# 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 : Ricardo Maronna**


**Fecha :**

# 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 trabajara 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 irradiación de los rayos gammas producida por el decaimiento de los isótopos del potasio presente en rocas sedimentarias de grano muy fino como la arcilla. 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 row_id y segmentación y comparación de labels. Se utilizara el algoritmo Random Forest por su simpleza y se pondrá el foco de la investigación en encontrar las correctas features que deberíamos generar para obtener un buen resultado.

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

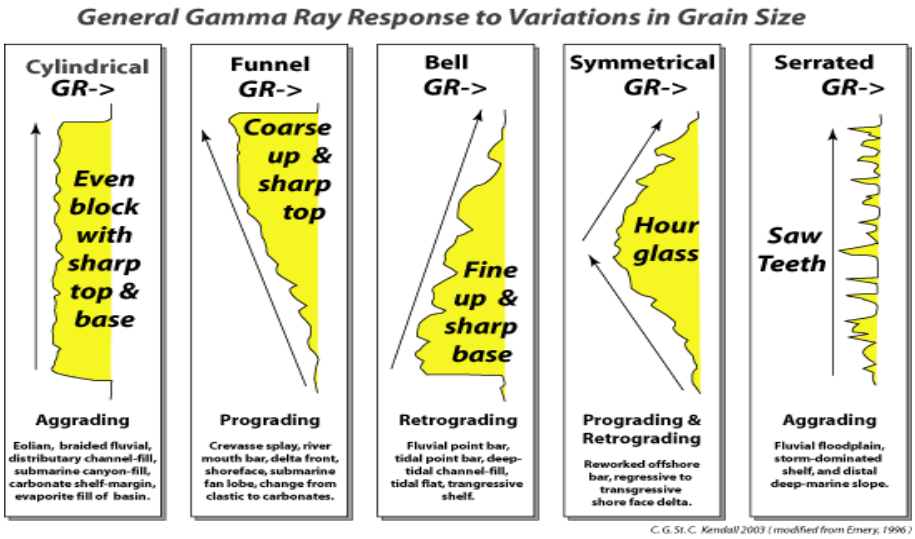


Fig. 1: Clasificaciones de estructuras arcillosas

## 2 Análisis exploratorio de los datos

El Dataset utilizado cuenta con 4.400.000 número de registros que ilustran la medición GR de 400 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 realizo 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 dataset.
- **well_id:** Número entero que identifica cada pozo.
- **GR:** Número real con la medición de la herramienta. Este valor genera la serie temporal 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 mas 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.

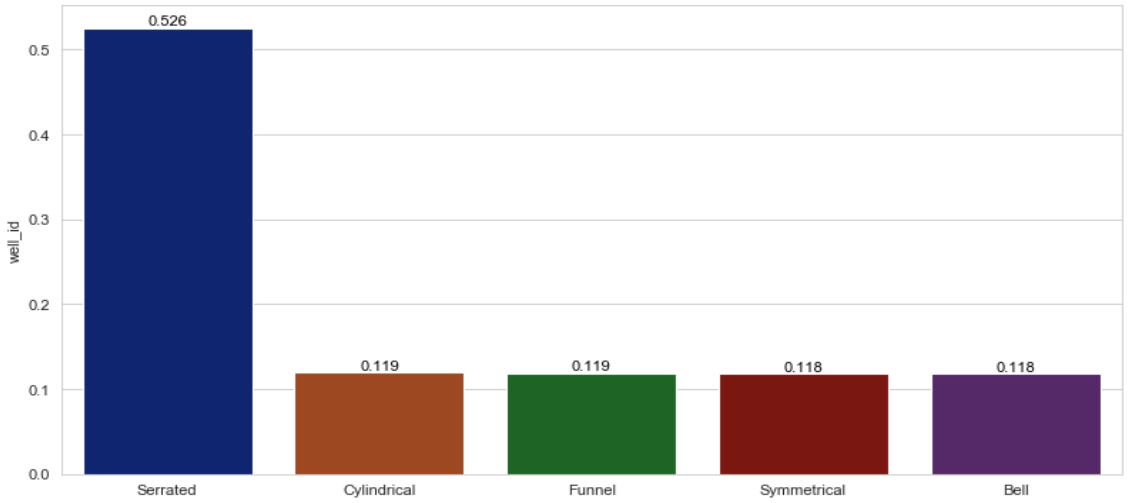


Fig. 2: Proporción de clases

## 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 2.

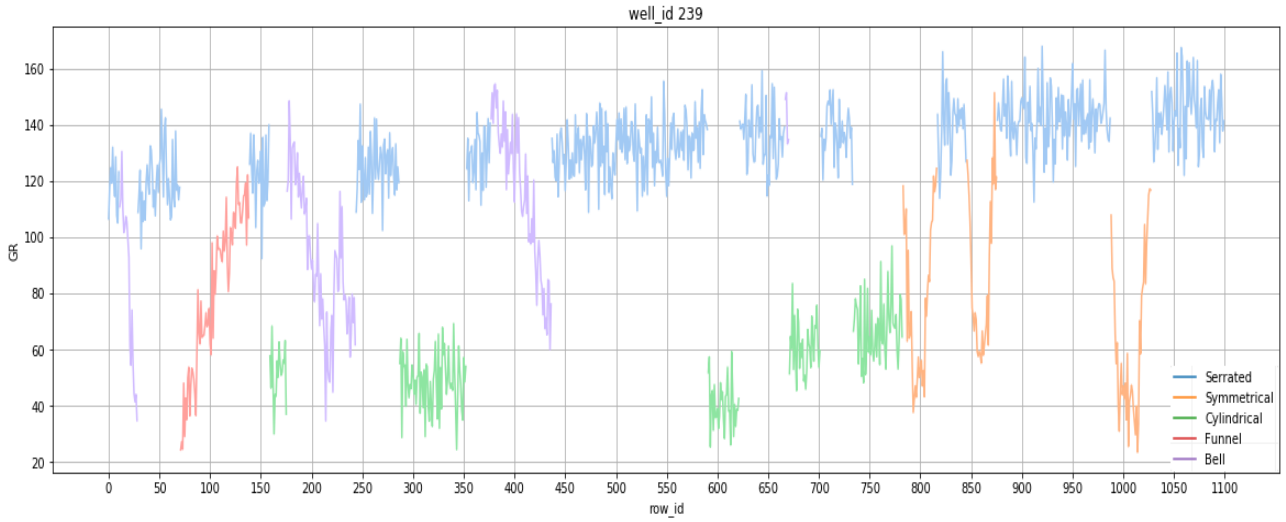


Fig. 3: Pozo ejemplo: well_id 239

A simple viste 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 la Bell del punto 370 vemos que hay una clase Funnel interconectando la misma clase Serrated de 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 todas 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 3.

En la Figura 5 ubicamos todos los pozos unos a lado del otro verticalmente y ordenamos los registros por label comprobamos la presencia mayoritaria de la clase Serrated en todos los pozos y


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

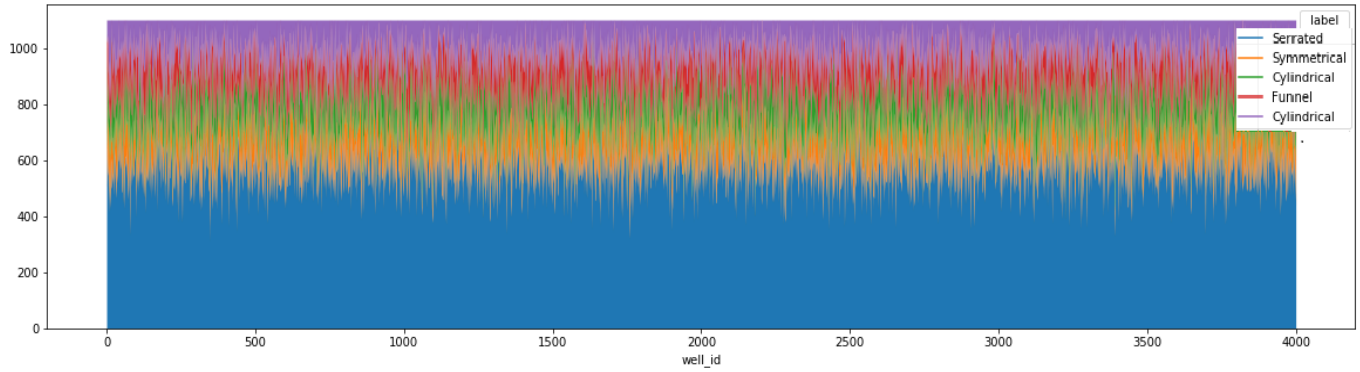


Fig. 5: Comparación de labels vertical entre los pozos

una proporción similar de las clases minoritarias entre si. Tener en cuenta que en la figura 5 los pozos se graficaron uno alado del otro según su well_id.

## 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

En la figura 7 se puede apreciar que efectivamente 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

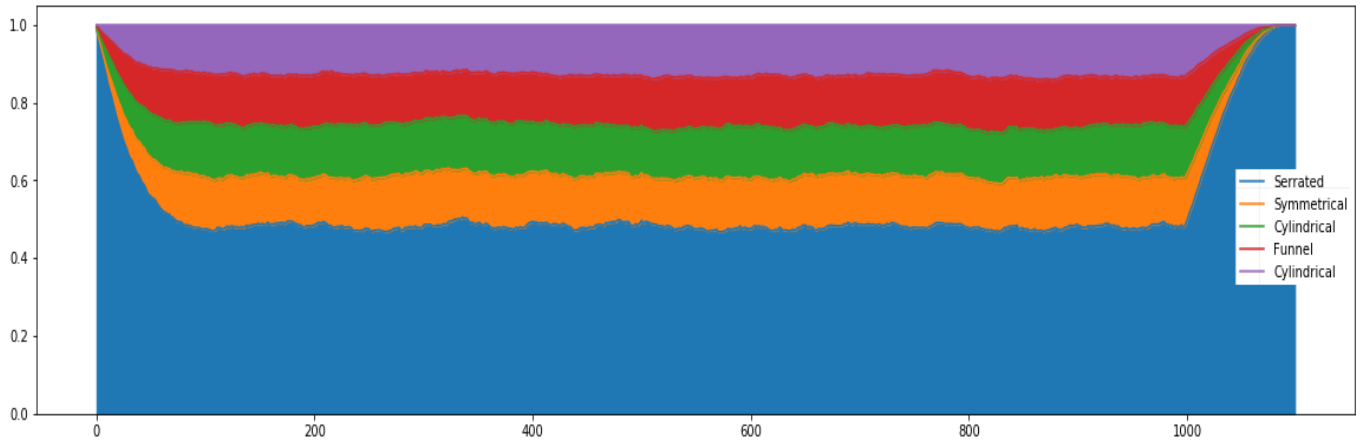


Fig. 6: Gráficos de Áreas labels por row_id

de hacer la predicción.

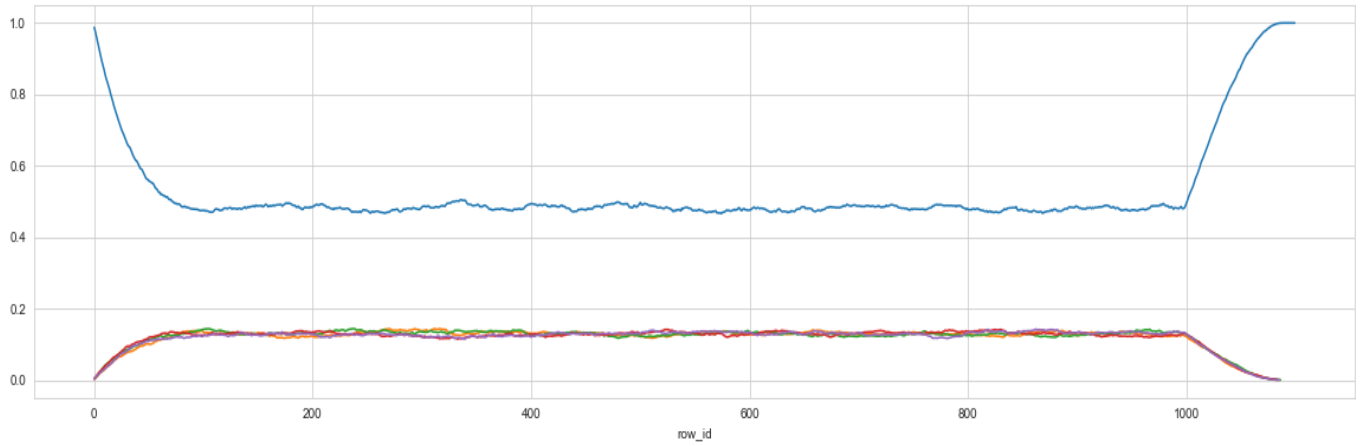


Fig. 7: Gráficos de probabilidades labels por row_id

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.Se observa que en este tramo la probabilidad de la clase Serrated es menor al 50% que se calculo en las proporciones totales de label, esto se debe al efecto borde 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 pertenecía a cada pozo, extraemos los intervalos de cada label y realizamos un análisis univariado con las 82.296 longitudes de cada uno por 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. Por otro lado la clase mayoritaria tiene una tendencia asimétrica por

mientras que la la clase mayoritaria se distingue por presentar longitudes mayores.

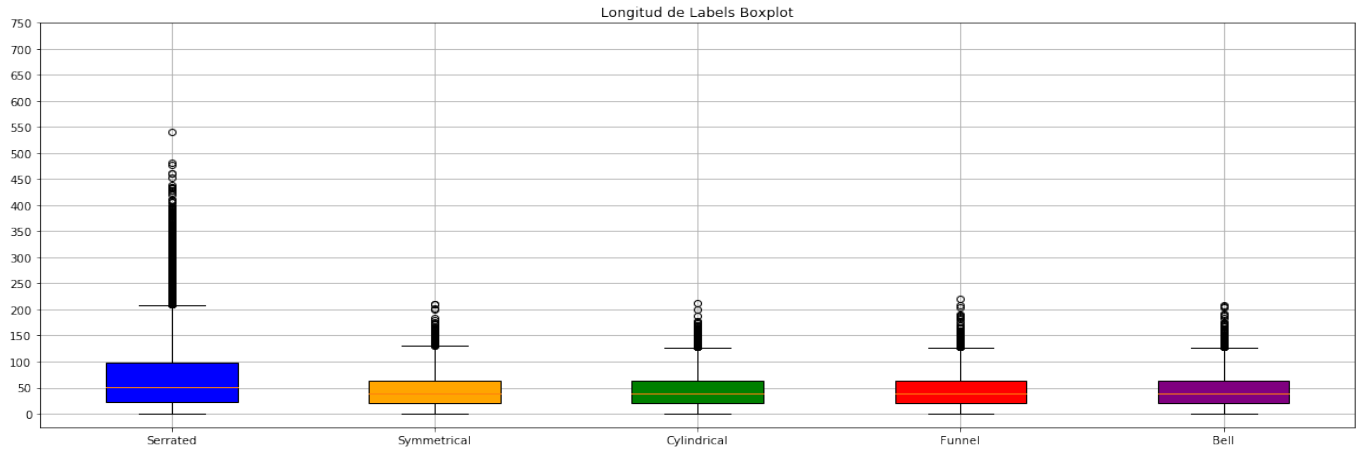


Fig. 8: Boxplot de longitudes de intervalo por label

| 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 importa 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 esta 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 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.

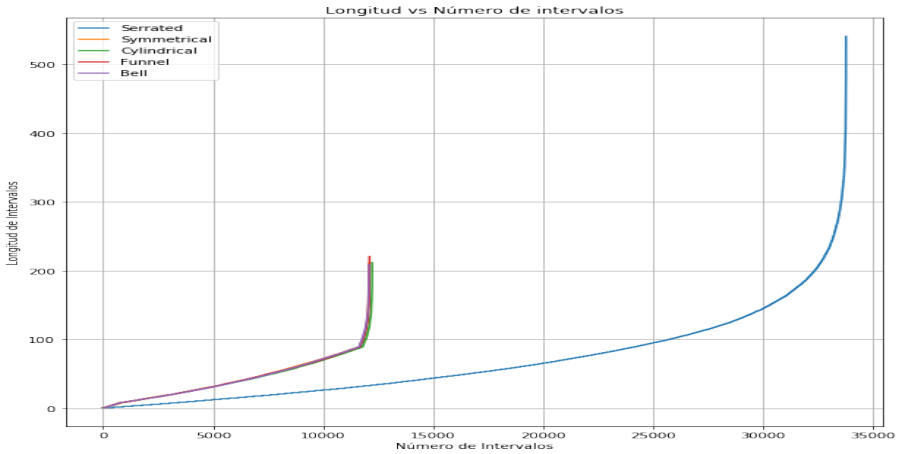


Fig. 9: Ranking ordenado de longitudes de intervalo por label.

En la Figura 10 muestreamos algunos intervalos de longitud similar a la media para visualizar el comportamiento normal de estos y confirmar el análisis técnico inicial.

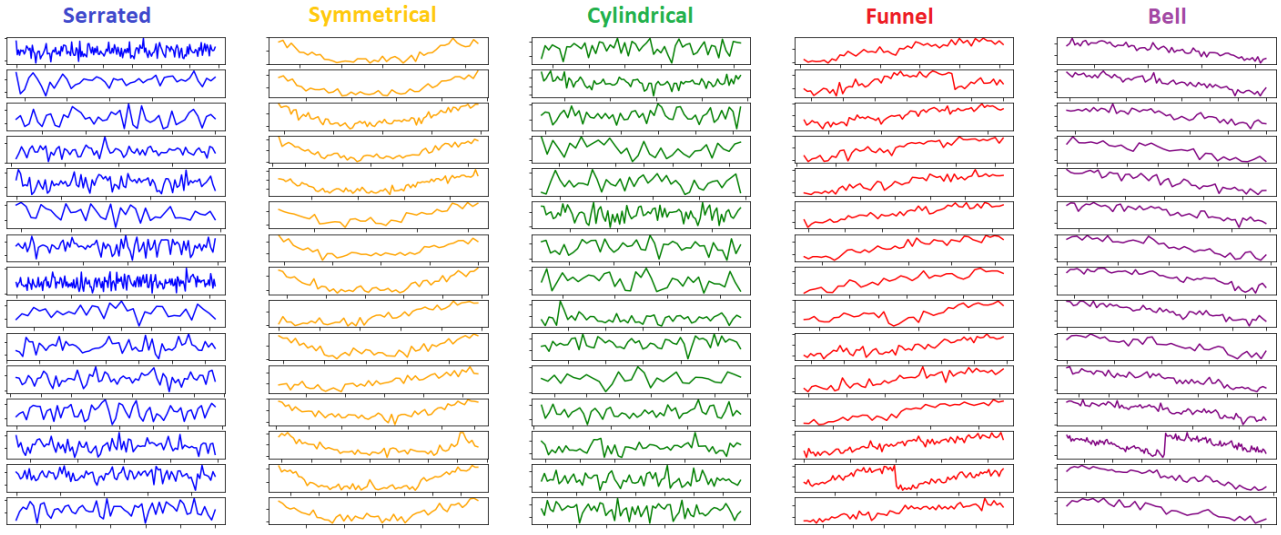


Fig. 10: Extracción al azar de varios segmentos de longitud superior a la media.

Pero por otro lado si elegimos segmentos de longitud muy baja entre 1 y 10 registros en la Figura

11 no se pueden distinguir ningún tipo de tendencia.

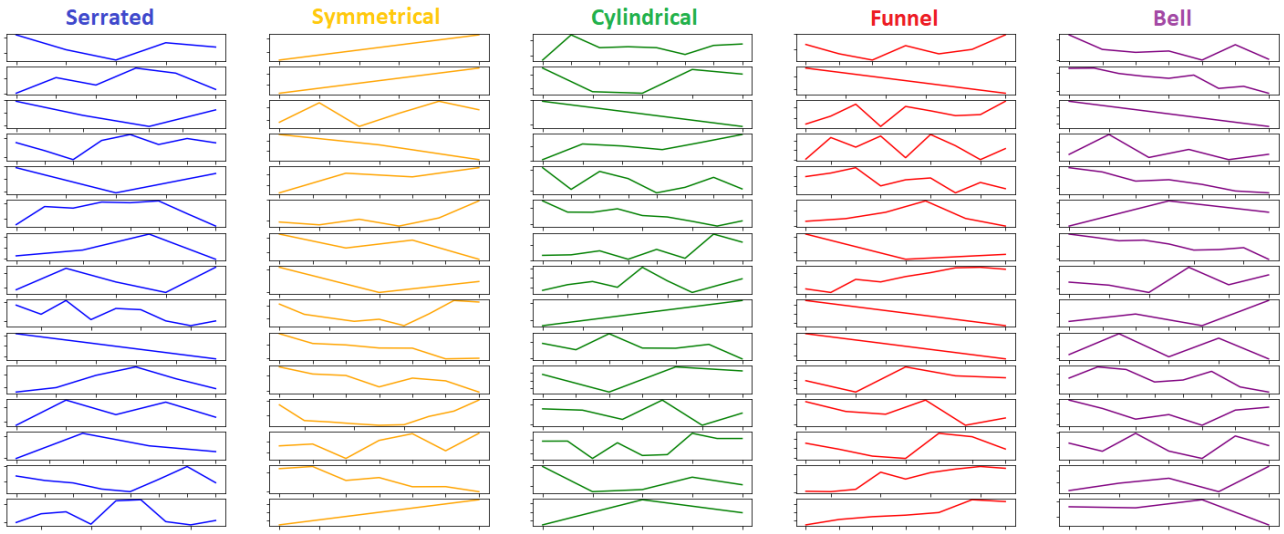


Fig. 11: Comparación de labels vertical entre los pozos

## 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 tiempo debemos crear variables temporales que le den a una registro información suficiente 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 el registro completo de valores.

Inicialmente se normalizo los valores de GR entre máximos y mínimos de todos los pozos dado a que se notaron desvíos generalizados entre pozo y pozo que podrían significar diferencias en las calibraciones en la herramienta, por lo que se reemplazo la variable GR por **GR_mnorm**

Por las características de las clases que tienen alta semejanza a patrones lineales cuadráticos en caso de la clase Symmetrical se decidió tomar para cada registro una ventana centrada en si mismo de 40 registros de longitud y aplicar un modelo de regresión lineal con el método Huber para evitar la interferencia por outliers y un modelo cuadrático. De estos modelos se extrajo los coeficientes y se utilizaron como features. Además se tomó 3 ventanas de medias de 20 , 30 y 40 para calcular la media y 2 ventanas de 15 y 30 para calcular la variaciones próxima al registro. Se tomaron estos valores de ventanas 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. En Resumen obtenemos las siguientes nuevas features como ventanas:

- **0huber:** Parámetro bias de la regresión lineal realizada en ventanas centradas de 35 registros de longitud.
- **1huber:** Parámetro pendiente de la regresión lineal realizada en ventanas centradas de 35 registros de longitud.
- **0polyfit:** Primer parámetro de la regresión cuadrática realizada en ventanas centradas de 35 registros de longitud.
- **1polyfit:** Segundo parámetro de la regresión cuadrática realizada en ventanas centradas de 35 registros de longitud.
- **2polyfit:** Tercer parámetro de la regresión cuadrática realizada en ventanas centradas de 35 registros de longitud.
- **20rolling_mean:** Promedio de una ventana centrada de 20 ponderando con una función de barthann (Distribución normal ancha)
- **30rolling_mean:** Promedio de una ventana centrada de 30 ponderando con una función de barthann (Distribución normal ancha)
- **40rolling_mean:** Promedio de una ventana centrada de 40 ponderando con una función de barthann (Distribución normal ancha)
- **20rolling_std:** Variación de una ventana centrada de 20 ponderando con una función de barthann (Distribución normal ancha)
- **40rolling_std:** Variación de una ventana centrada de 40 ponderando con una función de barthann (Distribución normal ancha)

Además se agrego 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 salteando de 2 en 2. De esta forma un registro ve valores desfasados en una ventana de 40 registros alrededor suyo para estas features.

Finalmente contamos con 60 nuevas features de realizar el desfasaje, 10 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 imitar las características principales del dataset original con respecto a la proporción de clases presente en cada pozo. Por eso aplicamos una regla para quedarnos con los pozos que tienen al menos un 40% de la clase mayoritaria y al menos un 11% de cada clase minoritaria. Conseguimos un dataset que tiene 115 pozos, aproximadamente un 3% de los datos originales, con 126500 registros.

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

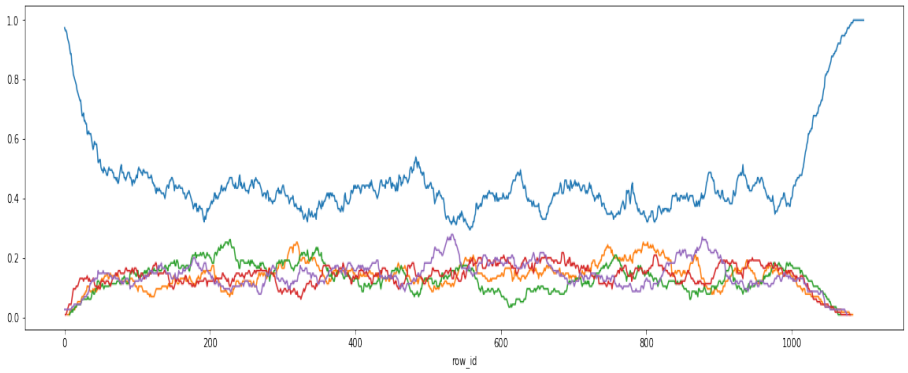


Fig. 12: Ranking ordenado de longitudes de intervalo por label

## 5 Predicción

Para realizar la selección del modelo se tomo un subsampling de la muestra general para optimizar la calibración de los hiperparametros del Random Forest. Se tomo solo un 30% de los registros y se procedió a hacer un Cross Validation para el cual la mayoría de los hiperparametros se ajustaron a manualmente salvo por la máxima profundidad del árbol que se eligió utilizando un random search grid. Es importante destacar que para partir el dataset en train y test no se puede hacerlo de forma

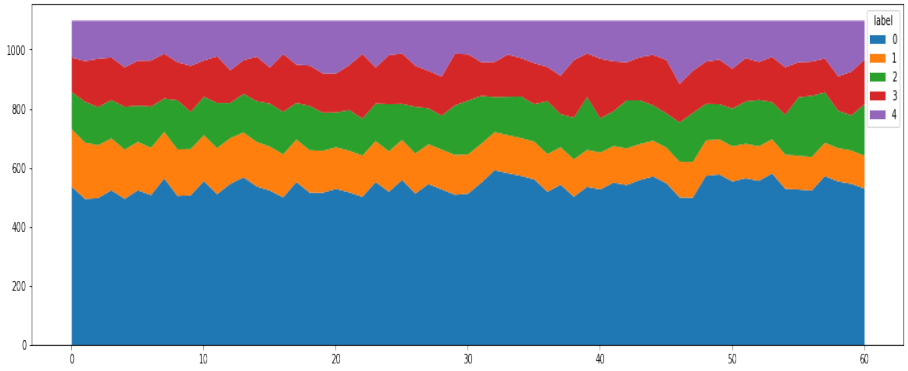


Fig. 13: Ranking ordenado de longitudes de intervalo por label

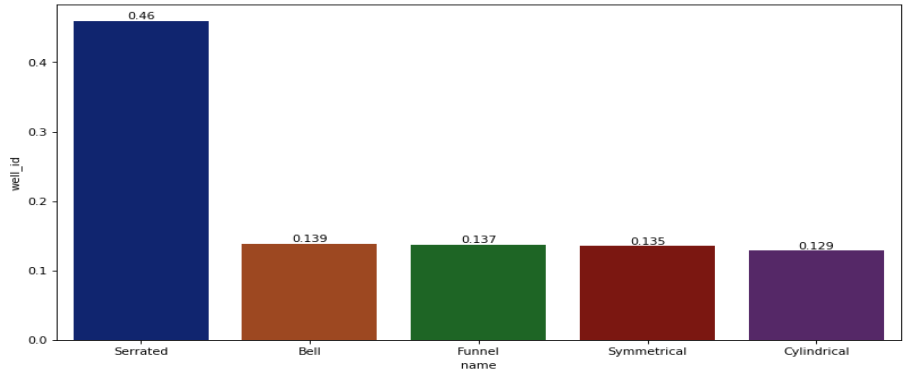


Fig. 14: Ranking ordenado de longitudes de intervalo por label

estratificada dado que uno no puede considerar cada pozo como una bolsa de datos sino que se debe entrenar y testear con pozos enteros sin particionar los datos de los pozos.

Se tomaron los siguientes hiperparametros para el modelo. bootstrap=True

criterion='gini'

max_depth=150

max_features='auto'

min_samples_leaf=2

min_samples_split=10

n_estimators=650

# 6 Resultados

En el entrenamiento final se realizo un train-test-split 70/30 con la totalidad de los datos y se obtuvieron los siguientes resultados:

| | 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 |

Se destaca que 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 sin mucho ruido pero poseen distintos valores medios por lo que es fácil identificarlas.

El Score accuracy general de la predicción es de 0.922.

Matriz de Confusión:

| | 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 |

Nuevamente se destaca las clases Serrated y Cylidrincal por tener alto numero relativo de verdaderos positivos y el numero de veces que confunde una clase con la otra es muy bajo.