Universidad de Buenos Aires

Facultad de Ciencias exactas físicas y naturales

Maestría en Explotación de Datos y Descubrimiento del Conocimiento

Trabajo de especialización

Predicción secuencial de series de tiempo en Registros de Pozo

Autor : Ingeniero Rodrigo Mauriño

Supervisor : Profesor Ricardo Maronna

Fecha : Junio 2020

Índice

1 Introducción Técnica del problema

En el industria del petróleo es muy importante entender las estructuras geológicas que yacen en el subsuelo para poder identificar áreas prospectivas con capacidad para albergar y transmitir hidrocarburos. Para ello se realizan mediciones de distintas propiedades del suelo a lo largo de un pozo llamadas Registro de Pozo. Las herramientas que se corren en un registro de pozo son variadas y cada una apunta a entender un aspecto distinto del subsuelo, en este estudio en particular se trabajará con la medición llamada Gamma Ray para identificar tipos de rocas sedimentarias. Es crucial entender que tipo de rocas están presentes en el pozo dado que las rocas arcillosa presentan poca porosidad y muy baja permeabilidad actuando usualmente como sello de reservorios en áreas productivas. En contraposición las areniscas presentan muy alta porosidad y permeabilidad por lo que son áreas muy interesantes para estudiar su carácter productivo.

La herramienta Gamma Ray mide la radiación de rayos gamma producida por el decaimiento de los isótopos del potasio presentes en arcillas. En caso de encontrarnos con una roca arenisca la herramienta leería un valor muy bajo de irradiación. En base a la amplitud de la señal se puede detectar distintas proporciones de arcilla en las paredes del pozo según su presencia o ausencia en las estructuras geológicas. Además si tenemos en consideración las mediciones contiguas dentro de un pozo podemos detectar gradientes de la señal que indicarían zonas mixtas que varían su composición de rocas sedimentarias en profundidad.

El objetivo del proyecto es analizar los registro Gamma Ray para predecir segmentos del pozo con areniscas, arcillas o zonas de mixtas. El análisis exploratorio se centrara en entender el comportamiento de las labels para obtener información interesante en el momento de crear nuevas features para el modelo predictivo. Para lograr esto analizaremos transversalmente el comportamiento de los pozos, fenómenos que puedan ocurrir en determinadas profundidades del pozo y la segmentación y comparación de intervalos de distintas tendencias. Se utilizará el algoritmo Random Forest por su simpleza y se pondrá el foco de la investigación en encontrar las features correctas que deberíamos generar para obtener una buena predicción.

Dentro de estos segmentos se pueden clasificar 5 casos distintos basándose en su forma y origen geológico.

- **Serrated:**Tiene forma de serrucho con base en valores altos de Gamma Ray y predomina la

presencia de arcillas Estos sistemas geológicos suelen ser depositados por ambientes de baja energía como planicies fluviales o plataformas costeras.

- **Funnel:** Se observa una tendencia linealmente progresiva en el aumento de la cantidad de arcilla y señal medida.
- **Bell:** De manera similar al Funnel, este tipo de estructuras tiene una tendencia linealmente descendente en la cantidad de arcilla en aumento.
- **Symmetrical:** En este caso vemos un descenso y posterior aumento en la señal y por lo tanto la cantidad de arcilla. Forma un banco de areniscas simétrico entre arcillas.
- **Cylidrincal:** Presenta forma de bloque de valores muy bajos de Gamma Ray por su predominante composición de rocas areniscas. Estas rocas sedimentarias suelen ser formadas en ambientes depositacionales de tipo eólico

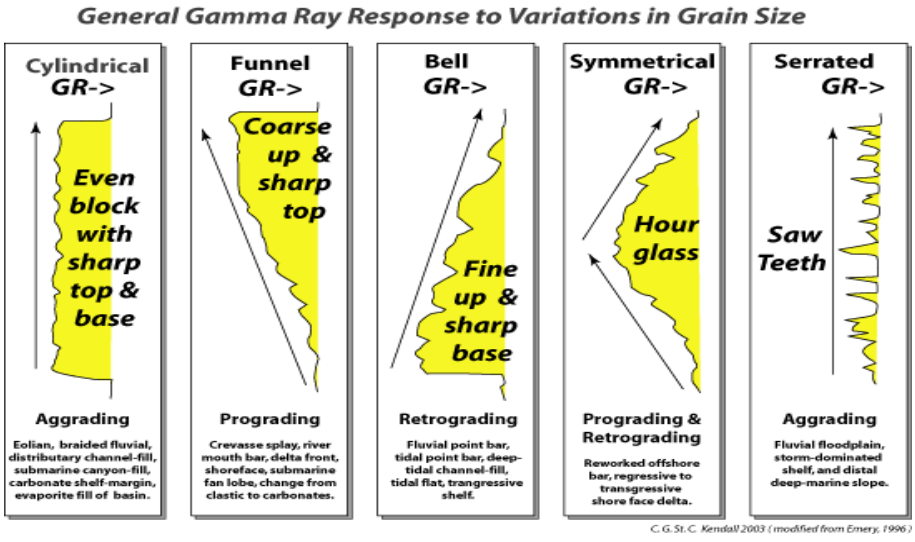

Fig. 1: Clasificaciones de estructuras arcillosas

2 Análisis exploratorio de los datos

El Dataset utilizado cuenta con 4.400.000 número de registros que ilustran la medición GR de 4000 pozos donde cada uno tiene 1100 mediciones a lo largo de su profundidad. Al no tener la

profundidad de cada pozo no podemos saber exactamente cada cuanto se realizó cada medición y esto puede ser fuente de ruido. Contamos con las siguientes columnas iniciales:

- **row_id:** Número entero de la medición a la largo del pozo. Está contenido entre 0 y 1100 para cada pozo en el database.
- **well_id:** Número entero que identifica cada pozo.
- **GR:** Número real con la medición de la herramienta. Este valor generará la serie de profundidad para cada **well_id** en cada una de sus **row_id**.
- **Label:** Clase utilizada que clasifica la **row_id** en base a su valor de **GR** en una de sus 5 clases posibles.

En la figura 2 vemos la proporción de clases en el dataset entero. La clase mayoritario es la Serrated con más del 52% de presencia mientras que las 4 clases restantes compartan el resto de las ocurrencias equitativamente con un 12% de proporción aproximadamente. Es importante tener estos valores en mente dado que nos plantea el baseline de las predicciones.

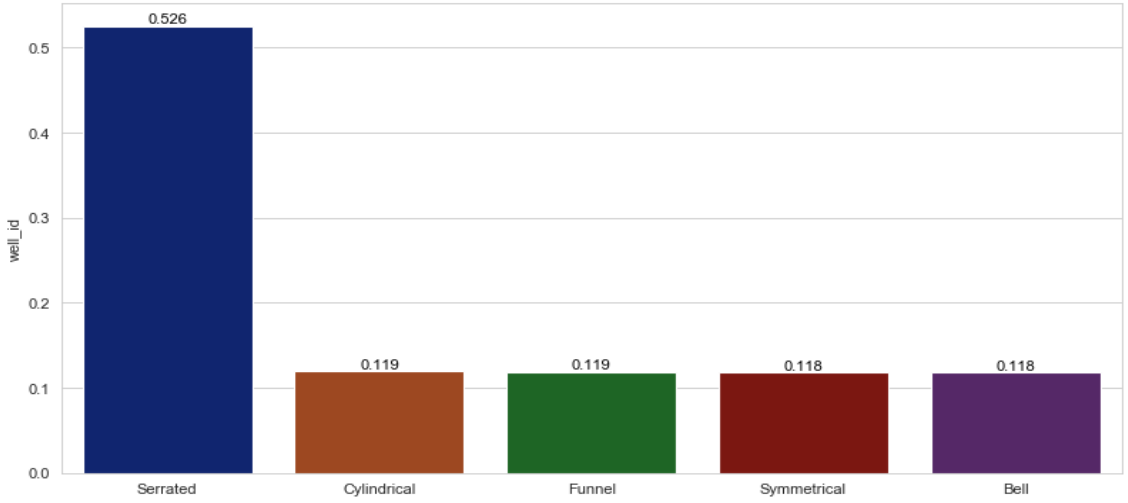

Fig. 2: Proporción de clases

2.1 Análisis enfocado a los pozos

Para ilustrar el comportamiento del **GR** tomamos como ejemplo el pozo de **well_id** 239 y graficamos todos sus valores mostrando la clasificación en la Figura 3.

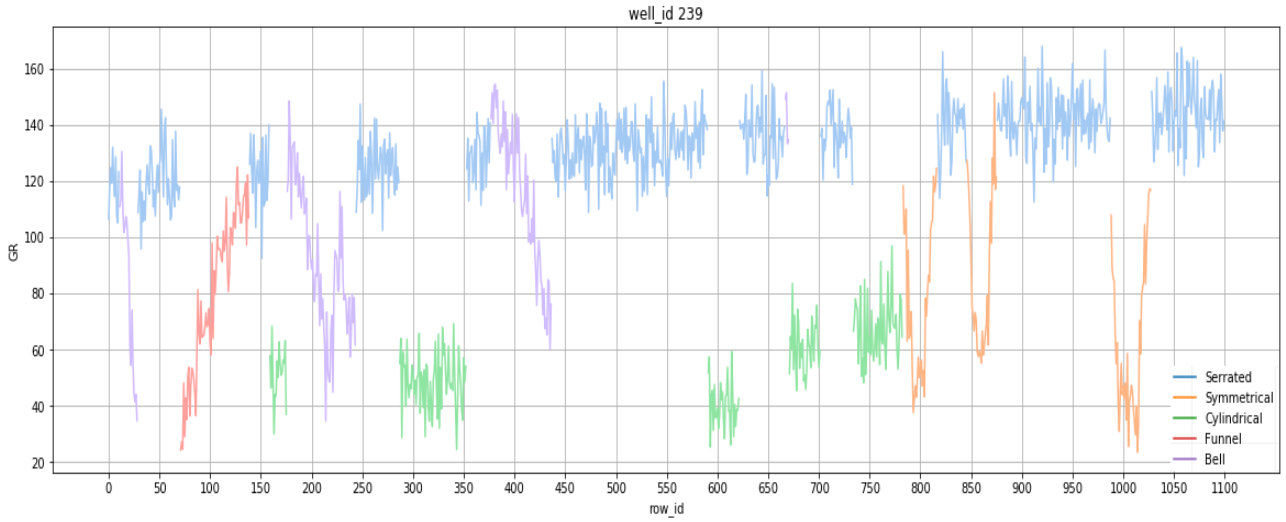

Fig. 3: Pozo ejemplo: well_id 239

A simple vista vemos que efectivamente las clases tienen un comportamiento muy similar al mencionado en el análisis técnico del problema en la Figura 1.Usaremos los mismos colores para representar las clases en todos los posteriores análisis para facilitar la lectura e interpretación. La clase Serrated es mucho más predominante en este pozo que el resto de las clases y posee una tendencia horizontal con ruido. De misma forma la clase Cylindrical presentan semejanzas en cuanto a la tendencia horizontal y al ruido pero vemos que tiene una media promedio mucho más baja que la clase Serrated. Funnel y Bell tienen un gradiente lineal ascendente y descendente respectivamente. Cabe destacar estos gradientes no necesariamente conectan continuidades de alta presencia y baja presencia de arcilla. Por ejemplo en el punto 370 vemos que hay una clase Funnel interconectando registros de clase Serrated en ambos extremos. En cuanto a la clase Symmetrical se puede decir que es una combinación de las clases Bell, Cylindrical y Funnel pero con menor ruido y pendiente más pronunciada. Veremos en los siguientes análisis que la gran mayoría de estas observaciones se extienden a todas las clases de todos los pozos.

Es Interesante destacar que no en todos los pozos se pueden observar las 5 clases. En 409 (10%) de los pozos al menos una clase se encuentra ausente y solamente 2 pozos faltan ocurrencias de 2

clases, como vemos en la Figura 4.

Fig. 4: Diagrama de Venn de Número de Pozo con alguna clase ausente en su Serie de datos Gamma Ray

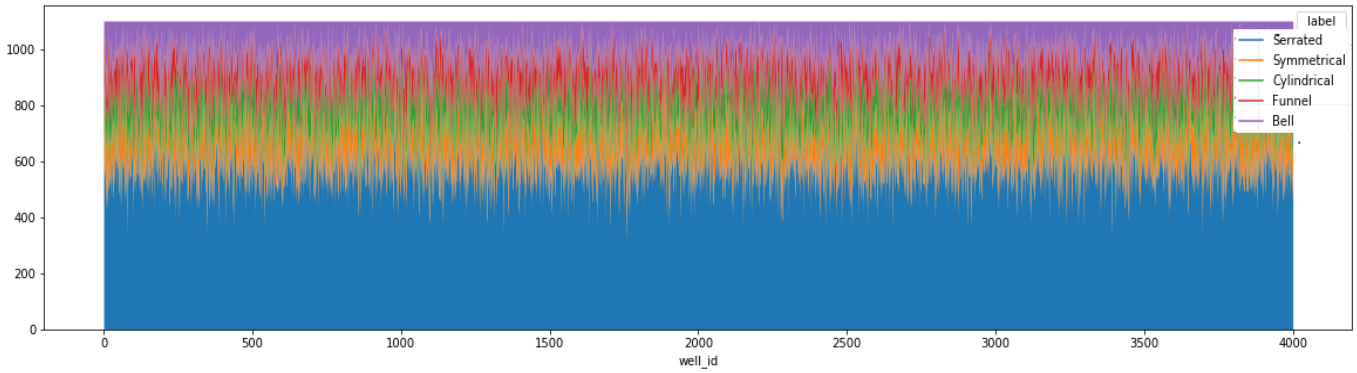

Fig. 5: Comparación de labels vertical entre los pozos

En la Figura 5 ubicamos todos los pozos unos a lado del otro verticalmente y ordenamos los registros por label comprobando la presencia mayoritaria de la clase Serrated en todos los pozos y una proporción similar de las clases minoritarias entre si.

2.2 Análisis por row_id

Es importante plantearse si la distribución de la label es dependiente de la variable row_id para apreciar si la ubicación longitudinal de la medición nos puede aportar valor predictivo.

Para analizar esta relación contamos la cantidad de veces que aparece cada label para cada row_id como vemos en la gráfico de áreas de la Figura 6

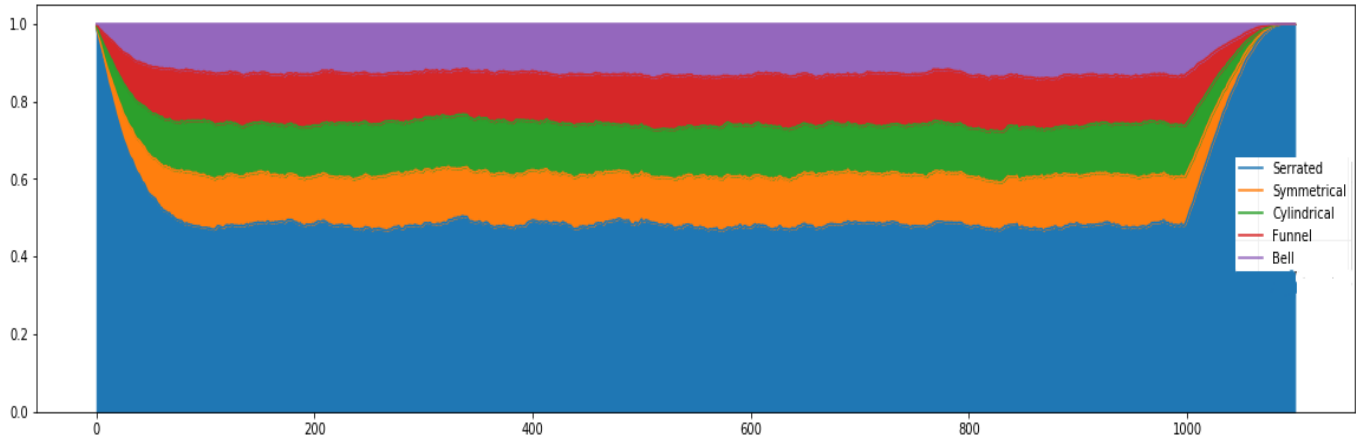

Fig. 6: Gráficos de Áreas labels por row_id

En la figura 7 se puede apreciar que la probabilidad de que una determinada label aparezca es dependiente de la row_id por lo que esta variable es importante conservarla a la hora de hacer la predicción.

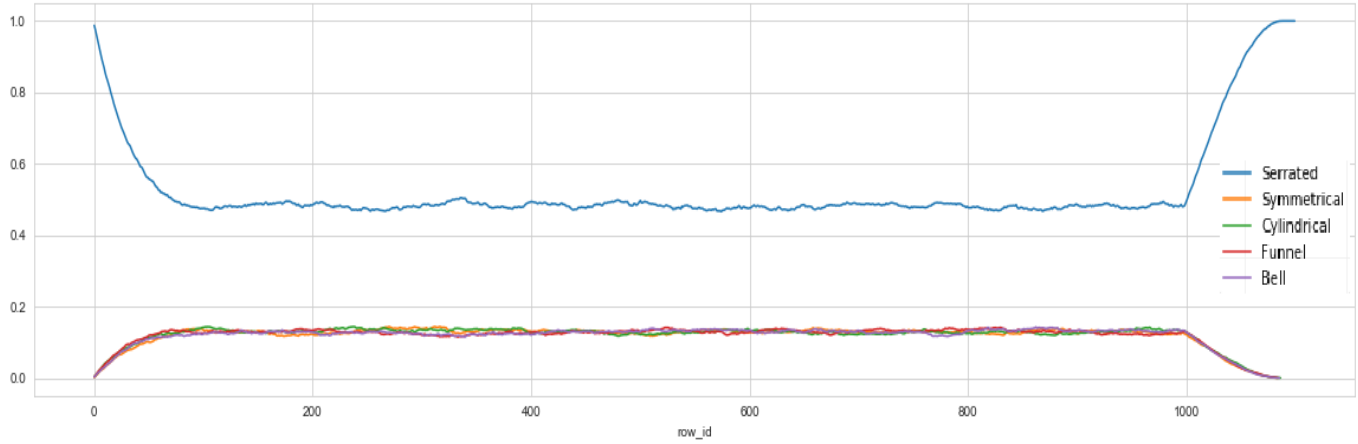

Fig. 7: Gráficos de probabilidades labels por row_id

Se destaca que los valores de row_id comprendidos entre 0 a 70 y 1000 a 1100 presentan una probabilidad muy aumentada que aparezca la clase Serrated disminuyendo gradualmente para los extremos cercanos al centro. En el tramo central las probabilidades son independientes de la row_id porque se mantienen constantes. En este tramo central la probabilidad de la clase Serrated es menor

al 50% que se calculo en las proporciones totales de label, esto se debe al efecto de los extremos que hay en los pozos.

2.3 Análisis univariados de los intervalos de labels

Para encarar el análisis de los segmentos de las label, más allá de su pertenencia a cada pozo, extraemos los intervalos de cada label y realizamos un análisis univariado con las 82.296 longitudes encontradas de cada label.

En el Boxplot de la Figura 8 notamos que todas las clases minoritarias tienen comportamientos similares, tienen valores de media, mediana, variación y distribución prácticamente iguales. La media de estas clases se encuentra centrada entre los cuartiles y podríamos estar viendo una distribución simétrica dejando de lado los outliers. La clase mayoritaria tiene una mayor tendencia a la asimetría.

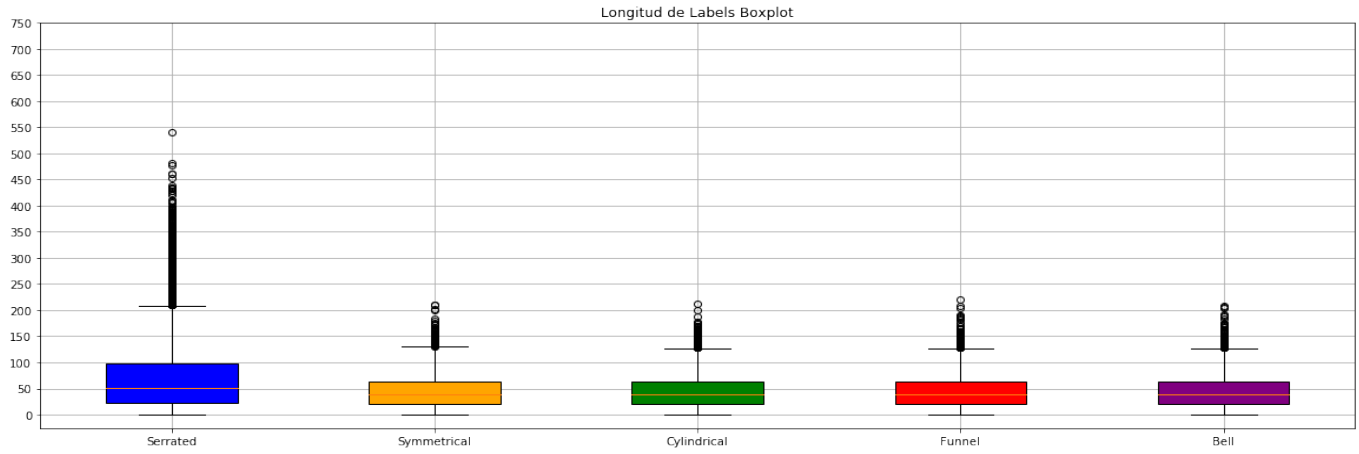

Fig. 8: Boxplot de longitudes de intervalo por label

Labels	Media	Desvío	Mediana	Máximo	Mínimo	Cantidad	Proporción
Serrated	68.4869	61.6081	51.0	541	1	33777	0.4104
Symmetrical	43.0937	28.0318	39.0	211	1	12084	0.1468
Cylindrical	42.9474	27.7330	39.0	212	1	12233	0.1486
Funnel	43.0951	27.8603	39.0	221	1	12116	0.1472
Bell	42.8975	28.1755	38.0	209	1	12086	0.1469

Tabla 1: Parámetros estadísticos

De la Tabla 1 es importante prestar atención al tamaño medio de los intervalos dado que nos ayuda a calibrar las ventanas de tiempo cuando creemos nuevas variables temporales. Si bien la label Serrated está presente en un 52% de los registros, solo posee posee un 41% de segmentos lo indicaría que efectivamente estos intervalos son más largos que los de las clases minoritarias.

En la Figura 9 vemos que salvo por unos outliers con longitud de segmento superior a 90 row_id las 4 clases presentan distribución uniforme en sus longitudes. La clase mayoritaria parece que tiene una distribución exponencial.

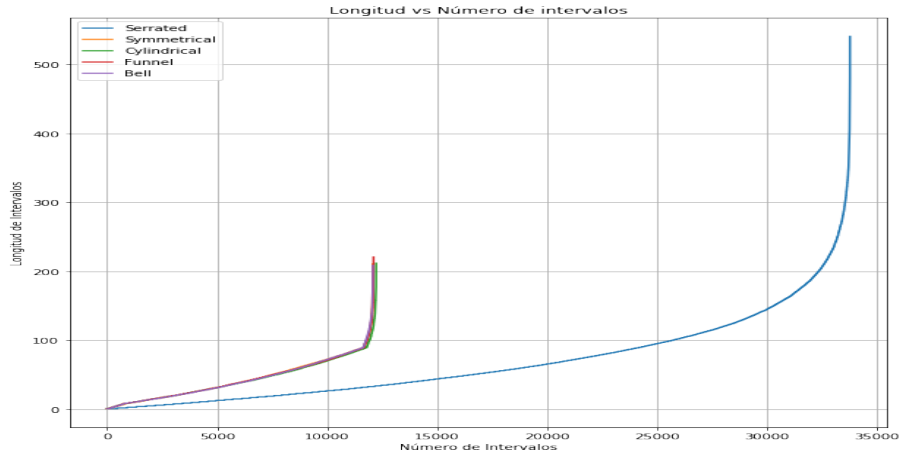

Fig. 9: Ranking ordenado de longitudes de intervalo por label.

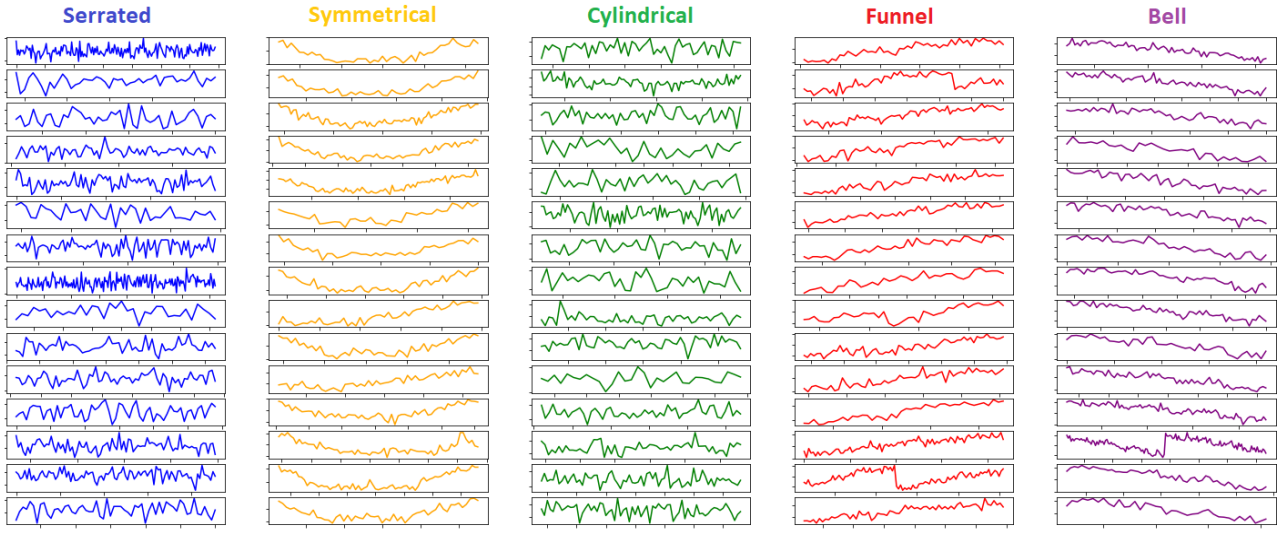

Fig. 10: Extracción al azar de varios segmentos de longitud superior a la media.

En la Figura 10 extrajimos al azar intervalos que solamente tienen una longitud similar a la media

de su clase y los visualizamos uno al lado de otro. En esta figura podemos comprobar la tendencia típica de cada clase según el análisis técnico. Se detecta que la clase Symmetrical presentan muy poco ruido dentro de su tendencia cuadrática para estas muestras tomadas, mientras que las clases Serrated y Cylindrical son las clases con más ruido dentro de su tendencia horizontal.

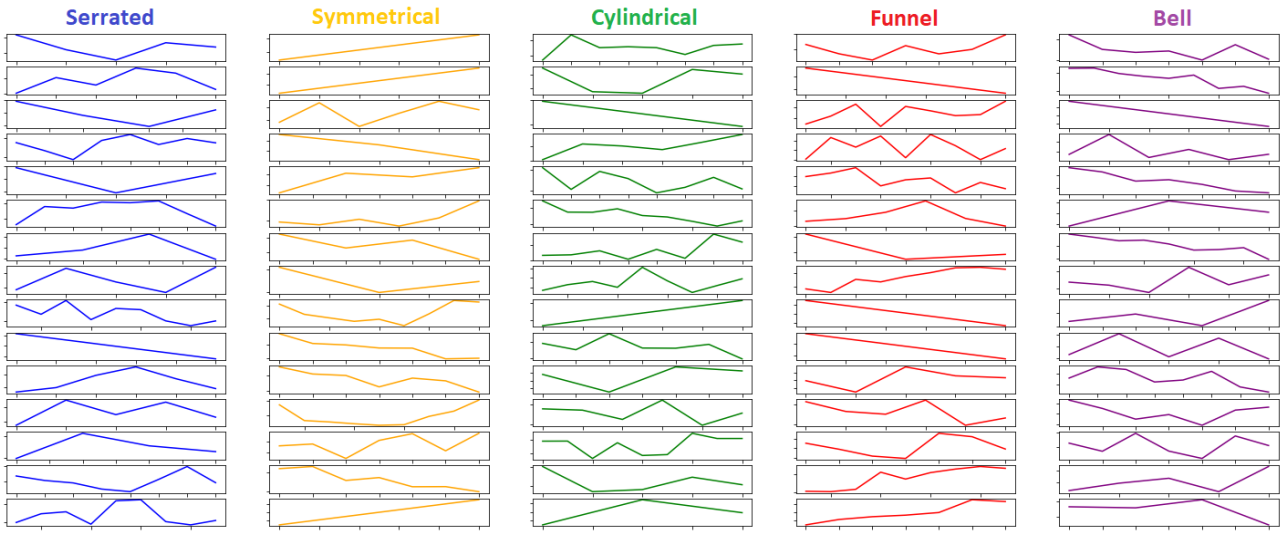

Fig. 11: Extracción al azar de varios segmentos de longitud muy inferior a la media.

Si hubiéramos elegido segmentos de longitud muy distintas a la media de su clase, como por ejemplo valores entre 1 y 10 registros, obtendremos la Figura 11. no se pueden distinguir ningún tipo de tendencia y estos serían los casos más difíciles de predecir dado que no podemos valernos de la tendencia del segmento.

3 Feature engineering

Inicialmente solo contamos con 2 variables predictoras row_id y valor de GR por lo que para encarar esta predicción de series de datos debemos crear variables que le den a un registro información contextual sobre los valores siguientes y posteriores dentro de un mismo pozo. Podemos buscar la información en ambas direcciones dado que no se predice en tiempo real sino que se cuenta con la serie de registros completa.

Primero se normalizo los valores de GR entre máximos y mínimos de todos los pozos dado a que se notaron diferencias generalizadas entre las series de tiempo que podrían significar variaciones en la calibración en la herramienta, por lo que se reemplazó la variable GR por **GR_mnorm**.

Por las características de las clases que presentan tendencias lineales y cuadráticas en algunos casos se planteó 2 modelos regresiones locales para poder captar la información de estas tendencias lo mejor posible. Por un lado se eligió una regresión lineal local Huber [1] que analizara en 2 ventanas de 20 y 40 registros de longitud centradas en cada punto a predecir. Se eligió esta modificación robusta de la regresión lineal clásica para ignorar los outliers generados por ruido que se presentan especialmente en los tramos más horizontales. Este coeficiente tiene la intención de representar la "derivada" local en cada punto de la serie de profundidad. Se pueden ver estos outliers en las muestras ejemplo de los casos de la Figura 10 en las clases Serrated y Cylindrical. Se tomarán los coeficientes lineales de ambas ventanas y se utilizaran como features de cada registro. En la Figura 12 vemos el caso de las regresiones locales robustas con ventanas de 40 unidades para el pozo 8,hay zonas que la recta describe muy bien el comportamiento de la serie pero en los valles no tanto.

Para captar la tendencia cuadrática presente, especialmente en la clase Symmetrical, se realizaran regresiones cuadráticas locales por mínimos cuadrados simples en ventanas de 30 registros de longitud. De cada una de estas ventanas extraeremos los 3 coeficientes de la ecuación cuadrática como features para cada registro. En este caso no se tomó percusiones para evitar deformaciones en los coeficientes por la presencia de outliers.

En adición a estas features generadas por modelos locales, se tomaron ventanas locales de medias y varianza de 30 unidades de largo utilizando una función de suavizado de Barthann para los bordes. Se utilizó este valor de ventana con la intención de simular la distribución de longitud de intervalo que tienen las clases minoritarias para captar la mayor cantidad de información importante sin agregar ruido. La función Barthann aplica pesos a los valores de cada ventana de manera de priorizar los valores cercanos al registro central con una curva similar a la de una normal, Figura 13.

En Resumen obtenemos las siguientes nuevas features como ventanas:

- **1huber:** Pendiente de la regresión lineal realizada en ventanas centradas de 20 registros de longitud.
- **2huber:** Pendiente de la regresión lineal realizada en ventanas centradas de 40 registros de longitud.

[1] https://link.springer.com/article/10.1007/BF01934700

Fig. 12: Ejemplo de regresión linea local Huber locales cada 40 espacios en ventanas de 40 unidades de longitud en el pozo 8

- **1polyfit:** Primer coeficiente de la regresión cuadrática realizada en ventanas centradas de 30 registros de longitud.
- **2polyfit:** Segundo coeficiente de la regresión cuadrática realizada en ventanas centradas de 30 registros de longitud.
- **3polyfit:** Tercer coeficiente de la regresión cuadrática realizada en ventanas centradas de 30 registros de longitud.
- **30rolling_mean:** Promedio de una ventana centrada de 30 ponderando con una función de barthann.
- **30rolling_std:** Variación de una ventana centrada de 30 ponderando con una función de barthann.

Finalmente se agregó para cada registro los valores futuros y pasados de los registros vecinos para distintas features. Cada registro cuenta con el valor de **GR_mnorm** , **1huber** y **40rolling_mean** de los 10 vecinos hacia adelante y hacia atrás saltando de a un valor. De esta forma cada registro cuenta con los valores desfasados en una ventana de 40 registros alrededor suyo para estas features.

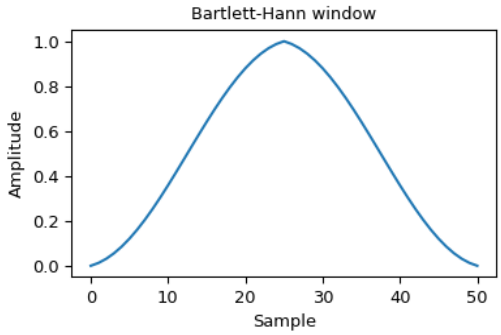

Fig. 13: Forma de la función de pesos aplicada Barthann

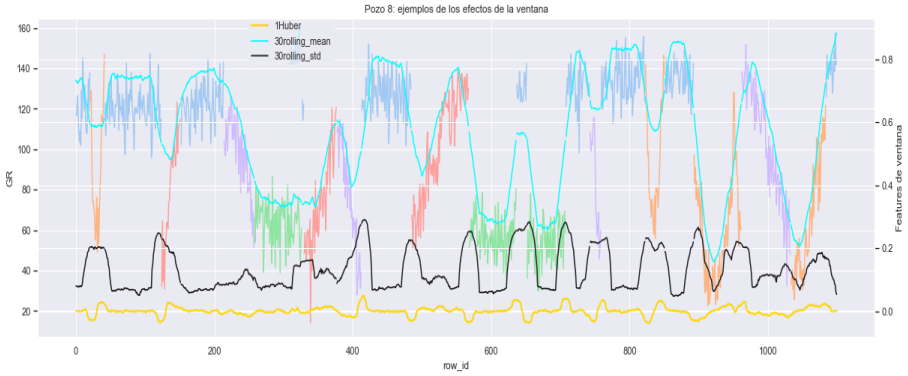

Fig. 14: Ejemplo de algunas ventanas de features en el pozo 8

Al final del feature engineering contamos con 69 features tras realizar el desfase de 60 variables, 7 features de la creación de propiedades de ventanas, el GR normalizado y la row_id.

4 Subsampleo de los datos

Debido a la gran cantidad de datos que tenemos y a los limitados recursos computacionales que contamos es conveniente realizar un subsampleo para realizar la toma de decisiones respecto a las variables generadas y a la elección de hiperparametros del modelo. Con este objetivo en mente realizamos una selección controlada de los datos tratando de hacer undersampling sobre la clase mayoritaria para igualar las proporciones. Por eso aplicamos una regla para quedarnos con los pozos que tienen menos de un 45% de la clase mayoritaria y al menos un 11% de cada clase minoritaria. Conseguimos un dataset que tiene 52 pozos, aproximadamente un 3% de los datos originales, con 57200 registros.

Podemos verificar que se mantuvieron las proporciones originales de clases al comparar:

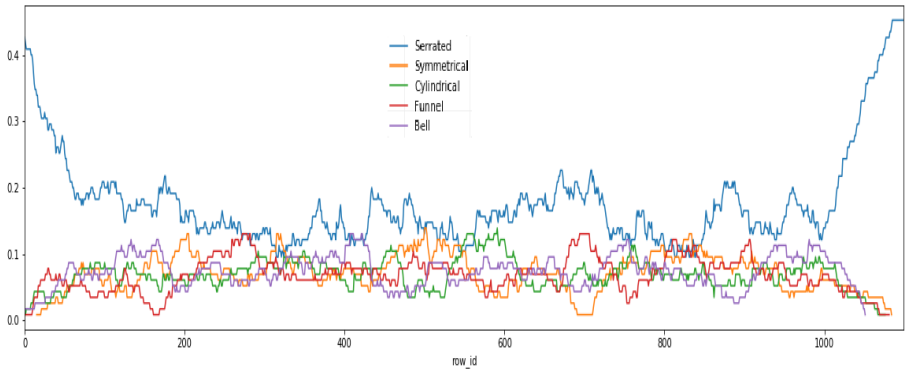

Fig. 15: Probabilidad de label a lo largo de la profundidad del subsampleo

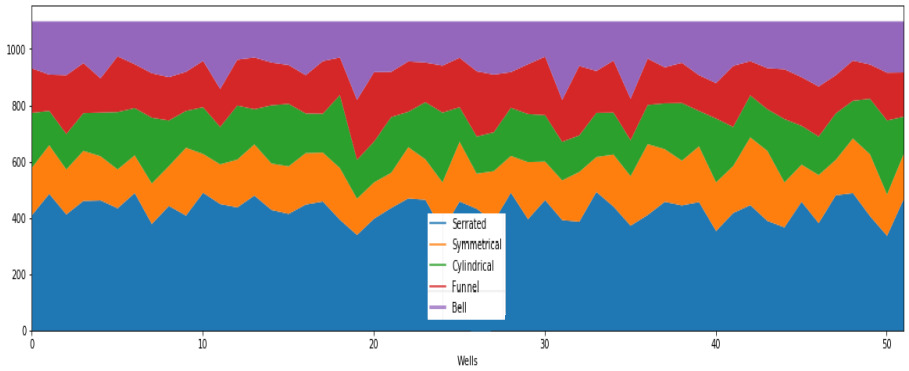

Fig. 16: Distribución de label en pozos verticales del subsampleo

5 Predicción

Es importante destacar que para este problema no se puede particionar el dataset de forma estratificada dado que se estarían particionando las series de datos y contando con información privilegiada a la hora de entrenar-testear. Se realizo un 5-Fold Cross validation sobre el dataset de subsampleo con 52 pozos para calibrar los hiperparametros del modelo con una optimización bayesiana[2] de 100 iteraciones teniendo como objetivo mejorar la accuracy.

Luego se realizó un entrenamiento con 2748 pozos y se evaluó sobre los 1200 pozos restantes.

[2] https://papers.nips.cc/paper/4522-practical-bayesian-optimization-of-machine-learning-algorithms.pdf

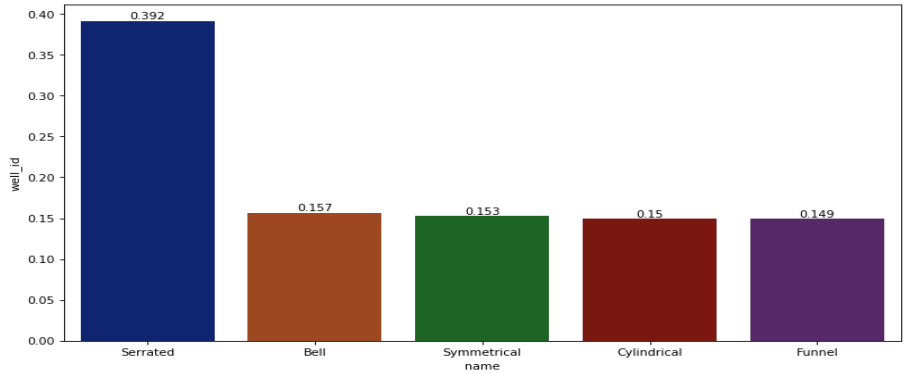

Fig. 17: Nueva proporción de labels del subsampleo

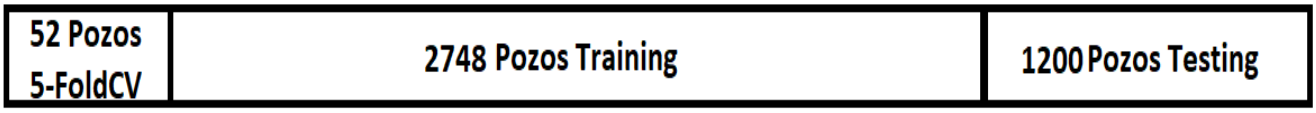

Fig. 18: Particiones dentro del Dataset

Se tomaron los siguientes hiperparametros para el modelo. bootstrap=True, criterion='gini', max_depth=156, max_features=0.4 y n_estimators=288

6 Resultados

Los resultados finales se obtuvieron de aplicar el modelo generado sobre los 1200 pozos con un total de 1320000 registros:

	precision	recall	f1-score	support
Serrated	0.954	0.979	0.966	697661.000
Symmetrical	0.829	0.844	0.837	155980.000
Cylidrincal	0.920	0.922	0.921	158133.000
Funnel	0.885	0.820	0.851	154791.000
Bell	0.902	0.844	0.872	153435.000
accuracy	0.922	0.922	0.922	0.922
macro avg	0.898	0.882	0.890	1320000.000
weighted avg	0.921	0.922	0.921	1320000.000

Tabla 2: Scores obtenidos

La métrica Precisión nos da información de cuantos elementos de los que se predijeron de una clase eran de verdad de esa clase, mientras que la métrica Recall nos dice cuántos elementos de los que había en una clase pudimos predecir bien. La F1 Score nos permite obtener un balance de la métrica Precisión y Recall y suele ser una buena métrica estándar para modelos predictivos. Para definir el mejor score para este caso se debería profundizar el problema técnico y entender que sería más importante. En nuestro análisis se tomó el Score accuracy como objetivo a mejorar para tener una idea general del poder predictivo del algoritmo pero se recomienda utilizar el F1 Score como una posible mejora. El support de la tabla 2 es la cantidad de registros por clase que hay en Test y muestra que las proporciones de clases que había en el dataset completo se mantuvieron en Test.

La clase mayoritaria Serrated y la clase Cylidrincal obtuvieron altos valores para todas las métricas. Estas 2 clases se comportan de forma muy similar, ambas tienen una orientación horizontal pero poseen distintos valores medios por lo que es fácil identificarlas. En particular en la clase Cylindrical tanto para su Precisión, Recall y F1Score tienen valores casi idénticos con un valor aproximado de 0.921. Las clases Symmetrical sufrió mayores inconvenientes para identificarse posiblemente dado a que su tendencia cuadrática es similar a la combinación de las clases Funnel y Bell. El Score accuracy general de la predicción es de 0.922.

	Serrated	Symmetrical	Cylidrincal	Funnel	Bell
Serrated	682775	3211	13	5932	5730
Symmetrical	7511	131673	6607	6487	3702
Cylidrincal	18	7594	145845	2188	2488
Funnel	13100	9900	2810	126875	2106
Bell	12520	6395	3201	1799	129520

Las clases Serrated y Cylidrincal tienen altos número de verdaderos positivos relativos a su presencia y es muy poco probable de confundirse estas labels entre si. Se puede agrupar también la clase Bell y Funnel por tener errores similares por su similitud en cuanto a sus tendencias lineales.

	Serrated	Symmetrical	Cylidrincal	Funnel	Bell
Serrated	0.978	0.005	0.000	0.008	0.008
Symmetrical	0.048	0.844	0.042	0.042	0.024
Cylidrincal	0.000	0.048	0.922	0.014	0.016
Funnel	0.083	0.063	0.018	0.819	0.014
Bell	0.081	0.041	0.002	0.021	0.844

Tabla 4: Matriz de confusión normalizada por label verdaderas.

7 Conclusiones

El Score obtenido es ampliamente superior a la alternativa baseline de clasificar todos los registros como Serrated, que hubiéramos obtenido solamente un Score del 52%. El algoritmo usado es relativamente simple y no se destaca por su aplicación en series de datos a menos que se ponga mucho énfasis en la creación de nuevas variables.

Durante el ajuste del modelo se tuve en consideración seleccionar tanto los mejores hiperparametros del algoritmo random forest como los mejores parámetros que definen la creación de nuevas features. Se noto que una buena selección de estos últimos tiene un mayor impacto en el resultado final de la predicción pero al ser un espacio de búsqueda tan extenso es difícil encontrar los absolutamente mejores. Por ejemplo podríamos haber utilizado una regresión cuadrática también robusta que tenga en cuenta la presencia outliers o haber incorporado una métrica sobre la calidad de las regresiones locales, como R2 o BIC, para incorporar información al algoritmo Random Forest de cuán bien se ajustan los modelos locales en cada registro.

Particularmente el tamaño de las ventanas de datos usado tanto para los modelos regresivos locales como para los resúmenes de variables estadísticos de media y varianza resultó ser un parámetro de gran impacto. Se pudo utilizar el análisis exploratorio para acotar el espacio de búsqueda de estas ventanas y utilizar tamaños de ventana similares a la longitud promedio de los intervalos de las labels pero como el random forest descarta las features sin poder predictivo, se podrían haber probado ventanas de tamaños más variados.

El desfase de datos posteriores y anteriores también tiene un papel muy importante en el ajuste, al contar con toda la medición del pozo se podría haber extendido mucho más para obtener más

información del contexto en el que se encuentra cada registro. Para este proceso no solo se limitó enormemente el rango del desfase sino que se tomaron valores que se saltaban cada un registro para reducir el poder de cómputo necesario.

Otras decisiones fueron más específicas y de menor impacto, como la utilización de la función de pesos Barthann en los resúmenes de variables estadísticas locales. No se vieron grandes cambios en la performance adicionando esta característica. Tal vez estos pesos cumplirían un rol más importante si se tomaron ventanas de datos más extensas.

8 Futuros análisis

Sería interesante plantearse si el orden de clasificación impacta en la certeza de la predicción de los registros vecinos, y por lo tanto, en la performance final. Si un registro se clasificaría primero, podría pasarle la información de su label a los registros cercanos otorgándoles cierta ventaja y mayor certidumbre en el momento de determinar su propia label. De esta forma explotaríamos la información de las distintas combinaciones de intervalos de labels, por ejemplo si después del intervalo de una label 1 la sigue usualmente un intervalo de la label 2 en una cantidad de casos relevantes estadísticamente o podríamos usar la caracteristica de continuidad que hay dentro de un mismo intervalo de label.

Para definir que labels se deberían clasificar primero se sugiere basarse en los que tengan mayor probabilidad de clasificación. Esta probabilidad, en el caso del Random Forest, proviene de la cantidad de votos que recibió por los distintos arboles de decisión. Una vez que los registros con altas probabilidad se hayan clasificado, usaríamos estos valores de label como features en los vecinos.

El desafió de este algoritmo es poder entrenar con suficientes casos que representen los estados que nos encontraríamos en la evaluación de test. Estos casos deberían mostrar situaciones donde a veces contaríamos con ninguna, algunas o todas las labels de los vecinos cercanos. Al mismo tiempo las labels vecinas que pasaremos como features, tienen que ser las labels de casos que hubieran tenido una alta probabilidad de ser clasificados de esa forma para simular correctamente el proceso. Dado que en este dataset contamos con la serie de datos completa del pozo y una abundancia de datos este algoritmo podría llegar a ser viable.

Otra forma posible de encarar este problema de predicción es mediante otros modelos más

populares en Series de Tiempo como redes RNN bidireccional[3] con la aplicación de LSTM para considerar características futuras y pasadas dentro del modelo en vez de la integración de Features históricas.

Un enfoque completamente distintos a las soluciones ya planteadas es concentrarse en predecir los puntos de "cambio de label" para predecir y segmentar únicamente las labels en intervalos. En una segunda etapa se puede usar el dataset de training para condensar las propiedades de cada intervalo de label en una solo curva característica por label. Y finalmente aplicar una métrica de similitud de series de tiempo como Dynamic Time Warping [4] entre los intervalos predichos y las 5 curvas características que obtuvimos previamente y clasificar el intervalo con la curva de mayor puntaje. Las labels en sí presentan muchas diferencias por lo que no debería haber muchos problemas en categorizar los intervalos por labels si se tiene su extracción. El desafió de este camino está en poder segmentar correctamente los registros por los puntos de cambio de label.

[3] https://link.springer.com/chapter/10.1007/11550907_126

[4] https://www.aaai.org/Papers/Workshops/1994/WS-94-03/WS94-03-031.pdf