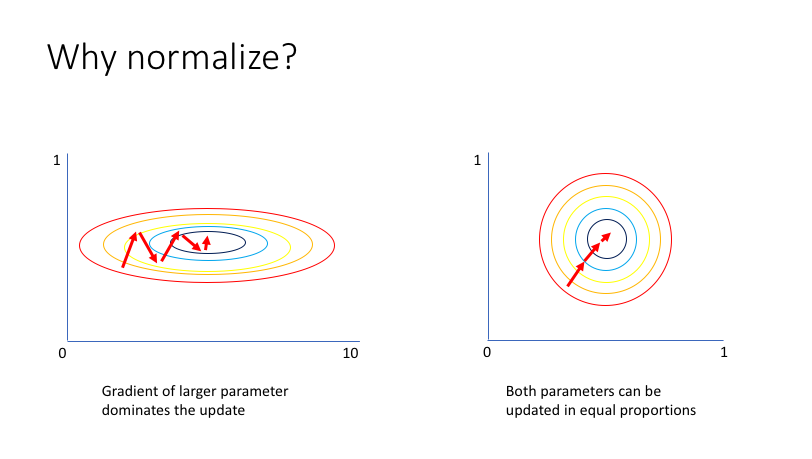
Observamos que la cantidad de variables asimétricas es más del 60% del total. Sin embargo, considerando que nuestra variable objetivo está entre ellos, optaremos por no transformar estas variables y dejar que el escalamiento de datos lo haga a través de una fórmula matemática que incluya la normalización de forma indirecta.

### Escalamiento de datos

También llamada normalización o estandarización, en este proceso transformamos nuestros datos de entrada al llevarlos a una misma escala (uniforme), esto nos beneficia ya que mejora la velocidad de procesamiento y de entrenamiento de nuestro modelo. Esto debido a que los pesos de la gradiente de nuestras variables de entrada son mucho menores, ya que el peso depende directamente de los valores de las mismas variables durante el proceso de entrenamiento.

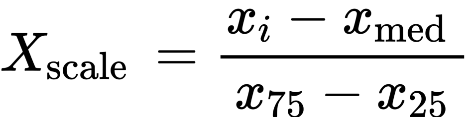
A continuación, se muestra una variable antes y despues de ser normalizada, en la imagen de la izquierda observamos como la gradiente actualiza el error usando vectores de mayor modulo cuando la variable tiene un amplio rango de valores, mientras que cuando la variable esta normalizada la actualización es en la misma proporción.

1. Comparación de una misma variable cruda y normalizada



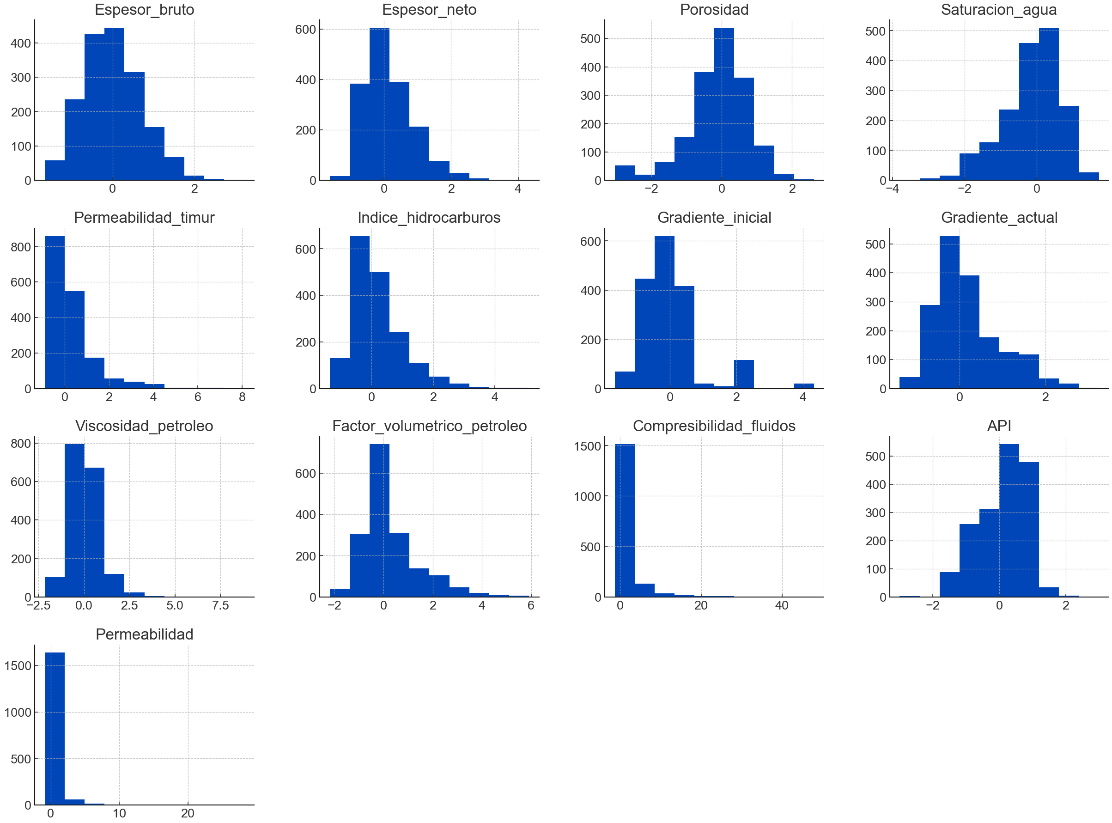
Fuente: Propia

Para el escalamiento de las variables numéricas usaremos la funcion RobustScaler de la librería scikit learn. Cuya funcion tiene la siguiente forma.



Esta funcion permite escalar atributos con valores atipicos. Esto debido a que no eliminamos todos los valores atípicos debido a que se realizó a criterio y algunos valores fueron dejados al representar los límites de algunos atributos como el caudal o espesor neto (medibles y probados). El algoritmo funciona removiendo la mediana y escalando la data de acuerdo al rango cuantil. Este rango cuantil suele ser el IQR, el cual es el rango intercuantil entre el primer y el tercer quantil.

1. Histograma de los atributos numéricos luego de ser escalados



Fuente: Propia

### Codificar Atributos No numéricos

En nuestros datos, tenemos variables categóricas tales como el nombre del yacimiento y la formación, estas seran binarizadas, convirtiendose en nuevas variables que tendran como valores unicos 1 y 0, 1 para los datos que coincidan con la variable en cuestión y 0 para los datos que no coincidan.

Luego de aplicar el algoritmo que realizara la binarización de las variables cualitativas (OneHotEncoder) tenemos todas nuestras variables o atributos en valor numérico y normalizados para la aplicación de los diferentes modelos de aprendizaje automático supervisado.

## Modelamiento.

Esta etapa estará conformada por la elección del algoritmo final que se usará para la predicción del caudal de producción, cada algoritmo será entrenado y validado para su uso en la predicción.

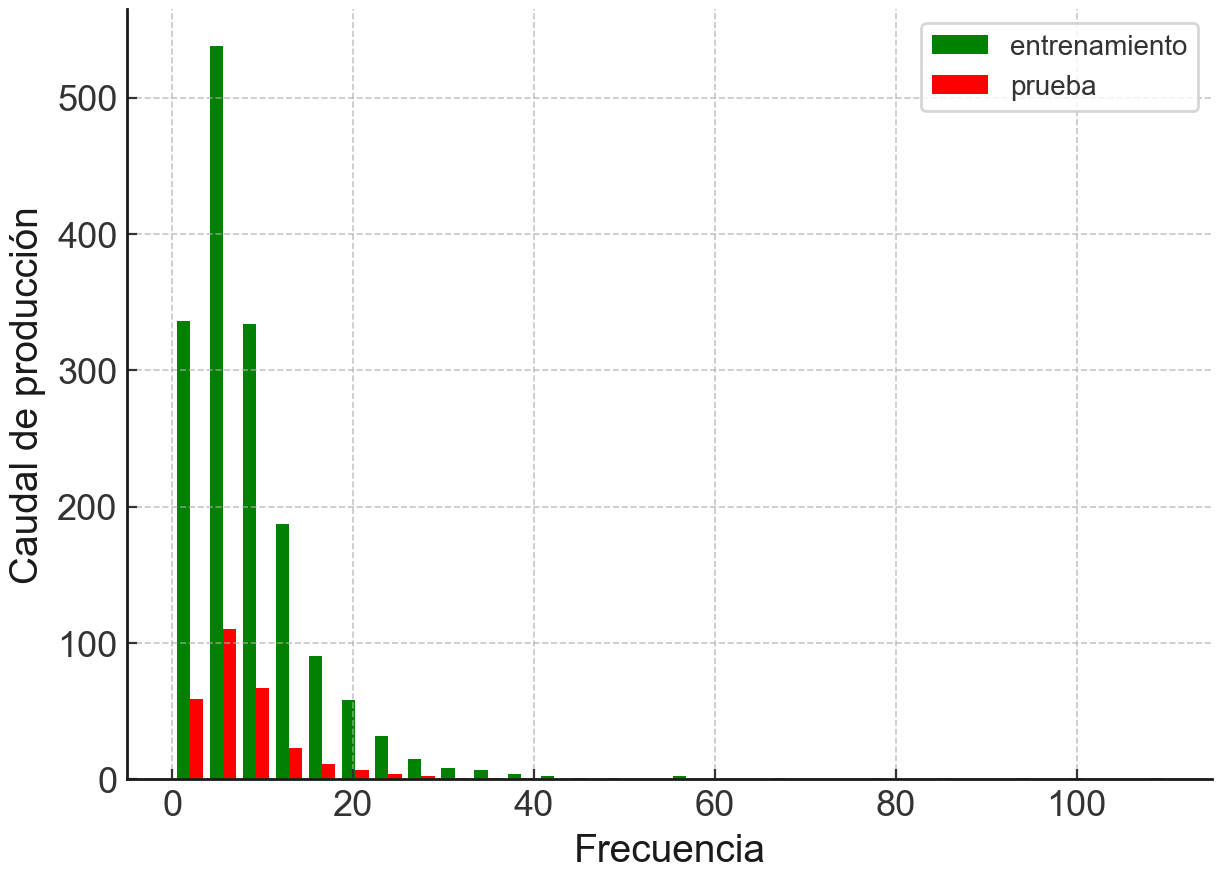
Luego de definir el flujo que seguirán los datos iniciales en el pre-procesamiento, se establece un pipeline que deben seguir los datos de pozos nuevos que usen los modelos para la predicción del flujo. Como resultado del pre procesamiento tenemos 1911 datos, los cuales usamos el 85% para entrenar el modelo y el 15% para comprobar su efectividad con data no vista.

Con el objetivo de que el caudal de los datos de entrenamiento y prueba tengan la misma distribución, los graficamos y evaluamos visualmente. Esto se realiza debido a que el modelo aprende de la data de entrenamiento, quiere decir que captura el comportamiento de la variable objetivo, entonces cuando evaluamos el desempeño de nuestra variable objetivo en la porción de data llamada de prueba tendrá un peor rendimiento si la distribución de la variable objetivo en la data de entrenamiento es diferente a la de prueba. Esto equivaldría a decir que hubo un mal aprendizaje por parte del algoritmo debido a la data de entrada.

Cabe mencionar que basta con que ambas distribuciones (variable objetivo de entrenamiento y de prueba) tengan similar asimetría.

A continuación, mostramos un gráfico que compara ambas distribuciones del caudal de producción en la porción de entrenamiento y de prueba, con esto corroboramos que la división se hizo correctamente.

1. Distribución del caudal en la data de entrenamiento y de prueba



Fuente: Propia

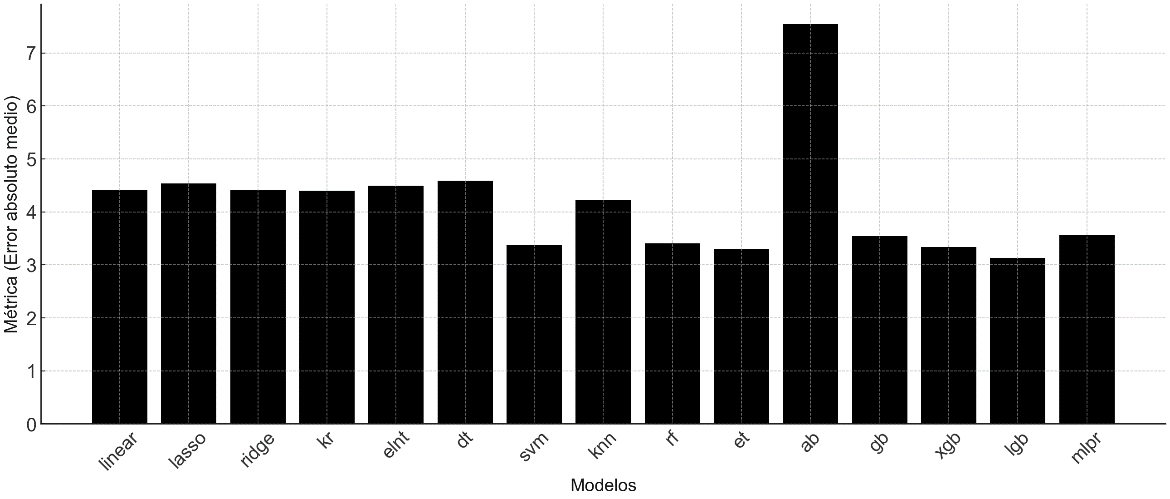
Para el modelado corrimos la mayor cantidad de algoritmos de regresión con el objetivo de comparar el desempeño, con la data de entrenamiento y prueba, de cada una de ellas. Realizamos esto debido a que no sabemos con exactitud que algoritmo tendría menor error en predecir el caudal de producción.

A continuación, listaremos todos los algoritmos que se usaran para el modelado, todos ellos en su versión de regresor, ya que el caudal de producción es una variable continua.

* Regresión lineal
* Lasso
* Ridge
* Kernel Ridge
* Red elástica
* Árbol de decisión
* Regresión de soporte vectorial (SVR)
* Vecinos cercanos
* Bosques aleatorios
* ExtraTreesRegressor
* AdaBoost
* GradientBoosting
* XGB
* LGBM

El valor numérico que usamos para la comparación es el error absoluto medio (MAE), el cual es una métrica para evaluar el desempeño de los algoritmos de aprendizaje automático supervisado, y se medirá con las mismas unidades del caudal (barriles por día). Esta métrica fue elegida por su fácil integración con una evaluación económica y ya que al estar en la misma escala que la variable objetivo, nos da una aproximación realista. A continuación, mostramos el resultado de correr los algoritmos antes mencionados en la data de entrenamiento.

1. Rendimiento de los modelos con sus parámetros por defecto.



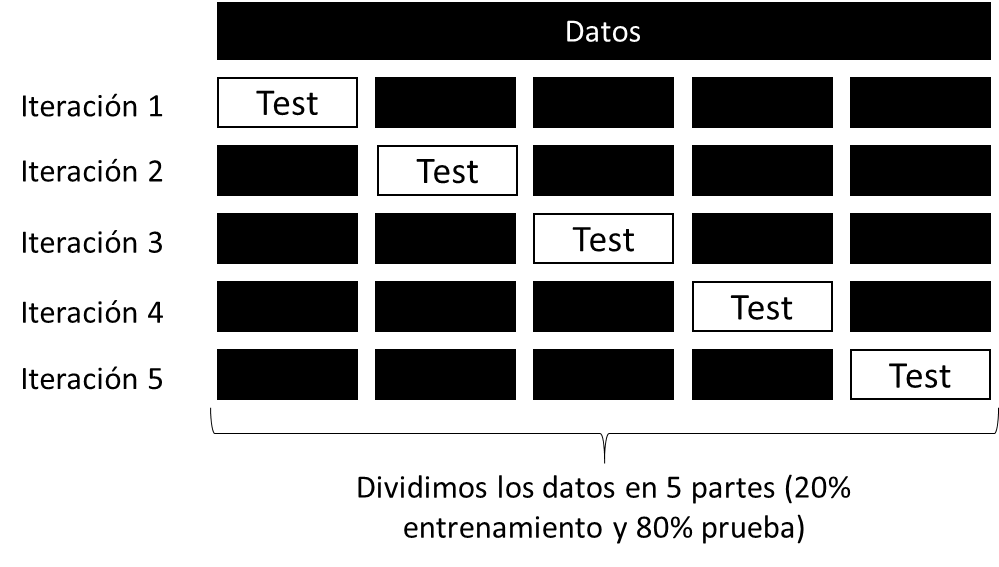
Fuente: Propia

Observamos que la mayoría de los algoritmos superan el 3.5 de error absoluto medio. Ello se debe a que los modelos fueron ajustados usando sus parámetros iniciales definidas por defecto.

No se obtuvo el resultado debido a que un modelo no puede predecir observaciones futuras que tengan como variables de entrada data nunca antes vista por el modelo. Por lo tanto, para evaluar mejor al modelo usamos estrategias de validación basadas en el re muestreo o también llamada validación cruzada.

La validación cruzada busca cumplir el objetivo del modelo, que es predecir en data no vista en la etapa de entrenamiento, y para lograrlo no hará uso de la data de prueba, ya que esta será la que al final decidirá el modelo más óptimo. A continuación, se muestra una imagen de cómo trabaja la validación cruzada.

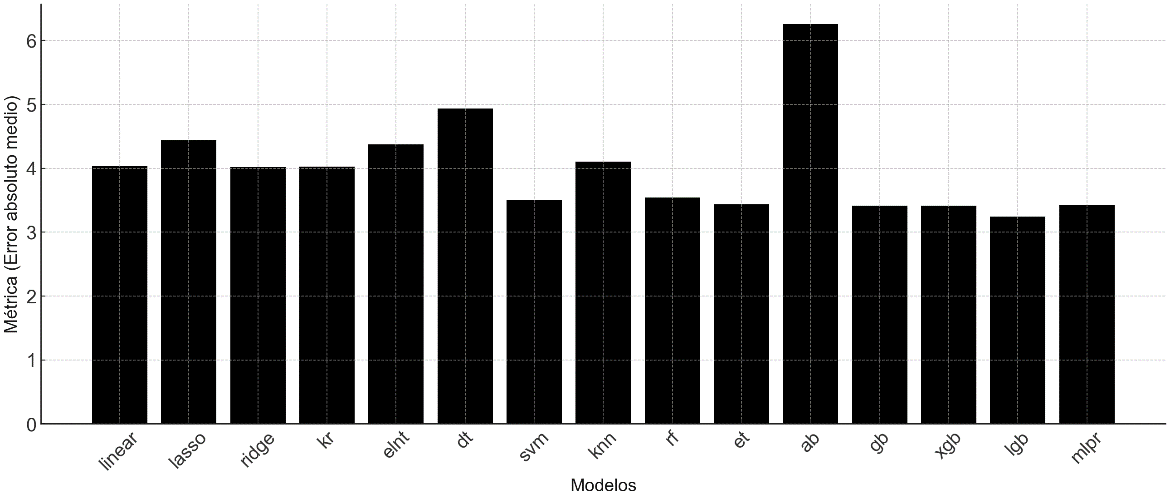
1. Ilustra la forma en cómo funciona la validación cruzada



Fuente: Propia

Para ello usamos funciones de la librería sklearn, el cual nos proporciona una en particular llamada *cross\_val\_score*. En esta función podemos especificar el modelo que usaremos, la data de entrenamiento, la métrica que se usará para evaluar al modelo y finalmente la división con la cual se hará la validación cruzada.

1. Rendimiento de los modelos con sus parámetros y usando la validación cruzada



Fuente: Propia

Comparamos el error absoluto medio obtenido de cada modelo en la data de prueba y observamos que varía de 4 a 7, también se puede ver que 5 de los algoritmos superan el 3.5 de error absoluto medio, mientras que los algoritmos basados en boosting apenas llega al 2.3. Esto quiere decir que los algoritmos que corrigen el error a través del incremento de pesos tienen un mejor rendimiento, en este caso la validación cruzada compensa el efecto de sobreajuste que caracterizan a este tipo de algoritmos, dándole más robustez.

De esta manera conseguimos un modelo entrenado de manera más robusta, pero aún puede serlo más, debido a que no hemos ajustado los parámetros que caracterizan a cada modelo, también llamados hiperparametros, cada uno tiene un valor por defecto y controla coeficientes, penalización o algún otro aspecto de la matemática del modelo. Y lo podemos controlar pasando estos valores como argumento a los algoritmos.

Se decidió usar un amplio rango de valores para la optimización, esto genera que el algoritmo de optimización demore en entrenar el modelo siendo conveniente usar una optimización aleatoria, pero el tiempo de computación al usar la optimización con cada combinación no pasa de los 10 minutos debido a que nuestro set de datos es pequeño, por lo tanto, se decidió usar este último.

A este algoritmo se le agregará la validación cruzada antes usada, solo que esta vez el modelo será más robusto con sus predicciones.

## Análisis de resultados

De la recolección de datos se obtuvo 1921 datos provenientes de intervalos producidos, con sus respectivos parámetros, algunos de estos generados sintéticamente al no tener la posibilidad de haber realizado pruebas específicas. Se realizó el proceso de limpieza de datos de acuerdo a la metodología, eliminando datos con dos o más campos ausentes, estos fueron divididos en 85% data de entrenamiento y 15% de prueba, el porcentaje correspondiente a la data de validación fue incluido con la validación cruzada realizada de por medio.

1. Porcentaje general de los datos usados en entrenamiento y prueba

Fuente: Propia

Antes de usar el modelo se realizó antes la normalización de algunos campos, esto mejoró el tiempo de demora de optimización de los modelos.

Finalmente, el algoritmo de mejor desempeño fue el Gradient Boosting, éste mismo fue usado para las comparaciones a nivel económico. Obteniendo una mejor aproximación al método tradicional.

Esta metodología puede ser encapsulada fácilmente en una función que se le puede correr a una base de datos bien estructurada y siempre servirá de ayuda para la evaluación económica que ya está sujeta a probabilidades en su determinación.

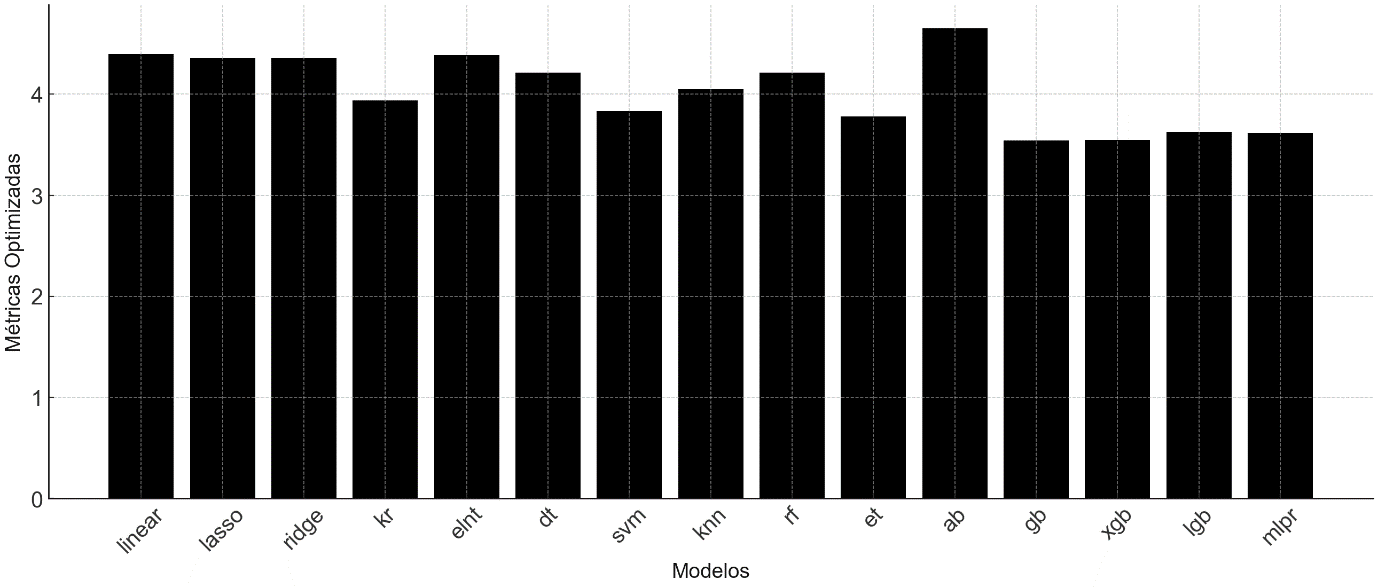
# CAPÍTULO IV: ANÁLISIS Y DISCUSIÓN DE RESULTADOS

Resultado de correr modelos con las todas las combinaciones posibles, obtenemos el error absoluto medio mínimo, extraemos los parámetros responsables de ese rendimiento y los guardamos para corroborar el desempeño del algoritmo en la data de prueba y realizar la comparación entre el caudal obtenido de la ecuación de flujo, el real y el obtenido por nuestro mejor modelo.

1. Muestra los algoritmos corridos con sus principales características.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Modelo | Validación cruzada |  | Combinaciones | iteraciones | Error absoluto medio |
| Regresión Lineal | 10 |  | 2 | 20 | 4.39 |
| Lasso | 10 |  | 24 | 240 | 4.35 |
| Ridge | 10 |  | 15 | 150 | 4.36 |
| Kernel | 10 |  | 48 | 480 | 3.94 |
| Red Elástica | 10 |  | 16 | 160 | 4.39 |
| Vector de soporte | 10 |  | 60 | 600 | 3.83 |
| KNN | 10 |  | 768 | 7680 | 4.05 |
| Arboles aleatorios | 10 |  | 36 | 360 | 4.21 |
| Gradient Boosting | 10 |  | 864 | 8640 | 3.54 |
| LGB | 10 |  | 9 | 90 | 3.63 |
| XGBoost | 10 |  | 9 | 90 | 3.55 |
| Adaboost | 10 |  | 30 | 300 | 4.65 |
| MLPRegressor | 10 |  | 10 | 100 | 3.62 |

1. Rendimiento de los modelos con sus parámetros optimizados y usando validación cruzada

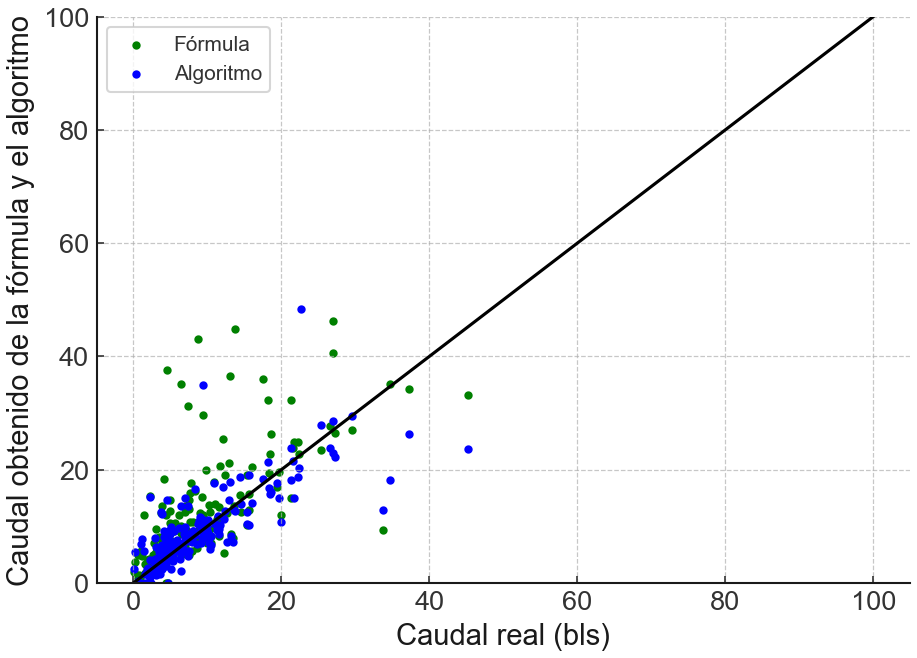


Fuente: Propia

A continuación, se graficaron las curvas de aprendizaje de cada modelo que fue usado en los anexos (ver Anexo A).

Graficamos el caudal real versus el caudal calculado por el algoritmo y versus el causal obtenido de la ecuación de flujo, esto lo realizamos usando la data de prueba (288 datos).

1. Ilustra el método que más se aproxima a la línea recta negra que representa al caudal real.



Fuente: Propia

Observamos que entre más los puntos se aproximen a la línea roja, más exacta será la aproximación realizada por este método. Por ende, se observa que la aproximación del algoritmo es más exacta que la realizada por la ecuación de flujo. Esto se evidencia al obtener el coeficiente de determinación respecto al cálculo del algoritmo es 0.500, mientras que respecto al cálculo por fórmula es 0.327.

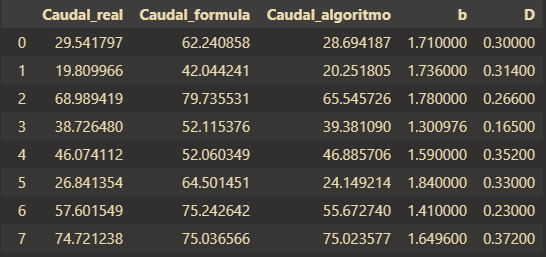
# CAPÍTULO V: ANÁLISIS ECONÓMICO

Para la evaluación económica debemos tener en cuenta que ésta se realiza por pozo y cada pozo posee en general de 3 a 6 formaciones abiertas y cada una corresponde a un dato de nuestra tabla, entonces se juntaran los datos correspondientes a un pozo y con el caudal agregado de estos se realizará la proyección de producción a través de la curvas de declinación junto con el flujo de caja para obtener el VAN y el TIR, estos parámetros serán los que compararemos para evidenciar el beneficio de usar algoritmos de machine learning como apoyo para este tipo de cálculos.

El código con la lógica de programación que se usó para este flujo de caja estará al final del presente trabajo.

A continuación, mostramos una de parte de la tabla que usaremos para la evaluación económica, cabe mencionar que los caudales están en barriles por día y el b y D son adimensionales.

1. Captura de la tabla con los datos necesarios para el pronóstico de producción

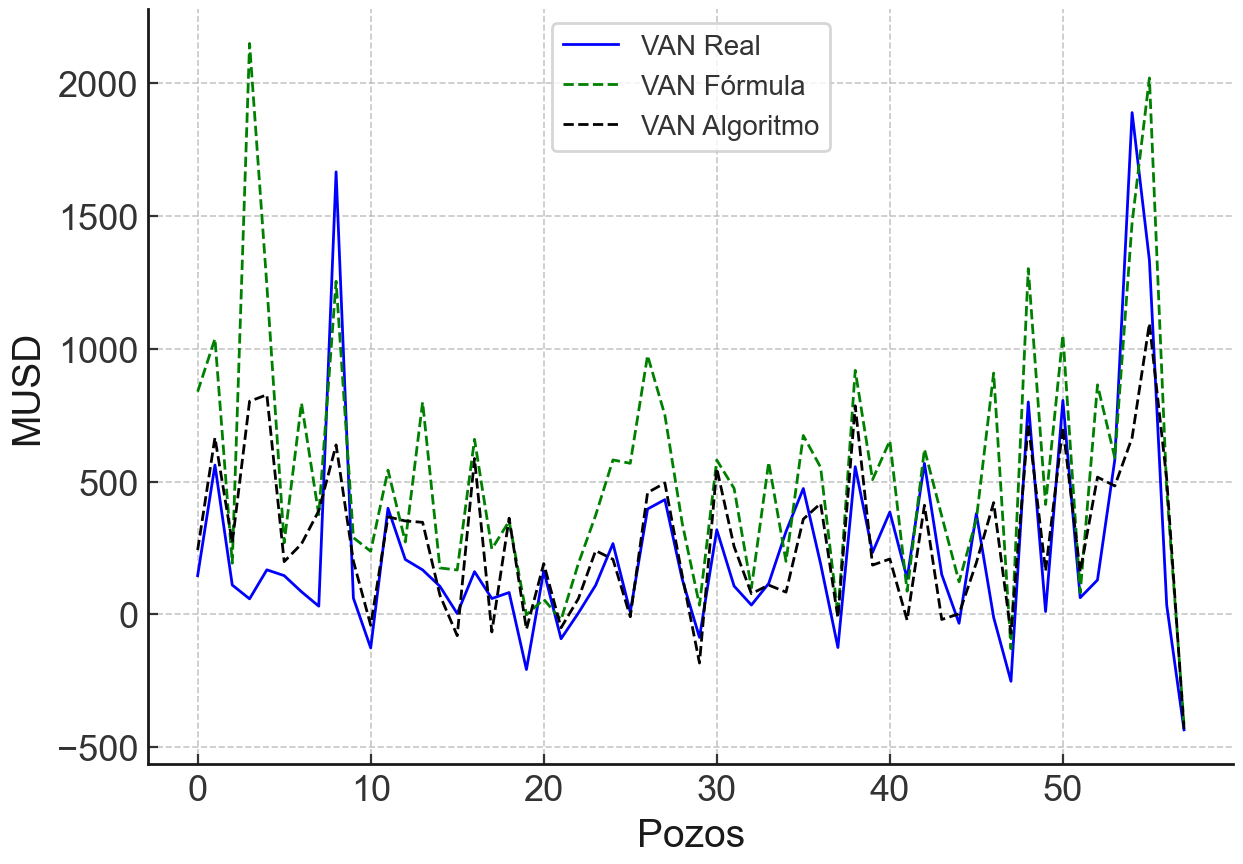


Fuente: Propia

Para la evaluación necesitamos la curva de producción pronosticada del pozo. Para este cálculo usamos el caudal inicial de producción y los parámetros de declinación con la fórmula de Arps, de esta manera podemos realizar el flujo de caja.

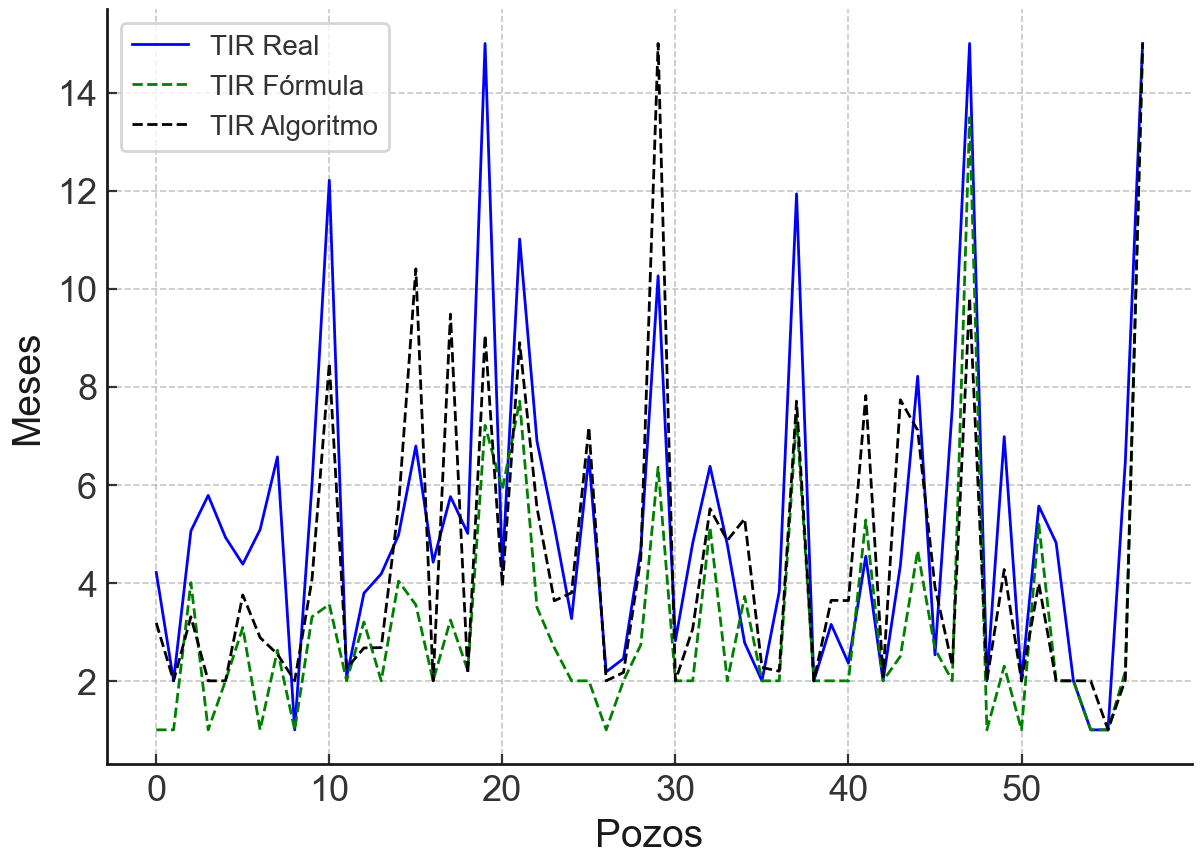
El precio del barril usado será de 60 dólares, así como la predicción de producción será a 15 años. La regalía por la producción de petróleo será de 7%. La comparación será usando el VAN de cada pozo y el TIR respectivo.

1. Comparación del VAN de cada pozo.



Fuente: Propia

1. Comparación del TIR de cada pozo.

**

Fuente: Propia

De los gráficos, observamos que los parámetros VAN y TIR que se acercan más a la realidad pertenecen al algoritmo. Usando el coeficiente de determinación de cada método para cada parámetro obtenemos.

|  |  |  |
| --- | --- | --- |
|  | ALGORITMO | FÓRMULA |
| VAN | 0.43740 | -0.35743 |
| TIR | 0.50538 | 0.28885 |

1. Los coeficientes de determinación que más se acercan a la unidad pertenecen al obtenido con el algoritmo.

*Nota.* Esta tabla muestra los coeficientes de determinación, donde se muestran valores negativos que indican un peor rendimiento de predicción.

# CAPÍTULO VI: CONCLUSIONES Y RECOMENDACIONES

* Se concluye que el uso de algoritmos de aprendizaje automatizado para predecir el caudal de producción de un pozo usando la misma información necesaria para el uso de la ecuación de flujo, obtiene resultados mucho más aproximados a los valores reales que se obtiene luego de producir la zona propuesta, ya que el error absoluto medio del algoritmo con mejor desempeño es de 3.54, mientras que el error de la ecuación de flujo es de 4.65.
* Para el uso de los algoritmos de aprendizaje automatizado basta con generar un script de Python que use una planilla de entrada (.xlsx, .sql, .csv, etc) con los atributos que requiera para la predicción, como dato de entrada, y este generara un archivo en el formato que desee con las predicciones respectivas.
* La implementación de estos algoritmos es a bajo costo, ya que hoy en día existen librerías de Python que implementan los modelos matemáticos complejos internamente y que fácilmente podemos usar en nuestro camino de obtener el modelo que mejor se ajuste.
* Es necesario resaltar que los usos de los resultados del algoritmo deben ser de apoyo al análisis principal usado en la organización, ya que estos son modelos matemáticos–estadísticos y están sujetos a errores producidos por la falta de información en variables claves como la probabilidad de encontrar fallas.
* A pesar de que la mejora en la aproximación al caudal real de producción no fue tan significativa, cuando las predicciones de varios pozos se juntan, notamos un incremento en la efectividad al predecir la rentabilidad del pozo.
* Finalmente, con las predicciones obtenidas de la ecuación de flujo y del algoritmo se realizó un análisis económico usando la producción total de los pozos en el tiempo, hasta los 30 años de producción y se obtuvo resultados más aproximados de VAN y TIR con el algoritmo.

# CAPÍTULO VII: REFERENCIAS BIBLIOGRÁFICAS (En Formato de Normas APA, 7ma edición)

David P. Craig and Michael J. Eberhard, Robert D. Barree. (2000) *Adapting High Permeability Leakoff Analysis to Low Permeability Sands for Estimating Reservoir Engineering Parameters*

Joaquín Amat Rodrigo. (2020) Machine learning con Python y Scikit-learn. [*https://www.cienciadedatos.net/*](https://www.cienciadedatos.net/)

José Luis Rivero S. (2007) Libro de Petróleo

McCain, W.D., (1990) *The Properties of Petroleum Fluids*. Segunda edición, EUA, PennWell Publishing Company

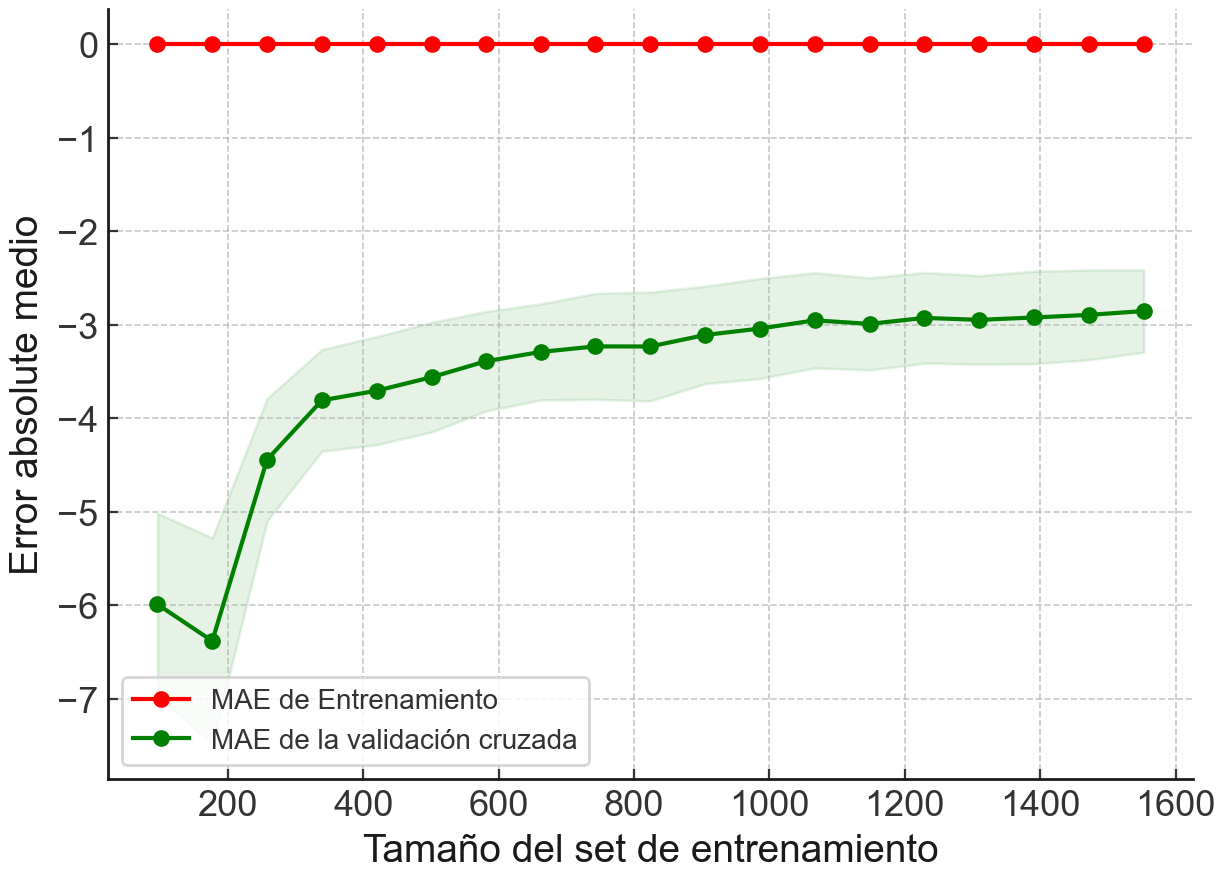
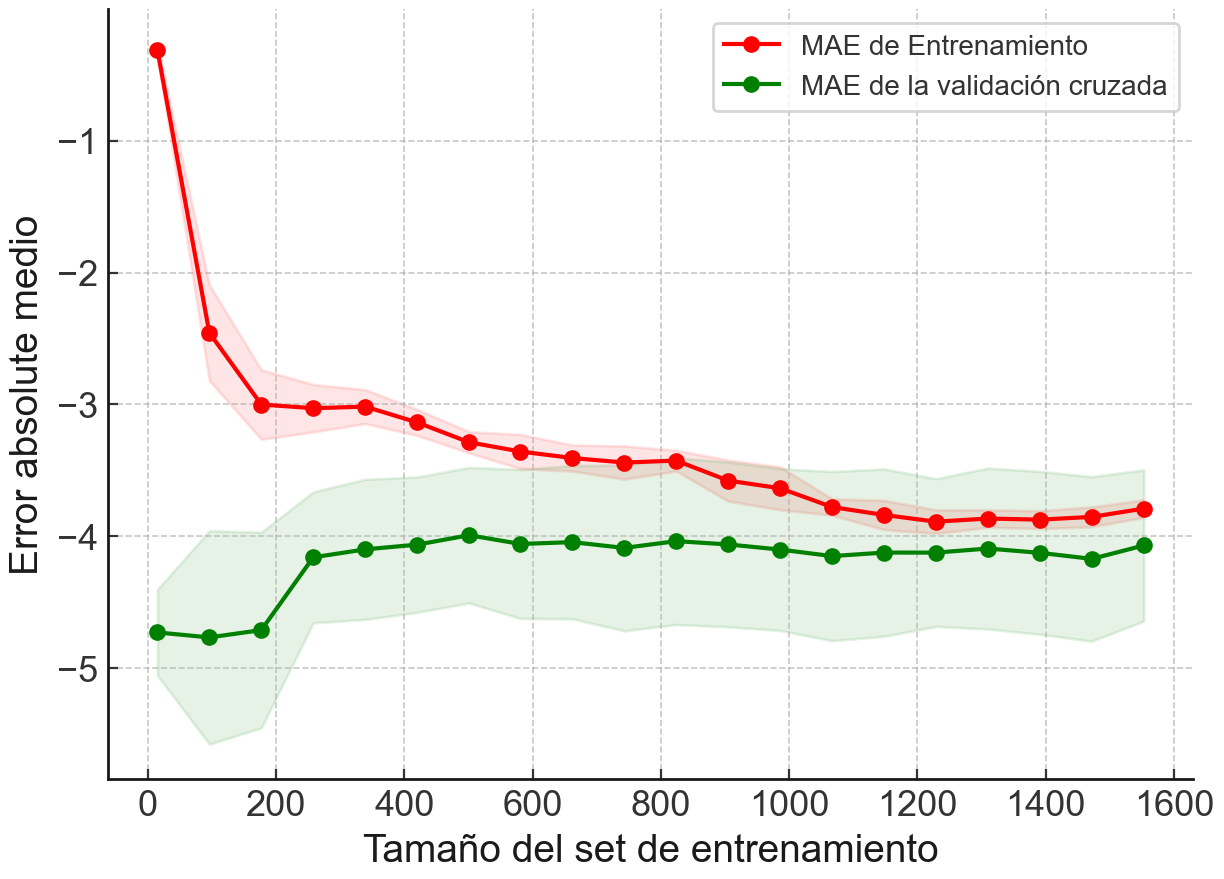
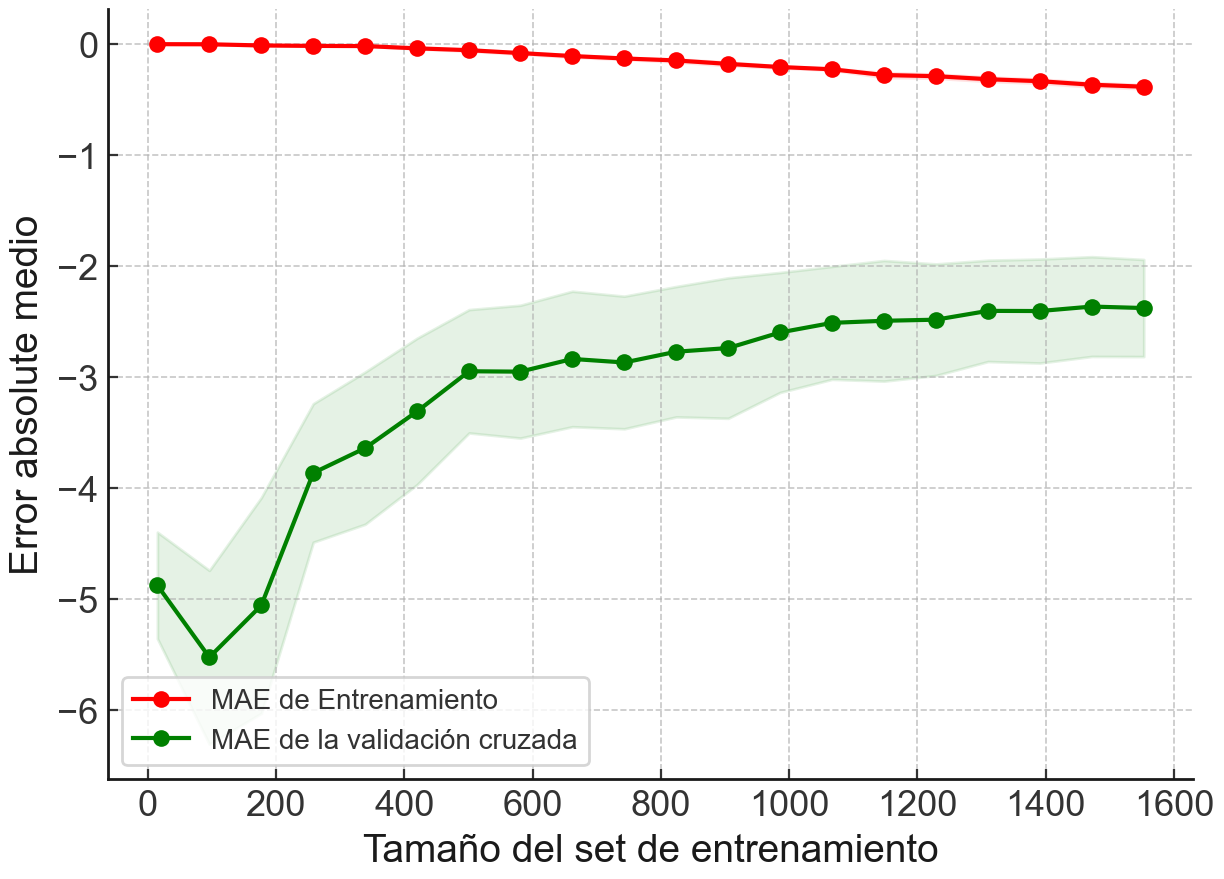
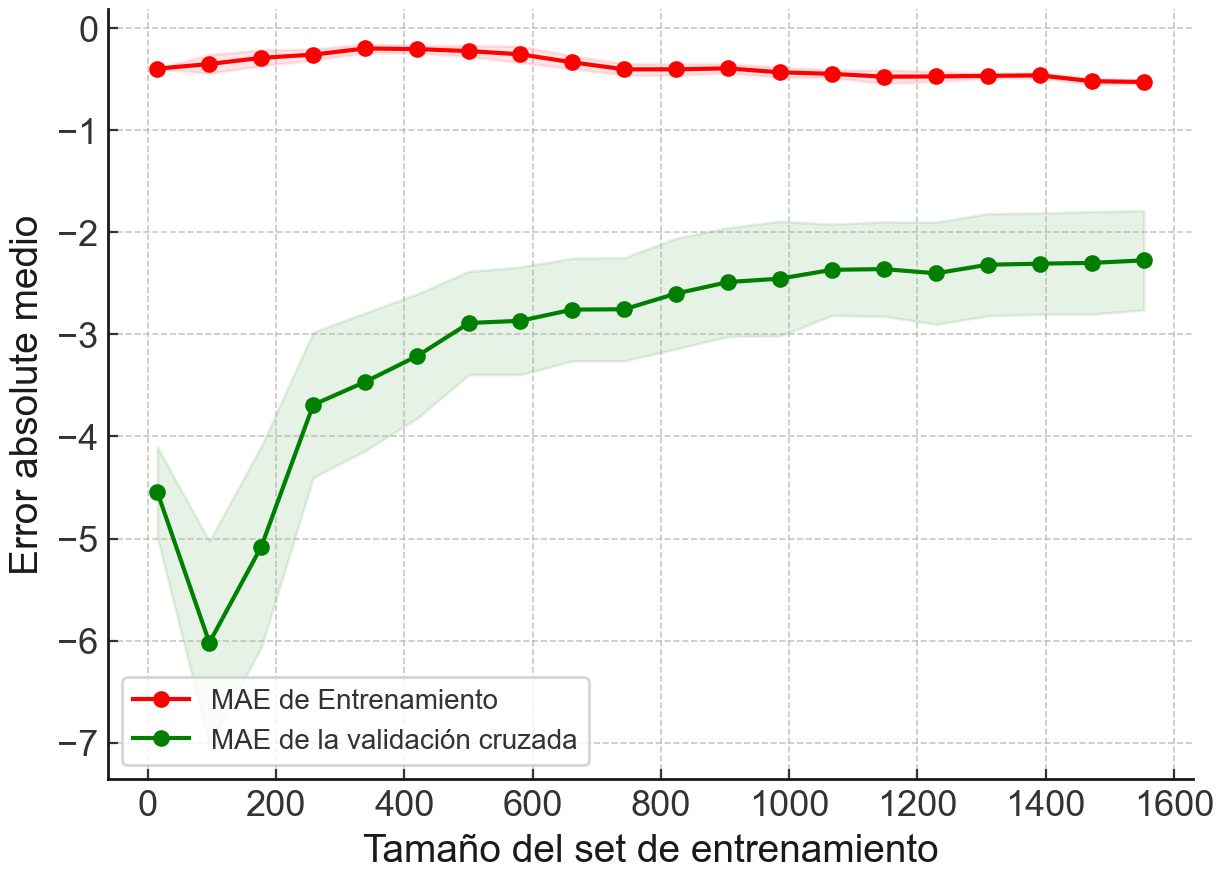
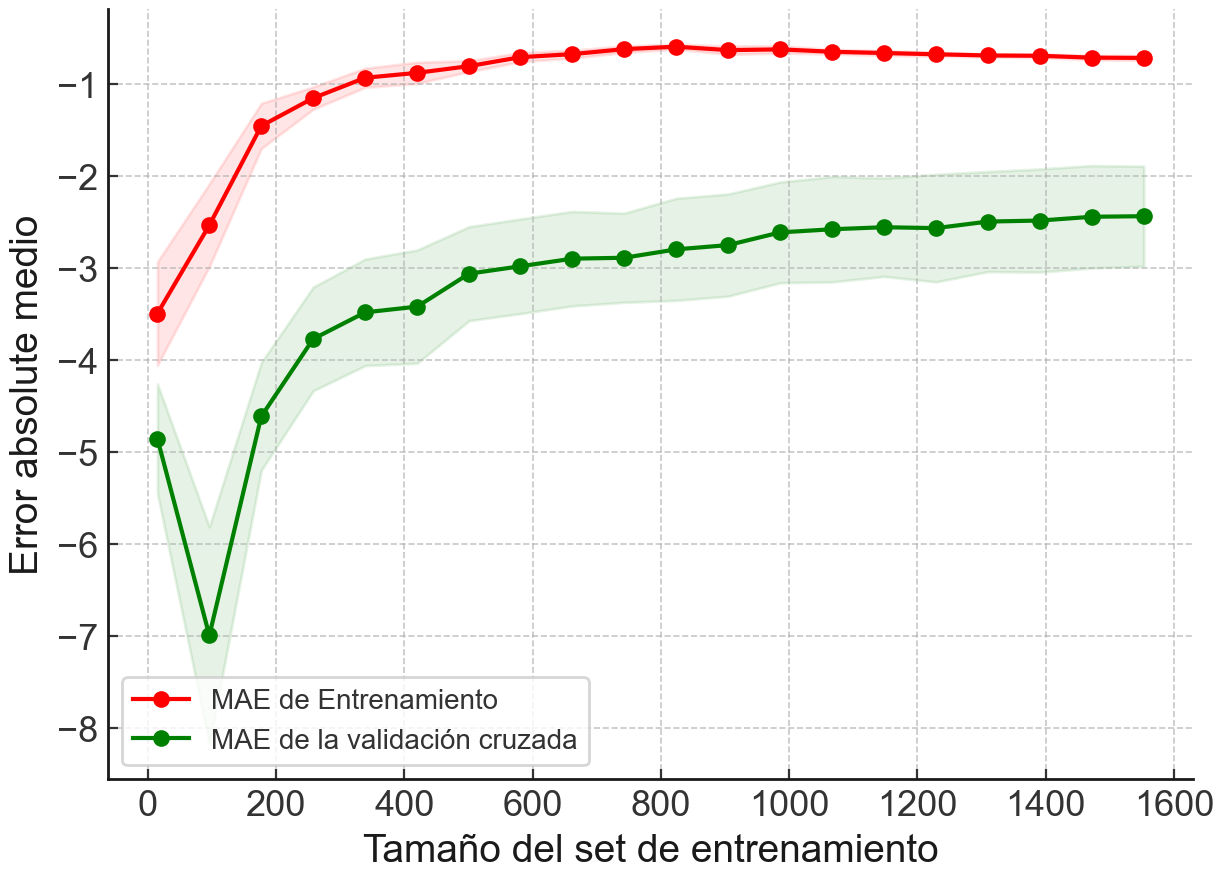
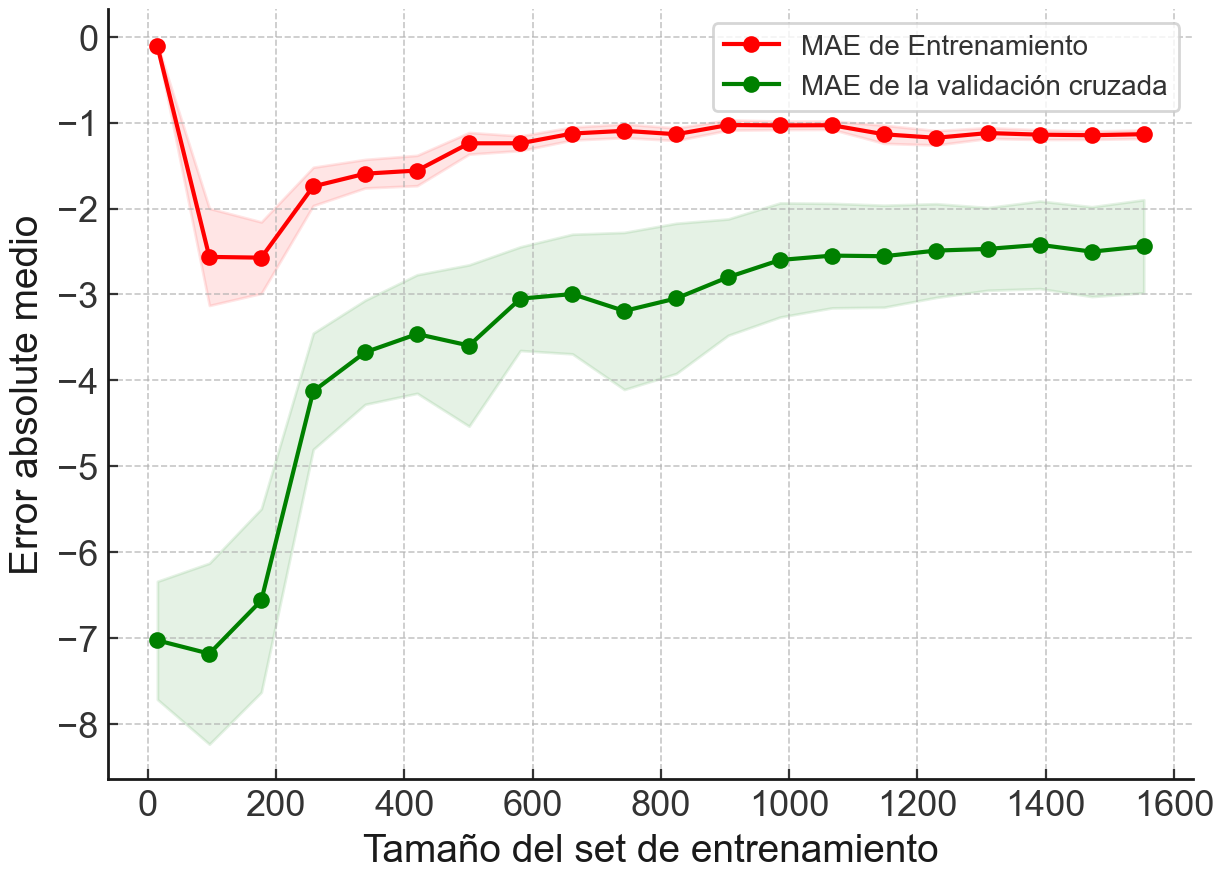
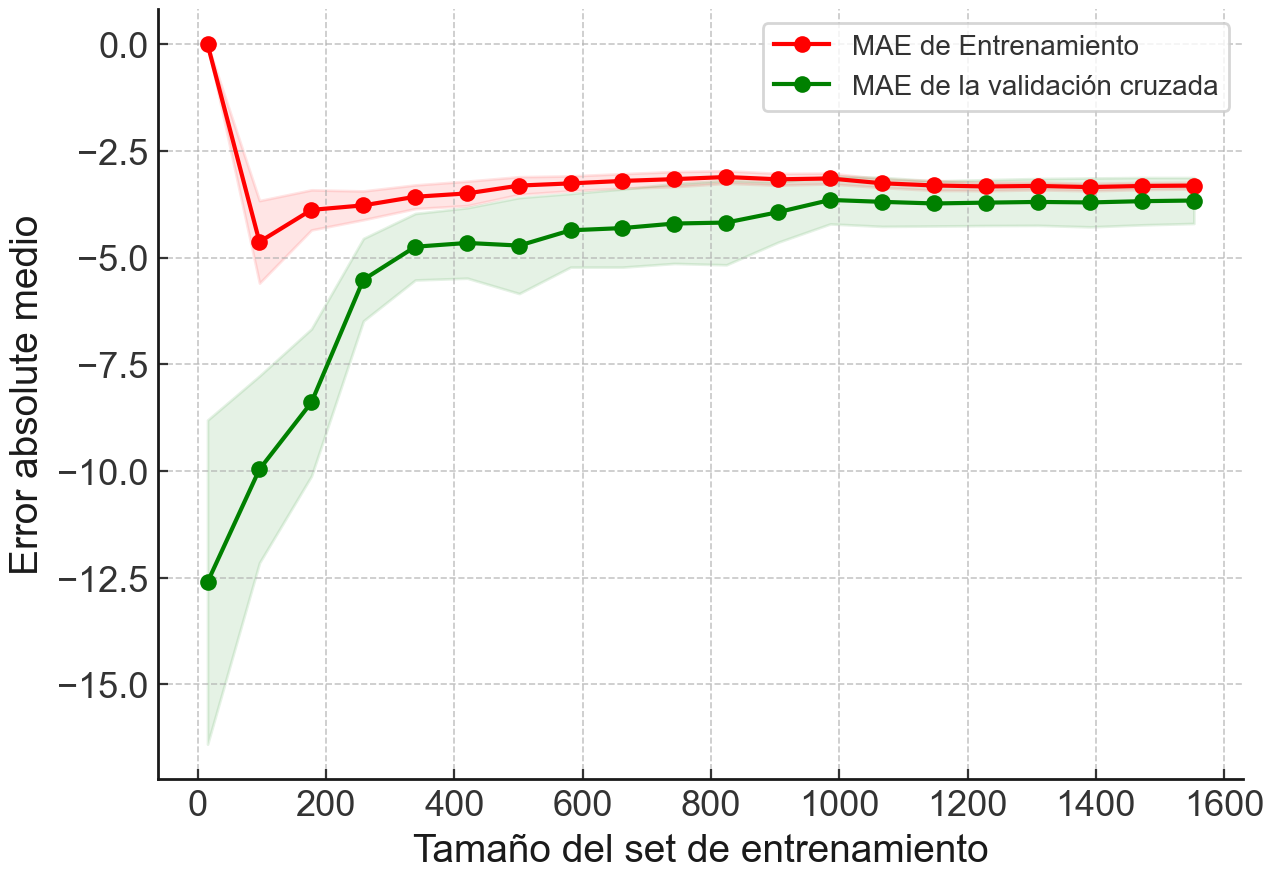
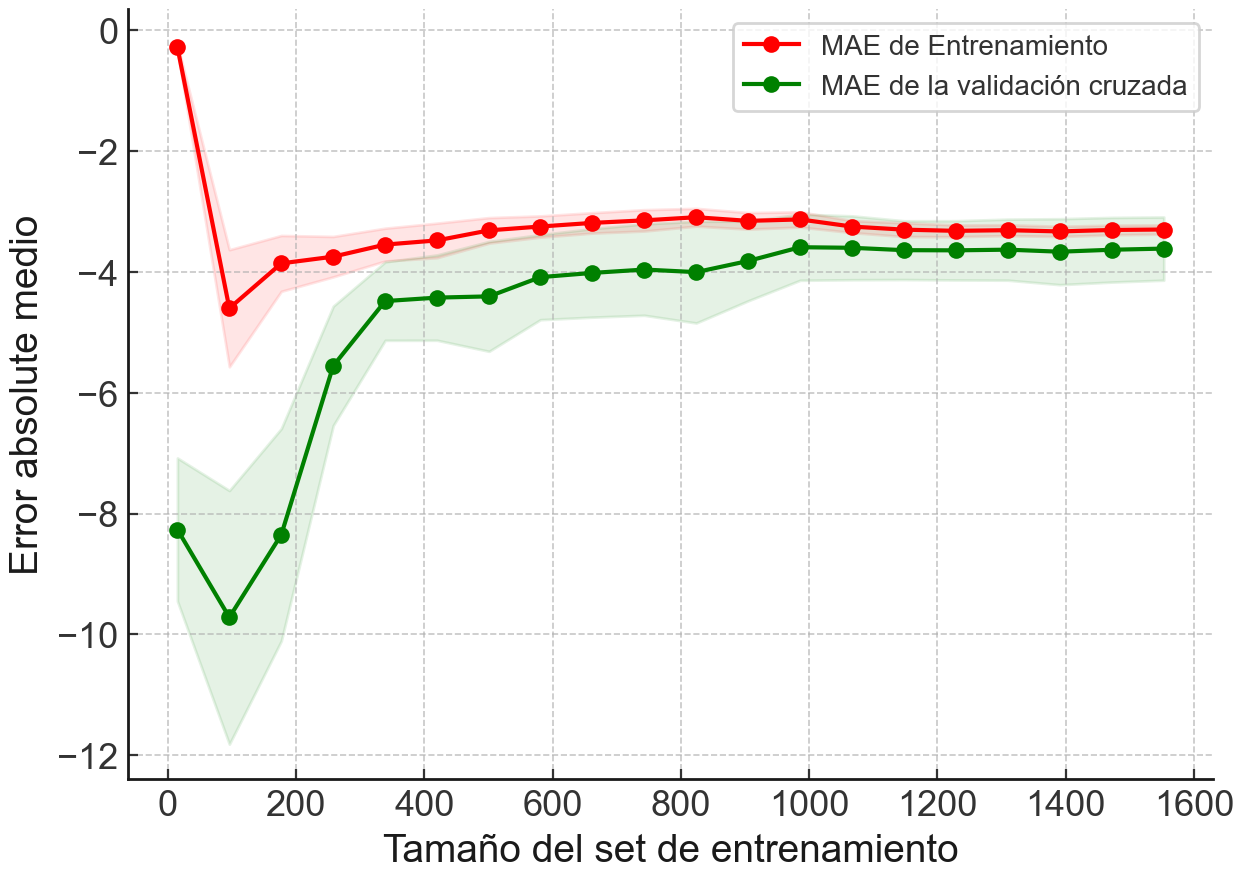
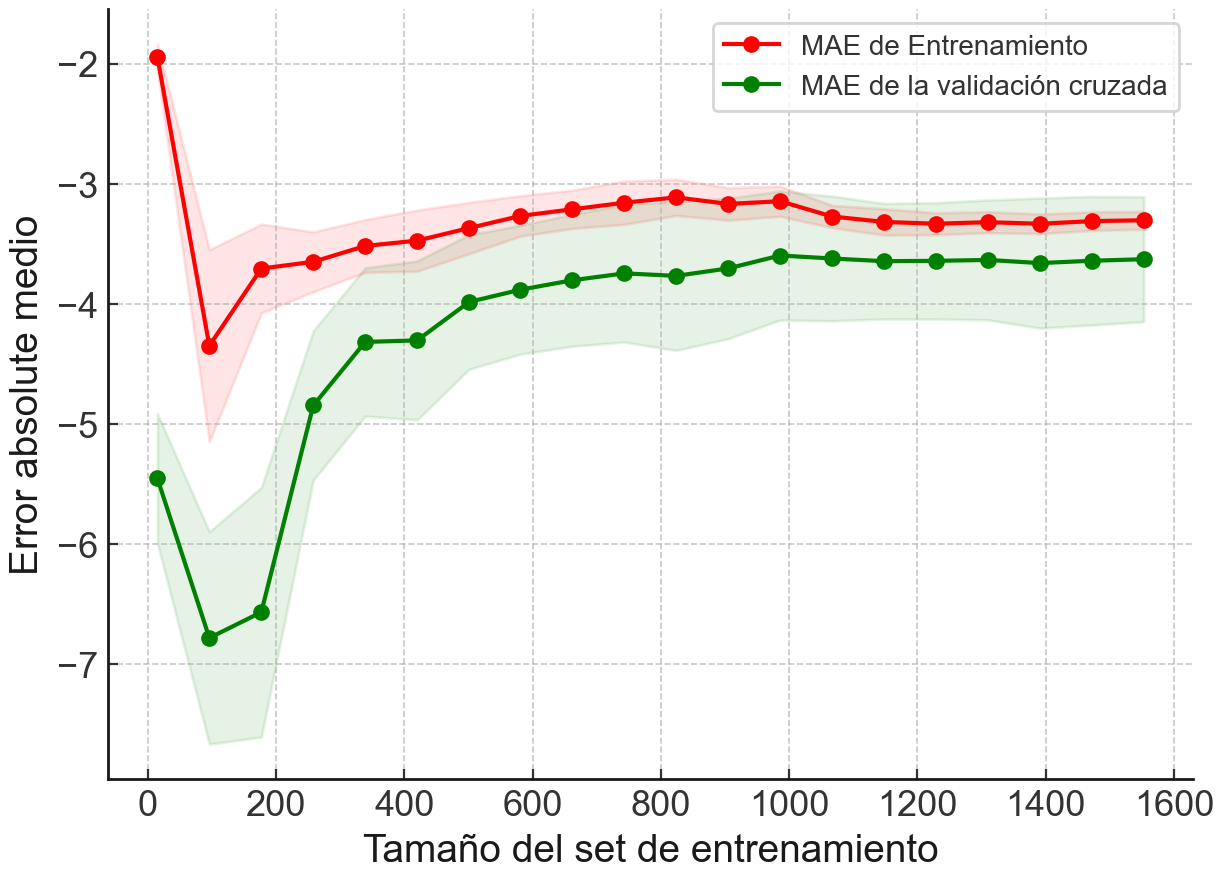
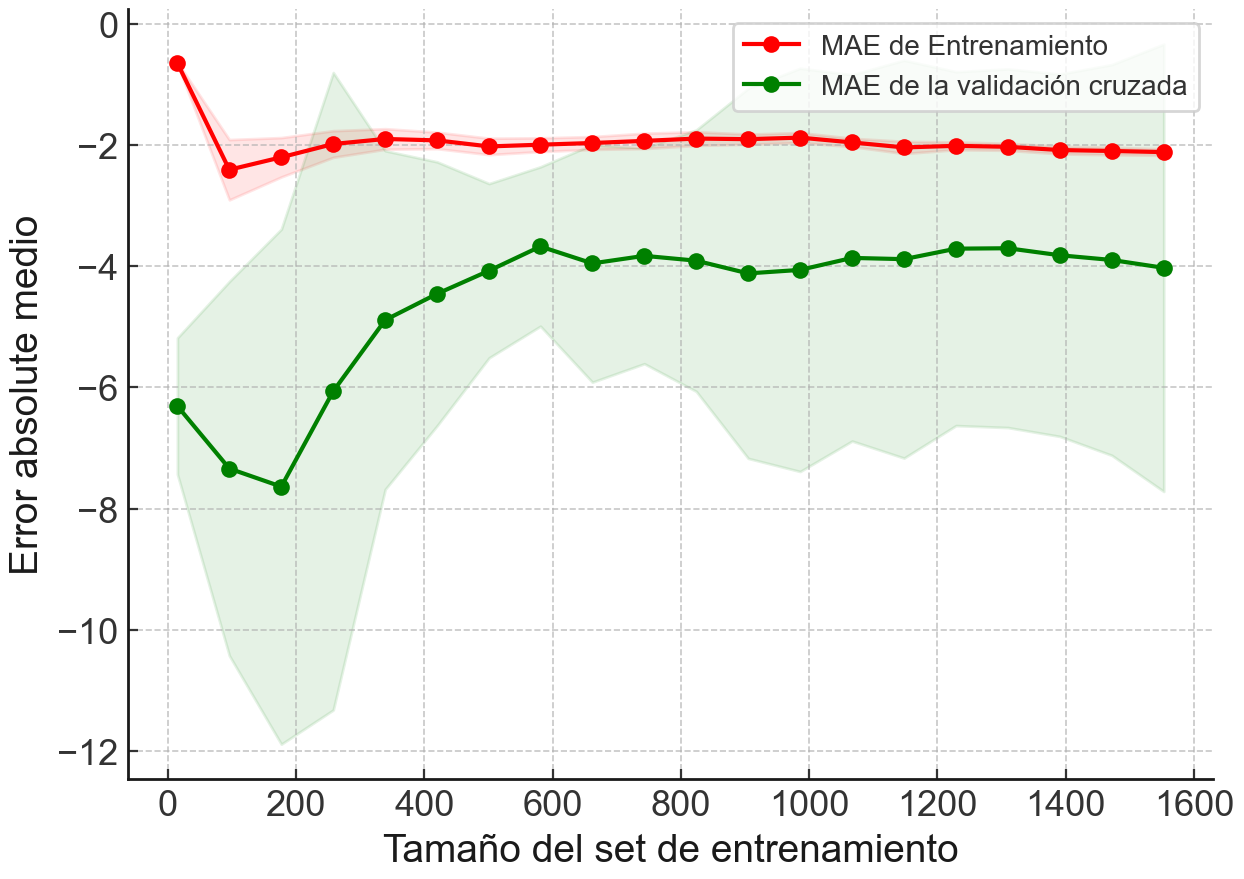
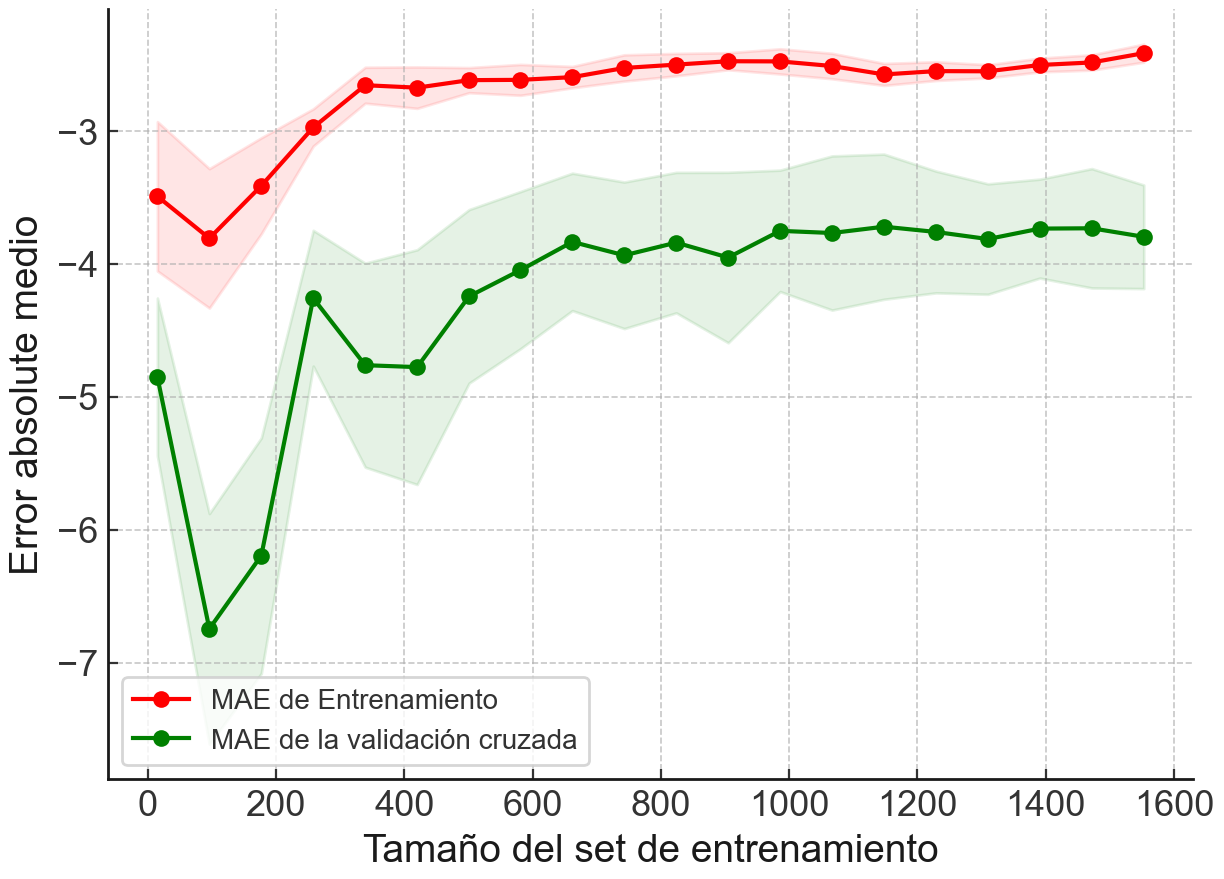
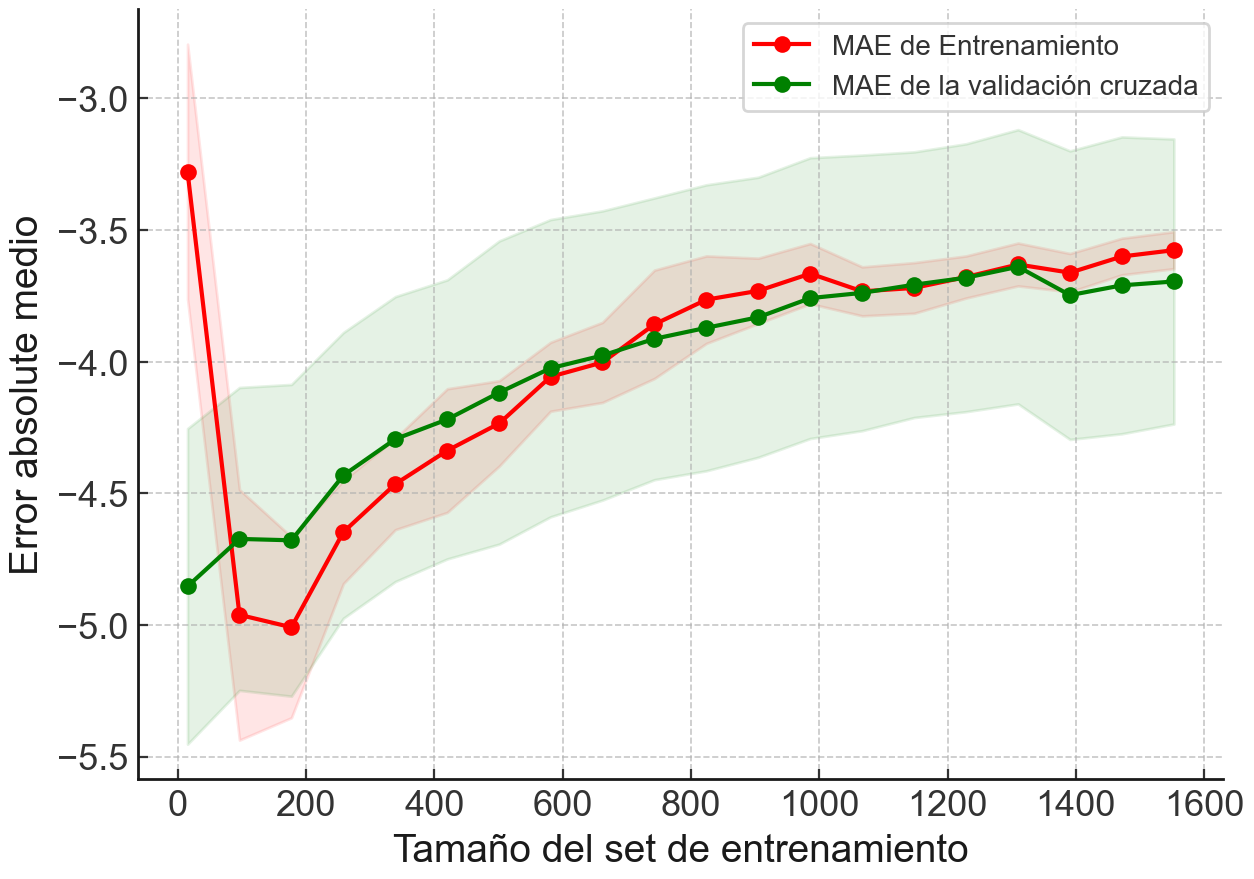
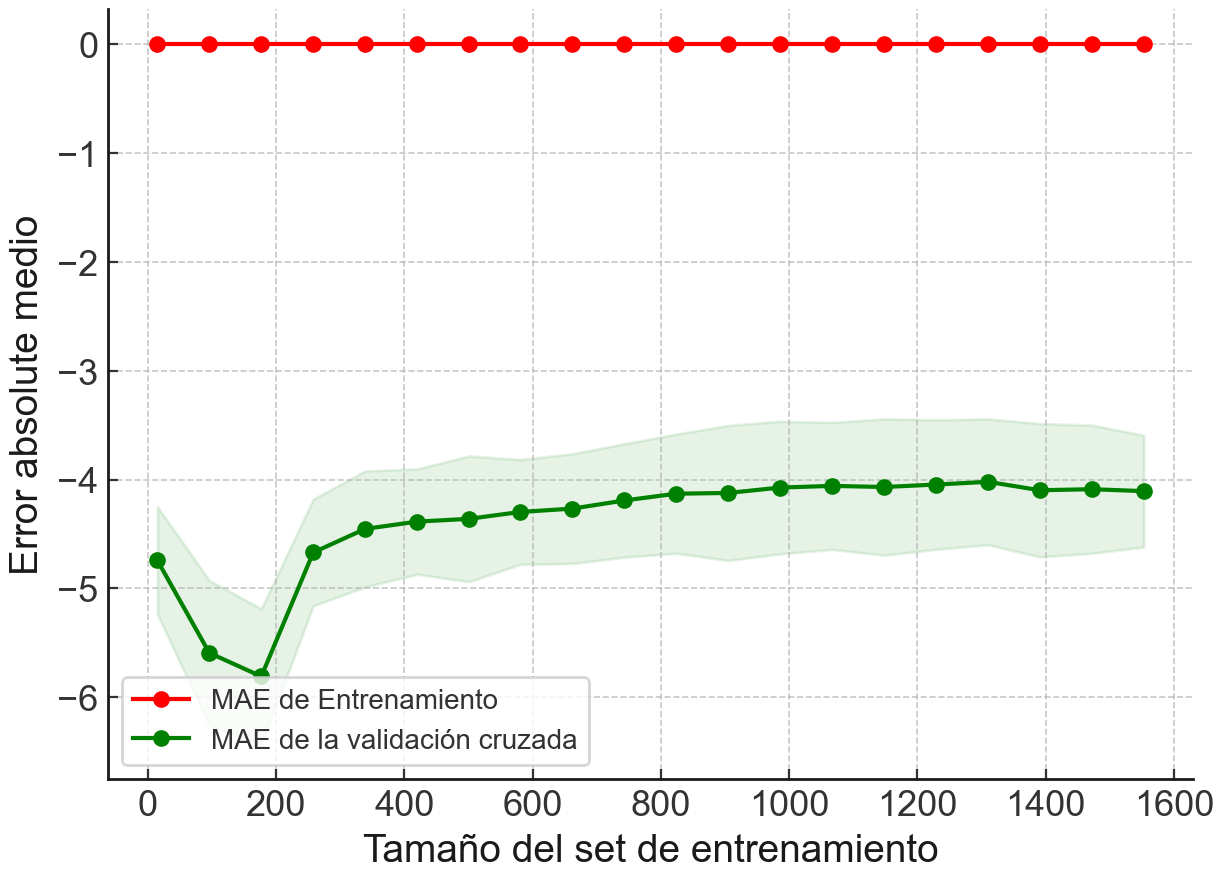
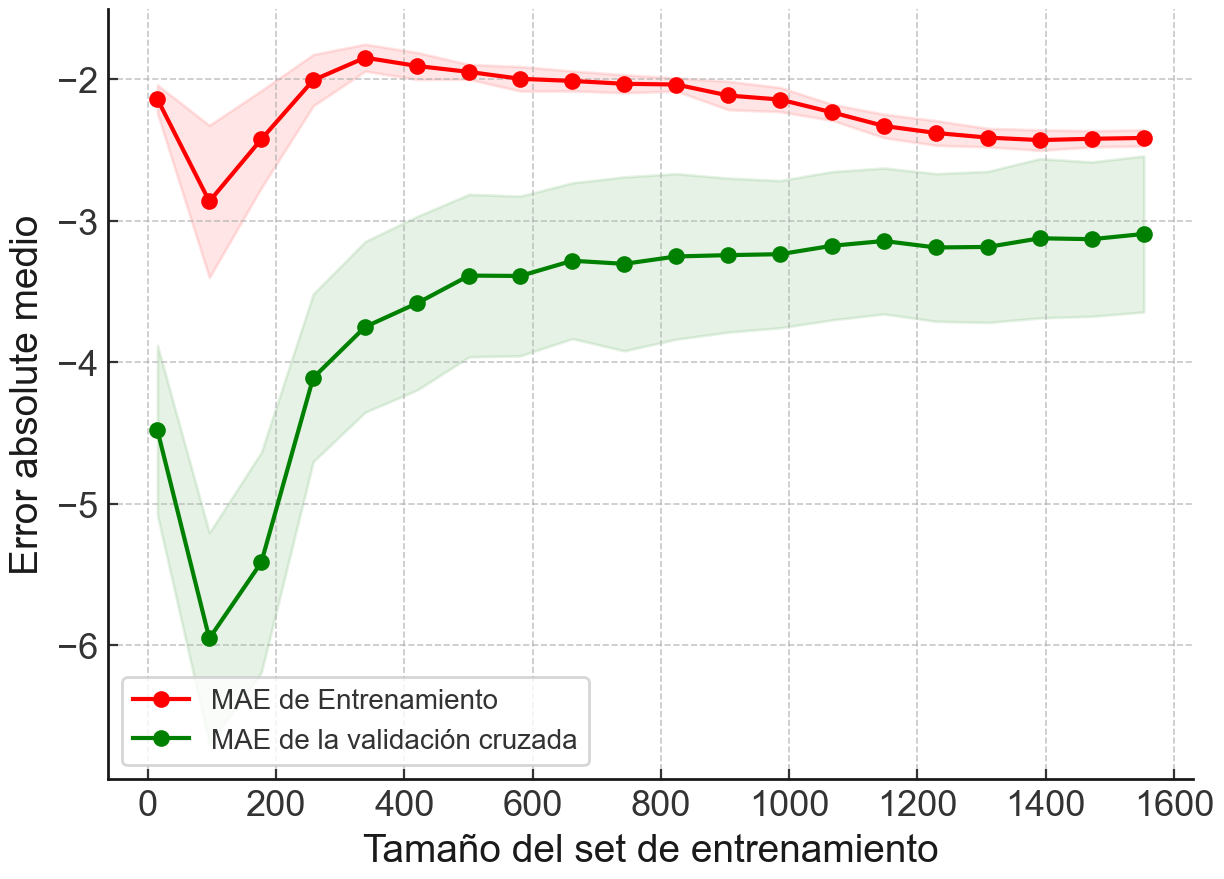
Tarek Ahmed. (2006) Reservoir Engineering. Tercera edición

Deepak Devegowda (2018) *Data Mining Course. The University of Oklahoma*

# ANEXOS

Anexo A.

}



Anexo B.

Fuente del código: https://github.com/PatrickTanta/tesis-pregrado