INFORME: EXAMEN FINAL DE PROGRAMACION 2025-1

AUTORES

Martin Pérez Betancourt

Samuel Ernesto Mercado Mercado

UNIVERSIDAD EIA

ENVIGADO

02-06-2025

1. PROBLEMA SELECCIONADO

El problema que escogimos es la identificación de episodios convulsivos de personas diagnosticadas con epilepsia.

Para ello, el dataset seleccionado fue *Epileptic Seizure Recognition,* que contiene datos de distintos electroencefalogramas (EEG) adquiridos de pacientes (Ver Anexo A).

El dataset cuenta con 11500 entradas de información. Esta se deriva de la siguiente manera:

Originalmente se contaban con los datos de **500** personas. Los EEG de cada uno de ellos tuvo una duración de 23 segundos, durante los cuales se tomaron un total de **4097** muestreos.

Estos 4097 muestreos se repartieron en **23** paquetes de **178** muestras para cada segundo del EEG, equivaliendo cada uno de ellos una fila de información asociado a cada sujeto, de modo que cada uno de los **500** individuos cuentan con **23** entradas, donde cada uno de estos, contando así un total de **11500** filas de datos.

Cada fila de datos tiene asociado una categoría entre **1** y **5.** Únicamente aquellos datos pertenecientes a la categoría 1 corresponden a personas con **epilepsia.** Es por esto por lo que para facilitar el análisis se sn las clasificaron : Los individuos con epilepsia pertenecen a la clase **1** y aquellos sin epilepsia a la clase **0.**

2. ANALISIS EXPLORATORIOS DE LOS DATOS

Los datos se analizaron usando gráficas y datos estadísticos representativos, a continuación, se describen los hallazgos encontrados.

Lo primero en analizarse fue la distribución de las clases al interior del dataset, a través de un gráfico de barras. Asi:

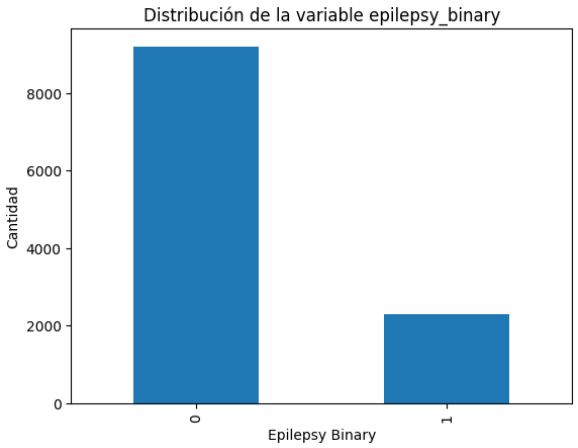

Figura 1. Gráfico de barras de las variables categóricas

En este se puede observar claramente que existe un desbalance considerable en la distribución de las categorías, habiendo alrededor de 3 veces más individuos catalogados como sanos con relación a aquellos que presentan epilepsia. Esto se debe a la binarización, pues las categorías 2,3,4 y 5 suman todas a el conteo de aquellos sin epilepsia, por lo cual inevitablemente terminan sumando más que aquellos categorizados como 1.

El segundo componente que se analizo es el comportamiento entre los individuos de distintas clases en su señal EEG. En las figuras 2 y 3, se muestran comparación de 5 personas con y sin epilepsia respectivamente.

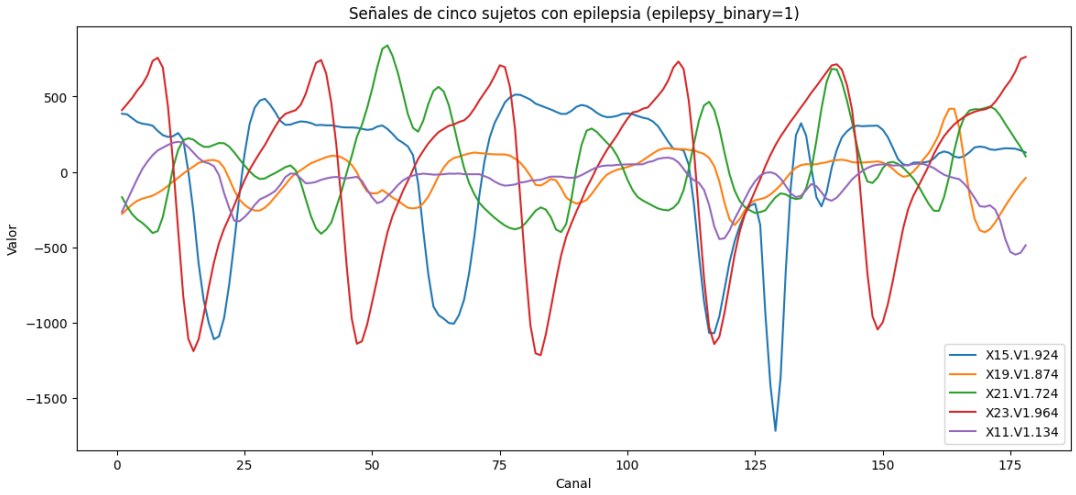

Figura 2. Señal de 5 personas con epilepsia

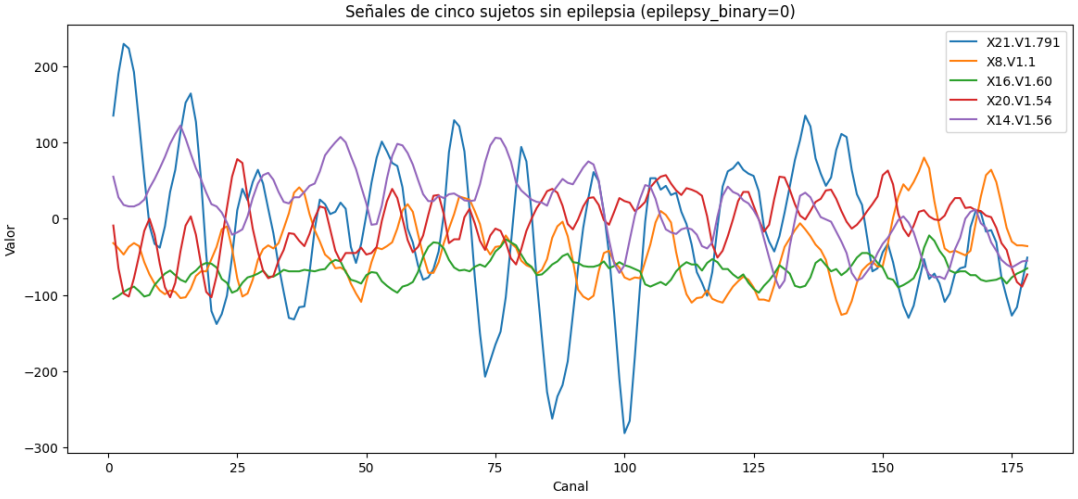

Figura 3. Señal de 5 personas sin epilepsia

En estas graficas es posible observar comportamientos propios de cada uno de los grupos de clasificación. Asi, por un lado, las señales en la figura 2 exhiben un comportamiento más errático, con picos y valles a lo largo de todo el muestreo. Esto sugiere que

las personas con epilepsia están asociadas a valores muy grandes o pequeños de la señal, y con cambios repentinos de gran magnitud de esta, además de ser muy dispersas las señales entre los pacientes.

Los sujetos sin epilepsia por su parte exhiben en un comportamiento inverso, sus señales son más estables, presentando menos variaciones, y se encuentran en un rango de valores más compacto, habiendo además menor separación entre las distintas muestras.

Por último, se analizó en general, la relación entre las dos clases, al graficar el promedio de las señales asociadas a cada una de ellas. En la figura 4 se muestran los resultados obtenidos.

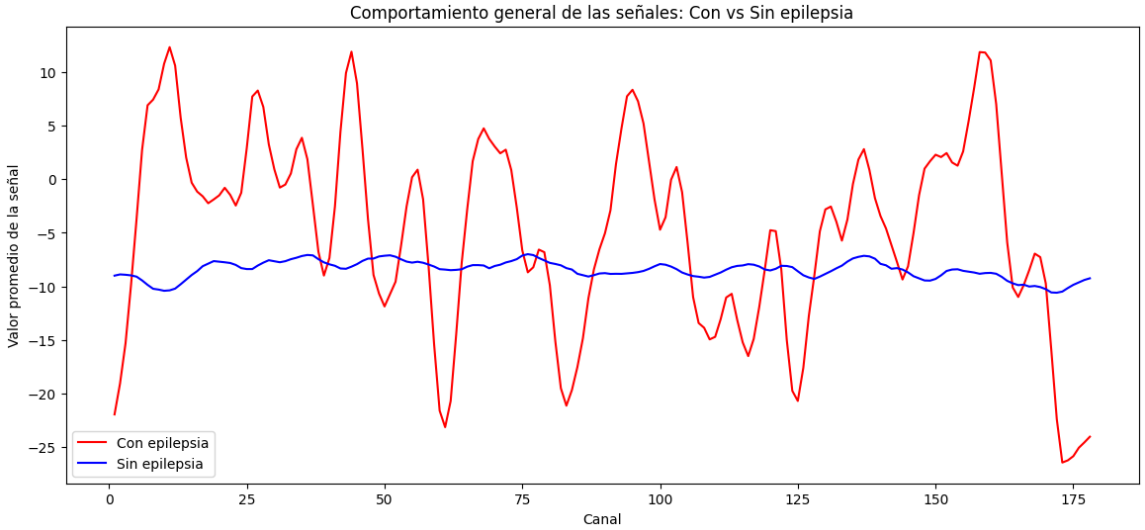

Figura 4. Señal promedio de las clases: Con Epilepsia (en rojo) y sin epilepsia (en azul).

Aquí se puede observar claramente que existen diferencias muy diferentes entre los distintas los sujetos con epilepsia y aquellos sin esta. Los epilépticos presentan los cambios abruptos en su señal, que oscila entre valles y picos constantemente, a diferencia de los no epilépticos, cuya señal es relativamente suave.

A partir de esto es bastante seguro que el dataset puede usarse para entrenar un modelo de clasificación para predecir si una persona está sufriendo o no de epilepsia, dado que es dataset cuenta con información que es suficientemente distintiva para separar las clases. Sin embargo, su distribución desigual y esas mismas diferencias tan marcadas, puede que produzcan sesgos y limiten que tambien pueden entrenarse los modelos para generalizar en situación no controladas de uso.

3. PROCESADO DE LOS DATOS

Antes de pasar al entrenamiento de los modelos, primero hizo falta procesar los datos.

En una primera etapa, como ya se mencionó anteriormente, se reclasificaron las clases. Las 2,3,4 y 5 se marcaron como clase 0, al no presentar epilepsia, y la clase 1, se marcó como 1, dado que si la presentan.

El dataset no contaba con datos NaN y tampoco se simularon los mismos, dado que los datos representan una señal continua y los NaN podrían generar inconsistencias en el análisis y el entrenamiento de los modelos.

Luego de la binarización, se procedió a la separación en dos conjuntos independientes, las variables de entrada (x) y las de salida (y), aplicándole un escalado por Minimo-Maximo (*MinMax Scaler*) a las variables de entrada.

Las variables de entrada corresponden a todas las instancias del dataset, excluyendo la etiqueta del individuo, las clases originales y la clase binaria. Mientras que las variables de salida solo contienen las etiquetas posteriores a la reclasificación.

4. ENTRENAMIENTO Y EVALUACION DE MODELOS

4.1. RandomForest Classifier

Para este primer modelo se definieron una serie de hiperparametros para evaluar, siendo estos los que se consideraron que tendrían mayor impacto en el desempeño del modelo:

- **N-estimators**: Número de árboles en el bosque, más árboles pueden mejorar el rendimiento, pero también aumentar el tiempo de cómputo.
- **Max depth**: Profundidad máxima de cada árbol, controla el sobreajuste.
- **Min simples split**: Número mínimo de muestras necesarias para dividir un nodo, valores más altos hacen que el modelo sea más considerado con las predicciones.
- **Min samples leaf**: Número mínimo de muestras requeridas en una hoja, evita hojas muy pequeñas, lo que ayuda a generalizar mejor.

Y se evaluaron en los siguientes valores haciendo uso de *GridSearchCV,* para poder encontrar las mejores combinaciones de hiperparametros en base a la precisión de cada una. Asi:

- **N-estimators**: [50,100,200]
- **Max depth**: [2,4,8]
- **Min samples split**: [2,5]
- **Min samples leaf**: [1,2]

Y se obtuvo la siguiente combinación como la mejor de todas ellas.

- **N-estimators:** 8
- **Max depth:**200
- **Min samples split:** 5

- **Min samples leaf:** 2

Cuyo modelo tuvo una precisión de 0.95 durante el entrenamiento. Asi pues, este fue el modelo que se escogió para el resto del análisis.

Sus curvas de aprendizaje y su curva de confusión se presentan en las figuras 5 y 6.

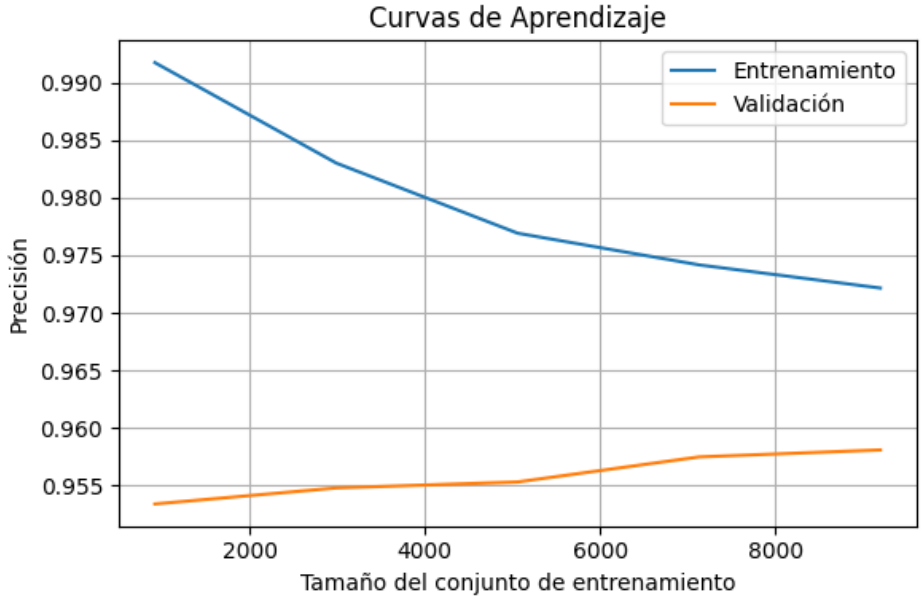

Figura 5. Curva de aprendizaje del modelo de RandomForest Classifier

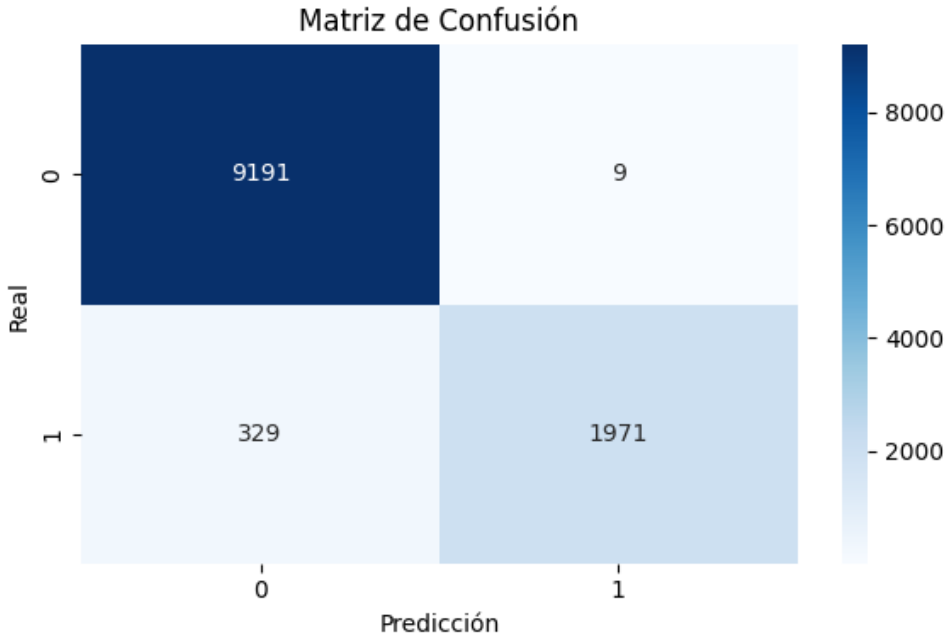

Figura 6. Matriz de confusión del modelo

En la primera figura, en la curva se observa que el modelo, a medida que aumenta el tamaño del conjunto de entrenamiento, su precisión tiende a disminuir, estabilizándose a medida que el conjunto aumenta más-, de manera similar a los que sucede con la validación, cuya precisión tienden a aumentar con el tamaño del conjunto.

El comportamiento durante el entrenamiento y la validación parece indicar que se están presentando problemas por overffiting, que hace el que el modelo aprenda muy bien los patrones de los datos que se le dan y no sea tan bueno generalizando cuando se le proporcionan nuevos datos, por lo cual, a medida que aumenta los volúmenes de datos de entrenamiento, tiene mayores extrapolando su conocimiento para poder clasificarlos.

Esto se ve reflejado tambien en la matriz de confusión, donde el modelo reconoció correctamente casi todo los positivos (1), pero presento muchos casos de falsos negativos (predijo un 0 pero 1), lo cual indica que su algoritmo parece estar buscando con mayor fuerza negativos que positivos, resultado en errores en la clasificación.

Ahora bien, esto es causado por la distribución desigual de las clases dentro del dataset, cuyo peso hace que el modelo les preste mucha atención a los casos donde no hay epilepsia en relación con los que si presentan la patología.

Asi pues, para mejorar esta situación es posible aplicar algunas técnicas que no impliquen necesariamente la recolección de más datos, pues esta sería una alternativa que podría solucionar muchos problemas, pero no siempre es posible disponer de ella.

De esta manera, el primer método de mitigación podría intentar reducir el número de entradas de la clase mayoritaria (sin epilepsia) para que ambos volúmenes de datos tuvieran un peso similar, esto, sin embargo, tiene la desventaja que puede reducir el poder predictivo del modelo, pues con menos datos el aprendizaje que este realiza es menor y tiene menor capacidad de generalizar.

Una segunda alternativa seria ajustar parámetros del modelo como la profundidad, disminuyéndola, o el min samples leaf, aumentándola, para mejorar su capacidad de generalización, aunque esto tiene un costo en cuanto a tiempo computacional del modelo, razón por la cual podría no ser particularmente la más viable de las opciones.

Por último, una tercera forma de mitigar el overffiting es ajustando el modelo para que le preste más atención a la clase minoritaria, esto tendría ventajas pues, no requiere quitar datos ni aumentar significativamente el tiempo computacional del entrenamiento del modelo, haciéndolo una alternativa muy buena. En RandomForestClassifier, se puede usar el parámetro *class_weight='balanced'* para este fin.

4.2. *Logistic Regression*

Para este segundo modelo se implementó una regresión logística. Este es un modelo comúnmente utilizado para clasificación binaria, como en este caso donde se busca diferenciar entre personas con epilepsia y personas sanas.

Se definieron varios hiperparametros que se consideraron relevantes para mejorar el desempeño del modelo, y en este caso, los principales fueron:

- **C**: valor de regularización inverso, controla qué tanto se penaliza la complejidad del modelo. Valores más bajos implican más regularización.
- **Penalty**: Este indica el tipo de penalización utilizada por el modelo para evitar un sobreajuste.
- **Solver**: este es un algoritmo utilizado para resolver la optimización de la función de coste.

Estos hiperparametros se evaluaron mediante una búsqueda con GridSearchCV con el fin de encontrar la mejor combinación posible que maximizara la precisión del modelo. Los valores evaluados fueron:

- **C**: [0.01, 0.1, 1, 10]
- **Penalty**: ['l2']
- **Solver**: ['lbfgs']

El resultado de esta búsqueda es que se encontró que la mejor combinación:

- **C**: 10
- **Penalty**: l2
- **Solver**: lbfgs

Este modelo, con esos parámetros, alcanzó una precisión de 0.812 en la validación cruzada, y una precisión total del 0.812 en todo el dataset.

Aunque la precisión global parece aceptable, al revisar la matriz de confusión se identificaron graves problemas en la clasificación de la clase minoritaria (personas con epilepsia). En total, de 2300 personas epilépticas, solo 143 se clasificaron correctamente, mientras que 2157 se clasificaron incorrectamente como sanas. Esto significa que el modelo no está logrando identificar correctamente los casos positivos, presentando un recall del 6.2% para la clase 1, lo que en contextos clínicos sería una tasa de error completamente inaceptable.

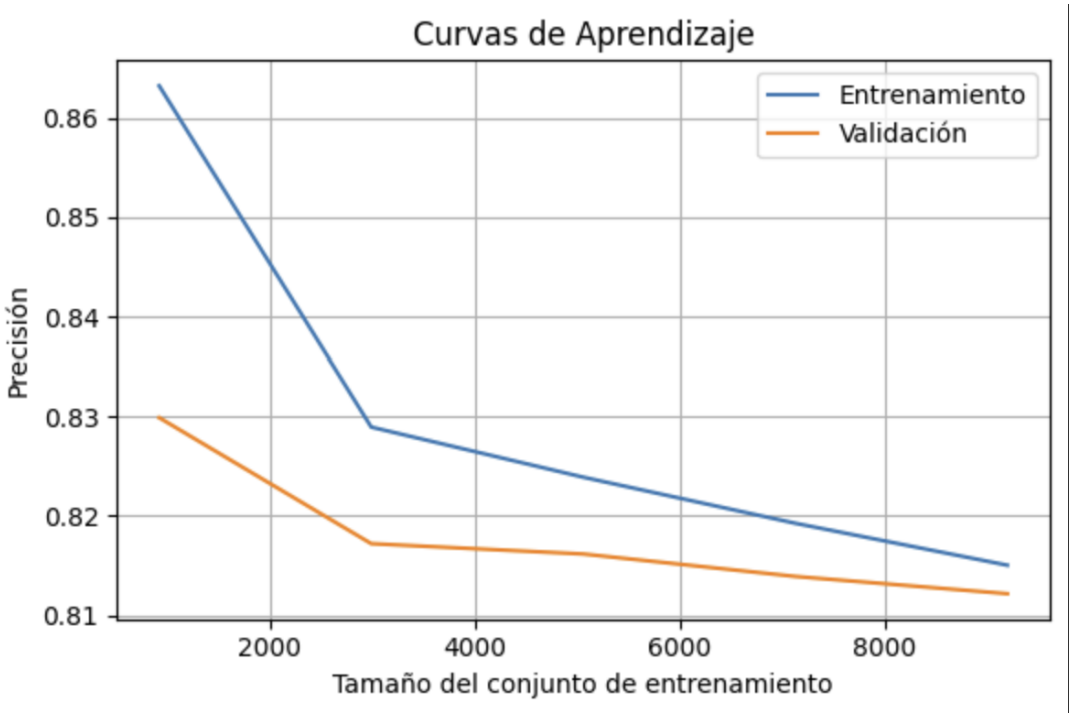

Figura 7. Curva de aprendizaje del modelo de Regresión Logística

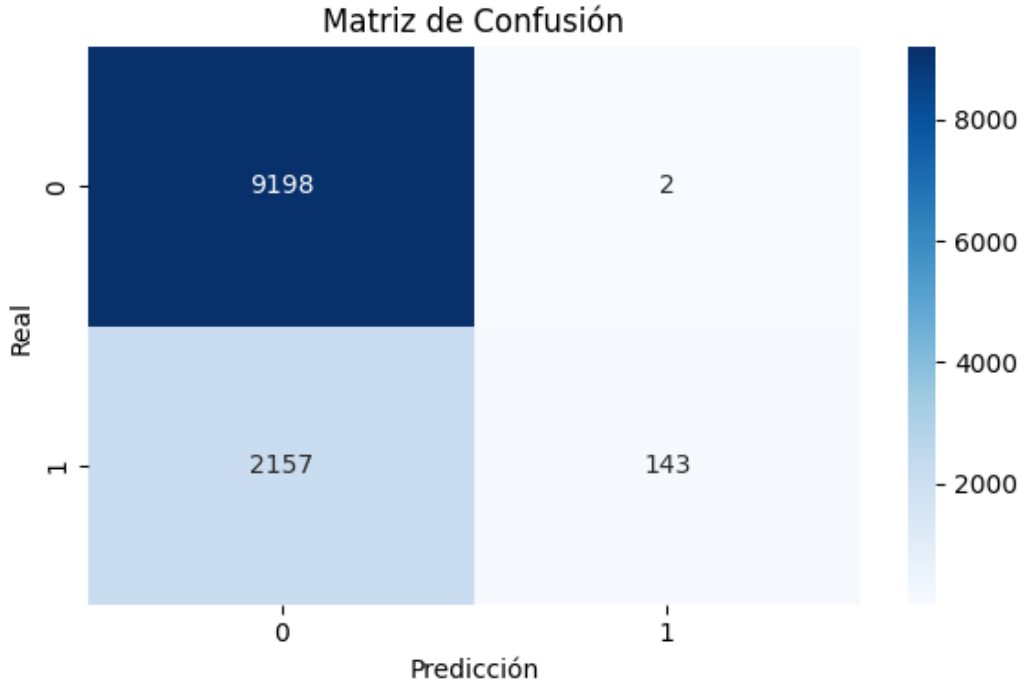

Figura 8. Matriz de confusión del modelo de regresión logística

En la figura 7 se muestra la curva de aprendizaje del modelo de regresión logística. En esta curva se observa que la precisión del conjunto de entrenamiento inicia alta (alrededor de 0.865) y va disminuyendo poco a poco mientras que se incrementa el tamaño del

conjunto de entrenamiento. Esto es un comportamiento esperado, ya que, entre más datos, el modelo se enfrenta a una mayor variedad de ejemplos y, por tanto, encuentra más difícil hacer un ajuste perfecto para todos.

La brecha entre entrenamiento y validación no es muy amplia, lo cual indica que el modelo no está sobre ajustado, pero sí que ha alcanzado un límite en su capacidad de mejora.

Esta forma de curva sugiere que el modelo ha aprendido todo lo que puede dadas sus características y la representación actual de los datos. La falta de mejora en la curva de validación no debe de estar en la cantidad de datos, sino más bien en las limitaciones del propio modelo o en el **desequilibrio de clases** presente en el dataset, porque hay muchas más personas sin epilepsia, que aquellas que si la tienen.

Esto también se ve reflejado en la matriz de confusión, donde el modelo predice muy bien los negativos (sanos), pero ignora por completo la clase epiléptica. Esto se debe a que, al estar tan desbalanceado el dataset, el modelo se vuelve muy bueno prediciendo ceros, pero muy malo prediciendo unos, porque en un primer lugar, hay muchos más ceros que unos.

Para mejorar esta situación se podría tener que modificar el muestreo, una opción podría ser aplicar sobre muestreo de la clase minoritaria (epilépticos), lo cual permitiría al modelo ver más ejemplos de estos y aprender mejor a identificarlos.

En general, aunque este modelo presenta un desempeño aceptable en términos generales, no es adecuado por sí solo para la tarea clínica que se busca resolver, ya que ignora casi completamente la clase de mayor relevancia en el diagnóstico.

5. COMPARACIÓN DE MODELOS

5.1. RESULTADOS DE PRECISIÓN GENERAL

El modelo de RandomForest alcanzó una precisión total de 97.0%, mientras que el modelo de Regresión Logística logró una precisión de 81.2%. Esta diferencia es un claro indicio de que el desempeño general del RandomForest es superior para este conjunto de datos. Sin embargo, estas estadísticas no deben considerarse concluyentes por sí solas.

5.2. COMPARACIÓN DE LAS CURVAS DE APRENDIZAJE

En la figura 9 se presentan las curvas de aprendizaje de ambos modelos, las cuales reflejan diferencias notables en su comportamiento:

El modelo RandomForest muestra claros signos de overffiting, presentando una brecha significativa entre la precisión de entrenamiento y validación. Pero, la curva de validación

mejora a medida que se incrementa el tamaño del conjunto, lo que indica que el modelo puede generalizar mejor con grandes volúmenes de datos.

Por otro lado, la Regresión Logística muestra curvas más cercanas entre sí, algo positivo, pues no se muestran señales de sobreajuste, pero también se muestra que no hay una mejora sustancial en la precisión a medida que se aumentan los datos, lo que sugiere que este modelo tiene una capacidad limitada de aprendizaje y que ya ha alcanzado su máximo rendimiento con la representación actual del dataset.

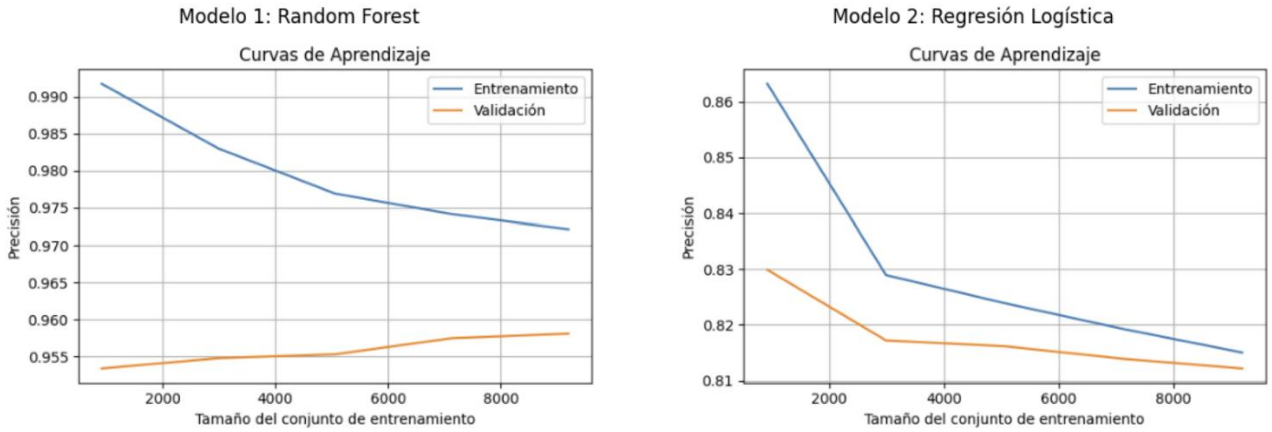

Figura 9. Curvas de aprendizaje lado a lado de RandomForest y Regresión Logística

5.3. COMPARACIÓN DE LAS MATRICES DE CONFUSIÓN

Las matrices de confusión de ambos modelos (figura 10) evidencian diferencias significativas en su capacidad de predicción, especialmente en los **falsos negativos**. Por un lado, el modelo de RandomForest logra identificar correctamente a la mayoría de las personas con epilepsia, presentando un número bajo de falsos negativos (considerado la distribución de las otras clases y el comportamiento del segundo modelo) , lo que muestra este modelo como un modelo ideal para predecir pacientes con epilepsia en un contexto clínico contraste, la Regresión Logística muestra una tendencia a clasificar erróneamente a muchas personas con epilepsia como sanas, generando una gran cantidad de falsos negativos, siendo casi inútil en contextos clínicos reales, con un recall de apenas 6.2% para la clase 1 (pacientes con epilepsia)

Ahora bien, ambos modelos siguen siendo buenos en encontrar las personas sin epilepsia, por lo cual no seria del todo correcto descartar a ninguno de los dos para este fin. Evidentemente la regresión logística necesita ajustarse mejor sus parámetros para mejorar sus tasas de predicción, que como se verá más adelante son apenas mejores que una decisión aleatoria.

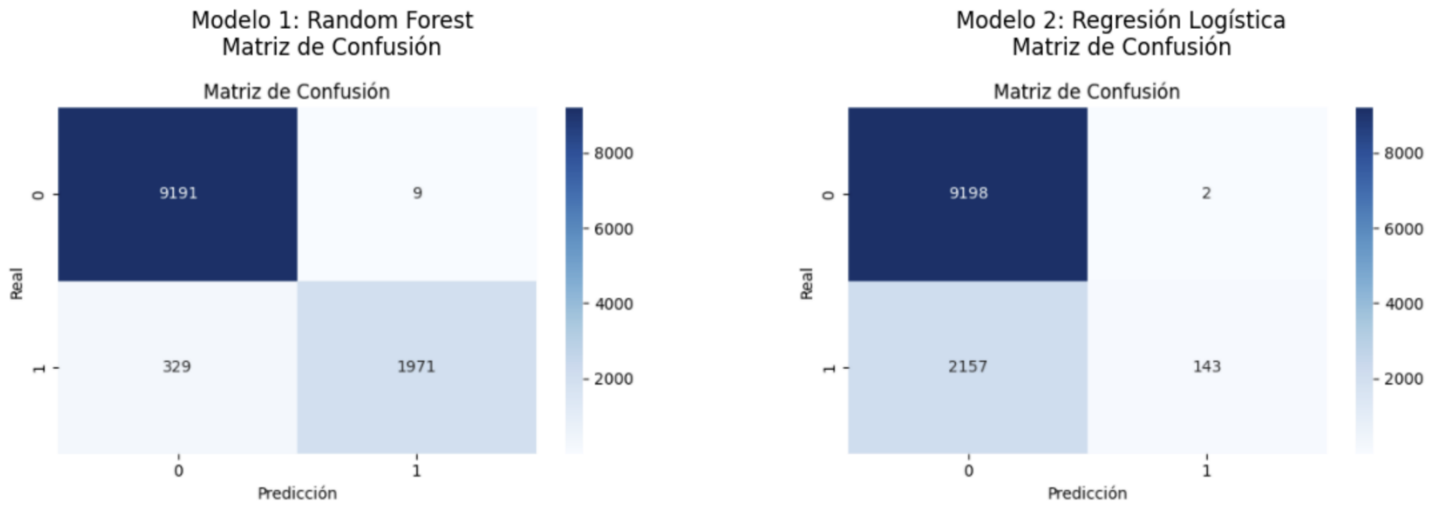

Figura 10. Matrices de confusión de ambos modelos lado a lado

5.4. COMPARACIÓN DE CURVAS ROC Y AUC

La curva ROC muestra que el modelo de regresión logística tiene un desempeño pobre (AUC = 0.57), apenas mejor que una predicción aleatoria, por lo cual no sería la mejor opción. En cambio, el modelo RandomForest alcanza un AUC de 0.998, lo que indica una capacidad de clasificación excelente. Sin embargo, un valor tan alto sugiere posible overfitting, por lo que se recomienda validar con técnicas como cross-validation y revisar el balance de clases para asegurar una evaluación justa.

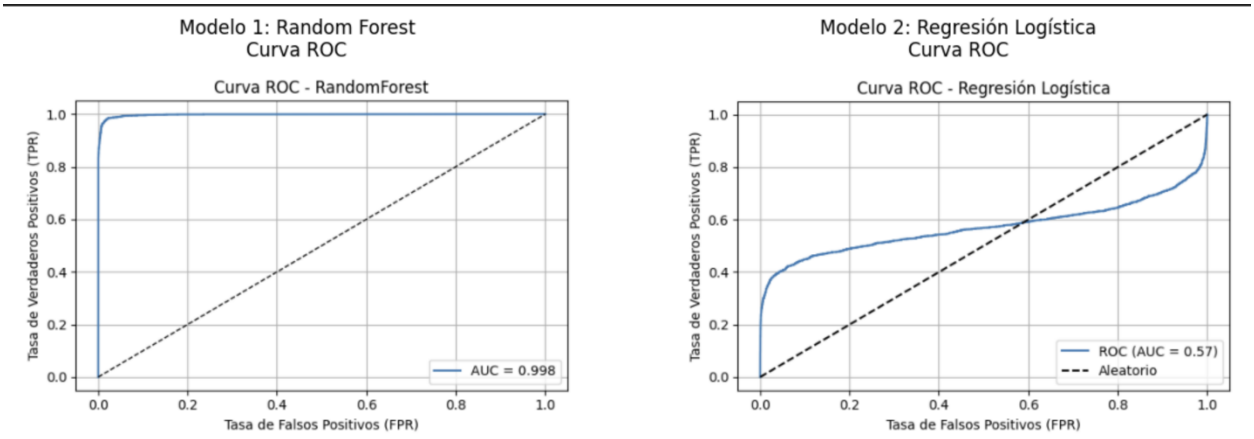

Figura 11. Curvas ROC de Random Forest y de Regresión Logística lado a lado

6. ANEXOS

A. Repositorio del dataset usado

https://www.kaggle.com/datasets/harunshimanto/epileptic-seizure-recognition/data