

Analisis de Data de Produccion del año 2019

Jonathan Avila - Juan Marcante

Indice de Contenido

- Generalidades del método de Producción y objetivo de la Investigación
- Fuentes de datos y variables
- Limpieza y exploración de data
- Análisis Univariado
- Análisis Multivariado
- Correlación entre variables
- Ajuste y entrenamiento del modelo de agrupamiento en base a un caudal de pozo esperado
- Ajuste y entrenamiento del modelo de agrupamiento en base a un corte de AyS de pozo esperado
- Ajuste y entrenamiento del modelo de estimación del caudal de pozo esperado
- Ajuste y entrenamiento del modelo de estimación del Ays de pozo esperado



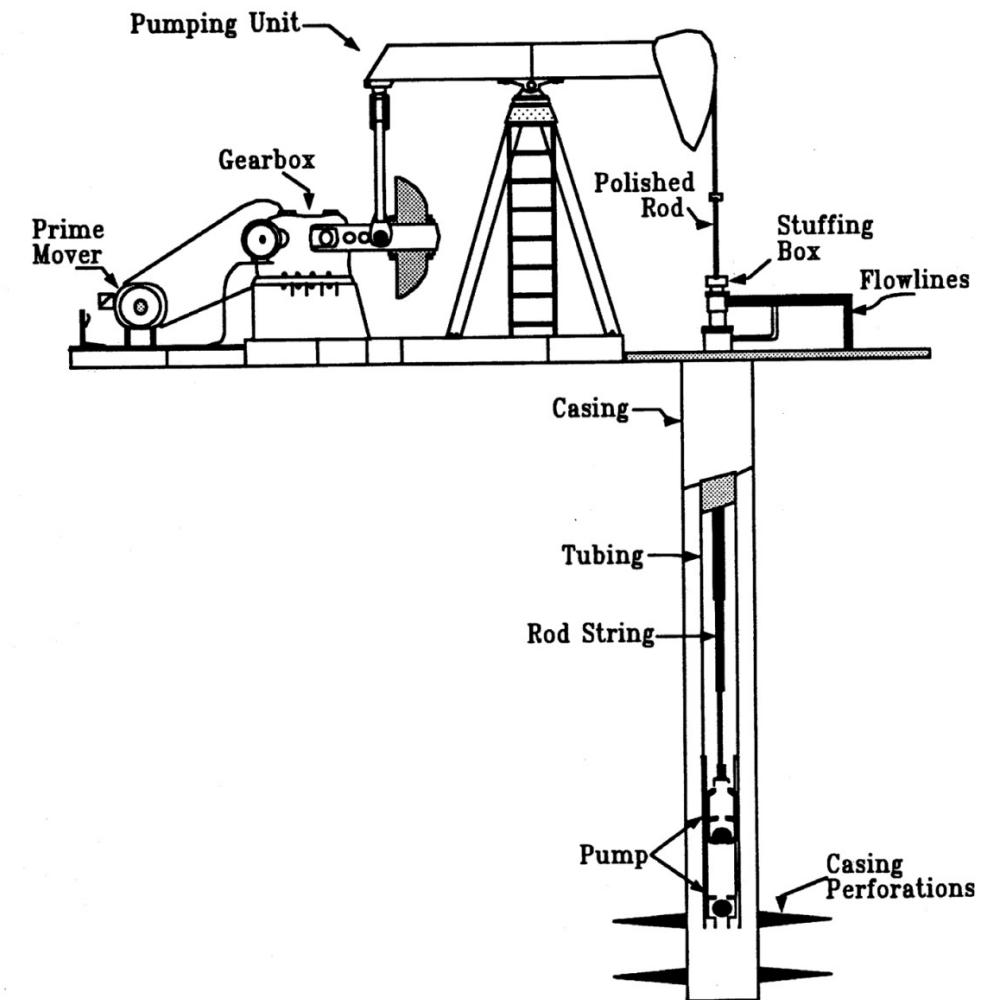
Generalidades del método de Producción y objetivo de la Investigación

Durante la vida productiva de un campo petrolero, los pozos petroleros normalmente se inician en una fase donde basta con conectar el subsuelo con superficie por medio de un pozo y el petróleo fluiría de forma natural, pero después de años de explotación de los yacimientos de subsuelo esto no siempre es posible y es cada vez mas escaso, es por eso que cuando la energía natural de un yacimiento se agota o baja, se hace necesaria la utilización de un Sistema de Levantamiento Artificial para elevar los fluidos a la superficie y la planificación de los sistemas de levantamiento es de suma importancia para alcanzar de manera eficiente y óptima la explotación de un yacimiento petrolero. Existen muchos tipos de levantamiento artificial, en esta investigación nos centraremos en un campo que tiene como método de levantamiento artificial el "Bombeo mecánico"

El Bombeo mecánico es el más común de los métodos de levantamiento artificial. Es el más antiguo y ampliamente usado método de levantamiento artificial costa adentro. Es usualmente el más económico y el sistema más fácil de mantener cuando es diseñado y operado apropiadamente.

Resulta critico en el diseño del sistema conocer el caudal bruto liquido que manejará el sistema así como el corte de agua esperado, es por eso que esta investigación tiene el siguiente objetivo:

¿Es posible estimar el caudal y corte de agua de un pozo nuevo para el campo?



Fuentes de datos y variables

Como fuente de datos de entrada se ha recopilado un dataset compuesto de 4494 registros, integrado por las pruebas de producción del año 2019 de los 100 pozos activos del campo, a cada pozo se le tomaron de 3 a 4 pruebas por pozo por mes, a estas pruebas se le agregaron los parámetros operativos relacionados al pozo durante la prueba.

En total el dataset cuenta con 17 columnas:

1	POZO	Identificador del pozo	texto
2	XCOORD	Coordenada relativa de longitud	Numero
3	YCOORD	Coordenada relativa de latitud	Numero
4	Prof	Profundidad del pozo	Numero
5	Fecha	Fecha en que culmino la prueba	Fecha
6	runlife	Cantidad de dias con la bomba trabajando a la fecha de culminacion de la prueba	Numero
7	ciclo	Numero de ciclos de inyeccion de vapor	Numero
8	Dpiston	Diametro del piston instalado en el pozo	Numero
9	Lon_Superficie	Longitud de la carrera de la unidad de superficie	Numero
10	SPM	Velocidad de la unidad superficie	Numero

11	RGP	Relación de gas disuelto en el petroleo en la prueba	Numero
12	AYS	Relación de agua y sedimentos en el petroleo en la prueba	Numero
13	Elasticity	Relación entre la carrera de superficie y la ultima medida en fondo	Numero
14	SP	Ultima carrera efectiva medida en fondo	Numero
15	Fillage	Llenado teórico medido el día de la prueba en carta de fondo	Numero
16	BBPD	Barrilaje bruto (agua + petroleo) medido en la prueba	Numero
17	BNPD	Barrilaje neto (solo petroleo) medido en la prueba	Numero

Limpieza y exploración de data

- Al cargar el dataset se valido que no existían datos nulos
- Seguidamente se buscaron valores outliers, en este caso como existen ecuaciones físicas que validan las medidas de producción se procedió a evaluar los valores existentes en base

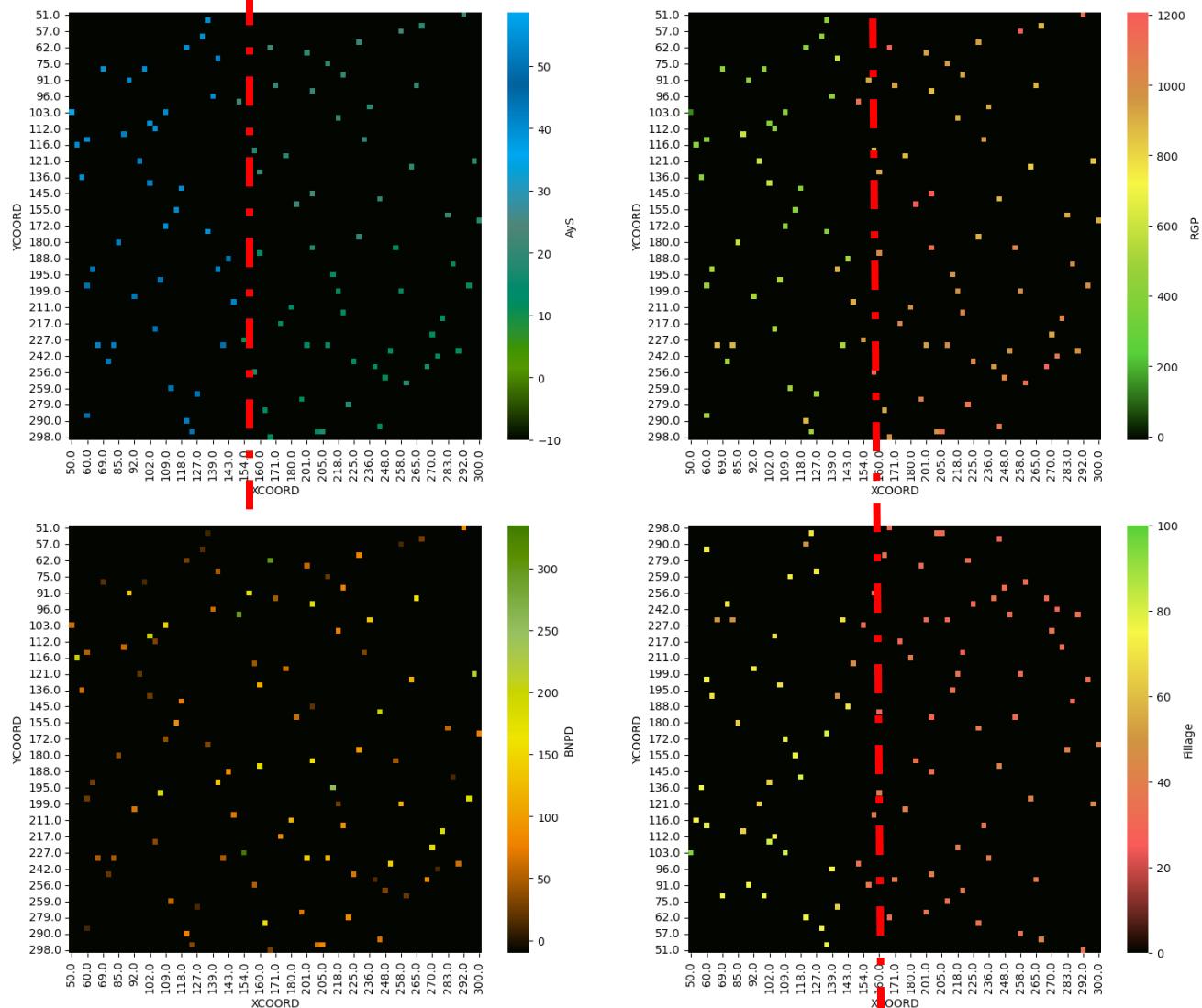
$$MaxTeorical = 0.1166 * Sp * dp^2 x SPM$$

- Se valido posteriormente que en ningún caso “BBPD” no superara el valor calculo como máximo.
- Seguidamente se procedió a realizar un agrupamiento de variables basados en casos de negocios .
 - Los pozos con profundidad menor a 7200 pies pertenecen a la arena “A” los mayores pertenecen a la arena “B”
 - Los pozos con una coordenada de longitud menor a 160 tienden a tener problemas de agua y mayores tienden a tener problemas de gas
 - Los pozos con producción bruta mayor a 200 bbd son considerados altos productores, los que están entre 100 y 200 son de media producción y los que producen menos de 100 son de baja producción
- El ultimo paso antes de realizar el análisis de la data se procedió a crear una tabla maestra o resumen, esta práctica es muy común en la rama petrolera pues debido a q toda data es de línea de tiempo y puede variar se procedió hacer una tabla resumen con los promedio anuales agrupados por pozos esto permite que los gráficos multivariados queden mas claro y es como se carga el pozo en el presupuesto anual.

Análisis Multivariado

1).- ¿Existe alguna sectorización del campo y de ser así en base a qué valores y cuáles serían su ubicación geográfica?

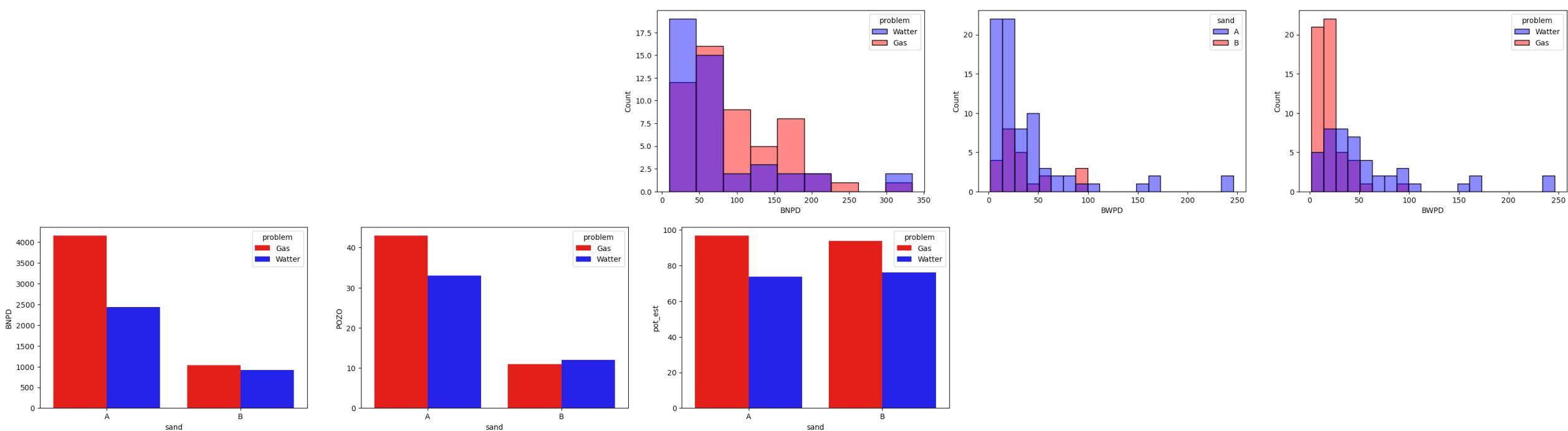
Con base en la comparación anterior, se observa que existe claramente una separación en 2 macro cuadrantes, dominados por varias condiciones, una de las cuales es la XCOORD, partiendo de 0-160 predominado por la producción de Agua y otro con XCOORD 160-300 afectado por la producción de gas. Aunque en el gráfico [1,0] donde vemos distribución espacial de producción de crudo no se observa una muy clara distribución categórica.



Análisis Multivariado

2).- ¿Qué tanto afecta la arena de producción de los pozos al caudal?

En función de la separación horizontal geográfica inicial de la data a la que llamaremos "Problem" con base en la predominancia de condición de producción asociada (agua / gas), introduciremos otra separación, la arena a la cual pertenece (en pozos con profundidad de 7200 ft o menos son de la arena A y mayores a 7200 son pozos nuevos pertenecientes a la arena B). En general se observa como la arena A presenta la mayor cantidad de pozos y por ende una mayor cantidad de producción total asociada, esto es congruente con el hecho de que la arena A tiene el doble de tiempo de operación que la arena B. Sin embargo, se observa como la proyección de la arena B "Pot-Est" es similar al de la arena A, es decir un pozo nuevo en cualquier arena podría esperarse cantidad similares de hidrocarburos (si lo vemos de forma promedio general), es por ello que en general la arena de la cual produzca el pozo no tiende a afectar mucho el caudal estimados, mas sin embargo en algunos pozo para ambas arenas se presentan algunos unos casos atípicos donde el caudal es elevado, para lo cual sería necesario realizar un estudio geológico un poco más a fondo. En resumen, la profundidad no tiende a afectar el caudal esperado.



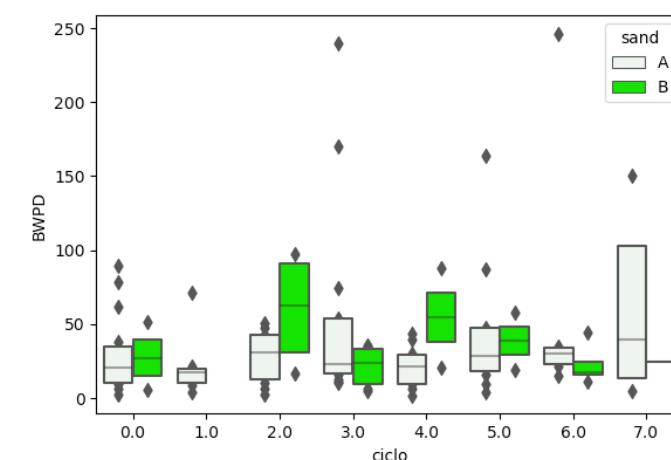
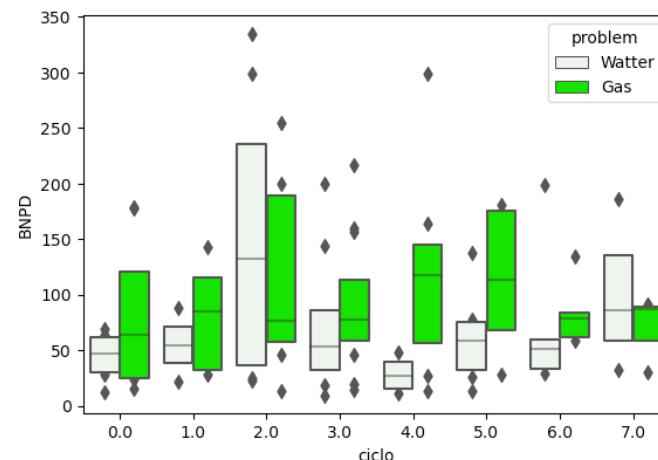
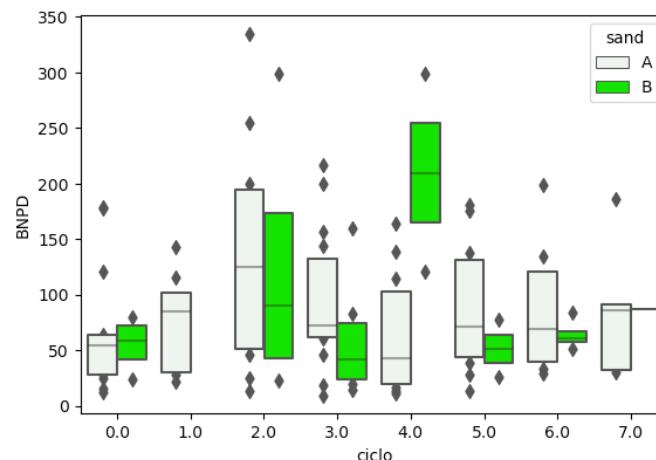
Análisis Multivariado

3).-¿Qué efectos tiene los ciclos de inyección en los pozos y hasta cuántos ciclos es recomendable inyectar?

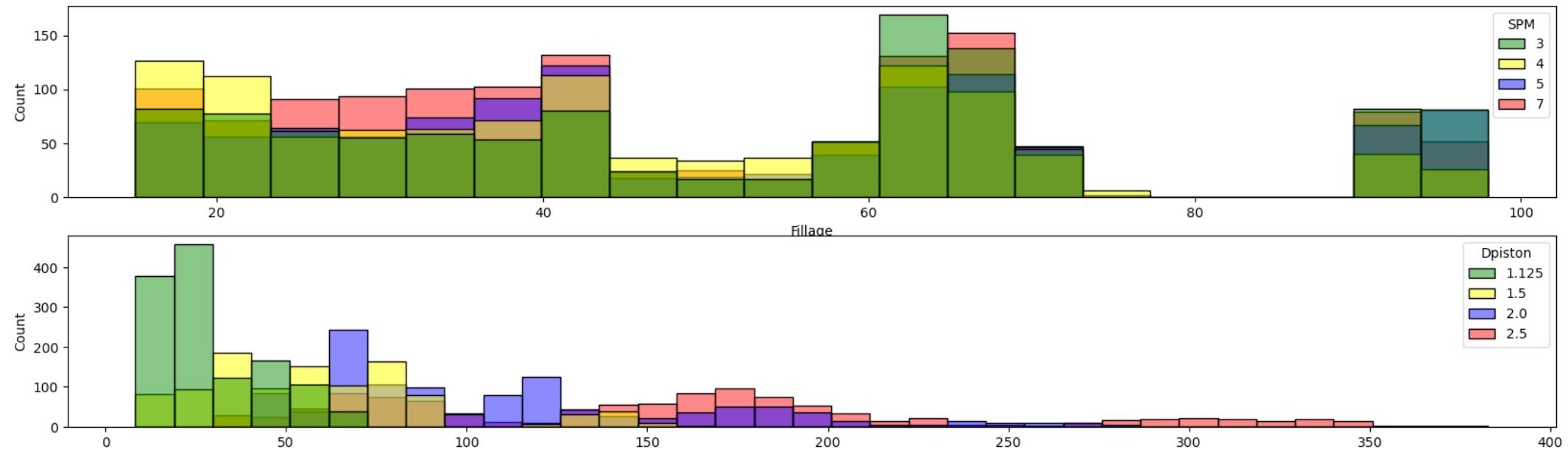
Referente al tema de los ciclos de inyección (los pozos cada cierto tiempo son inyectados con vapor caliente, esto ayuda a aumentar la temperatura de la arena y mejora la producción de hidrocarburo). En general, con base en la distribución espacial y data histórica de medidas en base a los ciclos podemos apreciar:

- Ciclos 0-2 : la arena A tiene un ligero mayor potencial esperado asociado a la inyección en comparación a los pozos de la arena B, ya en el segundo ciclo se observa como ambas arenas responden muy bien en los incrementos de producción, pero también es predominante que la inyección a largo plazo no agrega demasiada producción de agua.
- Ciclo 3 -4: en este grupo se observa que los pozos de la arena B tiende a responder de forma positiva, superior a los de la arena A, aumenta la producción de petróleo y no aumenta demasiado la de agua.
- Ciclo 5 -7: en general estos ya son pozos maduros con una varianza baja, son ya conocidos y no hay inestabilidad de producción. En estos ciclos vemos como la ya extendida inyección ha elevado la producción de agua.

Basado en los anterior se observa: es recomendable inyectar los pozos hasta un periodo de 3-4 ciclos, superior a estos el ganancial de hidrocarburo es poco.



Análisis Multivariado

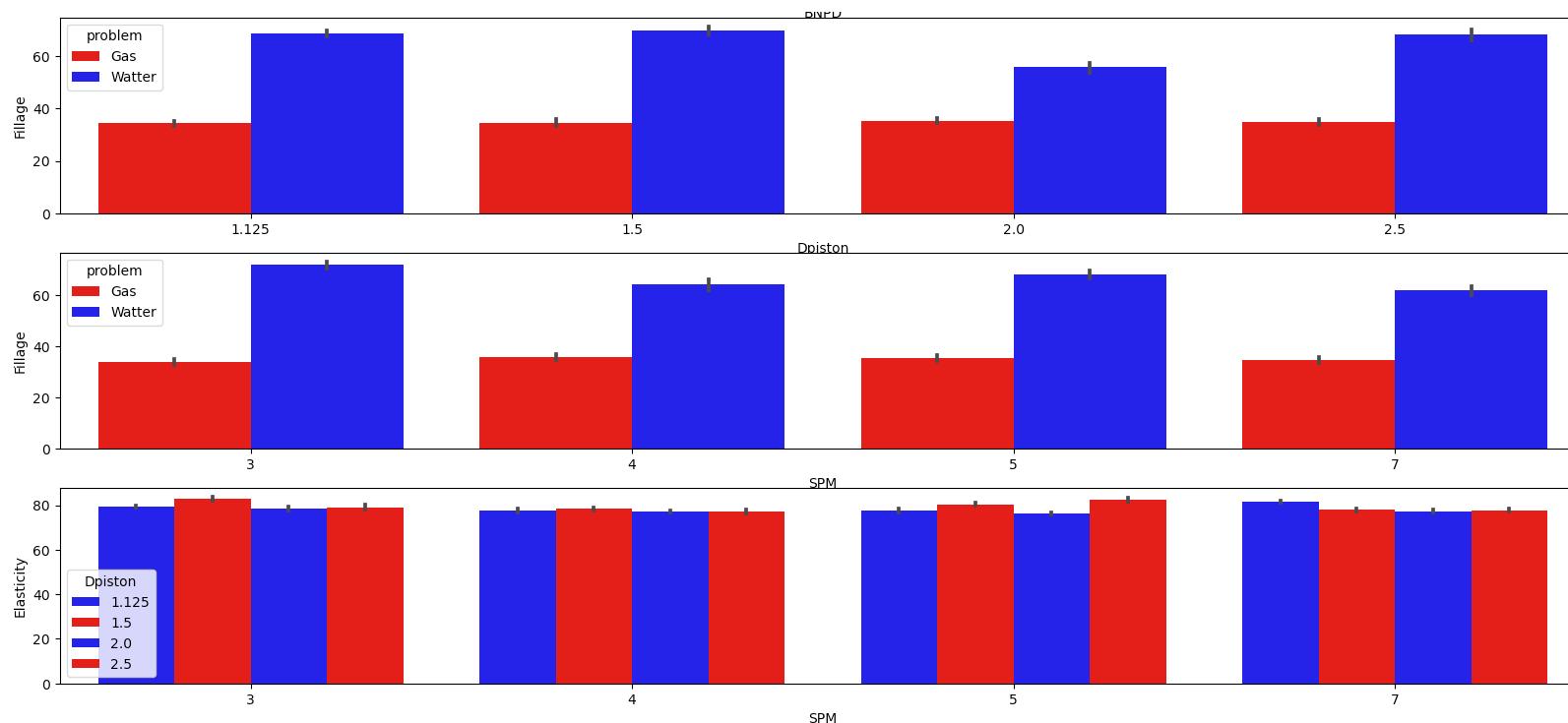


4).-¿Qué relación existe entre las variables operativas y qué tan optimizado está el campo?

Las 4 variables que miden qué tan optimizada está una bomba de subsuelo en este caso son: SPM (velocidad de revoluciones), Fillage (eficiencia de llenado en cada ciclo), Elasticity (que tan bien se transfiere el esfuerzo de superficie a fondo, esto se encuentra asociado a la eficiencia) y un parámetro adicional que es el pistón (tamaño del pistón):

- SPM-Fillage: lo ideal para tener mayor llenado es bajo SPM, en este sentido se observa como lo más predominante son pozos con 3 spm (el más bajo del campo). existen algunos pozos con spm (5-7) con llenado bajo (>40%) que deben ser evaluados a detalle, muy probablemente en estos casos sean pozos altos productores de gas, posiblemente en estos pozos de 5-7 spm se puedan bajar a 3 y mejoraría el llenado.
- SPM-Elasticity-Dpiston: en este caso la física dice que a mayor número de ciclos tiende a bajar la eficiencia por elasticidad, pero en general se observa que la elasticidad es constante entre los 0.75-0.85 sin importar la velocidad ni el diámetro del pistón. .

Análisis Multivariado

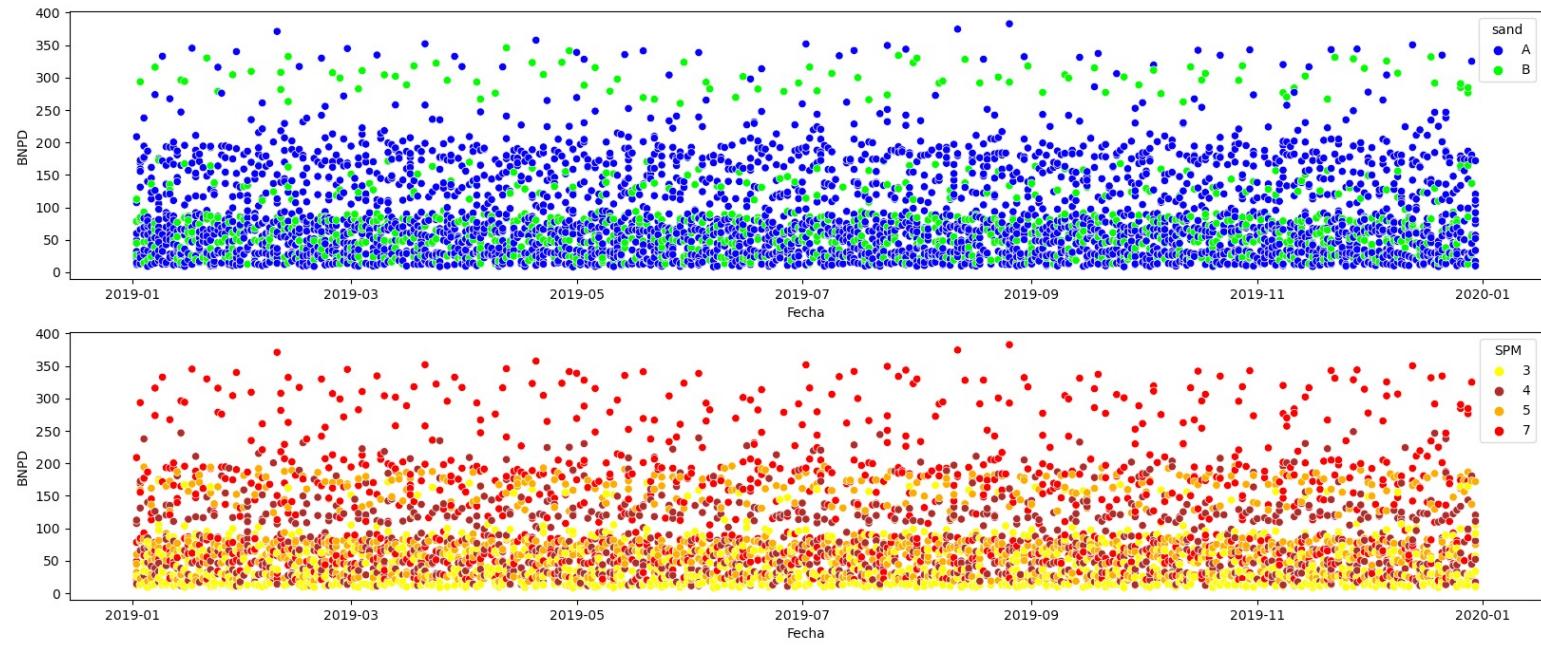


4).-¿Qué relación existe entre las variables operativas y qué tan optimizado está el campo? (Continuación)

- Fillage-Dpiston-Problem: se observa qué poco influye el pistón en el llenado efectivo de fondo y en general no se podría optimizar en este sentido, sólo resaltando los casos de pozos con alto SPM, pero son muy puntuales.

Concluyendo, se observa como el campo en general está constituido por pozos de baja producción, con pozos con pistones pequeños y bajos SPM, con llenados medios-bajos pero buena eficiencia elástica, se podrían probar bombas manejadoras de gas en zonas donde éste es un problema grande.

Análisis Multivariado

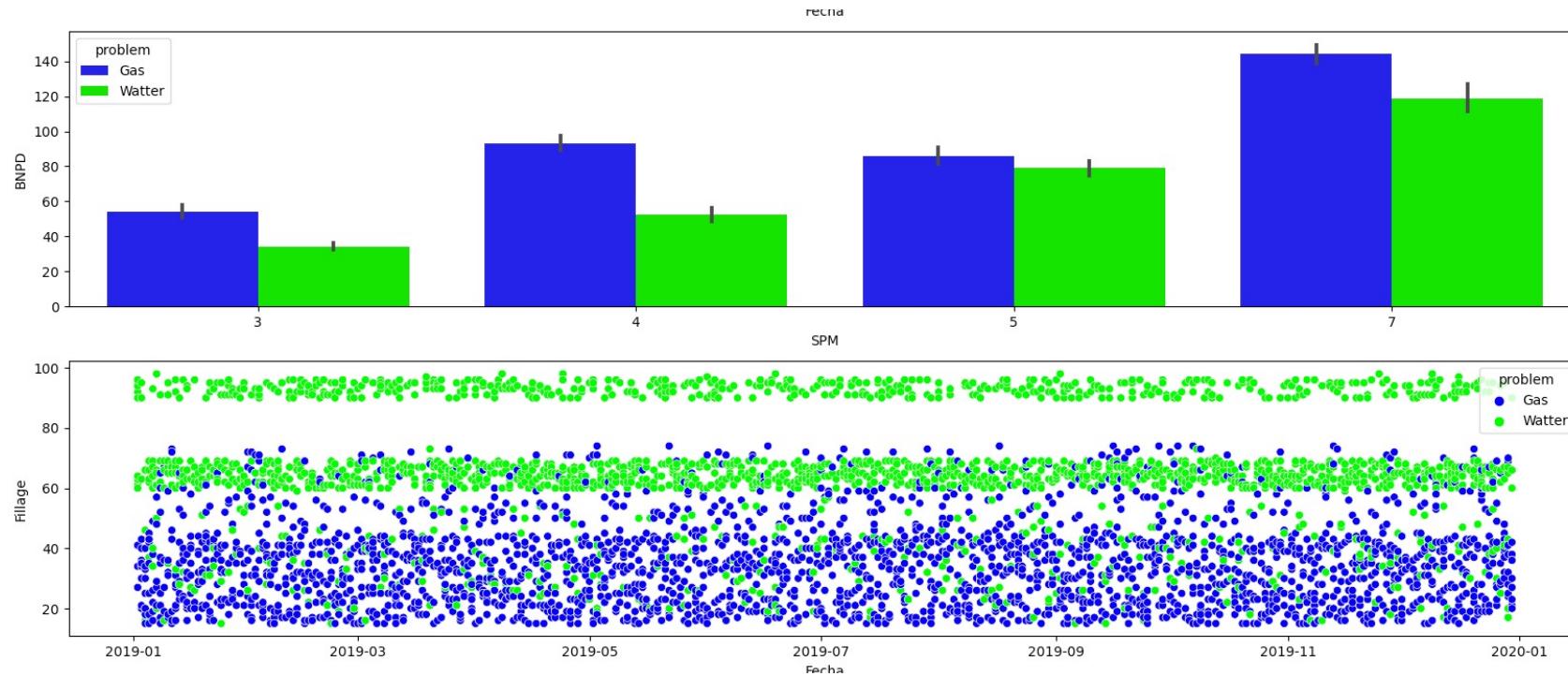


5).-¿Es homogéneo el comportamiento de los pozos a lo largo del campo o existen grupo basados a su tasa de producción, qué tan confiable sería un plan de explotación en el tiempo?

Evaluando la distribución histórica se observa como en general los mayores productores se encuentran distribuidos entre ambas arenas, lo cual apoya la idea inicial del potencial promedio igual en ambas arenas que vimos al inicio, pero es relevante que estos mayores productores son casos (muy pocos) que rompen la tendencia general.

En cuanto a la presunción de que la mayoría de los pozos estaban optimizados resulta parecer cierta pues hay una clara distribución de spm, donde los altos productores están entre 5-7 spm y la mayoría de los pozos con 3-4 spm están en zonas de baja producción, destacando que hay pozos con 7 spm en zonas de baja producción y probablemente sean los que vimos anteriormente con bajo llenado y estos podrían ser los pozos pendientes por optimizar, pero son pocos.

Análisis Multivariado



5).-¿Es homogéneo el comportamiento de los pozos a lo largo del campo o existen grupos basados a su tasa de producción, qué tan confiable sería un plan de explotación en el tiempo?

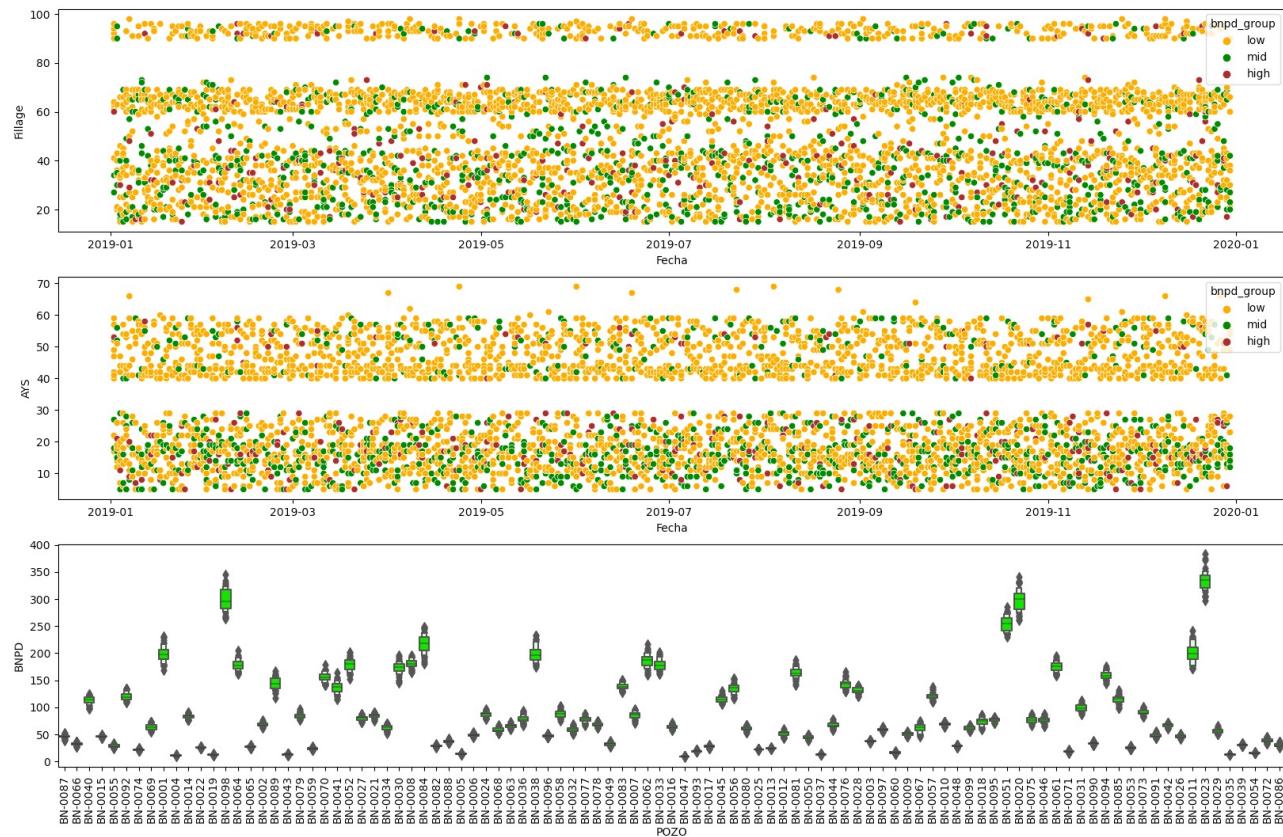
Resalta que el problema del gas está asociado a baja producción individual, pues en ningún caso se observa un buen llenado de bomba o buen potencial en pozos donde este problema predomina, en el caso del agua si se observa como existen grupos claramente definidos y algunos presentan buena producción (aquellos casos donde hay buen llenado), la relación entre buen llenado y altos productores es clara en pozos con velocidad entre 5-7 spm y en este caso ya se tiene un AYS superior a 40% (optimizados). Es relevante destacar que los productores de gas son los pozos que en conjunto acumulan el 60% de la producción del campo.

Análisis Multivariado

5).-¿Es homogéneo el comportamiento de los pozos a lo largo del campo o existen grupos basados a su tasa de producción, qué tan confiable sería un plan de explotación en el tiempo?

En resumen, vemos como en general los pozos son bajos productores con sus excepciones. Están bien sectorizados los bajos y altos productores y tienen baja variación en el tiempo por lo cual se pueden hacer claros planes de explotación.

Aun cuando en los análisis de arriba hemos visto una correlación entre variables debemos recalcar que ninguna variable presenta una cerrada correlación con otro (pues toda variable registrada es una consecuencia multicausal en el área de estudio (es decir siempre afectara el agua, el gas, incluso variables del área y roca que no hemos considerado))



Correlación entre variables

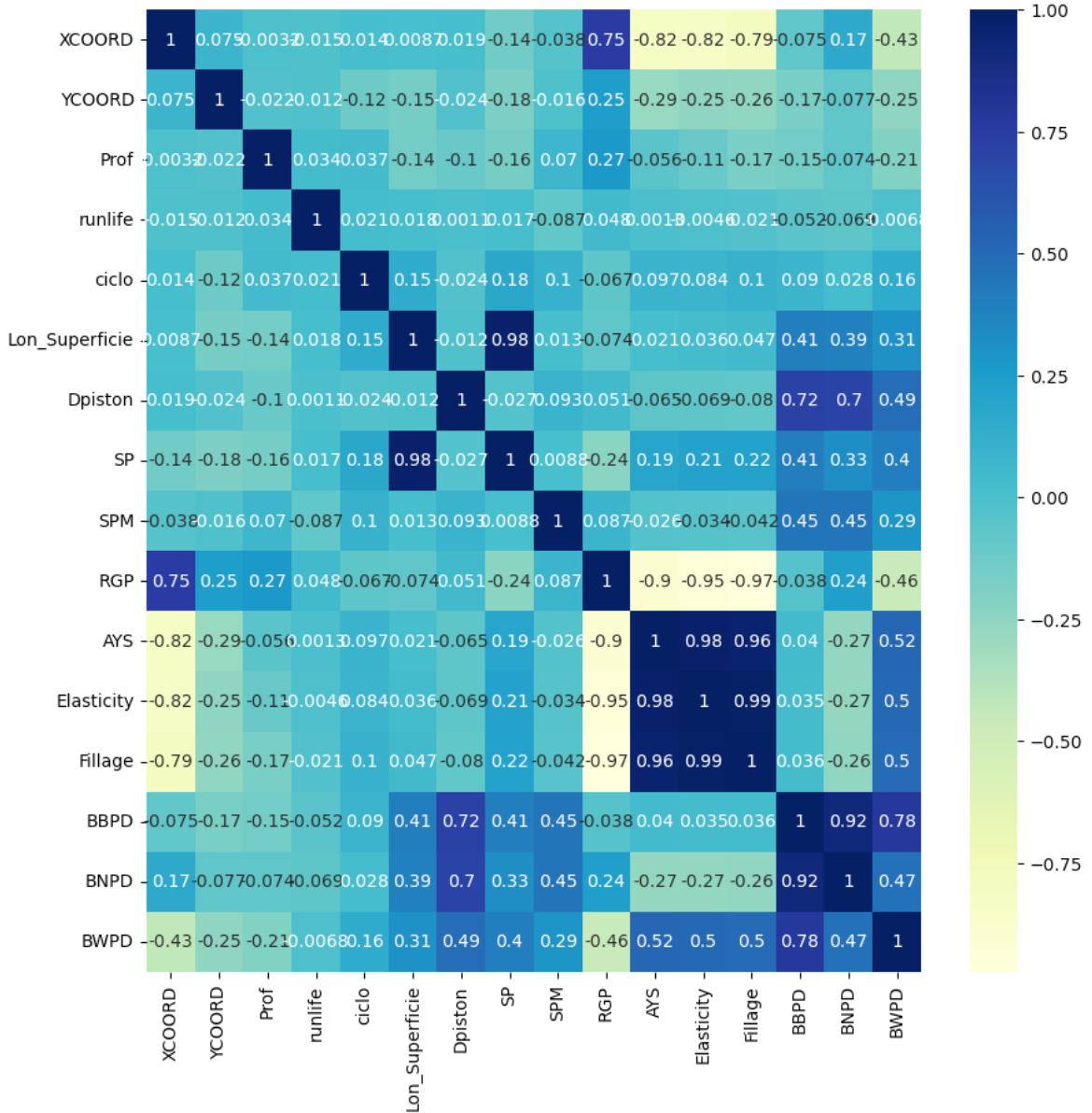
Antes de modelar hacemos una tabla de relación entre variables, es importante mencionar que de la misma podemos destacar las siguientes relaciones predominantes:

-

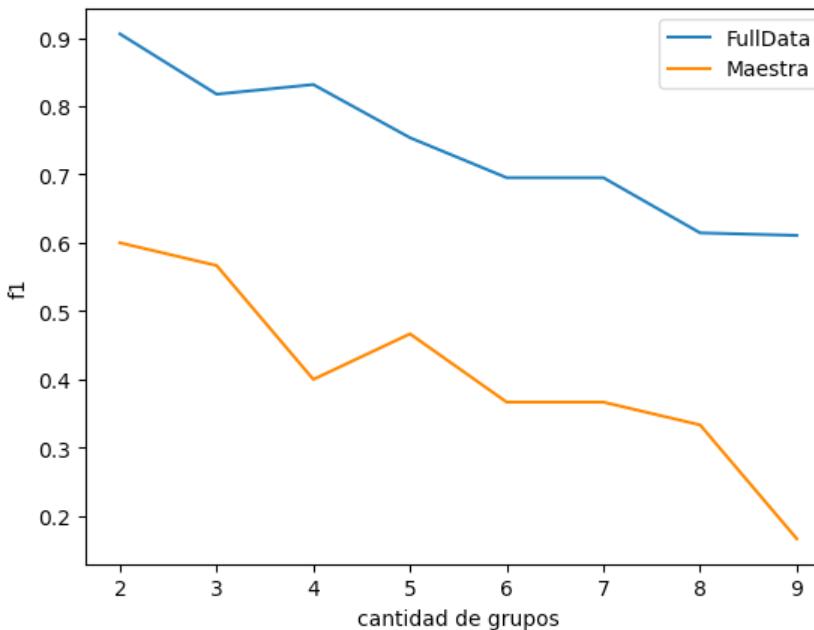
*BBPD : buena relación variables operacionales (Lon. superficie, Dpiston, Sp, SPM), de estas 3 escogeremos Dpiston para entrenar junto a las variables de ubicación geográficas, que aún cuando su correlación son bastante bajas, es una solicitud comercial introducirlas como data input.

-

*AYS : la relación a variables operacionales es baja por ello nos quedaremos sólo con las variables de ubicación geográficas que presentan buena correlación.



Ajuste y entrenamiento del modelo de agrupamiento en base a un caudal de pozo esperado



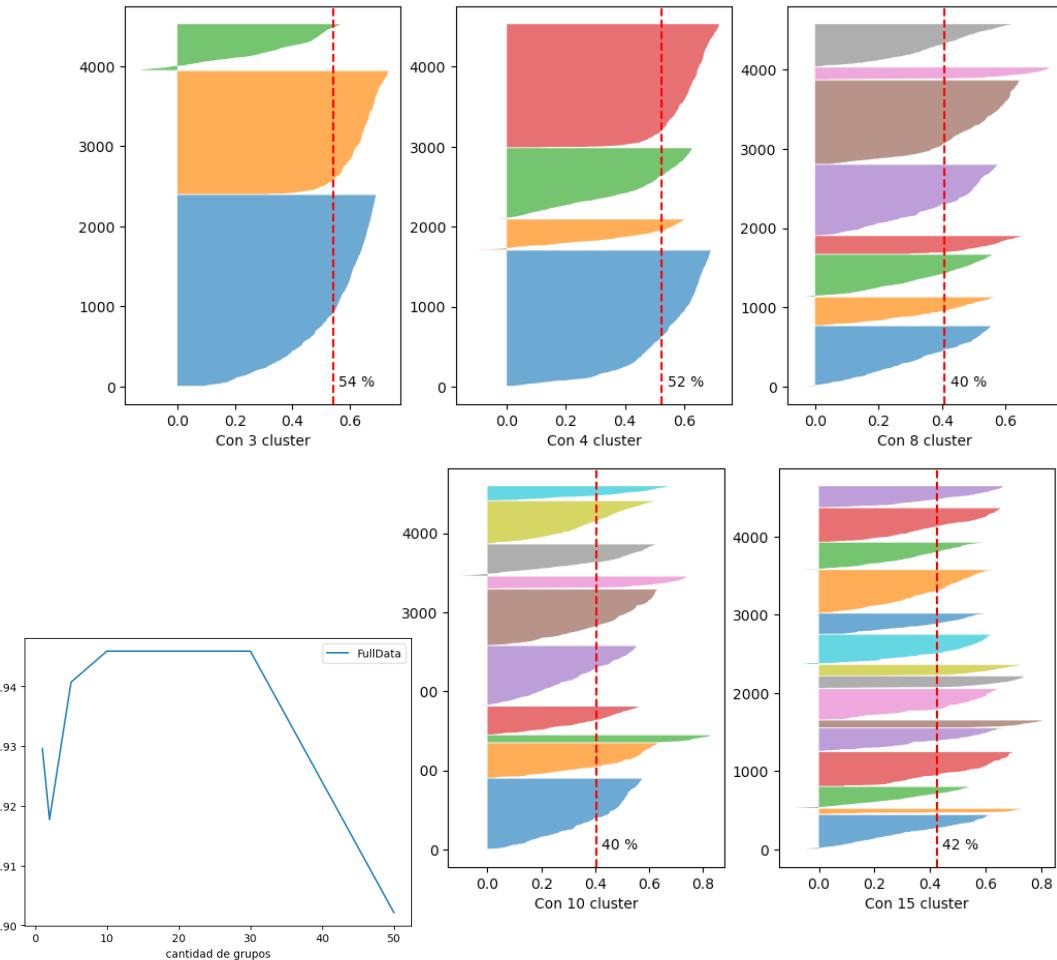
En este punto nos resulta llamativo si para modelos de data se utilizaría el df resumen que hemos llamada maestra (que se ha usado para el análisis univariado) o se utilizaría el df original, el resultado que podemos ver en el grafico es congruente con lo esperado (a mayor data mejores resultados para el modelo), por lo cual en adelante seguiremos usando el df original con toda la data.

En el gráfico podemos ver cómo, para efectos de modelado por árbol de decisión, resulta mucho más funcional la data cruda global no resumida de la tabla maestra. Adicionalmente se ve cómo al aumentar la cantidad de grupos entre los pozos baja la capacidad de los modelos para predecir el agrupamiento del pozo, el número ideal estaría entre 2 y 3 grupos, por razón de negocio nos quedaremos con 3 grupos sectorizados de forma manual (bajos, medios y altos).

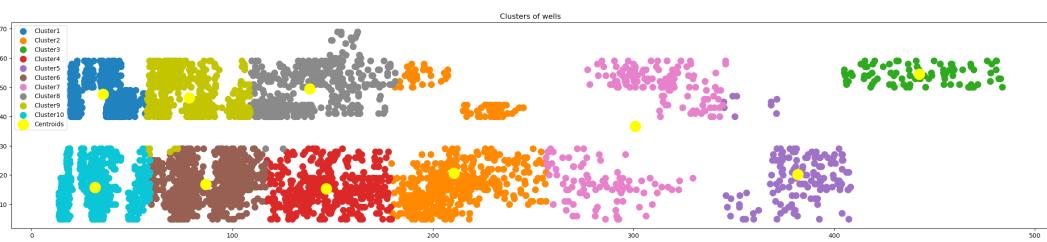
Ajuste y entrenamiento del modelo de agrupamiento en base a un caudal de pozo esperado

De los gráficos en esta sección podemos sacar los siguientes datos relevantes para modelado por KNN:

- Podríamos clasificar el cluster en 3 grupos (el mismo que trae negocio haciendo históricamente): si es cierto que el índice de silhouette muestra la clasificación mas alta al ver la clasificación visual de los cluster entre AYS vs BBPD vemos condiciones de barreras difusas y con un rango alto de posible caudal esperado donde se dificultaría determinar si un pozo es bajo, medio o alto y esto podría ser económicamente muy riesgoso, es por ello que descartamos este escenario
- Clasificar el cluster en 10 grupos: el estudio del codo muestra un F1 muy bueno, además el índice de silhouette muestra una clasificación aceptable (no tan buena pero aceptable), si ahora vemos la visualización física de AYS vs BBPD vemos clusters no tan extensos en rango, como en el caso anterior y aun cuando siempre se dificulta determinar que tanto dará un pozo alto productor los cluster difusos se encuentran por encima de los 200 BBPD donde se supera el mínimo de producción requerido para pagar operaciones de perforación, es por eso que esta opción resulta la mas idónea
- Esta clasificación por el diagrama del codo muestra poca mejora con respecto a los 10 grupos, adicionalmente a esto el índice silhouette mejora un poco pero no excesivamente y si sumamos que esto representaría mayor gasto computacional que no se vera reflejado en clusters de rango de producción mas bajo, se descarta esta opción para evitar sobrecomplejizar el caso de estudio



Basado en lo anterior se concluye que continuaremos con 10 grupos lo cual resulta en cluster bien separados y mas o menos homogéneos, esto estudiando la data cruda que nos aporta nuestro dataSet (todas las pruebas de producción), resalta que nuevamente el AyS resulta una variable de alto impacto especialmente en pozos entre los 180 hasta los 420 bbdp donde al acercarse a mayor caudal aumenta la presencia de agua y crea incluso subclusters especialmente en pozos entre (180-320 bbdp), se observa como las mayores medidas del campo se obtuvieron en presencia de alto corte de agua (>50%)



Ajuste y entrenamiento del modelo de agrupamiento en base a un caudal de pozo esperado

6).-¿Es posible clasificar un pozo nuevo como alto, medio o bajo productor y que método es el que resulta más apropiado para modelarlo?

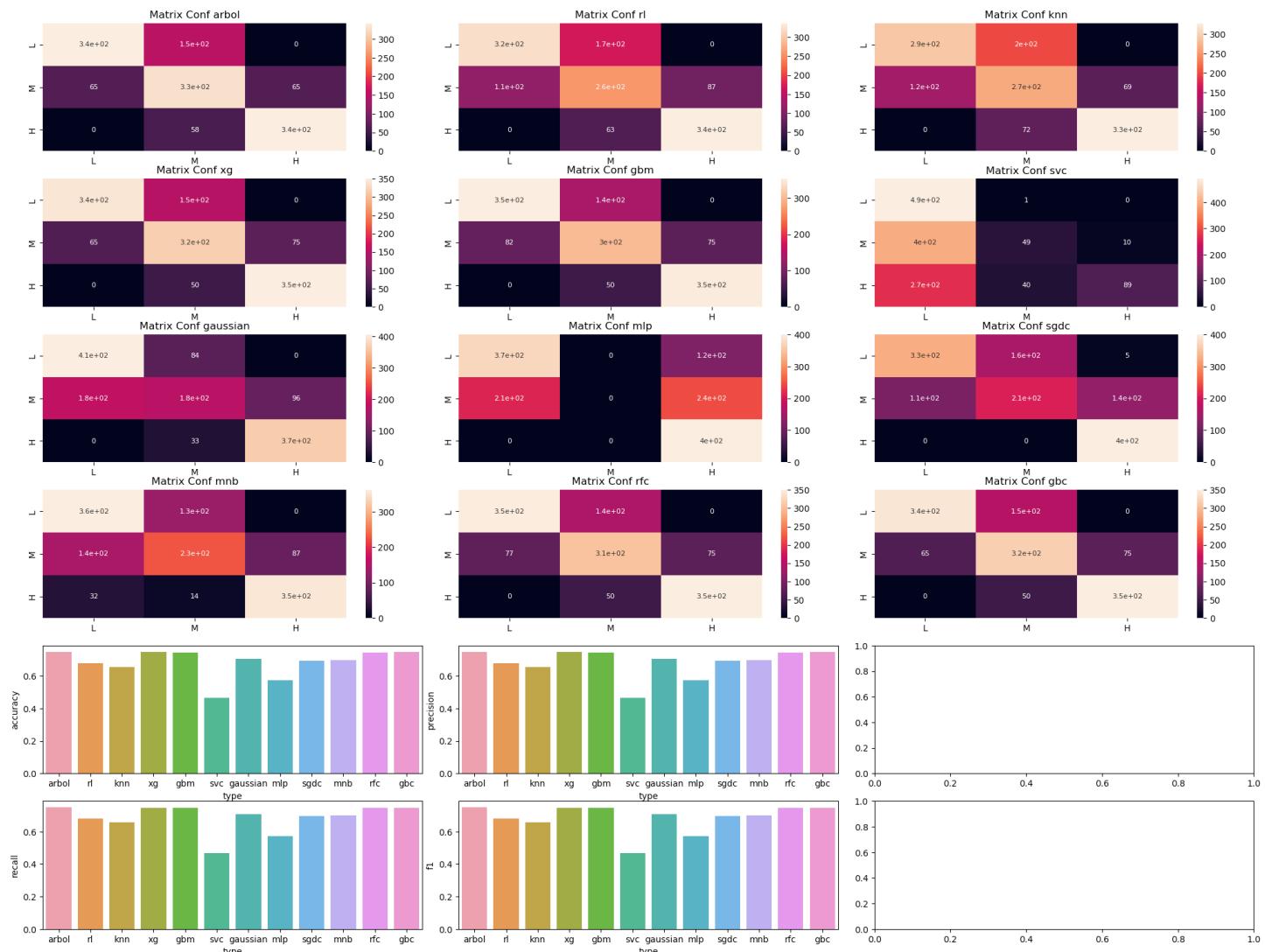
Con relativa buena precisión (>94%) es posible clasificar un pozo nuevo como alto, medio o bajo productor basado en condición espaciales y la velocidad de operación, los métodos más acertados son 'KNeighborsClassifier', 'XGBClassifier' y 'LGBMClassifier', resulta relevante como basado en matriz de confusión vemos como 'KNeighborsClassifier', 'XGBClassifier' y 'LGBMClassifier' se comportan de una forma casi idéntica, resaltan 19 casos de error críticos donde los pozos son clasificados como alto caudal y resultarían siendo de bajo y 10 de moderado cuidado donde son clasificados como altos pero son bajos, aunque en promedio estos 29 casos de cuidado tienen un contrapeso de 44 pozos que serian clasificados como bajo o medio y en realidad serian altos potencial.



Ajuste y entrenamiento del modelo de agrupamiento en base a un corte de AyS de pozo esperado

7).-¿Es posible clasificar un pozo nuevo como alto, medio o bajo corte de agua y que método es el que resulta más apropiado para modelarlo?

Con relativa media precisión ($\sim 74\%$) (usando Randomized Search XGB) es posible clasificar un pozo nuevo como alto, medio o bajo productor con base en condiciones espaciales y la velocidad de operación, el método más acertado es 'XGB', Es relevante como basados en matriz de confusión, en el caso de AYS el 'KNeighborsClassifier' tiende a subestimar en 1200 casos el corte de agua, clasificándolo como bajo pero que en realidad seria de medio caudal de AYS y en contraposición en 200 casos hace lo inverso clasifica como medio y seria bajo caudal de AyS, este parámetro es critico y se recomienda descartar este algoritmo y quedarse solo con XG y GBM para clasificar por producto de AYS.



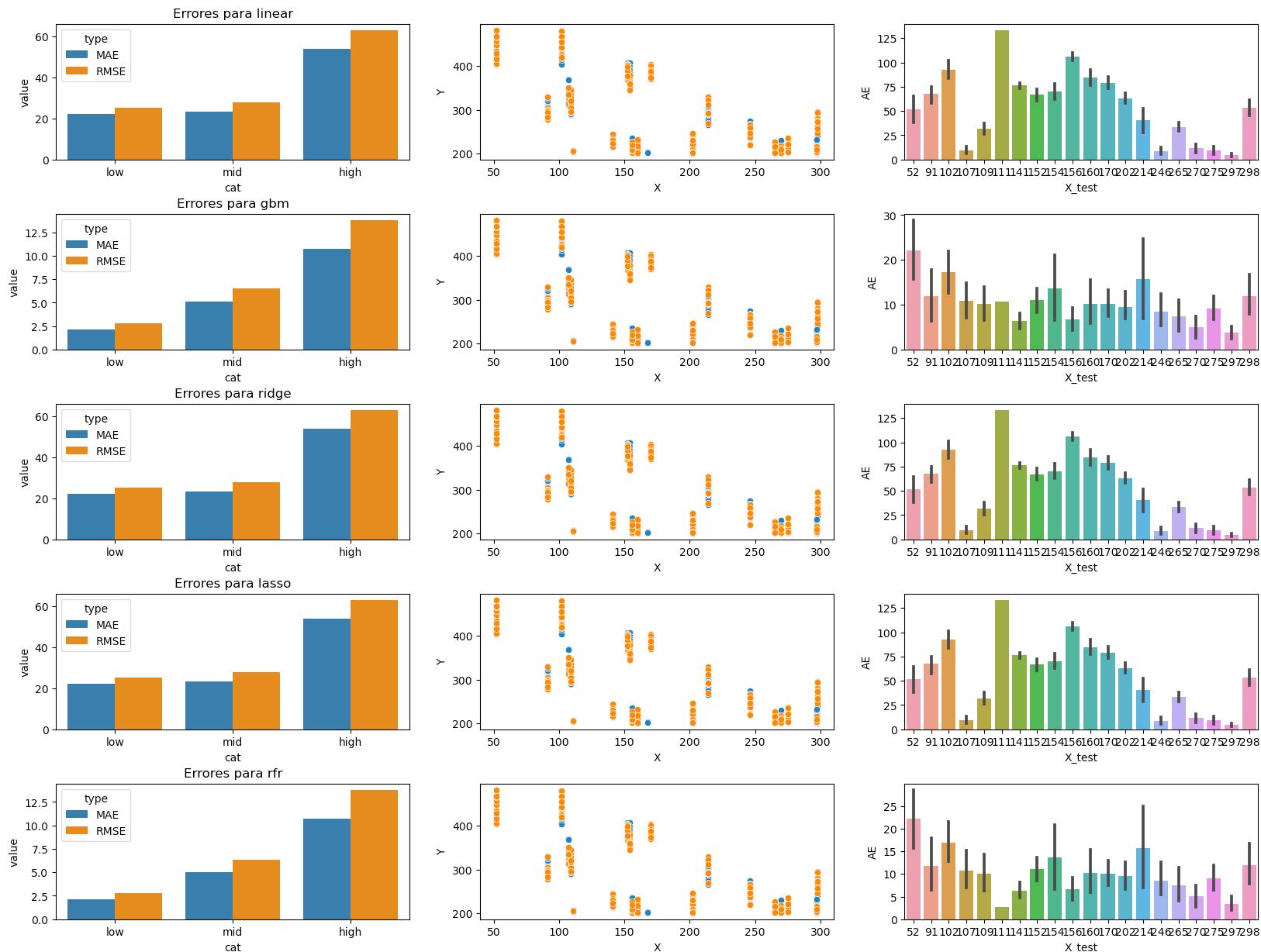
Ajuste y entrenamiento del modelo de estimación del caudal de pozo esperado

8).-¿Es posible predecir la tasa de producción bruta de un pozo nuevo y qué método resulta más apropiado?

La tasa final de un pozo será un parámetro multifactorial pero con base en una clasificación inicial que permite evaluar la inversión económica si se podría predecir un valor aproximado de la tasa de crudo, siempre que el pozo pueda ser previamente clasificado como alto, bajo o medio productor, el valor obtenido tendría una variación de:

- * ~ +/- 3 bbpd para pozos de baja tasa
- * ~ +/- 5 bbpd para pozos de media tasa
- * ~ +/- 11 bbpd para pozos de alta tasa

Esto utilizando el método de LGBMRegressor.



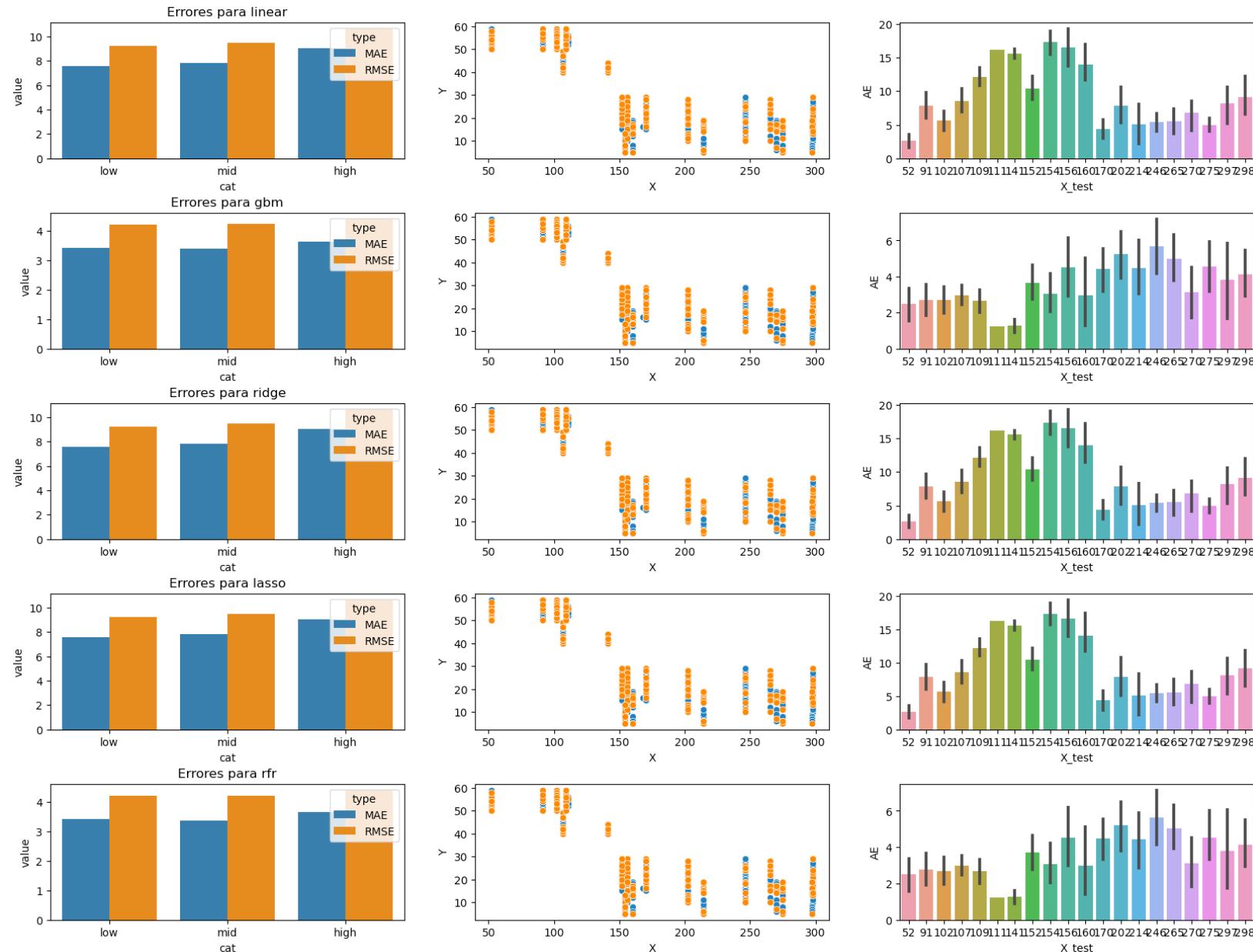
Ajuste y entrenamiento del modelo de estimación del Ays de pozo esperado

9).-¿Es posible predecir la tasa de agua de un pozo nuevo y que método resulta más apropiado?

La tasa final de agua de un pozo será un parámetro multifactorial pero con base en una clasificación inicial que permite evaluar inversión económica si se podría predecir un valor aproximado de la tasa de crudo, siempre que el pozo pueda ser previamente clasificado como alto, bajo o medio productor, el valor obtenido tendría una variación de:

* ~ +/- 4 para pozos de alto, baja y media tasa

Esto utilizando el método de LGBMRegressor.



Pase a producción del modelo

Basados en las buenas estadísticas obtenidas, el modelos se puso a correr en un server de testeo (render.com) en el siguiente endpoint:

<https://coder-house-final.onrender.com/predict>

Para probarlo es necesario hacerlo por medio de un request POST, en endpoint responderá :

```
{"bbpd_group": "mid", "bbpd_value": "159.48",
 "ays_group": "low", "ays_value": "13.91"}
```

La CURL para testeo es

```
curl --request POST \
--url https://coder-house-
final.onrender.com/predict \
--header 'Content-Type: application/json'
\ 
--data '{
  "XCOORD":79,
  "YCOORD":175,
  "Prof":7179,
  "Dpiston":1.5
}'
```

