



# REPORTE

---

## Reconocimiento de Ataques Epilépticos

Bryan Fallas L.



# INTRODUCCIÓN

## DESCRIPCIÓN GENERAL DEL PROBLEMA

El electroencefalograma (EEG) es un análisis que se utiliza para detectar anomalías relacionadas con la actividad eléctrica del cerebro. Este procedimiento realiza un seguimiento de las ondas cerebrales y las registra, mediante la colocación de discos metálicos con cables delgados (electrodos) sobre el cuero cabelludo y se envían señales a una computadora para registrar los resultados. La actividad eléctrica normal del cerebro forma un patrón reconocible. Por medio de un EEG es posible buscar patrones anormales que indiquen convulsiones u otros problemas, permitiendo así el diagnóstico y el control de los trastornos convulsivos.

En la actualidad, un neurólogo es el encargado de leer e interpretar los resultados obtenidos por el electroencefalograma, los cuales varían en complejidad y duración; ya que puede verse alterado por el comportamiento que tenga el paciente durante el examen. Debido a esto, no cabe duda de que el diagnóstico esté altamente ligado al criterio profesional del médico. Con el fin de reducir el error o la influencia de la interpretación humana en el diagnóstico del paciente, se propone realizar un modelo de clasificación que permita determinar si el paciente presenta durante la realización del examen, algún indicio de ataque convulsivo. De este modo, es posible desarrollar un "estándar", con una alta confiabilidad (superior al 85%), en el diagnóstico del paciente.

Este proyecto parte de la hipótesis de que es posible, mediante un modelo de clasificación, determinar si el sujeto se encuentra bajo una actividad convulsiva, esto basado en las lecturas registradas por el EEG; por lo que se tiene como objetivo el desarrollo de un modelo de clasificación que permita obtener un diagnóstico, superior al 85% en confiabilidad, para el paciente, partiendo de los resultados obtenidos en el EEG realizado.

## OBJETIVO DEL PROYECTO

Desarrollar un modelo que permita llevar el cabo el diagnóstico de un EEG con una precisión superior al 85% con el fin de automatizar este proceso de lectura del EEG (los cuales pueden ser muy extensos en ciertos casos) por parte del médico y reducir considerablemente el criterio humano el cual puede variar considerablemente de un médico a otro.

## CONJUNTO DE DATOS

El conjunto de datos para la realización del presente proyecto se obtuvo del Repositorio de Aprendizaje Automático de UCI:

<https://archive.ics.uci.edu/ml/datasets/Epileptic+Seizure+Recognition>

El conjunto de datos con el que se desarrolla el proyecto es una serie temporal que muestrea en 4097 puntos de datos un conjunto de referencia que consta de 5 carpetas diferentes, cada una con 100 archivos, donde cada archivo representa un solo sujeto o paciente. Cada uno de estos archivos del conjunto de datos de referencia es una grabación cerebral durante 23.6 segundos. Cada punto de datos es el valor de la grabación del electroencefalograma (EEG) en un punto diferente en el tiempo; por lo que el conjunto de datos con el que se trabaja corresponde a 4097 puntos de datos durante 23.5 segundos para cada uno de los 500 individuos en análisis.

Estos 4097 puntos de datos se dividen en 23 fragmentos que contienen 178 puntos de datos durante 1 segundo, donde cada uno de estos puntos de datos es el valor de la grabación del EEG en un momento diferente. Con esto, finalmente se tiene un conjunto de datos conformado por 11500 piezas de información (filas) y, cada una de estas contiene un total de 178 puntos de datos por 1 segundo conformando las columnas. Además, se cuenta con una columna adicional representada con la etiqueta "y" correspondiente a la variable de respuesta, conformada por valores numéricos enteros que indican lo siguiente:

- 5: indica que el paciente tenía los ojos abiertos en el momento en que se graba la señal de EEG.
- 4: indica que el paciente tenía los ojos cerrados en el momento en que se graba la señal de EEG.
- 3: si se identifica dónde estaba la región del tumor en el cerebro y registran la actividad EEG del área cerebral sana.
- 2: registran el EEG del área donde se localizó el tumor.
- 1: registro de la actividad convulsiva.

Para el caso en el que los pacientes se ubiquen en las clases 2, 3, 4 y 5; indica que son sujetos que no tuvieron registros de ataques epilépticos. Únicamente los sujetos que se ubican en la categoría 1 tuvieron crisis epilépticas.

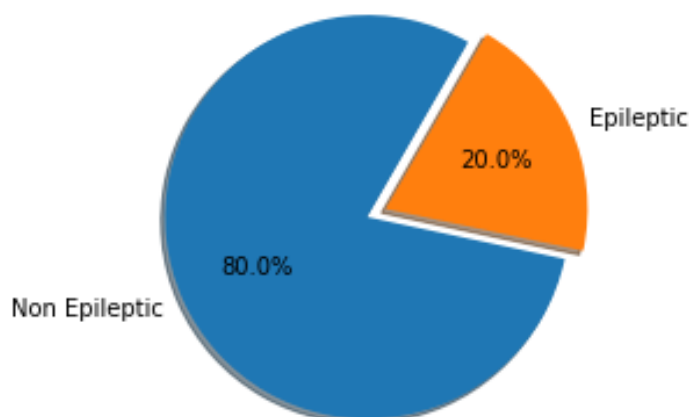
A pesar de que se cuentan con 5 clases distintas en la variable de respuesta, se va a realizar el análisis con una clasificación binaria que determine si el sujeto registró o no alguna actividad epiléptica.

# ANÁLISIS Y VISUALIZACIÓN DE DATOS

El conjunto de datos con el que se desarrolla el presente trabajo consiste en una serie temporal que muestrea las mediciones realizadas mediante un encefalograma a diferentes pacientes. A continuación, se resaltan aquellos hallazgos que se consideran pueden dar un potencial valor a este estudio.

## DISTRIBUCIÓN DE LOS CASOS EPILÉPTICOS Y NO EPILÉPTICOS

El conjunto de datos con el que se desarrolla el proyecto cuenta con únicamente un 20% de casos en los que se registran un ataque convulsivo; factor que puede perjudicar el entrenamiento de los modelos.



*Figura 1. Distribución de casos epilépticos y no epilépticos en el conjunto de datos*

## ANÁLISIS DEL PROMEDIO DE LOS VALORES REGISTRADOS POR EL EEG PARA CADA FRAGMENTO DE INFORMACIÓN

Con el fin de comprender mejor estadísticamente el conjunto de datos, se realiza un 'box plot' del promedio de los valores de los registros por el EEG para cada uno de los casos analizados: epilépticos y no epilépticos. Como se puede observar en la figura 2, el promedio de las mediciones de los casos que no presentan ataques convulsivos se concentra en un rango muy pequeño entre -10 y -5; mientras que para los casos que registran epilepsia, el promedio varía considerablemente, donde la mayor concentración de datos se encuentra entre -12 y 2 aproximadamente, llegando a alcanzar las mediciones inferiores a -25 o superiores a 10 en algunos casos.

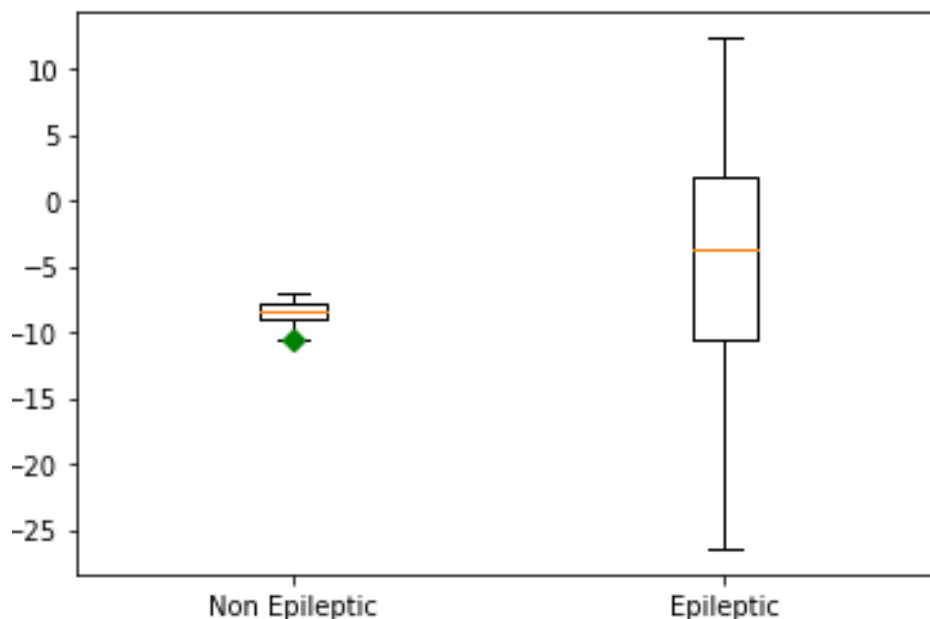


Figura 2. Box plot de promedio de los valores registrados por el EEG para cada uno de los fragmentos de información del conjunto de datos

## VISUALIZACIONES DE LA SERIE TEMPORAL PARA AMBAS CATEGORÍAS EN ANÁLISIS

Tomando 2 ejemplos de fragmentos de cada uno de los casos en estudio, se procede a graficar un caso en el que el paciente no presente señales de ataque convulsivo contra otro que sí con el fin de visualizar las diferencias entre ambas mediciones.

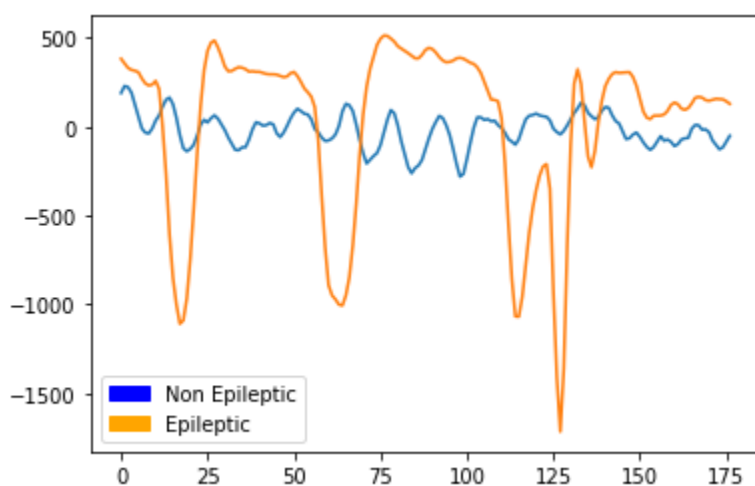


Figura 3. Ejemplo 1 de visualización de la serie temporal, caso de epilepsia vs caso normal

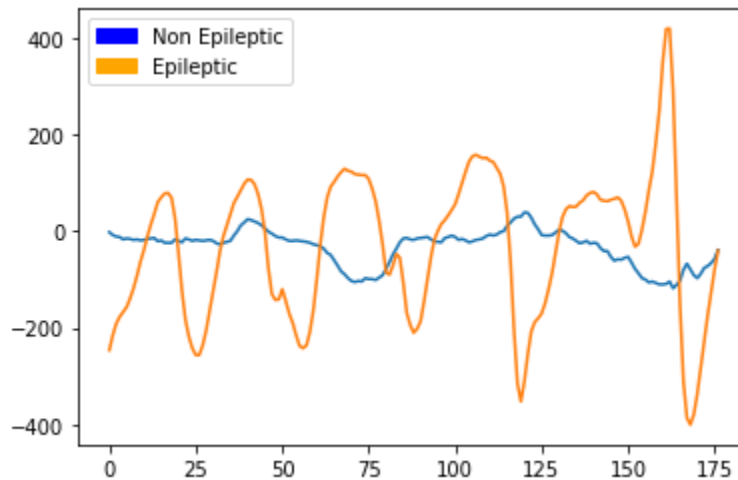


Figura 4. Ejemplo 2 de visualización de la serie temporal, caso de epilepsia vs caso normal

Ambas visualizaciones confirman lo antes analizado en el 'box plot' de la figura 1, en donde se puede observar claramente que para los casos en el que se registra algún ataque convulsivo los niveles de tensión varían entre rangos mucho mayores en comparación con aquellos que no lo presentan. Además, se observa una variación aleatoria y más pronunciada de las señales de medición a través del tiempo para los casos en el que se etiqueta como epiléptico.

### ANÁLISIS DE LA DESVIACIÓN ESTÁNDAR DE LOS VALORES REGISTRADOS POR EL EEG PARA CADA FRAGMENTO DE INFORMACIÓN

Partiendo de los resultados obtenidos de las visualizaciones de la serie temporal, se procede a analizar gráficamente la desviación estándar obtenida para cada uno de los puntos de datos los casos analizados.

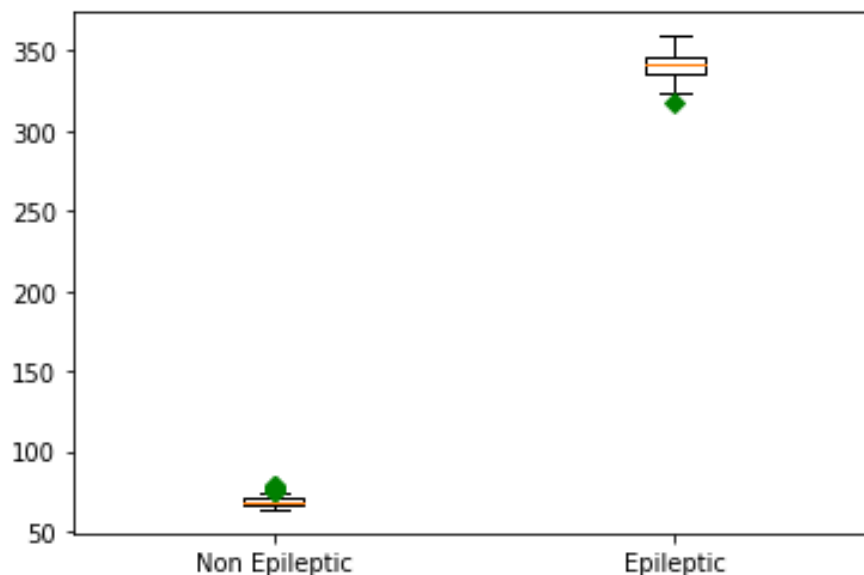


Figura 5. Box plot de la desviación estándar de los registros del EEG para cada uno de los fragmentos de información del conjunto de datos

Se puede observar una clara diferencia entre ambos casos en análisis. Para el caso de en el que el paciente experimenta un ataque convulsivo, la desviación estándar de los datos se ubica en valores cercanos a 340; mientras que para aquellos en el que el paciente no registra el padecimiento, este valor se encuentra alrededor de 70.

### ANÁLISIS CONSIDERANDO LOS VALORES MÁXIMOS Y MÍNIMOS DE LAS MEDICIONES DEL EEG PARA CADA FRAGMENTO DE INFORMACIÓN

Se analiza el valor registrado por el EEG para ambas categorías en estudio mediante un gráfico de dispersión, analizando los valores máximos (figura 6) y mínimos (figura 7) alcanzados en cada uno de los fragmentos en los que se divide el conjunto de datos.

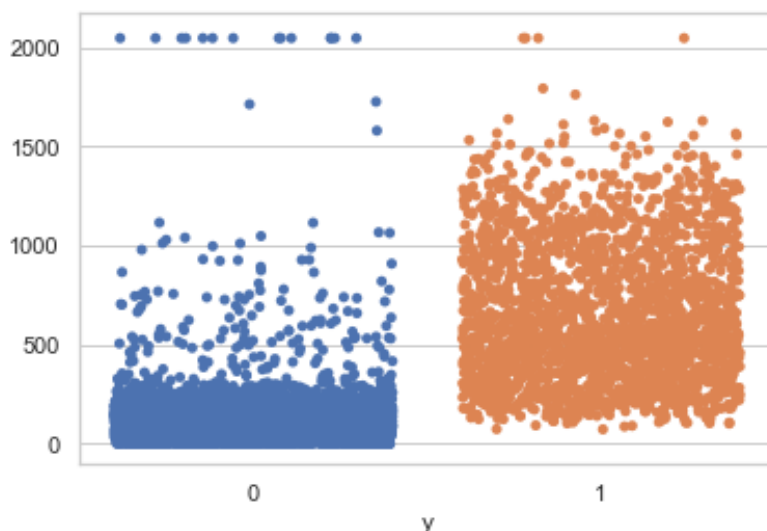


Figura 6. Gráfico de dispersión de los máximos registrados por el EEG en cada fragmento de información.

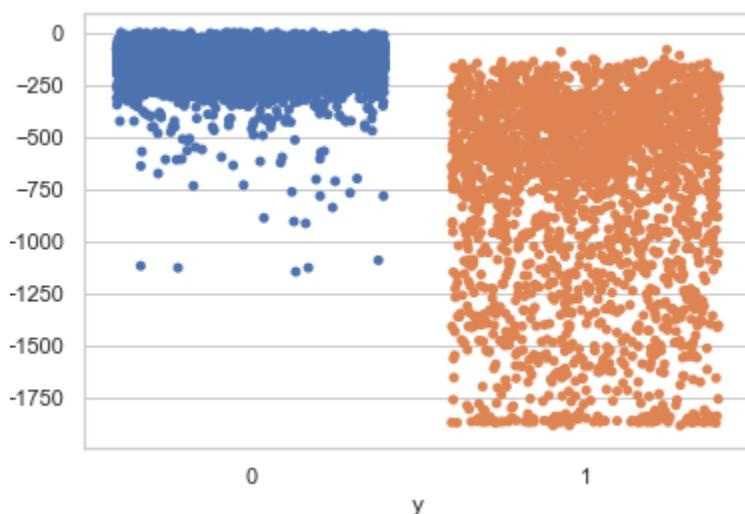


Figura 7. Gráfico de dispersión de los mínimos registrados por el EEG en cada fragmento de información.



Se puede observar una diferencia entre los valores máximos y mínimos alcanzados en las mediciones del EEG para las dos categorías en estudio; siendo la categoría que registra epilepsia aquella que alcanza los valores más altos y bajos en las mediciones.

Se procede a analizar la relación entre los valores máximos y mínimos para cada fragmento de información. La figura 8 resume los resultados obtenidos.

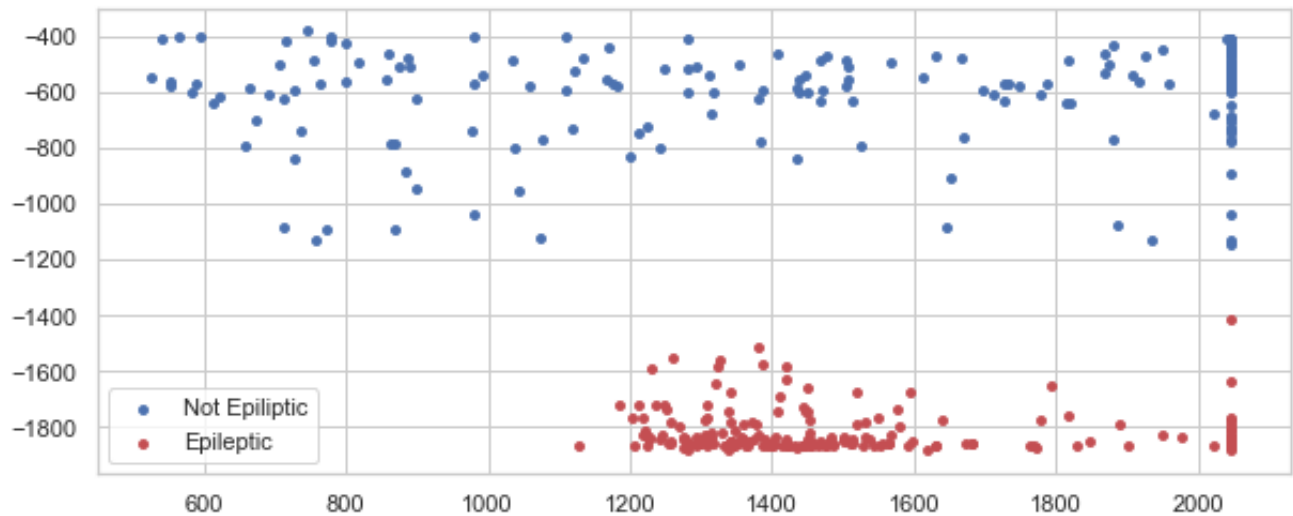


Figura 8. Gráfico de dispersión de los máximos vs mínimos registrados por el EEG en cada fragmento de información.

Se puede observar una clara separación en la relación entre ambas variables, lo cual puede ser de utilidad al momento de entrenar el modelo de clasificación. Debido a esto, se decide crear un nuevo atributo al conjunto de datos denominada 'Delta max-min' la cual corresponde al valor máximo registrado en ese fragmento menos el valor mínimo alcanzado.

Mediante el análisis de correlación se puede observar que esta variable creada es la que presenta una mayor correlación con la variable dependiente, 0.78 aproximadamente, confirmando lo observado anteriormente con la figura 8, siendo esta variable incluso la única que tiene una correlación significativa con la variable a predecir, ya que el resto de los atributos no superan el 0.1 de correlación.



# MODELOS DE PREDICCIÓN

Con el fin de reducir la dimensionalidad de los datos, el cual consiste en 178 atributos más 1 atributo agregado anteriormente, se decide utilizar PCA con un mínimo de componentes principales que contengan al menos un 95% de la varianza. Como resultado se obtiene un conjunto de datos de 39 atributos que se utiliza para el entrenamiento de los modelos.

En este caso, se lleva a cabo una separación del conjunto de datos que represente un 70% para un subconjunto de entrenamiento y un 30% para un subconjunto de prueba. Se lleva a cabo el desarrollo de 4 modelos de clasificación distintos basados en los algoritmos de: *Decision Tree Classifier*, *Random Forest Classifier*, *K-Nearest Neighbors* y *Support Vector Classifier*. La tabla 1 muestra el rendimiento de los modelos de clasificación en las predicciones.

Tabla 1. Parámetros de rendimiento de los modelos de clasificación.

Modelo	Precisión	Kappa
Decision Tree	96.6%	0.875
Random Forest	97.4%	0.922
K-Nearest Neighbors	91.5%	0.688
Support Vector Classifier	97.5%	0.923

De los resultados anteriores, se puede observar que aquel modelo que presenta un mejor rendimiento corresponde al 'Support Vector Classifier' con un 97.5% de precisión en las predicciones y el valor de kappa más alto.

# CONCLUSIONES

Es posible llevar a cabo el diagnóstico de un EEG con una precisión superior de un 85% mediante un modelo de clasificación, permitiendo así la reducción del factor humano en el diagnóstico dado, factor altamente influyente y que puede variar de un médico a otro.

Hay una clara diferencia en el comportamiento de las señales eléctricas durante un EEG entre pacientes que registran un ataque convulsivo y aquellos que no lo presentan.

Un factor determinante en la predicción corresponde a la diferencia entre el valor máximo alcanzado y el mínimo registrado en un período definido de tiempo, 1 segundo en este caso.