Mentoría Data Science aplicado a BCI

Diplomatura en Ciencia de Datos, Aprendizaje Automático y sus Aplicaciones

Edición 2021

Trabajo práctico 2 Análisis y Curación de datos

Integrantes 'Grupo 2'

- Iberra Yanina
- Junco Luis
- Wolfman Gabriel

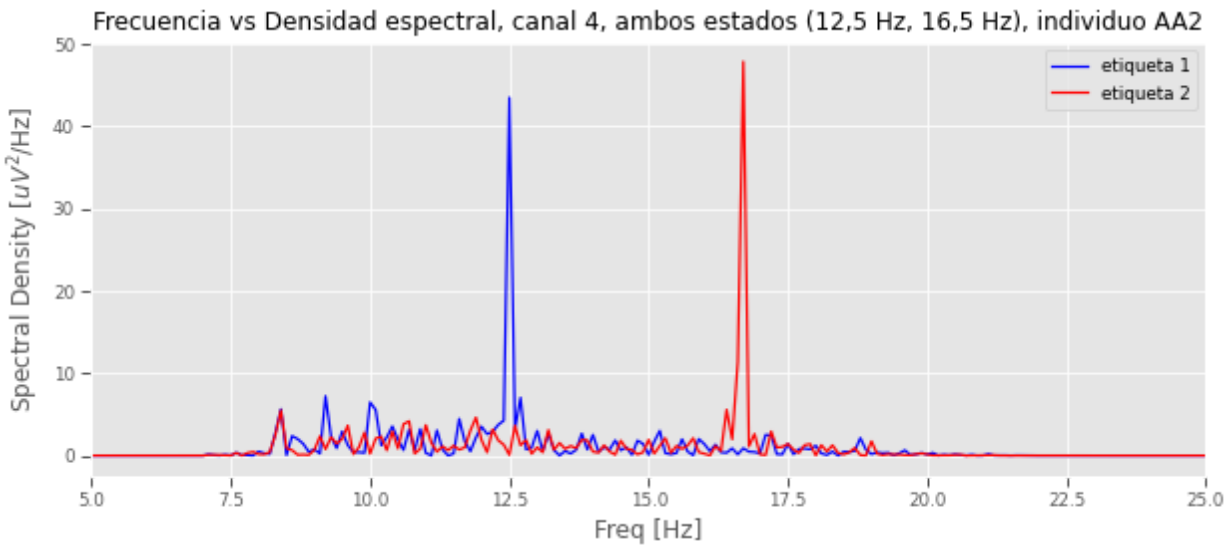
A) Nivel Segmento/Estado: Seleccione los datos correspondientes a un paciente y un canal, y para él defina un conjunto de datos para cada estado presente en el dataset. Para cada uno de ellos estudie los siguientes elementos y luego compárelos.

a) A partir de la frecuencia de muestreo previamente determinada y usando el teorema de Nyquist (f NQ = f o /2). ¿Qué frecuencia máxima pudo ser registrada en los datos disponibles?

La frecuencia máxima que pudo ser registrada con los datos disponibles es 100 Hz ya que es 200Hz/2 por el teorema de nyquist.

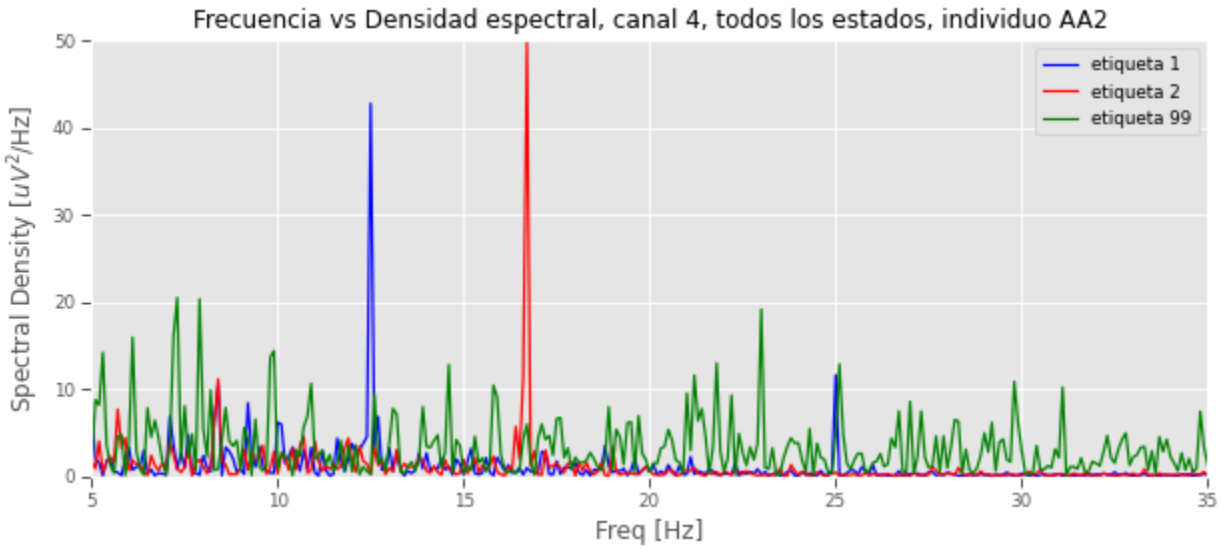
b) Compute y grafique la densidad espectral de potencia para cada segmento.
Elija las escalas apropiadas para su visualización (qué magnitudes están graficando? corresponde el uso de decibeles?).

Se está graficando la frecuencia en el eje x como variable independiente y la densidad espectral en el eje y como variable dependiente refiriéndose a la energía presente en cada frecuencia.



En este caso filtramos la señal de 50 Hz y la continúa. No hace falta el uso de decibeles en las visualizaciones ya que los filtros van a retirar las frecuencias que nos generan ruido. En los casos donde se quiera mantener estas frecuencias más preponderantes corresponde el uso de decibeles para una correcta visualización de los datos.

c) Determine la/las frecuencias con mayor presencia en cada caso. ¿Que puede concluir con esta nueva información? ¿Encuentra diferencias en este aspecto entre los tramos 99 y los tramos 1-2? ¿Hay frecuencias no asociadas con el fenómeno a detectar?



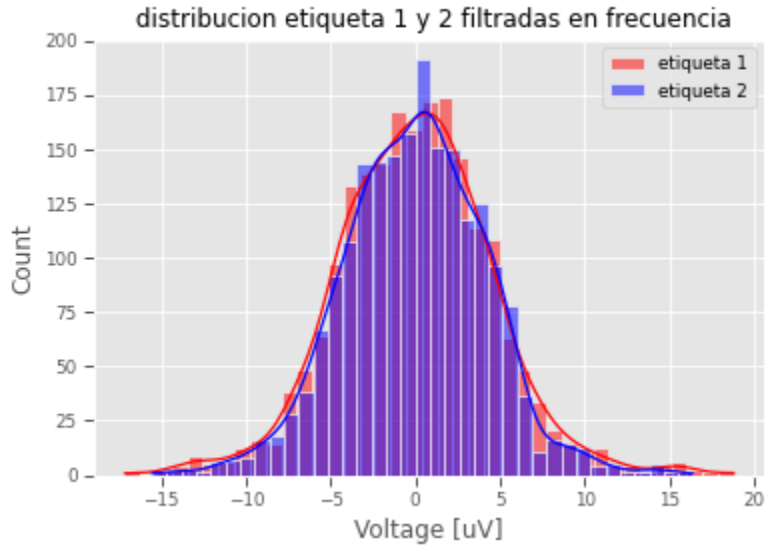
Para realizar dicho gráfico las señales fueron filtradas con un filtro pasa banda Butterworth entre 8 Hz y 20 Hz para representar mejor las frecuencias de interés, eliminando todo ruido que pueda llegar a tener como el de la continúa a 0 Hz y el de la transmisión de luz a 50 HZ.

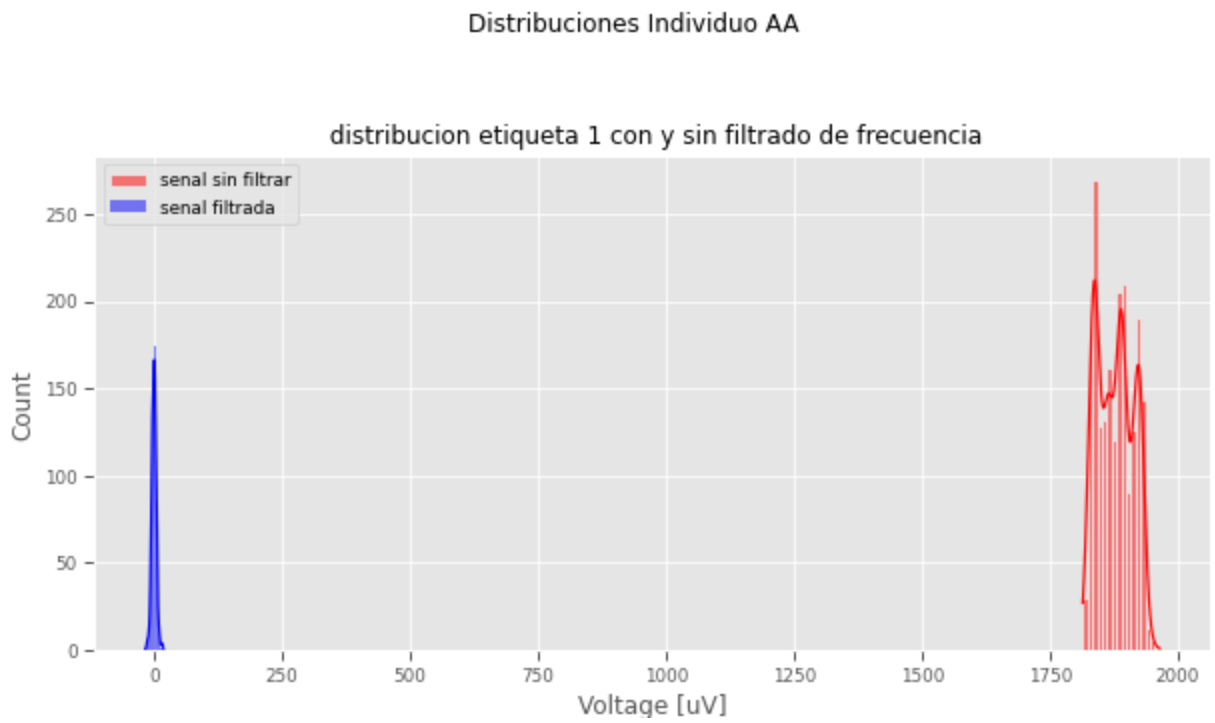
En estos casos hay mayor presencia de las frecuencias en estudio para las etiquetas 1 (12,5 Hz) y para las etiquetas 2 (16,5 Hz). También se pudieron observar en algunos canales la presencia de los armónicos de dichas frecuencias (25 HZ para etiqueta 1 y 33 Hz para etiqueta 2), pero estos no aparecen siempre. A su vez cabe destacar que no en todos los canales se observa la misma amplitud de voltaje en las frecuencias fundamentales. Se visualiza que aparecen amplitudes en frecuencias cercanas a las fundamentales ingresando ruido al sistema por el que fue filtrado posteriormente para una mejor visualización. Para el caso de las etiquetas 99 no se observa una frecuencia fundamental de estudio.

d) ¿Es útil el concepto de outliers o conviene usar otros criterios? ¿Se pueden descartar señales?

En el análisis de frecuencias al realizar el filtrado en las frecuencias de interés no es útil el concepto de outliers ya que primero todos los datos de esas frecuencias son relevantes para el estudio y segundo que si se retiran muestras como outliers puede llegar a producir saltos bruscos en la frecuencia generando ruido en la señal.

e) Comparar en el dominio del tiempo las señales con y sin filtrado de frecuencias indeseadas. ¿Qué diferencias encuentra?





Se observa que la filtrada es muy similar a una gaussiana, y la sin filtrar primero que tiene la componente de continua que hace que los valores están muy por encima de los rangos normales siendo un artefacto de la adquisición y segundo que su distribución tiende a una bimodal.

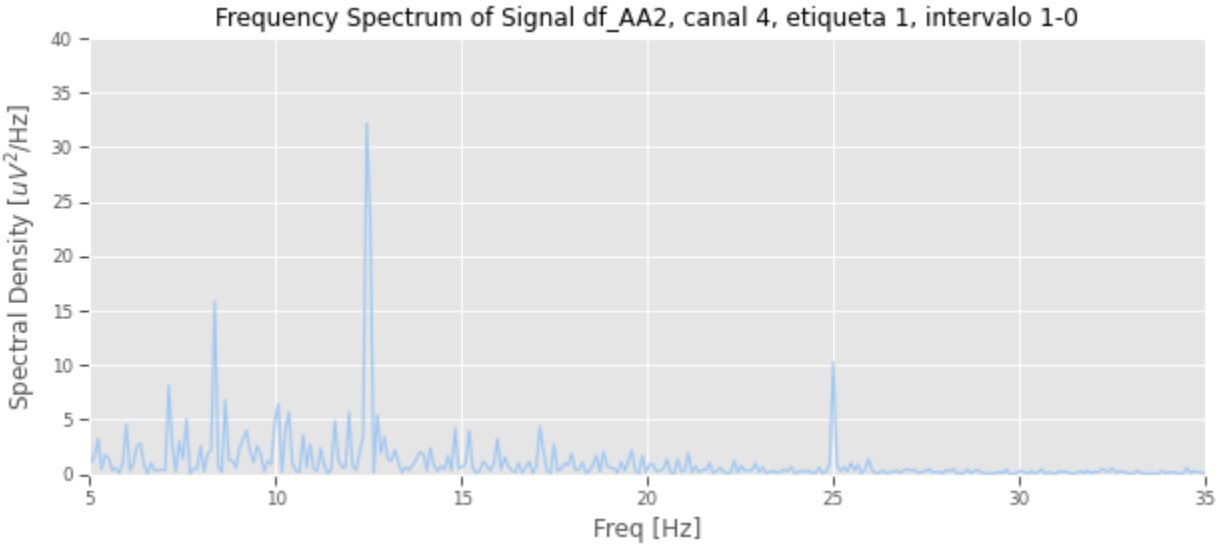
B) Nivel Paciente - un canal: Seleccione los datos correspondientes a un paciente y un canal de adquisición y para ese caso estudie los siguientes elementos:

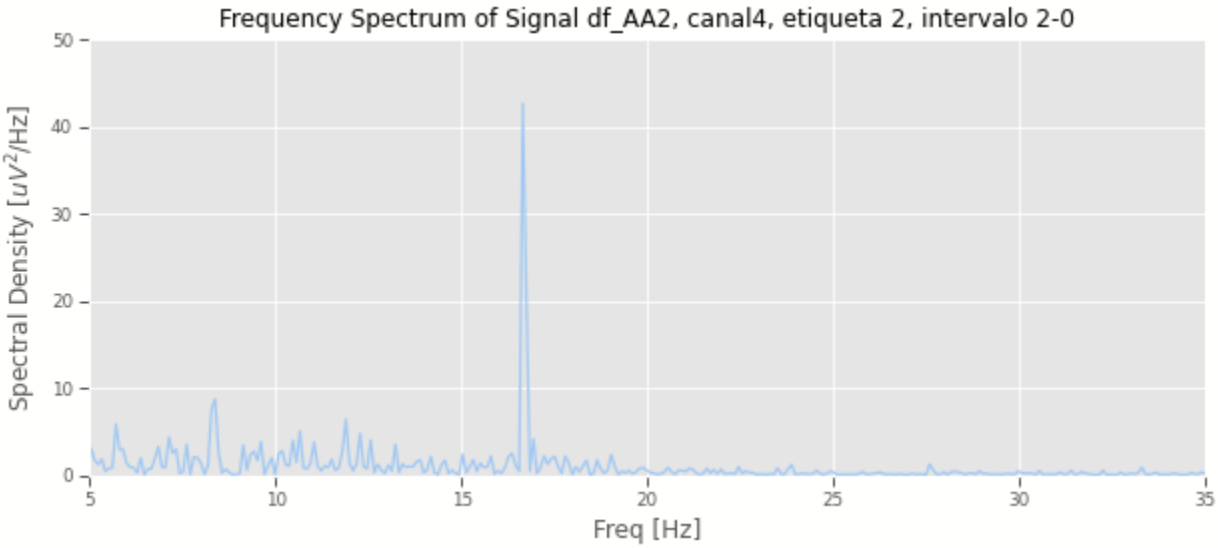
a) Para cada intervalo de adquisición de cada estado, determine la frecuencia de mayor presencia relevante al problema estudiado. Ej: Un canal tiene 10 intervalos 2, para cada uno de ellos la frecuencia máxima oscila alrededor de 12.5.

Se selecciona paciente AA2 canal4.

Luego de realizar la transformada de fourier de todos los intervalos de las señales de un canal, un individuo y graficar frecuencia vs. densidad espectral, observamos que los intervalos correspondientes a las Etiquetas 1 muestran una componente con mayor amplitud en 12,5 Hz y su armónica en 25 Hz, y las de Etiqueta "2" tienen una componente con mayor amplitud en la frecuencia de 16,5 Hz. En ambos casos coinciden con la frecuencia de la fuente de estimulación que se identifica con cada etiqueta.

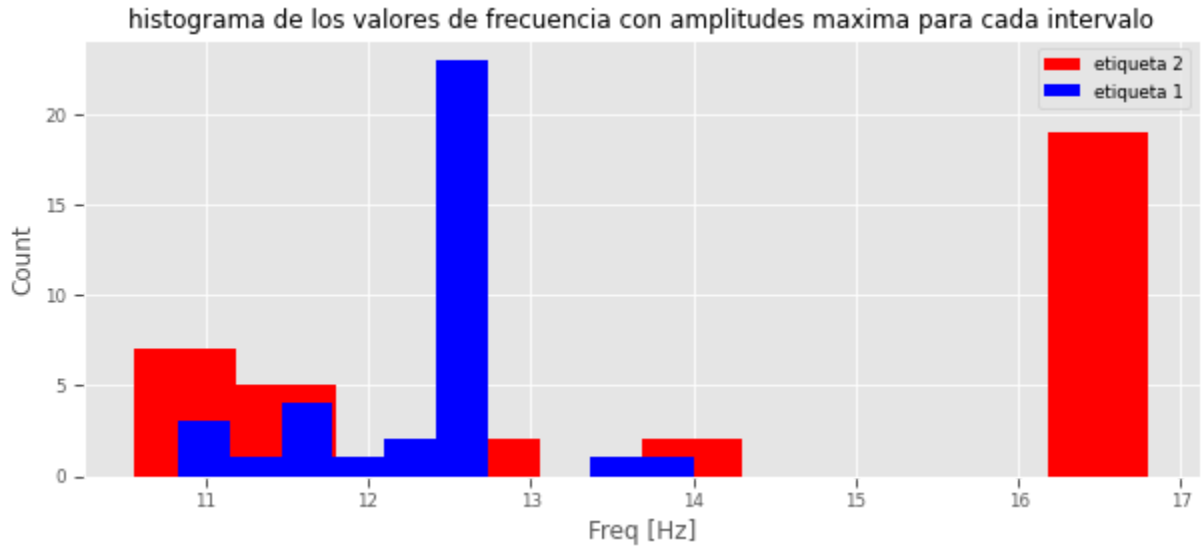
A continuación dos gráficos de densidad espectral de ejemplo:





b) Realice un análisis estadístico de dichas frecuencias a lo largo de todo el conjunto de intervalos para cada clase. Determine su distribución y su resumen estadístico según considere apropiado. Ej: Siguiendo el ejemplo anterior, cómo es la distribución de esas frecuencias alrededor de 12.5 en el intervalo 2 y de la misma manera para los otros.

Para el siguiente análisis estadístico se utilizan los valores de frecuencia donde se encuentra la máxima amplitud de la señal ya filtrada para cada intervalo. Para aumentar los datos se utilizan los 4 canales y así mejorar el análisis.
Se realiza primero un histograma de las dos etiquetas:



Con estos valores se calculó los siguientes estadísticos de posición central y dispersión.
cálculo de la media etiqueta 1= 12.28
cálculo de la mediana etiqueta 1= 12.45
cálculo del desvío estándar etiqueta 1= 0.63
cálculo de la varianza etiqueta 1= 0.4

cálculo de la media etiqueta 2 = 14.32
cálculo de la mediana etiqueta 2= 16.63
cálculo del desvío estándar etiqueta 2= 2.59
cálculo de la varianza etiqueta 2 = 6.7

Podemos ver que para las etiquetas 2 hay ruidos que ingresan en los valores máximos. Esto se da porque los valores de amplitudes no son tan grandes como los valores de etiqueta 1 generando que el ruido llegue a ser el valor máximo. Esto mismo se observa en el histograma como en la mayor varianza que tiene la etiqueta 2.

c) Determine si existe una diferencia estadísticamente significativa, entre los valores centrales de frecuencias para los estados existentes.

Se puede decir que los valores de mediana, se encuentran muy próximos a lo esperado para cada etiqueta. Los valores medios pueden estar influidos por valores atípicos de ruido como se observa en la etiqueta 2. Concluyendo que si hay diferencia estadísticamente significativa entre los distintos estados.
Aunque lo ideal sería realizar un estudio de diferencia de medias, con la poca cantidad de datos que se tiene (se tenía 9 intervalos por canal, por lo que al usar los cuatro canales para el estudio estadístico se obtienen 36 datos) resulta incorrecta la apreciación de dichos test. Por lo que se realiza solamente un análisis visual de los gráficos y de los estadísticos.

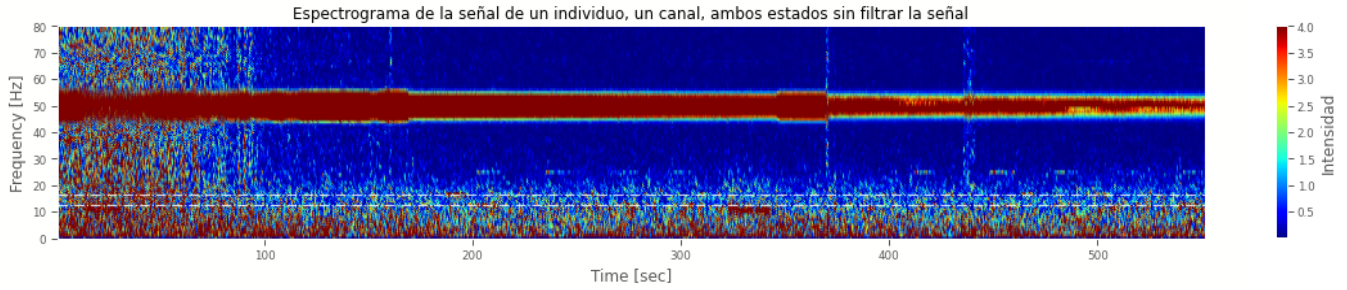
d) ¿Son variables independientes el estado registrado de la señal y la frecuencia mayoritaria presente en el registro? Use herramientas cuantitativas y cualitativas para justificar su respuesta.

Es evidente, de acuerdo a lo analizado hasta ahora, que son variables dependientes el estado (etiqueta) y la frecuencia de mayor presencia en amplitud en cada registro. La señales obtenidas en los registros con estimulación de luz 12,5Hz tienen como resultado una componente principal de frecuencia, o de mayor presencia en 12,5 Hz, lo mismo sucede en el caso que se estimulo con 16,5 Hz obteniendo las mayores amplitudes para la misma.
No fue posible utilizar herramientas estadísticas como matriz de correlación, heatmap o test de independencia de variables debido a la poca cantidad de muestras.

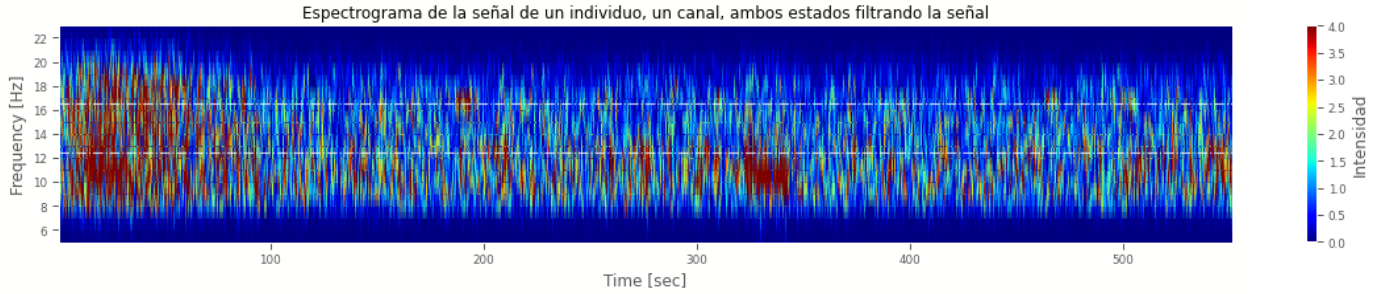
e) Compute y grafique el espectrograma del registro completo. Probar con el registro filtrado y sin filtrar y comentar las diferencias. Ver link informativo, si quedan dudas conceptuales no duden en preguntar.

El espectrograma es un gráfico en cuál se grafica tiempo en el eje independiente y la frecuencia en el eje dependiente. Luego se utiliza los distintos niveles de color para representar las distintas amplitudes de voltage a lo largo del tiempo en cada valor de frecuencia.

Una vez realizados ambos espectrogramas, observamos que en el caso que no filtramos la señal se visualiza con una amplia intensidad a lo largo de todo el registro la componente de 50Hz de transmisión de corriente eléctrica. También se observa con alta intensidad y a lo largo de todo el registro la componente de continua en 0 Hz debido al artefacto en la obtención de los datos.



En el espectrograma de la señal filtrada logran visualizarse ambas componentes de 12,5 Hz y 16,5 Hz (por periodos de 10 segundos). También algo de ruido al comienzo del registro correspondiente a etiqueta 99.



f) Resuma las principales conclusiones de este nivel de análisis.

Como principales conclusiones podemos mencionar:

- Afirmar que existe una relación entre la frecuencia de las fuentes de estimulación y las componentes de mayor amplitud obtenidas en el análisis en frecuencias de las señales del cerebro obtenidas del experimento.

- Que las señales obtenidas en uV en el experimento son muy sensibles a artefactos externos.

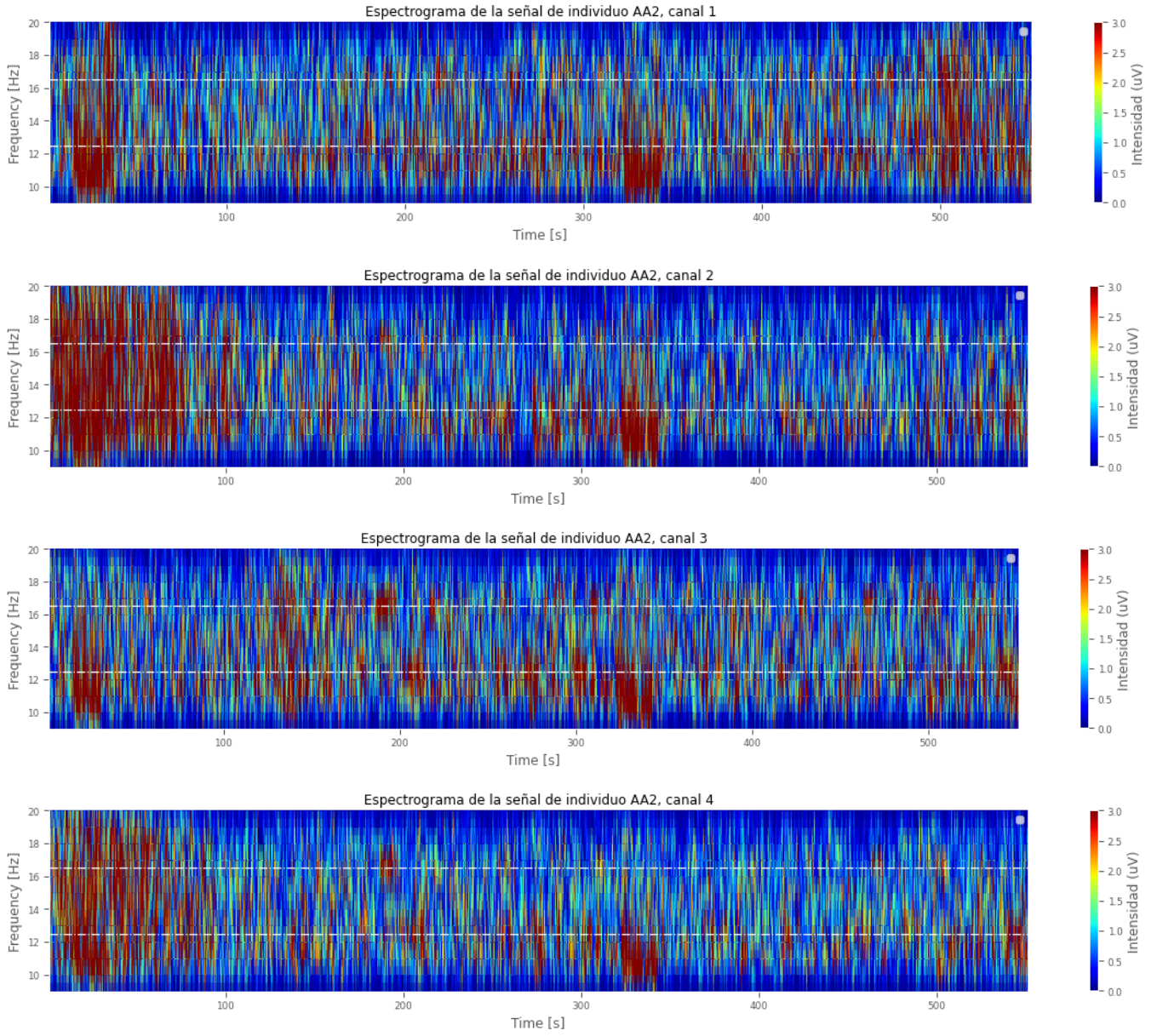
- Un buen método para identificar las componentes en frecuencia con mayor amplitud o intensidad en la señal analizada son: el espectrograma y el gráfico de frecuencia vs. densidad espectral.

- Un método eficiente en el dominio de la frecuencia para limpiar nuestros datos es el filtrado en frecuencia quedándonos con el segmento de frecuencia que nos interesa analizar de acuerdo al conocimiento de dominio.

C) Nivel Paciente - multi canal: Seleccione los datos correspondientes a un paciente y para ese caso estudie los siguientes elementos:

a) Compute el espectrograma para cada canal, y compárelos. ¿A simple vista, ¿Existe alguna correlación?

Se realizaron los espectrogramas del individuo AA tercera sesión para todos los canales.



Se observa que al introducir las señales con todas las etiquetas, se introduce ruido en las mismas en los primeros intervalos de tiempo. Se puede mencionar que a simple puede existir una correlación entre los canales al ser similares en la información brindada.

b) A partir de la respuesta anterior. ¿Considera relevante trabajar con todos los canales disponibles o podría quedarse con un subconjunto?

Se concluye que los canales presentan niveles de información similar entre ellos, por lo que se podría utilizar un solo canal para realizar el consecuente análisis de ciencia de datos.

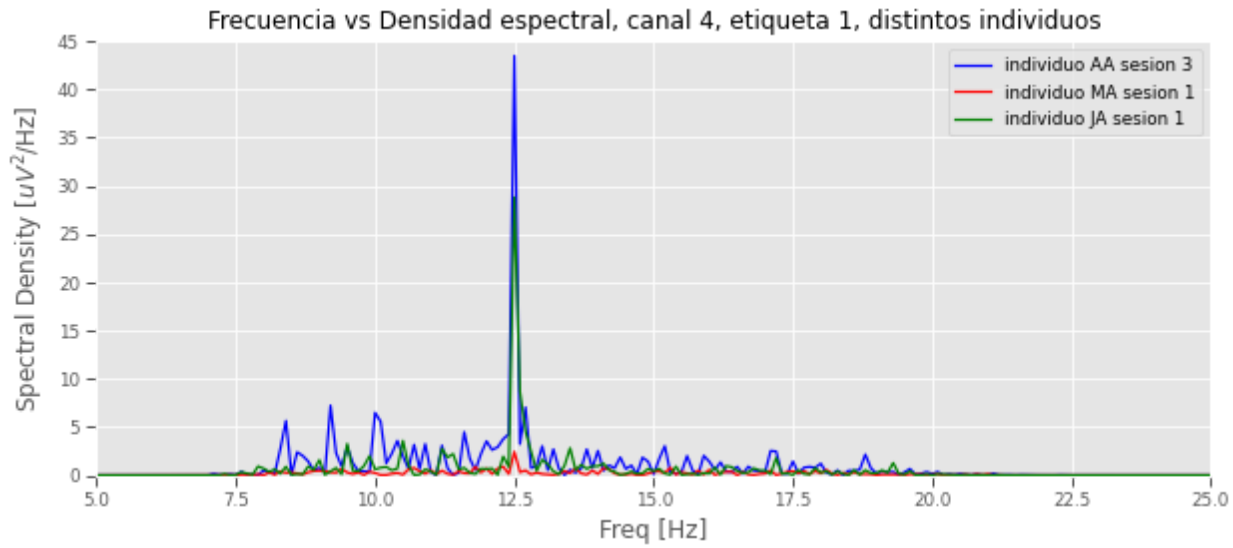
c) Resuma las principales conclusiones de este nivel de análisis.

Como conclusión comentamos nuevamente que se puede utilizar un solo canal para el análisis posterior de ciencia de datos.

D) Nivel Multi-Paciente.

a) A partir de las conclusiones extraídas de los niveles de análisis anteriores. Decida cuáles son los aspectos más importantes a analizar de los registros de un paciente y compárelos entre pacientes. ¿Encuentra diferencias significativas entre pacientes?

Para realizar una comparación entre pacientes se propone tomar un intervalo con el mismo estado y un canal, en distinto paciente. Con eso podemos comparar la amplitud de las señales en cada individuo. Se cree que si se toma todo un intervalo que contenga ambos estados se va a incluir etiquetas 99 donde aparezca mucho ruido para el análisis, por ende se prefiere hacerlo separado.



Se puede observar como varía la amplitud en la frecuencia de estimulación 12,5 Hz para cada paciente. Pero que en los mismos aparece el pico de señal a la frecuencia de estimulación.

E) Comparando con el Trabajo Práctico anterior.

a) ¿Cambia la complejidad del análisis según el dominio en el que se estudien los datos?

Creemos que es más complejo la comprensión de los datos en el dominio de la frecuencia para quien no comprende del tema, pero que una vez entendido este dominio ofrece mucha más información para diferenciar los estados (objetivo final), que en el dominio del tiempo.

b) Pensando en un problema de clasificación supervisada, ¿resulta más cómodo trabajar los datos en uno de los dos dominios?

Por la información brindada en el dominio de la frecuencia creemos que el problema de clasificación supervisada va a ser más eficiente utilizando como entrada los datos resultantes de este análisis. Ya que en el dominio del tiempo le va a ser más dificultoso encontrar diferencias entre los estados.