Propuesta de Proyecto: Herramienta de Aprendizaje de Máquinas para el diagnóstico de Distrofias Musculares

Camila Bergasa - Juan Pablo Cabeza - Sebastián García

Abstract

La distrofía muscular, corresponde a un grupo de enfermedades que provocan debilidad progresiva y pérdida de masa muscular. Existen varios tipos de distrofia muscular y los síntomas se presentan mayormente durante la infancia.

El objetivo de nuestro proyecto, es desarrollar una herramienta basada en modelos de clasificación, que sea capaz de reconocer patrones en los datos entregados y, por lo tanto, ayudar en el diagnóstico de la enfermedad. Estos metodos serán la base para mejorar el sistema a futuro.

1. Introducción

El trabajo a desarrollar se contextualiza en el curso de Aprendizaje de Máquinas y la motivación del grupo es aportar al desarrollo de herramientas que permitan facilitar y apoyar el diagnóstico de enfermedades, nos motiva aplicar lo aprendido en el curso en el área de la salud, entendiendo la importancia que suscita el uso de la tecnología al servicio de las personas.

Por otro lado, en general el diagnóstico de las distrofias musculares siempre se ha guiado por las prestaciones clínicas, las biopsias musculares y los datos que aporta el resonador magnético, que suguieren una correlación entre la infiltración de grasa muscular y el diagnóstico de la enfermedad. El problema de esto, es que aquellos patrones de infiltración de grasas muchas veces se traslapan entre sí y pueden ser fácilmente confundidos con otros trastornos, por lo que se necesitan patrones más precisos y específicos para diagnosticar la enfermedad.

2. Desarrollo

El Dataset fue obtenido desde la plataforma *Dryad*, la cual corresponde a un proyecto de código abierto impulsado por la comunidad que adopta un enfoque único para la publicación de datos y su preservación digital. El primer paso fue cargar el Dataset, donde se aprecian los distintos resultados de las resonancias magnéticas de los músculos inferiores, es decir, de las piernas, pelvis, glúteos y muslos de 976 pacientes con distintos tipos de distrofias musculares, todos con diagnósticos confirmados de la enfermedad.

Luego, se genera una función que permite vizualizar la media y la distribución de los 70 músculos a estudiar para cada paciente con su respectiva infiltración de grasa, en el eje-x se aprecian los nombres de los músculos (en latín) y en el eje-y se distingue la infiltración de grasa expresada en una escala de 0 a 4 (Escala de Mercury), donde 0 es nada de infiltración y 4 corresponde a la máxima infiltración.

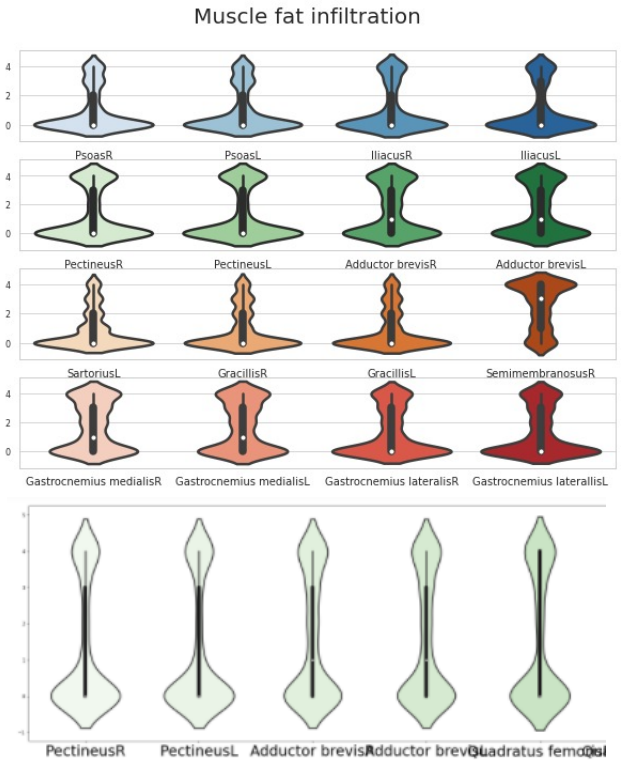

Figure 1. La imagen muestra algunos de los 70 músculos a estudiar y corresponde al promedio de todas las muestras

2.1. Metodología

Primero obtenemos y realizamos una exploración del Dataset, teniendo en cuenta la limpieza de los datos, que representa cada uno de los features y el significado de cada

variable, que en este caso son la musculatura inferior de los pacientes y el grado de infiltración de grasa. Luego, se detectan las variables de interés y su grado de correlación con el subtipo de la enfermedad y se aplican distintos algoritmos (en particular, nos centraremos en el uso de Naive-Bayes y Support Vector Machine), seleccionando aquellos que creemos se ajusten mejor, para finalmente proceder a la elección del modelo y a la interpretación de los resultados.

La metodología de trabajo a utilizar, se resume en el siguiente esquema:

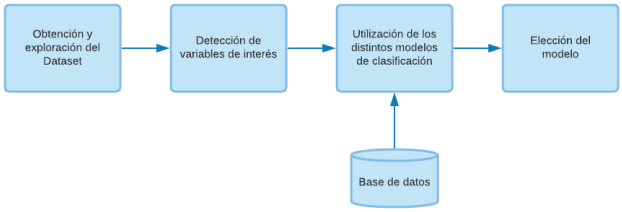

Figure 2. Esquema de la Metodología de trabajo

3. Resultados y Análisis

Debido las dimensiones de nuestro Dataset, se divieron los 976 pacientes en los 10 tipos de distrofia muscular existentes y fue necesario implementar tSNE desde la librería sklearn, esto con el fin visualizar en 2 o 3 dimensiones y explorar variables que puedan ser agrupadas por un mismo feature.

Para ello se realiza una selección de features implementados a partir de sklearn, donde para seleccionar los músculos con mayor importancia relativa para una futura clasificación, estos son ordenados jerarquicamente según su coeficiente de correlación de Pearson entre cada variable. Finalmente quedan con el siguiente orden[1]: Tensor fasciae latae R (Lado Derecho), Tensor fasciae latae L (Lado Izquierdo), Obturatorious externus L, Obturatorious externus R, etc.

[1]Gráfico se adjunta en Referencias

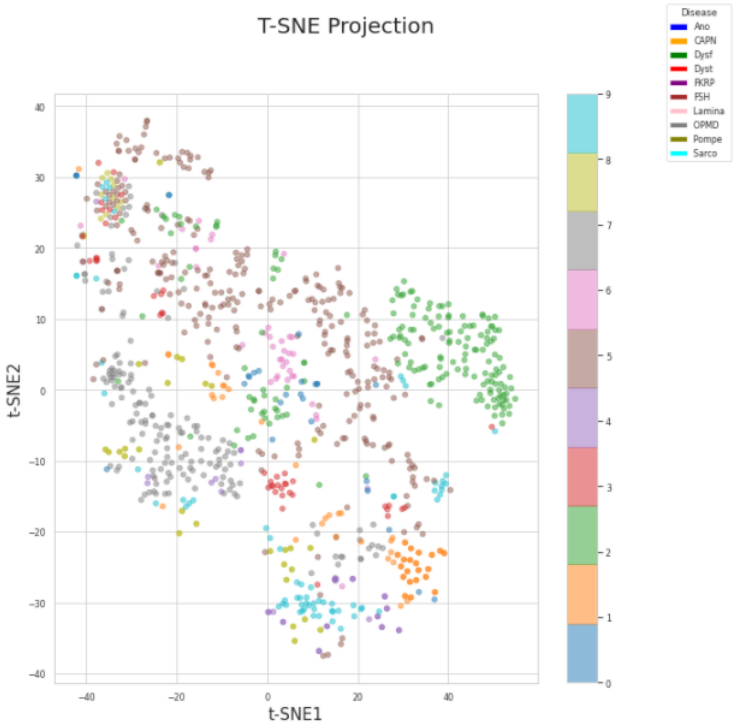

Figure 3. Proyección de tSNE a partir de los 10 tipos de Distrofia Muscular

Además, al ejecutar un modelo de clasificación predefinido por sklearn como Random Forest, como un primer acercamiento al análisis futuro, con los datos agrupados según cada tipo de distrofia muscular, se obtienen los siguientes resultados:

	Dystrophin	*LMNA*	Calpain	Dysferlin	Sarcoglycans	*FKRP*	*ANO5*	Pompe	FSHD	OPMD	Mean value
Accuracy	0.94	0.95	1	0.97	0.91	0.92	0.92	1	0.98	0.97	0.95
Sensitivity	0.88	0.90	1	0.96	0.83	0.8	0.85	1	1	0.97	0.91
Specificity	1	1	1	0.99	1	1	1	1	0.97	0.97	0.99
PPV	1	1	1	0.96	1	1	1	1	0.94	0.89	0.97
NPV	0.99	0.99	1	0.99	0.98	0.99	0.99	1	1	0.99	0.99

Figure 4. Analisis preliminar de clasificación agrupada por tipo de Distrofia

4. Trabajo a Futuro

Según lo expuesto anteriormente, se plantea la problemática de interiorizarse con el lenguaje técnico del dataset y de la literatura existente al respecto, ya que en ocasiones escapa a nuestra formación como ingenieros. Se plantea además que lo que se busca estudiar es la aplicación de un modelo que permita obtener una mayor presición de los patrones de infiltración de grasa muscular en los pacientes con distrofias musculares y para ello proponemos abordar el problema

aplicando los siguientes algoritmos vistos en clases: Inferencia Bayesiana (Naive Bayes), Support Vector Machine con una respectiva reducción de dimensionalidad, Clustering y Redes Neuronales con el objetivo de escoger mediante la selección de los parámetros relevantes aquel que mejor se ajuste a nuestro objetivo o bien el que nos proporcione un mayor poder interpretativo.

Otra problemática a abordar es cómo nos hacemos cargo de las infiltraciones (patrones) que puedan pertenecer a otro tipo de trastornos.

Se deben especificar además las métricas que serán de utilidad al momento de seleccionar el modelo, como presición o sensibilidad.

5. Conclusiones

Se requiere que se estudie con mayor detalle acerca de la enfermedad y sus subtipos con el objetivo de obtener un mayor poder interpretativo del problema al momento de aplicar los algoritmos mencionados.

A priori, concluimos que probablemente se requiera de opinión experta para probar la presición del diagnóstico y comparar aquella opinión con el diagnóstico que realice el software.

Se desprende además que las herramientas de Aprendizaje de Máquinas son útiles para apoyar el diagnóstico de cualquier tipo de enfermedad en general y por lo tanto, la toma de desiciones respecto del tratamiento a utilizar y la calidad de vida de las personas.

Software and Data

Se utiliza *Python* para exploración del Dataset y para la aplicación de los modelos mencionados en la sección de trabajo a futuro.

Referencias

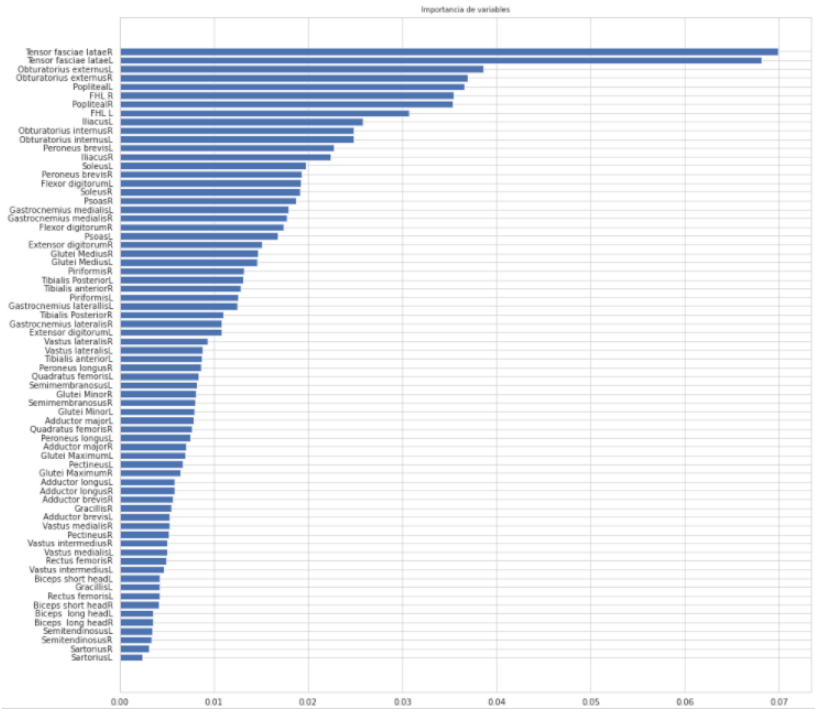

Figure 5. Importancia relativa de las variables