Proyecto de ML para Pregrado: Clasificación de Ritmo Cardíaco

Victor Cornejo – vcornejo19@alumnos.utalca.cl

Contexto

La fibrilación auricular es una condición en la que el corazón late de forma irregular y descoordinada. Detectarla a tiempo es fundamental para prevenir complicaciones. En este proyecto se utilizarán señales reales de electrocardiograma (ECG) extraídas del PhysioNet/Computing in Cardiology Challenge 2017, y se implementará un clasificador automático para distinguir entre ritmos normales y episodios de fibrilación auricular.

Objetivo

Desarrollar un modelo de Random Forest para clasificar segmentos de señales ECG como ritmo normal o fibrilación auricular (AFib), usando características estadísticas de los intervalos RR (Promedio, Desviación estándar, Asimetría, Curtosis).

Desarrollo

1.- Obtención y tratamiento del dataset

Se descarga el dataset del desafío PhysioNet, que contiene una carpeta con todos los ECG a analizar, formados por archivos .mat y .hea. Además, incluye el archivo REFERENCE.csv con las etiquetas de cada ECG.

Se desarrolló un script que procesa las señales y genera un CSV con características estadísticas basadas en los intervalos RR, útiles para entrenar el modelo junto con las etiquetas. Este CSV, llamado ecg_rr_features_completo.csv, contiene ECG normales y con AFib.

Para el proyecto se solicitó un subconjunto de 30 registros, por lo que se creó un nuevo script que selecciona 15 ECG normales y 15 con AFib, garantizando un subconjunto balanceado sin sesgo para el entrenamiento. Esto da como resultado un archivo llamado ecg_rr_features_curado.csv el cual servirá para entrenar el modelo.

2.- Análisis exploratorio de datos

Para el análisis exploratorio de datos se decidido crear 4 histogramas con curvas de densidad, uno para cada característica.

2.1.- Distribución del Promedio de intervalos RR (ms):

Este gráfico muestra la distribución del promedio de los intervalos RR (tiempo entre latidos sucesivos) para cada clase. Un promedio alto indica un ritmo cardiaco mas lento, lo cual evidencia que los ECG normales tienen un ritmo cardiaco más lento en comparación con los ECG normales.

2.2.- Distribución de la Desviación estándar de intervalos RR (ms):

Mide cuánto se dispersan los intervalos RR respecto a la media. Un ritmo cardiaco normal se ve reflejado con una baja variabilidad mientras que el AFib se muestra como todo lo contrario, lo cual se refleja en el gráfico.

2.3.- Distribución de la Asimetría de intervalos RR:

Mide si la distribución de los intervalos RR está sesgada hacia la izquierda o derecha. Una distribución cercana a 0 indica una distribución simétrica, esperada en los ECG normales, una distribución postitiva o negativa indica una simetría de datos, esperada en los casos de AFib.

2.4.- Distribución de la Curtosis de intervalos RR:

Indica si la distribución tiene colas más pesadas o picos más pronunciados que una distribución normal. En lo ECG normales se muestra una curtosis moderada mientras que en los casos AFib se ve una curtosis variable.

3.- Entrenar el modelo simple

Para el modelo simple, se decidió utilizar Random forest.

El modelo mostró un buen desempeño en la clasificación de ritmos normales y fibrilación auricular (AFib) usando estadísticas de los intervalos RR. Destaca su alta precisión global, buena detección de AFib (recall) y bajo nivel de falsos positivos (precision). El F1-score elevado refleja un buen equilibrio entre ambas métricas. Este enfoque es eficaz, rápido e interpretable, sin necesidad de procesar la señal ECG completa.

4.- Propuestas de mejora

4.1.- Extraer más características:

Incorporar métricas como frecuencia cardíaca promedio y HRV puede aportar información adicional para mejorar la detección de AFib.

4.2.- Probar modelos más complejos:

Algoritmos como SVM o XGBoost podrían aumentar la precisión al capturar relaciones más complejas.

4.3.- Ampliar el dataset:

Usar más señales ECG mejoraría la capacidad de generalización y reduciría el sobreajuste.

4.4.- Validación cruzada

Permite obtener resultados más confiables y robustos al evaluar el modelo en diferentes particiones de los datos.

Conclusiones

El modelo desarrollado demostró que es posible clasificar ritmos normales y fibrilación auricular (AFib) usando solo estadísticas simples derivadas de los intervalos RR, lo que permite construir un sistema rápido e interpretable. Sin embargo, el uso de un subconjunto pequeño y balanceado limita la generalización a situaciones reales, donde la variabilidad de los datos puede ser mucho mayor.

Desde el punto de vista ético, es fundamental recordar que estos sistemas no deben sustituir el diagnóstico médico, sino complementarlo. Un mal uso o la interpretación incorrecta de los resultados podría tener consecuencias graves para la salud. Además, se debe garantizar la confidencialidad de los datos de los pacientes y validar rigurosamente los modelos en entornos clínicos antes de su implementación.