Instituto de Biomecánica Clínica

Pre-Fall – Sistema inteligente para la prevención y predicción de caídas

E3.1 – Procedimiento de depuración y preprocesado de los datos

Proyecto	Pre-Fall – Sistema inteligente para la prevención y predicción de caídas
Entregable	E3.1 – Procedimiento de depuración y preprocesado de los datos

Contenido

Índice de tablas

Índice de figuras

1 Introducción

Este entregable está enmarcado en la tarea "T3.1: Depuración y preprocesado de los datos", perteneciente al paquete de trabajo "PT3 – Sistema experto de prevención de caídas" dentro del proyecto PRE-FALL. A lo largo del documento, se procederá a describir cada uno de los pasos a realizar con los datos recogidos en el paquete de trabajo previo "PT2 – Wearable de medición de la marcha humana", de forma previa al desarrollo del diseño y de los modelos de aprendizaje automático del mismo paquete aquí abordado.

En un primer nivel, se describirán los datos iniciales recogidos para el inicio del desarrollo de los modelos de *machine learning* posteriores. A continuación, los diferentes procesos para la adecuación de éstos, así como la metodología desarrollada para la identificación de las diferentes fases de la marcha.

2 Descripción de los datos

A lo largo de este apartado, se realiza un análisis descriptivo de los diferentes conjuntos de datos recogidos mediante la sensórica diseñada para tal fin en el paquete de trabajo previo "PT2 – Wearable de medición de la marcha humana" y recogidos en la tarea "T2.3: Captura de datos inicial" de dicho paquete.

Los datos aquí presentados han sido recogidos entre 2 individuos bajo un escenario que consiste en caminar a la velocidad natural de cada persona durante aproximadamente 5 segundos. En la Tabla 1 se presenta un resumen de los diferentes conjuntos de datos generados.

ID	Tamaño (KB)	Duración (segundos)
USER1	64,90	5,03
USER2	61,20	5,30
TOTAL	126,10	10,33

Tabla 1. Resumen de los datos obtenidos de las mediciones.

Dichos datos serán complementados por datos artificiales generados de cara a desarrollar posteriormente lo modelos de predicción y el sistema experto para las tareas "T3.2: Modelos de aprendizaje automático para el modelado de la marcha humana" y "T3.4: Diseño e implementación de un sistema experto de prevención y evaluación de la eficacia de la rehabilitación" respectivamente.

3 Pre-procesado

El primer paso del proceso de detección de las fases de la marcha consiste en la preparación de los datos. Estos consisten en mediciones de acelerómetro, giroscopio y magnetómetro en los 3 ejes (X, Y, Z) con una frecuencia de captación de datos de 100Hz por lo que cada elemento del conjunto se ha tomado cada 0,01 segundos.

Dichos datos se han generado como ficheros de texto con una estructura específica, por lo que ha sido necesario adaptarlos para poder realizar el pre-procesado. La siguiente imagen muestra la estructura de uno de los ficheros:

```
D:\Pacientes\TEST INERCIALES\Ana\ana_marcha.txt D:\Pacientes\TEST INERCIALES\Ana\ana_marcha.txt D:\Pacientes\TEST INERCIALES\Ana\ana_marcha.txt
D:\Pacientes\TEST INERCIALES\Ana\ana_marcha.txt D:\Pacientes\TEST INERCIALES\Ana\ana_marcha.txt D:\Pacientes\TEST INERCIALES\Ana\ana_marcha.txt D:\P
D:\Pacientes\TEST INERCIALES\Ana\ana_marcha.txt D:\Pacientes\TEST INERCIALES\Ana\ana_marcha.txt D:\Pacientes\TEST INERCIALES\Ana\ana_marcha.txt D:\P
    ACC_X    ACC_Y    ACC_Z   ACCL_X   ACCL_Y   ACCL_Z    GYR_X    GYR_Y    GYR_Z    MAG_X    MAG_Y    MAG_Z
    ANALOG   ANALOG   ANALOG   ANALOG   ANALOG   ANALOG   ANALOG   ANALOG   ANALOG   ANALOG   ANALOG   ANALOG
    ORIGINAL ORIGINAL ORIGINAL ORIGINAL ORIGINAL ORIGINAL ORIGINAL ORIGINAL ORIGINAL ORIGINAL ORIGINAL ORIGINAL
ITEM X        X        X        X        X        X        X        X        X        X        X
1    0.467781 -8.287636 -2.339395 0.088776 -1.006097 -0.740701 17.997953 3.420494  20.367746 -9.061678 -3.975446 -103.010811
2    0.666454 -8.208273 -2.122872 -0.131765 -1.097821 -0.923602 18.821432 7.100275  12.772273 -9.295527 -3.829290 -103.156967
3    0.826864 -8.089148 -1.919419 -0.300637 -1.227078 -1.097423 16.736357 12.304450 6.080693  -9.295527 -3.829290 -103.156967
```

Como se puede observar es necesario modificar la cabecera de los mismos para adecuar las columnas a los datos generados. Tras eliminar la línea inicial y colocar las columnas en sus respectivas posiciones el fichero puede leerse como csv y ser tratado con la librería Pandas de Python.

	ITEM	ax	ay	az	gx	gy	gz	mx	my	mz
0	1	1.274490	-9.410679	0.344289	10.824789	-43.355045	10.704852	-13.446361	-16.106401	-55.422390
1	2	1.425975	-9.379428	0.391114	12.094811	-44.634823	11.239221	-13.007893	-15.872552	-54.837765
2	3	1.464409	-9.281487	0.416243	11.988267	-45.888840	11.822715	-13.007893	-15.872552	-54.837765
3	4	1.566501	-9.301020	0.397836	11.803392	-46.485882	11.893462	-13.066355	-16.954107	-54.925461
4	5	1.713739	-9.444021	0.300039	12.393333	-45.603924	10.887078	-13.066355	-16.281788	-55.188541
...	...	...	...	...	...	...	...	...	...	...
493	494	-1.318581	-9.516723	-1.982684	-8.866790	14.584185	-26.990795	-5.583163	-11.429406	-52.996201
494	495	-0.936866	-10.842125	-2.113900	-6.259973	17.113083	-20.757370	-5.612394	-10.903245	-53.259281
495	496	-0.681724	-11.796433	-2.077904	-3.970574	17.363491	-11.631232	-5.875475	-11.078631	-53.756210
496	497	-0.485800	-12.225286	-1.961918	-1.170135	15.799289	-3.436223	-6.050862	-11.546331	-53.580822
497	498	-0.276537	-12.366166	-1.809829	3.211411	15.341651	3.140871	-6.050862	-11.546331	-53.580822

498 rows × 10 columns

Una vez tenemos los datos en el formato pandas deseado, podemos visualizar la información en forma de gráfica para obtener una idea general de los datos. La detección de la marcha se realizará sobre todo a partir de los datos del acelerómetro por lo que se ha decidido analizar fundamentalmente los datos del eje X, Y y Z de este sensor. A continuación, se muestra una gráfica con los valores de una de las muestras proporcionadas:

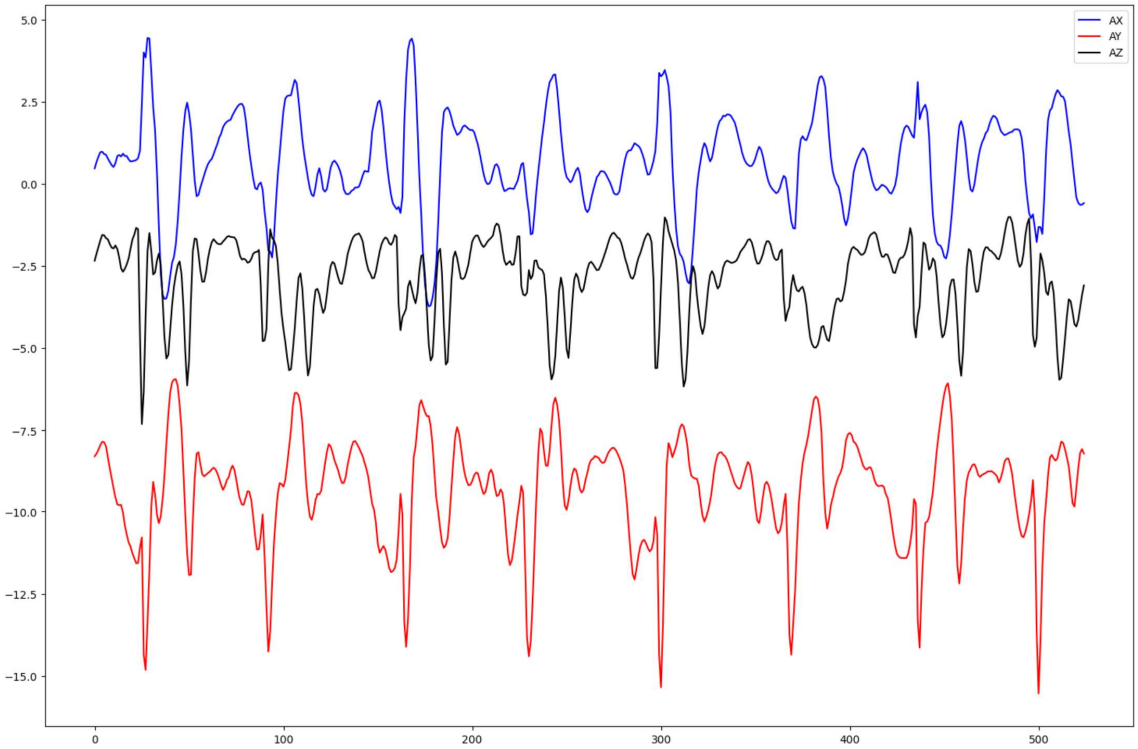


Figura 1. Datos recogidos mediante acelerómetro (Eje X en azul, eje Y en rojo y eje Z en negro)

3.1 Filtro de sensores

Dada la naturaleza del proceso de recogida de datos a través de sensores, es posible que existan ciertos valores anómalos en el conjunto de datos que afecten negativamente al rendimiento del modelo posteriormente. Por lo tanto, es necesario utilizar un filtro sobre el conjunto de datos para suprimir todos aquellos valores atípicos.

Para ello se ha considerado el siguiente intervalo de confianza:

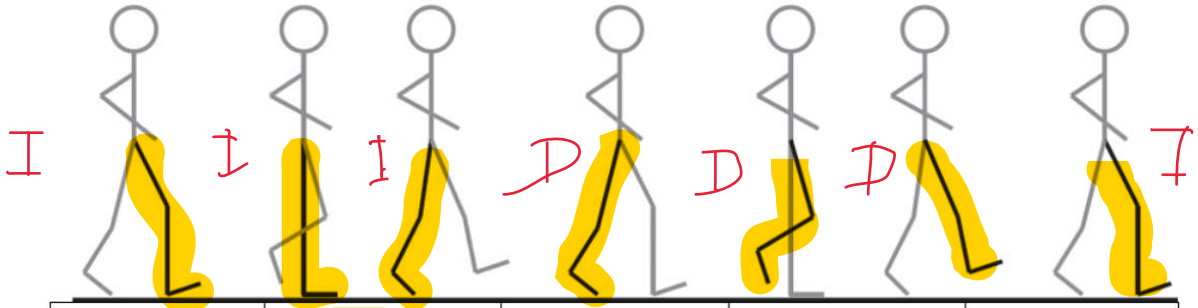
$$[\bar{M} - 8S_M, \bar{M} + 8S_M],$$

donde M representa la medida analizada en los tres ejes del acelerómetro, $\bar{M}$ es la media muestral, y S_M la desviación típica muestral.

Todos aquellos valores que se encuentren fuera de este intervalo serán sustituidos por otros obtenidos a partir de calcular la media con respecto a los valores anterior y posterior más cercanos y que no sean nulos.

4 Detección de la marcha

Las cuatro fases asociadas a la marcha humana consisten en las diferentes posiciones de las extremidades inferiores a lo largo del ciclo completo de la marcha. En la siguiente figura se presentan esquemáticamente cada una de dichas cuatro fases y cómo transicionan entre sí.



Soporte de ambas extremidades (extremidad izquierda delante)	Soporte de extremidad izquierda y balanceo de pierna derecha	Soporte de ambas extremidades (extremidad derecha delante)	Soporte de extremidad derecha y balanceo de pierna izquierda	Soporte de ambas extremidades (extremidad izquierda delante)
Fase 1	**Fase 2**	**Fase 3**	**Fase 4**	**Fase 1**

Figura 2. Posición corporal a lo largo de las diferentes fases de la marcha.

Tal y como se puede observar en dicho esquema, la transición de las cuatro fases ha de seguir la progresión marcada en éste (1 – 2 – 3 – 4 – 1). A partir de los datos recogidos a través del acelerómetro del sensor, se determinará en qué fase se encuentra el usuario en cada momento. Para ello, se detallan en los siguientes sub-apartados las diferentes fases de dicha identificación.

4.1 Apoyos (Fases 1/3 y 2/4)

De cara a detectar el número de apoyos del usuario en cada momento, es fundamental identificar si el usuario tiene ambas extremidades apoyadas en el suelo o si una de ellas se encuentra en fase de balanceo. Para ello se hace uso del eje Y del acelerómetro, que mide las aceleraciones verticales del usuario. En las fases de balanceo, se observan valores bajos de aceleración este eje, mientras que en las fases de doble soporte se presenta el comportamiento opuesto, lo que nos permite distinguir si el usuario se encuentra en las fases 1/3 (fases de doble soporte) o en las fases 2/4 (fases de soporte único).

Antes de trabajar con la serie temporal obtenida es necesario realizar un preprocesado de la misma. Este consiste en aplicar una función de suavizado con un número de grados de libertad que se ha definido en función del tamaño de la serie. En este caso se han definido 30 grados de libertad.

Una vez procesada la serie temporal, se pueden identificar los tramos en los que se distinguen ambos pares de fases siendo los tramos donde la serie suavizada sea superior al percentil 60 se

las de doble soporte, mientras que los que tengan valores inferiores se verán asociados a las fases de soporte único.

En la siguiente imagen, se puede observar la serie original del acelerómetro en el eje Y después de haber sido normalizada a media cero, así como la serie suavizada. Es importante destacar que se ha utilizado la asociación de colores para diferenciar ambas tipologías de fase: doble soporte (azul) y soporte único (rojo).

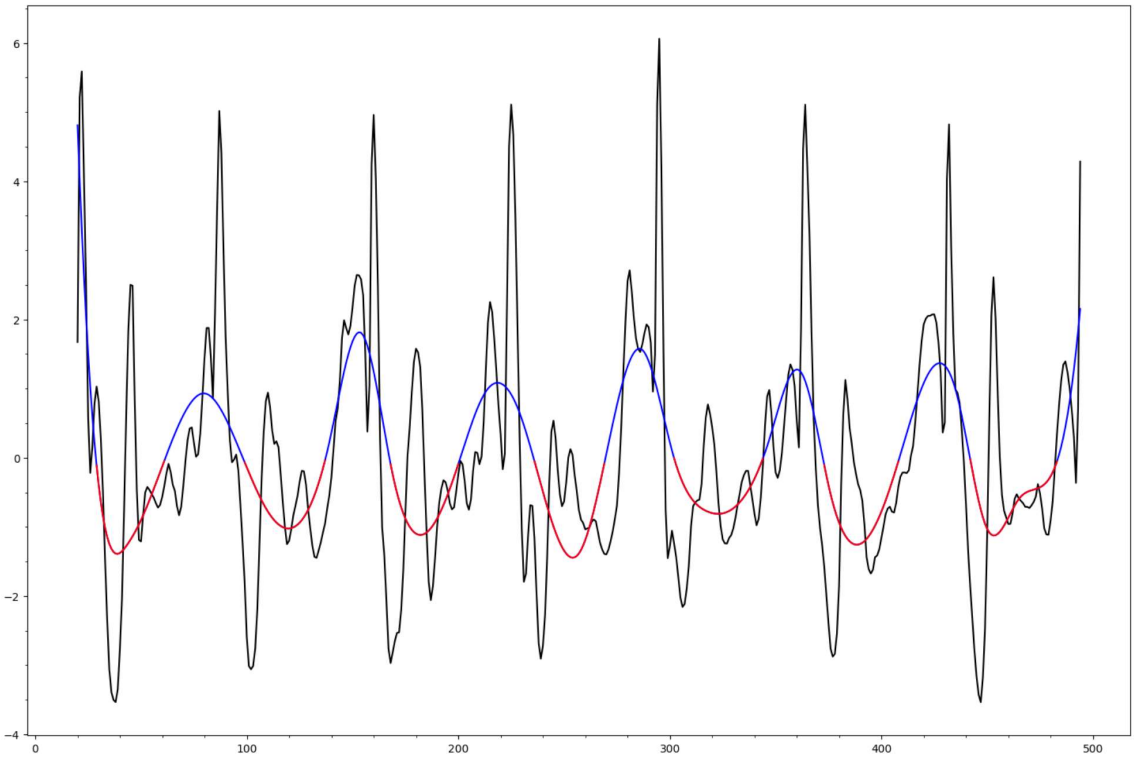


Figura 3. Eje Y del acelerómetro normalizado con la serie suavizada a 30 grados de libertad. En azul las fases 1/3 y en rojo las fases 2/4.

4.2 Movimiento (Fases 1/2 y 3/4)

El siguiente paso para poder realizar una distinción total de las cuatro fases es determinar el lateral del cuerpo de la fase. Para ello se toman las fases de apoyo identificadas en el punto anterior y se extrae el valor medio de los valores del eje X del acelerómetro en una ventana de 0.2 segundos centrada en el punto intermedio de cada fase.

Una vez obtenidos estos valores medios, se detecta aquél con el mayor valor absoluto y se utilizará este valor para definir la transición de estados en cada fase de manera que ésta no sea muy abrupta y tenga sentido.

En la siguiente figura se muestra la clasificación obtenida para los datos anteriores, donde se presenta la serie temporal de las aceleraciones del eje X, destacando en rojo las fases asociadas al lateral izquierdo (fases 1 y 2), y en azul las asociadas al lateral derecho (fases 3 y 4).

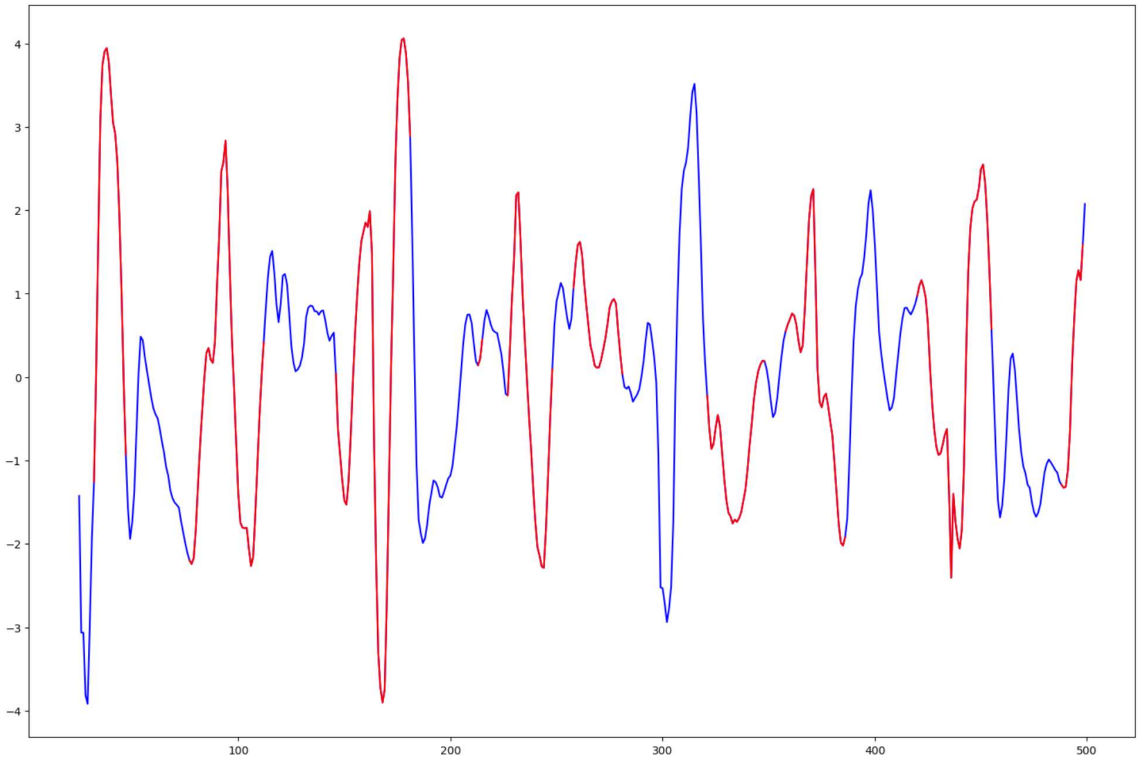


Figura 4. Eje X del acelerómetro normalizado. En rojo las fases de lateral izquierdo (1/2) y en azul las fases de lateral derecho (3/4).

4.3 Cruce de tipología (1 vs 2 vs 3 vs 4)

El siguiente paso tras detectar el número de apoyos (1/3 vs 2/4) y el lateral de referencia (1/2 vs 3/4) es realizar el cruce de ambas gráficas de manera que podamos identificar las cuatro fases de la marcha.

En la siguiente figura se muestran la información de los tres ejes del acelerómetro normalizados a media cero de la misma forma que en los ejemplos anteriores, incluyendo franjas de colores correspondientes a las cuatro fases de la marcha. El código de colores es el siguiente:

- Fase 1: rojo oscuro.
- Fase 2: rojo claro.
- Fase 3: azul oscuro.
- Fase 4: azul claro.

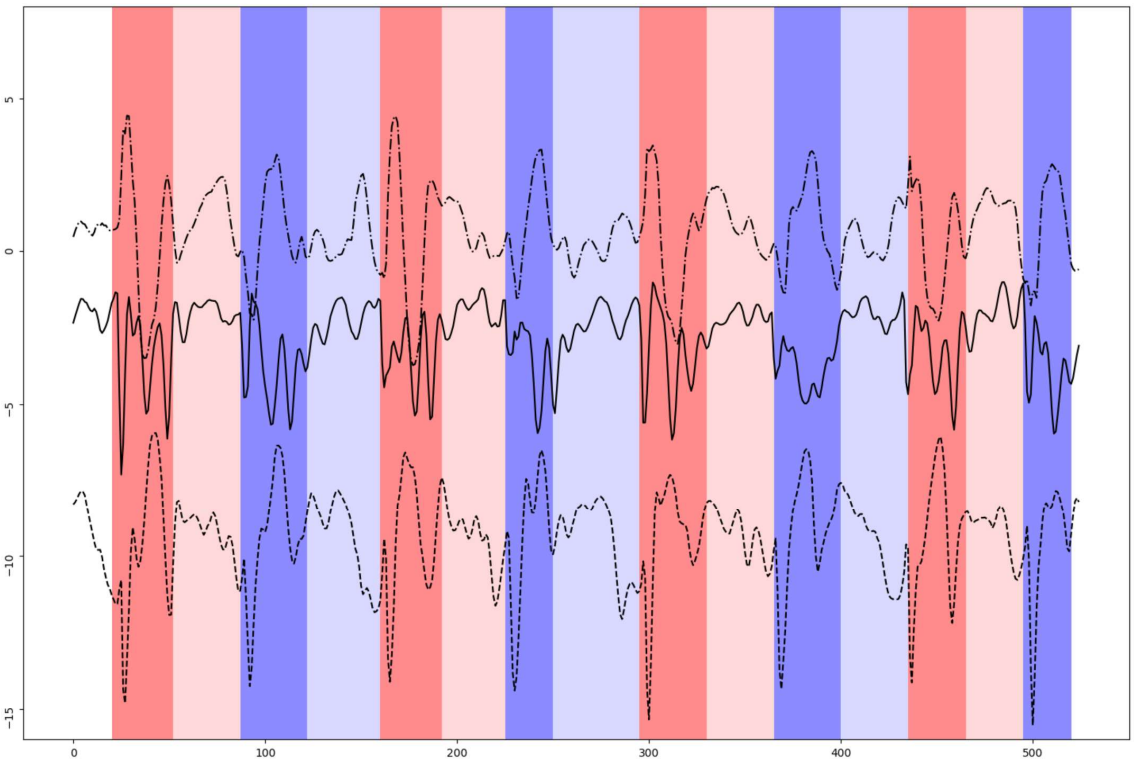


Figura 5. Valores del acelerómetro divido en franjas que representan las 4 fases de la marcha.

4.4 Definición de las fases a partir de las series temporales

Una vez se han definido en los puntos anteriores los procedimientos para identificar las diferentes fases de la marcha, se procede a aplicarlos sobre las series temporales para comprobar si estas se ajustan a los datos y poseen una estructura coherente con el esquema teórico de las fases de la marcha mencionado al comienzo de esta sección.

En la Figura 6 se han colocado de forma comparativa la gráfica original de una de las series temporales proporcionadas junto con la gráfica de las fases de la marcha detectadas. En esta se puede observar que las fases de la marcha siguen un patrón que concuerda con la marcha humana excepto en la zona central donde se observa que hay dos pisadas en un pequeño periodo de tiempo lo que concordaría con el momento el experimento en el que el paciente se gira sobre sí mismo para recorrer de nuevo el tramo.

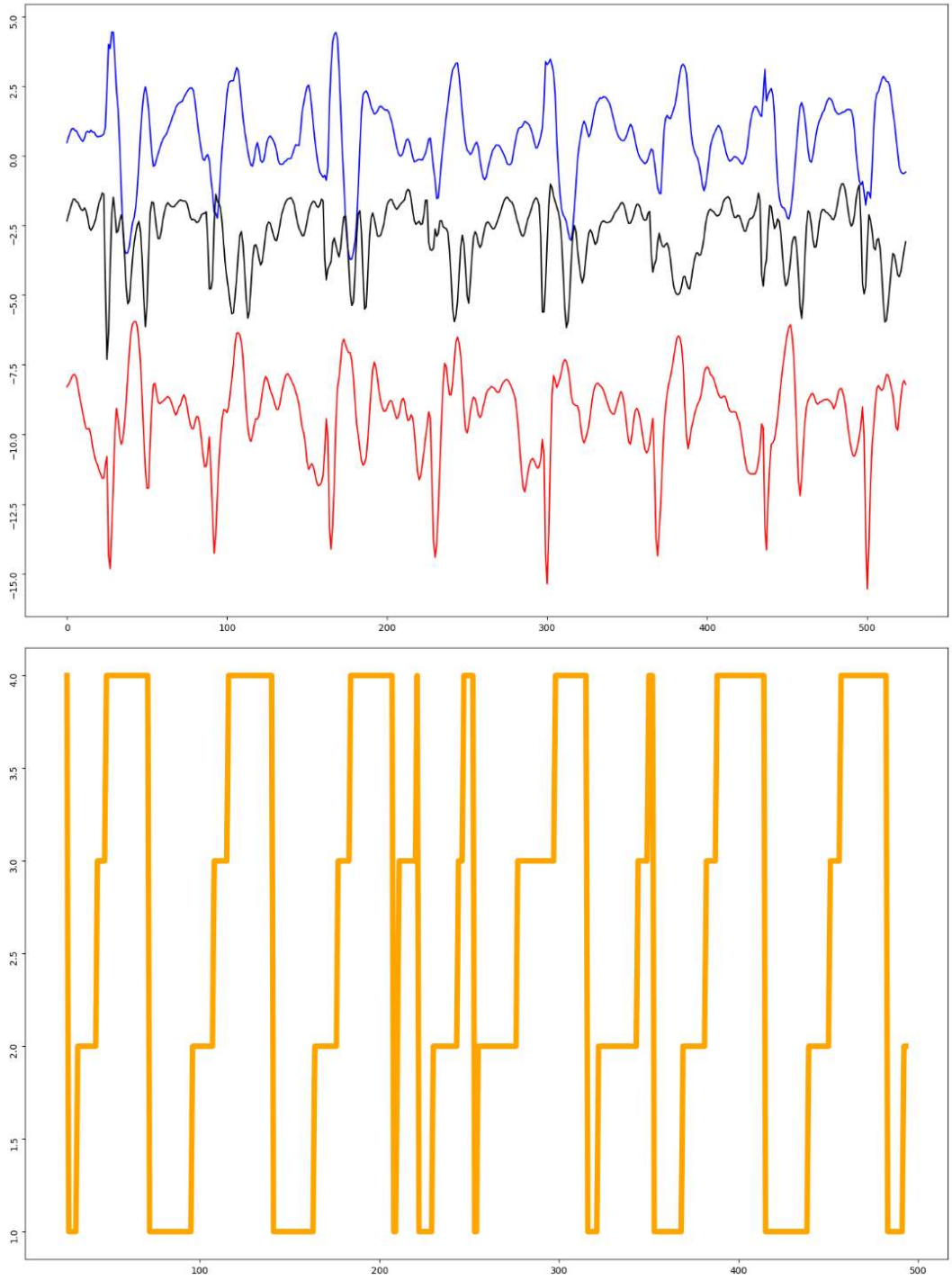


Figura 6. Gráfico de la evolución de las fases de la marcha junto con los datos originales.

4.5 Obtención de métricas de las fases de la marcha

Tras obtener las fases de la marcha a partir de los datos del acelerómetro, se ha procedido a extraer métricas de las mismas. Para ello se han tenido en cuenta todas las columnas del conjunto de datos y se han agrupado los elementos por la fase de la marcha a la que pertenecen.

Una vez agrupados se ha calculado la media y la desviación típica de cada uno de estos conjuntos para cada columna del dataset, es decir, de los valores X, Y y Z del acelerómetro, giroscopio y magnetómetro. Además, se incluye también la duración media de cada fase de la marcha.

De esta forma los valores de un fichero tendrán asociados los valores anteriores junto con la variable objetivo que indica si un paciente asociado a dicho fichero está en riesgo o no da caída formando un total de 81 variables para cada fichero de medición.

5 Generación de datos sintéticos

Una vez establecidas las técnicas para depurar los datos e identificar las diferentes fases de la marcha, deben aplicarse sobre los datos proporcionados para posteriormente entrenar un modelo que identifique el riesgo de caída. Sin embargo, actualmente sólo se dispone de 2 ficheros con mediciones de aproximadamente 5 segundos por lo que es necesario generar datos sintéticos adicionales para poder construir un modelo de aprendizaje que servirá de base para el futuro modelo alimentado con más datos reales.

Para generar estos datos se ha partido de los conjuntos de datos actuales y se ha optado por utilizar dos funciones de la librería pandas:

- *pandas.DataFrame.resample*: Utilizada para reconvertir las series actuales a otra escala temporal de acuerdo a los valores medios.
- *pandas.DataFrame.interpolate*: Para interpolar los datos de acuerdo a la escala que tenían previamente.

Esta combinación permite generar múltiples conjuntos de datos que conserven la naturaleza de los originales, pero que contengan diferente información.

Además de lo anterior, se han generado de manera aleatoria los diferentes valores de la variable de clasificación que asocie a cada conjunto de datos una clase que corresponda con la posibilidad o no de estar en riesgo de caída.

En la Figura 7 se observa una comparativa triple entre la serie temporal original de una de las muestras proporcionadas, una de las series temporales generada de forma artificial y las fases de la marcha detectadas para la muestra artificial. En esta se pueden hacer dos observaciones:

- La gráfica original y la artificial son muy similares y esto es debido a que se los datos sintéticos se han tenido que generar a partir de sólo dos series temporales reales, pero para este caso concreto nos interesa que mantenga una estructura similar más allá de la información en sí.
- La gráfica de fases de la marcha mantiene una estructura coherente lo que permite que estos datos se usen para entrenar un modelo posteriormente que se ajuste a los datos de los que disponemos.

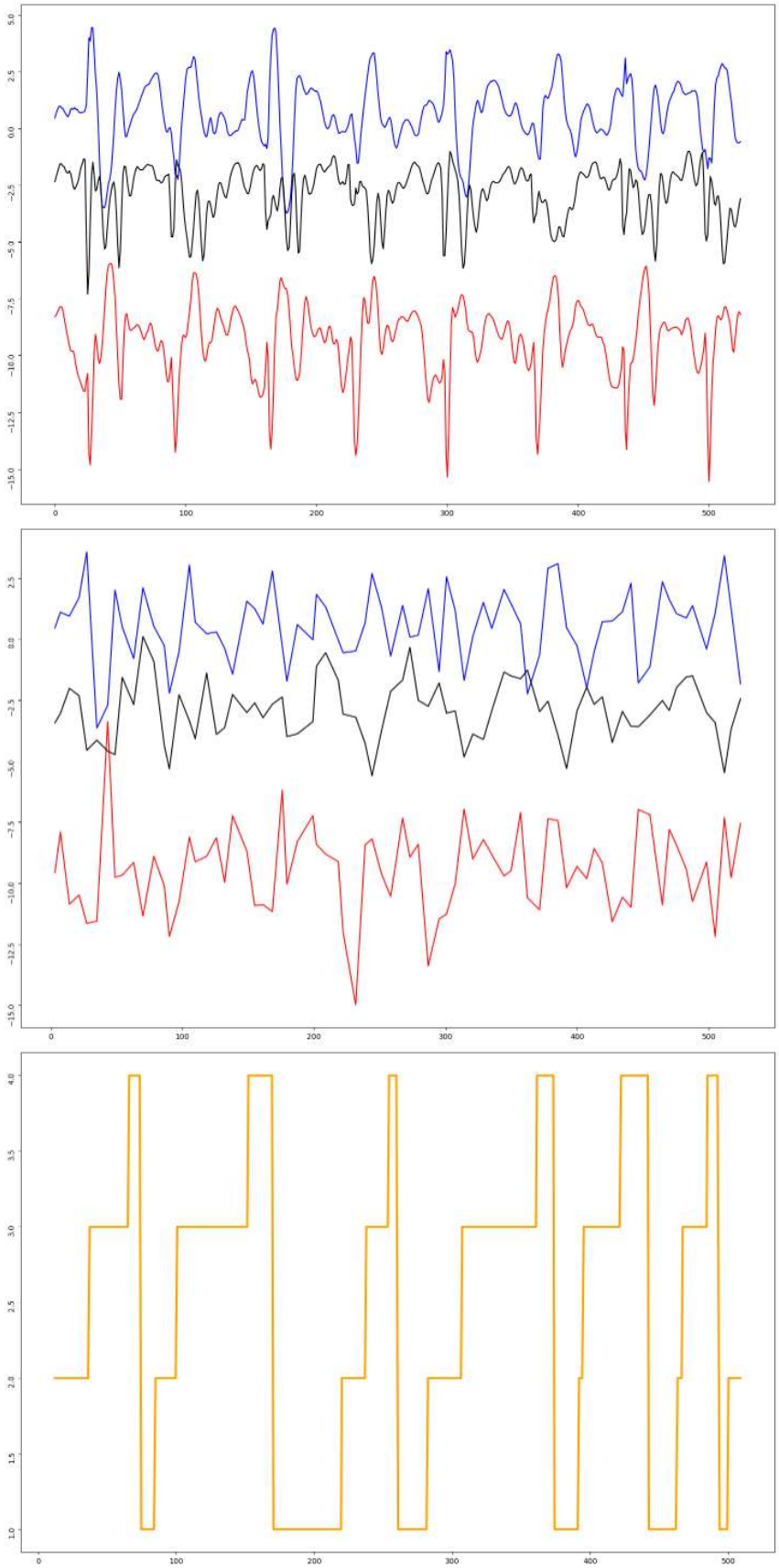


Figura 7. En la parte superior los datos originales. Debajo de este, los datos generados artificialmente a partir de los originales y en la parte inferior las fases de la marcha detectadas en los datos artificiales.

6 Conclusiones

Este entregable corresponde a la tarea "T3.1: Depuración y preprocesado de los datos", perteneciente al paquete de trabajo "PT3 – Sistema experto de prevención de caídas", abordando los pasos previos al desarrollo de los modelos de aprendizaje automático a generar a lo largo de éste.

En un primer nivel, se ha realizado un análisis descriptivo de los datos que existen actualmente y que únicamente constan de 2 ficheros con mediciones de aproximadamente 5 segundos.

Aun así, la información que contienen dichos ficheros es compleja y por lo tanto es de especial importancia realizar una fase de pre-procesado exhaustiva y detallada con varios pasos, tal y como se ha descrito en este documento. Por un lado, se ha diseñado un método para filtrar las aceleraciones recogidas por el acelerómetro, evitando así la distorsión de los resultados de los procedimientos posteriores a causa de valores atípicos o anómalos.

Tras dicho filtro, se ha procedido al desarrollo de un procedimiento para la detección de las cuatro fases que componen la marcha humana. Para ello, se ha detectado en un primer nivel el número de apoyos en cada instante del tramo temporal. Seguidamente, se analiza el lateral de movimiento del usuario en dicho momento. Mediante la combinación de ambas salidas, es posible obtener el estado de la marcha de entre los cuatro posibles de forma unívoca.

Debido a que el número de muestras medidas es reducida, se ha procedido a generar datos artificiales para poder entrenar un modelo de aprendizaje inicial que pueda entrenarse posteriormente con conjuntos de datos reales. Además, debido a que actualmente no existe un criterio para determinar el riesgo de caída a partir de los datos existentes, se han generado valores aleatorios de la clase para estos datos sintéticos.

De cara a la generación de los modelos de aprendizaje automático, se procede a generar una serie de variables que caractericen la marcha en estudio. Dichas variables están asociadas a cada uno de los ejes de las tres medidas recogidas (acelerómetro, giróscopo, magnetómetro) y a la duración de cada una de las cuatro fases.