Predicción de Riesgo Cardiovascular utilizando imágenes de electrocardiogramas
Prediction of Cardiovascular Risk Using Electrocardiogram Images

Trabajo de Fin de Grado
Curso 2024–2025

Autor
Noelia Barranco Godoy

Directores
Belén Díaz Agudo
Juan A. Recio García

Doble Grado en Ingeniería Inforomática y Matemáticas
Facultad de Informática
Universidad Complutense de Madrid

Predicción de Riesgo Cardiovascular utilizando imágenes de electrocardiogramas
Prediction of Cardiovascular Risk Using Electrocardiogram Images

Trabajo de Fin de Grado en Ingeniería Informática

Autor
Noelia Barranco Godoy

Director
Belén Díaz Agudo
Juan A. Recio García

Convocatoria: *Febrero 2025*

Doble Grado en Ingeniería Inforomática y Matemáticas
Facultad de Informática
Universidad Complutense de Madrid

11 de diciembre de 2024

Dedicatoria

{**TODO TODO TODO:** Hacer dedicatoria}

Agradecimientos

{**TODO TODO TODO:** Hacer agradecimientos}

Resumen

Predicción de Riesgo Cardiovascular utilizando imágenes de electrocardiogramas

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

Palabras clave

{**TODO TODO TODO:** Máximo 10 palabras clave separadas por comas}

Abstract

Prediction of Cardiovascular Risk Using Electrocardiogram Images

{**TODO TODO TODO:** An abstract in English, half a page long, including the title in English. Below, a list with no more than 10 keywords.}

Keywords

{**TODO TODO TODO:** 10 keywords max., separated by commas.}

Capítulo 1

Introducción

Resumen: En este capítulo pretendemos introducir los objetivos de este trabajo.

1.1. Motivación

El auge de la inteligencia artificial ha hecho posible crear aplicaciones que hace unos años nos parecían imposibles. Como es natural, uno de los campos que ha conseguido avances gracias a esto ha sido el de la medicina, ya que los avances en este campo permiten mejorar la calidad de vida de todo el mundo.

Una de las aplicaciones médicas de la inteligencia artificial que se han trabajado es el procesamiento automatizado de mediciones, creando algoritmos que permitan detectar si estas son normales o no con el fin de reducir la carga laboral de los profesionales de la salud.

Este trabajo se centrará en estudiar las posibilidades de un algoritmo que, a partir de las mediciones que se toman en un ECG, detectar posibles anomalías.

Además de todo esto, la comunidad científica lleva un tiempo dándole importancia a la explicabilidad de los algoritmos, por lo que en este trabajo no solo veremos cómo de eficientes son las distintas aproximaciones, sino que también evaluaremos el grado de explicabilidad de los modelos empleados.

1.2. Aproximaciones al problema

El modelo más conocido es el de Ribeiro et al. (2020), que toma como entrada los datos del ECG tal como se recogen, que en esencia son doce mediciones de la actividad eléctrica en distintos puntos del cuerpo (expandiremos más en este tópico en la siguiente sección).

Nosotros exploraremos la posibilidad de convertir el ECG en diagramas de frecuencia-tiempo, y aplicarle diferentes transformadas antes de entrenar a los modelos, y estudiaremos si esto mejora su eficiencia, así como su explicabilidad.

1.3. Objetivos

El objetivo de este trabajo es modificar distintos modelos de predicción para que tomen como entrada imágenes en lugar de los datos del ECG, y ver como esto afecta a su rendimiento y explicabilidad.

1.4. Plan de trabajo

En el siguiente capítulo tendremos que familiarizarnos con los datos que proporciona un ECG, así como con los modelos ya existentes que tendremos que modificar y con los datos que hay disponibles de manera pública.

Luego modificaremos los modelos para que acepten como entrada imágenes y los volveremos a entrenar con los datos transformados de varias maneras diferentes.

Una vez entrenados los modelos modificados, evaluaremos su rendimiento y explicabilidad, y los compararemos con los modelos originales para ver si hemos conseguido una mejora.

{**TODO TODO TODO:** Expandir esta sección cuándo haya hecho más cosas}

Capítulo 2

Estado de la Cuestión

Resumen: En este capítulo estudiaremos las distintas bases de datos que podemos utilizar, así como los modelos que ya hay.

2.1. Electrocardiogramas

Un ECG es una prueba médica en la que se mide la diferencia de potencial entre varios puntos del cuerpo (cada diferencia de potencial recibe el nombre de derivación). Para hacer un ECG estándar, se ponen diez electrodos en puntos concretos del cuerpo y se generan doce ondas que representan la diferencia de potencial entre doce pares de estos electrodos.

Como es natural, es posible tomar mediciones continuas, así que los ECGs funcionan a una frecuencia específica, lo que indica el número de muestras numéricas que toma el aparato medidor por segundo. Lo estándar es que los ECGs se tomen a 100Hz o a 500Hz dependiendo de la calidad del aparato medidor.

Con estas gráficas (podemos ver los datos de un ECG extraido de Wagner et al. (2022) y dibujado mediante la librería de python *ecg_plot* en la Figura 2.1), un médico puede detectar alteraciones en el ritmo cardíaco que indiquen diferentes patologías cardíacas.

Por tanto, la información de un ECG puede almacenarse simplemente en doce vectores de tamaño 100 o 500 por segundo de duración del ECG. Estos son los datos con los que funcionan los modelos de predicción actuales. Nosotros transformaremos cada una de las derivaciones (es decir, de los vectores) en una imagen (es decir, una matriz) utilizando las transformadas STFT y CWT con distintos parámetros. Luego entrenaremos modelos ligeramente modificados para clasificar anomalías a partir de estas imágenes.

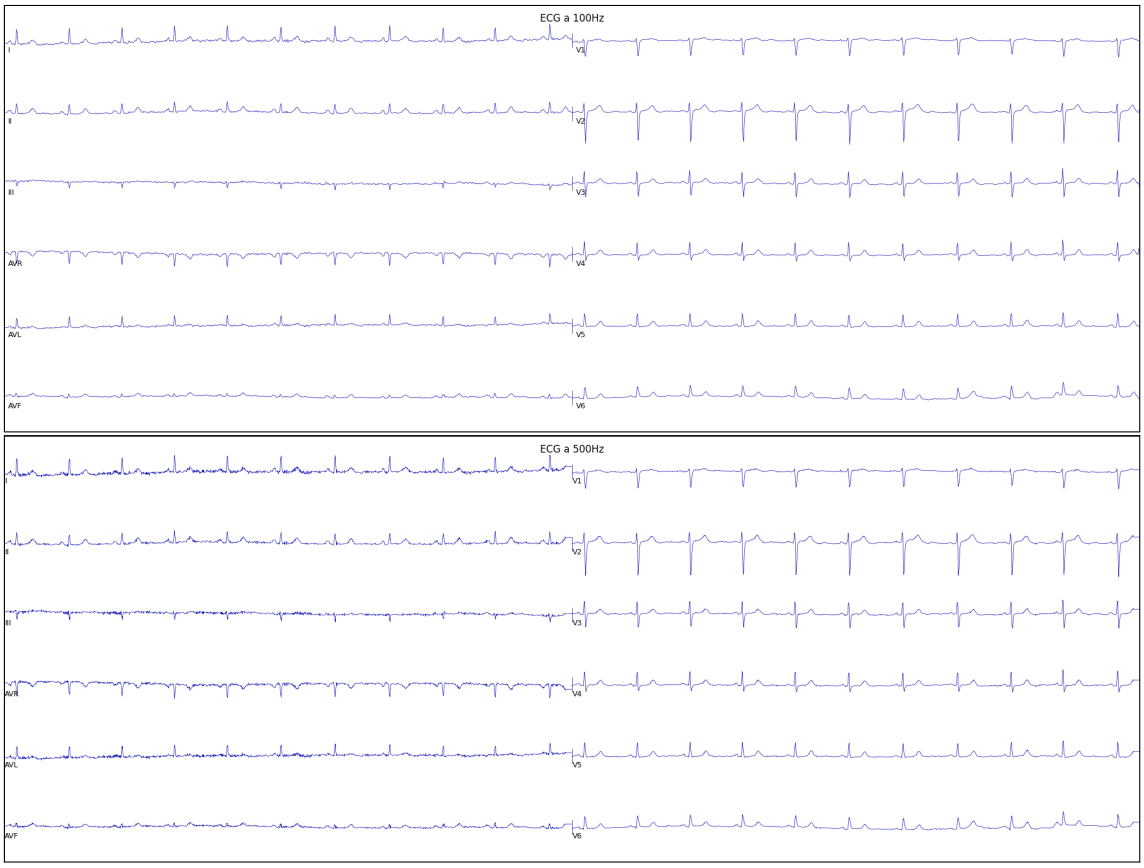

Figura 2.1: Mismo ECG de 12 derivaciones tomado a 100Hz (arriba) y 500Hz (abajo).

2.2. Bases de datos

Como queremos modificar y comparar el modelo de Ribeiro, lo ideal sería trabajar con la misma base de datos que este para entrenar y probar nuestros modelos, pero esta base de datos, llamada CODE (Ribeiro et al., 2021b), no es pública, por lo que esto no es una opción.

Existe una versión pública reducida de la base de datos llamada CODE-15 (Ribeiro et al., 2021a), que tiene alrededor de 350000 casos. No obstante, no podemos comparar un modelo entrenado con los datos de CODE-15 con otro entrenado con los datos originales, porque la ventaja en el conjunto de datos de entrenamiento haría imposible comparar la eficiencia y explicabilidad de manera equitativa. Por ello, vamos a volver a entrenar el modelo original de Ribeiro con una base de datos pública.

No obstante, en lugar de entrenar todos nuestros modelos con CODE-15, utilizaremos otra base de datos completamente pública, la PTB-XL (Wagner et al., 2022). Esta base de datos cuenta solo con unos 22000 casos de prueba, pero todos estos casos vienen con el mismo formato, por lo que procesar los datos va a ser mucho más sencillo con esta base de datos. Además, dado que el objetivo es comparar si las modificaciones de los modelos mejoran el rendimiento, es mucho más importante emplear bases de datos uniformes y realizar varios entrenamientos que usar una bse de datos más grande, lo que incrementaría significativamente el tiempo de entrenamiento.

Un posible trabajo futuro sería escoger el modelo que mejor nos haya funcionado y volver a entrenarlo con una base de datos mayor (incluso con CODE si se tuviera acceso a ella), y luego compararla con el modelo de Ribeiro original, no obstante, eso queda fuera del alcance de este documento.

2.3. Modelos

Como hemos mencionado ya varias veces, el modelo que nos va a servir de referencia en este trabajo es el de Ribeiro, por ser el que mejores resultados ha obtenido hasta ahora. No obstante, también probaremos con otros modelos (Comentario: Esto habrá que ver si es verdad al final).

En concreto, existe una modificación del modelo de Ribeiro que, con tan solo 3 derivaciones, consigue unos resultados muy parecidos (González Cabeza, 2024). Además, este modelo tiene la ventaja de estar ya entrenado con los datos de PTB-XL, por lo que no haría falta volver a entrenar el modelo original.

Capítulo 3

Cuestiones previas

Antes de empezar a entrenar modelos tenemos que decidir una serie de detalles que serán de gran importancia a lo largo del trabajo.

3.1. Análisis de los datos

El mayor problema de PTB-XL (Wagner et al., 2022), además de su tamaño relativamente pequeño, es el balanceo de clases, en la tabla 3.1 podemos ver la distribución de las mismas. Esto puede causar varios problemas en el modelo, como por ejemplo:

Número de registros	Superclase	Descripción	Porcentaje
9514	NORM	ECG Normal	43.64 %
5469	MI	Infarto de Miocardio	25.08 %
5235	STTC	Cambio ST/T	24.01 %
4898	CD	Transtorno de la conducción	22.46 %
2649	HYP	Hipertrofia	12.15 %

Tabla 3.1: Distribución de las superclases en PTB-XL

La información de esta tabla ha sido extraída directamente del repositorio de PTB-XL.

1. **Sobreajuste hacia la clase mayoritaria:** Al haber bastantes más datos de entrenamiento de una clase (NORM) y menos de otra (HYP), el modelo puede aprender mejor los patrones que identifican las clases mayoritarias, haciendo que sepa distinguir peor las minoritarias, lo que en este caso podría reducir notablemente su capacidad de predicción de anomalías raras (He y Garcia, 2009).

2. **Métricas no representativas:** Las métricas más comunes, como la exactitud, pueden ser poco informativas cuando hay un desbalance en los datos de prueba, ya que un modelo que predice siempre la clase mayoritaria puede tener una

exactitud alta. Esto puede dificultar la evaluación real del rendimiento del modelo (Yanminsun et al., 2011).

3. **Dificultad en el entrenamiento:** Las redes neuronales profundas requieren de grandes cantidades de datos de entrenamiento para poder entender patrones complejos. Si una de las clases tiene muy pocos ejemplos, es muy probable que el modelo no sea capaz de predecirla correctamente (Leevy et al., 2018)

Para abordar estos problemas se podrían considerar varias estrategias, como hacer *oversampling* o *undersampling*. El *oversampling* consiste en generar datos sintéticos a partir de los que ya tenemos para balancear las clases, pero esto no es una buena técnica cuándo los datos son complejos (como es el caso de un ECH), ya que no hay una técnica clara para crear datos sintéticos coherentes.

Por otro lado, el *undersampling* hace que todas las clases se queden con el mismo número de candidatos que la clase minoritaria, lo que no es una técnica adecuada cuándo los datos de entrenamiento son reducidos desde un principio.

3.2. Procesamiento de los datos

Como es habitual en el campo de la inteligencia artificial, antes de poder utilizar unos datos hay que hacer cierto procesamiento para asegurarnos que son adecuados.

Lo primero que habría que hacer es quitar los datos repetidos o con valores inválidos, incompletos o corruptos, pero afortunadamente la base de datos que estamos utilizando ya ha sido revisada por sus creadores, por lo que podemos obviar este paso.

En procesamiento de señales (especialmente en señales que son muy sensibles a determinadas perturbaciones, como es el caso de los ECGs) es muy importante aplicar determinados filtros antes de trabajar con las señales. En este trabajo utilizaremos los scripts que se utilizaron en el trabajo de González Cabeza (2024) (que nos han sido facilitados por el autor). En concreto, los datos se pre-procesan de la siguiente manera:

- Se normalizan todos para tener una frecuencia de 400Hz, que es con la que se entrenó al modelo original. Por tanto, tras hacer este procesamiento previo estaremos trabajando con vectores de 4096
- Se elimina el desplazamiento de la línea base. Como podemos ver en Chouhan y Mehta (2007), es muy importante hacer esto antes de analizar un ECG.
- Se elimina la interferencia de la línea de alimentación, lo que también es importante como podemos ver en González et al. (2005)

Por último, separamos los datos en tres conjuntos, siguiendo la división recomendada por la propia base de datos:

- **Entrenamiento (*train*):** El conjunto mayoritario (con un 80 % de los datos), que será usado para entrenar al modelo
- **Validación (*validation*):** Este conjunto (que representa el 10 % de los datos) se utilizará para ajustar los parámetros del modelo en el entrenamiento del mismo.
- **Pruebas (*test*):** Este conjunto (que está formado por el 10 % restante de los datos) es el que utilizaremos para obtener las diversas métricas de rendimiento del modelo.

3.3. Métricas

Antes de entrenar diversos modelos, tenemos que tener claro qué métricas estamos intentando maximizar, ya que no hay una métrica objetivamente mejor que las demás.

3.3.1. Métricas habituales

Entre las métricas más habituales podemos encontrar la *F-β Score*, *precision* y *recall*.

Precision (Precisión)

La precisión es la proporción de predicciones positivas que son realmente positivas, o más concretamente:

$$\text{precision} = \frac{\text{Verdaderos positivos}}{\text{Verdaderos positivos} + \text{falsos positivos}}.$$

Un valor alto de esta métrica indica que el modelo es bueno minimizando falsos positivos, es decir, cuándo el modelo predice que un dato no pertenece a una clase, esa predicción es fiable.

Recall (Sensibilidad)

La sensibilidad mide la proporción de casos positivos que el modelo predice correctamente, o más concretamente:

$$\text{recall} = \frac{\text{Verdaderos positivos}}{\text{Verdaderos positivos} + \text{falsos negativos}}.$$

Un valor alto de esta métrica indica que el modelo es bueno minimizando falsos positivos, es decir, cuándo el modelo predice que un dato pertenece a una clase, esa predicción es fiable.

F-β Score

El F-β Score es una media entre la precisión y el recall, la fórmula concreta es:

$$F_\beta = (1 + \beta^2) \times \frac{\text{precision} \times \text{recall}}{\beta^2 \times \text{precision} + \text{recall}}.$$

El valor más habitual para esto es $\beta = 1$, que nos da la media armónica y permite valorar tanto la fiabilidad del modelo cuando predice positivo como negativo.

Esto es adecuado cuándo, por la naturaleza de un problema, el coste de los falsos positivos es similar al de los falsos negativos, pero no es nuestro caso. En modelos aplicados a la salud, es mucho más importante predecir las anomalías correctamente (ya que de esto puede depende la salud de una persona) que predecir correctamente la ausencia de anomalías.

Los valores de $\beta = 0{,}5, 2$ hacen que tenga más peso la precisión y el recall respectivamente, por lo que la primera es más adecuada para cuándo los falsos positivos tienen un coste muy alto y la segunda para cuándo son los falsos negativos los que tienen el coste más alto.

3.3.2. Cálculo por clase vs. Promedio binario

Todas las métricas que hemos listado anteriormente están definidas para clasificadores binarios, pero nuestro clasificador es multietiqueta, por lo que es necesario adaptarlas. En este trabajo consideraremos dos enfoques, el cálculo por clase y el promedio binario.

Cálculo por clase

Este es el enfoque más sencillo de todos. Consideramos nuestro clasificador multietiqueta como uno binario para cada una de sus etiquetas, y calculamos las métricas para cada una de las clases.

Este enfoque permite ver el desempeño del modelo en cada una de sus clases, lo que permite entender mejor cuáles son sus debilidades y fortalezas. El principal problema que presenta este método es que no da un único valor para comparar modelos, por lo que puede ser difícil determinar qué modelo es el óptimo.

Promedio binario

Este enfoque (también conocido como *one-vs-rest*) consiste en hacer el promedio de las métricas obtenidas en el enfoque anterior para cada clase.

3.3.3. Métricas para nuestro problema

Tras realizar el análisis de las posibles métricas que implementar, hemos decidido calcular y mostrar varias métricas para cada modelo, y elegir una que consideramos

mejor para afirmar qué modelo es el mejor. Las métricas que mostraremos son las siguientes:

- Para cada una de las clases:
  - Precisión.
  - Recall.
  - F-1 Score.
- Precisión global calculada como promedio binario.
- Recall global calculado como promedio binario.
- F-1 Score global calculado como promedio binario.
- F Score ajustada, una métrica personalizada que definiremos a continuación.

Todas las métricas que mostraremos, salvo la personalizada, tienen el objetivo de entender mejor cómo funciona el modelo, pero no de compararlo. La F Score ajustada será la que utilizaremos para determinar qué modelo consideramos óptimo.

En nuestro problema tenemos cinco etiquetas. Una de ellas representa un ECH normal, mientras que las demás representan diversas anomalías. Dado que nuestro objetivo es que el modelo identifique lo mejor posible las anomalías (cuándo las haya), el coste de los falsos negativos en las etiquetas de anomalías es muy alto, mientras que el coste de los falsos positivos en la etiqueta normal es muy alto.

Por ello, definiremos la F Score ajustada como la media de la F-0.5 Score de la clase normal y las F-2 Score del resto de etiquetas. Esto nos permite tener una métrica que tiene en cuenta tanto reducir falsos negativos como falsos positivos en todas las etiquetas, pero dando más peso a los falsos negativos o positivos dependiendo de la etiqueta concreta.

La fórmula concreta de la métrica sería:

$$\text{F Score ajustada} = \frac{F - 0{,}5(\text{NORM}) + \sum_{i \neq \text{NORM}} F - 2(i)}{\text{Número de clases}}$$

3.4. Transformaciones

Capítulo 4

Conclusiones y Trabajo Futuro

Conclusiones del trabajo y líneas de trabajo futuro.

Antes de la entrega de actas de cada convocatoria, en el plazo que se indica en el calendario de los trabajos de fin de grado, el estudiante entregará en el Campus Virtual la versión final de la memoria en PDF.

Introduction

Introduction to the subject area. This chapter contains the translation of Chapter 1.

Conclusions and Future Work

Conclusions and future lines of work. This chapter contains the translation of Chapter 4.

Bibliografía

Y así, del mucho leer y del poco dormir, se le secó el celebro de manera que vino a perder el juicio.
(modificar en Cascaras\bibliografia.tex)

Miguel de Cervantes Saavedra

Chouhan, V. y Mehta, S. Total removal of baseline drift from ecg signal. En *2007 International Conference on Computing: Theory and Applications (ICCTA'07)*, páginas 512–515. 2007.

González Cabeza, S. Are 12-lead ecgs necessary for detecting heart anomalies? a comparison between 12-lead and 3-lead ecg classification using deep learning, 2024. No Publicado.

González, L. E. A., S., J. L. R. y Castellanos, G. Métodos para la eliminación de interferencia ac en ecg. *Scientia et Technica*, vol. 3(29), páginas 49–53, 2005. ISSN 0122-1701. Accedido el 5 de diciembre de 2024.

He, H. y Garcia, E. A. Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, vol. 21(9), páginas 1263–1284, 2009.

Leevy, J. L., Khoshgoftaar, T. M., Bauder, R. A. y Seliya, N. A survey on addressing high-class imbalance in big data. *Journal of Big Data*, vol. 5(1), página 42, 2018.

Ribeiro, A. H., Paixao, G. M., Lima, E. M., Horta Ribeiro, M., Pinto Filho, M. M., Gomes, P. R., Oliveira, D. M., Meira Jr, W., Schon, T. B. y Ribeiro, A. L. P. CODE-15 %: a large scale annotated dataset of 12-lead ECGs. `https://doi.org/10.5281/zenodo.4916206`, 2021a.

Ribeiro, A. H., Ribeiro, M. H., Paixão, G. M. M., Oliveira, D. M., Gomes, P. R., Canazart, J. A., Ferreira, M. P. S., Andersson, C. R., Macfarlane, P. W., Meira Jr., W., Schön, T. B. y Ribeiro, A. L. P. Automatic diagnosis of the 12-lead ECG using a deep neural network. *Nature Communications*, vol. 11(1), página 1760, 2020. Aviable at `https://doi.org/10.1038/s41467-020-15432-4`.

Ribeiro, A. H., Ribeiro, M. H., Paixão, G. M., Oliveira, D. M., Gomes, P. R., Canazart, J. A., Ferreira, M. P., Andersson, C. R., Macfarlane, P. W., Jr., W. M., Schön, T. B. y Ribeiro, A. L. P. CODE dataset. https://figshare.scilifelab.se/articles/dataset/CODE_dataset/15169716, 2021b.

Wagner, P., Strodthoff, N., Bousseljot, R.-D., Samek, W. y Schaeffter, T. PTB-XL, a large publicy available electrocardiography dataset. https://doi.org/10.13026/kfzx-aw45, 2022. (version 1.0.3). Último acceso: 28/09/2024.

Yanminsun, Wong, A. y Kamel, M. S. Classification of imbalanced data: a review. *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 23, 2011.

Apéndice A

Título del Apéndice A

Los apéndices son secciones al final del documento en las que se agrega texto con el objetivo de ampliar los contenidos del documento principal.

Apéndice B

Título del Apéndice B

Se pueden añadir los apéndices que se consideren oportunos.

Este texto se puede encontrar en el fichero Cascaras/fin.tex. Si deseas eliminarlo, basta con comentar la línea correspondiente al final del fichero TFGTeXiS.tex.

–¿Qué te parece desto, Sancho? – Dijo Don Quijote –
Bien podrán los encantadores quitarme la ventura,
pero el esfuerzo y el ánimo, será imposible.

Segunda parte del Ingenioso Caballero
Don Quijote de la Mancha
Miguel de Cervantes

–Buena está – dijo Sancho –; fírmela vuestra merced.
–No es menester firmarla – dijo Don Quijote–,
sino solamente poner mi rúbrica.

Primera parte del Ingenioso Caballero
Don Quijote de la Mancha
Miguel de Cervantes